



**UNIVERSIDADE ESTADUAL PAULISTA
“JÚLIO DE MESQUITA FILHO”
FACULDADE DE MEDICINA**

Nayane dos Santos Brito Silva

**Diversidade genética do gene *HLA-B* em populações
com diferentes perfis de ancestralidade**

Dissertação apresentada à Faculdade de
Medicina, Universidade Estadual Paulista “Júlio
de Mesquita Filho”, Câmpus de Botucatu, para
obtenção do título de Mestre(a) em Patologia.

Orientador: Prof. Dr. Erick da Cruz Castelli

Coorientadora: Profa. Dra. Camila Ferreira Bannwart Castro

**Botucatu
2021**

Nayane dos Santos Brito Silva

Diversidade genética do gene *HLA-B* em populações
com diferentes perfis de ancestralidade

Dissertação apresentada à Faculdade de
Medicina, Universidade Estadual Paulista
“Júlio de Mesquita Filho”, Câmpus de
Botucatu, para obtenção do título de
Mestre(a) em Patologia.

Orientador: Prof. Dr. Erick da Cruz Castelli

Coorientadora: Profa. Dra. Camila Ferreira Bannwart Castro

Botucatu
2021

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO TÉC. AQUIS. TRATAMENTO DA INFORM.
DIVISÃO TÉCNICA DE BIBLIOTECA E DOCUMENTAÇÃO - CÂMPUS DE BOTUCATU - UNESP
BIBLIOTECÁRIA RESPONSÁVEL: ROSEMEIRE APARECIDA VICENTE-CRB 8/5651

Silva, Nayane dos Santos Brito.

Diversidade genética do gene HLA-B em populações com diferentes perfis de ancestralidade / Nayane dos Santos Brito Silva. - Botucatu, 2021

Dissertação (mestrado) - Universidade Estadual Paulista "Júlio de Mesquita Filho", Faculdade de Medicina de Botucatu

Orientador: Erick da Cruz Castelli

Coorientador: Camila Ferreira Bannwart Castro

Capes: 21103003

1. Diversidade genética. 2. Variação biológica da população. 3. Genética de populações. 4. Antígenos HLA-B.

Palavras-chave: Brasil; HLA-B; NGS; Populações mundiais; Variabilidade.

Dedicatória

Dedico esse trabalho à minha professora Camila, que com tanto carinho e dedicação inspira os seus alunos. Por sempre acreditar em mim e por todo apoio que me deu desde o início.

“Todo profissional é prosa. Só o professor é poesia.”

- Leonardo Lisboa Bcena

Agradecimentos

“Diante da vastidão do tempo e da imensidão do universo, é um imenso prazer para mim dividir um planeta e uma época com você.”

-CARL SAGAN

Muitos foram aqueles que contribuíram para realização desse trabalho e para os quais deixo o meu agradecimento.

Primeiramente a Deus, pela oportunidade de estar no mundo. Pelo dom da vida e pelo seu amor infinito.

A minha mãe Josefa, que não terminou o ensino fundamental, mas já formou três filhas. Pela sua força e coragem, por toda a sua trajetória que não foi fácil, mas me permitiu chegar até aqui.

Ao meu querido padrasto Esmeraldo, que amou a mim e as minhas irmãs como se fossem suas próprias filhas e que sempre esteve presente. Já faz algum tempo que você partiu, mas eu não poderia deixar de te agradecer por todo amor, toda garra e por sempre acreditar em um futuro melhor.

Aos meus familiares, Cláudio Brescansim, Fernanda S. Nunes de Souza, Elaine C. dos Santos Almeida, Elaine C. dos Santos Marcusso, Eriton A. Marcusso, Hermano R. Almeida e José Antônio Brito Silva que me deram o suporte necessário para chegar até aqui. Em especial aos meus sobrinhos, Davi Santos de Almeida e Sara Santos Marcusso, por despertarem em mim o amor mais puro e os meus sorrisos mais sinceros.

Ao meu namorado Leonardo da Cunha Menezes Souza, pelo companheirismo, carinho e cuidado comigo, por tornar a vida e a quarentena mais leves, por todas as risadas que tirou de mim e lágrimas que limpou do meu rosto. Obrigada por ser sempre o meu anjo.

Ao meu querido amigo Italo Bertoni, que está comigo desde a graduação, por todas as conversas, todas as risadas e todos os desafios que passamos juntos. Trilhar mais essa etapa ao seu lado tornou tudo mais fácil. Obrigada pelo companheirismo que se começou na faculdade e se estendeu para vida.

À minha querida amiga Simone Mafra, por ter sentido comigo, todas as angustias e felicidades, acompanhando cada passo de perto, pela amizade e pela torcida positiva. Muito obrigada.

Aos meus amigos do GemBio – Raphaela Neto Pereira, Thállita H. Ayala Lima, Marília R. Passos, Emiliana Weiss, Hecttor S. Baptista, Arielle Lima da Rocha, Gabriela P. Carvalho, Gabi Sato e Isa – por todos os ensinamentos, pela amizade e pelas agradáveis lembranças que serão eternamente guardadas no coração. Em especial à Andréia Souza, por sua luz que ilumina todos a sua volta, por todo carinho e toda ajuda desde o começo; e à Helô S. Andrade, que sempre me ajudou e me incentivou, obrigada por tudo, você é muito especial para mim.

Ao meu orientador Prof. Dr. Erick Castelli, pela oportunidade e pela excelente orientação a qual se dedicou a ensinar e compartilhar seus conhecimentos. Obrigada pela atenção e paciência.

Ao programa de Pós-Graduação em Patologia da Faculdade de Medicina de Botucatu (FMB-UNESP), pela oportunidade de construir minha formação científica e a todo o corpo docente pelos ensinamentos nas disciplinas do programa.

Aos membros da banca avaliadora, por aceitarem o convite e se disponibilizarem para contribuir com a minha formação e melhoria desse trabalho.

Ao CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) pela bolsa de mestrado que possibilitou a dedicação integral ao programa de pós-graduação.

A todos que colaboraram direta ou indiretamente para concretização desse trabalho, o meu muito obrigada!

SUMÁRIO

Lista de Figuras.....	10
Lista de Tabelas.....	10
Lista de Abreviaturas e Siglas.....	11
RESUMO.....	13
ABSTRACT.....	14
Capítulo I – Revisão de Literatura.....	15
INTRODUÇÃO.....	16
Estrutura e função das moléculas HLA de classe I.....	19
HLA-B.....	24
JUSTIFICATIVA.....	28
OBJETIVOS.....	29
Objetivo Geral.....	29
Objetivos Específicos.....	29
REFERÊNCIAS.....	30
Capítulo II – Artigo.....	35
Abstract.....	37
1.Introduction.....	38
2.Methods.....	39
2.1.Description of each population and samples used to evaluate HLA-B.....	39
2.1.1. Brazilian samples and WGS sequencing	39
2.1.2.Amerindians samples.....	40
2.1.3. 1000genomes samples.....	40
2.2.Raw data processing.....	40
2.3.Genotype and haplotype calling.....	41
2.4. Allele calling in the genomic, exonic, and allotype levels.....	41
2.5. Statistical analysis.....	42
3.Results.....	42
3.1.Genotyping and haplotyping accuracy.....	42
3.2.HLA-B SNPs in different populations, SNP density, and nucleotide diversity across HLA-B.....	44
3.3.HLA-B allotype frequencies in different populations.....	47
3.4.HLA-B exonic and intronic sequences in different populations.....	48

3.5.HLA-B regulatory sequences.....	49
4.Discussion.....	49
References.....	56
Supplementary material.....	62
CONCLUSÃO.....	68

Lista de Figuras

Figura 1. Localização e organização do complexo HLA no cromossomo 6.	17
Figura 2. Estrutura das moléculas HLA Classe I.	19
Figura 3. Processamento e apresentação de antígeno via HLA de classe I.	21
Figura 4. Representação esquemática do reconhecimento dos resíduos polimórficos e peptídico do complexo peptídeo/HLA pelo receptor da célula T.	23

Lista de Figuras do Artigo

Figure 1. Genotype compatibility across HLA-B when comparing the hla-mapper workflow and the original 1000Genomes.	43
Figure 2. Proportion of variants.	44
Figure 3. Multidimensional scaling of FST values.	45
Figure 4. <i>HLA-B</i> gene diversity across different populations.	46
Figure 5. Multidimensional scaling of FST values among populations considering allotype frequencies.	48

Lista de Figuras do Artigo – Material Suplementar

Figure S1. Linkage Disequilibrium (LD) between pair of single nucleotide polymorphisms (SNPs) of the <i>HLA-B</i> locus.	63
Figure S2. Venn Diagram. Allotypes sharing among super populations.	64
Figure S3. Network of the <i>HLA-B</i> promoter sequences.	65
Figure S4. Network of the <i>HLA-B</i> 3'UTR sequences.	66
Figure S5. Distribution of <i>HLA-B</i> *27, <i>HLA-B</i> *53:01, <i>HLA-B</i> *46:01 and <i>HLA-B</i> *57:01 worldwide.	67

Lista de Tabelas

Tabela 1. Número de alelos identificados para os genes HLA de classe I.	18
--	----

Lista de Abreviaturas e Siglas

APC	Célula Apresentadora de Antígeno (do inglês, <i>Antigen Presenting Cell</i>)
ATP	Adenosina Trifosfato (do inglês, <i>Adenosine triphosphate</i>)
BiP	Proteína Ligação de imunoglobulina (do inglês, <i>Binding Immunoglobulin Protein</i>)
CTL	Linfócito T Citotóxico
ERAP	Aminopeptidase do Reticulo Endoplasmático (do inglês, <i>HLAendoplasmic reticulum aminopeptidase</i>)
ERp57	p57 do Reticulo Endoplasmático
HIV	Vírus da Imunodeficiência Humana (do inglês, <i>Human Immunodeficiency Virus</i>)
HLA	Antígeno Leucocitário Humano (do inglês, <i>Human Leukocyte Antigens</i>)
Ig	Imunoglobulina
IgA	Imunoglobulina do tipo A
IGHA	Cadeia pesada alfa da imunoglobulina (do inglês, <i>Immunoglobulin Heavy Constant Alpha</i>)
IMGT	International Immunogenetics Database
KIR	Receptor do tipo imunoglobulina (KIR, do inglês <i>Killer Cell Immunoglobulin Like Receptor</i>)
LTA	Linfotoxina alfa
LTB	Linfotoxina beta
MHC	Complexo Principal de Histocompatibilidade (do inglês, <i>Major Histocompatibility complex</i>)
mRNA	Ácido Ribonucleico mensageiro ou RNA mensageiro
NK	Célula Assassina Natural (do inglês, <i>Natural Killer</i>)
NT	Não Traduzida
Pb	Pares de bases
PCR	Reação em Cadeia da Polimerase (do inglês, <i>Polymerase Chain Reaction</i>)
RE	Retículo Endoplasmático
SNP	Polimorfismo de base única (do inglês, <i>Single Nucleotide Polymorphism</i>)
TAP	Transportador associado ao processamento de antígeno (do inglês, <i>Transporter Associated with Antigen Processing</i>)

TCR Receptor de célula T

TNF Fator de Necrose Tumoral

UTR Região não traduzida (do inglês, *Untranslated region*)

RESUMO

O gene *HLA-B* é possivelmente o mais variável do genoma humano, com quase oito mil sequências diferentes descritas no banco de dados IPD-IMGT/HLA. Este gene codifica uma molécula fundamental para a apresentação do antígeno aos linfócitos T CD8 + e modulação das células NK. Embora largamente estudado, a maioria dos estudos avaliaram apenas éxons, geralmente negligenciando sequências intrônicas e regulatórias em um contexto populacional. Da mesma forma, muitos dos estudos utilizaram métodos que detectam apenas variantes conhecidas, provavelmente subestimando a variabilidade deste gene. Neste estudo utilizamos um método computacional adequado para avaliar a variabilidade genética do gene *HLA-B* completo (incluindo todos os éxons, íntrons e regiões regulatórias), a partir de dados de sequenciamento de nova geração em 3.648 amostras de 29 populações mundiais diferentes, em todos os níveis: SNPs, haplótipos, sequências completas e de proteínas codificadas. Foram detectados 610 sítios de variação ao longo do gene *HLA-B*. A diversidade nucleotídica foi alta nos éxons 1, 2 e 3, e também no íntron 1. A região entre o íntron 3 e o éxon 7 é a mais conservada, com baixa diversidade nucleotídica no éxon 4 e sem nenhuma variante no éxon 6. Detectamos 535 sequências completas do gene *HLA-B* e 191 sequências codificadoras (apenas éxons), as quais codificam 171 proteínas distintas. Nosso método computacional se mostrou eficaz na obtenção de genótipos e haplótipos acurados do gene *HLA-B*. A maioria dos SNPs são compartilhados entre todas as populações, enquanto o oposto é observado para os haplótipos e sequências de proteínas. No entanto, essas populações compartilham as mesmas regiões regulatórias. Além disso, a amostra brasileira introduziu novos alelos e aumentou a frequência de alelos raros que estavam presentes no *1000Genomes*.

Palavras-chave: *HLA-B*, variabilidade, NGS, Brasil, populações mundiais, *1000Genomes*.

ABSTRACT

HLA-B is possibly the most variable gene in the human genome, with almost eight different sequences described in the IPD-IMGT/HLA database. This gene encodes a key molecule for the presentation of the antigen to CD8 + T lymphocytes and modulation of NK cells. Despite its widespread study, most population-based studies of the *HLA-B* gene mainly focused on exons, ignoring intronic and regulatory sequences. Likewise, many of the studies used methods that detect only known variants, probably underestimating the variability of this gene. In this study, we use an appropriate computational method to assess the genetic variability of the complete *HLA-B* gene (including all exons, introns, and regulatory regions), using new generation sequencing data in 3,648 samples from 29 different world populations, in all levels: SNPs, haplotypes, complete sequences and encoded proteins. We detected 610 sites of variation along the *HLA-B* gene. Nucleotide diversity was high in exons 1, 2 and 3, and also in intron 1. The region between intron 3 and exon 7 is the most conserved, with low nucleotide diversity in exon 4 and with no variant in exon 6. We detected 535 complete *HLA-B* gene sequences and 191 coding sequences (exons only), which encode 171 distinct proteins. Our computational method proved to be effective in obtaining accurate genotypes and haplotypes of the *HLA-B* gene. Most SNPs are shared among all populations, while haplotypes and protein sequences are not. However, these populations share the same regulatory regions. Furthermore, the Brazilian sample introduced new alleles and increased the frequency of rare alleles that were present in 1000Genomes.

Keywords: *HLA-B*, variability, NGS, Brazil, worldwide populations, 1000genomes.

Capítulo I – Revisão de Literatura

INTRODUÇÃO

Os Antígenos Leucocitários Humanos (HLA, do inglês *Human Leucocytes Antigens*) são glicoproteínas transmembrânicas que se associam a peptídeos de origem endógena ou exógena e os apresentam aos linfócitos T, desempenhando um papel fundamental na resposta imunitária adaptativa e na susceptibilidade a diversas doenças autoimunes e degenerativas. Os genes HLA estão localizados dentro do Complexo Principal de Histocompatibilidade (MHC, do inglês *Major Histocompatibility Complex*), no braço curto do cromossomo seis, em 6p21.3. Esse complexo é composto por aproximadamente 250 genes codificadores e não codificantes, cuja organização estrutural evoluiu gradualmente, envolvendo vários eventos de mutação, duplicação, deleção e rearranjo genômico¹. Na década de 70, esse complexo gênico foi dividido didaticamente em três classes ou regiões, denominadas classe I, II e III (Figura 1). Os genes de classe III não expressam moléculas que participam da apresentação antigênica (de histocompatibilidade), mas codificam componentes do sistema complemento (e.g., C2 e C4) e citocinas (e.g., Fator de Necrose Tumoral, Linfotóxina alfa e beta). A região de classe II codifica as cadeias α e β das moléculas HLA de classe II, que são divididos em clássicos: HLA-DR, HLA-DQ e HLA-DP; e não clássicos: HLA-DM e HLA-DO. Estas apresentam antígenos extracelulares aos linfócitos T CD4 e são expressas em células apresentadoras de antígenos (APC, do inglês *Antigen Presenting Cell*), como as células dendríticas, macrófagos e linfócitos B¹.

Os genes de classe I subdividem-se em: (a) genes clássicos (*HLA-A*, *HLA-B* e *HLA-C*), responsáveis pela apresentação antigênica aos linfócitos T citotóxicos (CTLs), extremamente variáveis e expressos de forma constitutiva na membrana celular de quase todas as células nucleadas; e (b) genes não clássicos (*HLA-G*, *HLA-E* e *HLA-F*), pouco variáveis e expressos em alguns tecidos ou em baixos níveis, de propriedades imunomodulatórias por meio da interação com receptores específicos de células do sistema imune, tais como células *Natural Killer* (NK)^{1,2}. Há também nessa região os *loci* *HLA-H*, *HLA-J*, *HLA-K* e *HLA-L*, que são pseudogenes³

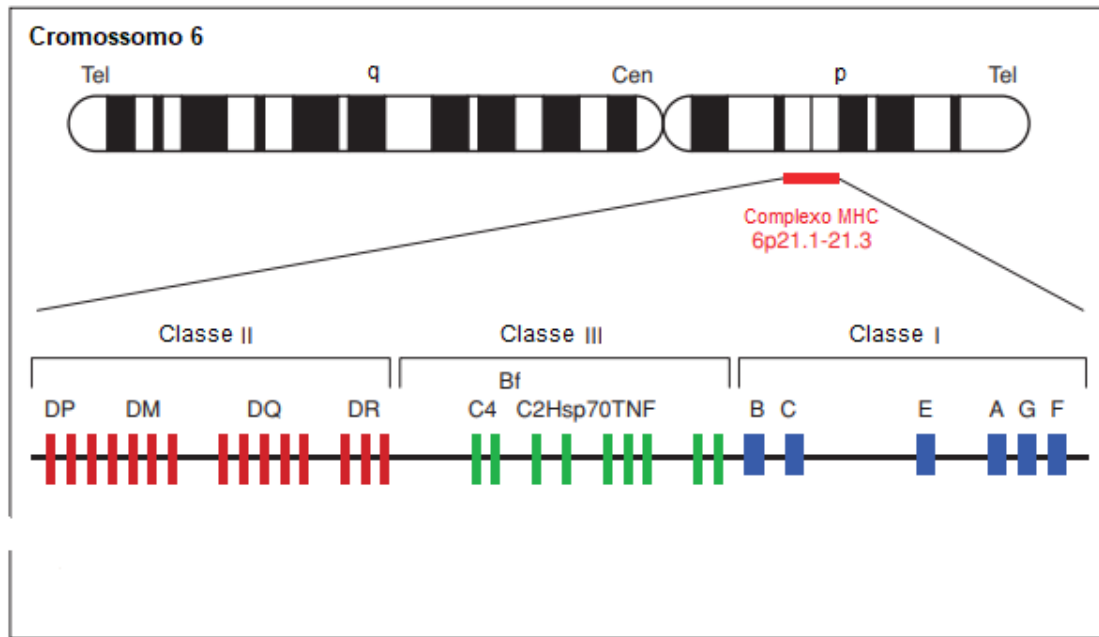


Figura 1. Localização e organização do complexo *MHC* no cromossomo 6. O complexo *MHC* é didaticamente dividido em três regiões: I, II e III. Nessa figura estão representados apenas os principais genes das diferentes classes. Fonte: Adaptado de Mehra e Kaur⁴

O complexo *HLA* compreende os genes mais polimórficos do genoma humano e essa variabilidade está principalmente contida nos genes de classe I clássicos e genes de classe II. Atualmente, há cerca de 20.897 alelos⁵ diferentes para genes de classe I depositados no banco de dados oficial para *HLA* (IPD-IMGT/*HLA* Database) (Tabela 1). Os níveis extremamente altos de polimorfismo e heterozigosidade dentro dessa região genômica fornece ao sistema imunológico uma vantagem seletiva contra diversos patógenos. A alta diversidade de alelos possibilita a apresentação de um amplo repertório de antígenos por um indivíduo, e consequentemente por uma população. Muitos estudos apoiam a hipótese de que essa diversidade alélica seja mantida por meio de seleção balanceadora mediada por micro-organismos. Até a realização deste estudo, pensava-se que essa variabilidade estava contida principalmente nos domínios $\alpha 1$ e $\alpha 2$, codificados pelos éxons 2 e 3, respectivamente, que formam a fenda de ligação ao peptídeo⁶. Se por um lado essa variabilidade é interessante do ponto de vista de adaptabilidade e resistência a patógenos, por outro lado, a necessidade ocasional de realização de enxertos alogênicos (transplantes) em humanos torna essa variabilidade problemática, pois a compatibilização desses polimorfismos é necessária para uma

boa aceitação do enxerto. Estudos clínicos demonstraram que a diferença de um único aminoácido em uma molécula HLA de classe I clássica pode causar rejeição aguda de transplante de medula óssea^{7,8}. No entanto, essa diversidade maior nos éxons 2 e 3 configura um viés de amostragem, visto que apenas esses dois éxons são exigidos para se incluir uma nova sequência no banco de dados de HLA.

Tabela 1. Número de alelos identificados para os genes *HLA* de classe I.

CLASSE I CLÁSSICOS		CLASSE I NÃO CLÁSSICOS	
Gene	Alelos	Gene	Alelos
<i>HLA-A</i>	6.766	<i>HLA-E</i>	271
<i>HLA-B</i>	7.967	<i>HLA-F</i>	45
<i>HLA-C</i>	6.621	<i>HLA-G</i>	82

Fonte: IPD-IMGT/HLA, Database 3.44.0

Os genes *HLA* podem ser definidos em diferentes resoluções. O primeiro campo da nomenclatura corresponde ao grupo ou família de alelos (ex. *HLA-B*35*). O segundo campo corresponde a proteína *HLA* específica (ex. *HLA-B*35:01*). Alelos cujos números diferem nos dois primeiros campos apresentam uma ou mais substituições de nucleotídeos que alteram a sequência de aminoácidos da proteína codificada. Alelos que diferem apenas por substituições sinônimas dentro da sequência codificadora são distinguidos pelo terceiro campo (ex. *HLA-B*35:01:01*). O quarto campo refere-se a polimorfismos dentro de íntrons e/ou nas regiões 5' e 3' não traduzidas (ex. *HLA-B*35:01:01:01*). A resolução máxima de 4 campos captura variantes alélicas localizadas dentro dos éxons, íntrons e regiões regulatórias

A predominância de certos alelos de genes *HLA* de classe I foi associada com doenças infecciosas, autoimunes e neoplásicas^{9,10}. Por exemplo, alelos do grupo *HLA-B27* foram associados com susceptibilidade à maioria das variantes da espondilite anquilosante¹¹; *HLA-A*01* e *HLA-A*24* com o Diabetes Mellitus¹²; *HLA-A*02* com doença de Alzheimer e Tireoidite de Hashimoto⁹; *HLA-B*35:02* e *HLA-B*35:03* com rápida progressão da infecção pelo HIV¹² e *HLA-A*24* com carcinoma de bexiga. Os mecanismos subjacentes à essas associações ainda não estão totalmente esclarecidos, porém acredita-se que determinadas variantes HLA não sejam capazes de acomodar antígenos tumorais ou virais, dificultando a

apresentação antigênica^{12,13}.

Estrutura e função das moléculas HLA de classe I

A molécula HLA de classe I (clássica e não clássica) é um heterodímero de superfície celular formado por uma cadeia pesada α codificada por um gene de classe I e uma cadeia leve β 2-microglobulina codificada no cromossomo 15, ligadas não-covalentemente (Figura 2)¹⁴. A cadeia α possui cinco domínios: dois domínios que constituem a fenda de ligação a peptídeos (α 1 e α 2), um domínio do tipo imunoglobulina (Ig), que interage com o co-receptor CD8 dos linfócitos (α 3), domínio transmembrana e domínio citoplasmático¹.

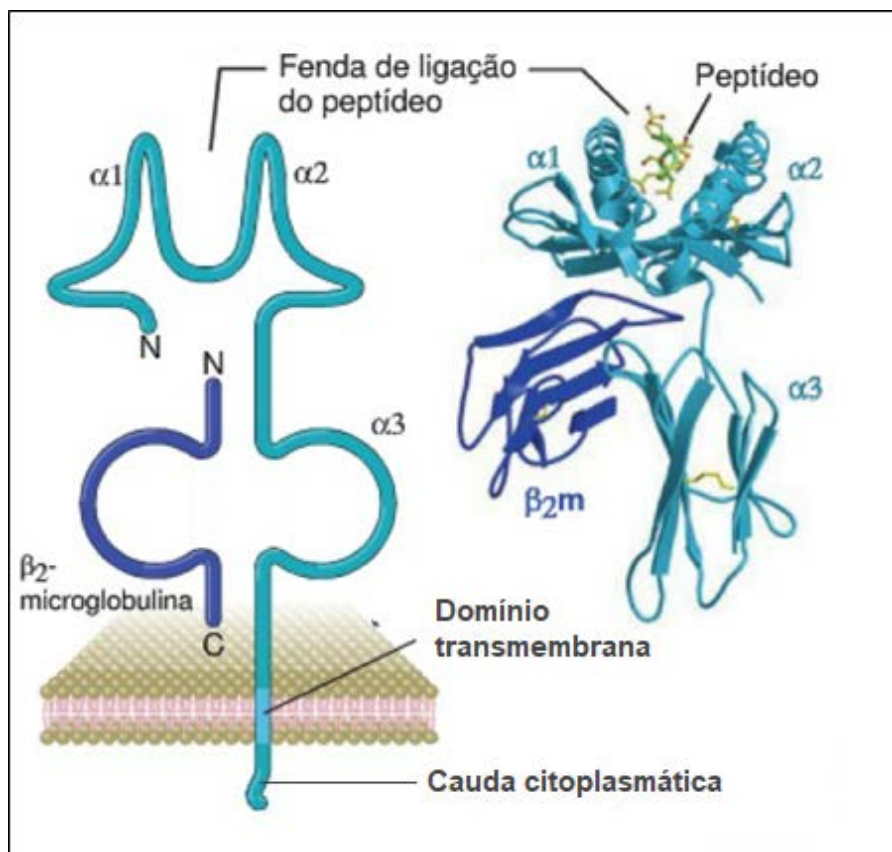


Figura 2: Estrutura das moléculas HLA Classe I.

A cadeia pesada da molécula de classe I possui dois domínios de ligação a peptídeos (α 1 e α 2), um domínio semelhante à imunoglobulina (α 3), o domínio transmembrana (TM) e a cauda citoplasmática. β 2-microglobulina é a cadeia leve da molécula de classe I. Fonte: Adaptado de Abbas¹⁵

A montagem da molécula HLA de classe I envolve uma via complexa que é iniciada pela formação do heterodímero, (associação das cadeias α e β 2m) guiada

por chaperonas¹⁶. Simultaneamente, os antígenos endógenos são degradados através de um complexo enzimático multiproteico denominado proteassoma, em um processo dependente de ubiquitina e ATP. Os peptídeos gerados são transportados para o lúmen do retículo endoplasmático (RE) através de uma molécula transportadora associada ao processamento de antígeno (TAP). O PLC (do inglês, *peptid-loading complex*), composto pelos heterodímeros TAP1 e TAP2, chaperonas, juntamente com a cadeia pesada de HLA I e a β 2-microglobulina, é essencial para o carregamento eficiente de peptídeos antigênicos na molécula de HLA de classe I¹⁷.

Algumas chaperonas, tais como a calnexina, calreticulina, p57 do retículo endoplasmático (ERp57), BiP (do inglês, *Binding immunoglobulin Protein*) e tapasina, estão envolvidas na montagem das moléculas de HLA de classe I. A ligação do peptídeo às moléculas do MHC classe I é um pré-requisito para a dissociação de complexos TAP-MHC. Essa dissociação mediada por peptídeos depende de sinais conformacionais do complexo TAP induzido pela ligação de ATP. Portanto, a atividade dinâmica do complexo TAP-MHC classe I é sincronizada com o ciclo de ligação e translocação de peptídeos da TAP^{18,19} (Figura 3). Na ausência de um ligante peptídico, as moléculas de HLA de classe I são geralmente instáveis e ficam retidas no RE. A ligação de um peptídeo libera as moléculas de HLA de classe I do PLC e permite o tráfego para a superfície celular via complexo de Golgi¹⁶.

Essas moléculas, em especial as de classe I clássicas, associam-se a um peptídeo de 8 a 12 resíduos de aminoácidos¹⁴. Peptídeos maiores precisam ser clivados para gerar antígenos de tamanho ideal para se ajustarem à fenda de ligação da molécula de HLA e as principais aminopeptidases responsáveis pelo corte de peptídeos N-terminais são ERAP1 e ERAP2 (do inglês, *Endoplasmic Reticulum Aminopeptidase*)²⁰.

Embora a maioria das moléculas de classe I formem complexos com peptídeos citosólicos, derivados de proteínas endógenas, virais ou aberrantes¹⁴, alguns micro-organismos exógenos podem ser reconhecidos pelos CTLs, visto que, muitos patógenos desenvolveram mecanismos para modular a via de apresentação direta, escapando das vesículas fagocíticas para o citosol. Os patógenos são

processados pela via comum de classe I, permitindo a geração de TCD8+ específico direcionado a esses patógenos^{15,21}.

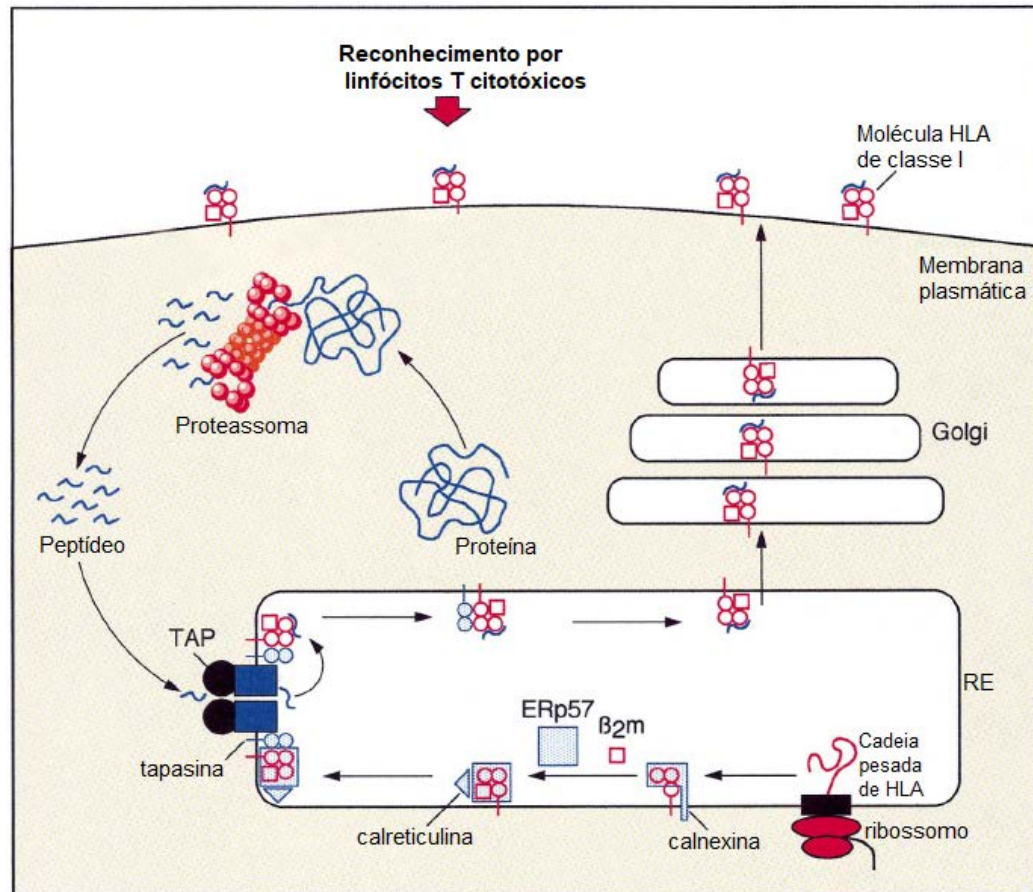


Figura 3. Processamento e apresentação de antígeno via HLA de classe I.

Proteínas endógenas são degradadas pelo proteassoma. Os peptídeos gerados são transportados para o retículo endoplasmático por meio do Transportador Associado ao Processamento de Antígenos (TAP). Muitas moléculas estão envolvidas na montagem e carregamento das moléculas HLA de classe I (maiores detalhes estão apresentados no texto). O complexo peptídeo/HLA estável deixa o ER e é transportado para a superfície celular via complexo de golgi. Fonte: Adaptado de Lankat-Buttgereit e Tampe¹⁹.

Algumas células dendríticas tem a capacidade de capturar e fagocitar células tumorais ou células infectadas por vírus e apresentar os seus antígenos para linfócitos T CD8 virgens, um processo denominado apresentação cruzada. Ele envolve a fusão de fagossomos contendo os antígenos ingeridos com o RE e que são translocadas para o citosol por vias pouco definidas. No citosol ocorre a degradação dessas proteínas via proteassoma. Os peptideos são transportados de volta pelo TAP para o RE onde são montados com moléculas do MHC de classe I^{15,22}.

Após serem transportadas à superfície celular, as moléculas de HLA de classe I interagem com o complexo receptor de células T (TCR) dos linfócitos CD8 ou com receptores específicos em células T e NK^{1,2}. Uma vez que os receptores dos linfócitos TCD8+ reconhecem tais peptídeos como partículas estranhas ao organismo, eles são ativados, diferenciando-se em linfócitos efetores (CTL) e podem então reconhecer qualquer célula alvo expressando os mesmos complexos HLA-peptídeo. Paralelamente, as interações HLA/peptídeo-TCR, HLA-CD8, moléculas de adesão e proteínas de sinalização intracelular entre as duas células, formam uma sinapse imunológica. A resposta citotóxica dos linfócitos estimula a destruição da célula alvo por meio da liberação de proteases no espaço sináptico ou pela indução da célula alvo à apoptose^{14,23}. No entanto, a presença de moléculas não clássicas na superfície celular pode inibir a ação de células T e NK.

Os TCRs (TCR, do inglês *T cell receptor*) reconhecem um peptídeo apresentado por apenas uma forma alélica específica de *HLA*. Essa especificidade ocorre através dos resíduos polimórficos presentes no topo das alfas hélices das moléculas de HLA²² (Figura 4). Ou seja, eles agem apenas quando o antígeno é apresentado por uma molécula de HLA do organismo do qual eles foram derivados. Esse fenômeno é denominado restrição do MHC^{22,24}. Isso acontece porque durante o desenvolvimento do timo, os timócitos, cujos receptores não são capazes de se ligar a um complexo peptídeo-MHC próprio, deixam de receber os estímulos necessários para sobreviver e morrem por uma via de apoptose padrão denominada morte por negligência, garantindo assim que os linfócitos sejam restritos ao próprio MHC^{23,25}. Já os timócitos, que expressam um TCR que se liga com uma baixa avidéz às moléculas de MHC próprio, são estimulados a sobreviver e continuar amadurecendo, processo conhecido como seleção positiva. Nessa seleção positiva também há restrição de MHC-I ou II, assegurando que linfócitos TCD8+ se acoplem a peptídeos apresentados pelo MHC-I e que células T CD4+ se liguem aos peptídeos expostos pela molécula de MHC-II. Se o TCR no timócito se ligar com alta avidéz a esses complexos de células dendríticas ou de células medulares do timo expondo peptídeos próprios, um sinal forte será transmitido, fazendo com que a célula sofra apoptose e evitando assim a maturação de linfócitos autorreativos, processo denominado de seleção negativa^{23,25,26}. Portanto, esse processo de seleção no timo garante que os linfócitos T periféricos cumpram dois

requisitos importantes: ativação por peptídeos estranhos ligados a moléculas de MHC e; tolerância a peptídeos próprios apresentados no mesmo contexto²⁷.

Vírus e células tumorais têm desenvolvido estratégias para inibir a expressão e/ou função das moléculas de HLA classe I para escapar do reconhecimento e destruição pelas células T CD8. Por outro lado, células com baixa expressão de moléculas de HLA classe I na superfície celular podem ser reconhecidas por células NK. Células infectadas ou neoplásicas frequentemente apresentam baixa ou total ausência da expressão de moléculas de HLA classe I, tornando-se susceptível à atividade das células NK induzida pelo processo de “*missing self*”²⁸, uma vez que a presença de HLA classe I na superfície celular é importante para a inibição da atividade de células NK, impedindo que sejam atacadas²⁹.

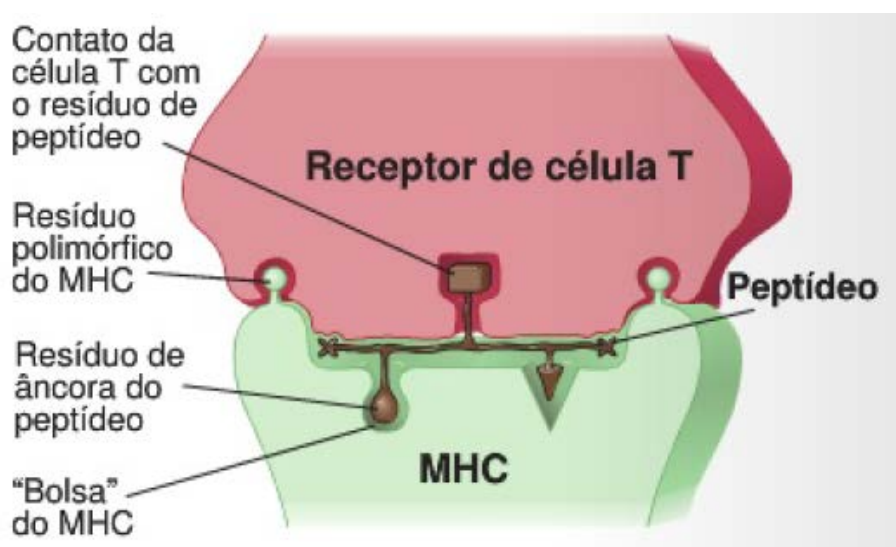


Figura 4. Representação esquemática do reconhecimento dos resíduos polimórficos e peptídico do complexo peptídeo/MHC pelo receptor da célula T. Fonte: Abbas¹⁵

As células NK expressam diversos receptores que modulam a sua função, incluindo os codificados pela família dos genes *KIR* (do inglês, *Killer cell Immunoglobulin like receptor*). Os KIRs são glicoproteínas transmembranares do tipo I com dois ou três domínios semelhantes a Ig e caudas citoplasmáticas de comprimentos variados, que podem ter função de ativar ou inibir a ação das NK. As moléculas de HLA classe I são reconhecidas pelos KIR e essa interação promove um equilíbrio de sinais que modulam a resposta das células NK contra células infectadas ou alteradas, através da sua ativação ou inibição²⁹. Cada célula NK possui pelo menos um receptor inibitório específico para uma molécula de classe I,

no intuito de evitar a auto reatividade²⁸. Um deles, *KIR3DL1*, se liga às moléculas HLA-B e algumas moléculas HLA-A contendo um epítipo Bw4, que é um motivo de sequência específico contido nos resíduos 77-83 da cadeia pesada^{30,31}; O receptor *KIR3DL2* reconhece HLA-B27, HLA-A11 e HLA-A3; já os receptores *KIR2DL1*, *KIR2DL2* e *KIR2DL3*, são amplamente específicos para variantes de HLA-C³¹⁻³³.

HLA-B

Os genes *HLA* de classe I são compostos por oito éxons o: primeiro éxon codifica o peptídeo líder, o segundo e o terceiro éxons codificam os domínios $\alpha 1$ e $\alpha 2$, respectivamente; o quarto éxon codifica o domínio $\alpha 3$, o quinto codifica o domínio transmembrana e os éxons 6 a 8 codificam os domínios citoplasmáticos da molécula. O éxon 8 contribui apenas com dois nucleotídeos para o último códon que codifica o resíduo C-terminal das moléculas *HLA-A* e *HLA-C*. Porém, para *HLA-B*, nenhum nucleotídeo do éxon 8 é traduzido por causa de um códon de parada no éxon 7. Portanto, para *HLA-B*, o éxon 8 configura a região 3' não traduzida do mRNA ³⁴.

O gene *HLA-B* é o mais variável entre os genes *HLA* (provavelmente o mais variável do genoma humano) e, desta forma, é observada uma alta heterozigose nesse gene, acarretando na produção de duas moléculas HLA-B distintas na superfície celular na maioria dos indivíduos. Este fato, associado com a elevada variabilidade e heterozigose de *HLA-A*, permite a apresentação de uma ampla gama de antígenos. São conhecidos milhares de alelos do gene *HLA-B* (Tabela 1), no entanto alguns alelos apresentam sequências muito semelhantes ou codificam moléculas idênticas. Neste contexto, por exemplo, mais de 100 alelos são categorizados como do subtipo *HLA-B27* e muitos deles estão associados com espondilite anquilosante¹¹.

Uma possível explicação para a maior variabilidade é que as substituições de aminoácidos afetam o repertório de ligação a peptídeos de forma mais significativa nas moléculas de HLA-B em comparação com outras moléculas de HLA. Os resultados de um estudo *in silico* sugeriram que mutações únicas afetam o repertório de ligação ao peptídeo das moléculas de HLA-B mais do que em moléculas de HLA-A³⁵. Ou seja, pode ser mais fácil gerar um novo motivo de ligação a peptídeos através de uma mutação pontual em uma molécula HLA-B do que em

uma molécula HLA-A. Considerando que uma nova molécula HLA com um repertório de peptídeos alterado seja mantida na população, novas moléculas de HLA-B poderiam evoluir facilmente através de mutações pontuais e, portanto, alcançar uma maior diversidade a nível populacional³⁵.

Além disso, o polimorfismo do gene *HLA-B* difere entre as populações. Dessa forma, nota-se que algumas variantes alélicas são altamente frequentes ou ausentes em determinadas populações enquanto outras apresentam variações em suas frequências³⁶, como demonstrado neste estudo (ver capítulo II). Por exemplo, o alelo *HLA-B*53:01* é mais frequente em populações africanas (> 10% em algumas populações), mas apresenta uma frequência muito menor em europeus e está ausente em algumas populações asiáticas³⁷. Em asiáticos, alelos *B*38:02*, *B*46:01*, *B*15:02*, *B*54:01*, *B*13:01* e *B*15:21* foram encontrados em frequências elevadas e parecem ser restritos a essas populações³⁸. Os alelos *B*07:02* e *B*08:01* são frequentes em populações europeias^{39,40}, mas também são comuns em outras populações^{37,41}. Alguns alelos *HLA-B*, como *B*39:09*, são restritos a populações nativo americanas ou miscigenadas.

As populações indígenas são caracterizadas por pequenos grupos que experimentam efeitos intensos de deriva genética e baixo fluxo gênico. Quando comparado a outras populações, as populações indígenas apresentam um polimorfismo alélico baixo dos genes *HLA*, o que reflete o isolamento geográfico dessas populações. Há evidências de que o gene *HLA-B* evoluiu mais rapidamente do que *HLA-A* e *HLA-C* nas tribos sul-americanas, visto que a maioria dos novos alelos encontrados nessas populações são do locus *HLA-B*. Em muitas tribos, os novos alelos são os mais frequentes, sugerindo que estes foram selecionados positivamente, provavelmente porque forneceram vantagens seletivas⁴².

Na população brasileira, que é caracterizada pela alta taxa de miscigenação entre europeus, africanos e indígenas, há uma diversidade maior de alelos *HLA*. De acordo com a predominância dessas etnias, a composição de alelos HLA da população pode variar de uma região para outra. A população do Piauí, por exemplo, apresenta alelos europeus e africanos em frequências intermediárias, e alelos de origem indígena são encontrados em baixas frequências⁴³. Um estudo mostrou uma predominância de alelos europeus em uma amostra de indivíduos autodeclarados brancos do Norte e Noroeste do Paraná⁴⁴. No Rio Grande do sul, o

alelo *HLA-B*35* foi encontrado em maior frequência, o alelo *HLA-B*53* foi o mais frequente em miscigenados e pretos e, alelos de origem europeia, como *HLA-B*35*, *HLA-B*44* e *HLA-B*51*, apresentaram frequências reduzidas em miscigenados e pretos⁴⁵. Além disso, diferenças na frequência dos alelos *HLA-B* podem contribuir com diferenças geográficas na suscetibilidade a muitas doenças. Por exemplo, o alelo *HLA-B*54:01* é pouco frequente na Europa e na África (< 0,5%), mas tem uma frequência de ~2,2–8,8% na Ásia³⁷, onde foi associados com a doença de Kawasaki em crianças coreanas⁴⁶.

O gene *HLA-B* têm sido amplamente estudados por seu papel na biologia do transplante, doenças humanas e diversidade populacional^{47–49}. Apesar disso, apenas alguns éxons são geralmente avaliados, principalmente aqueles que codificam a fenda de ligação ao peptídeo. Porém, o polimorfismo do gene *HLA-B* também ocorre em outras regiões e essa variação fora dos éxons é, provavelmente, sub-representada no banco de dados IPD-IMGT/HLA⁵. É possível que variantes dos domínios $\alpha 3$ (éxon 4), transmembrana (éxon 5) e cauda citoplasmática (éxons 6 e 7), que estão diretamente envolvidos na estabilidade e sinalização, também possam influenciar a atuação da molécula *HLA-B* durante a evolução de doenças infecciosas⁵⁰. Além disso, foi descrito um papel funcional para o locus *HLA-B* mediado por um microRNA (miR-6891-5p) codificado no íntron 4⁵¹. Um estudo funcional sugeriu que o miR-6891-5p pode afetar muitas vias metabólicas, incluindo as vias de resposta imune. Os principais transcritos afetados são os que codificam a cadeia pesada da IgA (IGHA1 e IGHA2), sugerindo que o miR-6891-5p regula a expressão dos genes IGHA1 e IGHA2 no nível pós-transcricional e que o aumento dos níveis de miR-6891-5p pode contribuir para a etiologia da deficiência seletiva de IgA⁵¹. Portanto, polimorfismos dentro dessas sequências podem afetar vários processos biológicos.

A principal região promotora dos genes *HLA* está no promotor proximal, até 300pb à montante da sequência de iniciação da transcrição⁵². Polimorfismos nas sequências da região promotora dos genes *HLA* afetam a montagem dos fatores gerais de transcrição e a transcrição geral do gene. Além disso, os genes *HLA* parecem ser alvos de um grande número de microRNAs⁵³. Neste contexto, tanto o perfil de microRNAs expressos em uma célula ou tecido (miRnoma), quanto a variabilidade da principal região alvo de microRNAs (a região 3' não traduzida ou

3'NT) influenciam a quantidade de mRNA disponível para tradução. Em todos os casos, combinações de variantes nos elementos regulatórios podem influenciar a quantidade de mRNA produzida.

Além disso, a natureza polimórfica dos genes *HLA* e a alta similaridade entre si, devido às suas origens parálogas⁵⁴, podem levar a erros de genotipagem quando o sequenciamento dos genes *HLA* é realizado por procedimentos NGS. Nesses casos, quando os algoritmos são usados com parâmetros padrão, muitas variantes verdadeiras não são identificadas porque estão presentes em haplótipos que diferem demasiadamente do genoma usado como referência e, portanto, as sequências geradas a partir dessas regiões não são alinhadas e são perdidas, levando a vieses de mapeamento⁵⁵.

Por isso, o tipo e a precisão do método de genotipagem usado para caracterizar as variantes genéticas desses genes tem um grande impacto na obtenção de um dado confiável.

Neste estudo apresentamos um pipeline de bioinformática adequado para avaliar a variabilidade do gene *HLA-B* em sequenciamentos do tipo NGS, considerando todos os éxons, íntrons e regiões regulatórias (promotora e 3'UTR). Aplicamos este método em uma amostra de 3648 indivíduos de 29 populações diferentes, relatando a diversidade genética completa do gene *HLA-B* em populações mundiais.

JUSTIFICATIVA

O gene *HLA-B* é o mais polimórfico dentre os genes que codificam moléculas HLA de classe I clássicas. Sua principal função está relacionada com a apresentação antigênica, sendo, portanto, importante em diversos contextos fisiológico e patológicos, como infecções, reação enxerto-hospedeiro, neoplasias e doenças autoimunes. A diversidade desse gene é subestimada, principalmente porque não há estudos populacionais que avaliaram todas as regiões desse gene, incluindo íntrons e promotores. Polimorfismos em regiões não codificantes do gene *HLA-B*, como íntrons e regiões regulatórias, não são expressos na proteína, mas podem ter efeitos significativos na função biológica da molécula, como mudanças nos níveis de expressão e padrões de edição do transcrito primário. Desse modo, a caracterização da estrutura e da variabilidade desse gene servirá de base para compreender a dinâmica entre a molécula *HLA-B* e diversas situações patológicas ou casos de transplantes. Além disso, informações sobre essas moléculas, tais como polimorfismo, frequência de alelos e haplótipos, também servirão de base para estudos populacionais.

OBJETIVOS

Objetivo Geral

Avaliar a variabilidade do gene *HLA-B* e identificar os polimorfismos e haplótipos do gene encontrados em amostras de 29 populações diferentes.

Objetivos Específicos

- Consolidar uma metodologia para avaliação da variabilidade do gene *HLA-B* por sequenciamento massivo paralelo (ou sequenciamento de nova geração);
- Detectar os sítios de variação encontrados para o gene *HLA-B* em amostras de 29 populações diferentes;
- Identificar as sequências completas do gene *HLA-B* em cada indivíduo, caracterizando regiões regulatórias, todos os éxons e íntrons.
- Estudar a diversidade do gene *HLA-B* em diferentes populações;

REFERÊNCIAS

1. Klein, J.; Sato A. The HLA System — First of Two Parts. *N Engl J Med.* 2000;343:702–9.
2. Donadi EA, Castelli EC, Arnaiz-Villena A, Roger M, Rey D, Moreau P. Implications of the polymorphism of HLA-G on its function, regulation, evolution and disease association. *Cell Mol Life Sci.* 2011;68(3):369–95.
3. Campbell RD, Milner CM. MHC genes in autoimmunity. *Curr Opin Immunol.* 1993;5(6):887–93.
4. Mehra NK, Kaur G. MHC-based vaccination approaches: Progress and perspectives. *Expert Rev Mol Med.* 2003;5(7):1–17.
5. Robinson J, Waller MJ, Parham P, Bodmer JG, Marsh SGE. IMGT/HLA Database - A sequence database for the human major histocompatibility complex. *Nucleic Acids Res.* 2001;29(1):210–3.
6. Tiercy JM. Molecular basis of HLA polymorphism: Implications in clinical transplantation. In: *Transplant Immunology.* Elsevier; 2002. p. 173–80.
7. Crispim JC, Mendes-Júnior CT, Wastowski IJ, Palomino GM, Saber LT, Rassi DM, et al. HLA Polymorphisms as Incidence Factor in the Progression to End-Stage Renal Disease in Brazilian Patients Awaiting Kidney Transplant. *Transplant Proc.* 2008;40(5):1333–6.
8. HUGHES, A. L; YEAGER M. Natural selection and the evolutionary history of major histocompatibility complex loci. *Front biosci.* 1998;3:509–16.
9. Shiina T, Inoko H, Kulski JK. An update of the HLA genomic region, locus information and disease associations: 2004. *Tissue Antigens.* 2004;64(6):631–49.
10. Thorsby E, Lie BA. HLA associated genetic predisposition to autoimmune diseases: Genes involved and possible mechanisms. *Transpl Immunol.* 2005;14(3-4 SPEC. ISS.):175–82.
11. Sampaio-Barros PD, Conde RA, Donadi EA, Bonfiglioli R, Costallat LTL, Samara AM, et al. Frequency of HLA-B27 and its alleles in patients with Reiter syndrome: Comparison with the frequency in other spondyloarthropathies and a healthy control population. *Rheumatol Int.* 2008;28(5):483–6.
12. Goulder PJR, Watkins DI. Impact of MHC class I diversity on immune control of immunodeficiency virus replication. Vol. 8, *Nature Reviews Immunology.* 2008. p. 619–30.
13. Castelli EC, Mendes CT, Viana De Camargo JL, Donadi EA. HLA-G polymorphism and transitional cell carcinoma of the bladder in a Brazilian population. *Tissue Antigens.* 2008 Aug;72(2):149–57.

14. FISCHER, G. F; MAYR WR. molecular genetics of the HLA complex. *Wien klin wochenschr.* 2001;113(20–21):814–24.
15. ABBAS, A. K; LICHTMAN, A. H; PILLAI SHI V. *Imunologia Celular e Molecular.* In: *Mólecúlas do Complexo Principal de Histocompatibilidade e Apresentação do Antígeno aos linfócitos T.* 7º. Rio de Janeiro: Elsevie; 2011. p. 109–38.
16. Blees A, Janulienė D, Hofmann T, Koller N, Schmidt C, Trowitzsch S, et al. Structure of the human MHC-I peptide-loading complex. *Nat Publ Gr.* 2017;551(7681):525–8.
17. Lehnert E, Tampé R. Structure and Dynamics of Antigenic Peptides in Complex with TAP. 2017;8(January):4–11.
18. Eggensperger S, Tampé R. The transporter associated with antigen processing: A key player in adaptive immunity. *Biol Chem.* 2015;396(9–10):1059–72.
19. Lankat-butgereit B, Tampe R, Tampe R. The Transporter Associated With Antigen Processing : Function and Implications in Human Diseases. 2019;187–204.
20. Saulle I, Vicentini C, Clerici M, Biasin M. An Overview on ERAP Roles in Infectious Diseases. *Cells.* 2020;1:1–23.
21. Niu TK, Princiotta MF, Sei JJ, Norbury CC. Analysis of MHC Class I Processing Pathways That Generate a Response to Vaccinia Virus Late Proteins. *ImmunoHorizons.* 2019;3(12):559–72.
22. Rock KL, Reits E, Neefjes J. Present Yourself! By MHC Class I and MHC Class II Molecules. *Trends Immunol.* 2016;37(11):724–37.
23. ALBERTS B et al. *Molecular Biology of the Cell.* In: *Journal of Chemical Information and Modeling.* 6th ed. New York: 2015; 1970. p. 1689–99.
24. Kourilsky P, Claverie JM. MHC restriction, alloreactivity, and thymic education: A common link? *Cell.* 1989;56(3):327–9.
25. Daley SR, Teh C, Hu DY, Strasser A, Gray DHD. Cell death and thymic tolerance. *Immunol Rev.* 2017;277(1):9–20.
26. Liu YJ. A unified theory of central tolerance in the thymus. *Trends Immunol.* 2006;27(5):215–21.
27. Wieggers GJ, Kaufmann M, Tischner D, Villunger A. Shaping the T-cell repertoire: A matter of life and death. *Immunol Cell Biol.* 2011;89(1):33–9.
28. Aptsiauri N, Cabrera T, Garcia-Lora A, Lopez-Nevot MA, Ruiz-Cabello F, Garrido F. MHC Class I Antigens and Immune Surveillance in Transformed Cells. *Int Rev Cytol.* 2007;256(07):139–89.

29. Augusto DG, Petzl-Erler ML. KIR and HLA under pressure: evidences of coevolution across worldwide populations. *Hum Genet.* 2015;134(9):929–40.
30. Béziat V, Hilton HG, Norman PJ, Traherne JA. Deciphering the killer-cell immunoglobulin-like receptor system at super-resolution for natural killer and T-cell biology. *Immunology.* 2017;150(3):248–64.
31. Pugh J, Nemat-Gorgani N, Djaoud Z, Guethlein LA, Norman PJ, Parham P. In vitro education of human natural killer cells by KIR3DL1. *Life Sci Alliance.* 2019;2(6):1–14.
32. Chou YC, Chen CH, Chen MJ, Chang CW, Chen PH, Yu MH, et al. Killer cell immunoglobulin-like receptors (KIR) and human leukocyte antigen-C (HLA-C) allorecognition patterns in women with endometriosis. *Sci Rep.* 2020;10(1):1–9.
33. Parham P, Moffett A. Variable NK cell receptors and their MHC class I ligands in immunity, reproduction and human evolution. *Nat Rev Immunol.* 2013;13(2):133–44.
34. Dubois V, Tiercy JM, Labonne MP, Dormoy A, Gebuhrer L. A new HLA-B44 allele (B*44020102S) with a splicing mutation leading to a complete deletion of exon 5. *Tissue Antigens.* 2004;63(2):173–80.
35. van Deutekom HWM, Keşmir C. Zooming into the binding groove of HLA molecules: which positions and which substitutions change peptide binding most? *Immunogenetics.* 2015;67(8):425–36.
36. Probst CM, Bompeixe EP, Pereira NF, Dalalio M. MO, Visentainer JE, Tsuneto LT., et al. HLA polymorphism and evaluation of European, African, and Amerindian contribution to the white and mulatto populations from Paran , Brazil - PubMed. *Hum Biol.* 2000;72(4):597–617.
37. Gonzalez-Galarza FF, Christmas S, Middleton D, Jones AR. Allele frequency net: A database and online repository for immune gene frequencies in worldwide populations. *Nucleic Acids Res.* 2011 Jan 1;39(SUPPL. 1):D913–9.
38. Cao K, Hollenbach J, Shi X, Shi W, Chopek M, Fern ndez-Vi na MA. Analysis of the frequencies of HLA-A, B, and C alleles and haplotypes in the five major ethnic groups of the United States reveals high levels of diversity in these loci and contrasting distribution patterns in these populations. In: *Human Immunology.* Elsevier Inc.; 2001. p. 1009–30.
39. Sanchez-Mazas A, Buhler S, Nunes JM. A New HLA Map of Europe: Regional Genetic Variation and Its Implication for Peopling History, Disease-Association Studies and Tissue Transplantation. *Hum Hered.* 2013;76:162–77.
40. Creary LE, Gangavarapu S, Mallemapati KC, Montero-Mart n G, Caillier SJ, Santaniello A, et al. Next-generation sequencing reveals new information about HLA

allele and haplotype diversity in a large European American population. *Hum Immunol.* 2019;80(10):807–22.

41. Shen J, Guo T, Wang T, Zhen Y, Ma X, Wang Y, et al. HLA-B*07, HLA-DRB1*07, HLA-DRB1*12, and HLA-C*03:02 strongly associate with BMI: Data from 1.3 million healthy Chinese adults. *Diabetes.* 2018 May 1;67(5):861–71.

42. Fernandez Vina MA, Hollenbach JA, Lyke KE, Sztein MB, Maiers M, Klitz W, et al. Tracking human migrations by the analysis of the distribution of HLA alleles, lineages and haplotypes in closed and open populations. Vol. 367, *Philosophical Transactions of the Royal Society B: Biological Sciences.* Royal Society; 2012. p. 820–9.

43. Carvalho MG, Tsuneto LT, Moita Neto JM, Sousa LCDM, Sales Filho HLA, Macêdo MB, et al. HLA-A, HLA-B and HLA-DRB1 haplotype frequencies in Piauí's volunteer bone marrow donors enrolled at the Brazilian registry. *Hum Immunol.* 2013 Dec 1;74(12):1598–602.

44. Reis PG, Ambrosio-Albuquerque EP, Fabreti-Oliveira RA, Moliterno RA, de Souza VH, Sell AM, et al. HLA-A, -B, -DRB1, -DQA1, and -DQB1 profile in a population from southern Brazil. *HLA.* 2018 Nov 1;92(5):298–303.

45. Bortolotto AS, Petry MG, da Silveira JG, da Fonte Raya AR, Fernandes SR, Neumann J, et al. HLA-A, -B, and -DRB1 allelic and haplotypic diversity in a sample of bone marrow volunteer donors from Rio Grande do Sul State, Brazil. *Hum Immunol.* 2012 Feb 1;73(2):180–5.

46. Kwon YC, Sim BK, Yu JJ, Yun SW, Yoon KL, Lee KY, et al. HLA-B*54:01 Is Associated With Susceptibility to Kawasaki Disease. Vol. 12, *Circulation. Genomic and precision medicine.* NLM (Medline); 2019. p. e002365.

47. Petersdorf EW, Malkki M, O'hUigin C, Carrington M, Gooley T, Haagenson MD, et al. High HLA-DP Expression and Graft-versus-Host Disease. *N Engl J Med.* 2015 Aug 13;373(7):599–609.

48. Klein, J.; Sato A. THE HLA SYSTEM - Second of Two Parts. *N Engl J Med.* 2000;343(11):782–6.

49. Nakaoka H, Inoue I. Distribution of HLA haplotypes across Japanese Archipelago: Similarity, difference and admixture. *J Hum Genet.* 2015;60(11):683–90.

50. Rizvi SM, Salam N, Geng J, Qi Y, Bream JH, Duggal P, et al. Distinct Assembly Profiles of HLA-B Molecules. *J Immunol.* 2014;192(11):4967–76.

51. Chitnis N, Clark PM, Kamoun M, Stolle C, Johnson FB, Monos DS. An expanded role for HLA genes: HLA-B encodes a microRNA that regulates IgA and other immune response transcripts. *Front Immunol.* 2017;8(MAY):1–11.

52. Handunnetthi L, Ramagopalan S V., Ebers GC, Knight JC. Regulation of major histocompatibility complex class II gene expression, genetic variation and disease. Vol. 11, *Genes and Immunity*. Nature Publishing Group; 2010. p. 99–112.
53. Porto IOP, Mendes-Junior CT, Felicio LP, Georg RC, Moreau P, Donadi EA, et al. MicroRNAs targeting the immunomodulatory HLA-G gene: a new survey searching for microRNAs with potential to regulate HLA-G. *Mol Immunol*. 2015 Jun;65(2):230–41.
54. Brandt DYC, Aguiar VRC, 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 Genes, Genomes, Genet*. 2015;5(5):931–41.
55. 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. *Hum Immunol*. 2018;79(9):678–84.

Capítulo II - Artigo

Immunogenetics of *HLA-B*: SNP and haplotype diversity in populations from different continents and ancestry backgrounds

Authors

Nayane dos Santos Brito Silva^{1,2}, Andreia S. Souza¹, Heloisa A. Andrade¹, Raphaela N. Pereira^{1,2}, Camila F. B. Castro^{1,3}, Marília O. Scliar⁴, Jaqueline Wang⁴, Maria Rita Passos-Bueno^{4,6}, Michel S. Naslavsky^{4,6,7}, Mayana Zatz^{4,6}, Diogo Meyer⁶, Eduardo A. Donadi⁸, Celso T. Mendes-Junior⁹, Erick C. Castelli^{1,2}

Affiliations

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), Pathology Program, School of Medicine, Botucatu, State of São Paulo, Brazil. ³ Centro Universitário Sudoeste Paulista – UniFSP. ⁴ Human Genome and Stem Cell Research Center, University of São Paulo, São Paulo, SP, Brazil. ⁶ Department of Genetics and Evolutionary Biology, Biosciences Institute, University of São Paulo (USP), São Paulo, State of São Paulo, Brazil. ⁷ Hospital Israelita Albert Einstein, São Paulo, State of São Paulo, Brazil. ⁸ Department of Medicine, School of Medicine of Ribeirão Preto, University of São Paulo (USP), Brazil. ⁹ Departamento de Química, Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, Universidade de São Paulo (USP), Ribeirão Preto, São Paulo, Brazil.

Contact

Erick da Cruz Castelli
Departamento de Patologia, Faculdade de Medicina de Botucatu, Unesp – Botucatu, SP
CEP: 18618970, Brazil
Phone: +55 14 3880-1696
E-mail address: erick.castelli@unesp.br

Abstract

The Human Leukocyte Antigen B (*HLA-B*) gene is possibly the most variable gene in the human genome, with almost eight thousand different sequences reported in the Immuno Polymorphism Database (IPD-IMGT/HLA). This gene encodes a key molecule for antigen presentation to T CD8⁺ lymphocytes and NK cell modulation. However, regulatory regions and introns are usually ignored in population surveys. Thus, *HLA-B* genetic variability is probably underestimated. Because of sequence similarity with other *HLA* genes and their polymorphic nature, the evaluation of *HLA-B* variability is quite challenging. We applied a molecular method and a bioinformatics pipeline suitable to evaluate the *HLA-B* variability in exons, introns, and regulatory regions, in 3,648 samples from 29 different worldwide populations. We detected 610 variable sites throughout *HLA-B*. Although most of these variable sites are found in the regulatory regions, they are usually low-frequency variants. Most of the SNPs are shared worldwide. However, the distribution of haplotypes is geographically structured. Overall, 535 genomic sequences (complete sequences) encompassing 191 coding sequences were found, which encode 171 different proteins. *HLA-B* gene diversity is higher in admixed populations, particularly in Brazilians. Our method has proven to be effective and necessary to obtain reliable *HLA-B* genotypes and haplotypes from whole-genome sequencing data. This *HLA-B* variation resource may improve HLA imputation accuracy and disease-association studies.

Keywords: *HLA-B*, variability, NGS, populations worldwide.

Introduction

The classical Human Leukocyte Antigens (*HLA*) class I genes, *HLA-A*, *HLA-B*, and *HLA-C*, encode critical molecules for the immunosurveillance against viruses or other intracellular pathogens^{1,2}. Self, foreign, or abnormal peptides of 8 to 12 amino acid residues may associate with HLA molecules after intracellular processing³. These complexes are loaded onto the cell surface where they interact with the T cell receptor complex (TCR) from Cytotoxic T Lymphocytes (CTL) or with specific receptors on T and Natural Killer (NK) cells^{1,2}. CTLs can recognize such peptides as foreign particles to the organism and induce the destruction of the target cell (e.g., parasitized or neoplastic cells)³.

The HLA complex and its orthologous loci comprise the most polymorphic genes in the genome of most vertebrates. This variability is mainly contained in classic class I and II genes². Many studies support the hypothesis that this allelic diversity is maintained through balancing selection mediated by microorganisms⁴. This variability is more accentuated in exons 2 and 3 in classical class I genes, such as *HLA-B*, and they encode the peptide-binding groove^{2,5}.

HLA-B is possibly the most variable in the human genome, with almost eight thousand different sequences reported in the Immuno Polymorphism Database (IPD-IMGT/HLA version 3.44.0)⁶. Thus, we usually observe a high heterozygosity in this locus and the production of two distinct HLA-B molecules on the cell surface of each individual. This diversity, associated with the high variability and heterozygosity levels observed on *HLA-A*, allows the presentation of a wide range of antigens⁷. Findings of an *in silico* analysis indicated that the peptide-binding repertoire of HLA-B is more sensitive to single mutation than of HLA-A, i.e., there is a larger probability that a single-nucleotide mutation in *HLA-B* will result in a different functional peptide-binding motif. Considering that a population may retain new HLA molecules associated with new peptide repertoires, *HLA-B* may easily evolve through point mutations and achieve greater diversity at the population level⁸.

Some class I alleles have been associated with susceptibility to infectious, autoimmune, and neoplastic diseases^{9,10}. For instance, *HLA-B*27* alleles are predisposition factors for ankylosing spondylitis⁷; *HLA-B*51* have been associated with Behçet syndrome¹¹, *HLA-A*01*, and *HLA-A*24* with type 1 Diabetes Mellitus¹²; *HLA-A*02* with Hashimoto's thyroiditis¹⁰; among many other associations. The

mechanisms underlying these associations are not yet fully understood. Nevertheless, a poor prognosis in neoplasms and infections may be associated with the lower capacity some *HLA* alleles have for presenting tumor or viral antigens^{13,14}.

In individuals infected with HIV, for instance, it has been shown that some *HLA-B* alleles are related to rapid progression^{13,15}, while others display the opposite effect^{16,17}. This association reflects the fact that variations in these molecules can impact the profile of peptides or antigens presented, their folding patterns, their expression on the cell surface, thus influencing the antigenic presentation pathways and the assemble of adaptive immune responses. It is possible that variants of the $\alpha 3$ (exon 4), transmembrane (exon 5), and cytoplasmic tail (exons 6 and 7) domains, which are directly involved in protein folding, protein stability, and signaling, may also affect the function of the *HLA-B* molecule during the outcome of infectious diseases¹⁸.

HLA-B has been extensively studied in different populations, mainly because it must be evaluated for donor/recipient compatibility^{19–21}. However, only exons are usually evaluated (particularly exons 2 and 3). Regions that control gene expressions, such as the promoter and the 3'UTR, intronic regions that guide mRNA folding, and exonic regions that encode other protein domains have been often left aside. Moreover, sequence similarity among *HLA* genes, especially between *HLA-B* and *HLA-C*, and the high levels of polymorphism when compared to the human reference genome jeopardize the evaluation of *HLA* genes when using massively parallel sequencing or next-generation sequencing (NGS) procedures^{22,23}. Here, we present a bioinformatics pipeline suitable to evaluate the complete *HLA-B* variability from targeted, whole-genome (WGS), and exome sequencing, in the SNP and haplotype levels. We applied this strategy to evaluate a large dataset of 3,658 individuals from 29 different populations, providing a comprehensive report concerning *HLA-B* genetic diversity in worldwide populations.

Methods

Samples used to evaluate *HLA-B*

Brazilian samples and WGS sequencing

We analyzed *HLA-B* of 1,171 samples from elderly (over 65 years old) individuals surveyed in the São Paulo city, the most populated city in South America, which are part of the SABE (Saúde, Bem Estar e Envelhecimento – Health, Well-Being, and Aging) project, conducted by the Center for Studies of the Human Genome and Stem Cells (CEGH-CEL) at the University of São Paulo (IB / USP)²⁶. The sequencing protocol is described elsewhere²⁷. We extracted all reads aligned to chromosome 6 and also unmapped reads using samtools²⁸, merging these reads in a BAM file. After applying the methods described below, 1,157 passed the quality control and were considered for further analysis.

The 1000 Genomes dataset

We analyzed *HLA-B* genetic diversity in 2,504 individuals from 26 different populations worldwide, from the 1000 Genomes dataset. These data were obtained from a high-coverage whole-genome sequencing (WGS) recently published by 1000Genomes Consortium²⁴. We downloaded the aligned BAM file using the ASPERA High-Speed protocol. After applying the methods described below, 2,476 passed the quality control and were considered for further analysis.

Amerindians samples

We analyzed *HLA-B* of 15 Amerindians from Brazil, from the Suruí ($n=6$) and Karitiana ($n=9$) tribes. These groups live in villages near the Mato Grosso-Rondônia states border and north of Rondônia, respectively. We downloaded the raw FASTQ data (high-coverage WGS) from the Human Genome Diversity Project (HGDP)²⁵. After, we aligned all reads to the human reference genome hg38 using BWA MEM with default parameters.

Raw data processing

We used *hla-mapper dna*, version 4.0, to extract all HLA-related reads and unmapped reads from the BAM files²², and align them to the reference genome hg38 avoiding mapping bias. Although we present here only the *HLA-B* data, all other *HLA* loci are included in the *hla-mapper* database. The *hla-mapper* program is available at www.castelli-lab.net/apps/hla-mapper. This software applies a set of filters to assign each pair of reads to the proper *HLA* locus using the hg38 reference

and evaluates whether these reads should be assigned to more than one locus (multi-mappings).

Genotype and haplotype calling

To call genotypes and haplotypes, we used a similar strategy as described in a previous manuscript addressing *HLA-C* genetic diversity in Brazil and Benin²⁹. Briefly, we called genotypes using the Genome Analysis Toolkit HaplotypeCaller (GATK, version 4.1.7) in the GVCF mode, joining all samples in a single G.VCF file³⁰. We annotated all variants considering dbSNP version 150. For variant refinement and selection, we used the vcfx package (available at www.castelli-lab.net/apps/vcfx), algorithms *checkad*, *checkpl*, and *evidence*, in this order. We manually checked each variant that was not annotated as "PASS" by inspecting the BAM files. To infer haplotypes within the *HLA-B* locus, we first estimate micro-haplotypes using GATK *ReadBackedPhasing* (RBP) with *phaseQualityThresh* set to 500. This step resulted in phase sets (micro-haplotypes) based on the co-occurrence of variants in the same reads. RBP does not phase indels or multiallelic loci. The final haplotypes were estimated by combining these phase sets and probabilistic models, using *Shapeit4*³¹ for bi-allelic variants, and phasing multi-allelic variants with *Beagle 4.1*³². This step is automated using a local program named *phasex* (available on request), which performs many independent runs comparing the results to estimate the most likely haplotype for each sample, as presented elsewhere²⁹.

The missing alleles (0.71% in the present study), both bi-allelic and multi-allelic, were imputed by *Shapeit4* and *Beagle*, respectively. We removed all singletons before the haplotyping stage, but we have re-introduced them, whenever possible, according to the RBP results using a local Perl script. The output is a phased VCF file with all variable sites (bi-allelic, multi-allelic, and indels).

Allele calling in the genomic, exonic, and allotype levels

We used *vcfx fasta* and *vcfx transcript* to create complete genomic and CDS sequences for each sample – one for each chromosome. This approach uses the phased VCF from the previous step, replacing the alleles from each variation site in the chromosome 6 reference sequence. As *HLA-B* is encoded on the GRCh38

chromosome 6 reverse strand, sequences have been reversed and complemented with *EMBOSS revseq*³³. We developed a Perl script to identify whether our sequences are identical to any sequence in the IPD-IMGT/HLA database (version 3.43.0), in the genomic (exons and introns), or exonic (only exons) levels. This script also used *EMBOSS transeq*³³ to translate the CDS sequences into protein sequences, which we used to define the allotypes.

Statistical analysis

Allele, genotype, and haplotype frequencies were calculated by direct counting. A multidimensional scaling (MDS) based on pairwise F_{ST} values was performed using the *PopGenReport* package in R and plotted using ggplot2. The GEN files for *PopGenReport* were generated using the *vcfx genepop* to convert the phased VCF into a genepop input file. We used the *Pegas* package in R to evaluate nucleotide diversity, for each segment. Linkage Disequilibrium (LD) among SNPs within the *HLA-B* locus was assessed using *Haploview* 4.2³⁴, considering only sites with a minor allele frequency (MAF) higher than 1% and *P*-value for Hardy-Weinberg departure cutoff of 0.001 (Figure S1). The PED and MAP files for *Haploview* were generated using *vcfx haploview*. We evaluated gene diversity and heterozygosity using *Arlequin* 3.5³⁵. The Arp file for *Arlequin* was generated using a local Perl script. To create the network, we used *PopArt* and the Median-Joining³⁶ algorithm, with further processing with Inkscape.

Results

We surveyed the *HLA-B* full genetic diversity in 3,648 samples from 29 different worldwide populations, using a new bioinformatic approach that allows genotyping in the SNP and haplotype levels and the definition of the complete *HLA-B* sequence in each sample. Here, we studied 5,606 nucleotides from *HLA-B*, from position 6:31.353.407-31.359.013 (hg38), comprising about 1,850 nucleotides upstream the first translated ATG, all exons and introns, including the full 3'UTR.

Genotyping and haplotyping accuracy

To evaluate the hla-mapper workflow's genotyping and haplotyping accuracies, we converted our *HLA-B* data from the *1000 Genomes* dataset (using

hla-mapper workflow) and the original data (without hla-mapper optimization) into complete CDS sequences, translating them to the allotypes that each sample encodes. Then, we compared the data with the results from Abi-Rached³⁷, which is largely used as a gold standard for *HLA* alleles in the *1000 Genomes* dataset, despite being analyzed in lower resolution (2-field). For this comparison, we used 2,471 samples, which have passed our haplotyping probability threshold and overlaps Abi-Rached data. The compatibility between the allotypes detected by the hla-mapper workflow and Abi-Rached is 98.75%. The original 1000Genomes data result in the compatibility of only 13.21%. These results demonstrate hla-mapper workflow's accuracy for detecting *HLA-B* SNPs, haplotypes, and *HLA-B* alleles in all resolutions in NGS data.

When we compare the final sequences obtained from the hla-mapper workflow and the original *1000 Genomes* data, we observed that some regions, such as exons 2 and 3, are prone to genotyping errors. The accuracy levels are as low as 10% in some regions (Figure 1). Other *HLA-B* regions also present high discordance rates, such as exon 4, some introns, the 3'UTR, and the proximal promoter. The comparison with Abi-Rached data indicates clearly that using a computational approach such as *hla-mapper* is mandatory to get reliable SNP and haplotype data for *HLA-B* in NGS studies.

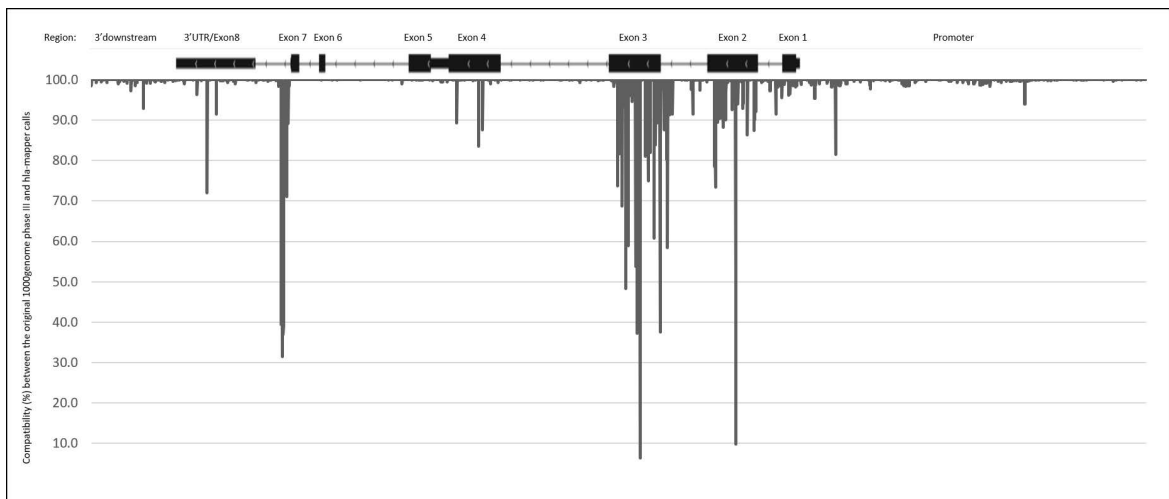


Figure 1: Genotype compatibility across *HLA-B* when comparing the hla-mapper workflow and the original 1000Genomes Phase III data. The *HLA-B* gene orientation observed in the hg38 reference genome is maintained in this illustration.

HLA-B SNPs in different populations, SNP density, and nucleotide diversity

We have found many variable sites throughout *HLA-B*, as depicted in Figure 2. Nucleotide diversity was high (more than 4%) in exons 2 and 3, which encodes the peptide-binding site, but also high in exon 1 (encodes the leader peptide) and intron 1 (Figure 2). Although 30% and 7% of all sites lay down on the promoter region and the 3'UTR, there are many low-frequency variants in these two gene regions, and they present intermediate nucleotide diversity indexes. The region between intron 3 and exon 7 (alpha-3, transmembrane, and cytoplasmic domains) is the most conserved, with low nucleotide diversity in exon 4 (alpha-3 domain) and no variants on exon 6 (transmembrane domain).

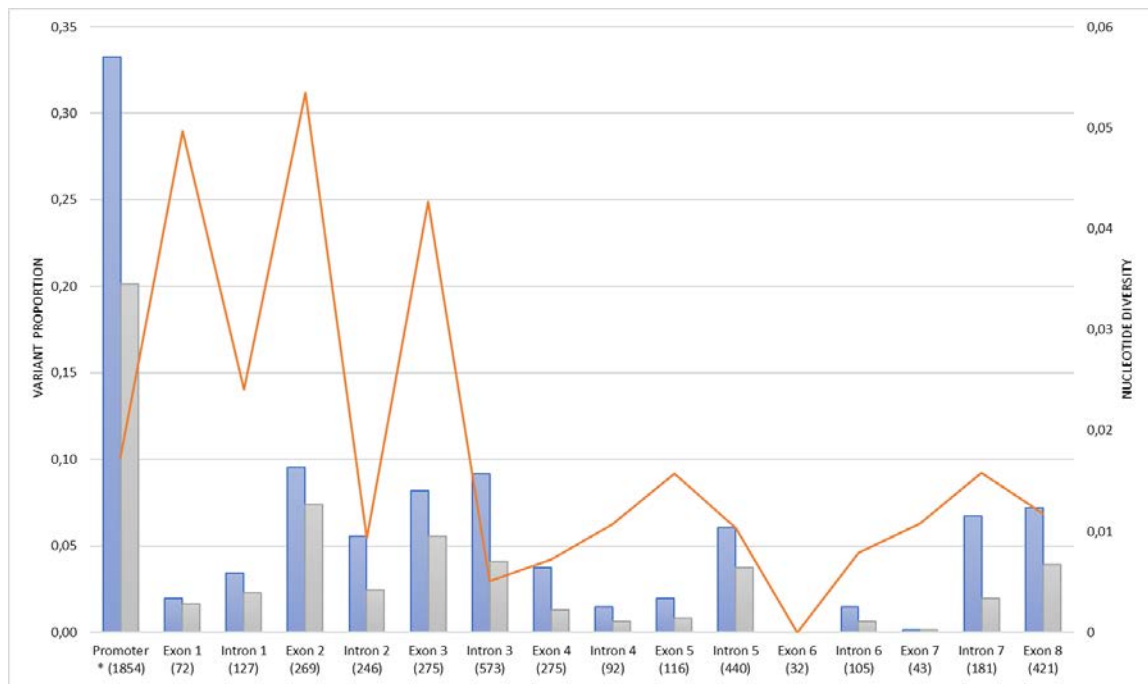


Figure 2: Primary Y-axis: the proportion of total variants (blue bars) and variants with MAF > 1% (gray bars) in each gene region. Secondary Y-axis: nucleotide diversity in each gene region (orange line). *Promoter region (1,854 nucleotides upstream the first translated ATG, including the 5'UTR). Exon 8 corresponds to 99.05% of the *HLA-B* 3'UTR.

We have classified all samples into super populations: Europeans (EUR), Africans (AFR), East Asians (EAS), South Asians (SAS), Americans (AMR), Brazilians (SABE), and Amerindians (Karitiana and Surui from South America). We opted to evaluate Brazilians separately from the remaining admixed Americans

because it configures the largest sample size in this dataset, it involves elderly individuals, and all samples came from the same city (São Paulo).

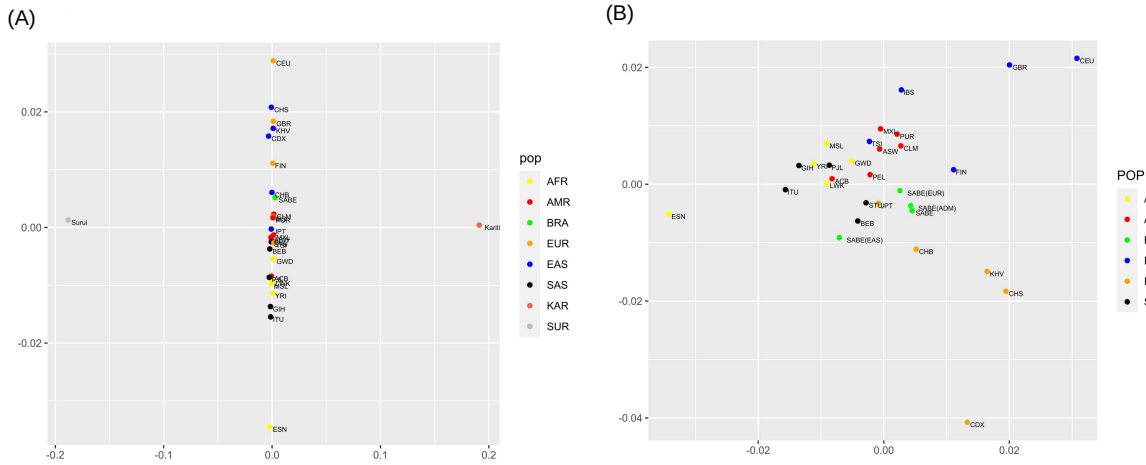


Figure 3. Multidimensional scaling of F_{ST} values obtained for *HLA-B* variation sites among populations of the 1000Genomes, HGDP, and from SABE cohort (Brazil). (A) It includes Amerindian populations. (B) Not include Amerindian populations. AFR: Samples from Africa. AMR: Samples from America. BRA: Brazilian samples from the SABE cohort. EAS: Samples from East Asia. EUR: Samples from Europe. SAS: Samples from South Asia.

Many variants are frequent in all super populations (359 out of 610), and most of them present similar frequencies. However, the frequencies of some variants vary among populations. For instance, the alternative alleles for rs2523608 (intron 5) and rs12721836 (exon 3) are less frequent among East Asians and more frequent among Africans, respectively.

To understand how each population differs regarding *HLA-B* variable sites, we produced an MDS based on the F_{ST} values among pairs of populations, using SNP data (Figure 3). Despite most of the *HLA-B* variants being frequent worldwide, we noticed at least a small population differentiation involving Europeans and East Asians. As anticipated, all American samples from the 1000Genomes dataset clustered between Europeans and Africans but unexpectedly closer to African populations. Because we have ancestry informative markers (AIMs) for all Brazilian samples (SABE)²⁷, we have stratified Brazilians according to their global ancestry into SABE(EUR), with more than 90% of European ancestry, SABE(EAS), with more than 90% of East Asian ancestry, and SABE(ADM), with all samples presenting a mix of different ancestries. SABE(EAS) represents the first generation of Asian

migrants to Brazil²⁷. Thus, it has clustered somewhat closer to Chinese and Japanese populations. SABE(EUR) and SABE(ADM) clustered close to other American samples but relatively distant from the European and African samples (Figure 3B). This pattern indicates that although *HLA-B* SNP frequencies are similar, possibly maintained in this way due to balancing selection, they still keep some demographic signatures from the human evolutionary history, leading to a poorly defined clustering of populations. Amerindian populations are genetically different from other populations and much more distinct from each other than urban populations among them. As a result, they are segregated from the rest of the MDS population (Figure 3A).

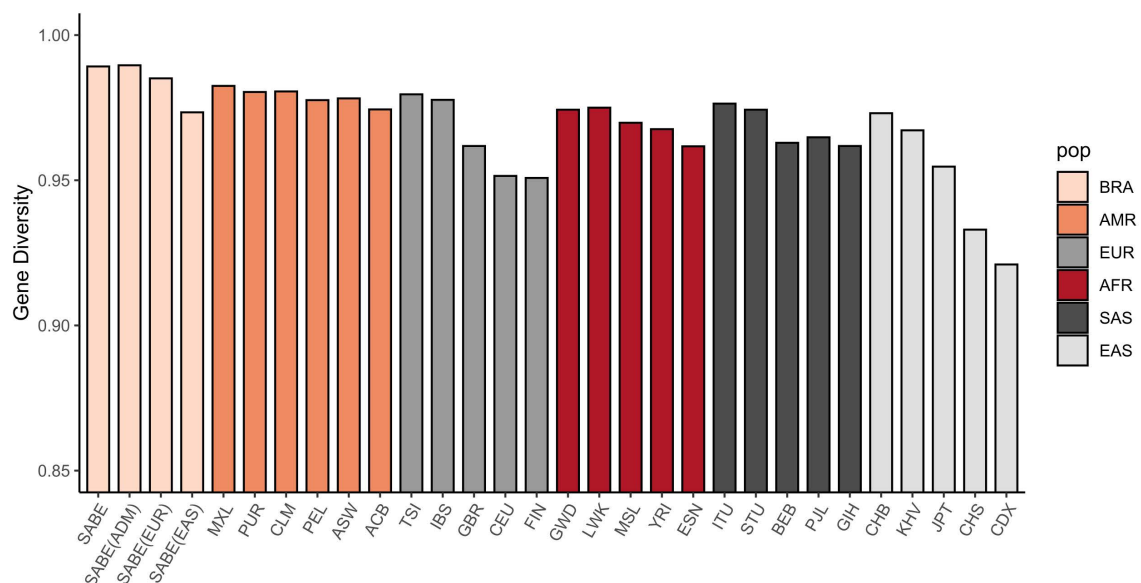


Figure 4. *HLA-B* gene diversity across different populations. AFR: Samples from Africa. AMR: Samples from America. BRA: Brazilian samples from the SABE cohort. EAS: Samples from East Asia. EUR: Samples from Europe. SAS: Samples from South Asia.

We have also calculated the *HLA-B* gene diversity in each population (Figure 4). Gene diversity was high in every population but higher among admixed samples (Brazilians and other populations from America). The lowest indexes were observed for some East Asian and European populations. We omitted Amerindians because of the low sample size for this group.

***HLA-B* allotype frequencies in different populations**

After translating the CDS sequences into protein sequences (the allotypes or 2-field resolution allele), we identified 171 different *HLA-B* allotypes, 162 of them already reported in the IPD-IMGT/HLA database version 3.43.0. The remaining ones (9 new allotypes) are rare and present a summed frequency of 0.18%. We have also detected copies of rare *HLA-B* allotypes, such as B*15:220, B*15:450, B*35:80, and others. A supplementary file with the allotypes for each sample is available upon request.

Some allotypes are exclusively associated with specific populations or are frequent in some groups and rare in others. The most frequent allotypes worldwide were B*51:01, B*35:01, and B*07:02, frequent in all super populations except Amerindians. While most of the SNPs are shared among super populations (359 out of 610), only 17 out of the 171 allotypes occur in all super populations (not including Amerindians), and they present a summed frequency of 46.08%. Thus, there is a weaker allotype sharing among super populations (Figure S2). When we produced an MDS considering the *HLA-B* allotypes, we observed that the Amerindian's population present allotypes that differ from each other and other populations, resulting in separation among these populations (Figure 5A). When we do not consider the Amerindian populations in this analysis, we also noticed a separation of African, East Asian, and European samples, with admixed samples (Brazil and American ones) between Europeans and Africans, and samples of South Asia between Europeans and East Asians (Figure 5B).

Some clinically relevant alleles present different frequencies in specific populations. B*53:01, frequent in Africa, is associated with protection against malaria infection^{38–40}. B*57:01, associated with protection against HIV infection^{16,41} and hypersensitivity to abacavir drug^{42,43}, is frequent in South Asia. *HLA-B**27, associated with protection against HIV infection and predisposition to ankylosing spondylitis^{44,45}, is frequent among Europeans. B*46:01 (in all levels) is highly frequent among East Asians, and it is associated with protection against leprosy infection and has been recently associated with Sars-CoV-2 as well^{46,47}.

***HLA-B* regulatory sequences**

We found 123 different promoter sequences considering 1,854bp upstream the first translated ATG. These sequences are clustered into 15 promoter lineages, each one associated with specific *HLA-B* allele groups (Figure S3). The only exceptions are promoters P09 and P15, both associated with alleles from the B*40 group but with different CDS sequences. The most frequent promoter sequence worldwide is P13 (28.45%), associated with B*31, B*51, B*52, B*58, and B*78 alleles.

The *HLA-B* promoter region includes at least 30% of the variants detected in this survey, and almost 20% of the frequent ones (Figure 1). It also presents an SNP density similar to the observed for other *HLA-B* regions. Unlike what was observed for *HLA-B* allotypes, the 15 shared promoters among all super populations present a high summed frequency of 83.17%.

There were 73 different 3'UTR sequences, with many shared variants and few mutational steps between them. Thus, these sequences were named following their global frequencies. The most frequent 3'UTR sequence was U01. This 3'UTR sequence is the most frequent one in all populations except EAS (Figure S4). The shared 3'UTR sequences (15 sequences) present a summed frequency of 87.63%.

Discussion

We present a molecular method to amplify the full *HLA-B* gene region, including all exons, introns, and regulatory sequences. We also present a bioinformatics pipeline to assess the full *HLA-B* genetic diversity in all levels (SNPs, alleles, allotypes, and regulatory/coding haplotypes) for whole-genome sequencing. Although not addressed here, this method is also suitable for targeted amplicon sequencing and exomes. We provide accurate genotypes and haplotypes from the *HLA-B* upstream promoter up to the complete 3'UTR in 3,648 samples from 29 populations in 4 continents. Our bioinformatics pipeline relies on the *hla-mapper* software, developed to minimize mapping bias and cross-mappings among *HLA* genes²², and also on the power of the *hla-mapper* + GATK HaplotypeCaller + vcfx + Phasex workflow to call genotypes and to detect haplotypes using both probabilistic models and read-aware phasing.

The consistency between our data and the Abi-Rached results³⁷, one of the most used resources for HLA alleles imputation, clearly validates our method's efficiency. We have detected the same allotypes in more than 98.7% of the samples. Moreover, the few inconsistencies seem to represent miscalled alleles, with two extremely rare alleles called by Abi-Rached, while our method called two different common alleles.

We have to consider two issues that prevent traditional aligning, SNP-calling, and haplotyping workflows, like the one used by the 1000Genomes consortium, for getting reliable *HLA-B* data. First, the original 1000Genomes phase III data (the phased VCF) was produced based on low coverage sequencing. Thus, the need for imputation to fill in the blanks might bias the final genotypes and haplotypes. This issue was circumvented with the recent publication of high-coverage sequencing data for all samples²⁴. Second, conventional aligners do not handle HLA reads adequately, producing cross-mappings between genes presenting high sequence similarities (e.g., *HLA-B* and *HLA-C*) and mapping failures when the sequenced alleles are too different from those of the reference genome^{22,23}. Thus, specific approaches such as *hla-mapper* can correct such alignments, allowing a more accurate genotype and haplotype call procedure. The new release of the 1000Genomes (high-coverage sequencing) did not employ an HLA-specific approach to call genotypes and haplotypes. The primary issue in this situation is mapping bias and not sequencing coverage. Therefore, using a computational approach such as the *hla-mapper* is essential to obtain reliable *HLA-B* SNP and haplotype calls when using NGS.

Here, we present the first study to address the full *HLA-B* genetic diversity in such a large dataset, using high-coverage NGS and reliable methods to call genotypes and haplotypes. Moreover, this new dataset includes more than a thousand highly admixed samples from Brazil, which is misrepresented in all known databases. In a previous study that used the data presented here²⁷, we demonstrated that imputation accuracy for *HLA-B* from SNP arrays increases by about 5% when using the 1000Genomes+SABE datasets. This initiative to improve HLA imputation is part of an international consortium (SHLARC) to build and share MHC-centric analysis in genomics⁴⁸. We propose using this new dataset as a reference for HLA imputation whenever necessary and that other samples

(particularly admixed ones) that undergo whole-genome sequencing (WGS) be processed using this same pipeline to increase the reference panels.

HLA-B is the most polymorphic *HLA* gene, with at least 7,967 alleles encoding 4,962 different proteins reported in the IPD-IMGT/HLA database so far. However, we detected only 231 different alleles out of these 7,562 sequences (3%) when evaluating more than 3,648 samples from worldwide populations. Thus, many of these alleles are extremely rare, and the majority have been described in just one individual.

At the SNP level, *HLA-B* is indeed one of the most variable genes in the genome. We also noticed that 58.8% of all the variable sites we have detected are shared among the super populations and present similar frequencies in different populations. They represent 92% of the most frequent variants, and the remaining are very rare nucleotide exchanges. Thus, the populations do not strongly differ in terms of the presence and frequency of variable sites. Although we have a separation of most of the European, and East Asian samples in the MDS displayed in Figure 3, this separation is not evident for samples from South Asia, or even from Italy (TSI), Japan (JPT), or Brazil (SABE-EUR and SABE-ADM). On the other hand, the separation among samples is better when we consider the haplotypes in Figure 5 (in this case, the allotypes) since only 17 out of 171 are shared among all super populations. We observed that the Surui and Karitiana populations are distinct and distant from the other populations. This position is not different from what was expected due to the demographic history of those populations. Europeans, Africans, and East Asians forming the vertices of an imaginary triangle, with the south Asians being placed between Europeans and East Asians, recalling their geographic positions, and with admixed samples between Europeans and Africans, recalling their recent demographic formation. Thus, although most of the variants are maintained and shared, these variants configure different haplotypes in each population. Despite the increased diversity of *HLA* alleles generated by population-specific events of intragenic recombination and gene conversion, which contributes to population differentiation, the maintenance of these different allotypes by an initial positive selection followed by balancing selection also contributes to preserving SNPs in worldwide populations⁴⁹. Furthermore, the role of conversion in *HLA*

polymorphism is most evident for the *HLA-B* locus⁵⁰. Therefore, the low differentiation in SNPs and high differentiation in *HLA* alleles is not unexpected⁵¹.

Gene diversity was high (greater than 0.9) in all populations, as depicted in Figure 4, but particularly high in the Brazilian samples. The admixed Brazilian samples (SABE-ADM) presented the highest gene diversity among all the evaluated populations. The Brazilian genetic background is quite heterogeneous, with major contributions from Europeans, Native Americans, and Africans, making the Brazilian population one of the most admixed populations in the world. This admixed nature increases its genetic diversity, with the presence of common (and sometimes rare) haplotypes from Europe, Africa, and Asia (mostly from Native Americans). Even the SABE subgroup of European descendants presents a higher gene diversity than observed for all European samples, possibly because the former is composed of different European subcontinental ancestries (Portuguese, Spanish, Italian, German) in addition to a smaller non-European contribution. Because of that, it is essential to include admixed samples in all reference panels used for imputation, especially for HLA imputation, as we have demonstrated elsewhere²⁷.

The *HLA-B*40:02* allele represents 41.6% of all HLA-B allotypes observed for Surui in this study, followed by B*40:03 (33.3%), B*35:05 (16.6%), and B*48:02 (8.3%). In Karitiana the *HLA-B*39:05* allele represents 33.3% of all allotypes, followed by B*35:05 (27.7%), B*39:06 (22.2%), B*39:02 (11.1%), and B*39:09 (5.56%). *HLA-B*40*^{52–54}, B*48^{52,53}, B*39, and B*35^{52,54,55} are common among Amerindians, and many subtypes are found in these populations. The alleles found in these populations include subtypes of six evolutionary lineage or serological groups (B5, B15, B35, B39, B40, and B48)⁵². The *HLA* allele repertoire of Amerindians is lower than other populations in several studies⁵³. This is consistent with the demographic and evolutionary history of ancient Native Americans that gave rise to Amazonian Amerindian populations, which involves founder effects and bottlenecks. In addition, current Amazonian Amerindian tribes are characterized by populations with small effective sizes that experience intense effects of genetic drift coupled with low levels of gene flow^{56,57}. However, *HLA-B* evolves faster in these populations than other *HLA* genes⁵⁸, and most of these new *HLA-B* alleles were originated from gene conversion involving Asian alleles⁵². For instance, in the Wichi tribe, from North-East Argentina, the origin of the *HLA-B*35:19* allele is probably the

result of gene conversion involving intra-locus recombination between the B*40:02 and B*35:01⁵⁴. Although we have only evaluated few Amerindian samples, this low *HLA-B* diversity was already described for Amerindians⁵³.

We found 17 alleles (*HLA-B*07:14*, *HLA-B*15:31*, *HLA-B*15:47*, *HLA-B*27:08*, *HLA-B*27:09*, *HLA-B*35:32*, *HLA-B*35:80*, *HLA-B*39:13*, *HLA-B*39:37*, *HLA-B*39:09*, *HLA-HLA-B*40:129*, *HLA-B*40:16*, *HLA-B*44:04*, *HLA-B*44:50*, *HLA-B*51:04*, *HLA-B*51:32*, *HLA-B*53:04*) exclusively in the Brazilian dataset or among Brazilians and Amerindians. Thus, there is not any copy of these alleles in the 1000 Genomes dataset, reinforcing the lack of data from admixed and Native American samples. This missing is due to undersampling and incomplete geographical coverage in the 1000 Genomes dataset³⁷. Furthermore, most of the samples in available reference panels are of European descents⁵⁹. Therefore, this data may be used to improve existing panels.

There is a positive correlation between a load of pathogens in each geographic region and the levels of diversity in the *HLA* loci, especially in *HLA-B*⁶⁰. The diversity of this gene is directly related to the ability to present different peptides, with a significant role in many disease outcomes^{16,38,42,44,61}. Thus, the *HLA*/pathogen coevolution can lead the locally adaptive *HLA* alleles to higher frequencies and thus cause an increase in population differentiation. These alleles are distributed in intermediate frequencies, taking the diversity indices and values higher than expected according to simulations that take into account the sample size and the number of alleles⁶². However, despite that, most of the allele frequencies do not exceed 10%, exception made to *HLA-B*53:01* among Africans, and *HLA-B*46:01* among East Asians. On the other hand, in Amerindians, the founder effect, genetic drift, and geographic isolation may be more determinant than selection in maintaining distinct allotypes.

In particular populations, some clinically significant alleles presented different frequencies (Figure S5), such as *HLA-B*53:01* (severe malaria resistance^{38,40}), which is frequent in Africa, *HLA-B*57:01* (HIV⁶³, hypersensitivity reaction to abacavir^{42,43}), frequent in South Asia, *HLA-B*27* (ankylosing spondylitis^{45,64}), frequent among Europeans, and *HLA-B*46:01* (resistance to leprosy^{46,47,65}), restricted to East Asians.

The most conserved *HLA-B* region was exon 6, which, together with exon 7, encodes the cytoplasmic tail. Indeed, no frequent variant in this region is reported in the IPD-IMGT/HLA, and the same pattern occurs with *HLA-C*²⁹. As expected, exons 2 and 3 presented high levels of nucleotide diversity (Figure 1). However, we notice high nucleotide diversity in exon 1 (the leader peptide), some introns such as intron 1, and in the promoter region. For instance, exon 1 encodes many different leader peptides, with at least six frequent non-synonymous mutations. The leader peptide (or signal peptide) is involved in targeting *HLA-B* to the cell membrane and is not part of the mature *HLA-B* molecule. Furthermore, the leader peptide of classical *HLA* class I molecules binds to *HLA-E* proteins, stabilizing *HLA-E* and allowing it to load to the cell surface, where it interacts with CD94/NKG2A receptors in NK cells⁵⁰. All *HLA-A* and *HLA-C* molecules carry P2-Methionine, while *HLA-B* molecules carry Methionine or Threonine in P2⁶⁶. P2-methionine has a higher *HLA-E* binding affinity than P2-threonine. Thus, individuals with *HLA-B* P2-Methionine alleles have higher *HLA-E* surface expression⁶⁷. *HLA-B* P2-Methionine haplotypes rarely encode *HLA-Bw4* and *HLA-C2* that bind to KIR receptors. Rather, they encode low-expression *HLA-C1* variants⁶⁸. Thus, there are two types of *HLA-B* haplotypes. The first one usually encodes CD94/NKG2A ligands and this generates an inhibitory signal for the NK cell. The second one encodes KIR ligands, which can generate an inhibitory or activator signal for these cells, changing the immune response profile, by cytokine environment, and recruiting an inflammatory response, favoring or not the development of the neoplastic, autoimmune, or infection disease^{68,69}. Some studies have suggested that these variations may influence the clinical outcome of immunotherapy in acute myeloid leukemia⁶⁷ and the outcome of HIV infections⁶⁶.

The *HLA-B* promoter region is more polymorphic than that of *HLA-A* or *HLA-C*^{29,70}. However, despite the high levels of nucleotide diversity found in this region, low variation has been observed in the *HLA-B* allele-specific mRNA expression, indicating that promoter polymorphisms do not significantly affect mRNA production for this locus⁷¹. Interestingly, each allele group seems to present different promoter sequences, and these sequences seem to be shared among super populations. However, *HLA-B* molecules with a more restricted peptide repertoire (e.g., *HLA-B*57:01*, *HLA-B*27:05*, and *HLA-B*07:02*) generally bind strongly to these peptides

and appears to be more stable and more expressed in the cell surface than others with a broader peptide repertoire (e.g., *HLA-B*35:01*). These same molecules were classified as being tapasin-dependent and tapasin-independent, respectively¹⁸. Because of that, HLA-B protein expression may be influenced by many different mechanisms, starting from the binding of transcription factors to specific promoter sequences, the binding of microRNAs to the highly polymorphic 3'UTR, but also peptide binding affinity, protein stability, and addressing.

In conclusion, our molecular and computational method proved to be effective to obtain reliable *HLA-B* genotypes and haplotypes when using NGS. We used this computational workflow to evaluate the entire *HLA-B* genetic diversity for 29 different worldwide populations. *HLA-B* gene diversity is higher in admixed populations, particularly in Brazilians, and this new dataset improves HLA imputation accuracy. The *HLA-B* promoter, exon 1, and the 3'UTR are highly variable, and each allele group seems to be in linkage disequilibrium with specific regulatory sequences. Most of the SNPs are shared worldwide (at least the frequent ones), while the opposite is observed for haplotypes. We also provide frequencies of each allotype (DNA and protein sequence), promoter sequences, 3'UTR sequences, and full allele sequences (genomic allele) for all samples we have studied.

Acknowledgments

This work was supported by Fundação de Amparo à Pesquisa do Estado de São - FAPESP/Brazil (Grants 2017/19223-0 and 2013/17084-2). This study was partially funded by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES)—Finance Code 001.

References

1. Donadi EA, Castelli EC, Arnaiz-Villena A, Roger M, Rey D, Moreau P. Implications of the polymorphism of HLA-G on its function, regulation, evolution and disease association. *Cell Mol Life Sci.* 2011;68(3):369–95.
2. Klein, J.; Sato A. The HLA System — First of Two Parts. *N Engl J Med.* 2000;343:702–9.
3. FISCHER, G. F; MAYR WR. molecular genetics of the HLA complex. *Wien klin wochenschr.* 2001;113(20–21):814–24.
4. HUGHES, A. L; YEAGER M. Natural selection and the evolutionary history of major histocompatibility complex loci. *Front biosci.* 1998;3:509–16.
5. Klein, J.; Sato A. THE HLA SYSTEM - Second of Two Parts. *N Engl J Med.* 2000;343(11):782–6.
6. Robinson J, Barker DJ, Georgiou X, Cooper MA, Flicek P, Marsh SGE. IPD-IMGT/HLA Database. *Nucleic Acids Res.* 2020 Jan 1;48(D1):D948–55.
7. Sampaio-Barros PD, Conde RA, Donadi EA, Bonfiglioli R, Costallat LTL, Samara AM, et al. Frequency of HLA-B27 and its alleles in patients with Reiter syndrome: Comparison with the frequency in other spondyloarthropathies and a healthy control population. *Rheumatol Int.* 2008;28(5):483–6.
8. van Deutekom HWM, Keşmir C. Zooming into the binding groove of HLA molecules: which positions and which substitutions change peptide binding most? *Immunogenetics.* 2015;67(8):425–36.
9. Thorsby E, Lie BA. HLA associated genetic predisposition to autoimmune diseases: Genes involved and possible mechanisms. *Transpl Immunol.* 2005;14(3-4 SPEC. ISS.):175–82.
10. Shiina T, Inoko H, Kulski JK. An update of the HLA genomic region, locus information and disease associations: 2004. *Tissue Antigens.* 2004;64(6):631–49.
11. Giza M, Koftori D, Chen L, Bowness P. Is Behçet's disease a 'class 1-opathy'? The role of HLA-B*51 in the pathogenesis of Behçet's disease. *Clin Exp Immunol.* 2018 Jan 1;191(1):11–8.
12. Noble JA, Valdes AM, Bugawan TL, Apple RJ, Thomson G, Erlich HA. The HLA class I A locus affects susceptibility to type 1 diabetes. *Hum Immunol.* 2002 Aug;63(8):657–64.
13. Goulder PJR, Watkins DI. Impact of MHC class I diversity on immune control of immunodeficiency virus replication. Vol. 8, *Nature Reviews Immunology.* 2008. p. 619–30.
14. Castelli EC, Mendes CT, Viana De Camargo JL, Donadi EA. HLA-G

polymorphism and transitional cell carcinoma of the bladder in a Brazilian population. *Tissue Antigens*. 2008 Aug;72(2):149–57.

15. Gao X, Nelson GW, Karacki P, Martin MP, Phair J, Kaslow R, et al. Effect of a single amino acid change in MHC class I molecules on the rate of progression to aids. *N Engl J Med*. 2001;344(22):1668–75.

16. Borghans JAM, Mølgaard A, de Boer RJ, Keşmir C. HLA alleles associated with slow progression to AIDS truly prefer to present HIV-1 p24. *PLoS One*. 2007;2(9).

17. Altfeld M, Kalife ET, Qi Y, Streeck H, Lichterfeld M, Johnston MN, et al. HLA alleles associated with delayed progression to AIDS contribute strongly to the initial CD8+ T cell response against HIV-1. *PLoS Med*. 2006;3(10):1851–64.

18. Rizvi SM, Salam N, Geng J, Qi Y, Bream JH, Duggal P, et al. Distinct Assembly Profiles of HLA-B Molecules. *J Immunol*. 2014;192(11):4967–76.

19. Petersdorf EW, Hansen JA, Martin PJ, Woolfrey A, Malkki M, Gooley T, et al. Major-Histocompatibility-Complex Class I Alleles and Antigens in Hematopoietic-Cell Transplantation. *N Engl J Med*. 2001 Dec 20;345(25):1794–800.

20. Ravazzi-Gauch C, Bajay MM, Caldas HC, Abbud-Filho M. HLA-A, -B, and -DRB1 allele and haplotype diversity in a cohort of Brazilian renal transplant candidates. Vol. 77, *Human Immunology*. Elsevier Inc.; 2016. p. 464–9.

21. Boquett JA, Bisso-Machado R, Zagonel-Oliveira M, Schüler-Faccini L, Fagundes NJR. HLA diversity in Brazil. *Hla*. 2020;95(1):3–14.

22. 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. *Hum Immunol*. 2018;79(9):678–84.

23. Brandt DYC, Aguiar VRC, 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 Genes, Genomes, Genet*. 2015;5(5):931–41.

24. Byrska-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, et al. High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *bioRxiv*. 2021;2021.02.06.430068.

25. Cann HM, Toma C De, Cazes L, Legrand M, Morel V, Piouffre L, et al. A Human Genome Diversity Cell Line Panel. 2016;8–10.

26. Naslavsky MS, Yamamoto GL, de Almeida TF, Ezquina SAM, Sunaga DY, Pho N, et al. Exomic variants of an elderly cohort of Brazilians in the ABraOM database. *Hum Mutat*. 2017 Jul 1;38(7):751–63.

27. Naslavsky MS, Scliar MO, Yamamoto GL, Yu Ting Wang J, Zverinova S, Karp T, et al. Whole-genome sequencing of 1,171 elderly admixed individuals from the largest Latin American metropolis (São Paulo, Brazil). *Wagner CS Magalhães*. 20AD Sep 16;10:24.
28. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009 Aug;25(16):2078–9.
29. Souza AS, Sonon P, Paz MA, Tokplonou L, Lima THA, Porto IOP, et al. HLA-C genetic diversity and evolutionary insights in two samples from Brazil and Benin. *HLA*. 2020 Oct 24;96(4):468–86.
30. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytzky A, et al. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20(9):1297–303.
31. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. *Nat Methods*. 2012;9(2):179–81.
32. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet*. 2007;81(5):1084–97.
33. Rice P, Longden L, Bleasby A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet*. 2000;16(6):276–7.
34. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics*. 2005 Jan 15;21(2):263–5.
35. Excoffier L, Laval G, Schneider S. Arlequin (version 3.0): An integrated software package for population genetics data analysis. *Evol Bioinforma*. 2005 Jan;1:117693430500100.
36. Bandelt HJ, Forster P, Röhl A. Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol*. 1999;16(1):37–48.
37. Abi-Rached L, Gouret P, Yeh J-H, Di Cristofaro J, Pontarotti P, Picard C, et al. Immune diversity sheds light on missing variation in worldwide genetic diversity panels. *DeAngelis MM*, editor. *PLoS One*. 2018 Oct 26;13(10):e0206512.
38. Hill AVS, Elvin J, Willis AC, Aidoo M, Allsopp CEM, Gotch FM, et al. Molecular analysis of the association of HLA-B53 and resistance to severe malaria. *Nature*. 1992;360(6403):434–9.
39. Hill AVS, Allsopp CEM, Kwiatkowski D, Anstey NM, Twumasi P, Rowe PA, et al. Common West African HLA antigens are associated with protection from severe malaria. *Nature*. 1991;352(6336):595–600.

40. Aidoo M, Lalvani A, Gilbert SC, Ting JH, Daubersies P, Hurt N, et al. Cytotoxic T-Lymphocyte Epitopes for HLA-B53 and Other HLA Types in the Malaria Vaccine Candidate Liver-Stage Antigen 3. Vol. 68, *INFECTION AND IMMUNITY*. 2000.
41. Klein MR, Van Der Burg SH, Hovenkamp E, Holwerda AM, Drijfhout JW, Melief CJM, et al. Characterization of HLA-B57-restricted human immunodeficiency virus type 1 Gag- and RT-specific cytotoxic T lymphocyte responses. *J Gen Virol*. 1998;79(9):2191–201.
42. Ostrov DA, Grant BJ, Pompeu YA, Sidney J, Harndahl M, Southwood S, et al. Drug hypersensitivity caused by alteration of the MHC-presented self-peptide repertoire. *Proc Natl Acad Sci U S A*. 2012;109(25):9959–64.
43. Dean L. Abacavir Therapy and HLA-B * 57 : 01 Genotype Drug : Abacavir. 2012;91(4):1–7.
44. Bowness P. HLA-B27. 2015;
45. Lin H, Gong YZ. Association of HLA-B27 with ankylosing spondylitis and clinical features of the HLA-B27-associated ankylosing spondylitis: a meta-analysis. *Rheumatol Int*. 2017;37(8):1267–80.
46. Hilton HG, McMurtrey CP, Han AS, Djaoud Z, Guethlein LA, Blokhuis JH, et al. The Intergenic Recombinant HLA-B*46:01 Has a Distinctive Peptidome that Includes KIR2DL3 Ligands. *Cell Rep*. 2017;19(7):1394–405.
47. Wang LM, Kimura A, Satoh M, Mineshita S. HLA linked with leprosy in southern China: HLA-linked resistance alleles to leprosy. *Int J Lepr Other Mycobact Dis*. 1999;67(4):403–8.
48. Vince N, Douillard V, Geffard E, Meyer D, Castelli EC, Mack SJ, et al. SNP-HLA Reference Consortium (SHLARC): HLA and SNP data sharing for promoting MHC-centric analyses in genomics. *Genet Epidemiol*. 2020 Oct 18;44(7):733–40.
49. Borghans JAM, Beltman JB, De Boer RJ. MHC polymorphism under host-pathogen coevolution. *Immunogenetics*. 2004;55:732–9.
50. Parham P, Ohta T. Population biology of antigen presentation by MHC class I molecules. *Science (80-)*. 1996 Apr 5;272(5258):67–74.
51. Meyer D, Vitor VR, Bitarello BD, Débora DY, Nunes K. A genomic perspective on HLA evolution. *Immunogenetics*. 2018;70(1):5–27.
52. Fernández-Viña MA, Lázaro AM, Marcos CY, Nulf C, Raimondi E, Haas EJ, et al. Dissimilar evolution of B-locus versus A-locus and class II loci of the HLA region in South American Indian tribes. *Tissue Antigens*. 1997;50(3):233–50.
53. Belich MP, Madrigal JA, Hildebrand WH, Zemmour J, Williams RC, Luz R, et al. Unusual HLA-B alleles in two tribes of Brazilian Indians. *Nature*. 1992;357:326–

9.

54. Marcos CY, Fernández-Viña MA, Lázaro AM, Moraes ME, Moraes JR, Stastny P. Novel HLA-A and HLA-B alleles in South American Indians. *Tissue Antigens*. 1999;53(5):476–85.

55. Marcos CY, Fernández-Viña MA, Lázaro AM, Nulf CJ, Raimondi EH, Stastny P. Novel HLA-B35 subtypes putative gene conversion events with donor sequences from alleles common in native americans (HLA-B*4002 or B*4801). *Hum Immunol*. 1997;53(2):148–55.

56. Fuselli S, Tarazona-Santos E, Dupanloup I, Soto A, Luiselli D, Pettener D. Mitochondrial DNA diversity in South America and the genetic history of andean highlanders. *Mol Biol Evol*. 2003 Jun 27;20(10):1682–91.

57. Tarazona-Santos E, Carvalho-Silva DR, Pettener D, Luiselli D, De Stefano GF, Labarga CM, et al. Genetic differentiation in South Amerindians is related to environmental and cultural diversity: Evidence from the Y chromosome. *Am J Hum Genet*. 2001 Jun;68(6):1485–96.

58. Watkins DI, McAdam SN, Liu X, Strang CR, Milford EL, Levine CG, et al. New recombinant HLA-B alleles in a tribe of South American Amerindians indicate rapid evolution of MHC class I loci. *Nature*. 1992;357:329–33.

59. Popejoy AB, Fullerton SM. Genomics is failing on diversity. Vol. 538, *Nature*. Nature Publishing Group; 2016. p. 161–4.

60. O'Callaghan CA. Molecular basis of human natural killer cell recognition of HLA-E (human leucocyte antigen-E) and its relevance to clearance of pathogen-infected and tumour cells. *Clin Sci*. 2000;99(1):9–17.

61. Nguyen A, David JK, Maden SK, Wood MA, Weeder BR, Nellore A, et al. Human Leukocyte Antigen Susceptibility Map for Severe Acute Respiratory Syndrome Coronavirus 2. *J Virol*. 2020 Apr 17;94(13).

62. Solberg OD, Mack SJ, Lancaster AK, Single RM, Tsai Y, Sanchez-Mazas A, et al. Balancing selection and heterogeneity across the classical human leukocyte antigen loci: A meta-analytic review of 497 population studies. *Hum Immunol*. 2008 Jul;69(7):443–64.

63. Yu XG, Lichterfeld M, Chetty S, Williams KL, Mui SK, Miura T, et al. Mutually Exclusive T-Cell Receptor Induction and Differential Susceptibility to Human Immunodeficiency Virus Type 1 Mutational Escape Associated with a Two-Amino-Acid Difference between HLA Class I Subtypes. *J Virol*. 2007;81(4):1619–31.

64. Brown MA, Pile KD, Kennedy LG, Calin A, Darke C, Bell J, et al. HLA class I associations of ankylosing spondylitis in the white population in the United Kingdom. *Ann Rheum Dis*. 1996;55(4):268–70.

65. Barber LD, Percival L, Valiante NM, Chen L, Lee C, Gumperz JE, et al. The inter-locus recombinant HLA-B*4601 has high selectivity in peptide binding and functions characteristic of HLA-C. *J Exp Med*. 1996 Aug 1;184(2):735–40.
66. Merino AM, Sabbaj S, Easlick J, Goepfert P, Kaslow RA, Tang J. Dimorphic HLA-B signal peptides differentially influence HLA-E- and natural killer cell-mediated cytotoxicity of HIV-1-infected target cells. *Clin Exp Immunol*. 2013 Dec;174(3):414–23.
67. Hallner A, Bernson E, Hussein BA, Sander FE, Brune M, Aurelius J, et al. The HLA-B 221 dimorphism impacts on NK cell education and clinical outcome of immunotherapy in acute myeloid leukemia. *Blood*. 2019 Mar 28;133(13):1479–88.
68. Horowitz A, Djaoud Z, Nemat-Gorgani N, Blokhuis J, Hilton HG, Béziat V, et al. Class I HLA haplotypes form two schools that educate NK cells in different ways. *Sci Immunol*. 2016 Sep 9;1(3):eaag1672–eaag1672.
69. Lee N, Llano M, Carretero M, Akiko-Ishitani, Navarro F, López-Botet M, et al. HLA-E is a major ligand for the natural killer inhibitory receptor CD94/NKG2A. *Proc Natl Acad Sci U S A*. 1998 Apr 28;95(9):5199–204.
70. Lima THA, Souza AS, Porto IOP, Paz MA, Veiga-Castelli LC, Oliveira MLG, et al. HLA-A promoter, coding, and 3'UTR sequences in a Brazilian cohort, and their evolutionary aspects. *HLA*. 2019 Feb;93(2–3):65–79.
71. Ramsuran V, Hernández-Sánchez PG, O'hUigin C, Sharma G, Spence N, Augusto DG, et al. Sequence and Phylogenetic Analysis of the Untranslated Promoter Regions for HLA Class I Genes. *J Immunol*. 2017;198(6):2320–9.

Supplementary Material

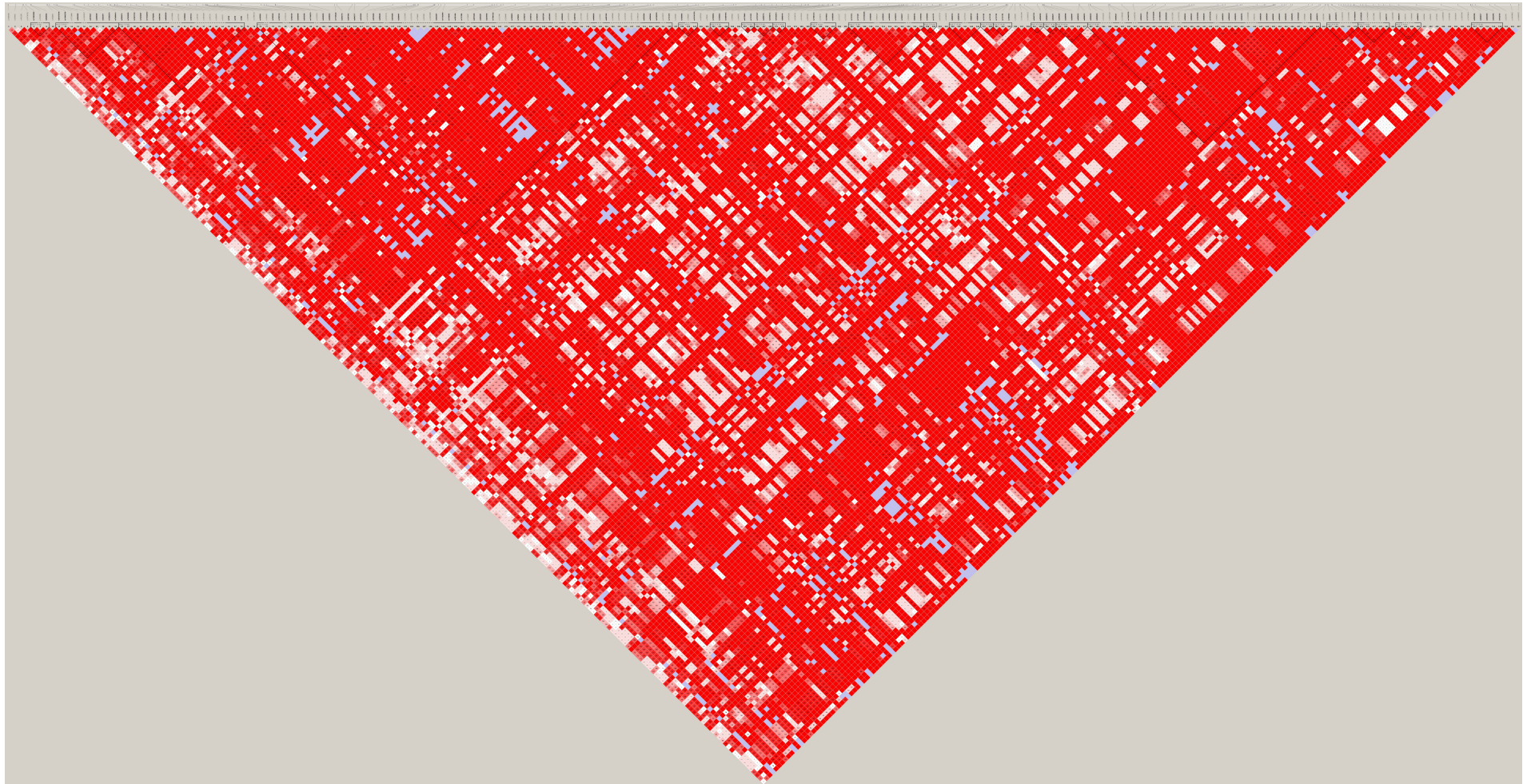


Figure S1. Linkage Disequilibrium (LD) between pair of single nucleotide polymorphisms (SNPs) of the *HLA-B* locus, considering variants with a minimum allele frequency (MAF) > 1% and P-value for Hardy-Weinberg departure cutoff of 0.001. Areas in red indicate strong LD ($\text{LOD} \geq 2$, $D' = 1$); areas in light red indicate moderate LD ($\text{LOD} \geq 2$, $D' < 1$); areas in blue indicate weak LD ($\text{LOD} < 2$, $D' = 1$); and white areas indicate very weak LD ($\text{LOD} < 2$, $D' < 1$).; LOD: Log of the odds – measure of confidence in the value of D' .

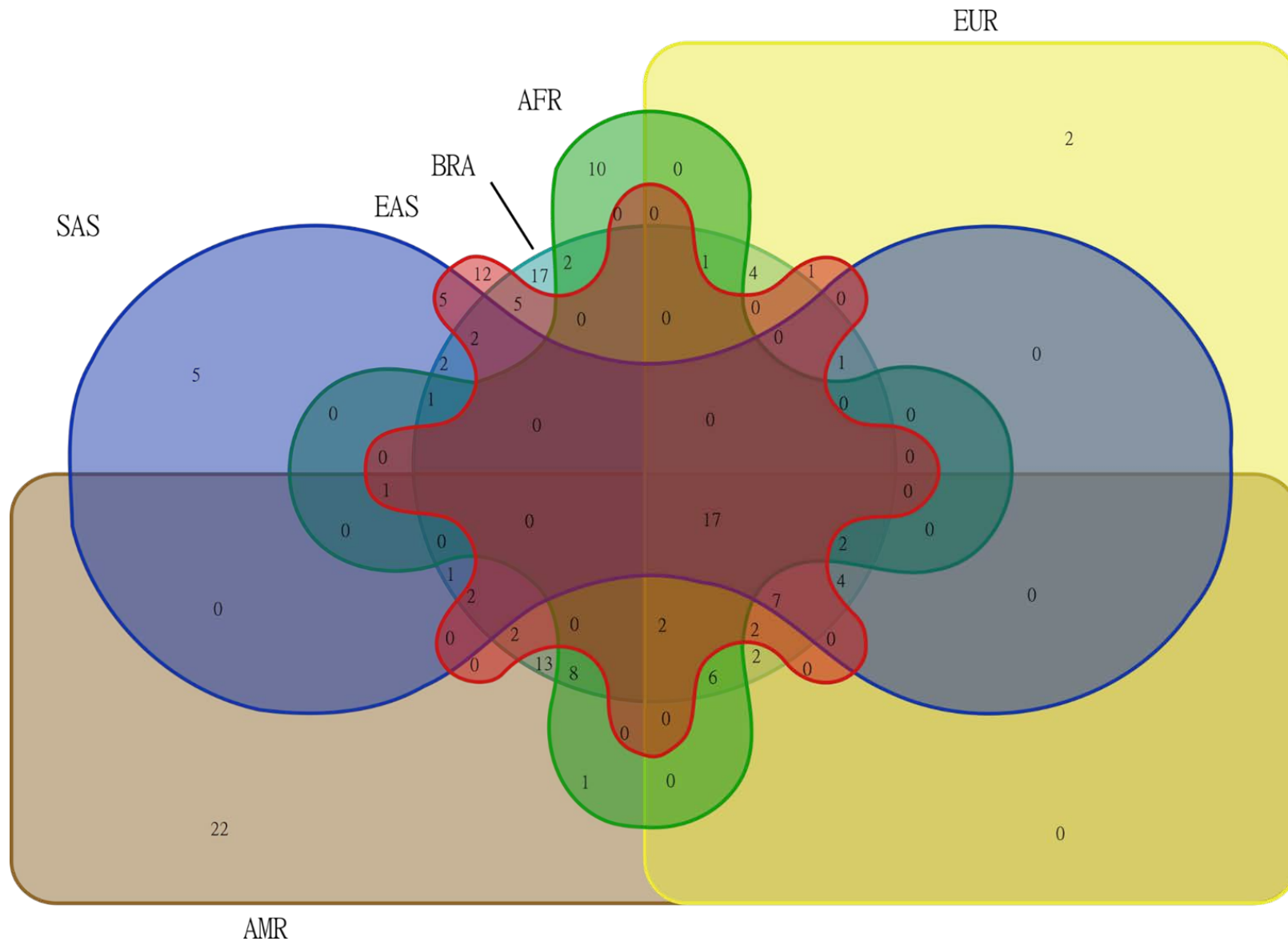


Figure S2. Venn Diagram. Allotypes sharing among super populations. Intersection among populations was calculated by Bioinformatics & Evolutionary Genomics (<http://bioinformatics.psb.ugent.be/webtools/Venn/>)

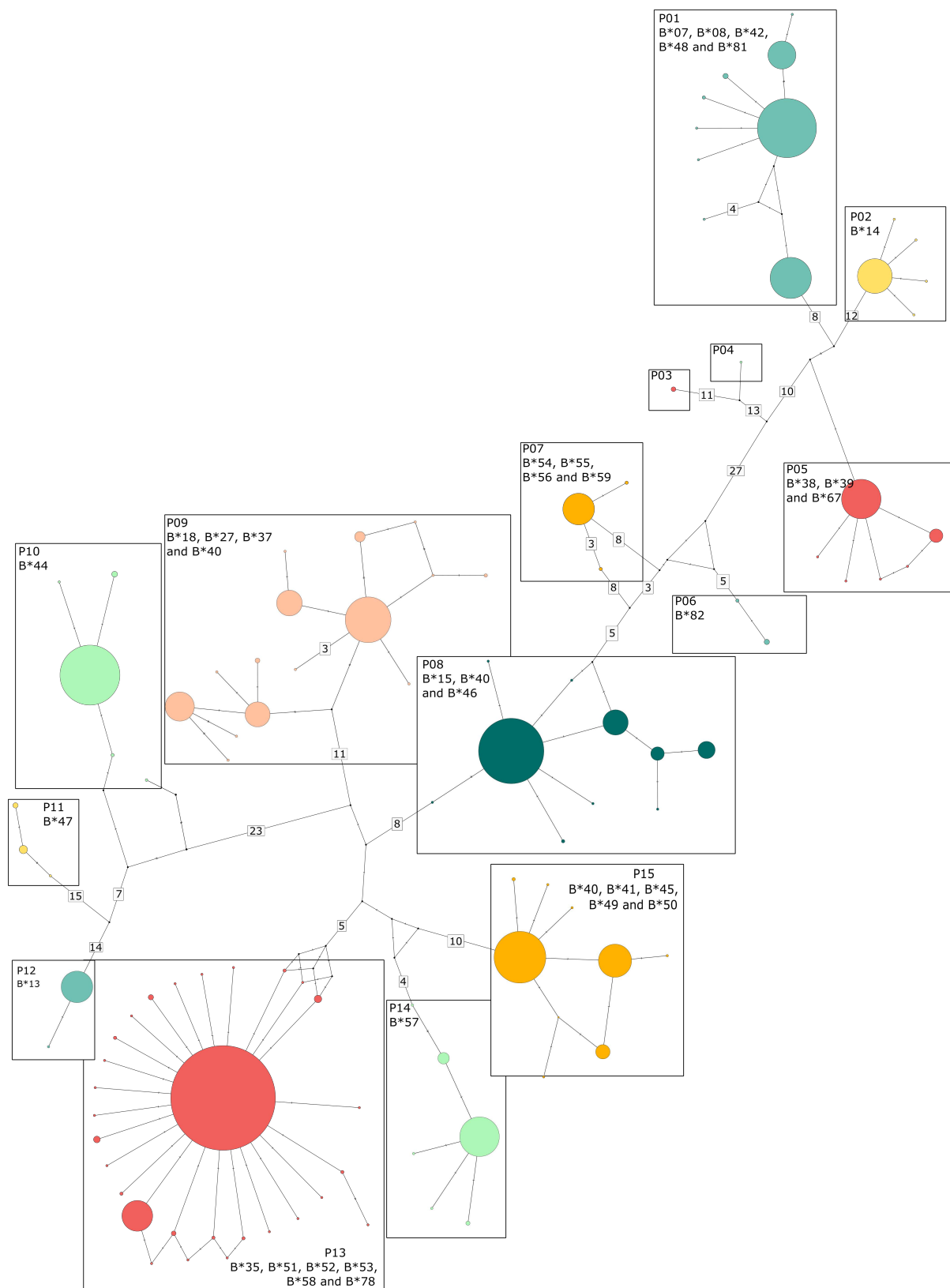


Figure S3. Network of the *HLA-B* promoter sequences (promoter sequences considering 1,850bp upstream the first translated ATG). Each lineage is represented within a square. The circles represent different sequences, and the size of circles corresponds to their frequencies. The numbers near the branches represent mutational steps separating sequences or lineages.

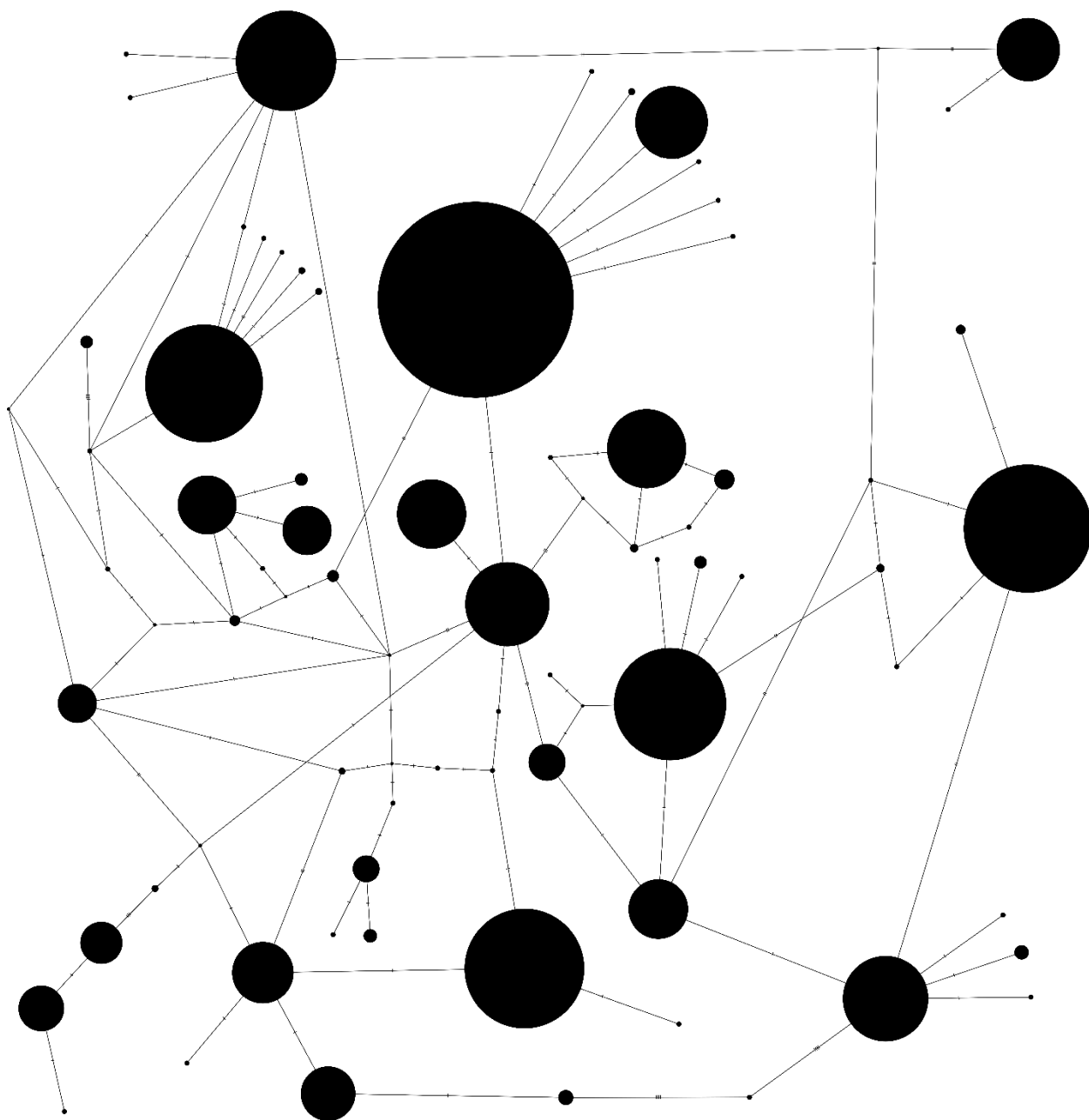


Figure S4. Network of the *HLA-B* 3'UTR sequences. There are 73 different 3'UTR sequences, with many shared variants among them, and few mutational steps among them. The circles represent different sequences, and the size of circles corresponds to sequence frequencies.

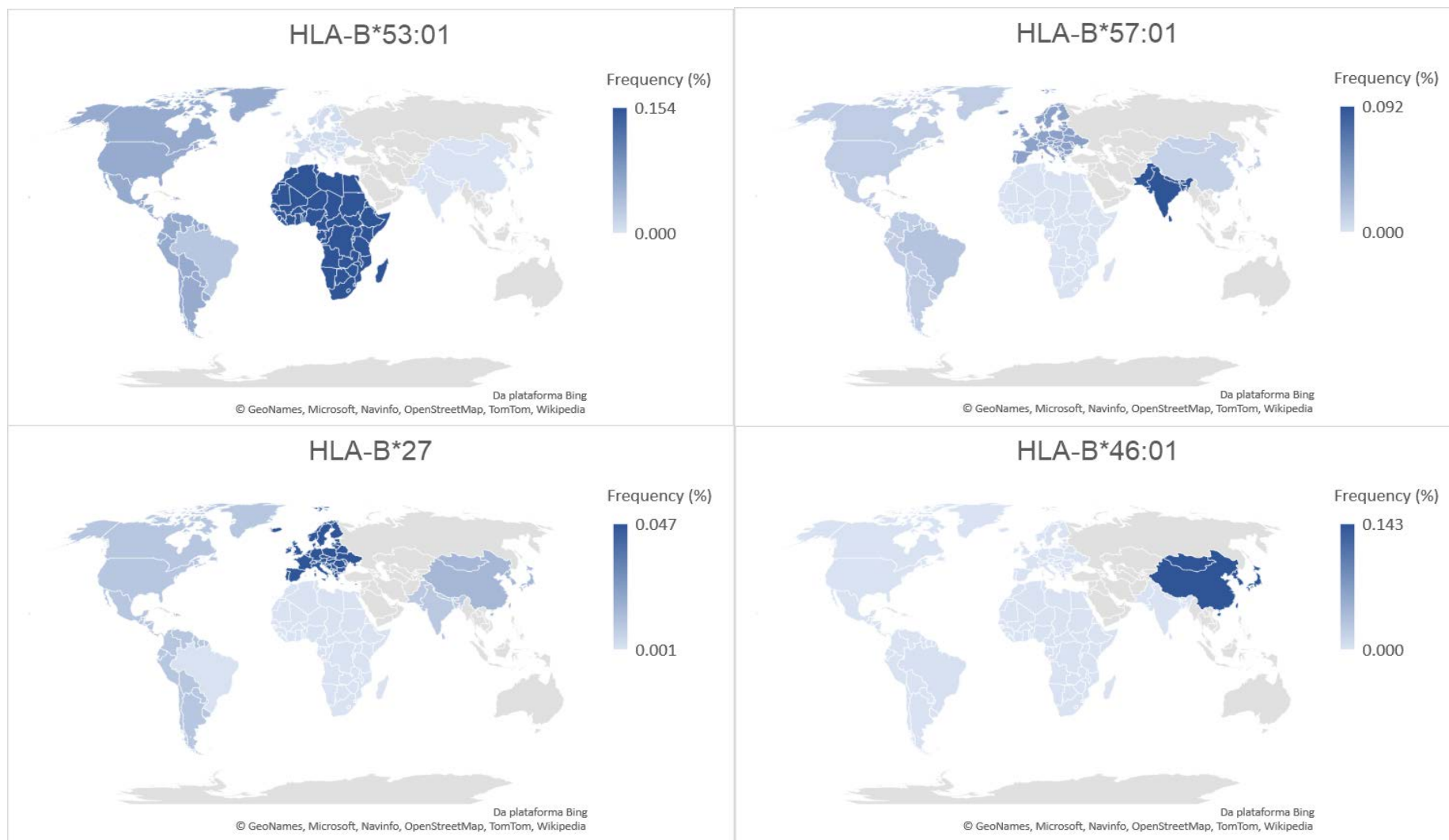


Figure S5. Distribution of HLA-B*27, HLA-B*53:01, HLA-B*46:01 and HLA-B*57:01 worldwide. HLA-B*27 represents the HLA-B*27:02, HLA-B*27:04 and HLA*27:05 alleles.

CONCLUSÃO

Aqui, apresentamos o primeiro estudo avaliando a diversidade genética do gene *HLA-B* em todos os éxons, íntrons, e regiões regulatórias, em amostras do mundo todo. A metodologia proposta nesse trabalho se mostrou eficaz para obtenção de genótipos e haplótipos de *HLA-B* a partir de sequenciamento de nova geração. Os resultados desse estudo mostraram que:

- ✓ A amostra brasileira introduziu novos alelos e também aumentou a frequência de alelos raros que estavam presentes no 1000genomes;
- ✓ A diversidade haplótípica do gene *HLA-B* é maior em populações miscigenadas, particularmente em brasileiros;
- ✓ O promotor de *HLA-B*, éxon 1 e a região 3'UTR são altamente variáveis, assim como os éxons 2 e 3;
- ✓ A maioria dos SNPs são compartilhados por todas as populações, enquanto o oposto é observado para os haplótipos e alótipos;
- ✓ Cada grupo alélico está relacionado com um grupo promotor específico;
- ✓ Todas as populações compartilham as mesmas sequências regulatórias;
- ✓ Como esperado, alelos clinicamente relevantes apresentam frequências diferentes entre as populações.