

ESTUDO DA VARIABILIDADE DO GENE *HLA-A* E DETECÇÃO
DE ASSINATURAS DE SELEÇÃO NATURAL ATUANDO EM
CADA SEGMENTO DO GENE

THÁLITTA HETAMARO AYALA LIMA

Botucatu, SP

2019

UNIVERSIDADE ESTADUAL PAULISTA

“Júlio de Mesquita Filho”

INSTITUTO DE BIOCIÊNCIAS DE BOTUCATU

ESTUDO DA VARIABILIDADE DO GENE *HLA-A* E DETECÇÃO
DE ASSINATURAS DE SELEÇÃO NATURAL ATUANDO EM
CADA SEGMENTO DO GENE

THÁLITTA HETAMARO AYALA LIMA

PROF. DR. ERICK DA CRUZ CASTELLI

Tese apresentada ao Instituto de Biociências, Campus de Botucatu, UNESP, para obtenção do título de Doutor no Programa de Pós-Graduação em Ciências Biológicas (Genética).

Botucatu, SP

2019

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

Lima, Thálitta Hetamaro Ayala.

Estudo da variabilidade do gene *HLA-A* e detecção de assinaturas de seleção natural atuando em cada segmento do gene / Thálitta Hetamaro Ayala Lima. - Botucatu, 2019

Tese (doutorado) - Universidade Estadual Paulista "Júlio de Mesquita Filho", Instituto de Biociências de Botucatu
Orientador: Erick da Cruz Castelli
Capes: 20205007

1. Antígenos *HLA*. 2. Haplotipos. 3. Seleção natural. 4. Polimorfismo (Genética).

Palavras-chave: *HLA-A*; Sequenciamento de Nova Geração; haplótipos; seleção balanceadora; seleção natural.

Dedicatória

Dedico esta tese aos dois maiores incentivadores que tive na vida...

Ao meu querido e grande “Avohai” – avô e pai - Sebastião Romel Ferreira de Alencar, por assegurar ao mundo que eu sempre tive e terei um lar e que aos “trancos e barrancos” estaríamos sempre juntos lutando pelos sonhos da nossa família. Sentados à mesa muitas vezes discutimos a necessidade de exercitar nosso lado racional diante à emoção, ainda quando criança exercitamos muitas vezes as situações hipotéticas onde tinha eu que tomar uma decisão, mal sabia eu que em alguns anos essas situações se tornariam uma realidade e fazer escolhas passariam a fazer parte do dia a dia. Ao seu lado aprendi lições sobre honestidade e retidão, sobre nunca dar prejuízos a ninguém e sobre um amor inabalável à família. Quando partiu em regresso à nossa verdadeira casa, percebi que sua missão havia findado aqui na Terra, e pedi a Deus que ele continuasse a me guiar agora do “outro lado”, não duvidei por nenhum momento que você está por aí sempre a me proteger. Ouço sempre na mente sua voz dizendo: “Isso aí é *conditio sine qua non*”, expressão em latim que significa “condição sem a qual não”, que de forma simplificada se traduz pela necessidade de elementos essenciais para a resolução de questões - problemas. Passo meus dias a indagar diante aos meus, qual é a *conditio sine qua non*? Hoje preste a realizar um sonho “nosso”, te vejo sentado, calmo, feliz e um semblante que expressa: - “É disdaí, minha filha” (É isto aí, minha filha).

À minha querida avó (“Pachuquinha”) Heloisa Helena Melo De Alencar, que sonhou com este momento como ninguém, professora por 26 anos me ensinou a arte de amar ensinar aquilo que aprendemos. Companheira fiel e fraterna, agradeço todos os dias com um saudoso “Bom dia, Sol” a possibilidade de ter compartilhado nossa afinidade intensamente nessa oportunidade que nos foi concedida. Ao seu lado desfrutei meus mais sinceros sorrisos e um sentimento de gratidão que ultrapassa esta vida. Obrigada por me receber em seus braços como uma filha, por não ter medido esforços na realização dos meus sonhos e por me educar na certeza que o investimento na educação e ciência são os recursos que temos para melhorar nosso país. Levo você na memória sempre com aquele jeito fantástico de lidar com os problemas, sempre forte, guerreira, disposta e grata à vida. Que exemplo de mulher, inteligência, mãe, filha, esposa, avó e amiga. Você lutou bravamente contra uma doença cruel, mas nem mesmo o câncer foi motivo para te entristecer. Daqui sinto uma saudade irreparável e ainda dolorida. Peço a Deus que continue a te guiar nos passos da evolução espiritual, daqui sigo certa que nosso amor invadirá milênios e milênios neste espaço físico e sideral.

Agradecimentos especiais

Ao Prof. Dr. Erick da Cruz Castelli, pela oportunidade e credibilidade no desempenho do projeto e por me confiar a responsabilidade deste trabalho, ainda pela amizade e possibilidade de uma formação de qualidade.

À Universidade Estadual Paulista “Júlio Mesquita Filho”, Campus de Botucatu, a qual reconheço como berço e casa da minha formação científica, aqui vivenciei as melhores experiências da trajetória acadêmica.

Ao programa de Pós-Graduação em Ciências Biológicas (Genética), pelo respaldo e infraestrutura para o desenvolvimento da pesquisa a qual me dediquei.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pela concessão da bolsa de doutoramento, imprescindível para a realização da pós-graduação.

À Fundação de Amparo e Apoio a Pesquisa (FAPESP) pelo financiamento do projeto de pesquisa realizado pelo grupo, que viabilizou o desenvolvimento em sua inteireza - FAPESP/Brasil (2013/17084-2).

Agradecimentos

Agradeço primeiramente à Deus por me conceder a oportunidade da vida e saúde e sobretudo pelo bom ânimo para desempenhar as tarefas do dia a dia. Sempre acreditei em uma força que nos mantém despertos para vivenciar as oportunidades maravilhosas que a vida nos concede, atribuo essa força ao plano espiritual que me serviu de acalento e paz durante o desenvolvimento da tese.

A minha querida e admirável mãezinha Bianca Ayala Melo Di Alencar, por ser meu porto seguro em qualquer momento, por ser a pessoa a quem confio meus sonhos e expectativas, por ser quem se curva em oração pedindo proteção e me desejando o melhor que este plano possa oferecer. Por me apoiar em qualquer decisão, por sempre terminar o discurso com a frase: “Filha, estamos juntas”. Em um mundo onde cada vez mais as pessoas se distanciam das responsabilidades emocionais, esse suporte faz com que eu me sinta abençoada de maneira especial.

Aos meus queridos tios Thiago Ayala Melo Di Alencar e Diego Ayala Melo Di Alencar por acreditarem nesse sonho e sobretudo por acreditarem que podemos mudar o mundo com a educação, pelo incentivo e amor de irmãos que sempre tiveram comigo. As frases: “Parabéns, Thaty!”, “É isso aí, Nega!” serviram de incentivo a cada conquista compartilhada com vocês.

As queridas tias Karinna Matias, Luciana Teixeira e Renata Teles pelas palavras de incentivo e carinho sempre direcionadas a mim. Sou feliz por ter pessoas como vocês ao meu lado, um bom combate se faz bem amparado por pessoas que acreditam e que torcem por nós.

Aos meus queridos primos Nicholas Ayala, Laura Ayala, Yana Ayala e Nathan Ayala, pelo amor gratuito e olhar fraternal que me fazem querer ser a cada amanhecer um ser humano melhor e servir de inspiração para que sigam sempre na trilha do amor, semeando o bem sempre.

Ao meu querido padrinho Pablo Ayala Melo Di Alencar, mais conhecido por “Blito” por ser minha referência de alma boa e forte, por sempre me ouvir com zelo e atenção nos maiores dilemas que enfrentei nos últimos anos, por ser a pessoa que sempre acreditou que estou no caminho certo e que não devo desistir, por ser a palavra certa que busco nos momentos de incerteza. Blito, você é “fera”.

Ao meu amável namorado Marcelo Ribeiro Begnini, por ser fonte de inspiração e compromisso, por cuidar de mim com tanto zelo, proteção e amor. Por me incentivar sempre

quando o assunto são meus sonhos, por sempre responder com firmeza quando a indagação é: “Você acha que vou conseguir? ”. A resposta é sempre “Claro que vai! ”. As pitadas de bom humor e gargalhadas que compartilhamos deixam os meus dias mais divertidos e alegres. Me sinto privilegiada por ter você ao meu lado, obrigada por toda força e carinho a mim dispensados. Você merece todo o sucesso do mundo!

Ao Laboratório de Genética Molecular e Bioinformática pela estrutura e por servir de terreno fértil às ideias ao longo do desenvolvimento do projeto.

Aos colegas de trabalho Michelle de Almeida Paz, Emiliana Weiss, Heloisa de Souza Andrade, Marília Rodrigues Silva Passos, Nayane dos Santos Brito e Silva e Camila Ferreira Bannwart pelo intercâmbio de conhecimentos científicos, pelo bom humor e companheirismo imprescindíveis para um bom trabalho em equipe. Vocês são demais e continuarão a brilhar sempre.

A minha querida e inesquecível amiga Andréia da Silva Souza, você foi o maior presente que ganhei no curso do meu doutoramento, sem você compartilhando esse caminho lado a lado afirmo com plena certeza que não teria conseguido me manter de cabeça erguida em diversas situações. Você foi pela palavra o que me amparou nos momentos de fraqueza, o empurrão quando foi a atitude que precisei tomar em muitas situações e o abraço silencioso que me fortaleceu quando não eram nem de as palavras nem de atitudes a necessidade. Agradeço a Deus todos os dias pela sua existência, e tenho certeza que o mundo está de braços abertos a te esperar e felizes serão os que estiverem ao seu lado. Mantenha essa sua fé inabalável e inquestionável. No Departamento de Genética ficamos conhecidas como “Dupla Dinâmica” e nos experimentos como “B1 e B2”, esse é o sentimento que sempre terei a seu respeito: alguém que posso contar e que não abandona nunca aqueles que estão ao seu lado.

A todos os amigos que a UNESP e Botucatu me concederam e que participaram do meu crescimento profissional e pessoal diretamente, servindo de ombro amigo, ou compartilhando as melhores risadas do mundo, Nathalia Souza, Luciana Bonome, Raquel Bezerra, Ana Paula Carvalheira, Carol Vaccari, Luana Sacramento e Paula Michaela, com vocês tive a certeza que o mundo pode ser sim o lugar que sonhamos e que nunca devemos desistir do que acreditamos.

Aos queridos amigos Iane de Oliveira Pires Porto, Leandro Prado Felício e Kaison Ernane, que concluíram seus trabalhos anteriormente no grupo e que se tornaram os melhores amigos que alguém poderia ter. Vocês são a prova viva que nada na vida acontece ao acaso,

estava programado para compartilharmos um pedaço da caminhada juntos e sigo firme que continuaremos compartilhando nossos melhores e piores momentos juntos, uma gratidão eterna é o que eu sinto pela amizade de vocês.

A todos mestres, doutores e docentes da pós-graduação, cada qual com seu modo singular de ensinar e compartilhar experiências, meu privilégio e gratidão são imensos e seria incapaz de descrever o quão feliz eu sou por ter sido e continuar a ser aprendiz desse seleto grupo.

A todos os profissionais com os quais tive contato no desenvolvimento desse trabalho, incluindo, técnicos, secretaria da pós-graduação, departamento de limpeza, que contribuíram cada um ao seu modo para que esse sonho se realizasse.

Aos alunos que pude exercitar a prática docente, pela receptividade e atenção ao longo das aulas, com vocês compartilhei o meu melhor como profissional e pessoa. Obrigada por me permitirem aprender e exercitar a arte do ensino ao lado de vocês.

A todos meus familiares, colegas e amigos que fizeram parte desta trajetória, todos possuem participações singulares e essenciais para a realização deste sonho. A todos que participaram direta ou indiretamente desta conquista, meu muito obrigada!

“Bom dia, Sol, é minha *conditio sine qua non*”

Sumário

Capítulo I	14
Introdução	15
Estrutura e função das moléculas HLA	16
O gene <i>HLA-A</i> e suas funções	22
Seleção natural e o complexo HLA	25
Entendendo os testes de assinatura de seleção natural	29
Referências	31
Justificativa	37
Hipótese do estudo	39
Objetivo geral	40
Objetivos específicos	40
Capítulo II	41
Artigo científico	42
Abstract	43
Introduction	44
Methods	46
Population sample	46
<i>HLA-A</i> amplification	46
Library preparation and sequencing	46
Raw data processing (mapping)	47
Genotype calling and processing	47
Diversity indexes, neutrality deviation tests and other analyses	49
Results	50
<i>HLA-A</i> coding region haplotypes	51
<i>HLA-A</i> promoter haplotypes.....	51
<i>HLA-A</i> 3' untranslated region haplotypes.....	52
Nucleotide diversity and natural selection signatures at the <i>HLA-A</i> locus.....	52
Discussion	53
Legends	59
Acknowledgements	60
References	68
Conclusões	72
Material suplementar	73
Figure S1. Network of the extended <i>HLA-A</i> haplotypes (associated promoters, coding and 3'UTR haplotypes).....	74

Figure S2. Linkage Disequilibrium across the <i>HLA-A</i> locus.....	75
Figure S3. <i>HLA-A</i> promoter haplotypes detected in Brazil.....	76
Table S1: List of variable sites at the <i>HLA-A</i> gene detected in Brazil.	77
Table S2: List of all <i>HLA-A</i> protein molecules detected in a Brazilian population sample from Southeast Brazil.	89

Lista de Abreviaturas e Siglas

VCF	Variant Call Format
µL	Microlitro (10 ⁻⁶ litro)
APC	Célula Apresentadora de Antígeno (APC, do inglês, <i>Antigen Presenting Cell</i>)
DAA	Vantagem do Alelo Divergente (DAA, do inglês <i>Divergent Allele Advantage</i>)
DNA	Ácido Desoxirribonucleico (DNA, do inglês, <i>Deoxyribonucleic acid</i>)
EDTA	Ácido Etileno Diamino Tetracético
HLA	Antígeno Leucocitário Humano (do inglês, <i>Human Leukocyte Antigens</i>)
IMGT	International Immunogenetics Database
Kb	Kilobase (10 ³ bases)
KIR	Receptor do tipo imunoglobulina (KIR, do inglês <i>Killer Cell Immunoglobulin Like Receptor</i>)
LD	Desequilíbrio de ligação (LD, do inglês <i>Linkage disequilibrium</i>)
Mb	Megabase (10 ⁶ bases)
MHC	Complexo Principal de Histocompatibilidade (do inglês, <i>Major Histocompatibility Complex</i>)
mL	Mililitro (10 ⁻³ Litro)
ng	Nanograma (10 ⁻⁹ grama)
NGS	Sequenciamento de Nova Geração (do inglês, <i>Next Generation Sequencing</i>)
NK	Célula Natural Killer (NK, do inglês <i>Natural Killer</i>)
UTR	Região não traduzida (do inglês, <i>Untranslated Region</i>)
°C	Graus Celsius
Pb	Pares de bases
PCR	Reação em Cadeia da Polimerase (PCR, do inglês, <i>Polymerase Chain Reaction</i>)
SNP	Polimorfismo de Base Única (SNP, do inglês, <i>Single Nucleotide Polymorphism</i>)
SP	São Paulo
TCR	Receptor de Célula T (TCR, do inglês <i>T Cell Receptor</i>)
TAP	Transportador Associado a Peptídeos (do inglês, <i>Transporter Associated with Antigen Processing</i>)
CTL	Linfócito T Citotóxico (CTL, do inglês <i>Cytotoxic T lymphocyte</i>)
B2-M	Beta-2 microglobulina

Resumo

O Complexo HLA (*Human Leukocyte Antigen Complex*) é a região mais variável do genoma humano. O gene *HLA-A* é o segundo mais polimórfico do grupo e responsável por codificar moléculas envolvidas no processo de apresentação antigênica e modulação das células *NK* (*Natural Killer*). Diversos estudos avaliaram a variabilidade do gene *HLA-A* em populações mundiais. No entanto, estes estudos focaram-se principalmente na variabilidade dos éxons, ignorando a variabilidade de íntrons e regiões regulatórias. A variabilidade do gene *HLA-A* é particularmente elevada no segmento que codifica a fenda de ligação a peptídeos (éxons 2 e 3), relacionado com a função de apresentação antigênica e alvo de seleção balanceadora. No presente estudo avaliamos a diversidade genética do gene *HLA-A* considerando um segmento contínuo que compreende o promotor estendido (1.5 Kb upstream do primeiro ATG traduzido), todos os éxons (incluindo a região 3' não-traduzida) e íntrons por meio de sequenciamento paralelo massivo. As estratégias computacionais empregadas no estudo utilizam-se de programas de livre acesso ou desenvolvidos localmente, viabilizando um mapeamento adequado das sequências produzidas e a genotipagem/haplotipagem para genes HLA. A variabilidade detectada para o gene *HLA-A* em uma amostra brasileira demonstrou que os segmentos regulatórios apresentam poucas sequências diferentes, porém as sequências existentes são bastante divergentes entre si. As variantes regulatórias apresentam alto desequilíbrio de ligação com variantes da região codificadora. Assim como os éxons, os íntrons do gene *HLA-A* são bastante variáveis. Além disso, padrões de diversidade molecular sugerem que a região promotora, juntamente com a região codificadora, estão sob o mesmo tipo de pressão seletiva, mas um cenário diferente é observado para o éxon 4 e para a região 3' não-traduzida.

Palavras-chave: seleção balanceadora, haplótipos, alelos, *HLA-A*, seleção natural, Sequenciamento de Nova Geração, polimorfismos, evolução.

Capítulo I

Introdução

O Complexo Principal de Histocompatibilidade (MHC, do inglês, *Major Histocompatibility Complex*) consiste de um grupo de genes localizados no braço curto do cromossomo seis (6p21.3), em um segmento de aproximadamente 4Mb (0.13% do genoma humano). Este complexo apresenta cerca de 200 genes (Klein e Sato, 2000a; b), representando ao todo cerca de 0.5% dos genes codificadores de proteínas descritos no genoma humano (Shiina *et al.*, 2009). Em sua maioria os produtos gênicos codificados pelo MHC estão de alguma forma envolvidos com a resposta imune, sendo eles: ligantes, receptores, fatores de sinalização celular, moléculas reguladoras do processo transcricional. Os produtos gênicos codificados pelo MHC estão relacionados a contextos imunes diversos, tais como: inflamação, processamento e apresentação antigênica (resposta imune adaptativa) e interações com receptores nas células *NK* (do inglês, *Natural Killer*), importantes na resposta imune inata (Shiina *et al.*, 2009). A região do MHC é considerada a região de maior densidade gênica do genoma humano, com uma média de 1 gene a cada 16Kb (Meyer *et al.*, 2018). Além disso, o MHC é a região mais variável do genoma humano e o elevado polimorfismo deve-se principalmente aos genes HLA (*HLA*, do inglês *Human Leucocyte Antigens*). O complexo HLA codifica moléculas responsáveis pelo processamento e apresentação de antígenos. Os genes do complexo HLA estão envolvidos nos processos de combate a patógenos, pois as moléculas que codificam apresentam antígenos aos linfócitos T (Klein e Sato, 2000a).

Didaticamente, os genes do complexo HLA estão divididos em dois grupos principais, denominados genes de classe I e II. Os genes de classe I, *HLA-A*, *HLA-B* e *HLA-C*, são também conhecidos como genes HLA de **classe I clássicos** e codificam glicoproteínas transmembrana expressas na maioria das células nucleadas. Essas moléculas atuam na apresentação de antígenos extracelulares aos linfócitos TCD8 durante a resposta imunitária adaptativa (Klein e Sato, 2000a; b; Goldberg e Rizzo, 2015). Os genes *HLA-G*, *HLA-E* e *HLA-F* são considerados genes **de classe I não clássicos** (Geraghty *et al.*, 1992; Klein e Sato, 2000a; b). Embora apresentem semelhança estrutural e molecular aos genes clássicos, possuem padrões de expressão distinto e estão relacionados principalmente com o processo de tolerância imunológica (Barrow e Trowsdale, 2008; Carosella *et al.*, 2008; Crispim, Duarte, *et al.*, 2008; Harada *et al.*, 2015; Martinez *et al.*, 2017). Em geral, as moléculas não clássicas ligam-se a receptores presentes em linfócitos T e células *NK*, modulando a ação destas células (Donadi *et al.*, 2011). Os genes de classe II, *HLA-DPA1*, *HLA-DPB1*, *HLA-DQA1*, *HLA-DQB1*, *HLA-DRB1* e *HLA-DRA1* também estão envolvidos com o processo de apresentação antigênica,

porém suas moléculas apresentam antígenos aos linfócitos TCD4 na superfície celular (Touraine *et al.*, 1989).

Os genes HLA são considerados os genes mais polimórficos do genoma humano, possuindo uma ampla diversidade alélica, ou seja, diversas sequências de DNA específicas para um determinado gene HLA. O banco de dados IPD-IMGT/HLA constitui um banco especializado em catalogar sequências HLA, contendo informações detalhadas sobre a origem e validação dessas sequências (Robinson *et al.*, 2016). Este banco de dados conta atualmente com um total de 17.344 sequências de alelos descritos. A gama de sequências HLA descritas extrapola mais de 16.000 sequências para genes HLA de classe I (Tabela 1) e por volta de 4.000 sequências para os genes HLA de classe II (Robinson *et al.*, 2016).

Tabela 1. Número de alelos identificados para os genes HLA de classe I (banco de dados IPD-IMGT/HLA, versão 3.36.0.1/20-02-2019).

HLA DE CLASSE I CLÁSSICOS		
Gene	Alelos	Proteínas
<i>HLA-A</i>	5018	3398
<i>HLA-B</i>	6096	4248
<i>HLA-C</i>	4852	3193
HLA DE CLASSE I NÃO CLÁSSICOS		
Gene	Alelos	Proteínas
<i>HLA-E</i>	30	9
<i>HLA-F</i>	44	8
<i>HLA-G</i>	68	19

Estrutura e função das moléculas HLA

As moléculas HLA de classe I clássicas são expressas na maioria das células nucleadas, ligam peptídeos de origem endógena e os apresentam na superfície celular aos receptores de linfócitos TCD8. Caso sejam reconhecidos como não-próprios, uma resposta citotóxica poderá ocorrer. No entanto, caso peptídeos próprios sejam reconhecidos como estranhos ao organismo, há possibilidade de desenvolvimento de doenças autoimunes (Matzaraki *et al.*, 2017). As moléculas de classe II são expressas em células apresentadoras de antígenos e em geral apresentam peptídeos de origem exógena às células TCD4 na superfície celular, levando a um processo de sinalização celular que estimula a multiplicação das células *T-helper*, levando, por

consequência, a estimulação das células B, que são responsáveis pela produção de anticorpos contra o antígeno que desencadeou a resposta (Rock *et al.*, 2016).

As moléculas HLA de classe I e II apresentam fragmentos protéicos às células TCD8 e TCD4, respectivamente. Essas moléculas são essenciais para a imunidade mediada por células e surgiram no início do sistema imune adaptativo cerca de 500 milhões de anos atrás (Flajnik e Kasahara, 2010). A estrutura de uma molécula de classe I, independentemente do gene que a codifica, compreende uma cadeia α pesada codificada no cromossomo 6, associada com uma cadeia β_2 -microglobulina codificada no cromossomo 15 (Figura 1) (Klein e Sato, 2000a). A cadeia α possui cinco domínios principais. Os domínios $\alpha 1$ e $\alpha 2$ dão origem à ‘fenda de ligação a peptídeos’ e estão envolvidos com o acoplamento de peptídeos. Esta fenda provavelmente evoluiu a partir da estrutura de chaperonas, uma vez que a) chaperonas ligam-se a fragmentos proteicos, b) diversas chaperonas possuem estrutura similar as moléculas MHC; c) existem diversas moléculas que se assemelham estruturalmente ao MHC e que atuam como chaperonas específicas no processo de apresentação antigênica restrita ao MHC (*i.e* tapasina) (Dong *et al.*, 2009; Rock *et al.*, 2016). As cadeias α possuem também um domínio semelhante à imunoglobulina ($\alpha 3$), que interage com co-receptores presentes nas células T, um domínio transmembrana e um domínio citoplasmático (Figura 1) (Klein e Sato, 2000a).

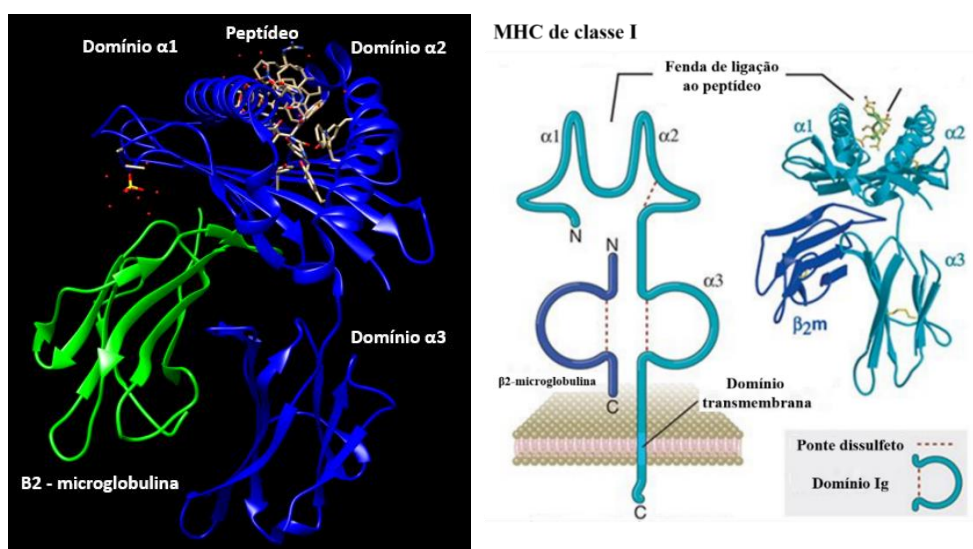


Figura 1. À esquerda, representação gráfica computadorizada da molécula HLA-A2, contendo todos os domínios estruturais e um peptídeo acoplado na fenda ligadora de peptídeos. **Fonte:** Protein Data Bank (<http://www.rcsb.org/structure/5E00>). À direita, ilustração da estrutura da molécula HLA de classe I, composta por três domínios α , região transmembrana acoplada a membrana plasmática, cauda citoplasmática e o domínio de β_2 -microglobulina, ligado por meio de ligações não-covalentes à cadeia α . **Fonte:** Adaptado de Abbas, *et. al.* Cellular and Molecular Immunology, 2008, p. 193.

Esta semelhança estrutural entre as diferentes moléculas de classe I se deve ao fato de que os genes desse complexo originaram-se por duplicações imperfeitas de blocos gênicos (Kulski *et al.*, 2000). O MHC tem sido descrito como uma região polimórfica composta por blocos conservados. Na região de classe I temos os blocos gênicos alfa e beta. O bloco alfa contém os genes *HLA-F*, *HLA-G*, *HLA-A* e pseudogenes como *HLA-H* e *HLA-J*. Dentro do bloco beta existem duas grandes duplicações de segmentos contendo os genes *HLA-B*, *HLA-C*, *MICA* e *MICB*. Essas duplicações contêm segmentos que compartilham aproximadamente 30 kb de sequências, incluindo sequências codificadoras e intergênicas, sendo que a maior parte das diferenças entre as duplicações estão relacionadas com a alta quantidade de inserções e deleções, contendo retroelementos, pseudogenes e outras sequências repetitivas inseridas nesses segmentos (Dawkins *et al.*, 1999; Gaudieri *et al.*, 1999; Kulski *et al.*, 2000). A origem comum evidenciada para os genes do MHC acarreta uma grande similaridade de sequência entre os vários genes HLA, dificultando abordagens para análise individual da estrutura e variabilidade de cada um desses genes.

No presente trabalho, avaliamos a variabilidade completa do gene *HLA-A*, incluindo suas regiões regulatórias e íntrons, por meio de sequenciamento paralelo massivo. Por conta da alta similaridade entre os genes do complexo HLA e do elevado polimorfismo, desenvolvemos um workflow para análise de dados utilizando ferramentas de bioinformática. Os métodos aplicados para a análise de dados minimizam os custos de genotipagem e definição dos haplótipos, uma vez que (a) utiliza bancos de dados e softwares gratuitos, (b) utiliza primers desenvolvidos localmente e kits de preparação de biblioteca de DNA comuns, sem custos adicionais além dos reagentes de amplificação e sequenciamento, e (c) permite a customização dos experimentos, uma vez que outros alvos (genes) podem ser avaliados em conjunto com o loci HLA. Além disso, a metodologia aqui desenvolvida avalia todo o gene, incluindo segmentos regulatórios, ao invés de focar apenas nas sequências codificadoras como comumente realizado na prática clínica (Sheldon e Poulton, 2006). Para minimizar os problemas de mapeamento errôneo de sequências de genes HLA, utilizamos o software ***hla-mapper***, que foi desenvolvido localmente por ocasião da realização deste estudo e de estudos envolvendo outros genes HLA (Castelli *et al.*, 2018).

As moléculas HLA de classe I e classe II são consideradas únicas no proteoma, uma vez que apresentam elevado polimorfismo (>16.000 diferentes alelos já catalogados) (Tabela 1). Essa elevada variabilidade possui consequências interessantes, pois a presença de polimorfismo no topo das *alpha* hélices está relacionada com modificação no padrão de interação entre a molécula

HLA e o TCR (do inglês, *T Cell Receptor*). Além disso, variações no sítio de ligação ao peptídeo estão relacionadas com o perfil de peptídeos apresentados (Rock *et al.*, 2016).

Os resíduos polimórficos presentes na fenda de ligação ao peptídeo são capazes de gerar mudanças na localização dos ‘bolsões’ (*pockets*). Esses *pockets* são preenchidos por cadeias laterais de aminoácidos variáveis complementares, também conhecidos como ‘resíduos âncoras’, com a consequência de que diferentes fragmentos de um determinado antígeno podem ser apresentados por diferentes moléculas HLA (Falk *et al.*, 1994; 2006). Próximo aos resíduos âncoras, outros aminoácidos preenchem o espaço livre, podendo ser qualquer um dos 20 aminoácidos existentes. Uma vez que essas moléculas possuem *pockets* com especificidade para somente algumas poucas cadeias laterais, permitindo que haja variação de 6 a 10 aminoácidos dentre todas as possibilidades, cada molécula HLA pode apresentar um alto repertório de peptídeos. Em teoria as moléculas HLA de classe I podem apresentar cerca de 6×10^6 peptídeos diferentes e HLA de classe II cerca de 12×10^{10} . Contudo, todo esse repertório de peptídeos não pode ser apresentado na superfície celular, uma vez que há uma estimativa de que cerca de 200.000 moléculas HLA de classe I e 20.000 de classe II seriam encontradas na superfície de uma célula T ou B (Eisen *et al.*, 2012).

Importante ressaltar que diferentes moléculas HLA apresentam um conjunto diferente de antígenos. Portanto, quando há mutações em um determinado patógeno que altera um de seus resíduos âncoras críticos para interação adequada com as moléculas HLA, isso prejudica a apresentação desse antígeno por um indivíduo portador das moléculas HLA que originalmente apresentariam este antígeno. Porém, essa apresentação não é prejudicada em indivíduos portadores de outras moléculas HLA (Schmid *et al.*, 2010). Desta forma, o polimorfismo das moléculas HLA é de extrema importância para a adaptabilidade de uma população.

As moléculas clássicas de classe I (HLA-A, HLA-B e HLA-C) associam-se a peptídeos antigênicos oriundos principalmente do meio intracelular. No entanto, alguns microrganismos são capazes de adentrar o meio intracelular e suas proteínas são processadas pelos mesmos mecanismos que processam as proteínas próprias do organismo. Portanto, nessas situações, as moléculas de classe I clássicas também podem se associar a antígenos virais, de microorganismos intracelulares ou fagocitados (Murphy *et al.*, 2012).

As moléculas MHC de classe I são montadas no retículo endoplasmático, onde ocorre a união das cadeias α e β em um processo que requer a ação de chaperonas, calreticulina, p57, calnexina e de uma proteína ligante muito importante na estabilização das moléculas HLA de

classe I recém-sintetizadas, denominada $\beta 2$ -M (Beta-2 microglobulina). Após a montagem, as moléculas HLA associam-se a um peptídeo e sua versão completa (complexo HLA + peptídeo) é exportada à superfície celular, podendo então interagir com os receptores de células T (Yang *et al.*, 1996; Klein e Sato, 2000a; Abbas *et al.*, 2007). A geração dos peptídeos que serão apresentados pelas moléculas de classe I envolve a participação de um complexo multiproteico conhecido como proteassomo. As proteínas destinadas a este complexo, de alta capacidade proteolítica, são marcadas com ubiquitinas e os peptídeos gerados no citosol são transportados ao retículo endoplasmático por meio do TAP. No retículo, esses peptídeos podem associar-se com moléculas HLA de classe I (Figura 2) (Yang *et al.*, 1996; York e Rock, 1996; Abbas *et al.*, 2007).

Na superfície celular, o complexo HLA + peptídeo interage com receptores TCR e CD8 dos linfócitos T citotóxicos (CTLs), ou ainda, com diversos outros receptores de células NK, que podem ser ativadores ou inibitórios. Na ocasião da interação entre MHC e TCR, uma série de proteínas estabilizam essa interação, permitindo com que ligante e receptor fiquem unidos tempo o suficiente para uma eventual ativação de um linfócito antígeno-específico (Gao *et al.*, 2000; Klein e Sato, 2000b) quando do encontro deste antígeno.

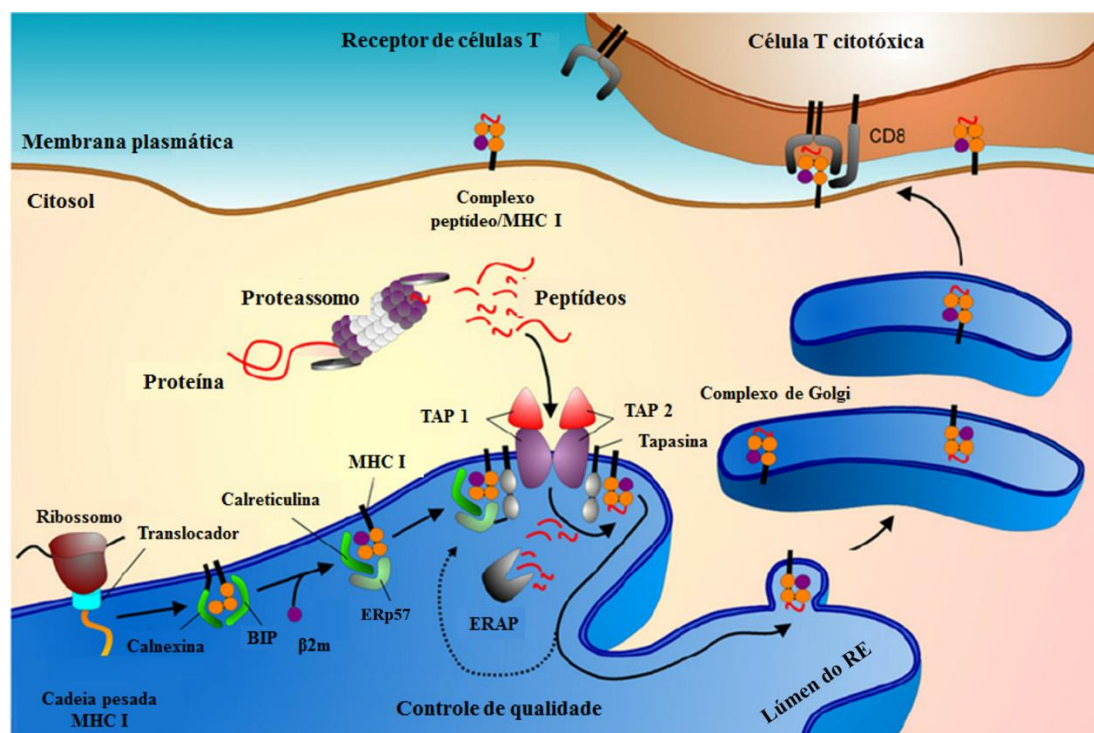


Figura 1. Via de processamento de antígenos das moléculas MHC de classe I. **Fonte:** Adaptado de SCHOLZ, C.; TAMPE, R. The intracellular antigen transport machinery TAP in adaptive immunity and virus escape mechanisms. J Bioenerg Biomembr, v. 37, n. 6, p. 509-15, Dec 2005.

Deve-se ressaltar que cada linfócito inativo presente na circulação possui um TCR com uma única especificidade, que foi determinada por um mecanismo de rearranjo gênico e que foi selecionado durante as fases de maturação dos linfócitos e seleção tímica. Em situações normais, esta seleção permite que apenas linfócitos que apresentam uma afinidade (porém baixa) pelas moléculas HLA do indivíduo e que não se associam fortemente com moléculas HLA do indivíduo carreando antígenos próprios são liberados na circulação. Desta forma, cada linfócito possui uma especificidade distinta, porém sua diversidade permite com que a gama de antígenos estranhos reconhecido por essas células do sistema imunitário seja muito grande (Lederberg, 1988; Abbas *et al.*, 2007; Lima *et al.*, 2016). Somente os linfócitos que encontraram um antígeno com o qual seu receptor é capaz de interagir poderão ser ativados a fim de proliferarem e se diferenciarem em células efetoras (Pardigon *et al.*, 1998). Além disso, a ativação das células T necessita também de sinais co-estimulatórios (Pardigon *et al.*, 1998). A presença ou ausência desses sinais pode acarretar na ativação do linfócito T ou a indução de tolerância (Sakaguchi *et al.*, 2008).

Embora a seleção negativa atue eliminando linfócitos auto-reativos durante seu desenvolvimento no timo (Kazansky, 2008), em algumas situações, as moléculas HLA apresentam antígenos próprios aos linfócitos, podendo levar ao desenvolvimento de doenças autoimunes (Caillat-Zucman, 2009; Lima *et al.*, 2016). O processo de seleção negativa consiste em um dos mecanismos mais importantes no processo de tolerância a auto-antígenos (Janeway Ca Jr, 2001; Abbas *et al.*, 2007). Por outro lado, a seleção positiva garante o desenvolvimento das células T cujos receptores interagem fracamente com moléculas MHC do hospedeiro. As células T maduras, cujos precursores sofreram seleção positiva no timo, agora são capazes de reconhecer antígenos não-próprios, apresentados pelas mesmas moléculas HLA do hospedeiro nas APCs nos tecidos periféricos (Janeway Ca Jr, 2001; Lima *et al.*, 2016).

As células NK pertencem ao grupo de células que compõe a imunidade inata, eliminando células tumorais ou infectadas por vírus sem a necessidade de sensibilização prévia. A redução ou ausência de moléculas HLA de classe I na superfície da célula alvo é um dos mecanismos que ativam a lise celular mediada pelas células NK. Contudo, as células NK apresentam uma variedade de receptores em sua superfície celular com capacidade de ligação com moléculas HLA de classe I, inibindo ou aumentando a citotoxicidade mediada por essas células (Davis *et al.*, 1999). O reconhecimento das moléculas HLA de classe I pelos receptores KIR (do inglês, *Killer-cell Immunoglobulin-like Receptors*) presente nas células NK é central no controle da função das células NK, bem como na modulação da ativação dos linfócitos T citotóxicos. Dentre

os receptores KIR, o *KIR3DL2* parece interagir com variantes do gene *HLA-A* conhecidas como *HLA-A3* e *HLA-A11*, contudo mais estudos ainda precisam ser realizados para a confirmação desses achados (Hansasuta *et al.*, 2004).

O gene *HLA-A* e suas funções

Considerando a alta variabilidade encontrada nos genes clássicos, é improvável que dois indivíduos sejam idênticos quanto a estes polimorfismos considerando-se todos os genes de classe I. Se por um lado esta variabilidade é interessante do ponto de vista populacional, pois aumenta a capacidade de apresentação antigênica, por outro lado, a eventual necessidade de realização de transplantes em humanos torna essa variabilidade problemática, uma vez que a compatibilização desses polimorfismos é crucial para uma boa aceitação do enxerto. A rejeição ao enxerto ocorre quando há um transplante de tecidos de um indivíduo para um receptor geneticamente diferente, levando a uma resposta imunológica direcionada ao órgão transplantado.

A montagem de respostas imunes adaptativas aos tecidos enxertados são o principal impedimento contra o sucesso de um transplante. A rejeição é causada por respostas imunes contra aloantígenos do órgão enxertado, que consistem em proteínas que variam entre indivíduos da mesma espécie e que passam a ser reconhecidas pelo sistema imune do hospedeiro como estranhas (resposta imune alorreativa). Os principais alvos no processo de compatibilização de órgãos para transplantes são as moléculas HLA de classe I clássicas e de classe II. Quando o doador e o receptor diferem quanto ao MHC, a resposta imune alorreativa é dirigida contra moléculas MHC não-próprias alogênicas presentes no órgão enxertado. Existem dois modos possíveis de apresentação dos aloantígenos para as células T do receptor. Na via conhecida como alorreconhecimento direto, as células apresentadoras de antígenos derivadas do doador podem migrar pelos linfonodos localizados próximos a região do enxerto e ativar as células T do hospedeiro. Desta forma, as células T alorreativas efetoras ativadas são direcionadas ao enxerto. O segundo mecanismo que pode desencadear o processo de alorreconhecimento envolve a captação de proteínas alogênicas por APCs do próprio receptor e sua posterior apresentação às células T, incluindo Treg(s), por moléculas MHC próprias, processo conhecido como alorreconhecimento indireto (Bradley, 1991; Sheldon e Poulton, 2006; Abbas *et al.*, 2007; Hansen *et al.*, 2008; Chinen e Buckley, 2010; Ayala Garcia *et al.*, 2012; Murphy *et al.*, 2012)

Alguns estudos já demonstraram que variações em um único aminoácido em uma dessas moléculas HLA clássicas pode levar a rejeição aguda em situações de transplante de medula óssea, pelos mecanismos expostos anteriormente. A análise dos segmentos das porções que acoplam peptídeos (domínios $\alpha 1$ e $\alpha 2$) justifica-se pela importância dessa região na compatibilização com receptores TCR que passaram pelo processo de seleção tímica (reconhecimento próprio e não-próprio) (Choo, 2007; Ayala Garcia *et al.*, 2012).

O gene *HLA-A*, alvo do presente estudo, apresenta atualmente um total de 5.018 alelos (Tabela 1), que codificam 3.398 proteínas distintas. Desta forma, o gene *HLA-A* é o segundo mais variável do genoma humano em termos do número de alelos diferentes, superado apenas pelo gene *HLA-B* (Tabela 1). Grande parte da variabilidade encontrada atualmente para esse gene concentra-se nos éxons 2 e 3, responsáveis pela codificação da fenda de ligação de peptídeos (domínios $\alpha 1$ e $\alpha 2$). Contudo, a variabilidade nesses segmentos justifica-se pela grande importância dada a esses domínios nos diversos contextos patológicos e fisiológicos. Em geral, estes são os segmentos avaliados pelos diversos laboratórios de histocompatibilidade que avaliam a compatibilidade entre doadores e receptores.

A estrutura da molécula *HLA-A* é de fundamental importância para o desempenho da sua função. Neste contexto o domínio $\alpha 3$, codificado pelo éxon 4, mostra-se importante para a manutenção da resposta imune adaptativa. A interação entre linfócitos TCD8 e seu alvo (moléculas HLA) requerem a interação física entre CD8 e porções conservadas no domínio $\alpha 3$ das moléculas HLA (Fayen *et al.*, 1995). A perda ou dificuldade de interação entre o domínio $\alpha 3$ e CD8 de linfócitos T reduz a sinalização e ativação de CTL (Wesley *et al.*, 1993). Contudo, polimorfismos no segmento $\alpha 3$ parecem influenciar não só a interação com o CD8, mas também com outros receptores. Um estudo de associação de variantes no domínio $\alpha 3$ do gene *HLA-B* encontrou uma associação entre um polimorfismo no códon 194 (acarreta troca de Isoleucina por Valina) e progressão lenta do vírus HIV, por alterar de forma significativa a interação da *HLA-B* com receptores LIR1 (ILT2/LILRB1/CD85j) (Grifoni *et al.*, 2013). Alterações nas sequências do éxon 5 (domínio transmembrana) do gene *HLA-A* podem modificar a conformação da molécula, uma vez que mutações nesse segmento podem originar moléculas solúveis (não aderidas a membrana plasmática) ou prejudicar sua ancoragem, com consequências na estabilidade dessas moléculas na superfície celular e no desempenho de uma resposta imunológica adequada. Alterações nas sequências dos éxons 6 e 7 podem levar a alterações na regulação do padrão de reciclagem e degradação de moléculas *HLA-A*. 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 prejudicar a apresentação de antígenos aos linfócitos, bem como inibir a ação das células NK, importantes para a resolução de determinados quadros patológicos (Schaefer *et al.*, 2008).

Embora a variabilidade nos éxons 2 e 3 seja bastante explorada, os demais segmentos do gene *HLA-A* (incluindo suas regiões regulatórias) são comumente negligenciados. Não está claro se a variabilidade nestes segmentos acompanha a variabilidade observada para os éxons 2 e 3, acarretando modificações funcionais como as descritas acima. Desta forma, a variabilidade genética do gene *HLA-A* (fora dos éxons 2 e 3) pode estar subestimada, apesar do elevado número de alelos registrado para o gene (tabela 1). Ainda, caso a variabilidade das regiões regulatórias do gene *HLA-A* seja tão elevada quanto a observada para os éxons 2 e 3, é provável que diferentes haplótipos encontrados nestes segmentos estejam relacionados com perfis transcricionais distintos. Polimorfismos presentes na região promotora poderiam influenciar a ligação de fatores de transcrição ou influenciar o perfil de metilação do gene. Polimorfismos presentes na região 3' não traduzida poderiam influenciar a ligação de microRNAs, modificando o perfil de transcrição de determinados haplótipos. Por fim, considerando a importância do gene *HLA-A* para a manutenção de uma reposta imunitária adaptativa, não está claro se o perfil de seleção balanceadora, que atua no segmento que codifica a fenda de ligação de peptídeo, também atua nos demais segmentos do gene.

Muitos polimorfismos dos genes clássicos foram associados com doenças autoimunes, degenerativas, infecciosas e susceptibilidade a determinadas neoplasias (Hughes e Yeager, 1998; Meyer e Thomson, 2001; Penn *et al.*, 2002; Bernatchez e Landry, 2003; Crispim, Mendes-Junior, *et al.*, 2008). Por exemplo, determinados alelos *HLA-B* relacionam-se com a progressão da infecção relacionada ao HIV, na qual se encontrou uma forte associação entre o alelo *HLA-B*58:01* e baixa carga viral e o alelo *HLA-B*58:02* e alta carga viral em uma população do Sul da África. Essas variantes *HLA-B* diferem-se por apenas três aminoácidos, destacando que pequenas diferenças entre moléculas HLA de classe I podem refletir em diferenças na especificidade de ligação de peptídeos (Kiepiela *et al.*, 2004). Um trabalho realizado com uma população de Mbeya, Tanzânia, demonstrou que o alelo *HLA-A*74:01* está associado com a proteção contra infecção pelo vírus HIV-1, porém progressão acelerada da doença (Koehler *et al.*, 2010). Os mecanismos de susceptibilidade a infecção e progressão não estão bem caracterizados e precisam ser melhor estudados, contudo a ausência ou presença de determinadas moléculas HLA podem estar relacionadas com diferentes capacidades de apresentação antigênica, influenciando assim os desfechos nos diversos quadros patológicos, a

exemplo das doenças infecciosas. O grupo de alelo *HLA-A*24* parece relacionar-se com a predisposição ao Diabetes mellitus do tipo 1 e os grupos *HLA-A*01*, *HLA-A*11* e *HLA-A*31* parecem desempenhar um papel protetor (Nejentsev *et al.*, 2007). Ainda, o alelo *HLA-A*03:01* parece estar relacionado com a predisposição (Fogdell-Hahn *et al.*, 2000) e *HLA-A*02:01* com a proteção à esclerose múltipla (Harbo *et al.*, 2004).

Em situações de câncer, onde há um descontrole nas vias que regulam a proliferação e morte celular, o desenvolvimento de uma célula tumoral acontece em função de uma sucessão de alterações de origem genética (Virani *et al.*, 2012; Ramsuran *et al.*, 2015) em genes críticos, entre eles aqueles responsáveis pelo controle do ciclo celular (Smyth *et al.*, 2001; Abbas *et al.*, 2007). As respostas imunológicas capazes de destruir células tumorais consistem em linfócitos T CD8 específicos a antígenos tumorais, células *NK* e macrófagos ativados. A ativação dos linfócitos T possui um papel essencial no desencadeamento da resposta imune, pois estas células podem desempenhar uma função de vigilância ao reconhecer e destruir células potencialmente malignas que expressam peptídeos derivados de proteínas celulares mutantes ou proteínas virais oncogênicas. Estes peptídeos são apresentados aos linfócitos T CD8+ associados a moléculas HLA de classe I, ou seja, peptídeos oriundos da célula tumoral são levados à superfície celular pelas moléculas HLA e o complexo HLA-peptídeo é reconhecido pelo receptor de células T. Neste contexto as moléculas HLA clássicas tem papel crucial (Cordon-Cardo *et al.*, 1991; Da Silva *et al.*, 2013). A ausência das moléculas HLA é ocorrência comum em transformações malignas (Abbas *et al.*, 2007).

Seleção natural e o complexo HLA

Darwin propôs em seus estudos que a seleção natural possui como princípio que características hereditárias que aumentam/promovem a aptidão do indivíduo tendem a aumentar em frequência nas populações com o passar do tempo e características que diminuem a aptidão tendem a ser eliminadas (Darwin, 1859). A manutenção ou eliminação dessas características operam por mecanismos distintos, podendo ser por meio da ação de seleção direcional (positiva ou negativa) ou de seleção balanceadora.

A seleção direcional positiva atua sobre uma mutação adaptativa aumentando sua frequência na população. A do tipo direcional negativa, também conhecida como purificadora, reduz a frequência da mesma ou elimina mutações deletérias na população. A seleção balanceadora é responsável pela manutenção dos polimorfismos na população e pode operar

por mecanismos diferentes (vantagem do heterozigoto, flutuação e seleção dependente de frequência). A vantagem do heterozigoto considera que alelos específicos são favorecidos em função da sua habilidade de promover proteção contra determinado patógeno, ou contrariamente quando resultam na susceptibilidade a um determinado patógeno (Penn *et al.*, 2002). A seleção dependente de frequência prediz que os valores de *fitness* não são fixados, mas são capazes de modificarem-se em proporção a frequência de alelos ou genótipos na população, *i.e* quando patógenos mutam determinada sequência em seu material genético, reduzem as chances de seus peptídeos serem apresentados pelas moléculas HLA que possuem elevada frequência, resultando em uma diminuição do *fitness* adaptativo, uma vez que a molécula HLA frequente não será capaz de reconhecer e apresentar esse antígeno. Na sequência, moléculas HLA de menor frequência capazes de reconhecer esse peptídeo aumentarão de frequência, até que uma nova onda de adaptação do patógeno aconteça e a frequência desse alelo diminua novamente, permitindo o aumento de outras moléculas menos frequentes (Meyer e Thomson, 2001). O mecanismo de flutuação considera que a intensidade de seleção natural varia conforme espaço e tempo. Se a frequência de patógenos muda constantemente, diferentes moléculas HLA serão selecionadas em tempos distintos, diferente do modelo dependente de frequência onde o valor do *fitness* muda conforme a frequência alélica, no modelo de flutuação o *fitness* varia em função da frequência de patógenos (Meyer e Thomson, 2001; Meyer *et al.*, 2018; Bitarello, 2011).

A expansão geográfica recente dos humanos, incluindo processos de colonização, expõe nossa espécie constantemente ao contato com diferentes agentes patogênicos. Patógenos potencialmente letais influenciaram sobremaneira as frequências alélicas em nosso genoma. Dessa forma, genes envolvidos com a imunidade apresentam assinaturas moleculares desse processo de adaptação da população humana a essas condições recentes (Ferrer-Admetlla *et al.*, 2008). Atualmente, destaca-se na literatura cerca de 91 genes que apresentam assinaturas de seleção positiva ou balanceadora, ou as duas formas, sendo que diversos desses estão envolvidos com a função imunológica e susceptibilidade a patógenos (Sabeti *et al.*, 2006). Por isso, estudos que esclareçam as assinaturas de seleção natural atuando nos genes HLA, podem trazer informações importantes acerca do passado evolutivo dessas moléculas, acrescentando informações importantes quanto a função desses genes.

A alta variabilidade associada aos genes HLA, principalmente nos éxons 2 e 3, que codificam a fenda de ligação ao peptídeo, levaram alguns autores a estudarem o perfil de seleção nesses genes (Garrigan e Hedrick, 2003; Meyer *et al.*, 2006; Buhler *et al.*, 2016). Uma das

hipóteses para a manutenção da alta diversidade alélica observada nesses segmentos parece ser a seleção balanceadora mediada por microorganismos, que acarreta em um aumento da heterozigotidade nesses segmentos. Diversos estudos têm buscado genes sob ação de seleção balanceadora (capaz de gerar um *fitness* adaptativo), regime que mantém diversas variantes na população com frequências intermediárias, fazendo com que cada alelo persista na população por tempo maior do que aqueles sob ação de seleção neutra (Andres *et al.*, 2009; Bitarello, 2011).

Neste caso, em uma população exposta a diversos patógenos, indivíduos heterozigotos são favorecidos quando comparados aos homozigotos, uma vez que o repertório de peptídeos apresentados pelos heterozigotos é maior (Doherty e Zinkernagel, 1975; Hughes e Yeager, 1998; Penn *et al.*, 2002; Bernatchez e Landry, 2003). Um estudo realizado por Solberg e colaboradores em 2008 considerando 497 populações, identificou um perfil de seleção balanceadora para o segmento codificador do gene *HLA-A*, sendo que as frequências alélicas observadas nas diferentes populações diferiram bastante. Estas diferenças podem ser justificadas pela ação de forças seletivas atuando nessas regiões, devido sua grande importância no processo de apresentação antigênica, sendo influenciado pelo contato com diversos microorganismos atuando nessas populações. No entanto, o perfil de seleção avaliado considerou basicamente polimorfismos presentes nos segmentos codificadores (Solberg *et al.*, 2008) e não considerou os segmentos relacionados com a regulação da expressão das moléculas HLA, tais como regiões promotoras e 3' não traduzidas. Além disso, padrões de migração, deriva genética e efeito fundador são outros fatores que podem levar a modificações nas frequências alélicas (Hartl e Clark, 2007).

Os genes do complexo HLA estão entre os alvos da seleção natural em humanos. Em especial destaca-se seleção balanceadora atuando nos genes clássicos, elevando a variabilidade nestes genes (Solberg *et al.*, 2008) e seleção purificadora atuando nos genes não clássicos. O mecanismo subjacente à baixa variabilidade detectada nas regiões codificadoras destes últimos (genes não clássicos) ainda não foi esclarecido, tornando-os virtualmente invariáveis quando comparados aos seus vizinhos *HLA-A*, *HLA-B* e *HLA-C* (Castelli *et al.*, 2011; Veiga-Castelli *et al.*, 2012; Felicio *et al.*, 2014). Também foram encontradas evidências de seleção balanceadora atuando nas regiões regulatórias 5' e 3' não traduzidas do gene *HLA-G*, mantendo uma alta heterozigose entre os haplótipos encontrados (Castelli *et al.*, 2011). Esta assinatura de seleção balanceadora atuando nas regiões regulatórias do *HLA-G*, porém não na região codificadora, foi encontrada em três diferentes populações Africanas: Senegal, Bêni e Congo (Gineau *et al.*,

2015). Embora seja consenso a atuação de seleção balanceadora na região que codifica a fenda de ligação ao peptídeo das moléculas clássicas (assim como nas regiões regulatórias do gene *HLA-G*), não está claro se esse mesmo perfil de seleção também ocorre nos demais segmentos dos genes clássicos ou em suas regiões regulatórias.

Levando em consideração a elevada variabilidade na fenda de ligação de peptídeos, alguns estudos discutem a possibilidade de Vantagem do Alelo Divergente (DAA, do inglês, *Divergent Allele Advantage*). De acordo com esta proposta, indivíduos heterozigotos que carregam alelos divergentes, teriam uma habilidade aumentada de apresentar uma maior gama de peptídeos comparado aos indivíduos carregando alelos próximos (Buhler *et al.*, 2016). No entanto, muitos estudos populacionais foram desenvolvidos nas últimas décadas e detectou-se uma observação contraditória a essas hipóteses: uma alta proporção de homozigotos em um ou mais genes HLA em populações pequenas e isoladas como as Ameríndias, Tailandesas aborígenes e ilhas do Pacífico, como resultado do rápido efeito da deriva genética durante suas histórias de migração (Black, 1992; Buhler *et al.*, 2016). Neste contexto, a homogeneidade genética dos Ameríndios poderia ser associada a uma alta taxa de mortalidade dessa população após a colonização européia, o que os deixariam susceptíveis a uma série de doenças (Black, 1992). Contudo, esse raciocínio não se confirma nessas populações, *i.e.*, populações tailandesas são extremamente populosas, apesar de serem consideradas as mais homogêneas quando levamos em consideração o *loci* HLA (Lin *et al.*, 2000).

Além disso, dados demográficos apontam que a população Ami e Paiwan exibem 43% e 76% de homogeneidade, respectivamente, para o *loci* *HLA-A*, com o alelo *HLA-A* 24:02* atingindo frequência de 64.0% e 86.3%, respectivamente (Buhler *et al.*, 2016). Os dados encontrados não indicam uma possibilidade de extinção dessa tribo, sugerindo então que outros mecanismos estão envolvidos para assegurar uma resposta imune adequada nessas populações (Buhler *et al.*, 2016). Quando outras espécies são analisadas, observa-se um decréscimo na expectativa de vida associado a homogeneidade das moléculas HLA. No entanto, isso não acontece para todas espécies, destacando o envolvimento de outras forças de seleção natural (Froeschke e Sommer, 2005; Huchard *et al.*, 2010; Worley *et al.*, 2010)

Recentemente um estudo demonstrou uma elevada frequência de variantes deletérias em genes não-HLA localizados no complexo MHC, sendo que a frequência dessas variantes aumenta conforme a proximidade com genes HLA, indicando uma dependência de ligação genética. Os resultados revelaram um custo indireto dessa diversidade genética mantida por meio da ação de seleção balanceadora, o que na maioria das vezes é visto como algo positivo,

trazendo implicações para o processo de recombinação, bem como para a epidemiologia de diversas doenças associadas ao complexo HLA (Lenz *et al.*, 2016).

Entendendo os testes de assinatura de seleção natural

Os testes para detecção das assinaturas de seleção natural baseiam-se em testes de desvio da neutralidade. Entende-se por neutralidade a condição em que o alvo da análise (*i.e* um determinado gene) sob observação não é afetado pela ação da seleção natural. Um ponto importante a se destacar é que os modelos neutros admitem a possibilidade de intensa ação purificadora contra mutações fortemente deletérias (Bitarello, 2011;2016). Assim que a teoria neutra foi proposta, permitiu-se o desenvolvimento de uma série de testes para a teoria nula da neutralidade, que considera que as mudanças evolutivas resultam exclusivamente por conta da ocorrência de mutações e deriva genética, quando a teoria neutra é rejeitada, considera-se hipóteses alternativas incluindo seleção positiva, purificadora ou balanceadora (Meyer *et al.*, 2018). Uma vez que a hipótese nula é baseada no modelo neutro da evolução, esses testes são chamados de testes de neutralidade. Entre eles, destacamos os testes baseados na ***no desequilíbrio de ligação e diversidade, testes baseados no padrão de distribuição de frequência alélica, testes de contraste de variação intra e inter espécies e polimorfismos compartilhados entre espécies*** (Meyer *et al.*, 2018).

Os testes baseados em **LD e diversidade** consideram que, sob forte ação de seleção positiva, as frequências aumentam rapidamente, reduzindo a diversidade de *locus* neutros ao redor. Para seleção balanceadora, onde duas ou mais variantes são mantidas, a variação genética local aumenta na região. Esses testes são excelentes para detectar seleção positiva recente e menos específicos para detectar seleção balanceadora antiga (Meyer *et al.*, 2018).

Os testes baseados nos **padrões de distribuição de frequência alélica** comparam a distribuição das frequências alélicas observadas com aquelas esperadas sob neutralidade. Incluem-se aqui o teste *D de Tajima*, *D de Fu e Li*, *F de Fu e Li*, dentre outros. Quando consideramos a seleção positiva, espera-se observar o aumento no número de variantes de baixa frequência, enquanto a seleção balanceadora aumenta o número de variantes com frequência intermediária acerca do sítio selecionado. Esses testes podem detectar a presença de seleção natural atuando ao longo de muitos períodos de forma comparativa, são mais eficientes quando comparados aos testes de *LD* e *Fst*, porém são influenciados por parâmetros demográficos (Fu e Li, 1993; Zhang *et al.*, 2015; Meyer *et al.*, 2018). No presente trabalho os testes descritos a

seguir foram alternativas empregadas para avaliar o desvio da neutralidade e determinar as assinaturas de seleção presentes no gene *HLA-A*, sendo eles (a) teste de Ewens-Watterson (Ewens, 1972; Watterson, 1978), que indica a possibilidade de neutralidade em determinado segmento, onde pode acontecer uma seleção direcionada (na qual identificamos um excesso de homozigose), ou atuação de seleção balanceadora, prevalecendo uma heterozigose maior que o esperado; (b) teste *D de Tajima* (Tajima, 1989), que compara as estimativas Θ derivado do número médio de diferenças observadas nas sequências (π) e o do número de sítios de segregação (S), utilizando um algoritmo de simulação de coalescência (Tajima, 1989; Excoffier e Lischer, 2010) e (c) teste de Fu e Li (Fu e Li, 1993), que compara a estimativa Θ derivado do número de sítios de segregação (S) e o número de mutações em singleton (alelos que aparecem apenas uma vez na amostra), bem como compara as diferenças nos pareamentos (π) e o número de mutações singleton. Este teste permite a avaliação de pontos de variação muito antigos (Fu e Li, 1993). O uso de diferentes testes de seleção evita resultados falso positivos, uma vez que variações nas frequências alélicas podem acontecer em função de efeitos demográficos, como deriva genética e efeito fundador (Gineau *et al.*, 2015).

Os testes contrastando **variação inter e intra espécies** incluem os testes de dN/dS (taxa de proporções de mutações não-sinônimas para sinônimas) e os testes *HKA* e *MK* (comparam polimorfismos e níveis de divergência). As assinaturas de seleção positiva e balanceadora são compatíveis com elevados valores de dN/dS , porém apresentam diferentes perfis quando os testes são aplicados. A seleção positiva apresenta baixo nível de polimorfismo comparado a divergência dentro de uma determinada espécie. A seleção balanceadora apresenta elevado polimorfismo em relação a divergência. Esses testes são excelentes para detectar seleção antiga, superior a 1 milhão de anos (Shapiro e Alm, 2008; Meyer *et al.*, 2018; Bitarello, 2011; 2016).

Referências

- ABBAS, A. K.; LICHTMAN, A. H.; PILLAI, S. **Cellular and molecular immunology**. 6th. Philadelphia: Saunders Elsevier, 2007. viii, 566 p.
- ANDRES, A. M. et al. Targets of balancing selection in the human genome. **Mol Biol Evol**, v. 26, n. 12, p. 2755-64, Dec 2009.
- AYALA GARCIA, M. A. et al. The major histocompatibility complex in transplantation. **J Transplant**, v. 2012, p. 842141, 2012.
- BARROW, A. D.; TROWSDALE, J. The extended human leukocyte receptor complex: diverse ways of modulating immune responses. **Immunol Rev**, v. 224, p. 98-123, Aug 2008.
- BERNATCHEZ, L.; LANDRY, C. MHC studies in nonmodel vertebrates: what have we learned about natural selection in 15 years? **J Evol Biol**, v. 16, n. 3, p. 363-77, May 2003.
- BITARELLO, B. D. **Natural selection on HLA genes: a molecular investigation of the location and timing of selection events**. 2011. (Mestrado). Instituto de Biociências, Universidade de São Paulo.
- BITARELLO, B. D. **Seleção balanceadora no genoma humano: relevância biológica e consequências deletérias**. 2016. (Doutorado). Instituto de Biociências, Universidade de São Paulo
- BLACK, F. L. Why did they die? **Science**, v. 258, n. 5089, p. 1739-40, Dec 11 1992.
- BRADLEY, B. A. The role of HLA matching in transplantation. **Immunol Lett**, v. 29, n. 1-2, p. 55-9, Jul 1991.
- BUHLER, S.; NUNES, J. M.; SANCHEZ-MAZAS, A. HLA class I molecular variation and peptide-binding properties suggest a model of joint divergent asymmetric selection. **Immunogenetics**, v. 68, n. 6-7, p. 401-416, Jul 2016.
- CAILLAT-ZUCMAN, S. Molecular mechanisms of HLA association with autoimmune diseases. **Tissue Antigens**, v. 73, n. 1, p. 1-8, Jan 2009.
- CAROSELLA, E. D. et al. HLA-G: from biology to clinical benefits. **Trends Immunol**, v. 29, n. 3, p. 125-32, Mar 2008.
- CASTELLI, E. C. et al. A comprehensive study of polymorphic sites along the HLA-G gene: implication for gene regulation and evolution. **Mol Biol Evol**, v. 28, n. 11, p. 3069-86, Nov 2011.
- CASTELLI, E. C. et al. Hla-mapper: An application to optimize the mapping of HLA sequences produced by massively parallel sequencing procedures. **Hum Immunol**, v. 79, n. 9, p. 678-684, Sep 2018.
- CHINEN, J.; BUCKLEY, R. H. Transplantation immunology: solid organ and bone marrow. **J Allergy Clin Immunol**, v. 125, n. 2 Suppl 2, p. S324-35, Feb 2010.
- CHOO, S. Y. The HLA system: genetics, immunology, clinical testing, and clinical implications. **Yonsei Med J**, v. 48, n. 1, p. 11-23, Feb 28 2007.

CORDON-CARDO, C. et al. Expression of HLA-A,B,C antigens on primary and metastatic tumor cell populations of human carcinomas. **Cancer Res**, v. 51, n. 23 Pt 1, p. 6372-80, Dec 1 1991.

CRISPIM, J. C. et al. Human leukocyte antigen-G expression after kidney transplantation is associated with a reduced incidence of rejection. **Transpl Immunol**, v. 18, n. 4, p. 361-7, Feb 2008.

CRISPIM, J. C. et al. HLA polymorphisms as incidence factor in the progression to end-stage renal disease in Brazilian patients awaiting kidney transplant. **Transplantation Proceedings**, v. 40, n. 5, p. 1333-1336, Jun 2008.

DA SILVA, G. B. et al. Expression of the Classical and Nonclassical HLA Molecules in Breast Cancer. **Int J Breast Cancer**, v. 2013, p. 250435, 2013.

DARWIN, C. **On the origin of species by means of natural selection, or, the preservation of favoured races in the struggle for life**. London: J. Murray, 1859. ix, 502 p., 1 folded leaf of plates

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. 1265-74, Apr 19 1999.

DAWKINS, R. et al. Genomics of the major histocompatibility complex: haplotypes, duplication, retroviruses and disease. **Immunol Rev**, v. 167, p. 275-304, Feb 1999.

DOHERTY, P. C.; ZINKERNAGEL, R. M. Enhanced immunological surveillance in mice heterozygous at the H-2 gene complex. **Nature**, v. 256, n. 5512, p. 50-2, Jul 3 1975.

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

DONG, G. et al. Insights into MHC class I peptide loading from the structure of the tapasin-ERp57 thiol oxidoreductase heterodimer. **Immunity**, v. 30, n. 1, p. 21-32, Jan 16 2009.

EISEN, H. N. et al. Promiscuous binding of extracellular peptides to cell surface class I MHC protein. **Proc Natl Acad Sci U S A**, v. 109, n. 12, p. 4580-5, Mar 20 2012.

EWENS, W. J. The sampling theory of selectively neutral alleles. **Theor Popul Biol**, v. 3, n. 1, p. 87-112, Mar 1972.

EXCOFFIER, L.; LISCHER, H. E. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. **Mol Ecol Resour**, v. 10, n. 3, p. 564-7, May 2010.

FALK, K. et al. Pool sequencing of natural HLA-DR, DQ, and DP ligands reveals detailed peptide motifs, constraints of processing, and general rules. **Immunogenetics**, v. 39, n. 4, p. 230-42, 1994.

FALK, K. et al. Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules. 1991. **J Immunol**, v. 177, n. 5, p. 2741-7, Sep 01 2006.

FAYEN, J. et al. Class I MHC alpha 3 domain can function as an independent structural unit to bind CD8 alpha. **Mol Immunol**, v. 32, n. 4, p. 267-75, Mar 1995.

- FELICIO, L. P. et al. Worldwide HLA-E nucleotide and haplotype variability reveals a conserved gene for coding and 3' untranslated regions. **Tissue Antigens**, v. 83, n. 2, p. 82-93, Feb 2014.
- FERRER-ADMETLLA, A. et al. Balancing selection is the main force shaping the evolution of innate immunity genes. **J Immunol**, v. 181, n. 2, p. 1315-22, Jul 15 2008.
- FLAJNIK, M. F.; KASAHARA, M. Origin and evolution of the adaptive immune system: genetic events and selective pressures. **Nat Rev Genet**, v. 11, n. 1, p. 47-59, Jan 2010.
- FOGDELL-HAHN, A. et al. Multiple sclerosis: a modifying influence of HLA class I genes in an HLA class II associated autoimmune disease. **Tissue Antigens**, v. 55, n. 2, p. 140-8, Feb 2000.
- FROESCHKE, G.; SOMMER, S. MHC class II DRB variability and parasite load in the striped mouse (*Rhabdomys pumilio*) in the Southern Kalahari. **Mol Biol Evol**, v. 22, n. 5, p. 1254-9, May 2005.
- FU, Y. X.; LI, W. H. Statistical tests of neutrality of mutations. **Genetics**, v. 133, n. 3, p. 693-709, Mar 1993.
- GAO, G. F. et al. Classical and nonclassical class I major histocompatibility complex molecules exhibit subtle conformational differences that affect binding to CD8alpha. **J Biol Chem**, v. 275, n. 20, p. 15232-8, May 19 2000.
- GARRIGAN, D.; HEDRICK, P. W. Perspective: detecting adaptive molecular polymorphism: lessons from the MHC. **Evolution**, v. 57, n. 8, p. 1707-22, Aug 2003.
- GAUDIERI, S. et al. Different evolutionary histories in two subgenomic regions of the major histocompatibility complex. **Genome Res**, v. 9, n. 6, p. 541-9, Jun 1999.
- GERAGHTY, D. E. et al. The HLA class I gene family includes at least six genes and twelve pseudogenes and gene fragments. **J Immunol**, v. 149, n. 6, p. 1934-46, Sep 15 1992.
- GINEAU, L. et al. Balancing immunity and tolerance: genetic footprint of natural selection in the transcriptional regulatory region of HLA-G. **Genes Immun**, v. 16, n. 1, p. 57-70, Jan-Feb 2015.
- GOLDBERG, A. C.; RIZZO, L. V. MHC structure and function - antigen presentation. Part 1. **Einstein (Sao Paulo)**, v. 13, n. 1, p. 153-6, Jan-Mar 2015.
- GRIFONI, A. et al. Role of HLA-B alpha-3 domain amino acid position 194 in HIV disease progression. **Mol Immunol**, v. 53, n. 4, p. 410-3, Apr 2013.
- HANSASUTA, P. et al. Recognition of HLA-A3 and HLA-A11 by KIR3DL2 is peptide-specific. **Eur J Immunol**, v. 34, n. 6, p. 1673-9, Jun 2004.
- HANSEN, J. A. et al. Genetics of allogeneic hematopoietic cell transplantation. Role of HLA matching, functional variation in immune response genes. **Immunol Res**, v. 41, n. 1, p. 56-78, 2008.
- HARADA, A. et al. Clinical implication of human leukocyte antigen (HLA)-F expression in breast cancer. **Pathol Int**, v. 65, n. 11, p. 569-74, Nov 2015.
- HARBO, H. F. et al. Genes in the HLA class I region may contribute to the HLA class II-associated genetic susceptibility to multiple sclerosis. **Tissue Antigens**, v. 63, n. 3, p. 237-47, Mar 2004.

- HARTL, D. L.; CLARK, A. G. **Principles of population genetics**. 4th. Sunderland, Mass.: Sinauer Associates, 2007. xv, 652 p.
- HUCHARD, E. et al. MHC, mate choice and heterozygote advantage in a wild social primate. **Mol Ecol**, v. 19, n. 12, p. 2545-61, Jun 1 2010.
- HUGHES, A. L.; YEAGER, M. Natural selection and the evolutionary history of major histocompatibility complex loci. **Front Biosci**, v. 3, p. d509-16, May 26 1998.
- JANEWAY CA JR, T. P., WALPORT M, ET AL. . Immunobiology: The Immune System in Health and Disease. **New York: Garland Science.**, v. 5th edition 2001.
- KIEPIELA, P. et al. Dominant influence of HLA-B in mediating the potential co-evolution of HIV and HLA. **Nature**, v. 432, n. 7018, p. 769-75, Dec 9 2004.
- KLEIN, J.; SATO, A. The HLA system. First of two parts. **N Engl J Med**, v. 343, n. 10, p. 702-9, Sep 7 2000a.
- KLEIN, J.; SATO, A. The HLA system. Second of two parts. **N Engl J Med**, v. 343, n. 11, p. 782-6, Sep 14 2000b.
- KOEHLER, R. N. et al. Class I HLA-A*7401 is associated with protection from HIV-1 acquisition and disease progression in Mbeya, Tanzania. **J Infect Dis**, v. 202, n. 10, p. 1562-6, Nov 15 2010.
- KULSKI, J. K.; GAUDIERI, S.; DAWKINS, R. L. Using alu J elements as molecular clocks to trace the evolutionary relationships between duplicated HLA class I genomic segments. **J Mol Evol**, v. 50, n. 6, p. 510-9, Jun 2000.
- LEDERBERG, J. Ontogeny of the clonal selection theory of antibody formation. Reflections on Darwin and Ehrlich. **Ann N Y Acad Sci**, v. 546, p. 175-82, 1988.
- LENZ, T. L. et al. Excess of Deleterious Mutations around HLA Genes Reveals Evolutionary Cost of Balancing Selection. **Mol Biol Evol**, v. 33, n. 10, p. 2555-64, Oct 2016.
- LEWONTIN, R. C.; KRAKAUER, J. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. **Genetics**, v. 74, n. 1, p. 175-95, May 1973.
- LIMA, T. H. et al. HLA-F coding and regulatory segments variability determined by massively parallel sequencing procedures in a Brazilian population sample. **Hum Immunol**, v. 77, n. 10, p. 841-53, Oct 2016.
- LIN, M. et al. Heterogeneity of Taiwan's indigenous population: possible relation to prehistoric Mongoloid dispersals. **Tissue Antigens**, v. 55, n. 1, p. 1-9, Jan 2000.
- MARTINEZ, J. et al. Expression of HLA-E molecules in the placental tissue of women infected with HIV-1 and uninfected women. **Placenta**, v. 49, p. 33-36, Jan 2017.
- MATZARAKI, V. et al. The MHC locus and genetic susceptibility to autoimmune and infectious diseases. **Genome Biol**, v. 18, n. 1, p. 76, Apr 27 2017.
- MEYER, D. et al. Signatures of demographic history and natural selection in the human major histocompatibility complex Loci. **Genetics**, v. 173, n. 4, p. 2121-42, Aug 2006.
- MEYER, D.; THOMSON, G. How selection shapes variation of the human major histocompatibility complex: a review. **Annals of Human Genetics**, v. 65, p. 1-26, Jan 2001.

- MEYER, D. et al. A genomic perspective on HLA evolution. **Immunogenetics**, v. 70, n. 1, p. 5-27, Jan 2018.
- MURPHY, K. et al. **Janeway's immunobiology**. 8th. New York: Garland Science, 2012. xix, 868 p.
- NEJENTSEV, S. et al. Localization of type 1 diabetes susceptibility to the MHC class I genes HLA-B and HLA-A. **Nature**, v. 450, n. 7171, p. 887-92, Dec 6 2007.
- PARDIGON, N. et al. Role of co-stimulation in CD8⁺ T cell activation. **Int Immunol**, v. 10, n. 5, p. 619-30, May 1998.
- PENN, D. J.; DAMJANOVICH, K.; POTTS, W. K. MHC heterozygosity confers a selective advantage against multiple-strain infections. **Proc Natl Acad Sci U S A**, v. 99, n. 17, p. 11260-4, Aug 20 2002.
- RAMSURAN, V. et al. Epigenetic regulation of differential HLA-A allelic expression levels. **Hum Mol Genet**, v. 24, n. 15, p. 4268-75, Aug 1 2015.
- ROBINSON, J. et al. The IPD-IMGT/HLA Database - New developments in reporting HLA variation. **Hum Immunol**, v. 77, n. 3, p. 233-237, Mar 2016.
- ROCK, K. L.; REITS, E.; NEEFJES, J. Present Yourself! By MHC Class I and MHC Class II Molecules. **Trends Immunol**, v. 37, n. 11, p. 724-737, Nov 2016.
- SABETI, P. C. et al. Positive natural selection in the human lineage. **Science**, v. 312, n. 5780, p. 1614-20, Jun 16 2006.
- SAKAGUCHI, S. et al. Regulatory T cells and immune tolerance. **Cell**, v. 133, n. 5, p. 775-87, May 30 2008.
- SCHAEFER, M. R. et al. A novel trafficking signal within the HLA-C cytoplasmic tail allows regulated expression upon differentiation of macrophages. **J Immunol**, v. 180, n. 12, p. 7804-17, Jun 15 2008.
- SCHMID, B. V.; KESMIR, C.; DE BOER, R. J. Quantifying how MHC polymorphism prevents pathogens from adapting to the antigen presentation pathway. **Epidemics**, v. 2, n. 3, p. 99-108, Sep 2010.
- SHAPIRO, B. J.; ALM, E. J. Comparing patterns of natural selection across species using selective signatures. **PLoS Genet**, v. 4, n. 2, p. e23, Feb 2008.
- SHELDON, S.; POULTON, K. HLA typing and its influence on organ transplantation. **Methods Mol Biol**, v. 333, p. 157-74, 2006.
- SHIINA, T. et al. The HLA genomic loci map: expression, interaction, diversity and disease. **J Hum Genet**, v. 54, n. 1, p. 15-39, Jan 2009.
- SMYTH, M. J.; GODFREY, D. I.; TRAPANI, J. A. A fresh look at tumor immunosurveillance and immunotherapy. **Nat Immunol**, v. 2, n. 4, p. 293-9, Apr 2001.
- SOLBERG, O. D. et al. Balancing selection and heterogeneity across the classical human leukocyte antigen loci: a meta-analytic review of 497 population studies. **Hum Immunol**, v. 69, n. 7, p. 443-64, Jul 2008.
- TAJIMA, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. **Genetics**, v. 123, n. 3, p. 585-95, Nov 1989.

- TOURAINÉ, J. L. et al. HLA class II antigens: structure, function, and expression in immunodeficiencies, autoimmune diseases, and allograft rejection. **Adv Nephrol Necker Hosp**, v. 18, p. 325-34, 1989.
- VEIGA-CASTELLI, L. C. et al. Non-classical HLA-E gene variability in Brazilians: a nearly invariable locus surrounded by the most variable genes in the human genome. **Tissue Antigens**, v. 79, n. 1, p. 15-24, Jan 2012.
- VIRANI, S. et al. Cancer epigenetics: a brief review. **ILAR J**, v. 53, n. 3-4, p. 359-69, 2012.
- WATTERSON, G. A. The homozygosity test of neutrality. **Genetics**, v. 88, n. 2, p. 405-17, Feb 1978.
- WESLEY, P. K. et al. The CD8 coreceptor interaction with the alpha 3 domain of HLA class I is critical to the differentiation of human cytotoxic T-lymphocytes specific for HLA-A2 and HLA-Cw4. **Hum Immunol**, v. 36, n. 3, p. 149-55, Mar 1993.
- WORLEY, K. et al. MHC heterozygosity and survival in red junglefowl. **Mol Ecol**, v. 19, n. 15, p. 3064-75, Aug 2010.
- YANG, Y.; SEMPE, P.; PETERSON, P. A. Molecular mechanisms of class I major histocompatibility complex antigen processing and presentation. **Immunol Res**, v. 15, n. 3, p. 208-33, 1996.
- YORK, I. A.; ROCK, K. L. Antigen processing and presentation by the class I major histocompatibility complex. **Annu Rev Immunol**, v. 14, p. 369-96, 1996.
- ZHANG, Q.; TYLER-SMITH, C.; LONG, Q. An extended Tajima's D neutrality test incorporating SNP calling and imputation uncertainties. **Stat Interface**, v. 8, n. 4, p. 447-456, Oct 1 2015.

Justificativa

O impacto dos polimorfismos das moléculas HLA no sucesso de transplantes de medula óssea e órgãos sólidos resultou no desenvolvimento de técnicas moleculares para a avaliação da variabilidade dessas moléculas e compatibilização entre doador e receptor (comumente conhecidas como “tipificação de HLA”). Esses métodos eram baseados originalmente em avaliações sorológicas por meio de anticorpos direcionados à grupos de moléculas HLA que compartilhavam resíduos polimórficos em comum. Com o advento das técnicas moleculares, métodos mais sofisticados e de detecção direta de variantes na molécula de DNA foram desenvolvidos, elevando assim a acurácia do procedimento para a tipificação dessas moléculas (Sheldon, 2006). A união dessas técnicas aos métodos sorológicos melhorou muito os resultados dos testes de histocompatibilidade realizados antes do transplante. No entanto, isso ainda é um desafio na clínica e na ciência, uma vez que esses genes apresentam alta similaridade entre si e um elevado polimorfismo. Ainda, como discutido anteriormente, os métodos de detecção da variabilidade dos genes HLA em geral avaliam somente alguns éxons do gene e negligenciam regiões intrônicas e regulatórias.

Neste sentido, o sequenciamento massivo paralelo (ou sequenciamento de nova geração) é uma técnica útil na caracterização da variabilidade de qualquer fragmento de DNA. No entanto, avaliar um gene HLA por NGS é um grande desafio, pois a similaridade de sequência entre os genes e o elevado polimorfismo dificultam o mapeamento correto das sequências geradas (Lima *et al.*, 2016). Por conta disto, poucos ensaios comerciais estão disponíveis para avaliação dos genes HLA por NGS, e estes ensaios não caracterizam as regiões regulatórias dos genes ou apenas reportam o nome dos alelos presentes em um indivíduo.

Considerando o papel chave da molécula de *HLA-A* nas respostas imunitárias adaptativas, importantes em diversos contextos fisiológicos e patológicos, destaca-se a importância de estudos que abordem a variabilidade completa do gene *HLA-A*, ou seja, considerando as regiões regulatórias e éxons que englobam os segmentos que codificam domínios diferentes da fenda de ligação a peptídeos. Desta forma, o estudo da variabilidade completa do gene *HLA-A* torna-se muito importante, uma vez que polimorfismos presentes nas regiões regulatórias podem afetar diretamente os níveis de expressão gênica e variações no domínio alfa-3, transmembrana e citoplasmático podem influenciar diretamente a função da molécula. Além disso a avaliação dos perfis de seleção para todos os segmentos do gene (regulatórias e codificadora) nos permitirá inferir as diferentes assinaturas de seleção associadas

ao gene *HLA-A*. A variabilidade observada em alguns segmentos do gene *HLA-A* pode ser mantida por seleção balanceadora, como já bem estabelecido para os éxons 2 e 3 (segmentos que codificam a fenda de ligação ao peptídeo), contudo estudos de assinatura de seleção natural nos demais segmentos do gene *HLA-A* não foram realizados (segmento completo, incluindo região promotora, codificadora e 3'NT), dados que somados ao estudo da variabilidade genética deste gene poderiam auxiliar bastante no entendimento da biologia do gene *HLA-A*.

Hipótese do estudo

A variabilidade da região promotora e 3' não traduzida do gene *HLA-A*, bem como dos segmentos que codificam o domínio $\alpha 3$, transmembrana e citoplasmático da molécula *HLA-A*, são tão variáveis quanto a região que codifica a fenda de ligação ao peptídeo (éxons 2 e 3). Esta variabilidade pode ser mantida por seleção balanceadora atuando nestes segmentos.

Objetivo geral

Avaliar a variabilidade completa do gene *HLA-A* por meio de sequenciamento paralelo massivo e inferir as assinaturas de seleção natural que atuam sobre este gene ao longo de toda sua extensão (região promotora, região codificadora e região 3' não traduzida).

Objetivos específicos

- Consolidar uma metodologia para avaliação do gene *HLA-A* por sequenciamento paralelo massivo, sem o uso de ferramentas comerciais;
- Caracterizar os pontos de variação e haplótipos encontrados para o gene *HLA-A* em uma amostra brasileira do estado de São Paulo;
- Elaborar um banco de dados de variação no gene *HLA-A* para futuros estudos de associação com doença (doenças autoimunes, infecciosas ou cânceres);
- Avaliar o perfil de seleção que atua em todos os segmentos do gene *HLA-A*;

Capítulo II

ARTIGO CIENTÍFICO

* O artigo científico foi publicado recentemente no periódico HLA journal e encontra-se disponível no link: <https://onlinelibrary.wiley.com/doi/abs/10.1111/tan.13474>, DOI: 10.1111/tan.13474. Em respeito as normas do periódico o texto anexado abaixo refere-se à versão enviada ao periódico sem qualquer revisão ou edição sugerida pela revista científica.

FULL TITLE: A COMPREHENSIVE STUDY OF POLYMORPHIC SITES ALONG THE ENTIRE *HLA-A* GENE: IMPLICATION FOR VARIANT TYPING, GENE REGULATION, AND EVOLUTION

SHORT TITLE: *HLA-A* PROMOTER, CODING AND 3'UTR HAPLOTYPES, AND EVOLUTIVE ASPECTS

Thálitta H. A. Lima^{1,2}, Iane O. P. Porto^{1,3}, Michelle A. Paz^{1,3}, Andreia S. Souza^{1,2}, Luciana C. Veiga-Castelli⁴, Maria Luiza G. Oliveira⁴, Eduardo A. Donadi⁵, Diogo Meyer⁶, Audrey Sabbagh⁷, Celso T. Mendes-Junior⁸, Erick C. Castelli^{1,2,3}

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

Contact

Erick C. Castelli

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

Conflicts of Interest: none

Abstract

HLA-A is the second most polymorphic loci at the Human Leucocyte Antigen (HLA) complex and encodes a key molecule for antigen presentation and *NK* cell modulation. Many studies have evaluated *HLA-A* variability in worldwide populations focusing mainly on exons, and the regulatory segments are poorly characterized. *HLA-A* variability is particularly high at the segment encoding the peptide-binding groove (exons 2 and 3), which is probably related to their antigen presentation function and balancing selection acting at these segments. Here we evaluate the variability and haplotypes of the *HLA-A* gene considering a continuous segment encompassing the extended promoter (1.5 Kb upstream the first translated ATG), all exons and introns, and the entire 3' untranslated region, by using massively parallel sequencing. To achieve this goal, we used a freely available bioinformatics workflow that optimizes read mapping for *HLA* genes and defined haplotypes using either the phase among variable sites observed directly in sequencing data or probabilistic models. The *HLA-A* variability detected in a highly admixed population sample from Brazil demonstrates that the *HLA-A* regulatory segments present few, but divergent haplotypes. The regulatory segments are in close association with the coding alleles. Introns segments are also quite variable. In addition, patterns of molecular diversity suggest that the promoter, in addition to the coding region, might be under the same selective regime, but a different scenario arises when it comes to exon 4 and 3'UTR segment.

Keywords: balancing selection, haplotypes, *HLA-A*, natural selection, Next Generation Sequencing - NGS, polymorphisms.

Introduction

The human leucocyte antigen (HLA) gene cluster includes some of the most polymorphic *loci* of the human genome. This gene complex is located at the short arm of chromosome 6, at 6p21.3, and most of the HLA genes are involved in antigen presentation, modulation of the immune response or both ¹⁻⁴. *HLA-A* is a classical class I gene that encodes a key molecule for self and non-self antigen presentation ⁵. *HLA-B* and *HLA-C* (other classical class I loci) and HLA class II *loci* (*HLA-DQA1*, *DQB1*, *HLA-DRA*, *HLA-DRB1*) encode molecules that are responsible for displaying self, foreign and damaged peptides on the cell surface, which in turn may be recognized by the T Cell Receptor (TCR) from CD8+ and CD4+ T cells, respectively, representing critical steps to trigger the adaptive immune responses ^{6,7}.

Some HLA alleles are associated with the outcome of autoimmune and infectious diseases as well as cancer ⁸⁻¹³. For instance, the *HLA-A*74:01* allele is associated with protection against HIV-1 infection and slow progression to AIDS (acquired immunodeficiency syndrome) ¹⁴. Although the mechanism underlying the association between HLA alleles and infectious diseases has not been completely elucidated, the ability of an HLA molecule to bind peptides depends on the structure of the peptide-binding groove, and of the presented pathogen peptide ^{15,16}. HLA genes also strongly influence grafted tissue acceptance and graft outcome, depending on the compatibility between donor and recipient ¹⁷⁻¹⁹. Achieving this compatibility can be challenging since the classical HLA class I genes (*HLA-A*, *HLA-B*, and *HLA-C*) are among the most variable ones of the human genome. Moreover, in the reproductive context, there is evidence that mate choice and pregnancy outcome may be influenced by HLA genotypes ²⁰⁻²³.

HLA-A, *HLA-B*, and *HLA-C* add up to 12,631 different alleles according to the last ImmunoGenetics Database (IPD-IMGT/HLA), version 3.30.0. The *HLA-A* locus itself has 3,997 different DNA sequence-level alleles, which encode 2,792 different *HLA-A* full-length proteins. Natural selection and genetic drift have shaped HLA worldwide sequence variability, allele diversity, and distribution of allele frequencies by different mechanisms ²⁴. It is well established that different HLA classical class I molecules can present a different subset of antigens ^{15,25}. Pathogen-driven selection may explain such high polymorphism observed for the classical class I molecules ^{24,26-28}. As a consequence of this balancing selection, HLA variability is enhanced by keeping alleles longer than expected under neutrality, which allows

the accumulation of new mutations in these variants, thus increasing HLA inter-allelic variability²⁹⁻³¹. The variability at *HLA-A*, *HLA-B*, and *HLA-C* is particularly high at the DNA segment encoding the peptide groove (exons 2 and 3)^{6,32}. Variability can also be maintained by selection that is variable over space and/or time, as well as by selection that is frequency-dependent³³. Most of the known *HLA-A* alleles differ by in exons 2 and 3. Since these two segments represent the minimum requirement to characterize an HLA classical class I allele, it is not clear if this high variability is restricted to exons 2 and 3, or whether other *HLA-A* segments follow the pattern observed at these two exons. In addition, the same allele group may present different promoter or 3'UTR segments, as observed for the *HLA-G* gene³⁴.

Studies addressing the complete variability of HLA genes are scarce and usually do not include regulatory and intronic segments. Ramsuran explored this variability extending nearly 2kb upstream from the transcription start site of HLA classical class I genes (*HLA-A*, *HLA-B*, and *HLA-C*) for few cell lines typed for the frequent HLA classical class I alleles³⁵. The promoter variability of 19 *HLA-A* alleles detected on homozygous cell lines was assessed through Sanger sequencing. Despite the presence of many variable sites at the *HLA-A* promoter, the core promoter was quite conserved. Ramsuran demonstrated a great allele-specific expression level variation, but there was no clear relationship between sequence similarities and expression levels³⁵. However, it is not clear whether such low variability at the core promoter would still be detected when evaluating larger population samples. This same research group demonstrated that the *HLA-A* mRNA expression levels correlate with allelic lineage and are continuously distributed, and this in part to differential DNA methylation levels mainly at the promoter segment³⁶. Considering that there is no population-level study to date that has addressed the variability at *HLA-A* for the entire gene and including regulatory segments, here we evaluated the nucleotide variability and haplotype structure of the *HLA-A* gene in a highly admixed Brazilian population sample, considering a continuous segment encompassing the extended promoter (1.5 kb upstream the first translated ATG), all exons and introns, and the entire 3' untranslated region (3'UTR), using massively parallel sequencing procedures. We observed that the regulatory segments present few, but divergent promoter and 3'UTR haplotypes, which are in close association with the coding regions. The analysis of this population-level dataset indicates that the promoter, together with the coding regions, might be under selection.

Methods

Population sample

HLA-A variability was surveyed in 408 unrelated samples (82.84% female, a mean age of 31.3 years old) from the state of São Paulo, Southeastern Brazil. All participants signed an informed consent before blood withdraw and this study protocol was reviewed and approved by the Human Research Ethics Committee from the School of Medicine (Unesp/Brazil). From those, 106 randomly selected samples had their ancestry evaluated by using Ancestry Informative Markers (AIMs). Ancestry composition of the population sample was 75.9% European, 16.7% African, and 7.3% Amerindian/Asian. These proportions are expected for a Brazilian sample from the São Paulo State. DNA was obtained by a salting out procedure, quantified using Qubit Broad Range Assays (Thermo Fisher Scientific Inc., Waltham, MA) and normalized to a final concentration of 50 ng/μL.

***HLA-A* amplification**

HLA-A amplification was performed as a single amplicon of 5,883 nucleotides (excluding primers), encompassing the segment between nucleotides 29,908,285 and 29,914,168 based on the sequence available for chromosome 6 (human genome assembly hg19). Amplification primers were HAPR.F1 (5'-GCTACTTGGGACTGAGACAG-3') and HAUT.R1 (5'-GCCTTATGGGTACAATTTTCC-3') in a final volume of 50 μL, containing 2.0 mM of dNTPs (Invitrogen, EUA), 15 pmol of each primer, 1.25 units of DNA polymerase (PrimeStar GXL, Takara) and 1X the PCR buffer supplied with the DNA polymerase. Cycling conditions were: 30 cycles of 98°C for 10 seconds, 60°C for 15 seconds and 68°C for 6 minutes. Amplification was evaluated on 1% agarose gel stained with GelRed® (Biotium™, Hayward, USA). Prior to library preparation, amplicons were purified using Illustra ExoProStar (GE Healthcare), quantified using Qubit dsDNA High-Sensitivity Assays (ThermoFisher Scientific, USA) and normalized to the proper concentration. The sequencing was performed in the same assay as for other *HLA* class I genes, with the amplicons pooled together. A bioinformatics approach (presented below) allowed the discrimination of sequences from each locus.

Library preparation and sequencing

Amplicons were sequenced using the MiSeq Platform (Illumina, Inc.). Sequencing libraries were prepared using Nextera XT Library Preparation Kit, multiplexed with the Nextera XT Index Kit (both from Illumina, Inc.). Libraries were quantified by qPCR using Kapa (Kapa

Biosystems, Wilmington, USA), and normalized to the recommended concentration. The library fragmentation pattern was assessed with High-Sensitivity DNA BioAnalyser chips (Agilent Technologies, CA, USA). Sequencing was performed using MiSeq Reagent Kit (V2, 500 cycles, 2 x 250bp), as recommended by Illumina Inc. The paired-end sequencing data was evaluated using freely available and locally developed software, as described below. Here we aimed to achieve a minimum sequencing depth of about 100 for the entire *HLA-A* segment and, with the exception of part of intron 2 and part of intron 3 (discussed hereinafter), average sequencing depth was over 100 for all samples.

Raw data processing (mapping)

Prior to mapping, the sequences (reads) from all individuals were trimmed on both ends for primer and adapter sequences and for low-quality sequences using seqtk trimfq (-q 0.01) (<https://github.com/lh3/seqtk>). A major challenge when performing HLA sequencing by NGS procedures is to obtain a reliable mapping of the reads ^{34,37-40}, a problem aggravated for HLA classical class I genes because of the high level of sequence similarity among them, which may lead to misalignment (mapping to the wrong reference sequence). Moreover, the high level of polymorphism may lead to alignment failure. Both these issues contribute to errors in genotype calls. In addition, to avoid unnecessary costs, usually, more than one HLA gene is sequenced at the same time (as was done here), magnifying the read mapping issues.

To circumvent these issues, we used *hla-mapper* (version 2.0, database version 002.1, function *dna* using default parameters) to map the *HLA-A* derived reads to a reference genome (version hg19), and record the results in an *HLA-A* specific alignment file (BAM format) ⁴⁰. This software is available at www.castelli-lab.net/apps/hla-mapper and was previously introduced when evaluating the *HLA-E*, *HLA-F*, and *HLA-G* genes in different population samples ^{34,37,38,40}.

Genotype calling and processing

The Genome Analysis Toolkit (GATK, version 3.6) HaplotypeCaller (GVCF model with default parameters) was used to infer SNP genotypes ⁴¹⁻⁴³, and chromosome 6 from the hg19 genome draft as a reference (the same used by *hla-mapper*). Next, GATK GenotypeGVCFs combined all .G.VCF files into a multi-sample VCF (variant call format) file. At this moment, each variable site was annotated using dbSNP version 146. The VCF file was then processed by *vcfx*, function *checkpl* (with the minimum genotype likelihood set to

99.99%), which guarantees that only high-quality genotypes are retained for the imputation step. *vcfx* is available at <http://www.castelli-lab.net/apps.vcfx>. This approach was mandatory because Nextera kits produce bias in the fragmentation pattern, so that part of intron 2 and part of intron 3 are usually underrepresented, and these segments might be prone to genotyping errors, especially the calling of untrue homozygous genotypes in segments with low sequencing depth. Thus, missing alleles replaced alleles presenting low likelihood, and they were further imputed as discussed later. The use of *vcfx* *checkpl* introduced missing alleles in 0.66% of genotypes, considering the entire *HLA-A* segment.

The GATK ReadBackedPhasing inferred the phase among variable sites, with a minimal Phase Quality Threshold of 2000 and minimum base quality set to 20. This ensures that only alleles from closely located variable sites (present in a same read or library fragment) are directly phased. Considering that variable sites may sometimes be quite far from each other, not all variable sites were phased in this way. Considering all samples together, 77.65% of the heterozygous sites were directly phased using GATK. For the remaining 22.35%, we used the PHASE algorithm^{44,45}, which also imputed the 0.66% of missing alleles observed after the *vcfx* *checkpl* treatment. Most of the variants that were not directly phased by the ReadBackedPhasing algorithm represent indels or multiallelic *loci*, which are not supported by the aforementioned algorithm.

To perform the PHASE analysis, we converted the partially GATK-phased VCF file (obtained with the ReadBackedPhasing algorithm) into an input file for PHASE and into an accessory file containing the known phases between variable sites. In many cases, there were blocks with known phases, but with unknown phase between them. The PHASE algorithm was used to infer haplotypes, requiring several different runs. For each run, we considered a block with known phases, and a final run was also performed disregarding the GATK information. Samples presenting the same pair of haplotypes with the highest *P* value for all runs were maintained, as long as the pair of haplotypes was in agreement with the known phases detected by GATK (which happened for all samples, with no exception). The most probable haplotype pair of each sample was then considered to create a phased VCF file. All singletons were removed to perform the haplotyping step using the PHASE algorithm because they are by definition ambiguous. Singletons were only included in the phased VCF file manually, if they met the following conditions: (a) there were no missing alleles for any sample regarding that particular position, and (b) there was a clear relationship between the singleton and a close heterozygous site visualized using the BAM file of that sample or it was the only heterozygous

site detected on that particular sample. This haplotyping approach is described in details at <https://github.com/erickcastelli/phase-readbackedphasing>.

This phased VCF file was converted into complete *HLA-A* sequences using the hg19 chromosome 6 as a draft and replacing the correct nucleotide at each position, thus creating two sequences per sample (one for each chromosome), by using application *vcfx* function *fasta* (www.castelli-lab.net/apps/vcfx). This same approach was used to create two coding sequences (CDS, only exons, from the ATG to the stop codon) for each sample. Then, we used a local BLAST server with databases containing all known HLA alleles described so far, either for genomic sequences and CDS sequences, downloaded from the IPD-IMGT/HLA database ³², version 3.30.0, to define the closest known *HLA-A* coding allele for each different sequence observed. In cases in which the sequence observed was identical to one already described, this sequence was named as the known allele. In cases in which the sequence was not identical to anyone already reported, the closest allele was identified, and its name was accompanied with observed differences using the official IMGT/HLA alignment as a reference. For the promoter and 3'UTR segments, haplotypes were named manually as described later. The frequency of each *HLA-A* haplotype (for each segment) was computed by the direct counting method. Each CDS sequence was translated using *EMBOSS transeq* and the protein molecules were counted directly and named according to the IMGT/HLA database. The *HLA-A* alleles called by the bioinformatics approach described above were compared with the ones obtained by using the NGSEngine (from GenDX).

Diversity indexes, neutrality deviation tests and other analyses

Nucleotide diversity and gene diversity at specific segments, Tajima's *D*, and Ewens-Watterson were calculated using Arlequin 3.5.1.2, using aligned sequences because of the presence of indels ⁴⁶. Tajima's *D* *P*-value was computed by comparing the statistic to a distribution obtained from 5,000 simulations, and represents the probability of obtaining a simulated Tajima's *D* larger than that observed ⁴⁶. The dn/ds ratio was calculated using Mega version 7.0.20 ⁴⁷, considering one copy of each Brazilian sequence, or one copy of each IMGT/HLA sequence. The network was created using PopArt (<http://popart.otago.ac.nz>) and the Median-Joining algorithm, and further processed using Inkscape. The nucleotide diversity plot was computed by using VariScan ⁴⁸ with a sliding window of 150 bp, and a step size of 1. The linkage disequilibrium (LD) plot was generated using Haploview (standard color scheme,

D'/LOD)⁴⁹. Tagging was performed using the same software with the aggressive tagging mode (two-marker haplotypes), with $r^2 \geq 0.8$ and LOD threshold set to 3.0.

Results

In order to evaluate the phylogenetic/evolutionary relationship among the extended *HLA-A* haplotypes, we attempted to identify different lineages using quantitative criteria. We considered as lineages groups of alleles separated by less than 25 differences, whereas branches with a higher number of differences were used to delimit distinct lineages. In this way, we identified 10 different lineages (Figures 1 and S1) that were named as “Lineage 01 to Lineage 10”, starting from the top right part to the lower left (Figure S1). The single occurrence of a unique promoter sequence, the *A*80:01:01:02* coding allele, and a unique 3'UTR sequence was then named as lineage 10.

Since almost all lineages present a single 3'UTR haplotype, we named the 3'UTR haplotypes following the lineage name, as *HLA-A*U01* to *HLA-A*U10*. The naming of the full promoter haplotypes followed the same rule, in which a single haplotype or few closely related ones were detected for each *HLA-A* lineage. Thus, they were also named following the lineage name, in which all the promoters detected within lineage 01 were named as *HLA-A*P01*, and the specific promoters were named consecutively (e.g., *HLA-A*P01-01* and *HLA-A*P01-02*, etc.). A version of this full *HLA-A* network presenting the identification of each of the extended haplotypes, including the promoter names, coding alleles and 3'UTR haplotype names, is available in Figure S1.

We identified the following properties for the *HLA-A* lineages: (a) coding sequences showed high similarity, usually belonging to the same *HLA-A* allele group, or representing alleles that were previously recognized as a single *HLA-A* antigen; (b) a single 3' UTR sequence or haplotype, with the exception of lineage 4 which has two different but closely related 3'UTR haplotypes; and (c) few different but closely related promoter haplotypes.

The pattern depicted in Figures 1 and S1 is also apparent from the high LD along the entire *HLA-A* segment (Figure S2). The entire region behaves essentially as a single segregation block, from the promoter region (1,500 nucleotides upstream the first translated ATG) to the 3'UTR. Because of this strong LD, the 453 variable sites detected along the entire *HLA-A* segment were distributed in only 94 extended haplotypes, with many variable sites in complete ($D' = 1$, $r^2 < 1$) or perfect ($D' = 1$, $r^2 = 1$) LD. The complete list of all variable sites detected,

including their frequencies and their IPD-IMGT/HLA relative position, is available as a supplementary file (Table S1). Because of this LD, it is possible to infer the *HLA-A* extended haplotype by evaluating a smaller subset of variable sites. For instance, only 58 variable sites are necessary to capture all the polymorphic sites (MAF > 1%) at the promoter, exons, introns, and 3'UTR ($r^2 \geq 0.8$, multi-allelic loci not included), while only 25 are necessary to infer the promoter and 3'UTR haplotypes (and thus the *HLA-A* lineage).

***HLA-A* coding region haplotypes**

When we investigated the *HLA-A* variability between nucleotides 29,910,031 (IMGT/HLA position -300, considered as 5'UTR) and 29,913,562 (IMGT/HLA position +3,203, considered as 3'UTR), using chromosome 6 from the hg19 human genome draft as reference, which corresponds to the segment tracked by the IMGT/HLA database, 90 different *HLA-A* coding sequences were detected (Table 1). Among them, 54 are identical to previously described sequences, and the remaining 36 are new *HLA-A* coding sequences not yet described in the IPD-IMGT/HLA database version 3.30. Although eight of these new alleles were detected more than once, the cumulative population frequency of the new *HLA-A* alleles was only 6.69%, the remaining 93.31% representing previously reported *HLA-A* sequences. In addition, many of these new *HLA-A* alleles encode known *HLA-A* protein molecules, and only four new sequences encode new *HLA-A* molecules (1.22%) (Table S2).

The most frequent coding sequences detected in this Brazilian sample are *A*02:01:01:01* (19.49%), followed by *A*24:02:01:01* (7.97%), *A*03:01:01:01* (7.48%) and *A*01:01:01:01* (7.26%), and 21 alleles among the 90 detected (23.3%) present frequencies higher than 1% (Table 1). Some rare *HLA-A* alleles previously described in the IPD-IMGT/HLA database were detected in the present sample, including alleles *A*29:02:01:04*, *A*02:01:01:08*, *A*31:01:02:04*, *A*68:17*, *A*02:20:01* and *A*80:01:01:02*, and others. Altogether, we detected 44 different full-length protein molecules (Table S2).

***HLA-A* promoter haplotypes**

The nucleotide variability of the extended *HLA-A* promoter segment was surveyed between nucleotide chr6:29,908,726 (IMGT/HLA position -1605) and chr6:29,910,330 (IMGT/HLA position -1). Only the last 300 bases out of these 1,605 nucleotides are included in the IMGT/HLA database, and were also included as part of the coding sequence described in the previous subsection. We detected 100 variable sites within this segment, arranged into

25 different full promoter haplotypes (Figure S3). The low number of haplotypes contrasts with the high number of variable sites, a consequence of the high LD detected along the promoter segment, with many variable sites in perfect LD with each other (Figure S3).

As mentioned earlier, these full promoter haplotypes were named according to the main *HLA-A* lineage they belong to (Figures 1 and S1), and 14 out of 25 haplotypes were at a frequency higher than 1% (Table 2). The most frequent promoter haplotype was *HLA-A*P06-01*, which is associated with several coding alleles from the *HLA-A*02* group (Table 2). Each main *HLA-A* lineage presents only one or at most a few closely related promoter haplotypes, as indicated in Figures 1 and S1.

***HLA-A* 3' untranslated region haplotypes**

The *HLA-A* variability at the last *HLA-A* exon, between nucleotide chr6:29,913,228 (IMGT/HLA position +2899) and chr6:29,913,661 (IMGT/HLA position +3332) was considered as encompassing the *HLA-A* 3'UTR. Only 70% of this segment is in the IMGT/HLA database, and was also included as part of the coding sequence described in a former subsection. We detected 21 variable sites along this segment, arranged into 11 different haplotypes (Table 3).

The majority of *HLA-A* 3' UTR haplotypes (9 out of 11) present a frequency higher than 5% in the present series (Table 3). The most frequent 3'UTR haplotype detected in this sample was *HLA-A*U06*, which, like the P06 promoter, is associated with several coding alleles from the *HLA-A*02* group (Table 3 and Figure S1). Each main *HLA-A* lineage presented only one 3'UTR haplotype (with the exception of lineage 4), as indicated in Figures 1 and S1.

Nucleotide diversity and natural selection signatures at the *HLA-A* locus

The nucleotide diversity at the *HLA-A* locus, considering the extended promoter, coding and 3'UTR segments, was 0.02467 (Table 4). Considering each segment separately, nucleotide diversity ranged from 0.00753 at the 3'UTR to 0.04731 at intron 1. Intron segments were on average as variable as the exons. The promoter segment presented a nucleotide diversity of 0.01716.

When nucleotide diversity was evaluated across the *HLA-A* genomic sequence (Figure 2), the highest values were detected at the segments encompassing (a) the promoter around position -900, (b) exon 1, (c) exon 2, (d) the first half of exon 3, (e) the end of intron 3, (f) exon 5 and (g) exon 7. A somewhat different scenario is observed when only the mRNA sequence is

taken into account (Figure 2). The nucleotide diversity peaks observed at the genomic sequence for exons 1, 5 and 7 (Figure 2 upper panel) are no longer evident when evaluating the mRNA (Figure 2 middle panel), since they are related with the highly polymorphic introns surrounding these exons (Table 4).

Considering the genomic sequence, the lowest values were observed for exon 4 and the 3'UTR segment, especially for the latter. Considering the mRNA sequence, the lowest values were observed for the segment after exon 6, mainly at the 3'UTR. Despite the relative conservation at the core promoter, nucleotide diversity at the promoter segment rises above 3% around position -900 (Figure 2).

Neutrality tests were applied to detect signatures of natural selection along the *HLA-A* locus. We detected positive Tajima's *D* for all segments of *HLA-A* when evaluated separately (Table 4). Significant Tajima's *D* were detected for exons 2, 3, 4, introns 1, 3, 5, and the promoter segment, as well as for the full segment, and the CDS, and borderline significant values were observed for intron 2 and exon 5. As previously observed for exons 2 and 3^{50,51}, these positive Tajima's *D* values are indicative of balancing selection. In addition, significant Ewens-Watterson values were also detected for exons 2, 5 and the 3'UTR segment; except for the full sequence, all observed *F* was lower than the expected *F*.

An excess of synonymous mutations was detected at exon 4 ($ds-dn = +2.721$, $P=0.0037$, Table 5), suggesting the action of purifying selection for a conservation of this coding sequence. This is consistent with the lower nucleotide diversity detected at this exon (Figure 2). A similar pattern was observed when considering the sequence data from all the known *HLA-A* alleles described in the IMGT/HLA database (data not presented here). Most of the variable sites at this segment are quite frequent, with 76% of them presenting MAF higher than 5%, which may explain the high Tajima's *D* value detected for exon 4 (Table 4).

Discussion

In the present study, we use massively parallel sequencing and a bioinformatics workflow to quantify the variability of the entire *HLA-A* genomic region, including regulatory segments and introns. Our methods minimize the costs to obtain individual genotypes and haplotypes, because (a) they rely on either publicly available software⁴⁰, (b) they use locally developed primers and common library preparation kits, without any additional costs beyond the sequencing and amplification reagents, and (c) they are customizable, allowing the

evaluation of other targets together with the class I HLA loci ⁴⁰. In addition, this methodology evaluates the whole gene, including regulatory segments, instead of focusing only on the coding sequence, which are the main targets of most HLA typing methodologies. We were thus able to generate data on *HLA-A* variability for full-length haplotypes including exons, introns, the entire 3'UTR and at least 1,605 nucleotides upstream the first translated ATG defining the extended promoter segment.

The methodology presented here relies on the *hla-mapper* software, locally developed to minimize cross mapping between similar HLA sequences. This software applies a series of filters to assign each pair of sequences (from a paired-end sequencing) to their proper locus, based on the similarity between the pair of sequences and known HLA sequences described in the IPD-IMGT/HLA database and a local database of known sequences ⁴⁰. After using *hla-mapper* and inferring SNP genotypes with GATK HaplotypeCaller (GVCF mode), 99.64% of the genotypes presented a likelihood higher than 99.99%. Then, the GATK ReadBackedPhasing algorithm directly phased 77.65% of all heterozygous sites, and the remaining sites, most of them indels and multiallelic loci, which are not supported by the ReadBackedPhasing algorithm, were phased using the PHASE algorithm.

Overall, 60% (54/90) of the *HLA-A* coding sequences detected here were identical to those already described in the IPD-IMGT/HLA database (version 3.30.0) and presented a cumulative frequency of 93.25%. Thus, the new alleles detected in this study represent rare sequences and may be attributed to the specific origin of the sample evaluated, since Brazilians are one of the most admixed populations in the world ^{52,53}, carrying African, European and Native American ancestry. Furthermore, the allele group frequencies detected here (Table 1) are consistent with the ones already reported for the Southeastern Brazilian population ^{54,55} (www.allelefrequencies.net). In addition, when the *HLA-A* alleles called by the method applied here were compared with the results obtained by NGSEngine, we found a compatibility of 100% for allele groups, 98.11% for the encoded protein and 92.45% for the full-length (8-digit) allele. However, it should be pointed that most of the incompatibilities were related to the presence of new *HLA-A* alleles detected when using our method.

Our study aimed to evaluate the nucleotide variability, haplotype structure and signatures of natural selection along the entire *HLA-A* gene in a Brazilian population sample. We found 421 variable sites encompassing the promoter, coding, introns and 3'UTR segments. Most of these variable sites were in complete or perfect LD, and a strong LD block structure

was detected encompassing the whole *HLA-A* gene (Figure S2). The low number of haplotypes in the regulatory segments confirmed the high level of LD, particularly in the promoter region where the 100 variable sites were arranged into only 25 promoter haplotypes. A similar pattern was observed for the 3'UTR segment. In addition, each *HLA-A* allele group was usually associated with a single (or closely related subset of) promoter and 3'UTR haplotypes. The number of regulatory haplotypes is similar to the one observed for *HLA-G*, despite the fact that *HLA-G* is a more conserved gene^{34,56-58}. This low number of haplotypes for the regulatory segments and the high LD throughout *HLA-A* can also be visualized by the network at Figures 1 and S1, in which few very divergent *HLA-A* lineages were observed. This network is compatible with balancing selection because of the presence of highly divergent haplotypes (closest lineages differing by between 29 and 125 nucleotides) occurring at high frequencies, which is in fact expected (and well documented) for the *HLA-A* coding region, especially for exons 2 and 3 that encode the peptide-binding groove^{50,59}. Since the *HLA-A* gene is a key locus for triggering adaptive immune responses, this locus is expected to be a major target for natural selection.

Balancing selection is an important force that maintains advantageous genetic diversity in populations, including variation sites that are responsible for long-term adaptation to the environment⁶⁰. Considering our data, we found evidence of balancing selection along the entire *HLA-A* locus, with significant Tajima's *D* for promoter, exons 2, 3 and 4 (and also for some intron segments) and significant Ewens-Watterson tests for exons 2, 5 and 8 (Table 4). Although demographic events may result in signatures that resemble those arising from natural selection, this possibility may be discarded for *HLA-A*, since the results from at least exons 2 and 3 have been systematically observed in populations with very distinct demographic histories^{59,61,62}.

The majority of studies evaluating *HLA-A* variability have shown high polymorphism at sites directly involved in antigen presentation, mainly the exons encoding the peptide-binding groove (exons 2 and 3), thus providing a straightforward link between functional knowledge and population variation²⁴. Neutral evolution cannot explain the high polymorphism found in these segments. Nucleotide diversity at exons 2 and 3 was indeed very high, at about 50 times the human genome average^{50,63}. However, nucleotide diversity at other exons was also high, from 0.01072 for exon 7 (14 times the human genome average) to 0.02896 for exon 5 (38 times the human genome average)^{50,63}, and nucleotide diversity in some introns was sometimes higher than the ones detected for exons (Table 4). This higher diversity at introns and other

exons that do not encode the peptide-binding groove might be a consequence of a hitchhiking effect, arising from the action of balancing selection at exons 2 and 3 and the high level of LD along the entire *HLA-A* locus (Figure S2). This phenomenon was actually already described, with nucleotide variation decreasing with distance from the classical HLA class I genes^{24,64}. We have detected signatures of balancing selection in almost all *HLA-A* segments, including the promoter segment and some introns (Table 4). While balancing selection at the peptide binding groove can be explained by the fitness benefits resulting from an increased peptide repertoire that can be presented at the cell surface^{24,50,65}, a rationale for balancing selection in other *HLA-A* coding regions and in noncoding ones is not straightforward.

Besides the high nucleotide diversity along the *HLA-A* locus when compared to the human genome average, the lowest value within *HLA-A* coding sequence (CDS) was at exon 4. This segment presents a significantly positive Tajima's *D* value, but it also presents evidence of purifying selection because of a higher proportion of synonymous substitutions compared to nonsynonymous ones (Table 5). These results suggest that different nucleotide sequences encoding the same (or similar) amino acid sequence have been maintained throughout evolution as a result of balancing selection. A certain level of conservation of the protein sequence is indeed expected at exon 4 in order to favor the $\alpha 3$ domain interaction with CD8 at T lymphocytes^{66,67}. It is well established that three clusters of amino acids encompassing residues 222-245 are involved in this interaction, but the binding of HLA molecules to T CD8+ cells involves mainly a conserved and negatively charged loop domain in the $\alpha 3$ domain (residues 223 to 229)⁶⁸. For example, residues 227 (Asp, encoded at positions 29,912,030 to 29,912,032 at the hg19) and 228 (Thr, encoded at positions 29,912,033 to 29,912,035 at the hg19) seem to be important for this interaction, since double amino acid substitutions such as Asp/Thr to Lys/Ala at these positions prevent the HLA binding to T CD8+ cell⁶⁹. Using available crystallographic structures from the Protein Data Bank database of some HLA-A (3V5D) and HLA-B (1JGD and 1JGE) molecules, we noticed that residues 223-229 at the $\alpha 3$ domain, encoded by exon 4, are indeed involved in the HLA/CD8 interaction. This segment is conserved in our data since no variable sites were detected (Table S1). Interestingly, by using the 1000 Genomes Phase 3 data (available at <http://phase3browser.1000genomes.org/index.html>), we observed a signature of purifying selection in exon 2 of the *CD8A* gene ($ds - dn = 1.859$, $P=0.033$). It is possible that the presence of divergent nucleotide sequences (but similar amino acid sequences) occurring at similar frequencies results from a hitchhiking effect because of the high LD between this segment and exons 2 and 3.

As *HLA-A* is widely expressed, some level of conservation is expected for the regulatory regions, in order to ensure an adequate expression profile. Thus, signatures of purifying selection, minimizing the presence of variants that might modify the *HLA-A* expression pattern, are expected at the regulatory segments. However, the *HLA-A* extended promoter was found to be very polymorphic, with at least 100 variable sites within approximately 1,600 nucleotides. Nucleotide diversity across the extended promoter may reach 3% or more around position -900, but the lowest diversity was indeed observed at the core promoter (Figure 2), maintaining the key elements for the transcription machinery. This phenomenon was previously observed when studying homozygous cell lines ³⁵, but it also became evident here studying a larger batch of promoter haplotypes from a highly admixed and large population sample. In addition, this former study also detected the same nucleotide diversity pattern across the extended promoter as the one detected here (Figure 2).

Previous studies demonstrated different *HLA-A* allele-specific mRNA levels, but closely related sequences presented a divergent expression pattern ^{35,36}. The expression profile of each evaluated *HLA-A* coding allele was apparently related to both nucleotide variations and different methylation profiles ^{35,36}. In fact, our data indicate that at least 22 variable sites at the *HLA-A* promoter would modify a potential CpG site (such as rs9260072) and that the high expressing alleles *HLA-A**23 and *A**24 ^{35,36} would present a promoter sequence with fewer CpG sites. However, in both studies, *HLA-A* expression levels were measured in specific cell lines. Hence, it is not clear whether the observed pattern would also be detected in different tissues. We hypothesize that the high variability observed at the *HLA-A* regulatory (distal) promoter might influence the binding of transcription factors, and also affect the promoter methylation profile, therefore influencing the gene expression profile according to the microenvironment. In this scenario, the evidence of balancing selection acting on the promoter segment, and the presence of few divergent haplotypes, might (a) result from a hitchhiking effect between the promoter and the coding region, or (b) be related to the maintenance of divergent promoters that respond to different transcription factors and present a different methylation profile as discussed earlier, which might promote an adequate *HLA-A* expression level in different tissues. A confirmatory and comprehensive functional study is required to evaluate these two possibilities.

The most conserved *HLA-A* segment was the 3'UTR (Figure 2, Tables 3 and 4), which is frequently involved in post-transcriptional regulation of gene expression. The lowest

nucleotide diversity indices were indeed detected in the first half of the *HLA-A* 3'UTR. It is not clear why this segment is particularly conserved when compared to the entire *HLA-A* locus, but this phenomenon might be related to microRNA binding. Considering that the *HLA-A* molecule must be expressed in almost all nucleated cells, which in turn present different microRNA environments, some conservation at this segment is indeed expected. In this scenario, it is plausible to consider that the variable sites at the *HLA-A* 3'UTR segment would not influence post-transcriptional regulation (even by modifying microRNA binding or mRNA stability). In fact, the 3'UTR segment is quite conserved among different primates (aligned files upon request), including humans, chimpanzees, gorillas, orangutans, rhesus and others, with many of the variable sites detected at the human 3'UTR segment also detected in non-human primates.

Hence, in conclusion, we were able to develop a reliable methodology for analyzing *HLA-A* variability and *HLA-A* typing considering the entire gene, including the distal promoter and the complete 3'UTR segment. The *HLA-A* extended promoter is very polymorphic, with the presence of few but divergent sequences (Figure S3). The variable sites along the entire *HLA-A* segment are tightly linked to each other, behaving essentially as a single segregation block. As a consequence, each *HLA-A* coding allele group (or closely related coding allele groups) constitutes a distinct *HLA-A* lineage, each characterized by a unique promoter and 3'UTR sequence. As expected, nucleotide diversity was very high at exons 2 and 3, but the same pattern was also observed for the promoter segment (considered as a whole), other exons and many introns. Assessing the evolutionary forces acting on this gene was also surprising since balancing selection was not expected to be the main force acting on the whole *locus* (exception made, maybe, for the 3'UTR segment). However, since these gene regions encompass a very short DNA segment, the effects of natural selection over one region will definitely be experienced by the surrounding ones. Disentangling the effects of natural selection on neighboring regions is a task of a great burden. In order to address this issue, and to completely understand *HLA-A* biology and expression patterns, functional evaluations are mandatory. Meanwhile, we can hypothesize that *HLA-A* expression may be finely regulated to ensure its ubiquitous expression. This fine regulation might be achieved because the *HLA-A* core promoter is quite conserved, maintaining the key elements for the transcription machinery, while the extended promoter presents many variable sites, which in turn may modulate the binding of different transcription factors and the methylation profile, in tissues with different microenvironments. The *HLA-A* 3'UTR is the most conserved *HLA-A* segment, which is also

probably related to the maintenance of an adequate HLA-A expression pattern via an important post-transcription regulation. Although balancing selection would be the main force acting at *HLA-A*, here we also detected evidence of purifying selection at exon 4, maintaining an invariable or similar $\alpha 3$ domain to assure a proper HLA-A/CD8 interaction.

Conflicts of interest

The authors declare that they have no conflict of interest.

Legends

Figure 1: Network from *HLA-A* extended haplotypes (promoter, coding, and 3'UTR). The network was calculated by Median-Joining method using PopArt and further processed using Inkscape. We considered lineages groups of alleles separated by less than 25 differences, whereas branches with more than this number of differences were used to delimit distinct lineages. The number of mutational steps between different haplotypes is indicated as numbers inside small white squares (when higher than 1 mutational step) or not indicated in cases that there is only one mutational step. Each *HLA-A* lineage presents a single 3'UTR haplotype (indicated at each lineage group).

Figure 2: Nucleotide diversity plot across the *HLA-A* locus, considering the genomic sequence from -1605 at the promoter segment up to the end of the 3'UTR segment (upper panel), the mRNA sequence (central panel) and the extended promoter (lower panel). Nucleotide diversity was assessed using a sliding-window approach, with a window size of 150 bp and a step size of 1. Core indicates the region of the *HLA-A* core promoter.

Figure S1: A network of *HLA-A* extended haplotypes (promoter, coding, and 3'UTR). The network was calculated using the Median-Joining method implemented in PopArt (popart.otago.ac.nz) and it was further processed using Inkscape. We considered lineages groups of alleles separated by less than 25 differences, whereas branches with more than this number of differences were used to delimit distinct lineages. The number of mutational steps between different haplotypes is indicated as numbers inside small white squares (when higher than 1 mutational step) or not indicated in cases where there is only one mutational step. Each *HLA-A* lineage presents a single 3'UTR haplotype (indicated at each lineage group). Each

haplotype is named following the IMGT/HLA nomenclature (or as new *HLA-A* alleles) and their associated promoter haplotypes.

Figure S2: Linkage disequilibrium across the *HLA-A* locus. The image was generated in the Haploview 4.2 program using variable sites with a frequency higher than 1%. Areas in dark red or dark gray indicate strong LD ($\text{LOD} \geq 2$, $D' = 1$), shades of pink or shades of gray indicate moderate LD ($\text{LOD} \geq 2$, $D' < 1$), blue or light gray indicate weak LD ($\text{LOD} < 2$, $D' = 1$), and white indicates no LD ($\text{LOD} < 2$, $D' < 1$). D' values different from 1.00 are represented inside the squares as percentages. LOD, log of the odds; D' , pairwise correlation between variable sites. The haplotype block was defined by the confidence intervals method implemented in Haploview.

Figure S3: *HLA-A* promoter haplotypes detected in Brazil.

Acknowledgements

This work was supported by *Fundação de Amparo à Pesquisa do Estado de São Paulo* – FAPESP/Brazil (Grant# 2013/17084-2).

Figure 1. Network of the extended *HLA-A* haplotypes (associated promoters, coding and 3'UTR haplotypes).

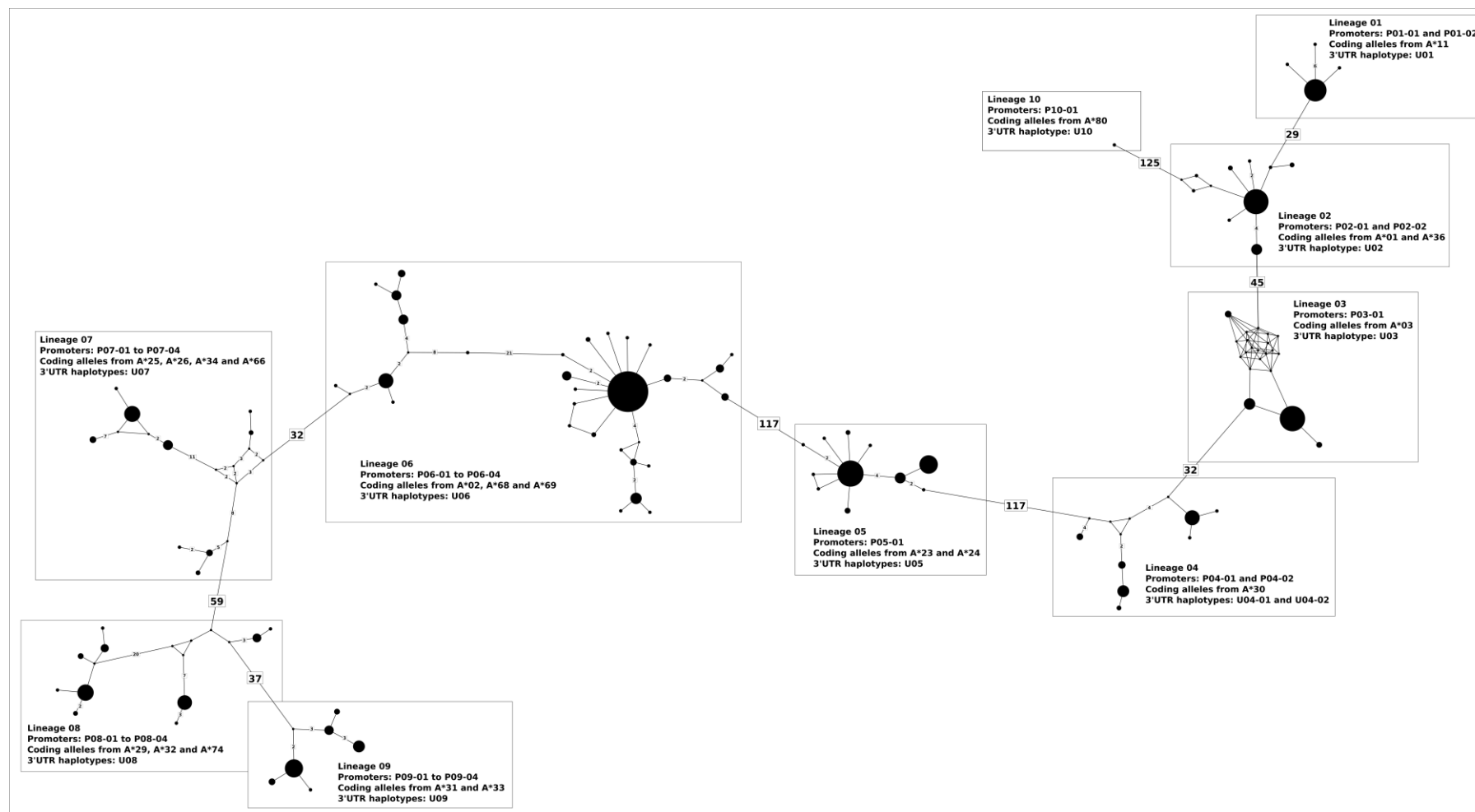


Figure 2: Nucleotide diversity across the entire *HLA-A* gene.

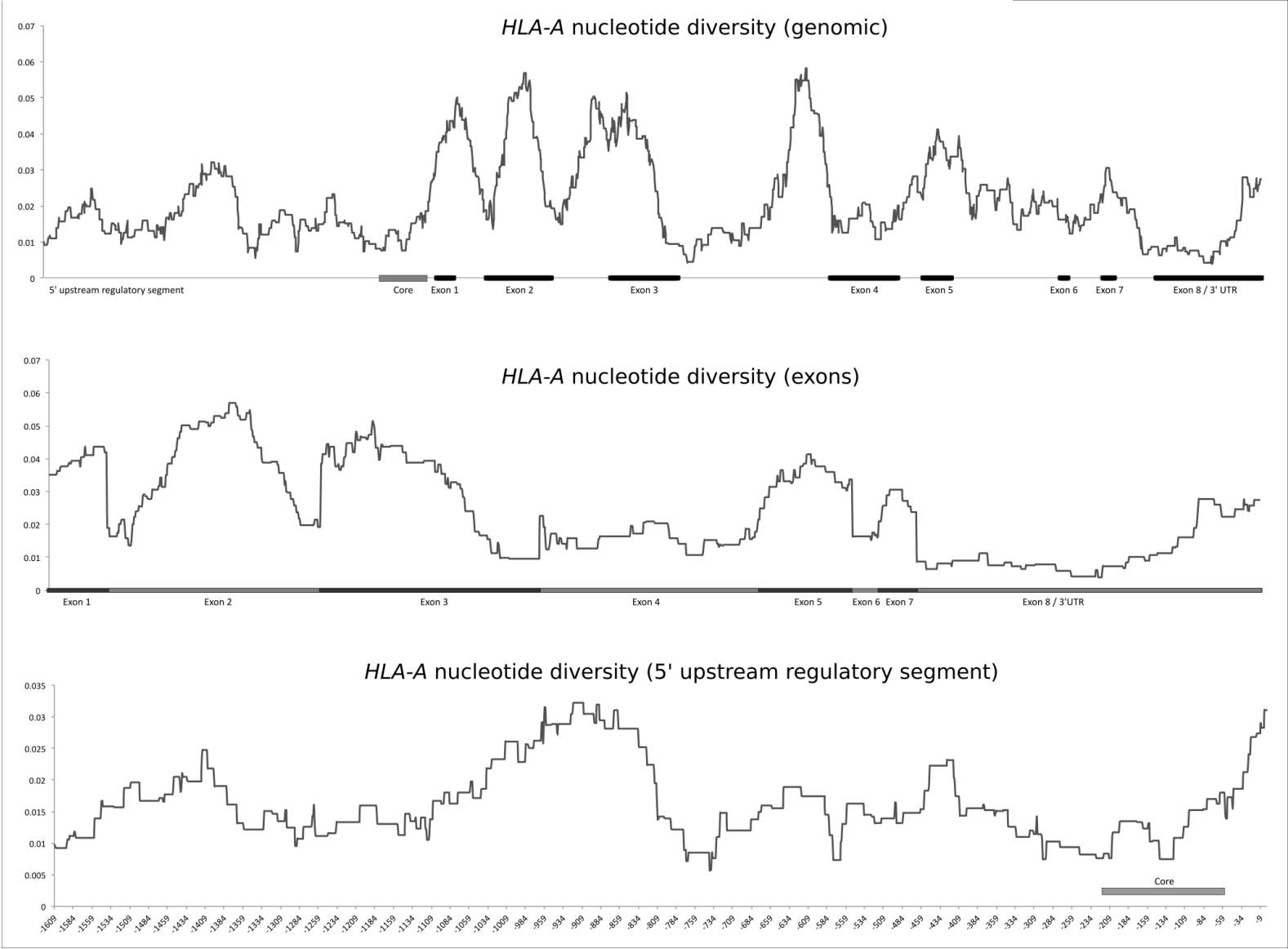


Table 1: List of all *HLA-A* coding alleles detected in a Brazilian population sample from Southeast Brazil.

<i>HLA-A</i> allele ^a	Frequency <i>2n=816</i>	<i>HLA-A</i> allele (continued)	Frequency (continued) <i>2n=816</i>
A*02:01:01:01	0.1949	A*24:02:01:02L	0.0025
A*24:02:01:01	0.0797	A*29:02:01:03	0.0025
A*03:01:01:01	0.0748	A*30:02:01:03	0.0025
A*01:01:01:01	0.0723	A*66:02	0.0025
A*11:01:01:01	0.0576	A*01:01:01:01 ^(127G, 130C)	0.0012
A*23:01:01:01	0.0392	A*01:01:01:01 ^(148C, 152T)	0.0012
A*31:01:02:01	0.0380	A*01:01:01:01 ^(166C)	0.0012
A*29:02:01:01	0.0306	A*01:01:01:01 ^(670C)	0.0012
A*26:01:01:01	0.0294	A*01:01:01:01 ^(99G, 127G)	0.0012
A*30:01:01	0.0257	A*02:01:01:01 ^(different indel at +1256)	0.0012
A*32:01:01:01	0.0245	A*02:01:01:01 ^(1643C)	0.0012
A*68:02:01:01	0.0245	A*02:02:01:01 ^(691T)	0.0012
A*30:02:01:01	0.0159	A*02:02:01:01 ^(726A)	0.0012
A*33:01:01:01	0.0159	A*02:02:01:01 ^(different indel at +1256)	0.0012
A*02:05:01:01	0.0147	A*02:01:01:08	0.0012
A*03:01:01:05	0.0147	A*02:05:01:01 ^(699T)	0.0012
A*23:17	0.0135	A*02:05:01:01 ^(700T)	0.0012
A*36:01	0.0135	A*02:06:01:01	0.0012
A*66:01:01:01	0.0110	A*02:17:02 ^(670C, 707G, 734A, 739A)	0.0012
A*68:01:01:02	0.0110	A*02:20:01	0.0012
A*68:01:02:01	0.0110	A*11:01:01:01 ^(127G, 130C, 138A, 148C, 152T, 873G)	0.0012
A*02:11:01	0.0098	A*11:01:01:01 ^(-245T)	0.0012
A*33:03:01:01	0.0098	A*11:01:01:01 ^(2095G)	0.0012
A*74:01:01	0.0086	A*23:17 ^(734A, 739A)	0.0012
A*02:17:02 ^(707G, 734A, 739A)	0.0074	A*24:02:01:01 ^(670G)	0.0012
A*29:02:01:02	0.0074	A*24:02:01:01 ^(670T)	0.0012
A*34:02:01	0.0074	A*24:02:01:04	0.0012
A*02:04	0.0061	A*24:02:13	0.0012
A*02:17:02 ^(707G)	0.0061	A*24:03:01:01	0.0012
A*30:02:01:02	0.0061	A*26:01:01:01 ^(1004T)	0.0012
A*68:01:02:02	0.0061	A*29:02:01:04	0.0012
A*03:02:01	0.0049	A*29:02:03	0.0012
A*25:01:01:01	0.0049	A*30:01:01:01 ^(107T)	0.0012
A*30:04:01	0.0049	A*30:01:01:01 ^(99G)	0.0012
A*31:01:02:04	0.0049	A*31:01:02:01 ^(-224G)	0.0012
A*02:01:01:01 ^(699T)	0.0037	A*32:02	0.0012
A*02:01:01:01 ^(700T)	0.0037	A*34:02:01 ^(675A, 681T)	0.0012
A*02:02:01:01	0.0037	A*66:03 ^(2390G, 2394C)	0.0012
A*03:01:01:03	0.0037	A*68:02:01:01 ^(2165T, 2147G, 2266G)	0.0012
A*24:02:01:01 ^(673A)	0.0037	A*68:02:01:01 ^(681T)	0.0012
A*29:01:01:01	0.0037	A*68:02:01:03	0.0012
A*33:03:01:01 ^(-196G)	0.0037	A*68:17	0.0012
A*01:01:01:01 ^(670C, 673A)	0.0025	A*69:01:01:01	0.0012
A*01:01:01:01 ^(-172A)	0.0025	A*74:01:01 ^(89G, 91indel)	0.0012
A*02:01:04	0.0025	A*80:01:01:02	0.0012

^a Coding sequences were named following the closest known *HLA-A* allele described in the IPD-IMGT/HLA database version 3.30.0, followed by any divergences eventually observed. The *HLA-A* coding alleles are listed in decreasing order of frequency. The complete sequence of each of these coding alleles is available upon request.

Table 2: Frequency and associated coding alleles of all *HLA-A* promoter haplotypes detected in a Southeastern Brazilian population sample.

Promoter ^a	Frequency 2n=816	Associated <i>HLA-A</i> coding allele groups
P01-01	0.0600	A*11
P01-02	0.0012	A*11
P02-01	0.0944	A*01 and A*36
P02-02	0.0025	A*01 and A*36
P03-01	0.0980	A*03
P04-01	0.0282	A*30
P04-02	0.0294	A*30
P05-01	0.1458	A*23 and A*24
P06-01	0.2623	A*02
P06-02	0.0012	A*02
P06-03	0.0025	A*02
P06-04	0.0588	A*68 and A*69
P07-01	0.0037	A*66
P07-02	0.0061	A*34
P07-03	0.0025	A*34
P07-04	0.0466	A*26
P08-01	0.0257	A*32
P08-02	0.0098	A*74
P08-03	0.0343	A*29
P08-04	0.0123	A*29
P09-01	0.0527	A*31 and A*33
P09-02	0.0012	A*31
P09-03	0.0159	A*33
P09-04	0.0037	A*33
P10-01	0.0012	A*80

^a For a list of variable sites and the structure of each of these promoter haplotypes, please refer to Figure S3.

Table 3: Frequency and associated *HLA-A* coding alleles of all *HLA-A* 3' untranslated region haplotypes detected in a Southeastern Brazilian population sample.

	hg19 chr6 position ^a	29913266	29913298	29913343	29913344	29913365	29913367	29913374	29913386	29913430	29913434	29913449	29913483	29913501	29913515	29913539	29913542	29913554	29913588	29913630	29913638	29913639	Frequency n = 816	<i>HLA-A</i> coding associated group
<i>HLA-A</i> 3' untranslated region haplotypes	U03	G	A	C	G	G	G	C	C	C	A	C	T	C	C	C	T	A	G	A	G	A	0.098	A*03
	U04-01	G	A	C	G	G	G	C	C	C	A	C	T	C	C	C	G	A	G	A	G	A	0.0527	A*30 ^b
	U04-02	G	A	C	G	G	G	C	C	C	A	C	T	C	C	C	G	C	G	A	G	A	0.0049	A*30 ^c
	U06	G	A	C	G	G	G	C	C	C	A	C	T	C	C	C	G	A	G	A	C	A	0.3248	A*02, A*68 and A*69
	U05	G	A	C	A	G	G	C	C	C	A	G	T	C	C	C	G	A	G	A	C	A	0.1458	A*23 and A*24
	U07	T	A	C	T	G	G	C	C	C	A	C	C	C	T	C	G	A	G	A	C	A	0.0588	A*25 and A*66
	U08	T	A	C	T	G	G	C	C	C	A	C	C	C	C	C	G	A	G	A	C	A	0.0821	A*29, A*32 and A*74
	U09	T	T	C	T	G	G	C	C	C	A	C	C	C	C	C	G	A	G	A	C	A	0.0735	A*31 and A*33
	U01	G	A	C	G	G	G	C	G	T	A	C	T	C	C	C	G	A	G	C	C	G	0.0613	A*11
	U02	G	A	C	G	G	G	G	C	T	A	C	T	C	C	C	G	A	G	T	C	G	0.0968	A*01 and A*36
	U10	G	A	G	G	A	A	C	C	C	G	C	T	T	T	A	G	A	C	A	C	A	0.0012	A*80 ^d

^a For a list of variable sites and their IMGT/HLA related position, please refer to Table S1.

^b A*30:01:01, A*30:02:01

^c A*30:04:01

^d A*80:01:01

Table 4: Diversity and neutrality deviation tests for each *HLA-A* segment

<i>HLA-A</i> segment	Sequence Size	Number of Codons	Segregation Sites	Nucleotide Diversity	Gene Diversity	Tajima's <i>D</i>	Ewens-Watterson
Promoter	1608	-	98	0.0172 ± 0.0083	0.8764 ± 0.0063	2.4513, <i>P</i> = 0.0108	<i>Fo</i> = 0.1247, <i>Fe</i> = 0.1702, <i>P</i> = 0.2272
Exon 1	73	24	6	0.0195 ± 0.0132	0.7753 ± 0.0057	1.2665, <i>P</i> = 0.1050	<i>Fo</i> = 0.2256, <i>Fe</i> = 0.4926, <i>P</i> = 0.0778
Intron 1	130	-	18	0.0473 ± 0.0249	0.8507 ± 0.0055	3.2053, <i>P</i> = 0.0030	<i>Fo</i> = 0.1504, <i>Fe</i> = 0.2343, <i>P</i> = 0.7890
Exon 2	270	90	40	0.0379 ± 0.0192	0.8941 ± 0.0055	2.2638, <i>P</i> = 0.0154	<i>Fo</i> = 0.1070, <i>Fe</i> = 0.1638, <i>P</i> = 0.0345
Intron 2	241	-	27	0.0247 ± 0.0131	0.8812 ± 0.0054	1.4999, <i>P</i> = 0.0608	<i>Fo</i> = 0.1199, <i>Fe</i> = 0.1451, <i>P</i> = 0.7901
Exon 3	276	92	37	0.0401 ± 0.0202	0.9100 ± 0.0051	3.0634, <i>P</i> = 0.0017	<i>Fo</i> = 0.0911, <i>Fe</i> = 0.1271, <i>P</i> = 0.1592
Intron 3	601	-	38	0.0349 ± 0.0171	0.8624 ± 0.0048	2.5694, <i>P</i> = 0.0088	<i>Fo</i> = 0.1386, <i>Fe</i> = 0.2917, <i>P</i> = 0.1282
Exon 4	276	92	25	0.0267 ± 0.0138	0.7950 ± 0.0075	2.8155, <i>P</i> = 0.0056	<i>Fo</i> = 0.2060, <i>Fe</i> = 0.3340, <i>P</i> = 0.2259
Intron 4	102	-	5	0.0249 ± 0.0148	0.6270 ± 0.0088	0.8490, <i>P</i> = 0.1780	<i>Fo</i> = 0.3737, <i>Fe</i> = 0.5399, <i>P</i> = 0.5234
Exon 5	117	39	14	0.0290 ± 0.0164	0.7650 ± 0.0079	1.6751, <i>P</i> = 0.0510	<i>Fo</i> = 0.2359, <i>Fe</i> = 0.4494, <i>P</i> = 0.0400
Intron 5	444	-	43	0.0320 ± 0.0159	0.8378 ± 0.0081	2.5823, <i>P</i> = 0.0089	<i>Fo</i> = 0.1632, <i>Fe</i> = 0.2454, <i>P</i> = 0.1719
Exon 6	33	11	3	0.0262 ± 0.0204	0.4328 ± 0.0122	1.4778, <i>P</i> = 0.0916	<i>Fo</i> = 0.5677, <i>Fe</i> = 0.7539, <i>P</i> = 0.5761
Intron 6	142	-	16	0.0172 ± 0.0103	0.6642 ± 0.0156	0.2491, <i>P</i> = 0.6668	<i>Fo</i> = 0.3366, <i>Fe</i> = 0.4145, <i>P</i> = 0.2226
Exon 7	48	16	2	0.0107 ± 0.0100	0.4723 ± 0.0167	0.9928, <i>P</i> = 0.1419	<i>Fo</i> = 0.5283, <i>Fe</i> = 0.7562, <i>P</i> = 0.0729
Intron 7	169	-	17	0.0286 ± 0.0155	0.8007 ± 0.0065	2.2155, <i>P</i> = 0.0190	<i>Fo</i> = 0.2003, <i>Fe</i> = 0.3592, <i>P</i> = 0.2621
Exon 8 / 3'UTR	337	-	21	0.0075 ± 0.0043	0.8331 ± 0.0079	0.3165, <i>P</i> = 0.3059	<i>Fo</i> = 0.1679, <i>Fe</i> = 0.3561, <i>P</i> = 0.0097
CDS	1095	365	127	0.0318 ± 0.0153	0.9209 ± 0.0049	2.8495, <i>P</i> = 0.0044	<i>Fo</i> = 0.0803, <i>Fe</i> = 0.0826, <i>P</i> = 0.4869
Full segment	4964	-	408	0.0247 ± 0.0117	0.9345 ± 0.0047	2.5612, <i>P</i> = 0.0082	<i>Fo</i> = 0.0667, <i>Fe</i> = 0.0363, <i>P</i> = 0.9999

Table 5: Synonymous and non-synonymous nucleotide substitution test of neutrality, positive and purifying selection considering the *HLA-A* sequences observed in Brazilian samples.

Location	Size	Codons	<i>H</i> _a = Neutrality (<i>dn</i> ≠ <i>ds</i>)	<i>H</i> _a = Positive Selection (<i>dn</i> > <i>ds</i>)	<i>H</i> _a = Purifying Selection (<i>dn</i> < <i>ds</i>)
Exon 1	73	24	<i>dn</i> – <i>ds</i> = -0.544, <i>P</i> =0.588	<i>dn</i> – <i>ds</i> = -0.575, <i>P</i> =1.000	<i>ds</i> – <i>dn</i> = +0.545, <i>P</i> =0.293
Exon 2	270	90	<i>dn</i> – <i>ds</i> = +0.421, <i>P</i> =0.675	<i>dn</i> – <i>ds</i> = +0.413, <i>P</i> =0.340	<i>ds</i> – <i>dn</i> = -0.411, <i>P</i> =1.000
Exon 3	276	92	<i>dn</i> – <i>ds</i> = +1.654, <i>P</i> =0.100	<i>dn</i> – <i>ds</i> = +1.540, <i>P</i> =0.063	<i>ds</i> – <i>dn</i> = -1.571, <i>P</i> =1.000
Exon 4	276	92	<i>dn</i>–<i>ds</i> = -2.780, <i>P</i>=0.006	<i>dn</i> – <i>ds</i> = -2.685, <i>P</i> =1.000	<i>ds</i>–<i>dn</i> = +2.721, <i>P</i>=0.004
Exon 5	117	39	<i>dn</i> – <i>ds</i> = -0.044, <i>P</i> =0.965	<i>dn</i> – <i>ds</i> = -0.043, <i>P</i> =1.000	<i>ds</i> – <i>dn</i> = +0.042, <i>P</i> =0.433
Exon 6	33	11	<i>dn</i> – <i>ds</i> = -0.557, <i>P</i> =0.578	<i>dn</i> – <i>ds</i> = -0.576, <i>P</i> =1.000	<i>ds</i> – <i>dn</i> = +0.588, <i>P</i> =0.279
Exon 7	48	16	<i>dn</i> – <i>ds</i> = -0.417, <i>P</i> =0.677	<i>dn</i> – <i>ds</i> = -0.421, <i>P</i> =1.000	<i>ds</i> – <i>dn</i> = +0.417, <i>P</i> =0.338
CDS	1095	365	<i>dn</i> – <i>ds</i> = -1.216, <i>P</i> =0.226	<i>dn</i> – <i>ds</i> = -1.189, <i>P</i> =1.000	<i>ds</i> – <i>dn</i> = +1.195, <i>P</i> =0.117

References

1. Krensky AM, Clayberger C. Structure of HLA molecules and immunosuppressive effects of HLA derived peptides. *International reviews of immunology*. 1996;13(3):173-185.
2. Maffei A, Harris PE. Peptides bound to major histocompatibility complex molecules. *Peptides*. 1998;19(1):179-198.
3. Shiina T, Hosomichi K, Inoko H, Kulski JK. The HLA genomic loci map: expression, interaction, diversity and disease. *J Hum Genet*. 2009;54(1):15-39.
4. Howell WM, Carter V, Clark B. The HLA system: immunobiology, HLA typing, antibody screening and crossmatching techniques. *Journal of clinical pathology*. 2010;63(5):387-390.
5. Game DS, Lechler RI. Pathways of allorecognition: implications for transplantation tolerance. *Transplant immunology*. 2002;10(2-3):101-108.
6. Klein J, Sato A. The HLA system. Second of two parts. *The New England journal of medicine*. 2000;343(11):782-786.
7. Vyas JM, Van der Veen AG, Ploegh HL. The known unknowns of antigen processing and presentation. *Nature reviews Immunology*. 2008;8(8):607-618.
8. Hayashi M, Jin Y, Yorgov D, et al. Autoimmune vitiligo is associated with gain-of-function by a transcriptional regulator that elevates expression of HLA-A*02:01 in vivo. *Proc Natl Acad Sci U S A*. 2016;113(5):1357-1362.
9. Jansen ME, Brankovic I, Spaargaren J, Ouburg S, Morre SA. Potential protective effect of a G>A SNP in the 3'UTR of HLA-A for Chlamydia trachomatis symptomatology and severity of infection. *Pathogens and disease*. 2016;74(2).
10. Monteiro SP, Brasil PE, Cabello GM, et al. HLA-A*01 allele: a risk factor for dengue haemorrhagic fever in Brazil's population. *Memorias do Instituto Oswaldo Cruz*. 2012;107(2):224-230.
11. Mosaad YM, Farag RE, Arafa MM, et al. Association of human leucocyte antigen Class I (HLA-A and HLA-B) with chronic hepatitis C virus infection in Egyptian patients. *Scand J Immunol*. 2010;72(6):548-553.
12. Samandary S, Kridane-Miledi H, Sandoval JS, et al. Associations of HLA-A, HLA-B and HLA-C alleles frequency with prevalence of herpes simplex virus infections and diseases across global populations: implication for the development of an universal CD8+ T-cell epitope-based vaccine. *Human immunology*. 2014;75(8):715-729.
13. Vandiedonck C, Raffoux C, Eymard B, et al. Association of HLA-A in autoimmune myasthenia gravis with thymoma. *Journal of neuroimmunology*. 2009;210(1-2):120-123.
14. Koehler RN, Walsh AM, Saathoff E, et al. Class I HLA-A*7401 is associated with protection from HIV-1 acquisition and disease progression in Mbeya, Tanzania. *The Journal of infectious diseases*. 2010;202(10):1562-1566.
15. Fontaine Costa AI, Rao X, Lechenadec E, van Baarle D, Kesmir C. HLA-B molecules target more conserved regions of the HIV-1 proteome. *AIDS*. 2010;24(2):211-215.
16. Celerino da Silva R, Rodrigues de Moura R, Victor Campos Coelho A, et al. HLA-C single nucleotide polymorphism associated with increased viral load level in HIV-1 infected individuals from Northeast Brazil. *Current HIV research*. 2017.
17. Sorensen BS, Szer J, Shaw B, et al. Inadvertent completely HLA-mismatched allogeneic unrelated bone marrow transplant: lessons learned. *Bone marrow transplantation*. 2016;51(7):1016-1018.
18. Rosansky SJ. The effectiveness of HLA matching in improving transplant organ survival. *Transplantation*. 1984;37(4):427.
19. Bosserman LD, Murray C, Takvorian T, et al. Mechanism of graft failure in HLA-matched and HLA-mismatched bone marrow transplant recipients. *Bone marrow transplantation*. 1989;4(3):239-245.
20. Ober C, Aldrich CL, Chervoneva I, et al. Variation in the HLA-G promoter region influences miscarriage rates. *American journal of human genetics*. 2003;72(6):1425-1435.

21. Winternitz J, Abbate JL, Huchard E, Havlicek J, Garamszegi LZ. Patterns of MHC-dependent mate selection in humans and nonhuman primates: a meta-analysis. *Molecular ecology*. 2017;26(2):668-688.
22. Kromer J, Hummel T, Pietrowski D, et al. Influence of HLA on human partnership and sexual satisfaction. *Scientific reports*. 2016;6:32550.
23. Hviid TV. HLA-G in human reproduction: aspects of genetics, function and pregnancy complications. *Human reproduction update*. 2006;12(3):209-232.
24. Meyer D, Thomson G. How selection shapes variation of the human major histocompatibility complex: a review. *Annals of Human Genetics*. 2001;65:1-26.
25. Klopper HN, Stryhn A, Harndahl M, et al. HLA-B*57 Micropolymorphism shapes HLA allele-specific epitope immunogenicity, selection pressure, and HIV immune control. *Journal of virology*. 2012;86(2):919-929.
26. Penn DJ, Damjanovich K, Potts WK. MHC heterozygosity confers a selective advantage against multiple-strain infections. *Proc Natl Acad Sci U S A*. 2002;99(17):11260-11264.
27. Prugnolle F, Manica A, Charpentier M, Guegan JF, Guernier V, Balloux F. Pathogen-driven selection and worldwide HLA class I diversity. *Current biology : CB*. 2005;15(11):1022-1027.
28. Cagliani R, Sironi M. Pathogen-driven selection in the human genome. *Int J Evol Biol*. 2013;2013:204240.
29. Yasukochi Y, Satta Y. Current perspectives on the intensity of natural selection of MHC loci. *Immunogenetics*. 2013;65(6):479-483.
30. Gineau L, Luisi P, Castelli EC, et al. Balancing immunity and tolerance: genetic footprint of natural selection in the transcriptional regulatory region of HLA-G. *Genes and immunity*. 2015;16(1):57-70.
31. Takahata N, Nei M. Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. *Genetics*. 1990;124(4):967-978.
32. Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SG. The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Res*. 2015;43(Database issue):D423-431.
33. Meyer D, VR CA, Bitarello BD, DY CB, Nunes K. A genomic perspective on HLA evolution. *Immunogenetics*. 2018;70(1):5-27.
34. Castelli EC, Gerasimou P, Paz MA, et al. HLA-G variability and haplotypes detected by massively parallel sequencing procedures in the geographically distinct population samples of Brazil and Cyprus. *Mol Immunol*. 2017;83:115-126.
35. Ramsuran V, Hernandez-Sanchez PG, O'Huigin C, et al. Sequence and Phylogenetic Analysis of the Untranslated Promoter Regions for HLA Class I Genes. *Journal of immunology*. 2017;198(6):2320-2329.
36. Ramsuran V, Kulkarni S, O'Huigin C, et al. Epigenetic regulation of differential HLA-A allelic expression levels. *Human molecular genetics*. 2015;24(15):4268-4275.
37. Lima TH, Buttura RV, Donadi EA, Veiga-Castelli LC, Mendes-Junior CT, Castelli EC. HLA-F coding and regulatory segments variability determined by massively parallel sequencing procedures in a Brazilian population sample. *Human immunology*. 2016;77(10):841-853.
38. Castelli EC, Mendes-Junior CT, Sabbagh A, et al. HLA-E coding and 3' untranslated region variability determined by next-generation sequencing in two West-African population samples. *Human immunology*. 2015.
39. Brandt DY, Aguiar VR, Bitarello BD, Nunes K, Goudet J, Meyer D. Mapping Bias Overestimates Reference Allele Frequencies at the HLA Genes in the 1000 Genomes Project Phase I Data. *G3 (Bethesda)*. 2015;5(5):931-941.
40. Castelli EC, Paz MA, Souza AS, Ramalho J, Mendes-Junior CT. hla-mapper: an application to optimize the mapping of HLA sequences produced by massively parallel sequencing procedures. *Human immunology*. 2018;(in press).
41. DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*. 2011;43(5):491-498.
42. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*. 2010;20(9):1297-1303.

43. Van der Auwera GA, Carneiro MO, Hartl C, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics*. 2013;11(1110):11 10 11-11 10 33.
44. Stephens M, Smith NJ, Donnelly P. A new statistical method for haplotype reconstruction from population data. *American journal of human genetics*. 2001;68(4):978-989.
45. Stephens M, Donnelly P. A comparison of bayesian methods for haplotype reconstruction from population genotype data. *American journal of human genetics*. 2003;73(5):1162-1169.
46. Excoffier L, Lischer HE. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol Ecol Resour*. 2010;10(3):564-567.
47. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol Biol Evol*. 2016;33(7):1870-1874.
48. Vilella AJ, Blanco-Garcia A, Hutter S, Rozas J. VariScan: Analysis of evolutionary patterns from large-scale DNA sequence polymorphism data. *Bioinformatics*. 2005;21(11):2791-2793.
49. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*. 2005;21(2):263-265.
50. Meyer D, Single RM, Mack SJ, Erlich HA, Thomson G. Signatures of demographic history and natural selection in the human major histocompatibility complex Loci. *Genetics*. 2006;173(4):2121-2142.
51. Hughes AL, Nei M. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature*. 1988;335(6186):167-170.
52. Lins TC, Vieira RG, Abreu BS, Grattapaglia D, Pereira RW. Genetic composition of Brazilian population samples based on a set of twenty-eight ancestry informative SNPs. *Am J Hum Biol*. 2010;22(2):187-192.
53. Alves-Silva J, da Silva Santos M, Guimaraes PE, et al. The ancestry of Brazilian mtDNA lineages. *American journal of human genetics*. 2000;67(2):444-461.
54. Rosales T, Guilherme L, Chiarella J, et al. Human leukocyte A and B antigen, gene and haplotype frequencies in the population of the city of Sao Paulo in Brazil. *Braz J Med Biol Res*. 1992;25(1):39-47.
55. Salvadori Lde C, Santana FC, Marcos EV. Frequency of alleles and haplotypes of the human leukocyte antigen system in Bauru, Sao Paulo, Brazil. *Rev Bras Hematol Hemoter*. 2014;36(2):108-114.
56. Hviid TV, Rizzo R, Melchiorri L, Stignani M, Baricordi OR. Polymorphism in the 5' upstream regulatory and 3' untranslated regions of the HLA-G gene in relation to soluble HLA-G and IL-10 expression. *Human immunology*. 2006;67(1-2):53-62.
57. Castelli EC, Ramalho J, Porto IO, et al. Insights into HLA-G Genetics Provided by Worldwide Haplotype Diversity. *Front Immunol*. 2014;5:476.
58. Castelli EC, Mendes-Junior CT, Veiga-Castelli LC, Roger M, Moreau P, Donadi EA. A comprehensive study of polymorphic sites along the HLA-G gene: implication for gene regulation and evolution. *Mol Biol Evol*. 2011;28(11):3069-3086.
59. Hedrick PW, Thomson G. Evidence for balancing selection at HLA. *Genetics*. 1983;104(3):449-456.
60. Andres AM, Hubisz MJ, Indap A, et al. Targets of balancing selection in the human genome. *Mol Biol Evol*. 2009;26(12):2755-2764.
61. Black FL, Hedrick PW. Strong balancing selection at HLA loci: evidence from segregation in South Amerindian families. *Proc Natl Acad Sci U S A*. 1997;94(23):12452-12456.
62. Solberg OD, Mack SJ, Lancaster AK, et al. Balancing selection and heterogeneity across the classical human leukocyte antigen loci: a meta-analytic review of 497 population studies. *Human immunology*. 2008;69(7):443-464.
63. Sachidanandam R, Weissman D, Schmidt SC, et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*. 2001;409(6822):928-933.
64. Gaudieri S, Dawkins R, Habara K, Kulski J, Gojobori T. Nucleotide diversity within the human major histocompatibility complex: function of hitchhiking effect, duplications, indels and recombination. In: Kasahara M, ed. *The Major Histocompatibility Complex: Evolution, Structure, Function*. Tokyo: Springer-Verlag; 2000:186-200.

65. Bernatchez L, Landry C. MHC studies in nonmodel vertebrates: what have we learned about natural selection in 15 years? *J Evol Biol.* 2003;16(3):363-377.
66. Fayen J, Huang JH, Meyerson H, et al. Class I MHC alpha 3 domain can function as an independent structural unit to bind CD8 alpha. *Mol Immunol.* 1995;32(4):267-275.
67. Littman DR. The structure of the CD4 and CD8 genes. *Annual review of immunology.* 1987;5:561-584.
68. Wesley PK, Clayberger C, Lyu SC, Krensky AM. The CD8 coreceptor interaction with the alpha 3 domain of HLA class I is critical to the differentiation of human cytotoxic T-lymphocytes specific for HLA-A2 and HLA-Cw4. *Human immunology.* 1993;36(3):149-155.
69. Pittet MJ, Rubio-Godoy V, Bioley G, et al. Alpha 3 domain mutants of peptide/MHC class I multimers allow the selective isolation of high avidity tumor-reactive CD8 T cells. *Journal of immunology.* 2003;171(4):1844-1849.

Conclusões

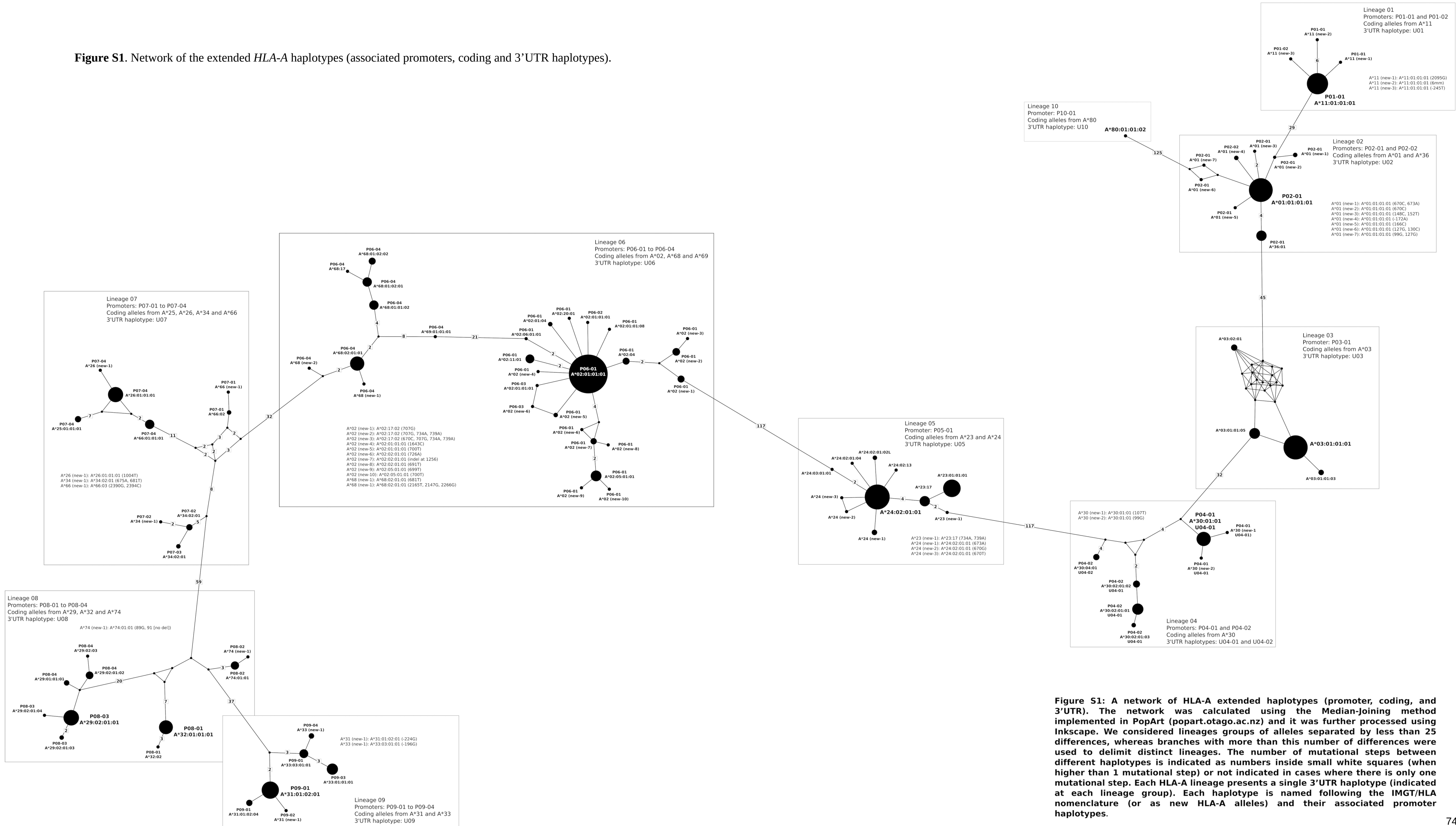
Em conclusão, nós desenvolvemos uma metodologia altamente eficaz para analisar a variabilidade dos genes HLA, considerando todo seu segmento, incluindo o promotor distal e o segmento completo da região 3' não-traduzida. A região do promotor estendido do gene *HLA-A* é altamente polimórfica, com a presença de poucas sequências, mas altamente divergentes (Figura S3). Os sítios de variação detectado ao longo de todo o segmento gênico estão altamente associados em função do alto LD observado. Como consequência do alto LD, cada grupo de alelos codificadores do gene *HLA-A* (ou grupo de alelos próximos relacionados) constitui uma linhagem *HLA-A* distinta, cada uma caracterizada por um único promotor e uma única sequência 3' não-traduzida. Como esperado, uma elevada diversidade nucleotídica foi observada para os éxons 2 e 3, mas o mesmo padrão também foi observado para o segmento promotor completo, outros éxons e regiões intrônicas.

Quanto as assinaturas de seleção natural atuando neste gene, foi surpreendente detectar o perfil de seleção balanceadora atuando em todo segmento, uma vez que apostávamos na possibilidade de perfis de seleção diferentes atuando nos segmentos regulatórios do gene *HLA-A*. Avaliações funcionais são necessárias para o entendimento completo da biologia e padrão de expressão deste gene, contudo aqui supomos que o *HLA-A* possui um padrão de regulação da expressão muito fino o que garante que seu padrão de expressão seja ubíquo. Essa regulação fina provavelmente se explica pela conservação das regiões do core do promotor, mantendo os elementos chaves para a ação da maquinaria de transcrição, enquanto o promotor estendido apresenta muitos sítios de variação, o que pode modular a expressão deste gene, possibilitando a ligação de diferentes fatores de transcrição a depender do tecido e do microambiente em que se encontra. A região 3' não-traduzida é o segmento mais conservado, provavelmente relacionado com a manutenção de níveis de expressão adequados da molécula HLA-A. Além disso, o domínio $\alpha 3$ do *HLA-A*, que interage com CD8, também parece apresentar certo grau de conservação, de certa forma garantindo uma adequada interação entre HLA-A/CD8, e a porção 3'NT é altamente conservada, o que provavelmente favorece a expressão diferencial do gene *HLA-A* nos tecidos com microRNAs e microambientes diferentes.

Material suplementar

Tabelas e figuras suplementares. A resolução das figuras podem ficar comprometidas em função do formato do arquivo, mas todas encontram-se disponíveis por meio de solicitação (thalitta.ayala@unesp.br).

Figure S1. Network of the extended *HLA-A* haplotypes (associated promoters, coding and 3'UTR haplotypes).



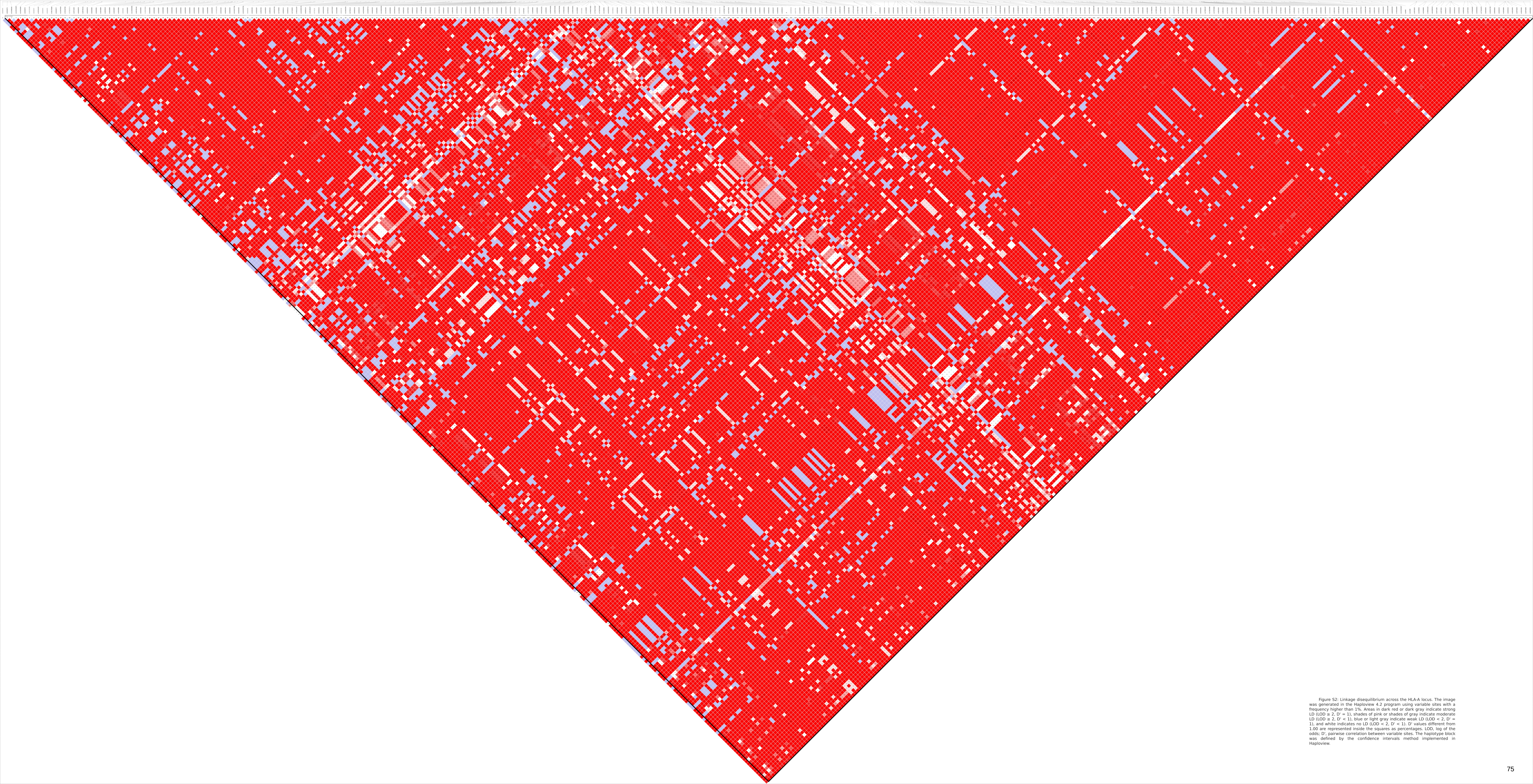


Figure S2: Linkage disequilibrium across the HLA-A locus. The image was generated in the Haploview 4.2 program using variable sites with a frequency higher than 1%. Areas in dark red or dark gray indicate strong LD ($LOD \geq 2$, $D' = 1$), shades of pink or shades of gray indicate moderate LD ($LOD \geq 2$, $D' < 1$), blue or light gray indicate weak LD ($LOD < 2$, $D' = 1$), and white indicates no LD ($LOD < 2$, $D' < 1$). D' values different from 1.00 are represented inside the squares as percentages. LOD, log of the odds; D' , pairwise correlation between variable sites. The haplotype block was defined by the confidence intervals method implemented in Haploview.

Table S1: List of variable sites at the *HLA-A* gene detected in Brazil.

Chromossome position (hg19)	IMGT- relative position	SNPid	Reference Allele	Frequency 2n = 816	First Alternative Allele	Frequency 2n = 816	Second Alternative Allele	Frequency 2n = 816	Third Alternative Allele	Frequency 2n = 816	Notes
29908726	-1605	rs9260070	C	0,9522	G	0,0478					promoter
29908742	-1589	rs9260071	A	0,9154	C	0,0846					promoter
29908753	-1578	rs72498365	G	0,9265	C	0,0735					promoter
29908789	-1542	rs112519287	T	0,9289	C	0,0711					promoter
29908804	-1527	rs13203830	G	0,6630	A	0,3370					promoter
29908838	-1493	rs9260072	G	0,3137	C	0,6863					promoter
29908869	-1462	rs143741490	C	0,9534	A	0,0466					promoter
29908891	-1440	rs9260073	C	0,9522	T	0,0478					promoter
29908892	-1439	rs79761237	G	0,8456	A	0,1544					promoter
29908895	-1436	rs9260074	G	0,9522	A	0,0478					promoter
29908901	-1430	rs9260075	A	0,9522	C	0,0478					promoter
29908928	-1403	rs67901303	GAC	0,3615	G	0,3015	AAC	0,3370			promoter
29908937	-1394	rs78179089	C	0,6985	T	0,3015					promoter
29908954	-1377	rs9260077	G	0,3137	A	0,6863					promoter
29908967	-1364	rs13207054	G	0,6630	A	0,3370					promoter
29908976	-1355	rs74218766	C	0,9265	T	0,0735					promoter
29909014	-1317	rs375623213	T	0,9657	C	0,0343					promoter
29909021	-1310	rs2571420	C	0,9032	G	0,0968					promoter
29909033	-1298	rs9260078	A	0,2855	G	0,7145					promoter
29909044	-1287	rs28749141	G	0,6728	C	0,3272					promoter
29909070	-1261	rs9260079	T	0,6164	C	0,3836					promoter
29909071	-1260	rs9260080	G	0,8419	A	0,1581					promoter
29909111	-1220	rs191021452	C	0,9988	G	0,0012					promoter
29909153	-1178	rs9260081	G	0,3137	T	0,6863					promoter
29909181	-1150	rs1632889	A	0,1556	C	0,8444					promoter
29909197	-1134	rs1655919	G	0,0980	C	0,9020					promoter
29909205	-1126	rs1632888	C	0,8419	T	0,1581					promoter
29909217	-1114	rs1632887	T	0,8419	G	0,1581					promoter
29909219	-1112	rs1632886	T	0,8419	C	0,1581					promoter
29909238	-1093	rs9260082	T	0,9645	C	0,0355					promoter
29909249	-1082	rs9260083	A	0,8444	C	0,1556					promoter

29909251	-1080	rs112664400	T	0,3137	TA	0,6863			promoter
29909279	-1052	rs28749142	T	0,7341	A	0,2659			promoter
29909339	-992	rs9260084	G	0,4706	A	0,5294			promoter
29909353	-978	rs140625882	C	0,9534	G	0,0466			promoter
29909360	-971	rs9260085	G	0,8444	A	0,1556			promoter
29909372	-959	rs9260086	C	0,4706	G	0,5294			promoter
29909376	-955	rs9260087	C	0,3137	A	0,6863			promoter
29909391	-940	rs9260088	C	0,8431	A	0,1569			promoter
29909409	-922	rs9260089	C	0,8444	T	0,1556			promoter
29909426	-905	rs9260090	T	0,8431	C	0,1569			promoter
29909440	-891	rs9260091	A	0,8824	G	0,1176			promoter
29909447	-884	rs9260092	G	0,4706	C	0,5294			promoter
29909454	-877	rs9260093	T	0,8824	C	0,1176			promoter
29909472	-859	rs9260094	G	0,3137	C	0,6863			promoter
29909499	-832	rs9260095	A	0,3150	T	0,6850			promoter
29909510	-821	rs9260096	C	0,3150	T	0,6850			promoter
29909520	-811	rs9260097	G	0,3150	A	0,6850			promoter
29909523	-808	rs9260098	A	0,3150	C	0,6850			promoter
29909524	-807	rs9260099	G	0,3150	T	0,6850			promoter
29909533	-798	rs576018139	G	0,9841	C	0,0159			promoter
29909542	-789	rs9260100	A	0,8444	G	0,1556			promoter
29909559	-772	rs9260101	G	0,4706	A	0,5294			promoter
29909562	-769	rs9260102	C	0,8444	A	0,1556			promoter
29909593	-738	rs9260103	T	0,3137	G	0,6850	C	0,0012	promoter
29909597	-734	rs114157794	T	0,9265	G	0,0735			promoter
29909616	-715	rs9260104	T	0,3137	C	0,6863			promoter
29909633	-698	rs141035764	A	0,9424	AT	0,0576			promoter
29909674	-657	rs151250402	A	0,9988	G	0,0012			promoter
29909675	-656	rs9260105	G	0,9645	T	0,0355			promoter
29909714	-617	rs9260106	T	0,8824	C	0,1176			promoter
29909745	-586	rs9260107	A	0,3137	G	0,6863			promoter
29909751	-580	rs2905764	T	0,4706	C	0,5294			promoter
29909756	-575	rs9260109	G	0,3137	C	0,5404	T	0,1458	promoter
29909769	-562	rs67139508	AAAC	0,7831	A	0,2157	CAAC	0,0012	promoter
29909797	-534	rs3129022	C	0,8419	T	0,1581			promoter
29909807	-524	rs3094141	T	0,0980	G	0,9020			promoter
29909811	-520	rs35333729	G	0,9167	T	0,0833			promoter

29909840	-491	rs2734904	G	0,4694	A	0,5306	promoter
29909916	-415	rs9260111	A	0,3137	T	0,6863	promoter
29909918	-413	rs9260112	T	0,3137	C	0,6863	promoter
29909924	-407	rs2734903	G	0,6164	A	0,3836	promoter
29909957	-374	rs79231337	G	0,9265	C	0,0735	promoter
29909970	-361	rs116395178	T	0,9375	A	0,0625	promoter
29909988	-343	rs9260114	C	0,2574	T	0,7426	promoter
29909999	-332	rs35158227	C	0,8542	T	0,1458	promoter
29910022	-309	rs2735115	C	0,9534	A	0,0466	promoter
29910025	-306	.	T	0,9988	C	0,0012	promoter
29910027	-304	rs9260115	C	0,3150	G	0,6850	promoter
29910033	-298	rs41541221	G	0,9387	A	0,0613	promoter
29910034	-297	rs2735114	G	0,5882	A	0,4118	promoter
29910057	-274	rs41272547	C	0,9265	A	0,0735	promoter
29910083	-248	rs3132688	A	0,0980	G	0,9020	promoter
29910086	-245	.	C	0,9988	T	0,0012	promoter
29910105	-226	rs9260116	G	0,9534	A	0,0466	promoter
29910107	-224	.	A	0,9988	G	0,0012	promoter
29910122	-209	rs11964254	C	0,9424	A	0,0576	promoter
29910132	-199	rs187514143	G	0,9988	A	0,0012	promoter
29910135	-196	rs115743151	T	0,9963	G	0,0037	promoter
29910159	-172	.	C	0,9975	A	0,0025	promoter
29910167	-164	rs2735113	C	0,9179	G	0,0821	promoter
29910176	-155	rs9260118	C	0,3137	T	0,6863	promoter
29910189	-142	rs9260119	A	0,3137	T	0,6863	promoter
29910227	-104	rs193269784	C	0,9988	T	0,0012	promoter
29910265	-66	rs41545520	G	0,9387	T	0,0613	promoter
29910276	-55	rs41560714	T	0,4596	G	0,5404	promoter
29910277	-54	rs9260120	G	0,9387	C	0,0613	promoter
29910286	-45	rs2916801	C	0,1556	T	0,8444	promoter
29910324	-7	rs114945359	G	0,9387	C	0,0613	promoter
29910330	-1	rs112877416	G	0,9914	A	0,0086	promoter
29910343	13	rs41541913	G	0,9988	C	0,0012	exon 1
29910358	28	rs1143146	C	0,4706	G	0,5294	exon 1
29910371	41	rs2230954	C	0,8444	T	0,1556	exon 1
29910378	48	rs41546314	C	0,7341	T	0,2659	exon 1
29910397	67	rs41564012	T	0,9902	A	0,0098	exon 1

29910402	72	rs1136656	G	0,8529	A	0,1471			exon 1
29910419	89	rs9260122	G	0,8456	T	0,1544			intron 1
29910420	91	rs9278464	GA	0,8456	G	0,1544			intron 1
29910429	99	rs9260123	C	0,4669	G	0,5331			intron 1
29910437	107	rs9260124	C	0,6152	T	0,3848			intron 1
29910450	120	rs17885299	G	0,7341	C	0,2659			intron 1
29910457	127	rs9260125	T	0,3100	G	0,6900			intron 1
29910460	130	rs9260126	T	0,7230	C	0,2770			intron 1
29910468	138	rs41557221	G	0,9522	A	0,0478			intron 1
29910473	143	rs41561116	G	0,8542	A	0,1458			intron 1
29910478	148	rs9260127	G	0,3113	C	0,6887			intron 1
29910482	152	rs9256980	G	0,3113	A	0,5306	T	0,1581	intron 1
29910492	162	rs9260128	G	0,8824	T	0,1176			intron 1
29910496	166	rs2904759	C	0,0993	G	0,9007			intron 1
29910506	176	rs41549214	G	0,9424	T	0,0576			intron 1
29910507	177	rs41554614	C	0,9559	T	0,0441			intron 1
29910508	178	rs9260129	A	0,3137	G	0,6863			intron 1
29910510	180	rs41541212	G	0,9240	A	0,0760			intron 1
29910522	192	rs9260130	G	0,3137	C	0,6863			intron 1
29910527	197	rs17879990	C	0,5882	T	0,4118			intron 1
29910538	208	rs1136657	C	0,7341	T	0,2659			exon 2
29910557	227	rs1136659	T	0,8799	A	0,1201			exon 2
29910558	228	rs2075684	T	0,4792	C	0,3235	A	0,1973	exon 2
29910562	232	rs1059418	A	0,8027	C	0,1789	T	0,0184	exon 2
29910566	236	rs41545116	G	0,9718	A	0,0282			exon 2
29910568	238	rs41555114	G	0,9988	C	0,0012			exon 2
29910581	251	rs1059423	C	0,9424	A	0,0576			exon 2
29910583	253	rs1059426	C	0,9424	T	0,0576			exon 2
29910586	256	rs12721672	G	0,9424	A	0,0576			exon 2
29910604	274	rs12721675	C	0,6752	A	0,3248			exon 2
29910623	293	rs1059438	A	0,9988	T	0,0012			exon 2
29910636	306	rs1059441	G	0,9988	A	0,0012			exon 2
29910640	310	rs3173429	C	0,9179	T	0,0821			exon 2
29910660	330	rs41559117	A	0,9755	G	0,0245			exon 2
29910663	333	rs707910	G	0,9032	A	0,0968			exon 2
29910679	349	rs199474404	G	0,9534	A	0,0466			exon 2
29910698	368	rs1059449	G	0,8983	A	0,1017			exon 2

29910699	369	rs1059451	G	0,9988	A	0,0012			exon 2
29910700	370	rs1059452	G	0,7341	T	0,2659			exon 2
29910703	373	rs41541222	G	0,8983	T	0,1017			exon 2
29910716	386	rs1059455	C	0,5870	G	0,4130			exon 2
29910717	387	rs1064588	A	0,5404	G	0,4130	T	0,0466	exon 2
29910719	389	rs2230991	G	0,8064	A	0,1471	C	0,0466	exon 2
29910721	391	rs199474424	G	0,8529	C	0,1471			exon 2
29910725	395	rs199474430	C	0,8542	G	0,1458			exon 2
29910730	400	rs199474436	T	0,5895	A	0,4105			exon 2
29910731	401	rs79361534	G	0,9032	A	0,0968			exon 2
29910742	412	rs78306866	G	0,3150	C	0,6850			exon 2
29910750	420	rs199474457	C	0,9167	T	0,0833			exon 2
29910752	422	rs1136683	G	0,7426	C	0,2561	A	0,0012	exon 2
29910759	429	rs1071742	T	0,6189	A	0,2059	C	0,1752	exon 2
29910761	431	rs1136688	G	0,6189	A	0,3811			exon 2
29910762	432	rs2231004	A	0,9694	G	0,0306			exon 2
29910767	437	rs1136689	G	0,8235	C	0,1765			exon 2
29910770	441	rs755077203	ACCC	0,8235	A	0,1765			exon 2
29910774	445	rs778137446	TG	0,8235	T	0,1765			exon 2
29910779	451	rs775073538	G	0,8235	GCTCC	0,1765			exon 2
29910801	471	rs1136692	C	0,7855	A	0,2145			exon 2
29910816	486	rs41562315	G	0,9265	A	0,0735			intron 2
29910817	487	rs41559412	G	0,9755	C	0,0245			intron 2
29910820	490	rs9260142	G	0,1654	C	0,8346			intron 2
29910822	492	rs41541123	G	0,9963	T	0,0037			intron 2
29910830	501	rs281864745	GTCAGGACC	0,9988	G	0,0012			intron 2
29910834	504	rs63225861	G	0,1078	C	0,8909	*	0,0012	intron 2
29910839	509	rs9260144	C	0,5870	T	0,4130			intron 2
29910849	519	rs377364793	C	0,9988	T	0,0012			intron 2
29910861	531	rs41555717	A	0,8076	G	0,1924			intron 2
29910865	535	rs41540614	C	0,9387	G	0,0613			intron 2
29910866	536	rs41543916	G	0,9265	A	0,0735			intron 2
29910892	562	rs9260145	A	0,6164	G	0,3836			intron 2
29910910	580	rs9260146	C	0,9032	T	0,0968			intron 2
29910923	593	rs9260147	C	0,8456	A	0,0576	T	0,0968	intron 2
29910947	617	rs41545815	T	0,8542	A	0,1458			intron 2
29910949	619	rs189287094	C	0,9988	G	0,0012			intron 2

29910980	650	rs41552627	C	0,9424	T	0,0576			intron 2
29910986	656	rs17882350	G	0,7328	A	0,2672			intron 2
29910990	660	rs539098093	T	0,9988	G	0,0012			intron 2
29911000	670	rs9256982	T	0,1569	G	0,4694	C	0,3738	intron 2
29911003	673	rs9260148	G	0,8297	A	0,1703			intron 2
29911005	675	rs41547618	G	0,9375	A	0,0625			intron 2
29911011	681	rs41562213	C	0,9350	T	0,0650			intron 2
29911019	689	rs9260149	T	0,6238	C	0,3762			intron 2
29911029	699	rs9260150	G	0,8873	T	0,1127			intron 2
29911030	700	rs9260151	C	0,8309	T	0,1691			intron 2
29911037	707	rs9260152	G	0,6397	C	0,3603			intron 2
29911038	708	rs41554914	G	0,9975	A	0,0025			intron 2
29911056	726	rs1071743	A	0,5882	C	0,1838	G	0,2279	exon 3
29911063	733	rs199474485	T	0,6752	G	0,3248			exon 3
29911064	734	rs1136695	A	0,3333	G	0,6667			exon 3
29911069	739	rs1136697	A	0,8493	T	0,1507			exon 3
29911077	747	rs199474490	G	0,9963	C	0,0037			exon 3
29911086	756	rs1136700	T	0,7279	C	0,2721			exon 3
29911092	762	rs1136702	G	0,7328	T	0,2672			exon 3
29911098	768	rs1059488	T	0,9645	C	0,0355			exon 3
29911103	773	rs41563718	C	0,9755	G	0,0245			exon 3
29911111	782	rs757867739	ACC	0,9424	A	0,0576			exon 3
29911114	784	rs3173420	G	0,3419	T	0,0576	A	0,6005	exon 3
29911115	785	rs1059498	G	0,5012	GAA	0,0576	C	0,4412	exon 3
29911119	789	rs3173419	G	0,5012	C	0,0576	T	0,4412	exon 3
29911120	790	rs41557212	A	0,9988	T	0,0012			exon 3
29911124	794	rs3173418	C	0,9412	T	0,0588			exon 3
29911149	819	rs1059506	C	0,8444	T	0,1556			exon 3
29911154	824	rs1059509	C	0,5294	A	0,4706			exon 3
29911178	848	rs41542714	G	0,9975	A	0,0025			exon 3
29911190	860	rs1059514	G	0,5172	A	0,4828			exon 3
29911198	868	rs1059516	T	0,6752	C	0,3248			exon 3
29911203	873	rs1059517	A	0,6728	C	0,3260	G	0,0012	exon 3
29911207	877	rs1059520	G	0,6752	A	0,3248			exon 3
29911218	888	rs1059526	G	0,9412	A	0,0588			exon 3
29911222	892	rs1059535	C	0,9032	T	0,0968			exon 3
29911225	895	rs1059536	A	0,7377	G	0,2623			exon 3

29911227	897	rs1059537	G	0,9461	T	0,0282	C	0,0257			exon 3
29911228	898	rs9256983	A	0,1520	G	0,0539	T	0,6360	C	0,1581	exon 3
29911239	909	rs9260155	T	0,7426	C	0,2574					exon 3
29911240	910	rs9260156	T	0,6066	G	0,2341	A	0,1593			exon 3
29911246	916	rs9260157	C	0,9032	T	0,0968					exon 3
29911256	926	rs1059542	T	0,0980	G	0,9020					exon 3
29911260	930	rs543623321	A	0,8039	C	0,1912	G	0,0049			exon 3
29911261	931	.	C	0,8039	G	0,1912	A	0,0049			exon 3
29911271	941	rs3098019	G	0,7708	C	0,2292					exon 3
29911272	942	.	T	0,7708	G	0,2292					exon 3
29911284	954	rs41563116	T	0,9841	C	0,0159					exon 3
29911326	986	rs2517725	A	0,9534	G	0,0466					intron 3
29911334	994	rs570061466	C	0,9988	T	0,0012					intron 3
29911344	1004	rs17878838	C	0,7953	T	0,0588	A	0,1458			intron 3
29911356	1016	rs3132687	G	0,8419	A	0,1581					intron 3
29911360	1020	rs17878860	A	0,8542	G	0,1458					intron 3
29911419	1079	rs2735112	A	0,3885	G	0,6115					intron 3
29911431	1091	rs186459750	G	0,9988	C	0,0012					intron 3
29911457	1118	rs140855897	G	0,0980	GT	0,9020					intron 3
29911466	1127	rs559107954	C	0,9988	T	0,0012					intron 3
29911505	1166	rs41543412	C	0,9265	G	0,0735					intron 3
29911510	1171	rs41553212	A	0,9755	T	0,0245					intron 3
29911522	1187	rs562709947	T	0,9988	TA	0,0012					intron 3
29911534	1195	rs75439900	A	0,8542	C	0,1458					intron 3
29911538	1199	rs2735111	G	0,3137	A	0,6863					intron 3
29911545	1206	rs76475517	G	0,8542	A	0,1458					intron 3
29911571	1232	rs29028728	G	0,8542	C	0,1458					intron 3
29911583	1245	rs139109856	G	0,7843	GACAC	0,2157					intron 3
29911587	1249	rs9280789	C	0,6777	CA	0,3223					intron 3
29911588	1246	rs371311437	C	0,6752	CACA	0,3248					intron 3
29911592	1253	rs549627736	A	0,9988	G	0,0012					intron 3
29911604	1265	rs77467746	C	0,8542	A	0,1458					intron 3
29911627	1288	rs182296210	T	0,9988	C	0,0012					intron 3
29911650	1311	rs188062958	G	0,9988	A	0,0012					intron 3
29911684	1345	rs9260164	C	0,8407	T	0,1593					intron 3
29911685	1346	rs17879103	C	0,7574	T	0,2426					intron 3
29911727	1388	rs29028878	T	0,3137	C	0,6863					intron 3

29911728	1389	rs41561516	C	0,4596	T	0,5392	G	0,0012	intron 3
29911732	1397	.	G	0,4608	GGGACTAGAATTTTCCAC	0,5392			intron 3
29911740	1401	rs2508038	A	0,4608	G	0,5392			intron 3
29911770	1432	rs549444296	GGT	0,9988	G	0,0012			intron 3
29911790	1451	rs9260166	T	0,4608	C	0,5392			intron 3
29911803	1464	rs9260167	G	0,4596	A	0,5404			intron 3
29911818	1479	rs188231435	T	0,9988	C	0,0012			intron 3
29911824	1485	rs193073597	G	0,9988	A	0,0012			intron 3
29911827	1488	rs9260168	G	0,4608	A	0,5392			intron 3
29911836	1497	rs9260169	C	0,4608	T	0,5392			intron 3
29911837	1498	rs17879578	G	0,8542	A	0,1458			intron 3
29911844	1505	rs9260170	T	0,4608	A	0,5392			intron 3
29911857	1518	rs70944751	G	0,9265	T	0,0735			intron 3
29911872	1533	rs374722929	C	0,9988	G	0,0012			intron 3
29911879	1540	rs9260171	T	0,4608	G	0,5392			intron 3
29911880	1541	rs17885212	T	0,6752	A	0,3248			intron 3
29911893	1554	rs9260172	C	0,4608	T	0,5392			intron 3
29911895	1556	rs9260173	T	0,4608	A	0,5392			intron 3
29911901	1562	rs1136741	C	0,5343	G	0,4657			exon 4
29911908	1569	rs41557315	A	0,9841	G	0,0159			exon 4
29911909	1570	rs1136746	G	0,6752	A	0,3248			exon 4
29911912	1573	rs1136749	A	0,4608	G	0,5392			exon 4
29911921	1582	rs1136754	C	0,4608	T	0,5392			exon 4
29911924	1585	rs45520243	C	0,9988	T	0,0012			exon 4
29911928	1589	rs1059563	C	0,4608	G	0,5392			exon 4
29911930	1591	rs1059564	C	0,4608	T	0,5392			exon 4
29911931	1592	rs9260179	A	0,4608	G	0,5392			exon 4
29911945	1606	rs12721753	G	0,6752	A	0,3248			exon 4
29911951	1612	rs1059570	C	0,8529	T	0,1471			exon 4
29911957	1618	rs12721755	G	0,8542	A	0,1458			exon 4
29911970	1631	rs9260180	G	0,4596	A	0,5404			exon 4
29911972	1633	.	C	0,9988	G	0,0012			exon 4
29912041	1702	rs9260181	C	0,7721	T	0,2279			exon 4
29912068	1729	rs145353606	A	0,9534	G	0,0466			exon 4
29912085	1746	rs148794540	C	0,9424	T	0,0576			exon 4
29912086	1747	rs199474634	G	0,8542	A	0,1458			exon 4
29912087	1748	rs1059632	G	0,7855	T	0,2145			exon 4

29912098	1759	rs142475339	G	0,8529	A	0,1471			exon 4
29912108	1769	rs2231095	G	0,4596	C	0,5392	A	0,0012	exon 4
29912114	1775	rs1059644	C	0,9988	A	0,0012			exon 4
29912147	1808	rs1059686	C	0,6752	T	0,3248			exon 4
29912149	1810	rs114818671	G	0,9265	C	0,0735			exon 4
29912153	1814	rs1136835	A	0,9988	G	0,0012			exon 4
29912179	1840	rs567795363	G	0,9988	A	0,0012			intron 4
29912188	1849	rs9260184	T	0,6164	C	0,3836			intron 4
29912204	1865	rs2276447	C	0,4608	T	0,5392			intron 4
29912208	1869	rs9260186	G	0,9645	C	0,0355			intron 4
29912227	1890	rs146444598	TGGA	0,4608	T	0,5392			intron 4
29912249	1910	rs188245242	G	0,9988	A	0,0012			intron 4
29912280	1941	rs1136903	T	0,3137	C	0,6863			exon 5
29912281	1942	rs1136917	G	0,8529	A	0,1471			exon 5
29912297	1958	rs1136949	A	0,8542	G	0,1458			exon 5
29912301	1962	rs41558618	C	0,9608	A	0,0392			exon 5
29912305	1966	rs41555922	C	0,9988	T	0,0012			exon 5
29912326	1987	rs1059846	G	0,8444	A	0,1556			exon 5
29912333	1994	rs3179982	C	0,4608	T	0,5392			exon 5
29912342	2003	rs1059896	G	0,9645	A	0,0355			exon 5
29912345	2006	rs3180278	A	0,8444	T	0,1556			exon 5
29912348	2009	rs1137078	A	0,7843	G	0,2157			exon 5
29912368	2029	rs1141891	C	0,3150	T	0,6850			exon 5
29912373	2034	rs1137110	T	0,8444	G	0,1556			exon 5
29912382	2043	rs41550119	G	0,9988	A	0,0012			exon 5
29912386	2047	rs1137160	G	0,8542	C	0,1458			exon 5
29912395	2057	rs376806238	T	0,8444	TGG	0,1556			intron 5
29912398	2059	rs9260191	A	0,4608	G	0,3836	T	0,1556	intron 5
29912405	2066	rs9260192	G	0,8591	A	0,1409			intron 5
29912410	2071	rs41551619	G	0,9387	A	0,0613			intron 5
29912413	2078	rs9278466	TG	0,4608	T	0,5392			intron 5
29912424	2085	.	A	0,9988	G	0,0012			intron 5
29912451	2112	rs9260193	C	0,4608	A	0,5392			intron 5
29912457	2118	rs9260194	C	0,4608	G	0,5392			intron 5
29912463	2124	rs9260195	A	0,4608	G	0,5392			intron 5
29912473	2134	rs9260196	T	0,4608	C	0,5392			intron 5
29912476	2137	.	C	0,9988	G	0,0012			intron 5

29912477	2138	rs9260197	A	0,4608	G	0,5392			intron 5
29912480	2141	rs183484257	A	0,9988	C	0,0012			intron 5
29912493	2154	rs1655894	G	0,0980	A	0,7439	T	0,1581	intron 5
29912494	2155	rs2975059	T	0,6752	C	0,3248			intron 5
29912502	2163	rs9260198	C	0,4608	T	0,5392			intron 5
29912506	2167	rs9260199	G	0,9412	A	0,0588			intron 5
29912510	2171	rs1632885	T	0,6949	G	0,3039	C	0,0012	intron 5
29912517	2178	rs1655895	T	0,0980	C	0,9020			intron 5
29912562	2223	rs17880291	T	0,8529	G	0,1471			intron 5
29912564	2225	rs17885126	A	0,8529	C	0,1471			intron 5
29912594	2255	rs17882753	C	0,7794	T	0,2206			intron 5
29912595	2256	rs2975058	G	0,6765	A	0,3235			intron 5
29912598	2259	rs41551518	G	0,9706	A	0,0294			intron 5
29912613	2274	rs189337983	T	0,9951	A	0,0049			intron 5
29912614	2275	rs41545117	C	0,9387	A	0,0613			intron 5
29912649	2310	rs17887246	G	0,8542	C	0,1458			intron 5
29912652	2314	rs67131684	CAGA	0,4608	C	0,5392			intron 5
29912656	2317	rs17885779	C	0,8542	T	0,1458			intron 5
29912657	2318	rs9260200	C	0,4608	T	0,5392			intron 5
29912668	2329	rs9260201	T	0,3137	G	0,6863			intron 5
29912669	2330	rs9260202	A	0,4596	G	0,5404			intron 5
29912680	2341	rs17884187	A	0,8542	G	0,1458			intron 5
29912687	2348	rs17883986	C	0,8542	T	0,1458			intron 5
29912713	2374	rs17879496	C	0,8493	T	0,1507			intron 5
29912728	2389	rs141912766	G	0,9988	T	0,0012			intron 5
29912738	2399	rs150666121	T	0,9914	A	0,0086			intron 5
29912754	2415	rs3823323	G	0,4608	A	0,5392			intron 5
29912763	2424	rs548326797	A	0,9988	G	0,0012			intron 5
29912766	2427	rs3823324	A	0,4608	T	0,5392			intron 5
29912767	2428	rs3823325	G	0,9988	C	0,0012			intron 5
29912796	2457	rs2735103	G	0,6716	A	0,3284			intron 5
29912810	2471	rs184394934	T	0,9988	G	0,0012			intron 5
29912815	2476	rs200725098	C	0,9951	T	0,0049			intron 5
29912830	2491	rs2571381	T	0,1556	C	0,8444			intron 5
29912836	2497	rs41551512	A	0,9988	T	0,0012			exon 6
29912852	2513	rs1137296	T	0,3137	C	0,6863			exon 6
29912856	2517	rs2231119	A	0,3137	T	0,6863			exon 6

29912885	2546	rs139910379	C	0,9265	A	0,0735			intron 6
29912890	2551	rs41546014	G	0,9265	C	0,0735			intron 6
29912892	2553	rs3823329	C	0,9265	A	0,0735			intron 6
29912893	2554	rs3823330	T	0,9265	A	0,0735			intron 6
29912896	2557	rs41549018	G	0,9265	A	0,0735			intron 6
29912897	2558	rs41555715	G	0,9265	A	0,0735			intron 6
29912900	2561	rs41558113	C	0,9265	A	0,0735			intron 6
29912901	2562	rs41561414	T	0,9265	C	0,0735			intron 6
29912914	2575	rs541705597	T	0,9988	G	0,0012			intron 6
29912925	2586	rs2735102	A	0,7855	G	0,2145			intron 6
29912935	2596	rs2916802	C	0,0833	T	0,9167			intron 6
29912938	2599	rs147947139	C	0,8542	A	0,1458			intron 6
29912944	2605	rs2916803	T	0,0980	C	0,9020			intron 6
29912968	2629	rs3823339	C	0,9534	G	0,0466			intron 6
29913001	2662	rs2735101	T	0,7855	C	0,2145			intron 6
29913007	2668	rs200140272	C	0,9988	T	0,0012			intron 6
29913037	2698	rs1137631	G	0,7855	A	0,2145			exon 7
29913042	2703	rs14288	C	0,0980	T	0,9020			exon 7
29913067	2728	rs3823342	T	0,4596	C	0,5404			intron 7
29913074	2735	rs9260205	A	0,1556	G	0,8444			intron 7
29913078	2739	rs41546815	A	0,1556	G	0,8444			intron 7
29913098	2759	rs3129015	C	0,9020	T	0,0968	G	0,0012	intron 7
29913099	2760	rs41553916	G	0,9755	A	0,0245			intron 7
29913111	2772	rs3823343	A	0,7855	G	0,2145			intron 7
29913127	2788	rs367651478	G	0,9988	A	0,0012			intron 7
29913135	2796	rs28749156	G	0,6752	C	0,3248			intron 7
29913137	2798	rs3132686	A	0,9032	G	0,0968			intron 7
29913171	2832	rs2975057	G	0,3137	A	0,6863			intron 7
29913175	2836	rs114353087	T	0,9975	A	0,0025			intron 7
29913179	2840	rs2975056	C	0,7855	T	0,2145			intron 7
29913191	2852	rs2975055	T	0,3137	C	0,6850	G	0,0012	intron 7
29913192	2853	rs9260210	G	0,3137	A	0,6850	T	0,0012	intron 7
29913197	2858	rs2975054	T	0,4608	C	0,5392			intron 7
29913201	2862	rs41557912	T	0,8542	G	0,1458			intron 7
29913213	2875	rs35830199	TA	0,8542	T	0,1458			intron 7
29913214	2875	.	ATT	0,8529	AT	0,0012	*	0,1458	intron 7
29913266	2927	rs1061156	G	0,7855	T	0,2145			3'UTR

29913298	2959	rs1061235	A	0,9265	T	0,0735			3'UTR
29913343	3004	rs578027483	C	0,9988	G	0,0012			3'UTR
29913344	3005	rs13395	G	0,6397	A	0,1458	T	0,2145	3'UTR
29913365	3026	rs181754482	G	0,9988	A	0,0012			3'UTR
29913367	3028	rs113042145	G	0,9988	A	0,0012			3'UTR
29913374	3035	rs1061815	C	0,9032	G	0,0968			3'UTR
29913386	3047	rs41545022	C	0,9387	G	0,0613			3'UTR
29913430	3091	rs112863520	C	0,8419	T	0,1581			3'UTR
29913434	3095	rs186532503	A	0,9988	G	0,0012			3'UTR
29913449	3110	rs66585529	C	0,8542	G	0,1458			3'UTR
29913483	3144	rs79244404	T	0,7855	C	0,2145			3'UTR
29913501	3162	rs144149732	C	0,9988	T	0,0012			3'UTR
29913515	3176	rs112525507	C	0,9400	T	0,0600			3'UTR
29913539	3200	rs577937205	C	0,9988	A	0,0012			3'UTR
29913542	3203	rs2499	T	0,0980	G	0,9020			3'UTR
29913554	3215	rs9385	A	0,9951	C	0,0049			3'UTR
29913588	3249	rs573809399	G	0,9988	C	0,0012			3'UTR
29913630	3291	rs2975053	A	0,8419	T	0,0968	C	0,0613	3'UTR
29913638	3299	rs13488	G	0,1556	C	0,8444			3'UTR
29913639	3300	rs2975052	A	0,8419	G	0,1581			3'UTR

Table S2: List of all *HLA-A* protein molecules detected in a Brazilian population sample from Southeast Brazil.

<i>HLA-A</i> coding allele (only CDS) ^a	Frequency <i>2n=816</i>
A*02:01	0.2071
A*03:01	0.0931
A*24:02	0.0907
A*01:01	0.0833
A*11:01	0.0600
A*31:01	0.0441
A*29:02	0.0429
A*23:01	0.0392
A*26:01	0.0306
A*30:01	0.0282
A*68:01	0.0282
A*68:02	0.0282
A*30:02	0.0245
A*32:01	0.0245
A*02:05	0.0172
A*33:01	0.0159
A*23:17	0.0135
A*33:03	0.0135
A*36:01	0.0135
A*66:01	0.0110
A*02:11	0.0098
A*74:01	0.0098
A*02:04 (codon119:V/L)	0.0086
A*34:02	0.0086
A*02:02	0.0061
A*02:04	0.0061
A*02:17	0.0061
A*03:02	0.0049
A*25:01	0.0049
A*30:04	0.0049
A*29:01	0.0037
A*66:02	0.0025
A*02:01 (codon170:S/R)	0.0012
A*02:02 (codon119:L/I)	0.0012
A*02:06	0.0012
A*02:20	0.0012
A*11:248	0.0012
A*23:17 (codon121:M/I)	0.0012
A*24:03	0.0012
A*32:02	0.0012
A*66:03	0.0012
A*68:17	0.0012
A*69:01	0.0012
A*80:01	0.0012

^a *HLA-A* protein molecules were named according to the IPD-IMGT/HLA database version 3.30.0, followed by any divergences eventually observed. Codon positions refer to the first translated Methionine as codon 1.