UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
FACULDADE DE MEDICINA

Michelle Almeida da Paz

VARIABILIDADE DOS DOMÍNIOS ALPHA-3, TRANSMEMBRANA E CAUDA CITOPLASMÁTICA DE HLA-C E DETECÇÃO DE VARIANTES QUE PODEM MODIFICAR SUA FUNÇÃO

Dissertação apresentada à Faculdade de Medicina, Universidade Estadual Paulista "Júlio de Mesquita Filho", Campus de Botucatu, para obtenção do título de Mestra em Patologia.

Orientador: Prof. Dr. Erick da Cruz Castelli

BOTUCATU, 2018

Michelle Almeida da Paz

VARIABILIDADE DOS DOMÍNIOS ALPHA-3, TRANSMEMBRANA E CAUDA CITOPLASMÁTICA DE HLA-C E DETECÇÃO DE VARIANTES QUE PODEM MODIFICAR SUA FUNÇÃO

Dissertação apresentada à Faculdade de Medicina, Universidade Estadual Paulista "Júlio de Mesquita Filho", Campus de Botucatu, para obtenção do título de Mestra em Patologia.

Orientador: Prof. Dr. Erick da Cruz Castelli

BOTUCATU, 2018

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

Paz, Michelle Almeida da.
Variabilidade dos domínios Alpha-3, transmembrana e cauda citoplasmática de HLA-C e detecção de variantes que podem modificar sua função / Michelle Almeida da Paz. - Botucatu, 2018

Dissertação (mestrado) - Universidade Estadual Paulista "Júlio de Mesquita Filho", Faculdade de Medicina de Botucatu
Orientador: Erick da Cruz Castelli
Capes: 21103003

1. Antígenos HLA-C. 2. Antígenos de histocompatibilidade HLA. 3. Polimorfismo (Genética). 4. Proteínas de Membrana. 5. Imunologia celular.

Palavras-chave: HLA-C; cauda citoplasmática; domínio α3 ; domínio transmembrana; variabilidade.

Michelle Almeida da Paz

VARIABILIDADE DOS DOMÍNIOS ALPHA-3, TRANSMEMBRANA E CAUDA CITOPLASMÁTICA DE HLA-C E DETECÇÃO DE VARIANTES QUE PODEM MODIFICAR SUA FUNÇÃO

Dissertação apresentada à Faculdade de Medicina, Universidade Estadual Paulista "Júlio de Mesquita Filho", Campus de Botucatu, para obtenção do título de Mestra em Patologia.

Orientador: Prof. Dr. Erick da Cruz Castelli

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

_____
Profa. Dra. Agnes Alessandra Sekijima Takeda
Universidade Estadual Paulista (UNESP)

_____
Prof. Dr. Eduardo Antônio Donadi
Universidade de São Paulo (USP)

Botucatu, 23 de fevereiro de 2018

Agradecimentos

Primeiramente, agradeço a Deus. Agradeço ao orientador e a todos os companheiros do Laboratório de Genética Molecular e Bioinformática. À Fundação de Amparo à Pesquisa do Estado de São Paulo (Processo FAPESP nº 2016/03622-0). Ao corpo docente e funcionários do Programa de Pós-Graduação em Patologia da Faculdade de Medicina de Botucatu. A todos amigos e familiares.

RESUMO

O Complexo Principal de Histocompatibilidade (MHC) é um complexo gênico que está intimamente envolvido com a regulação do sistema imune. Esse complexo comporta o sistema de Antígenos Leucocitários Humano (HLA), cuja principal importância está relacionada com o reconhecimento do que é próprio ou não do organismo. *HLA-C* é o gene polimórfico menos variável dos genes HLA clássicos e o que tem menor expressão nos tecidos, exceto na interface materno-fetal, em que é o único gene clássico expresso. A molécula codificada por esse gene possui significante função na apresentação antigênica e regulação da atividade de células NK, o que permite uma íntima associação com situações fisiológicas, como gestação, e patológicas, como doenças infecciosas, autoimunes, inflamatórias, neoplasias e rejeições a enxertos transplantados. Sua porção gênica mais estudada é a que codifica a fenda de ligação a peptídeos antigênicos, devido sua destacada importância na apresentação de antígenos a células T citotóxicas. No entanto, outras regiões do gene, que são negligenciadas nos estudos de variabilidade, também merecem destaque por influenciarem na sinalização e modulação da citotoxicidade de células efetoras, na ancoragem e estabilidade da molécula na membrana plasmática e na internalização e reciclagem da molécula HLA-C. Desta maneira, nós exploramos a variabilidade dos segmentos que codificam α3 (éxon 4), transmembrana (éxon 5) and cauda citoplasmática (éxon 6 and éxon 7) da molécula HLA-C em uma população miscigenada do Sudeste do Brasil, a fim de entender a influência que polimorfismos presentes nessas regiões podem impactar na estrutura e função da molécula HLA-C.

Palavras chave: *HLA-C*, Sequenciamento de Nova Geração, domínio α3, domínio transmembrana, cauda citoplasmática, variabilidade

ABSTRACT

The Major Histocompatibility Complex (MHC) is a gene complex closely involved in the regulation of the immune system. This complex includes the Human Leukocyte Antigen (HLA) system, whose main role is related to the recognition of self/non-self structures of humans. *HLA-C* is the least variable polymorphic gene of classical HLA genes and has the lowest expression in tissues, except at the maternal-fetal interface, where it is the only classical HLA class I expressed gene. The molecule encoded by this gene has a significant role in the antigen presentation and regulation of NK cells activities, which allows an intimate association with physiological conditions, such as pregnancy, and pathological conditions like infectious, autoimmune, and inflammatory diseases, cancer, and transplantation rejection. The most studied *HLA-C* portion is that encoding the peptide-binding groove, due to its outstanding importance in presentation of antigens to cytotoxic T cells. However, other regions of the gene, which are neglected in the variability studies, are also important in influencing the signaling and modulation of effector cell cytotoxicity, in the anchorage and stability of the molecule on the cell surface, and in the internalization and recycling of the HLA-C molecule. Here, we explore the variability of the segments encoding the α3 (exon 4), transmembrane (exon 5) and cytoplasmic tail (exon 6 and exon 7) domains of the HLA-C molecule in an admixed population sample from Southeastern Brazil, to understand the influence that polymorphisms present in these regions may impact the structure and function of HLA-C molecule.

Keywords: HLA-C, Next Generation Sequencing, α3 domain, transmembrane domain, cytoplasmic tail, variability.

SUMMARY

Capítulo I – Revisão de Literatura

1. REVISÃO DE LITERATURA

O Complexo Principal de Histocompatibilidade (MHC, do inglês *Major Histocompatibility Complex*) é reconhecido como uma das regiões mais polimórficas do genoma humano, compreende aproximadamente 4 Mb do braço curto do cromossomo 6 (6p21.3) e contém mais de 200 loci gênicos relacionados com o controle de respostas imunológicas. O complexo é dividido didaticamente em três regiões, denominadas classe I, II e III (KLEIN; SATO, 2000).

As classes I e II compreendem os genes do sistema de Antígenos Leucocitários Humano (HLA, do inglês *Human Leucocyte Antigens*), cujos produtos gênicos estão relacionados com o reconhecimento daquilo que é próprio ou não do organismo. Os genes de classe III codificam algumas citocinas e moléculas do sistema complemento, entre outras moléculas importantes para o sistema imunitário.

O sistema HLA de classe I é dividido em genes clássicos ou Ia (genes *HLA-A*, *HLA-B* e *HLA-C*), bastante polimórficos e que codificam moléculas expressas em praticamente todas as células somáticas para o desempenho de funções de apresentação antigênica e controle de respostas imunes; e não clássicos ou Ib (genes *HLA-E*, *HLA-F* e *HLA-G*), que desempenham função de modulação de respostas imunes, são menos polimórficos, com expressão restrita a determinados tecidos e menor nível de expressão quando comparado aos clássicos, embora o nível de expressão varie dependendo do tecido (BECK et al., 1999; SHIINA; INOKO; KULSKI, 2004).

Os genes *HLA* de classe I clássicos e não clássicos codificam cinco domínios organizados em uma cadeia polipeptídica pesada α, que faz ligação com β2-microglobulina codificada por um gene localizado fora do complexo gênico do sistema HLA, mas no cromossomo 15, formando uma glicoproteína transmembrana (KLEIN; SATO, 2000) (Figura 1). Os éxons 2 e 3 desses genes *HLA* são responsáveis pela codificação dos domínios α1 e α2, onde a fenda de ligação a peptídeos está localizada. Esses peptídeos são antígenos citosólicos, que são associados a uma molécula HLA e apresentados às células T CD8+ do sistema imunitário. Essas células T CD8+ tornam-se citotóxicas (CTLs) quanto ativadas e deflagram citólise ou estimulação de apoptose da célula alvo. No entanto, as moléculas de classe I clássicas são capazes de apresentar diversos peptídeos diferentes, em parte devido a grande variabilidade na região da fenda de ligação ao peptídeo, enquanto que as moléculas não clássicas associam-se a apenas poucos peptídeos diferentes.

Para facilitar e restringir a sinalização e ativação de CTL, o co-receptor CD8 presente na superfície de linfócitos T liga-se ao domínio α3 semelhante à imunoglobulina da molécula HLA, que é codificado pelo éxon 4 dos genes *HLA* de classe I. O éxon 5 codifica o domínio transmembrana que está ancorado na membrana celular. A cauda citoplasmática é codificada pelos éxons 6 e 7 e está relacionada com a reciclagem da molécula MHC. O códon de terminação do gene está localizado nas primeiras bases do éxon 8.

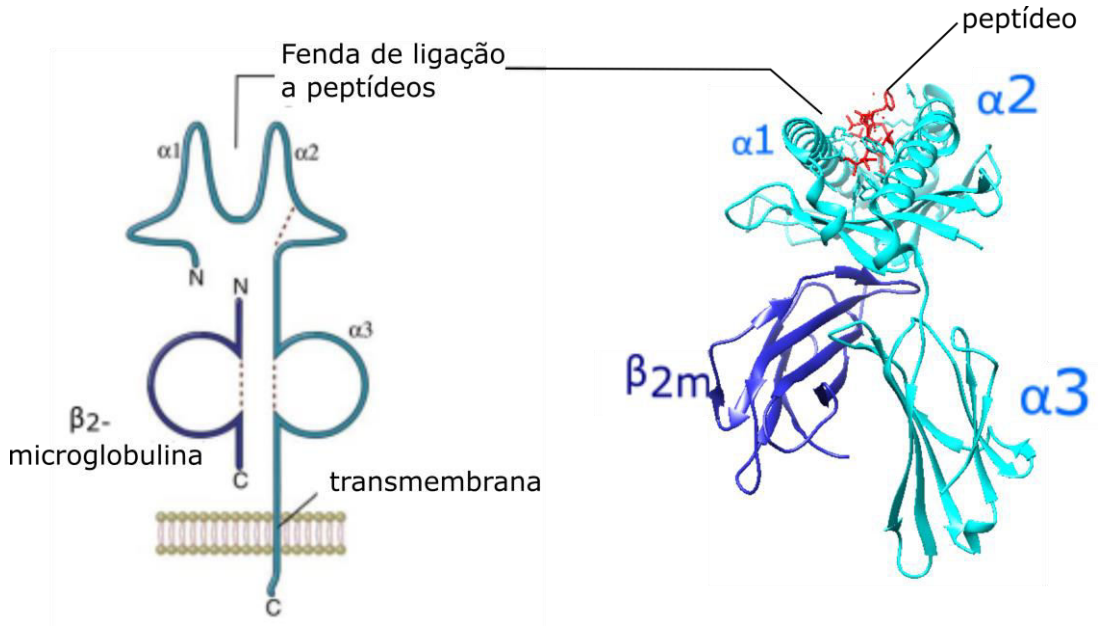


Figura 1. Na esquerda, representação esquemática da estrutura da molécula HLA de classe I indicando a cadeia pesada, seus domínios e β2-microglobulina (Adaptado de Abbas, et. al. Cellular and Molecular Immunology, Ed. 6, 2008, p. 193). Na direita, estrutura 3D da molécula de PDB ID: 4NT6 e suas porções correspondentes.

O banco de dados *International Immunogenetics Database* (IMGT/HLA), versão 3.30.0 (https://www.ebi.ac.uk/ipd/imgt/hla/), descreve grande parte da variabilidade conhecida dos genes *HLA*. Conforme descrito, são reconhecidos oficialmente 3605 alelos, que dão origem a 2497 proteínas codificadas pelo gene *HLA-C*. Esta variabilidade é reduzida quando comparada a dos demais genes clássicos (*HLA-A*: 3997 alelos e 2792 proteínas e *HLA-B*: 4859 alelos e 3518 proteínas).

O gene *HLA-C* é o menos variável e de menor expressão quando comparado ao *HLA-A* ou *HLA-B* e sua fenda de ligação a peptídeos foi extensivamente estudada, pois é uma porção extremamente variável da molécula devido sua ligação a diversos peptídeos antigênicos que são apresentados aos linfócitos TCD8+ (SANCHES-MAZAS, 2007). Em infecções por agentes intracelulares, a célula infectada é capaz de degradar as proteínas codificadas pelo genoma integrado do agente infectante e expressá-las na superfície celular como um complexo peptídeo – MHC de classe I.

Desta maneira, vem sendo descrita a importância que as moléculas de HLA-C possuem no controle das doenças infecciosas, como acontece para o vírus da imunodeficiência humana (HIV) (KULPA; COLLINS, 2011), vírus da hepatite B (HBV) e C (HCV) (MOZER-LISEWSKA et al., 2015; MORRISON et al., 2010).

A molécula HLA-C também possui relevância em doenças autoimunes, como a esclerose múltipla, já que, nessas situações, linfócitos T citotóxicos são reativos a auto antígenos apresentados por essa molécula, resultando em citotoxicidade e deflagração da doença (MIRSHAFIEY; KIANIASLALI, 2013); como também é importante no câncer, uma vez que o genoma instável de células neoplásicas produz antígenos tumorais que podem não ser reconhecidos como próprios do organismo (BLAIS; DONG; ROWLAND-JONES, 2011); bem como possui papel na resposta a aloenxertos (MOYA-QUILES et al., 2003).

O aloreconhecimento de enxertos transplantados ocorre por meio do reconhecimento de moléculas de MHC intactas do doador (direto), ou depois de serem processadas e apresentadas como peptídeos antigênicos (indireto) (MOYA-QUILES et al., 2003), ou então quando células próprias adquirem MHC de células do doador por meio de transferência célula a célula ou exossomos (aloreconhecimento semidireto) (GÖKMEN; LOMBARDI; LECHLER, 2008).

A função da molécula HLA-C não deve ser abordada apenas sob o ponto de vista de apresentação antigênica. Além desta função, a molécula desempenha importante habilidade em interagir com receptores inibitórios KIR (do inglês, *Killer cell Immunoglobulin like Receptor*) expressos por células NK (do inglês, *Natural Killer cell*), provocando a regulação dessas células (PARHAM, 2005).

Apesar da participação de receptor TCR de linfócitos T CD4+ e CD8+ ser mais relevante no processo de rejeição a células do enxerto, em uma situação de transplante, células NK podem ter papel na aloreatividade a células do enxerto, visto que as células do enxerto não apresentam ligantes MHC próprios aos receptores KIR do organismo receptor (hipótese missing-self), pode ocorrer rejeição ao enxerto (MOYA-QUILES et al., 2003). Assim, a comparação da variabilidade entre as moléculas MHC do doador e receptor do transplante mostra-se importante para a seleção de doadores em transplantes (BENTLEY et al., 2009; PETERSDORF, 2008).

A ligação de HLA-C com receptores do tipo KIR também é muito significativa em doenças inflamatórias, como a vasculite reumatoide, em que foi observado que existe suscetibilidade quando o paciente possui uma descompensação para aumento de quantidade de receptores KIR ativatórios, mesmo que haja baixa afinidade de ligação da

molécula HLA-C com receptor KIR ativatório (KULKARNI; MARTIN; CARRINGTON, 2008). Além dessas condições clínicas, o *HLA-C* possui importante papel na gestação, consistindo no único gene HLA clássico expresso na interface materno-fetal.

Durante a formação da placenta na gravidez, células trofoblásticas do feto, que expressam tanto *HLA-C* materno quanto paterno, podem interagir potencialmente com receptores KIR expressos em células NK maternas circulantes no útero. As células NK maternas podem reconhecer o HLA-C paterno como algo estranho ao organismo e podem desencadear uma resposta contra estas células, causando diversos problemas na gestação, incluindo o fenômeno de pré-eclâmpsia e parto prematuro. Em condições fisiológicas, a interação de HLA-C com receptores KIR inibitórios mostra-se essencial para a modulação da atividade de células NK e pode favorecer o sucesso da gestação, com efeitos no desenvolvimento da placenta. Provavelmente, este seja o principal motivo pelo qual o gene *HLA-C* é o único gene MHC altamente polimórfico expresso no citotrofoblasto durante a gravidez (CHAZARA; XIONG; MOFFETT, 2011).

A inibição de células NK por moléculas HLA-C não ocorre suficientemente apenas pela expressão na superfície celular dos domínios extracelulares de HLA-C. Porções transmembrana e citoplasmáticas de moléculas *HLA-C* também exercem importante função (DAVIS et al., 1999) O domínio transmembrana pode ser responsável por parte da modulação da atividade de NK, por poder controlar a mobilidade lateral da HLA-C na membrana celular (DAVIS et al., 1999).

Manter a inibição da atividade de NK pode limitar a vigilância imunitária normal. Desta forma, a baixa expressão de *HLA-C* na superfície celular da maioria dos tecidos, comparada a de outros alótipos clássicos (*HLA-A* e *HLA-B*), mostra-se relevante. Pode haver circunstâncias em que rapidamente ocorre regulação positiva de *HLA-C* a fim de aumentar a capacidade de apresentar certos tipos de antígenos ou diminuir a resposta imune inata por aumentar a sinalização para KIR inibitório. Desta forma, a manutenção de um baixo nível de expressão de *HLA-C* pode permitir um balanço entre sinais necessários para apresentação de antígenos e inibição adequada de células NK, sem que sinais dominantemente fortes aumentem excessivamente o limiar de ativação.

Mecanismos complexos existem para manter níveis significantes de *HLA-C* enquanto limita, mas não elimina, sua expressão na superfície celular. A cauda citoplasmática parece ter papel fundamental nesses mecanismos por afetar sinais de internalização da molécula e sinalização de direcionamento lisossomal, orientando as

moléculas para a degradação em organelas com enzimas, regulando o padrão de reciclagem dessas moléculas de MHC (SCHAEFER et al., 2008).

A estrutura extracelular da molécula HLA-C também é bastante relevante para o desempenho de suas funções e o domínio α3 mostra-se importante para a manutenção da resposta imune adaptativa, pois a troca de aminoácidos na região de loop onde ocorre a interação com CD8 pode causar efeito na perda ou dificuldade de interação com CD8 de linfócitos T (FAYEN et al., 1995; WESLEY et al., 1993) e reduzir, por sua vez, a sinalização e ativação de CTL.

Considerando a variabilidade descrita para a região que codifica o sítio de ligação a peptídeos, estabelece-se a hipótese que outros segmentos da molécula, i.e., domínios α3, transmembrana e citoplasmático também são variáveis e possuem variantes que podem modificar a estrutura e função da molécula HLA-C. Logo, a avaliação de variabilidade nesses domínios pode sugerir mudanças na estrutura e no comportamento da molécula em diferentes situações fisiológicas e patológicas e indicar possíveis novos alvos terapêuticos para perfis proteicos diferentes.

2. JUSTIFICATIVA

O *HLA-C* é o único gene associado com apresentação antigênica que possui expressão na interface materno-fetal. É o menos variável e de menor expressão nos demais tecidos quando comparado aos outros genes clássicos (*HLA-A* e *HLA-B*). A molécula HLA-C possui duplo papel no organismo: ao promover apresentação de peptídeos antigênicos a linfócitos T citotóxicos, está participando da resposta imune adaptativa e, ao ligar-se com receptores KIR desempenhando funções tolerogênicas, está colaborando com a resposta imune inata.

A variabilidade do gene *HLA-C* é muito estudada para éxon 2 e éxon 3, responsáveis pela codificação dos domínios pesados α1 e α2 respectivamente e, consequentemente, pelo domínio de ligação ao peptídeo. A variabilidade desses éxons, que é conhecida e anotada nos bancos de dados públicos, é a principal base para *softwares* de genotipagem que detectam variantes em *HLA* e permitem a seleção adequada de doadores em transplantes. A exploração destes éxons é justificada pela importância que a fenda de ligação a peptídeos antigênicos possui na compatibilização com receptores TCR de linfócitos T que sofreram processo de seleção tímica para o reconhecimento de antígenos no contexto das moléculas MHC do indivíduo. Variantes alélicas nestas regiões do gene podem fazer com que a molécula não acomode

adequadamente peptídeos não reconhecidos como próprios do organismo e isto pode influenciar na suscetibilidade a diversas doenças.

As funções desempenhadas pela molécula HLA-C permitem que haplótipos específicos sejam associados com suscetibilidade ou resistência a doenças infecciosas, autoimunes, inflamatórias, neoplasias, transplantes e à gestação (APANIUS et al., 1997). A variabilidade inerente das regiões que codificam outros domínios da molécula HLA-C, como o domínio α3, transmembrana e citoplasmático, não foi adequadamente explorada. Variantes nestas regiões poderiam influenciar diretamente na atuação das moléculas HLA-C de diferentes formas.

O mecanismo de interação do domínio α3, codificado pelo éxon 4, com o co-receptor CD8 presente em linfócitos T citotóxicos não está muito bem esclarecido, embora se conheça a importância para a sinalização e ativação de linfócitos T CD8+. A perda ou dificuldade de interação entre domínio α3 e o CD8 pode reduzir a citotoxicidade, levando a uma queda da resposta imune adaptativa. Variações no domínio α3 poderiam influenciar diretamente na interação deste domínio com o co-receptor CD8.

Alterações nas sequências do éxon 5 do gene *HLA-C*, que codifica o domínio transmembrana, podem modificar a conformação da molécula e a sua ancoragem/mobilidade na membrana, afetando a sua estabilidade na superfície celular. A perda de fixação do HLA-C na superfície celular pode reduzir respostas imunes adaptativas e inatas, comprometendo grande parte do sistema imune do organismo.

A regulação do padrão de reciclagem e degradação de moléculas HLA-C é afetada por sinais da cauda citoplasmática, codificada pelo éxon 6 e éxon 7. A variabilidade desses éxons pode influenciar no nível de sinais de regulação positiva ou negativa destas moléculas na superfície celular e, dependendo da situação patológica, pode inibir excessivamente a atividade de células NK ou reduzir demais a apresentação de peptídeos antigênicos a linfócitos T citotóxicos importantes para resolução de doenças.

Assim, torna-se necessário um estudo detalhado da variabilidade de outros domínios que compõem a cadeia polipeptídica pesada α codificada pelo gene *HLA-C* (domínio α3, transmembrana e cauda citoplasmática), além dos domínios peptídeo-ligante α1 e α2, para o entendimento integrado destas porções na estrutura da molécula HLA-C e função imunológica.

3. OBJETIVOS

O objetivo geral é avaliar a variabilidade genética dos segmentos que codificam os domínios α3 (éxon 4), transmembrana (éxon 5) e cauda citoplasmática (éxons 6 e 7) da molécula HLA-C, por meio de sequenciamento de nova geração, em uma amostra de população brasileira do estado de São Paulo, e verificar a diversidade proteica da molécula HLA-C codificada.

A partir do objetivo geral, foram propostos os objetivos específicos que se fundamentam em verificar variações que alteram o perfil de aminoácidos que compõem a proteína; modelagem e análise integrada da estrutura proteica da molécula HLA-C e determinação de influências que polimorfismos nessas regiões estudadas podem causar na estrutura e função da molécula HLA-C.

4. REFERÊNCIAS BIBLIOGRÁFICAS

APANIUS, V. et al. The Nature of Selection on the Major Histocompatibility Complex. **Critical Reviews' in Immunology**, v. 17, p. 179–224, 1997.

BECK, S. et al. Complete sequence and gene map of a human major histocompatibility complex. **Nature**, v. 401, p. 921–923, 1999.

BENTLEY, G. et al. High-resolution, high-throughput HLA genotyping by next-generation sequencing. **Tissue Antigens**, 2009.

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.

CHAZARA, O.; XIONG, S.; MOFFETT, A. Maternal KIR and fetal HLA-C: a fine balance. **Journal of Leukocyte Biology**, p. 703–716, 2011.

DAVIS, D. M. et al. The Transmembrane Sequence of Human Histocompatibility Leukocyte Antigen (HLA)-C as a Determinant in Inhibition of a Subset of Natural Killer Cells. **J. Exp. Med**, v. 189, n. 8, p. 1265–1274, 1999.

FAYEN, J. et al. Class I MHC alpha 3 domain can function as an independent structural unit to bind CD8?? **Molecular Immunology**, v. 32, n. 4, p. 267–275, 1995.

GÖKMEN, M. R.; LOMBARDI, G.; LECHLER, R. I. The importance of the indirect pathway of allorecognition in clinical transplantation. **Current Opinion in Immunology**, 2008.

KLEIN, J.; SATO, A. The HLA system First of Two Parts. **The New England Jounal of Medicine**, p. 702–709, 2000.

KULKARNI, S.; MARTIN, M. P.; CARRINGTON, M. **The Yin and Yang of HLA and KIR in human diseaseSeminars in Immunology**, 2008.

KULPA, D. A.; COLLINS, K. L. **The emerging role of HLA-C in HIV-1 infectionImmunology**, 2011.

MIRSHAFIEY, A.; KIANIASLANI, M. Autoantigens and autoantibodies in multiple sclerosis. **Iranian Journal of Allergy, Asthma and Immunology**, v. 12, n. 4, p. 292–303, 2013.

MORRISON, B. A. et al. Multiple sclerosis risk markers in HLA-DRA, HLA-C, and IFNG genes are associated with sex-specific childhood leukemia risk. **Autoimmunity**, v. 43, n. 8, p. 690–7, 2010.

MOYA-QUILES, M. R. et al. Human Leukocyte Antigen–C in Short-and Long-Term Liver Graft Acceptance. **Liver Transplantation**, v. 9, p. 218–227, 2003.

MOZER-LISEWSKA, I. et al. Genetic (KIR, HLA-C) and Some Clinical Parameters Influencing the Level of Liver Enzymes and Early Virologic Response in Patients with Chronic Hepatitis C. **Archivum Immunologiae et Therapiae Experimentalis**, v. 64, n. 1, p. 65–73, 2015.

PARHAM, P. MHC class I molecules and kirs in human history, health and survival. **Nature Reviews Immunology**, p. 201–214, 2005.

PETERSDORF, E. W. **Optimal HLA matching in hematopoietic cell transplantationCurrent Opinion in Immunology**, 2008.

SANCHEZ-MAZAS, A. **An apportionment of human HLA diversity**. Tissue Antigens, 2007

SCHAEFER, M. R. et al. A novel trafficking signal within the HLA-C cytoplasmic tail allows regulated expression upon differentiation of macrophages. **Journal of immunology (Baltimore, Md. : 1950)**, v. 180, n. 12, p. 7804–17, 2008.

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.

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.

Capítulo II – Artigo

5. ARTICLE

HLA-C ALPHA-3, TRANSMEMBRANE AND CITOPLASMIC TAIL DOMAINS VARIABILITY AND DETECTION OF FUNCTIONAL VARIANTS

Michelle Almeida da Paz[1,2], Andréia da Silva Souza[1,3], Erick C. Castelli[1,2,3]

[1]Molecular Genetics and Bioinformatics Laboratory, Experimental Research Unit (UNIPEX), Medical School of Botucatu, São Paulo State University – UNESP, Brazil.

[2]Graduate Program on Pathology, Medical School of Botucatu, São Paulo State University – UNESP, Brazil.

[3]Graduate Program on Biological Sciences (Genetics), Bioscience's Institute of Botucatu, São Paulo State University – UNESP, Brazil.

Acknowledgements
This work was supported by São Paulo Research Foundation (FAPESP) (Grant#2013/17084-2) and M.A.P is supported by FAPESP (Grant#2016/03622-0).

Contact
Michelle Almeida da Paz
Departamento de Patologia, Faculdade de Medicina de Botucatu, UNESP, Botucatu, SP
CEP: 18618970, Brazil
Phone: +55 14 3880-1696
E-mail address: michelleapaz@outlook.com

6. ABSTRACT

The Human Leucocyte Antigen-C (*HLA-C*) gene belongs to the Major Histocompatibility Complex (MHC), and along with its classical counterparts (*HLA-A* and *HLA-B*) comprises the most polymorphic genes of the human genome. Besides playing a pivotal role in endogenous antigen presentation to T cells, the molecule encoded HLA-C has been described as an important immune modulatory molecule, influencing cancer, inflammatory, autoimmune and infectious diseases outcomes, as well as the outcome of grafted tissues and pregnancy. The most studied HLA-C portion is the peptide-binding groove, due to its outstanding importance in presentation of antigens to cytotoxic T cells. However, other regions of the molecule, which are neglected in the variability studies, are also important in influencing the signaling and modulation of effector cell cytotoxicity, in the anchorage and stability of the molecule on the cell surface, and in the internalization and recycling of the HLA-C molecule. Here, we aimed to explore the variability of the segments encoding the α3 (exon 4), transmembrane (exon 5) and cytoplasmic tail (exon 6 and exon 7) domains of the HLA-C molecule in 400 individuals of an admixed population sample from Southeastern Brazil, by using Next Generation Sequencing approach. We identified 37 different full-length proteins presenting high frequencies. The studied domains have a certain level of conservation and the residue variations found do not alter the protein secondary structure, usually. Variable residues presented in critical regions of receptor interaction are in low frequency or they are rare. In conclusion, the minor impact on the function of the HLA-C molecule is important to maintain its immunological functions.

Keywords: HLA-C, Next Generation Sequencing, α3 domain, transmembrane domain, cytoplasmic tail, variability.

7. INTRODUCTION

The Human Leucocyte Antigen C (*HLA-C*) gene is a classical Human Leucocyte Antigen (HLA) class I locus encoded at the short arm of chromosome 6, within the human Major Histocompatibility Complex (MHC). According to the International Immunogenetics Database (IPD-IMGT/HLA), version 3.30.0, the *HLA-C* locus presents 3,605 different coding allele sequences encoding 2,497 full-length protein molecules, which together with its classical counterparts (*HLA-A* and *HLA-B*) configure the most polymorphic genes of the human genome (KLEIN; SATO, 2000). However, the *HLA-C* gene variability is reduced in the group (*HLA-A*: 3,997 alleles and 2,792 proteins, and *HLA-B*: 4,859 alleles and 3,518 protein molecules), but these results might be biased since *HLA-A* and *HLA-B* polymorphisms are usually explored when it comes to transplantation procedures.

The structure of HLA class I genes is essentially composed of 8 exons, each of which encodes practically only one domain of the heavy chain that is anchored in the cell membrane. Exon 1 encodes the leader peptide. Exon 2 and 3 encode the α1 and α2 domains, respectively, and thus the peptide binding groove. Exon 4 encodes the α3 immunoglobulin-like domain. Exon 5 encodes the transmembrane domain. Exon 6 and exon 7 encode together the cytoplasmic tail domain, while the stop codon is located at the first bases of exon 8.

Besides, *HLA-C* is the least polymorphic of the classical HLA class I genes, it presents lower expression levels on the cell surface of nucleated cells when compared to *HLA-A* and *HLA-B* (APPS et al., 2015). In addition, the *HLA-C* is the only classical gene expressed in the maternal-fetal interface (CHAZARA; XIONG; MOFFETT, 2011), along with the non-classical *HLA-E* and *HLA-G* genes performing a pivotal role in pregnancy success (CHAZARA; XIONG; MOFFETT, 2011; VARLA-LEFTHERIOTI, 2004). The maintenance of a lower level of HLA-C molecule at the cell surface appears to be fundamental for the balance between signals necessary to maintain the performance of its two distinct functions. The dual immunological role of HLA-C is based on the its ability to bind and present a range of intracellular peptides to cytotoxic $CD8^+$ T cells to trigger an adaptive immune response, as well as regulate innate immune responses by interacting with killer cell immunoglobulin-like receptors (KIR) expressed on natural killer (NK) cells (PARHAM, 2005).

Some *HLA-C* specific haplotypes have been associated to resistance or susceptibility to the HIV-1 virus (KULPA; COLLINS, 2011) and other infectious

diseases (MOZER-LISEWSKA et al., 2016), autoimmune diseases, such as type 1 diabetes (ZHI et al., 2014) and multiple sclerosis (MORRISON et al., 2010), inflammatory disorders like Crohn's disease (APPS et al., 2013), and cancer (BLAIS; DONG; ROWLAND-JONES, 2011). Also, HLA-C diversity is an important determinant in influencing clinical outcomes of unrelated hematopoietic cell transplantation (PETERSDORF, 2008) and pregnancy success (CHAZARA; XIONG; MOFFETT, 2011).

Whereas the variability of the segment encoding the peptide binding groove (exon 2 and exon 3) are significant for the molecule to accommodate a range of different peptides, the amino acid sequence of this molecule portion may maintain a pattern to not influence the recognition of antigens by T cells which suffered the Thymic selection process and also the NK cell interaction through KIR receptors. Therefore, the variability in the region has already been extensively studied (SANCHEZ-MAZAS, 2007). However, the inherent variability in regions that encode other domains of the molecule HLA-C should be explored, since variations in other segments may impact the structure and function of the molecule in the context of physiological and pathological conditions.

For instance, exon 1 variability may change the leader peptide constitution and impact on the export of the HLA-C molecule to the cell surface, and also the stabilization of other HLA class I molecules such as HLA-E (DI MARCO et al., 2017). Taking in account that the α3 domain, encoded by exon 4, has a significant role in facilitating the signaling and activation of cytotoxic T lymphocytes by interaction with the CD8 co-receptor found on their surface (WESLEY et al., 1993), non-synonymous substitutions in the loop region of the α3 domain may hinder or avoid the α3/CD8 interaction. The main role of the transmembrane domain, encoded by exon 5 of the *HLA-C* gene, is to anchor the HLA-C molecule on the cell surface (DAVIS et al., 1999). Alterations in this segment may impair the appropriate anchorage and flexibility of the molecule, affecting its stability on the cell surface and, thus, reducing immunological responses. The cytoplasmic tail domain, encoded by exon 6 and exon 7, appears to play a key role in regulating the pattern of recycling and degradation of HLA-C molecules in order to maintain significant levels of expressed HLA-C, while limiting, but does not eliminate its expression on the cell surface (SCHAEFER et al., 2008). Therefore, variations in the conserved regions of the cytoplasmic domain may affect the signals for internalization and lysosomal targeting of the HLA-C molecule for its degradation.

Accordingly, it is imperative to evaluate the variability of the other domains composing the heavy chain of the HLA-C molecule (α3, transmembrane and cytoplasmic tail domains), in addition to the peptide-binding domains formed by α1 and α2 domains, in order to obtain an integrated understanding of those portions in the structure of the HLA-C molecule. Here, we evaluate the variability of the segments encoding the α3 (exon 4), transmembrane (exon 5) and cytoplasmic tail (exon 6 and exon 7) domains, together with the variability of exon 1 which encodes the leader peptide and exons 2 and 3 that encode, respectively, the α1 and α2 domains for the HLA-C molecule, by using a massively parallel sequencing and bioinformatic approach, in an admixed Brazilian population sample from the State of São Paulo. For the purpose of verify the variations which alter the profile of amino acids that compose the HLA-C protein, correlate the variants found and determine the influences that polymorphisms may cause on the structure and function of the HLA-C molecule.

8. MATERIAL AND METHODS

This study was reviewed and approved in its ethical aspects by the Human Research Ethics Committee of this Institution (School of Medicine, UNESP/Brazil), according to Protocol #24157413.7.0000.5411.

8.1. DNA samples and amplification

We collected peripheral blood of 400 unrelated individuals from the State of São Paulo, Southeastern Brazil. Prior to blood withdrawal, all participants signed an informed consent. DNA samples analyzed were obtained from white blood cells by salting out procedure. Each sample was quantified using Qubit dsDNA Broad Range Assays (Thermo Fisher Scientific Inc., Waltham, MA, USA) and normalized to 50 ng/μL using ultra-pure water.

The *HLA-C* gene segment was amplified in a unique amplicon of approximately 5,712 nucleotides, encompassing its promoter segment, the complete coding sequence with all introns, and the complete 3'UTR. Polymerase Chain Reaction (PCR) technique was performed using 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 recommended by DNA polymerase were carried out with an initial denaturation at 98º C for 10 s, annealing at 60º C for 15 s and extension at 68º C for 6 min, repeated for 30 cycles, and finalizing with 4º C. Each amplicon was confirmed by fragment length of about 6 Kb on 1% agarose gel stained with GelRed® (Biotium, Inc. Hayward, CA, USA), quantified using Qubit dsDNA High-Sensitivity (HS) Assays (Thermo Fisher Scientific Inc., Waltham, MA, USA) and normalized to 3 ng/µL using ultra-pure water.

Although only *HLA-C* variability is presented, other HLA class I genes were sequenced together. Amplicons for the same sample were pooled together and purified by Illustra ExoProStar (GE Healthcare Life Sciences). The amplicon pool was quantified using Qubit dsDNA HS Assays. Lastly, the pool was normalized to 0.2 ng/µL using ultra-pure water, which is the recommended concentration for sequencing library preparation using Nextera XT kits (Illumina, Inc., San Diego, CA, USA).

8.2. Library preparation and sequencing

Sequencing libraries were prepared using Nextera XT Sample Preparation Kit multiplexed with Nextera DNA Preparation Index Kit (both from Illumina, Inc., San Diego, CA, USA). Library quantification was achieved by qPCR using Kapa (Kapa Biosystems, Wilmington, USA) and the fragmentation pattern was estimated using High-Sensitivity DNA BioAnalyzer chips (Agilent Technologies, CA, USA). Then, libraries were normalized to 4 nM and sequenced using the MiSeq system (V2 500 cycles, 2×250 pb – Illumina, Inc.).

8.3. Raw data processing (mapping)

All sequences contained in the fastq files (reads) were trimmed on both ends for primer and adapter sequences. Furthermore, reads were evaluated by seqtk trimfq (https://github.com/lh3/seqtk) to remove low-quality bases from both ends, with the error rate threshold set to 0.02.

The main goal when performing HLA sequencing by NGS procedures is to get a reliable mapping of the produced reads against a reference genome. The low mapping accuracy is due to the highly polymorphic nature of HLA genes and the sequence similarity among these paralogous genes (BRANDT et al., 2015). To circumvent mapping issues, we used *hla-mapper* function *dna* (and database version 2.0), freely

available at www.castelli-lab.net/apps/hla-mapper. This strategy was crucial to obtain a reliable read mapping, which consisted in assign each pair of reads to the HLA gene they were derived from, based on sequence similarities with known HLA sequences described at the IPD-IMGT/HLA database (ROBINSON et al., 2015) and known HLA haplotypes. The hla-mapper software produces BAM files for each HLA class I gene, containing sequences aligned to the reference genome (hg19 or hg38).

8.4. Genotype calling and processing

We used the Genome Analysis Toolkit (GATK, version 3.7) HaplotypeCaller routine GVCF mode with the minimum base quality score set to 20 and not using soft-clipped bases (MCKENNA et al., 2010) to infer genotypes from each *HLA-C*-specific BAM file. GVCF files were joined into a multi-sample VCF file using GATK GenotypeGVCFs algorithm, considering the chromosome 6 sequence of the human genome draft version hg19 as reference and variant annotation based on the dbSNP database version 146.

Genotypes inferred at low coverage segments would be prone to errors. To certify the presence of only high-quality genotypes, the original VCF file was treated by vcfx using the subroutine checkpl, available at www.castelli-lab.net/apps/vcfx. This application introduced missing alleles on genotypes with likelihood lower than 99.9%, always keeping the most likely allele. Then, all monomorphic sites eventually detected after the vcfx treatment were removed and the VCF file was manually curated to remove low quality genotypes.

8.5. Haplotype inference

The associations among variable sites that occur on the same read were determined by the GATK routine ReadBackedPhasing (MCKENNA et al., 2010), using a minimal phase quality threshold of 2,000 (100x higher than the default) and minimum base quality set to 20. Nonetheless, the ReadBackedPhasing does not perform phase inference on indels, multi-allelic loci, and also at genotypes presenting missing alleles, and neither among distant variable sites. Therefore, the PHASE algorithm (STEPHENS; SMITH; DONNELLY, 2001) was used to infer the phase of the 18.13% of the heterozygous sites that were not directly phased by GATK. In addition, the PHASE algorithm was used to impute the remaining missing alleles.

To continue with haplotype inference by PHASE, all variable sites detected in only one individual and in a heterozygous state (singletons) were removed from the GATK-phased VCF file. The resulted file was converted into an input file for PHASE. The known phase information obtained by ReadBackedPhasing was organized into an accessory PHASE file and the haplotypes were inferred for each sample, fixing known haplotype blocks in each run. The haplotype pair selected for each sample was used to construct a final VCF file with all variable sites phased. The scripts used to convert VCF to the PHASE input and to create the accessory file from ReadBackedPhasing is available at https://github.com/erickcastelli/phase-readbackedphasing. Whenever possible, singletons were manually introduced in the final VCF file by evaluating the BAM file. This phased VCF file was used to create *HLA-C* complete sequences and CDS sequences using vcfx function fasta (www.castelli-lab.net/apps/vcfx), and the sequences were reversed and complemented by using emboss revseq tool (RICE; LONGDEN; BLEASBY, 2000).

8.6. Naming process of *HLA-C* coding alleles and encoded proteins

We used local BLAST server with a database containing all known *HLA-C* alleles downloaded from IPD-IMGT/HLA database, version 3.30.0, to define the closest or identical known *HLA-C* coding allele for each sequence. Each *HLA-C* allele was named according to the WHO Nomenclature Committee for Factors of the HLA System (MARSH et al., 2010) and divergences eventually detected were listed after the name of the closest known allele.

The current nomenclature defines that the digits before the first colon are related to the serological type. The second set of digits distinguishes alleles that differ in one or more nucleotide substitution that change the amino acid sequence of the encoded protein. The use of the third set of digits discriminate alleles that differ only in synonymous nucleotide substitutions at exons segments, and the fourth set of digits can distinguish alleles that only differ by sequence polymorphism in the introns and regulatory segments.

The encoded protein sequences were obtained by the corresponding coding allele sequences, which were translated by using emboss transeq tool (RICE; LONGDEN; BLEASBY, 2000). The Protein BLAST was used to search the closest or identical known HLA-C protein deposited in the IPD-IMGT/HLA, version 3.30.0. And the direct counting method calculated the frequency of each HLA-C protein sequence.

8.7. Template selection and molecular model construction

The prediction of the HLA-C molecule structure was performed for the most frequent protein sequence found in the Brazilian sample population – in this case, HLA-C*04:01 (Table 1). Physical-chemical parameters were computed by using ProtParam (GASTEIGER et al., 2005), which classifies the protein as stable with molecular weights (MV) and theoretical isoelectric point (pI) values 41.00 kDa and 6.04, respectively. The homology detection and selection template was achieved by the HHpred (SÖDING; BIEGERT; LUPAS, 2005), considering the search in a database built using the Uniclust30 database (MIRDITA et al., 2017). The crystallographic structure deposited in the Protein Data Bank (PDB) database (BERMAN et al., 2000) of the *Homo sapiens* HLA-C*08:01 obtained by X-ray diffraction at 1.84 Å resolution (PDB ID: 4NT6, Chain A) (CHOO et al., 2014) was chosen as a template. The identity of 94.87% between the primary structure of the chosen template and the selected sequence of HLA-C*04:01 from the analyzed Brazilian sample was confirmed by using the SWISS-MODEL (BIASINI et al., 2014).

The output of HHpred was forwarded to MODELLER 9.19 (WEBB; SALI, 2017) to construct the three-dimensional structure model based on homology detection method. A hundred models were generated and the most suitable model was selected based on the stereochemical parameters using the program RAMPAGE (LOVELL et al., 2003), with 98.5% of the residues in favored regions and 1.5% (corresponding to four residues) in allowed region and no residue found in the outlier region (Figure S2).

8.8. Evaluation of variable regions, electrostatic potential and secondary structure prediction

The evolutionary conservation analysis of the amino-acid positions and the detection of variable amino-acid residues were performed by using the ConSurfServer (ASHKENAZY et al., 2010), and all figures were generated by using the UCSF Chimera program (PETTERSEN et al., 2004).

The electrostatic potential was computed using the interface Adaptive Poisson-Boltzmann Solver (APBS) of UCSF Chimera program including ion charge to the system in a concentration similar to physiological. Positive ion charge of +1 electron unit, 0.15 M of concentration and 1.47 Å of radius. Negative ion charge of -1 electron unit, 0.15 M of concentration and 1.36 Å of radius.

The protein secondary structure was predict to leader peptide, transmembrane and cytoplasmic tail using Quick2D (ZIMMERMANN et al., 2017), since these portions have no defined tridimensional structure to be compared by homology.

8.9. Other analysis

To evaluate the linkage disequilibrium (LD) pattern, the D', LOD scores were calculated and LD plots were visualized using Haploview 4.2 (BARRETT et al., 2005), considering variable sites with a minor allele frequency (MAF) of 1% and the fraction of strong LD in informative comparison set to 0.9. We used the SnpEff software (CINGOLANI et al., 2012) to classify possible effects of variable sites. The phylogenetic tree was created using the Maximum Likelihood method and visualized by Mega version 7 (KUMAR et al., 2008).

9. RESULTS

9.1. Relationship between the *HLA-C* CDS variants and encoded HLA-C proteins

A strong LD across the *HLA-C* locus was detected, as observed by the formation of a unique segregation block in the segment covered by the IPD-IMGT/HLA database version 3.30.0, i.e., from position -283 (6: 31,240,128) to +3066 (6: 31,236,775), using hg19 as genome reference (Figure S1), but excluding all insertion/deletion (indel) and multiallelic loci which can not be supported by Haploview.

Here, only the data for all HLA-C exons will be presented, although introns were included in the original sequencing. We detected 110 variable sites (Table S1) arranged in 41 different haplotypes or coding alleles, of which 39 alleles are identical to those already described by the IPD-IMGT/HLA database, while two might be considered new sequences (HLA-C*03:04:01$^{(cds1032T)}$ and C*12:02:02$^{(cds1023G)}$), both occurring only once in the considered sample (Table 1).

Considering that 67.27% of the variable sites here detected configured non-synonymous mutations, based on the prediction by SnpEff using the transcript NM_001243042.1.7, the 41 different CDS sequences detected in our population sample encode 37 different HLA-C full-length protein molecules, including a null allele (HLA-C*04:09N) with frequency of 0.0025 (Table 1).

***HLA-C* coding alleles** [a]	**Coding allele Frequency (2*n*=800)**	**Encoded HLA-C protein** [b]	**Protein Frequency (2*n*=800)**
C*01:02:01	0.02625	C*01:02	0.02625
C*02:02:02	0.04000	C*02:02	0.04000
C*02:10:01	0.01625	C*02:10	0.01625
C*02:14:02	0.00250	C*02:14[compatible]	0.00250
C*03:02:02	0.00750	C*03:02	0.00750
C*03:03:01	0.04375	C*03:03	0.04375
C*03:04:01 C*03:04:01[(cds1032T)] C*03:04:02	0.03000 0.00125 0.01000	C*03:04	0.04125
C*03:309	0.00125	C*03:309	0.00125
C*04:01:01	0.16875	C*04:01	0.16875
C*04:07	0.00125	C*04:07	0.00125
C*04:09N	0.00250	C*04:09N	0.00250
C*05:01:01	0.04250	C*05:01	0.04250
C*06:02:01	0.09250	C*06:02	0.09250
C*07:01:01 C*07:01:02	0.08750 0.01500	C*07:01	0.10250
C*07:02:01	0.08500	C*07:02	0.08500
C*07:04:01	0.01125	C*07:04	0.01125
C*07:18	0.01750	C*07:18	0.01750
C*08:01:01	0.00125	C*08:01	0.00125
C*08:02:01	0.05375	C*08:02	0.05375
C*08:03:01	0.00125	C*08:03	0.00125
C*08:04:01	0.00125	C*08:04	0.00125
C*12:02:02[(cds1023G)] C*12:02:02	0.00125 0.00750	C*12:02	0.00875
C*12:03:01	0.05625	C*12:03	0.05625
C*14:02:01	0.03125	C*14:02	0.03125
C*14:03	0.00375	C*14:03	0.00375
C*15:02:01	0.02875	C*15:02	0.02875
C*15:08	0.00125	C*15:08[compatible]	0.00125
C*15:09	0.00125	C*15:09	0.00125
C*15:13	0.00375	C*15:13	0.00375
C*16:01:01	0.04250	C*16:01	0.04250
C*16:02:01	0.01125	C*16:02	0.01125
C*16:04:01	0.00500	C*16:04	0.00500
C*17:01:01	0.02500	C*17:01	0.02500
C*17:03:01	0.00500	C*17:03	0.00500
C*17:38	0.00125	C*17:38	0.00125
C*18:01	0.00875	C*18:01	0.00875
C*18:02	0.00625	C*18:02	0.00625

Table 1. List of *HLA-C* coding alleles and the corresponding encoded protein detected at Southeastern Brazil population sample.

[a] Coding sequences according to the closest known *HLA-C* allele at the IPD-IMGT/HLA database version 3.30.0, followed by the divergences observed for the given haplotype, considering all the nucleotide sequences of concatenated exons (no introns included). The word "compatible" indicates sequences found in our population sample which are only partially characterized on IPD-IMGT/HLA database.

[b] The translated full-length HLA-C proteins.

The phylogenetic tree was inferred using the HLA-C protein sequences detected in the Brazilian population sample toward evaluate the association among them (Figure

2). It can be noticed that there are many protein molecules that are divergent mostly at the peptide-binding groove. For instance, all the C*03 and C*14 molecules here detected present the same leader peptide, α3, transmembrane and cytoplasmic tail, but divergent peptide-binding grooves demonstrated by the distance between the C*03 and C*14 group. This same pattern can be observed between C*08 and C*05 molecules, or even between the C*06 and C*12 groups. It can also be observed that distant protein molecules, such as C*15:02 and C*08:02, share some segments (in this particular case, the same α3 and cytoplasmic tail), which might be a consequence of gene conversion giving rise to divergent *HLA-C* sequences. Nevertheless, there are many highly divergent HLA-C protein molecules presenting high frequencies.

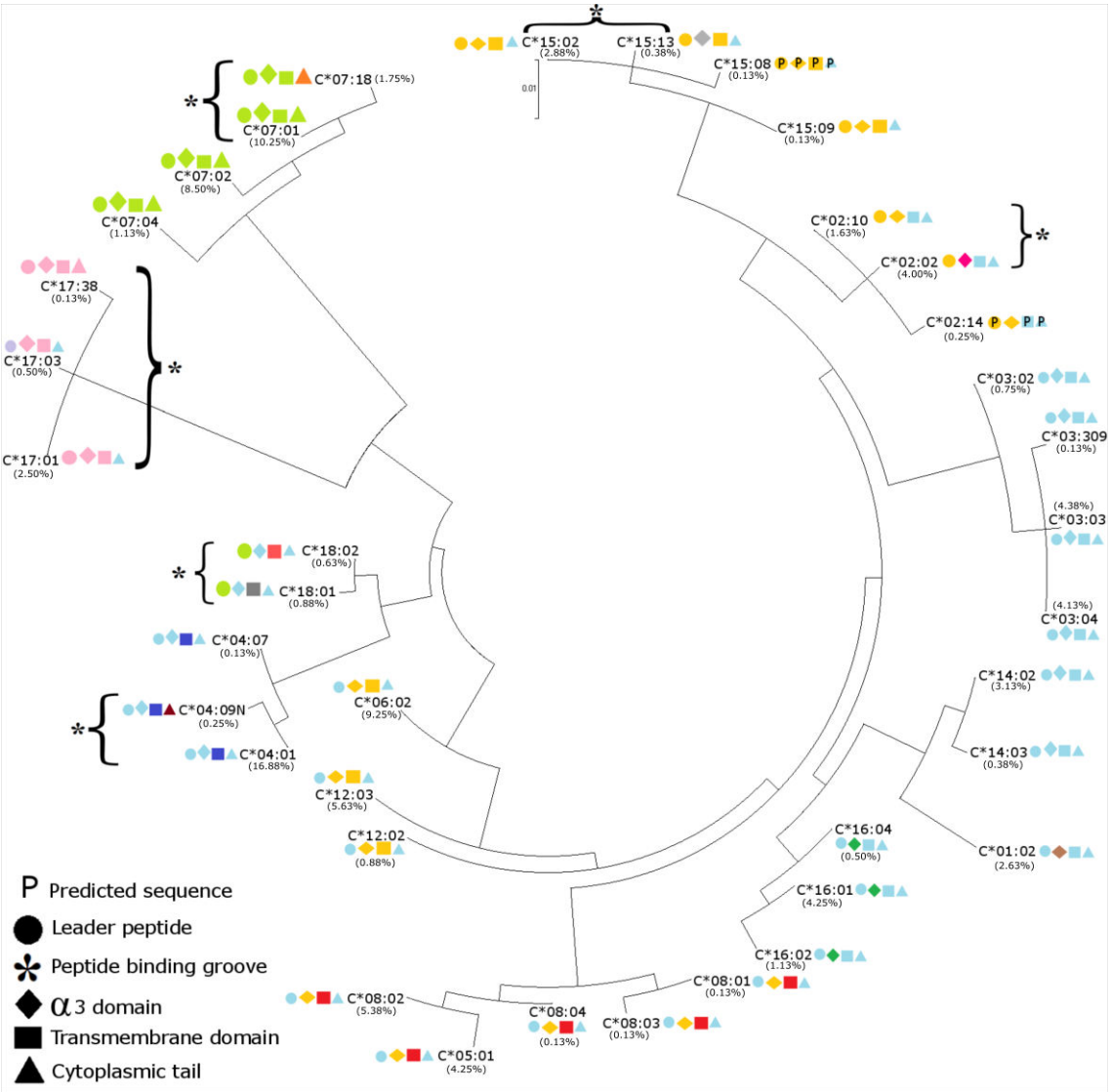


Figure 2. Phylogenetic tree of 37 HLA-C protein sequences detected in a Brazilian population sample from the State of São Paulo. The tree is drawn to scale, with branch length measured in the number of substitutions per site. All positions containing gaps were eliminated. HLA-C molecules presenting a geometric figure of the same color share the same sequence of the segment identified in the legend. All HLA-C proteins have a different peptide binding groove, except for those represented by asterisks. The term "predicted sequence" indicates protein sequences found in our population sample, but they are partially characterized on the IPD-IMGT/HLA database.

9.2. Variability in the HLA-C leader peptide

We identified 5 different protein sequences for the HLA-C leader peptide. The amino acid sequences of the two most frequent leader peptides (I and III – Figure 3A) represent 87.5% of all sequences detected. The most frequent one (64.37%) encompasses the VMAP**RT**L**I**L motif and it is shared among the HLA-C*01, -C*03, -C*04, -C*05, -C*06, -C*08, -C*12, -C*14 and -C*16 encoded protein found (I – Figure 3B). The second most frequent (23.12%) containing the VMAP**RA**L**L**L motif and it is derived from the HLA-C*07 and -C*18 (III – Figure 3B), followed by VMAP**RT**L**L**L (9.38%) derived from HLA-C*02 and -C*15 (II – Figure 3B), and VMAP**QA**L**L**L (3.13%) derived from two different sequences of HLA-C*17 (IV and V – Figure 3B).

Apparently, the residue variations found do not alter the protein secondary structure, which is maintained in alpha-helix, except for sequence III that was predicted containing beta-strands (III – Figure 3B).

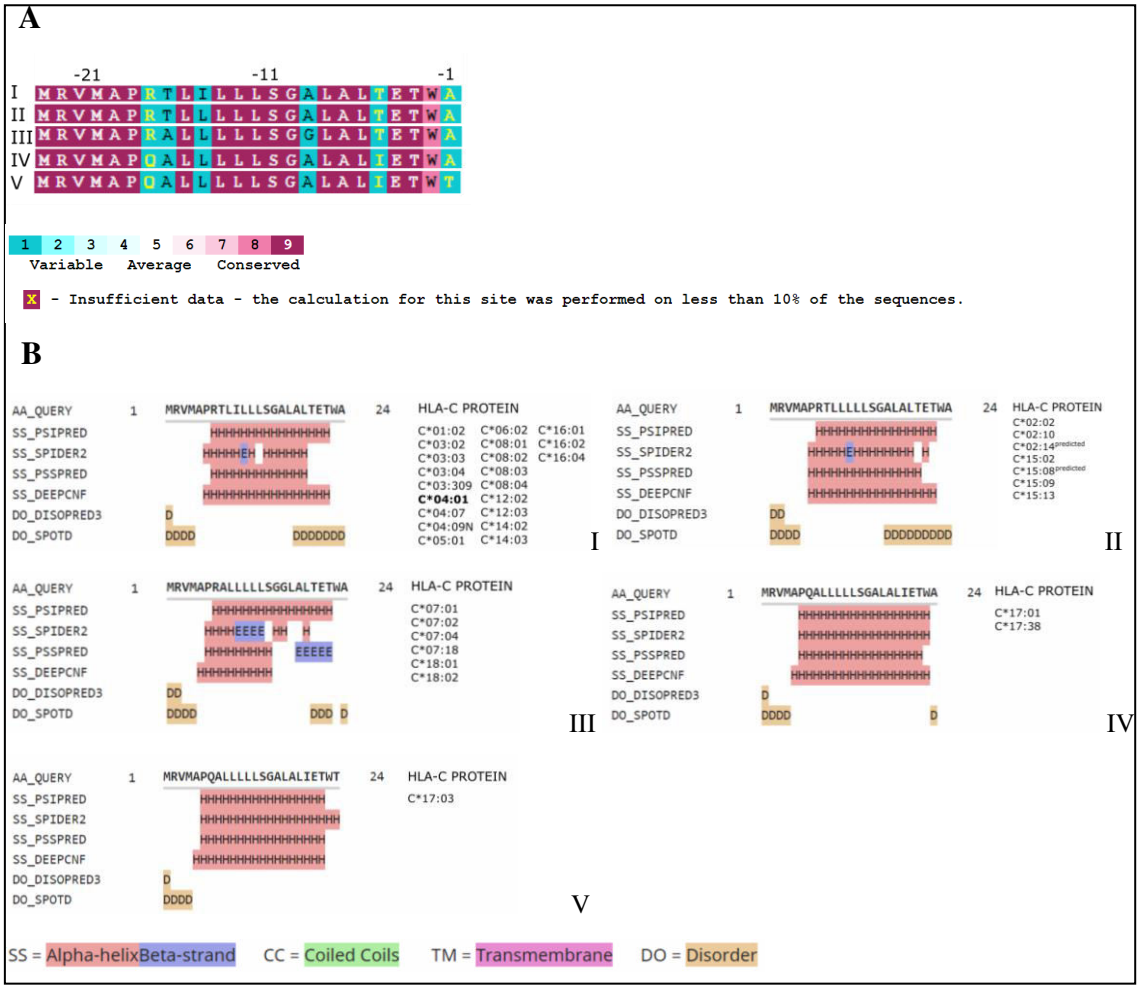


Figure 3. (A) Conservation of the HLA-C leader peptide sequences. The protein sequence residues were colored according to the level of conservation. The colored residues in pink denoted higher level of conservation among the found sequences and the most variable residues were colored in blue, according to the legend. The word "predicted" indicates protein sequences found in our population sample, which are only partially characterized on IPD-IMGT/HLA database. However, they could be predicted from the corresponding HLA-C coding sequence. **(B) Secondary structure prediction for each leader peptide sequence.** Representation of the combination of predictor output of secondary structure features such as alpha-helices, beta-strands, coiled coils (PSIPRED (JONES, 1999), SPIDER2 (ZHOU et al., 2017), PSSpred (YAN et al., 2013), DeepCNF-SS (WANG et al., 2016)) and intrinsically disordered regions (DISOPRED3 (JONES; COZZETTO, 2015), SPOT-Disorder (HANSON et al., 2017)).

9.3. Variability in the HLA-C peptide binding groove

For the peptide-binding groove, 30 different protein sequences were found (Figure 4A), and they can be divided in two groups based on a dimorphism at position 80 of the α1 domain. The subtype (or allotype) HLA-C1 (HLA-C*01, -C*03, -C*07, -C*08, -C*12, -C*14, -C*16:01 and -C*16:04), is characterized by the presence of an asparagine Asn^{80} (a polar and uncharged amino acid), while HLA-C2 (HLA-C*02, -C*04, -C*05, -C*06, -C*15, -C*16:02, -C*17, -C*18) is characterized by a lysine, Lys^{80} (a positively charged amino acid), as noted in Figures 4C and 4D. The frequency of both C1 (54.125%) and C2 (45.875%) groups were quite similar. In addition, the $Asn^{80}Lys$ co-occurred with the variant $Ser^{77}Asn$ (they are in absolute LD), and they occurred very closely in the same alpha-helix structure (Figure 4B). Interestingly, only these two variations related to C1/C2 allotypes differentiate proteins C*05:01 and -C*08:02, with similar frequencies (4.250% and 5.375%, respectively).

Figure 4B shows some polymorphic residues at the α1 and α2 domains detected in our population sample. Among these, attention is drawn to the residue in the position 97, which might be positively charged (Arg^{97}, Figure 4C and 4D) or uncharged (Trp^{97}). Among HLA-C residues interacting with the peptide C terminus, preferably hydrophobic and aromatic residues, such as those located at 74, 77, 80, 81, 84 positions of α1 domain and 95, 97, 116, 118, 143 and 147 positions of α2 domain (KAUR et al., 2017; VAN DEUTEKOM; KEŞMIR, 2015), only those located at 77, 80, 95, 97, 116 and 147 positions are actually polymorphic. Other important positions of the pocket remained monomorphic as Tyr^{7}, Phe^{22}, Tyr^{67}, among others. Residues which interact with TCR receptors of CD8+ T cells, such as 58, 62, 65, 69, 72, 76, 145, 149, 150, 151, 154, 158, 162 and 166 (VAN DEUTEKOM; KEŞMIR, 2015) are conserved in our population sample, exception for residue at position 163 that can also interact with peptide repertoire (VAN DEUTEKOM; KEŞMIR, 2015).

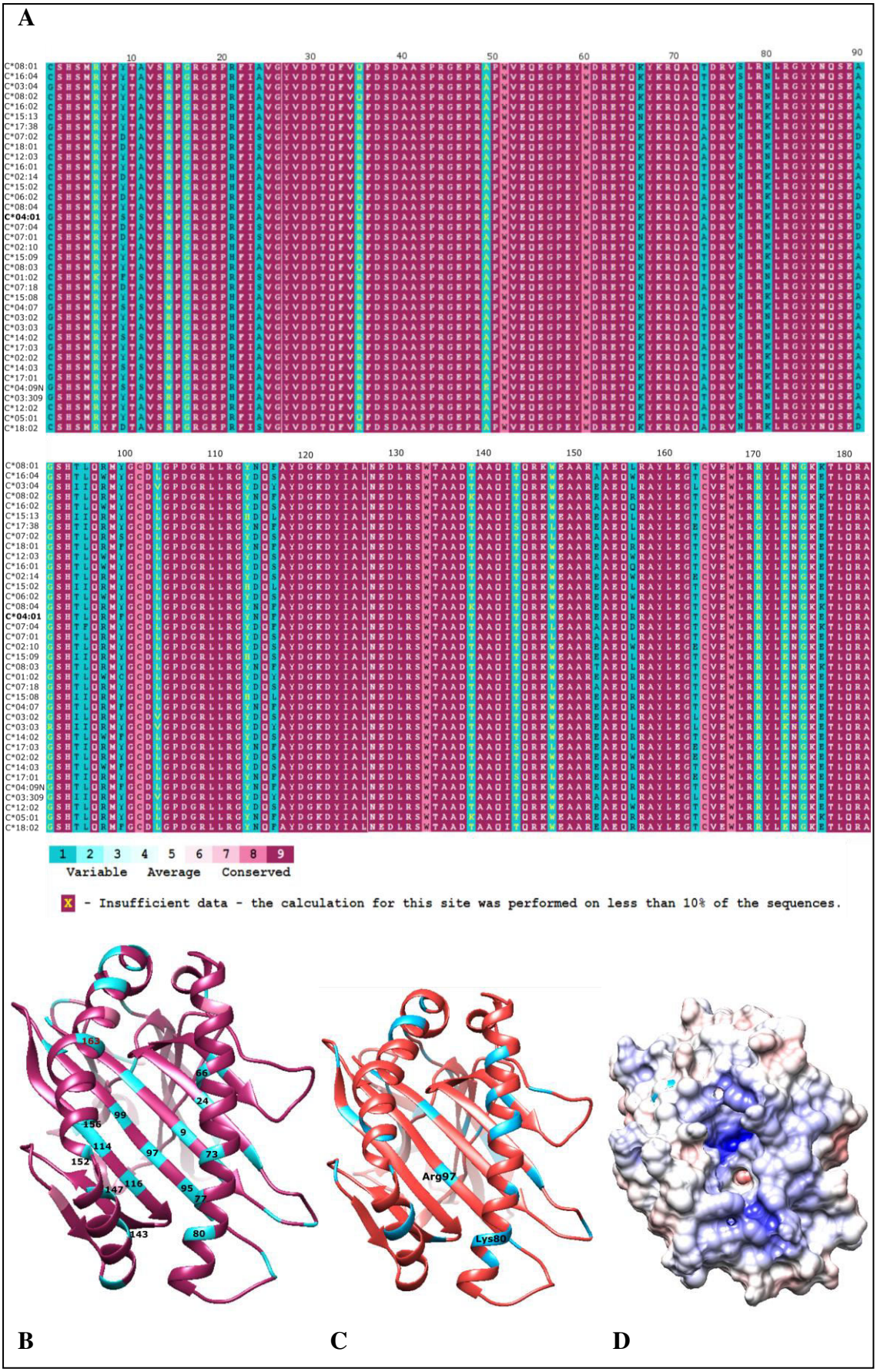


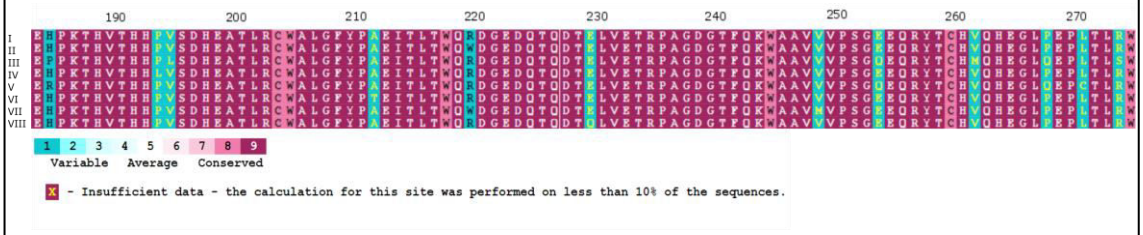
Figure 4. (A) Conservation of the HLA-C peptide binding groove sequences. The α1 domain encompasses the residues from 1 to 90 and α2 domain includes residues ranging from 91 to 182. **(B) Polymorphic residues which may influence the peptide binding** (BJOKMAN et al., 1987; VAN DEUTEKOM; KEŞMIR, 2015). The multiple protein sequences alignment was colored according to the level of conservation. The colored residues in pink denoted higher level of conservation among the found

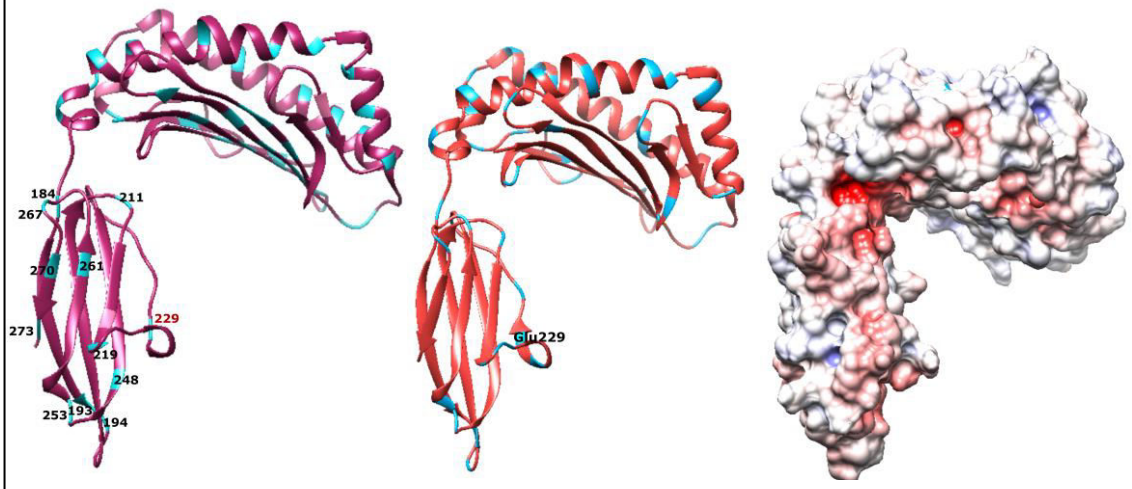
sequences and the most variable residues were colored in blue, according to the legend. **(C) Representation of positively charged amino acid residues in the modeled HLA-C*04:01 molecule.** Residues positively charged was colored in blue and in red was designated uncharged or negatively charged residues. **(D) Illustration of electrostatic surface potential of the HLA-C peptide binding groove in the modeled HLA-C*04:01 molecule.** In blue was represented the electrically positive regions, electrically negative regions were colored in red and neutral regions were in white.

9.4. Variability in the HLA-C α3 domain

The α3 domain presented 8 different protein sequences (Figure 5A). The residues Asp^{227} and Thr^{228}, respectively), which are responsible for the interaction with CD8 co-receptor, were conserved considering all protein sequences predicted at our sample. Although an adjacent residue is variable for the HLA-C*15:13 protein, its frequency is quite low in Brazil (0.375%, Table 1). This amino acid exchange ($Glu^{229}Gln$) is related to a non-synonymous rare substitution (rs41547622, Table S1). Figure 5B shows all polymorphic residues found in the HLA-C α3 domain for our population sample. It can be noticed that no deletion or insertion was detected for this segment.

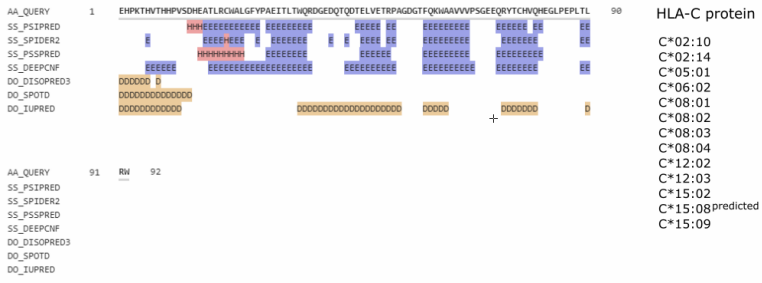
A





B **C** **D**

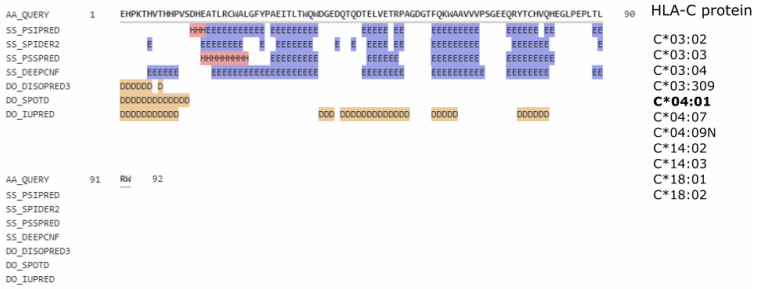
E



I

HLA-C protein

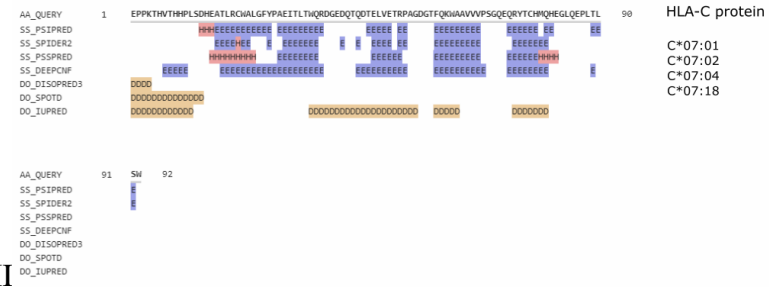
C*02:10
C*02:14
C*05:01
C*06:02
C*08:01
C*08:02
C*08:03
C*08:04
C*12:02
C*12:03
C*15:02
C*15:08 predicted
C*15:09



II

HLA-C protein

C*03:02
C*03:03
C*03:04
C*03:309
C*04:01
C*04:07
C*04:09N
C*14:02
C*14:03
C*18:01
C*18:02



III

HLA-C protein

C*07:01
C*07:02
C*07:04
C*07:18

…continued

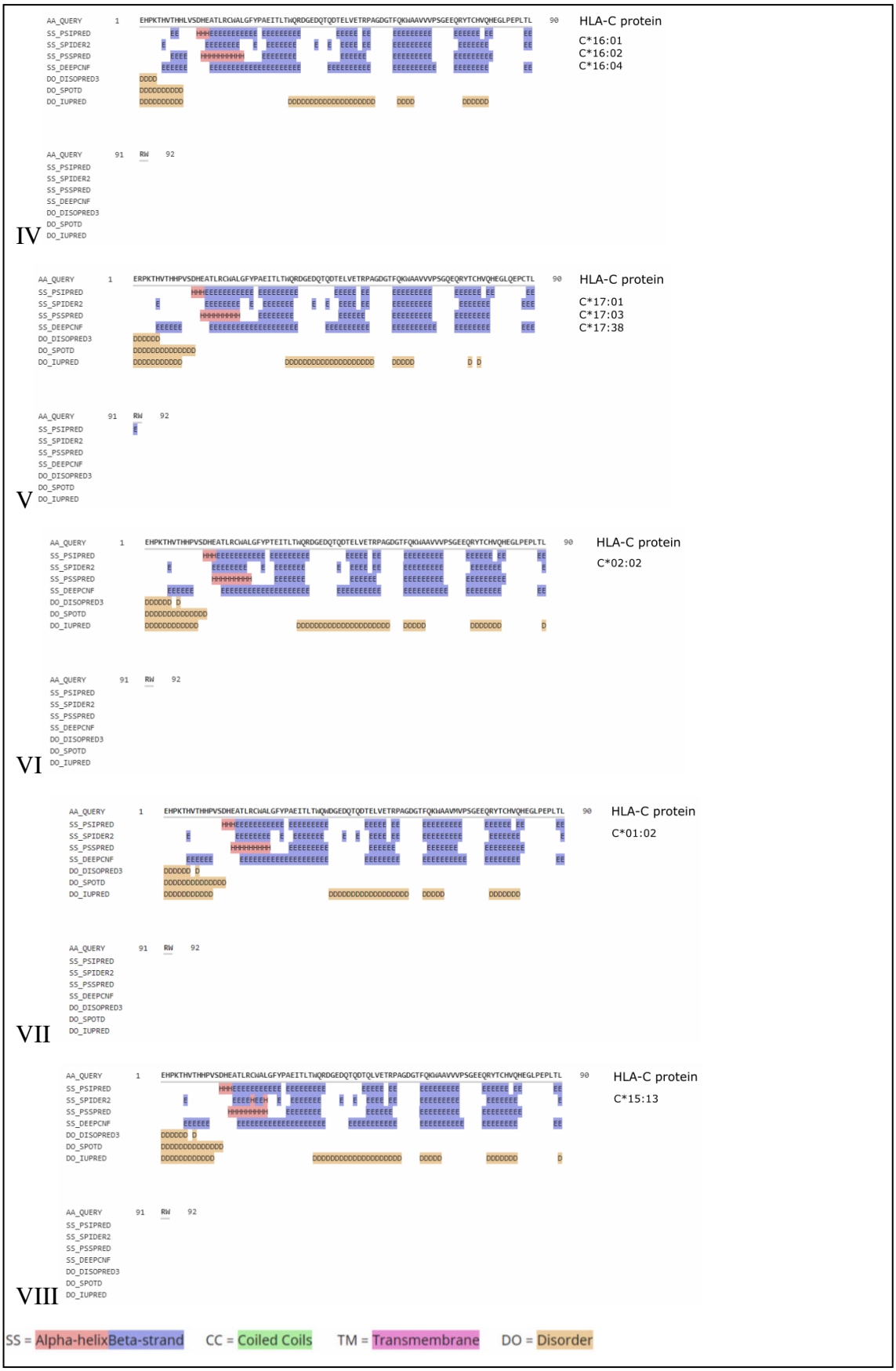


Figure 5. (A) Conservation of the HLA-C α3 domain sequences. The word "predicted" indicates protein sequences found in our population sample which are only partially characterized on IPD-IMGT/HLA database, however they could be predicted from the corresponding HLA-C coding sequence. **(B) Polymorphic residues and their respective positions.** The multiple protein sequences alignment was colored according to the level of conservation. The colored residues in pink denoted higher level of conservation among the found sequences and the most variable residues were colored in blue, according

to the legend. **(C) Representation of positively charged amino acid residues in the modeled HLA-C*04:01 molecule.** Residues positively charged was colored in blue and in red was designated uncharged or negatively charged residues. **(D) Illustration of electrostatic surface potential of the HLA-C α3 domain in the modeled HLA-C*04:01 molecule.** In blue was represented the electrically positive regions, electrically negative regions were colored in red and neutral regions were in white.
(E) Secondary structure prediction for each leader peptide sequence. Representation of the combination of predictor output of secondary structure features such as alpha-helices, beta-strands, coiled coils (PSIPRED (JONES, 1999), SPIDER2 (ZHOU et al., 2017), PSSpred (YAN et al., 2013), DeepCNF-SS (WANG et al., 2016)) and intrinsically disordered regions (DISOPRED3 (JONES; COZZETTO, 2015), SPOT-Disorder (HANSON et al., 2017), IUPred (DOSZTÁNYI et al., 2005)).

9.5. Variability in the HLA-C transmembrane domain

We detected 8 different protein sequences for the transmembrane domain (Figure 6A). In addition, an insertion of 18-bp (CAGCTGTCCTGGCTGTCC) at the IMGT/HLA position +2028 within exon 5 (Table S1) results in a transmembrane domain with six amino acids (PAVLAV) longer (CEREB; HUGHES; YANG, 1997) (VI – Figure 6A). This variant characterizes the HLA-C*17 serological type and it was detected in our sample associated with proteins HLA-C*17:01, -C*17:03 and -C*17:38 in our population sample. The repetition of the LAV motif associated with this insertion, and also all other variants detect at this segment, do not affect the protein secondary structure, as demonstrated at Figure 6B-VI.

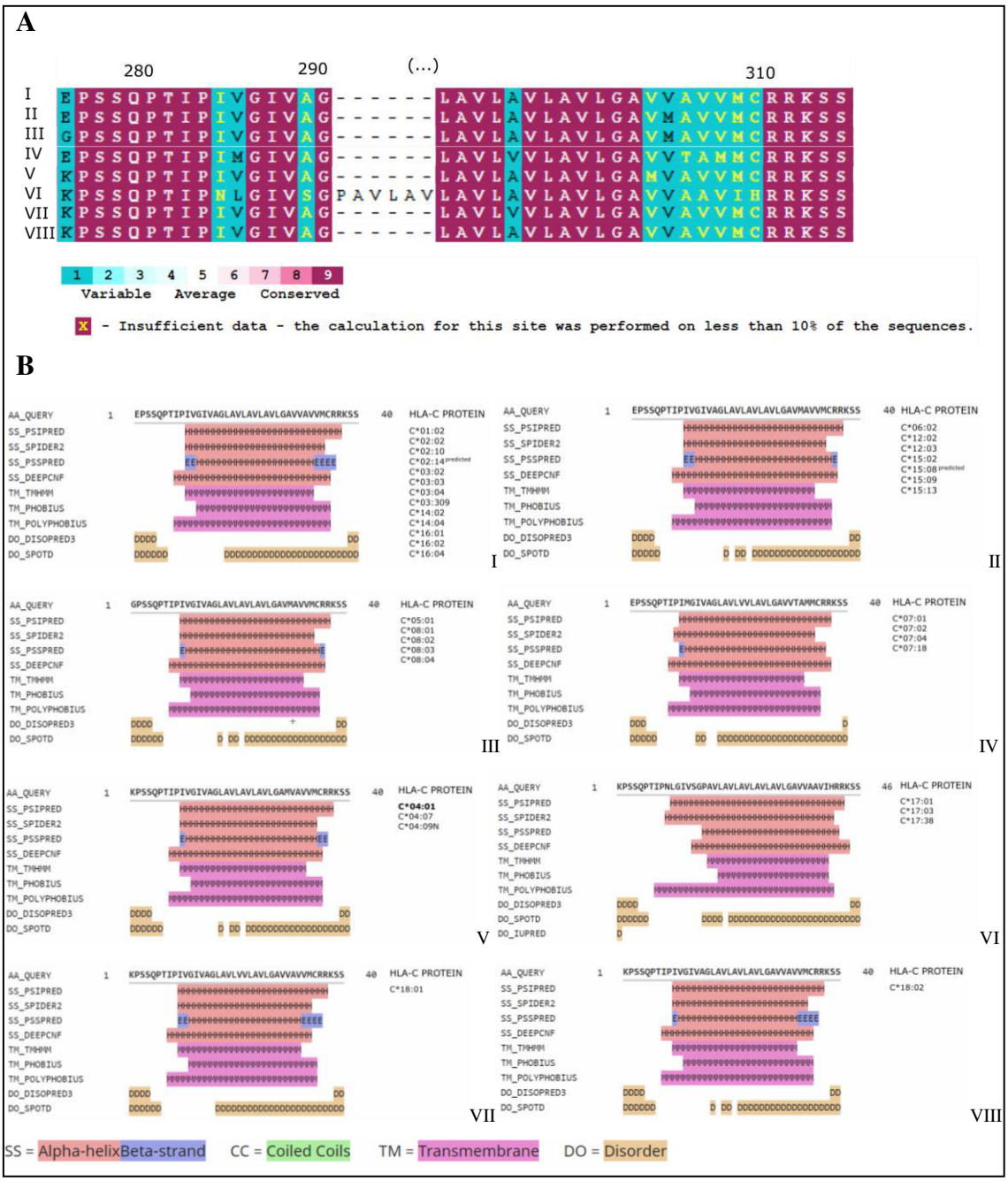


Figure 6. (A) Conservation of the HLA-C transmembrane domain sequences detected at Brazil. The protein sequence residues were colored according to the level of conservation. The colored residues in pink denoted higher level of conservation among sequences detected in Brazil and the most variable residues were colored in blue, according to the legend. The word "predicted" indicates protein sequences found in our population sample which are partially characterized on IPD-IMGT/HLA database. **(B) Secondary structure prediction for each transmembrane domain sequence.** Representation of the combination of predictor output of secondary structure features such as alpha-helices, beta-strands, coiled coils (PSIPRED (JONES, 1999), SPIDER2 (ZHOU et al., 2017), PSSpred (YAN et al., 2013), DeepCNF-SS (WANG et al., 2016)), transmembrane helices (TMHMM (KROGH et al., 2001), Phobius (KÄLL; KROGH; SONNHAMMER, 2004), PolyPhobius (KÄLL; KROGH; SONNHAMMER, 2005)) and intrinsically disordered regions (DISOPRED3 (JONES; COZZETTO, 2015), SPOT-Disorder (HANSON et al., 2017)).

9.6. Variability in the HLA-C cytoplasmic tail

The cytoplasmic tail domain presented 5 different protein sequences (Figure 7A) and the predicted secondary structure is intrinsically disordered (Figure 7B), i.e., regions which act in unfolded states. Nevertheless, this segment is quite conserved, since only the cytoplasmic tails I and II are frequent, and they differ by only two variable sites. The deletion of an Adenine at IMGT/HLA position +2726 associated with allele C*04:09N changes the reading frame, adding 32 additional amino acids to the cytoplasmic tail of the HLA-C*04:09N molecule. This molecule is retained (WANG et al., 2002), not being expressed in the cell membrane, but nothing is known about its function.

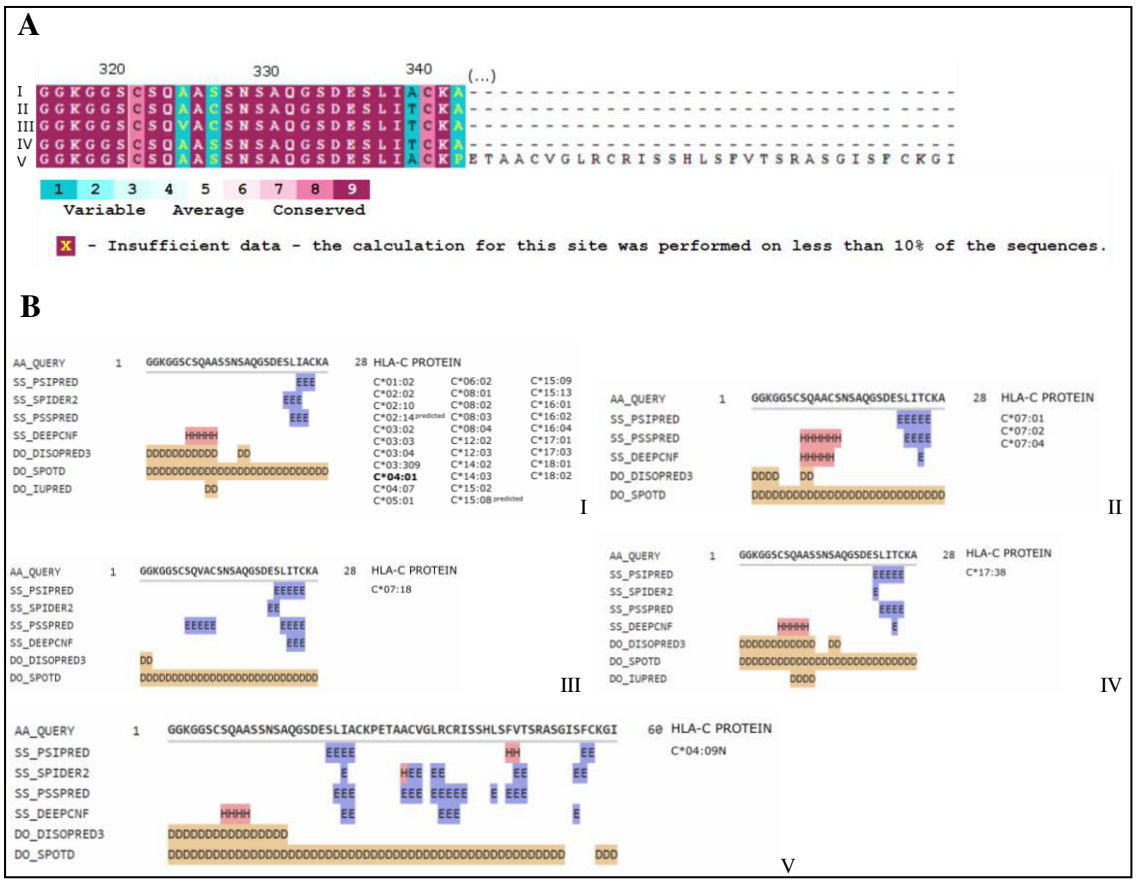


Figure 7. (A) Conservation of the HLA-C cytoplasmic tail sequences. The protein sequence residues were colored according to the level of conservation. The colored residues in pink denoted higher level of conservation among the found sequences and the most variable residues were colored in blue, according to the legend. The word "predicted" indicates protein sequences found in our population sample which are only partially characterized on IPD-IMGT/HLA database, however they could be predicted from the corresponding HLA-C coding sequence. **(B) Secondary structure prediction for each cytoplasmic tail sequence.** Representation of the combination of predictor output of secondary structure features such as alpha-helices, beta-strands, coiled coils (PSIPRED (JONES, 1999), SPIDER2 (ZHOU et al., 2017), PSSpred (YAN et al., 2013), DeepCNF-SS (WANG et al., 2016)) and intrinsically disordered regions (DISOPRED3 (JONES; COZZETTO, 2015), SPOT-Disorder (HANSON et al., 2017)).

10. DISCUSSION

The two most frequent leader peptide sequences contain motifs of nine amino acids that can stabilize the two most frequent HLA-E molecules in our sample population, HLA-E*01:01 and HLA-E*01:03 (CASTELLI et al., 2015; RAMALHO et al., 2017). The VMAP**RT**L**I**L (I – Figure 3A) cleaved peptide and the VMAP**RT**L**L**L peptide (II – Figure 3A) can bind HLA-E*01:03 most efficiently (CELIK et al., 2016). The first motif might be mimicked by the UL40 protein of the human cytomegalovirus (HCMV), promoting the upregulation of HLA-E molecule, and forming the HLA-$E^{VMAPRTLIL}$ complex. This complex is able to circumvent the NK cell-mediated lysis by serving as a ligand for the CD94/NKG2A or NKG2C inhibitory receptor (KRAEMER et al., 2015; KRAEMER; BLASCZYK; BADE-DOEDING, 2014).

Likewise, HLA-E*01:01 is stabilized by VMAP**RA**L**L**L motif (III – Figure 3A) (CELIK et al., 2016), derived from the leader peptide which has a secondary structure slightly different from the others (III – Figure 3B). The change in the structure may be a potential factor for the binding profile with different molecules. Nevertheless, the least frequent VMAP**QA**L**L**L (IV and V – Figure 3A) found were not presented by HLA-E*01:01 (DI MARCO et al., 2017), although it only contains only one amino acid change in position that contribute less to the proper interaction.

When considering the cluster into C1 and C2 groups of the HLA-C molecules in the context as ligands for KIR receptors on NK cells, the group C1 (HLA-C^{Asn80}) provide the ligand for the inhibitory KIR2DL2/KIR3DL3 and the activating KIR2DS2. Group C2 (HLA-C^{Lys80}) are ligand for inhibitory KIR2DL1 and activating KIR2DS1 (HIBY et al., 2004; MARTÍNEZ-LOSADA et al., 2017). Probably, the electrostatic potential surface is altered by the residue 80 exchange and this modify the KIR receptor profile to which the HLA-C molecule can bind. Studies indicate that patients may be assigned to a low or a high-risk group according to their C1/C2 ligand, once ligands to the C1 group (KIR2DL2 and KIR3DL3) are expressed earlier than specific C2-specific KIR ligand (KIR2DL1) in certain pathological situations (FISCHER et al., 2007, 2012). Thus, C1/C1 patients have a higher number of immunocompetent NK cells, whereas C2/C2 patients have a higher frequency of NK cells with delayed recognition (FISCHER et al., 2007).

The amino acid variation at position 80, which is attended by the mutation at position 77, might be the result of the same mutational event of multinucleotide mutation (MNM) during the gene replication, belonging to the spectrum of the error-

prone DNA polymerase (HARRIS; NIELSEN, 2014). This suggestion is supported by the characteristic transversion mutation, corresponding frequency, and proximity among the mutation points.

The preference for nonpolar and aromatic residues at the position 9, such as the bulky amino acids tyrosine and phenylalanine may restrict the size of the N-terminal-binding pocket. The same for the tyrosine located at position 99 of the HLA-C α2 domain, which is also situated at the bottom of the peptide-binding groove near the N-terminal end of the peptide (KAUR et al., 2017). A certain level of polymorphism has been found in the peptide binding groove, including positions 9, 99, and position 97, which might change the affinity with the peptide. However, the variations do not cause significant impact on the molecule function, since the exchange of amino acids is usually by chemically similar residues. Accordingly, it may be important for the restricted repertoire of peptides presented by HLA-C comparing with HLA-A and HLA-B molecules. The conserved regions found in the peptide binding groove might be essential to the overlap the HLA-peptide complex recognition by TCR receptors of $CD8^+$ T cells and KIR receptors of NK cells.

The interaction of T $CD8^+$ cells and HLA-C molecules involves essentially conserved segments and a negatively charged loop of the α3 domain (WESLEY et al., 1993). A $Glu^{229}Gln$ variation presented by HLA-C*15:13 molecule exchanges a negatively charged amino acid (Figure 5C and 5D) by a polar and uncharged amino acid. This exchange might decrease the negative charge of the loop and hinder the interaction among α3 domain and CD8 co-receptor (CONNOLLY et al., 1990). In our population sample, the HLA-C*15:13 is a rare molecule which can be offset by another co-expressed HLA-C molecules (-C*01:02, -C*08:02 or -C*15:02), with the negatively charged loop at α3 domain. Therefore, the maintenance of activation of cytotoxic T cells is not harmed.

The longer alpha-helix transmembrane domain, characteristic of HLA-C*17 serological type, seems to be functional and accompanied the presence of histidine rather than cysteine at position 309 for these molecules. Cysteine at position 309, as well as at positions 321 and 340, is generally conserved, because it facilitates in some aspect the recognition by NK cells via interaction of the inhibitory receptor LIR1 (also called ILT2) with α3 domain adjacent to this transmembrane/ cytoplasmic tail region (DAVIS et al., 1999; GRUDA et al., 2007). Cysteines also influence the post-translational modification named palmitoylation, which increases the hydrophobicity of

the HLA-C protein and contribute to its membrane anchoring, also influences the endocytosis of the molecule (DAVIS et al., 1999).

We observed that positions 321 and 340 were conserved. However, the $Cys^{309}His$ variation for HLA-C*17:01, -C*17:03 and -C*17:38 molecules might alter the structural conformation of the transmembrane domain and it could influence the α3 domain structure preventing the LIR1 binding. Beyond the HLA-C*17 molecules hinder the inhibition of NK cells by the structural modification at position 309 of transmembrane domain, their leader peptide proteins do not bind to HLA-E molecule, hindering the inhibition of NK cell activity by this way. Therefore, the HLA-C*17 molecule may have a peculiar function disfavoring the immunotolerance in the microenvironment in which it is present.

The most divergent cytoplasmic tail sequence found belongs to the null allele (HLA-C*04:09N), which was found in low frequency and in heterozygosity, reducing the impact of the absence of these HLA-C molecules performing its functions on the cell surface. However, the impact of the variation can not easily quantified, since the molecule is not expressed. In addition, considering all the HLA-C protein molecules here described, the conserved DESLI motif in the final portion of the cytoplasmic tail (334-338 positions) maintains the lysosomal targeting signal (SCHAEFER et al., 2008) and recycling of HLA-C proteins. Beyond the conservation of the HLA-C cytoplasmic tail sequences, the intrinsically disordered secondary structure guarantees a flexibility in the protein which reduces the impact on the function of the HLA-C molecule.

11. CONCLUSION

In general, even in an admixed and heterogeneous Brazilian population sample, the HLA-C proteins present a certain level of conservation. Although some residues are variable, they exchange for other residues of chemically similar profile. Even those residues which potentially might alter the function of the HLA-C molecule presented in critical regions of receptor interaction are in low frequency or they are rare. This minor impact on the function of the HLA-C is important to maintain the immunological role in physiological process such as pregnancy and pathological as HIV infection.

12. REFERENCES

APPS, R. et al. Influence of HLA-C Expression Level on HIV Control. **Science**, p. 1–10, 2013.

APPS, R. et al. Relative Expression Levels of the HLA Class-I Proteins in Normal and HIV-Infected Cells. **The Journal of Immunology**, v. 194, n. 8, p. 3594–3600, 2015.

ASHKENAZY, H. et al. ConSurf 2010: Calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. **Nucleic Acids Research**, v. 38, n. SUPPL. 2, p. 529–533, 2010.

BARRETT, J. C. et al. Haploview: Analysis and visualization of LD and haplotype maps. **Bioinformatics**, v. 21, n. 2, p. 263–265, 2005.

BERMAN, H. M. et al. The protein data bank. **Nucleic acids research**, v. 28, n. 1, p. 235–242, 2000.

BIASINI, M. et al. SWISS-MODEL: Modelling protein tertiary and quaternary structure using evolutionary information. **Nucleic Acids Research**, v. 42, n. W1, p. 252–258, 2014.

BJOKMAN, P. J. et al. The foreign antigen binding site and T cell recognition regions of class I histocompatibility antigens. **Nature**, v. 329, p. 512–518, 1987.

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. C. et al. Mapping Bias Overestimates Reference Allele Frequencies at the *HLA* Genes in the 1000 Genomes Project Phase I Data. **G3: Genes|Genomes|Genetics**, 2015.

CASTELLI, E. C. et al. HLA-E coding and 3' untranslated region variability determined by next-generation sequencing in two West-African population samples. **Human Immunology**, v. 76, n. 12, p. 945–953, 2015.

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, 2016.

CEREB, N.; HUGHES, A. L.; YANG, S. Y. Cw*1701 a new HLA-C allelic lineage with an unusual transmembrane domain. v. 115, p. 252–255, 1997.

CHAZARA, O.; XIONG, S.; MOFFETT, A. Maternal KIR and fetal HLA-C: a fine balance. **Journal of Leukocyte Biology**, p. 703–716, 2011.

CHOO, J. A. L. et al. The Immunodominant Influenza A Virus M158-66 Cytotoxic T Lymphocyte Epitope Exhibits Degenerate Class I Major Histocompatibility Complex Restriction in Humans. **Journal of Virology**, v. 88, n. 18, p. 10613–10623, 2014.

CINGOLANI, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. **Fly**, 2012.

CONNOLLY, J. M. et al. Recognition by CD8 on cytotoxic T lymphocytes is ablated by several substitutions in the class I alpha 3 domain: CD8 and the T-cell receptor recognize the same class I molecule. **Proceedings of the National Academy of Sciences of the United States of America**, v. 87, n. 6, p. 2137–41, 1990.

DAVIS, D. M. et al. The transmembrane sequence of human histocompatibility leukocyte antigen (HLA)-C as a determinant in inhibition of a subset of natural killer cells. **The Journal of experimental medicine**, v. 189, n. 8, p. 1265–74, 1999.

DI MARCO, M. et al. Unveiling the Peptide Motifs of HLA-C and HLA-G from Naturally Presented Peptides and Generation of Binding Prediction Matrices. **The Journal of Immunology**, p. ji1700938, 2017.

DOSZTÁNYI, Z. et al. IUPred: Web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. **Bioinformatics**, v. 21, n. 16, p. 3433–3434, 2005.

FISCHER, J. C. et al. Relevance of C1 and C2 epitopes for hemopoietic stem cell transplantation: role for sequential acquisition of HLA-C-specific inhibitory killer Ig-like receptor. **Journal of immunology** (Baltimore, Md. : 1950), v. 178, n. 6, p. 3918–23, 2007.

FISCHER, J. C. et al. The impact of HLA-C matching depends on the C1/C2 KIR ligand status in unrelated hematopoietic stem cell transplantation. **Immunogenetics**, v. 64, n. 12, p. 879–885, 2012.

GASTEIGER, E. et al. Protein Identification and Analysis Tools on the ExPASy Server. **The Proteomics Protocols Handbook**, p. 571–607, 2005.

GRUDA, R. et al. Intracellular cysteine residues in the tail of MHC class I proteins are crucial for extracellular recognition by leukocyte Ig-like receptor 1. **Journal of immunology** (Baltimore, Md. : 1950), v. 179, n. 6, p. 3655–61, 2007.

HANSON, J. et al. Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks. **Bioinformatics**, v. 33, n. 5, p. 685–692, 2017.

HARRIS, K.; NIELSEN, R. Error-prone polymerase activity causes multinucleotide mutations in humans. **Genome Research**, v. 24, n. 9, p. 1445–1454, 2014.

HIBY, S. E. et al. Combinations of Maternal KIR and Fetal HLA-C Genes Influence the Risk of Preeclampsia and Reproductive Success. **The Journal of Experimental Medicine**, v. 200, n. 8, p. 957–965, 2004.

JONES, D. T. Protein secondary structure prediction based on position-specific scoring matrices. **J. Mol. Biol.**, v. 292, p. 195–202, 1999.

JONES, D. T.; COZZETTO, D. DISOPRED3: Precise disordered region predictions with annotated protein-binding activity. **Bioinformatics**, v. 31, n. 6, p. 857–863, 2015.

KÄLL, L.; KROGH, A.; SONNHAMMER, E. L. L. A combined transmembrane topology and signal peptide prediction method. **Journal of Molecular Biology**, v. 338, n. 5, p. 1027–1036, 2004.

KÄLL, L.; KROGH, A.; SONNHAMMER, E. L. L. An HMM posterior decoder for sequence feature prediction that includes homology information. **Bioinformatics**, v. 21, n. SUPPL. 1, p. 251–257, 2005.

KAUR, G. et al. Structural and regulatory diversity shape HLA-C protein expression levels. **Nature Communications**, v. 8, n. May, p. 15924, 2017.

KLEIN, J.; SATO, A. The HLA system Second of Two Parts. **The New England Jounal of Medicine**, p. 782–786, 2000.

KRAEMER, T. et al. HLA-E : Presentation of a Broader Peptide Repertoire Impacts the Cellular Immune Response — Implications on HSCT Outcome. **Stem Cells International**, v. 2015, p. 1–5, 2015.

KRAEMER, T.; BLASCZYK, R.; BADE-DOEDING, C. HLA-E: A novel player for histocompatibility. **Journal of Immunology Research**, v. 2014, p. 1–6, 2014.

KROGH, A. et al. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. **Journal of Molecular Biology**, v. 305, n. 3, p. 567–580, 2001.

KULPA, D. A.; COLLINS, K. L. **The emerging role of HLA-C in HIV-1 infectionImmunology**, 2011.

KUMAR, S. et al. MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences. **Briefings in Bioinformatics**, v. 9, n. 4, p. 299–306, 2008.

LOVELL, S. C. et al. Structure validation by C alpha geometry: phi,psi and C beta deviation. **Proteins-Structure Function and Genetics**, v. 50, n. August 2002, p. 437–450, 2003.

MARSH, S. G. E. et al. Nomenclature for factors of the HLA system, 2010. **Tissue Antigens**, p. 291–455, 2010.

MARTÍNEZ-LOSADA, C. et al. Patients lacking a KIR-ligand of HLA group C1 or C2 have a better outcome after umbilical cord blood transplantation. **Frontiers in Immunology**, v. 8, n. JUL, p. 1–8, 2017.

MCKENNA, A. et al. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. **Genome Research**, p. 1297–1303, 2010.

MIRDITA, M. et al. Uniclust databases of clustered and deeply annotated protein sequences and alignments. **Nucleic Acids Research**, v. 45, n. D1, p. D170–D176, 2017.

MORRISON, B. A. et al. Multiple sclerosis risk markers in HLA-DRA, HLA-C, and IFNG genes are associated with sex-specific childhood leukemia risk. **Autoimmunity**, v. 43, n. 8, p. 690–7, 2010.

MOZER-LISEWSKA, I. et al. Genetic (KIR, HLA-C) and Some Clinical Parameters Influencing the Level of Liver Enzymes and Early Virologic Response in Patients with Chronic Hepatitis C. **Archivum Immunologiae et Therapiae Experimentalis**, v. 64, n. 1, p. 65–73, 2016.

PARHAM, P. MHC class I molecules and kirs in human history, health and survival. **Nature Reviews Immunology**, p. 201–214, 2005.

PETERSDORF, E. W. Optimal HLA matching in hematopoietic cell transplantation**. Current Opinion in Immunology**, 2008.

PETTERSEN, E. F. et al. UCSF Chimera - A visualization system for exploratory research and analysis. **Journal of Computational Chemistry**, v. 25, n. 13, p. 1605–1612, 2004.

RAMALHO, J. et al. HLA-E regulatory and coding region variability and haplotypes in a Brazilian population sample. **Molecular Immunology**, v. 91, n. September, p. 173–184, 2017.

RICE, P.; LONGDEN, I.; BLEASBY, A. EMBOSS: The European Molecular Biology Open Software Suite. **The European Molecular Biology Open Software Suite**, v. 16, n. 6, p. 276–277, 2000.

ROBINSON, J. et al. The IPD and IMGT/HLA database: Allele variant

databases. **Nucleic Acids Research**, v. 43, n. D1, p. D423–D431, 2015.

SANCHEZ-MAZAS, A. **An apportionment of human HLA diversity**. Tissue Antigens, 2007

SCHAEFER, M. R. et al. A novel trafficking signal within the HLA-C cytoplasmic tail allows regulated expression upon differentiation of macrophages. **Journal of immunology** (Baltimore, Md. : 1950), v. 180, n. 12, p. 7804–17, 2008.

SÖDING, J.; BIEGERT, A.; LUPAS, A. N. The HHpred interactive server for protein homology detection and structure prediction. **Nucleic Acids Research**, v. 33, n. SUPPL. 2, 2005.

STEPHENS, M.; SMITH, N. J.; DONNELLY, P. A New Statistical Method for Haplotype Reconstruction from Population Data. **The American Journal of Human Genetics**, p. 978–989, 2001.

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.

VARLA-LEFTHERIOTI, M. Role of a KIR/HLA-C allorecognition system in pregnancy. **Journal of Reproductive Immunology**, v. 62, n. 1–2, p. 19–27, 2004.

WANG, S. et al. Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields. **Scientific Reports**, v. 6, n. January, p. 1–11, 2016.

WANG, Z. C. et al. Molecular characterization of the HLA-Cw*0409N Allele. **American Journal Of Veterinary Research**, v. 8859, n. 2, p. 1457–1461, 2002.

WEBB, B.; SALI, A. Comparative Protein Structure Modeling Using MODELLER. **Curr Protoc Bioinformatics**, v. 54, p. 1–55, 2017.

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.

YAN, R. et al. A comparative assessment and analysis of 20 representative sequence alignment methods for protein structure prediction. **Scientific Reports**, v. 3, 2013.

ZHI, D. et al. Killer cell immunoglobulin-like receptor along with HLA-C ligand genes are associated with type 1 diabetes in Chinese Han population. **Diabetes/Metabolism Research and Reviews**, v. 32, n. 30, p. 13–23, 2014.

ZHOU, Y. et al. SPIDER2: A Package to Predict Secondary Structure, Accessible Surface Area, and Main-Chain Torsional Angles by Deep Neural Networks. [s.l: s.n.]. v. 1484

ZIMMERMANN, L. et al. A completely reimplemented MPI Bioinformatics Toolkit with a new HHpred server at its core. **Journal of Molecular Biology**, 2017.

13. SUPPLEMENTARY DATA

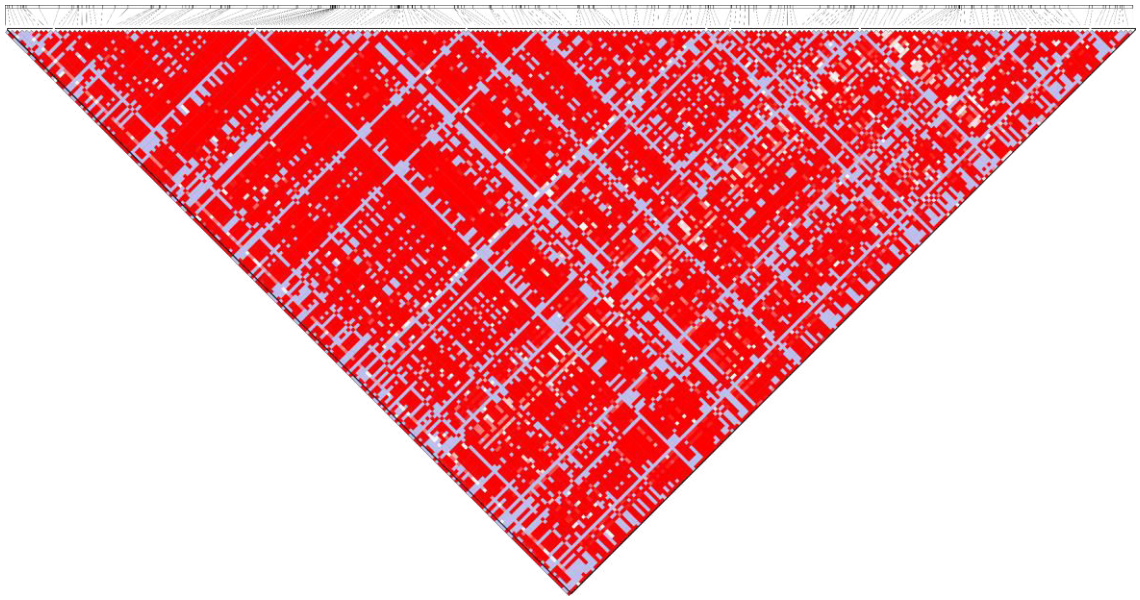


Figure S1. Linkage Disequilibrium (LD) between pair of single nucleotide polymorphisms (SNPs) considering the coding allele segment (from -283 to +3033 IMGT/HLA database positions). LD plot generated by Haploview 4.2, considering variable sites with a minor allele frequency (MAF) ≥ 1% and the fraction of strong LD in informative comparison set to 0.9. Areas in red color indicate strong LD [logarithm of odds (LOD score) ≥ 2, pairwise correlation between SNPs (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 white indicate no LD (LOD<2, D'<1)

Chr6 Position [a]	SNPid	*HLA-C* segment	Notes [b]	IMGT/HLA relative position [c]	Reference allele [d]	Frequency (2*n*=800)	First alternative	Frequency (2*n*=800)	Second alternative	Frequency (2*n*=800)
31237115	rs41544716	Exon 7	C,D,I	2726	CT	0.9975	C	0.0025		
31237124	rs1130838	Exon 7	B,I (p.Thr339Ala)	2717	T	0.2175	C	0.7825		
31237162	rs35708511	Exon 7	B,D,I (p.Cys326Ser)	2679	C	0.2163	G	0.7838		
31237275	rs41559915	Exon 6	B,I (p.Ala324Val)	2566	G	0.9825	A	0.0175		
31237286	.	Exon 6	A,F,G	2555	G	0.9988	A	0.0013		
31237295	.	Exon 6	A,F,G	2546	T	0.9988	C	0.0013		
31237760	rs1143551	Exon 5	B,I (p.Cys309His)	2082	C	0.9688	T	0.0313		
31237761	rs41560617	Exon 5		2081	A	0.9688	G	0.0313		
31237762	rs41540416	Exon 5	B,I (p.Met308Ile)	2080	C	0.9688	T	0.0313		
31237765	rs2308650	Exon 5		2077	C	0.9413	A	0.0588		
31237767	rs1130935	Exon 5	B,I (p.Met307Val)	2075	T	0.2163	C	0.7838		
31237769	rs1050105	Exon 5	B,I (p.Ala306Val)	2073	G	0.2475	A	0.7525		
31237771	rs1050106	Exon 5	A,I	2071	G	0.2163	A	0.7838		
31237773	rs1130947	Exon 5	B,I (p.Thr305Ala)	2069	T	0.2163	C	0.7838		
31237774	rs41540512	Exon 5	A,I	2068	G	0.2475	C	0.7525		
31237776	rs1050118	Exon 5	B,I (p.Val304Met)	2066	C	0.7075	T	0.2925		
31237779	rs146911342	Exon 5	B,I (p.Val303Met)	2063	C	0.8275	T	0.1725		
31237786	rs41556617	Exon 5	A	2056	A	0.2163	T	0.7838		
31237802	rs1050147	Exon 5	B (p.Val295Ala)	2040	A	0.2250	G	0.7750		
31237814	rs148706212	Exon 5	E,I	2028	A	0.9688	AGGACAGCCAGGACAGCTG	0.0313		
31237821	rs41545712	Exon 5	B,I (p.Ala289Ser)	2021	C	0.9688	A	0.0313		
31237833	rs1050180	Exon 5	B,I (p.Met285Val/Met285Leu)	2009	T	0.2163	C	0.7525	A	0.0313
31237835	rs1065600	Exon 5	B,I (p.Ile284Asn)	2007	A	0.9688	T	0.0313		
31237846	rs11757919	Exon 5	A,I	1996	G	0.9850	A	0.0150		
31237858	rs34794906	Exon 5	A,I	1984	T	0.6263	C	0.3738		
31237862	rs9264621	Exon 5	B,D,I (p.Glu275Gly)	1980	T	0.9000	C	0.1000		
31237987	rs41556321	Exon 4		1858	C	0.7813	T	0.2188		
31237991	rs2308628	Exon 4	B,I (p.Ser273Arg)	1854	G	0.2163	T	0.7838		
31238001	rs41546713	Exon 4	B,I (p.Leu270Cys)	1844	A	0.9688	C	0.0313		
31238002	rs41558713	Exon 4		1843	G	0.9688	A	0.0313		
31238009	rs1131014	Exon 4	A,I	1836	T	0.4738	C	0.5263		
31238010	rs1131015	Exon 4	B,I (p.Gln267Pro)	1835	T	0.2475	G	0.7525		
31238027	rs41540117	Exon 4	B,I (p.Met261Val)	1818	C	0.8275	A	0.1725		
31238029	rs2308622	Exon 4		1816	T	0.2163	C	0.7838		
31238048	rs41542914	Exon 4	A,I	1797	C	0.9688	T	0.0313		
31238053	rs707908	Exon 4	B (p.Gln253Glu)	1792	G	0.2475	C	0.7525		
31238068	rs1050276	Exon 4	B,I (p.Val248Met)	1777	C	0.9738	T	0.0263		
31238125	rs41547622	Exon 4	B,I (p.Glu229Gln)	1720	C	0.9963	G	0.0038		
31238126	rs2308618	Exon 4	A	1719	G	0.9063	A	0.0938		
31238135	rs1050317	Exon 4	A,I	1710	G	0.2163	A	0.7838		
31238138	rs1050320	Exon 4	A,I	1707	C	0.2163	T	0.7838		
31238147	rs1050326	Exon 4	A	1698	C	0.5900	G	0.4100		
31238155	rs1050328	Exon 4	B,I (p.Arg219Trp)	1690	G	0.6575	A	0.3425		
31238179	rs41562012	Exon 4	B,I (p.Ala211Thr)	1666	C	0.9600	T	0.0400		
31238230	rs1050716	Exon 4	B,I (p.Leu194Val)	1615	G	0.2163	C	0.7838		
31238232	rs1050343	Exon 4	B,I (p.Pro193Leu)	1613	G	0.9413	A	0.0588		
31238234	rs1050344	Exon 4	A,I	1611	G	0.2163	A	0.7838		
31238259	rs1131096	Exon 4	B,I (p.Pro184His/Pro184Arg)	1586	G	0.2163	T	0.7525	C	0.0313
31238851	rs2308604	Exon 3	A	994	T	0.2050	C	0.7950		
31238868	rs1131103	Exon 3	B (p.Glu177Lys)	977	C	0.8888	T	0.1113		
31238874	rs41552417	Exon 3	B,F,I (p.Gly175Arg)	971	C	0.9988	T	0.0013		
31238875	rs1131104	Exon 3	A	970	G	0.9063	A	0.0938		
31238880	rs1050357	Exon 3	B (p.Glu173Lys)	965	C	0.9063	T	0.0938		
31238889	rs2308598	Exon 3	B,I (p.Arg170Gly)	956	T	0.9688	C	0.0313		
31238909	rs1050686	Exon 3	B,I (p.Thr163Leu/Thr163Glu)	936	G	0.8163	A	0.0938	T	0.0900
31238910	rs1050685	Exon 3		935	T	0.8163	G	0.0938	C	0.0900
31238928	rs766416978	Exon 3	C	917	TCA	0.9888	T	0.0113		
31238930	.	Exon 3	E,I (p.Leu156Arg/Leu156Gln)	915	AG	0.3675	CG	0.5675	TG	0.0538
31238931	rs697743	Exon 3	C,I	914	G	0.7675	A	0.2213	GTC	0.0113
31238942	rs2308590	Exon 3	B,I (p.Ala152Thr/Ala152Glu)	903	G	0.2788	T	0.7213		
31238943	rs41552817	Exon 3		902	C	0.9975	T	0.0025		
31238957	rs1050366	Exon 3	B (p.Leu147Trp)	888	A	0.2475	C	0.7525		
31238970	rs142570222	Exon 3	B,I (p.Thr143Ser)	875	T	0.9688	A	0.0313		
31238983	rs2308585	Exon 3	A,I	862	G	0.2425	C	0.7175	T	0.0400
31238984	rs2308584	Exon 3	B (p.Thr138Asn)	861	G	0.9025	T	0.0975		
31238992	rs41550715	Exon 3	A,I	853	G	0.9600	A	0.0088	C	0.0313
31238995	rs41553316	Exon 3	A	850	G	0.9438	A	0.0563		
31239010	rs1050371	Exon 3	A	835	G	0.6838	A	0.3163		
31239016	rs1050373	Exon 3	A	829	G	0.9000	A	0.1000		
31239049	rs1065406	Exon 3	A,I	796	G	0.9663	T	0.0338		
31239050	rs713032	Exon 3	B,I (p.Ser116Phe/Ser116Tyr/Ser116Leu)	795	G	0.5238	A	0.3638	T	0.1125
31239057	rs2308575	Exon 3	B,I (p.Asp114Asn)	788	C	0.6813	T	0.3188		
31239060	rs2308574	Exon 3	B,I (p.Tyr113His)	785	A	0.9650	G	0.0350		
31239082	rs1050384	Exon 3	A,I	763	G	0.8125	C	0.1875		
31239090	rs34592426	Exon 3	B,I (p.Leu103Val)	755	G	0.9063	C	0.0938		
31239100	rs1131114	Exon 3	A,I	745	A	0.9225	G	0.0775		
31239101	rs1131115	Exon 3	B,I (p.Ser99Tyr/Ser99Phe/Ser99Cys)	744	G	0.0850	T	0.6663	A	0.2225
31239108	rs1131118	Exon 3	B,I (p.Arg97Trp)	737	T	0.7313	A	0.2688		
31239114	rs1071649	Exon 3	B,I (p.Leu95Ile/Leu95Phe)	731	G	0.8363	T	0.1525	A	0.0113
31239116	rs1131119	Exon 3	B (p.Thr94Ile)	729	G	0.8713	A	0.1288		
31239124	rs41543218	Exon 3	A,I	721	C	0.9688	A	0.0313		
31239376	rs1131122	Exon 2	B,D,I (p.Gly91Arg)	473	C	0.9563	T	0.0438		
31239378	rs1131123	Exon 2	B,I (p.Asp90Ala)	471	T	0.4963	G	0.5038		
31239407	rs17408553	Exon 2	B,I (p.Asn80Lys)	442	G	0.5413	T	0.4588		
31239417	rs2308557	Exon 2	B,I (p.Ser77Asn)	432	C	0.5413	T	0.4588		
31239430	rs41543814	Exon 2	B,I (p.Ala73Thr)	419	C	0.5925	T	0.4075		
31239449	rs28626310	Exon 2	B,I (p.Lys66Asn)	400	C	0.8450	G	0.1550		
31239501	rs1050409	Exon 2	B,I (p.Ala49Glu)	348	G	0.8288	T	0.1713		
31239506	rs1050414	Exon 2	A,I	343	C	0.8813	G	0.1188		
31239518	rs1050420	Exon 2	A,I	331	C	0.5063	T	0.4938		
31239543	rs1050428	Exon 2	B,I (p.Arg35Gln)	306	C	0.9000	T	0.1000		
31239577	rs707911	Exon 2	B,I (p.Ser24Ala)	272	A	0.3500	C	0.6500		
31239585	rs1050437	Exon 2	B,I (p.Arg21His)	264	C	0.8088	T	0.1913		
31239593	rs41542719	Exon 2	A,I	256	T	0.6988	C	0.3013		
31239601	rs151341100	Exon 2	B,I (p.Gly16Ser)	248	C	0.9413	T	0.0588		
31239602	rs281860336	Exon 2	A,I	247	G	0.9738	A	0.0263		
31239607	rs41542423	Exon 2	B,I (p.Arg14Trp)	242	G	0.8275	A	0.1725		
31239614	rs1050444	Exon 2	A,I	235	G	0.8125	A	0.1875		
31239616	rs1050445	Exon 2	B,I (p.Ala11Ser)	233	C	0.7663	A	0.2338		
31239617	rs1050446	Exon 2	A,I	232	G	0.7663	T	0.2338		
31239621	rs1071650	Exon 2	B,I (p.Asp9Tyr/Asp9Ser/Asp9Phe)	228	T	0.7663	G	0.2075	A	0.0263
31239622	rs9264668	Exon 2		227	C	0.3263	A	0.6738		
31239630	rs1131151	Exon 2	B,I (p.Arg6Lys)	219	C	0.9738	T	0.0263		
31239776	rs2074493	Exon 1	B,D,I (p.Cys1Gly)	73	A	0.7025	C	0.2975		
31239779	rs41553415	Exon 1	B,I (p.Ala(-1)Thr)	70	C	0.9950	T	0.0050		
31239790	rs41549413	Exon 1	B,I (p.Thr(-5)Ile)	59	G	0.9688	A	0.0313		
31239802	rs1050451	Exon 1	B,I (p.Gly(-9)Ala)	47	C	0.2313	G	0.7688		
31239821	rs2308527	Exon 1	B,I (p.Leu(-15)Ile)	28	G	0.3563	T	0.6438		
31239827	rs2308525	Exon 1	B,I (p.Ala(-17)Thr)	22	C	0.2625	T	0.7375		
31239829	rs41548123	Exon 1	B,I (p.Arg(-18)Gln)	20	C	0.9688	T	0.0313		

Table S1. List of variable sites detected in all exons of *HLA-C* gene.

The variations not found in the IMGT/HLA database, version 3.30.0, are marked in gray.

There is a third allele (a cytosine with frequency of 0.0263) at position 31,239,101.

The deletion at position 31,238,928 spans the position 31,238,930 with a frequency of 0.0113.

[a] Chromosome 6 position considering the human genome draft version hg19.

[b] Notes:

(A) Synonymous mutation on exon;
(B) Non-synonymous mutation on exon;
(C) Frameshift variant;
(D) Splice region variation in exon;
(E) Conservative inframe insertion;
(F) Singleton with a defined haplotype;
(G) New variable site on a region covered by the IPD-IMGT/HLA database, version 3.30.0;
(H) Variable site on a region not covered by the IPD-IMGT/HLA database, version 3.30.0;
(I) Variable site also detected by the 1000 Genomes Project, Phase 3, considering all the 2504 individuals.

The non-synonymous mutation on exon (B) are accompanied with notes about the amino acid change. Ala: alanine; Arg: arginine; Asn: asparagine; Asp: aspartic acid; Cys: cysteine; Gln: glutamine; Glu: glutamic acid; Gly: glycine; His: histidine; Ile: isoleucine; Leu: leucine; Lys: lysine; Met: methionine; Phe: phenylalanine; Pro: proline; Ser: serine; Thr: threonine; Trp: tryptophan; Tyr: tyrosine; Val: valine.

[c] The IMGT/HLA relative position, considering the Adenine of the first translated ATG as nucleotide +1.

[d] The reference allele considered at the human genome draft version hg19.

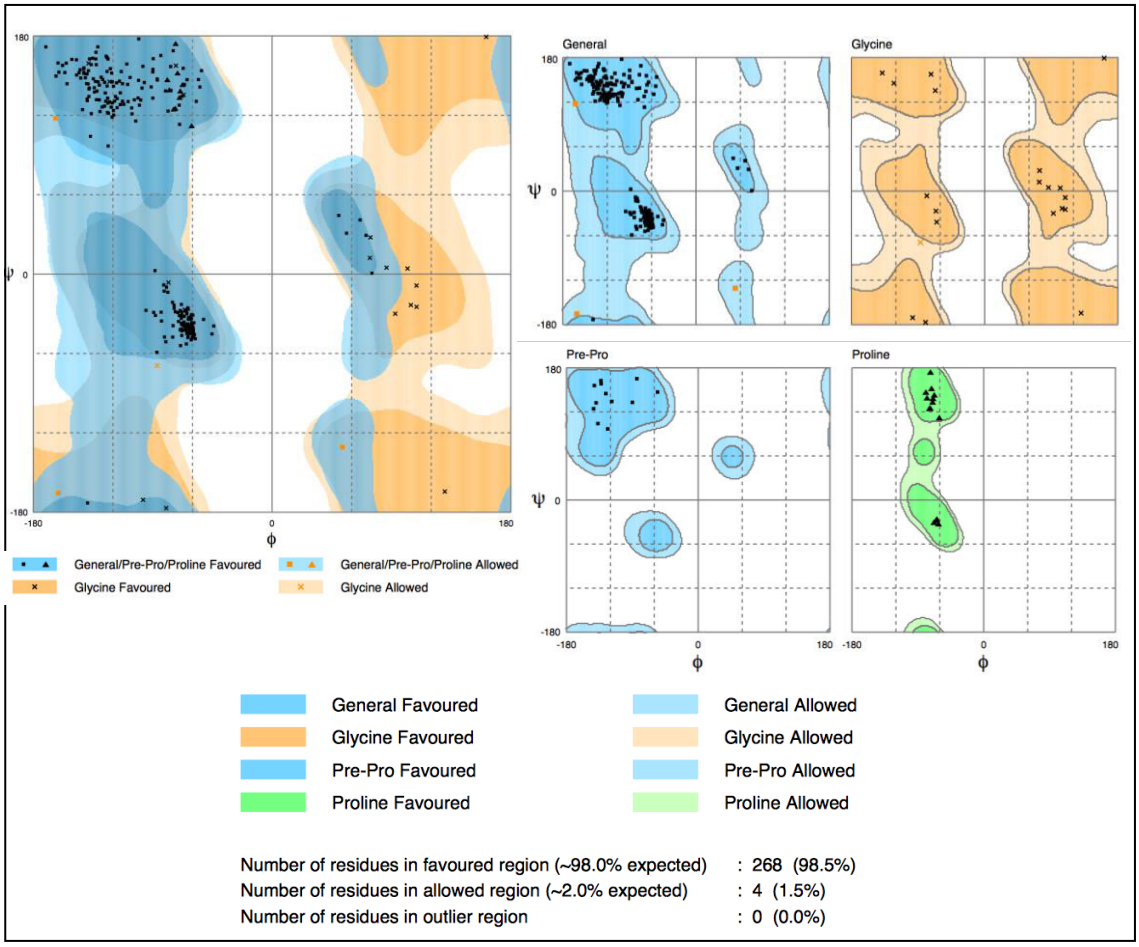


Figure S2. Ramachandran plot

14. CONCLUSÃO

Até mesmo considerando uma população miscigenada e heterogênea como a brasileira, os domínios α3, transmembrana e citoplasmático das proteínas HLA-C apresentam certo nível de conservação. As proteínas diferentes encontradas compartilham alguns segmentos entre si e isso pode ser consequência da conversão de genes, dando origem a divergentes sequências de HLA-C que apresentam altas frequências. Ainda que alguns resíduos se apresentam variáveis, geralmente a troca ocorre entre aminoácidos de perfis quimicamente semelhantes. De regra, as variações de resíduos que foram encontradas nos domínios estudados não alteram a estrutura secundária da proteína. Mesmo os resíduos, que potencialmente podem alterar a função da molécula HLA-C por estarem localizados em regiões críticas de interação com receptores, estão em baixa frequência ou são raros. Essa conservação característica nos os domínios α3, transmembrana e citoplasmático mostra-se importante para sustentar toda a função imunológica da molécula HLA-C e sua importância clínica.