**Carga genética por efeito carona sob diferentes
regimes seletivos no genoma humano**

Hitchhiking influencing genetic load in human genomes
under different selective regimes

Luiz Carlos Machado de Oliveira Junior

UNIVERSIDADE ESTADUAL DE SÃO PAULO

INSTITUTO DE BIOCÊNCIAS

PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA E BIOLOGIA EVOLUTIVA

São Paulo

2018

Luiz Carlos Machado de Oliveira Junior

**Carga genética por efeito carona sob diferentes
regimes seletivos no genoma humano**

Hitchhiking influencing genetic load in human genomes
under different selective regimes

Dissertação de mestrado apresentada ao
Programa de Pós-graduação do Instituto de
Biociências da Universidade de São Paulo
como parte dos requisitos para a obtenção
do título de Mestre em Ciência Biológicas.

Área de concentração: Ciência Biológicas

Orientador: Diogo Meyer

São Paulo

2018

Machado, Luiz Carlos

Carga genética por efeito carona sob diferentes regimes seletivos no genoma humano.

79 páginas

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

1. Carga Genética 2. Reamostragem 3. Efeito Carona 4. SNPs deletérios I. Universidade de São Paulo. Instituto de Biociências. Departamento de Genética e Biologia Evolutiva.

Comissão Julgadora

Prof(a). Dr(a).

Prof(a). Dr(a).

Prof(a). Dr(a).

Prof. Dr. Diogo Meyer

Este trabalho é dedicado aos meus queridos pais, Fátima e Luiz.

Agradecimentos

Queria agradecer a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes) pelo apoio financeiro que possibilitou este trabalho.

Queria agradecer imensamente ao Diogo Meyer pela atenção, incentivo, paciência, compreensão, e lógico, por todas as direções que ele me deu ao longo de todo mestrado.

Aos professores Tatiana Torres, Cristina Miyaki e Sergio Matioli pelas contribuições dadas durante a banca de qualificação deste trabalho.

A todo pessoal do laboratório: Kelly, Vitor, Jonatas, Bárbara e Débora, que sempre me ajudaram nas dúvidas sobre evolução e R. Também pelas polêmicas tardes que passamos juntos discutindo assuntos cotidianos.

Um obrigado especial à Bárbara, por estar sempre disposta a tirar minhas dúvidas, mas sobretudo, pelas discussões sobre música e sobre a vida.

Ao Vitor, por ser esse poço de bom senso e generosamente compartilha-lo com todos.

Ao Jonatas, pela fundamental ajuda com os códigos.

Agradeço a todos do Porão pelas conversas de corredor sobre todos os assuntos possíveis, regados ao melhor café da Biociências.

Agradeço a todos os meus amigos da biologia, em especial à Síndrome por todas as

noites que passamos discutindo evolução e que sempre, sempre se empolgou com este trabalho. Ao Bitch por estar sempre presente para me lembrar as coisas realmente importantes da vida. Ao Meg por todos os perrengues, reclamações e alegrias que passamos juntos. Ao Max por ser o amigo que eu mais admiro e tenho como referência pra minha vida. Também gostaria de agradecer ao não biólogo quase biólogo Daniel, pelas cervejas, risadas e torcidas no futebol. Todos vocês têm um pedaço neste trabalho.

Ao pessoal do samba do matão, do fute campo ICBIO-JnP e todos os Nuzofers, pelos momentos de distrações que me proporcionaram.

À toda minha família que sempre apoiou minhas decisões e sempre esteve perto para me ajudar nos momentos difíceis. Meus pais, Fátima e Luiz, que são as pessoas que mais admiro na vida, por me educarem da forma que me educaram, com liberdade, amor e carinho. Meus irmãos, Beta e Henrique, que sempre me serviram de exemplo, tenho muito de vocês em tudo que eu faço.

À Raquel e ao Pedro, por serem minha segunda família.

À Gabi, minha companheira e melhor amiga, que diariamente conviveu com todos meus problemas e angustias. Que gastou horas discutindo questões evolutivas, computacionais e burocráticas, que foram fundamentais para este trabalho. E pessoa com quem mais gosto de gastar minhas horas de descanso.

Sumário

Lista de ilustrações	1
Lista de tabelas	4
RESUMO	6
<i>ABSTRACT</i>	8
INTRODUÇÃO GERAL	10
Regimes Seletivos	10
História Demográfica Humana	12
Carga genética	13
Estrutura da dissertação	15
1 METODOLOGIA DE REAMOSTRAGEM PARA COMPARAR CARGA GENÉTICA EM POPULAÇÕES HUMANAS	16
1.1 Contextualização	16

1.2	Métodos tradicional para comparar carga genética intrapopulacional	17
1.3	Metodologia proposta	18
1.4	Avaliando a metodologia proposta	20
1.5	Quando utilizar cada critério de reamostragem	24
1.6	Conclusão	25
2	CARGA GENÉTICA POR EFEITO DE CARONA NA ADJACÊNCIA DE SÍTIOS SOB SELEÇÃO POSITIVA	26
	Resumo	26
	Abstract	27
2.1	Introdução	27
2.2	Dados e Métodos	30
2.2.1	Banco de dados	30
2.2.2	Identificando seleção positiva	31
2.2.3	Identificando sinais de seleção positiva em escala continental	33
2.2.4	Diversidade ao redor de sítios sob seleção positiva	34
2.2.5	Definição da janelas de estudo	34
2.2.6	Quantificação de carga genética	35
2.2.7	Reamostragem em janelas sob seleção positiva	36
2.2.8	Análises Interpopulacionais	36

2.3	Resultados	37
2.4	Discussão	42
2.4.1	Maior carga genética em regiões em que há efeito carona	43
2.4.2	Carga genética somente nos sítios sob seleção	43
2.4.3	Padrão de SFS encontrado	45
2.4.4	Escolha conservadora dos genes sob seleção	45
2.4.5	Contribuições do uso de reamostragem	46
2.5	Conclusão	46
3	CARGA GENÉTICA EM GENES DO HLA POR EFEITO CARONA DE SÍTIOS SOB SELEÇÃO BALANCEADORA .	48
	Resumo	48
	Abstract	49
3.1	Introdução	49
3.2	Dados e Métodos	52
3.2.1	Banco de dados	52
3.2.2	Identificando os sitios não-ARS	53
3.2.3	Enriquecimento de polimorfismo nos sítios não-ARS	53
3.2.4	Quantificação de carga genética	54
3.2.5	Reamostragem em janelas sob seleção balanceadora	55

3.3	Resultados	56
3.4	Discussão	59
3.4.1	Enriquecimento de SNPs por efeito carona	59
3.4.2	Maior carga genética por efeito carona	60
3.4.3	Genes sob seleção balanceadora	62
3.4.4	Conclusão	63
	CONSIDERAÇÕES FINAIS	64
	Diferente desequilíbrio de ligação entre regiões genômicas	64
	Diferentes assinaturas de seleção e diferentes reamostragens	65
	REFERÊNCIAS	66
	Material Suplementar do capítulo 2	73
	Material Suplementar do capítulo 3	79

Lista de ilustrações

- Figura 1.1 – *Viés da distribuição de p valores para cada estatística avaliada (serão explicadas nos capítulos 2 e 3) e métodos de reamostragens (SFS-Gene-Boot, STD-Gene-Boot, SFS-Boot e STD-Boot).* 21
- Figura 1.2 – *Distribuições de SNPs deletérios separadas (a) pelos quatro métodos de reamostragem: STD-Boot, SFS-Boot, STD-Gene-Boot e SFS-Gene-Boot e (b) por classes de reamostragens que controlam para SNPs e genes (ClasseGene: STD-Gene-Boot e SFS-Gene-Boot) ou apenas para SNP (ClasseSNP: STD-Boot, SFS-Boot) pelo número aproximado de genes.* 23
- Figura 2.1 – *Baixa diversidade (medida pela heterozigosidade média dos SNPs na região) na região ligada aos SNPs HighD formada com janelas cumulativas em (a) leste asiáticos (EAS) e (b) europeus (EUR). Os retângulos preenchidos indicam o tamanho da janela para a qual há uma diversidade significamente menor do que a média genômica, conforme teste feito utilizando a reamostragem STD-Gene-Boot.* 39
- Figura 2.2 – *Espectro de Frequência Alélica (SFS) dos grupos positivo e controle, para europeus e leste asiáticos. SFS de todos os SNPs codificantes para o grupo positivo e controle em (a) europeus (EUR) e (b) leste asiáticos (EAS); SFS separando por grupos de SNPs codificantes (não sinônimo, N; sinônimos, S) para o grupo positivo e controle para (c) europeus (EUR) e (d) leste asiáticos (EAS).* 40

- Figura 2.3 – *Análise de outlier para as estatísticas P_{del} e P_{del}/P_{neu} . Os violin plots indicam as regiões sob seleção positiva para (a) leste asiáticos (EAS) e (b) europeus (EUR), utilizando o genoma de grupo continental como controle. Os valores acima de cada violin plot são os p-valores da respectiva análise de outlier. A cor verde indica os casos em que há significância seguindo o critério de valor crítico de 0,05, e vermelho indica não significância, pelo mesmo critério.* 41
- Figura 2.4 – *Diferença entre a carga genética (estatística P_{del}/P_{neu}) ao redor dos SNPs HighD e sua média genômica, para diferentes distâncias em relação ao SNP HighD. As linhas indicam os diferentes grupos continentais utilizando os mesmos genes que estão na adjacência de SNPs HighD em (a) leste asiáticos (EAS) e (b) europeus (EUR).* 42
- Figura 3.1 – *Maior número de SNPs da região não-ARS do gene HLA se comparada com uma distribuição do número de SNPs de três genes sorteados do genoma, para todos os grupos continentais: europeu (EUR), leste asiático (EAS), sul asiático (SAS), africano (AFR) e todos amalgamados (ALL). O grupo continental EAS apresenta o menor enriquecimento de SNPs, apresentando contagem de SNPs superior a 99,3% dos sorteios.* 57
- Figura 3.2 – *Espectro de Frequência Alélica (SFS) de SNPs codificantes dos grupos não-ARS e controle para os continentes africano (AFR), europeu (EUR), leste asiático (EAS), sul asiático (SAS) e todos amalgamados (ALL).* 58
- Figura 3.3 – *Análise de outlier para as estatísticas P_{del} e P_{del}/P_{neu} para o grupo de SNPs não-ARS de genes HLA clássico de classe I não-ARS para o continente africano (AFR), europeu (EUR) e asiático (EAS e SAS) e com as populações amalgamadas (ALL), utilizando o genoma de cada grupo como controle. Os valores acima de cada violin plot são os p-valores da respectiva análise de outlier. A cor verde indica os casos em que há significância seguindo o critério de valor crítico de 0,05, e vermelho indica não significância, pelo mesmo critério.* 59

Figura S2.1—Representação esquemática dos critérios de inclusão de genes na análise. A cor verde indica genes incluídos na análises: (1) genes que continham SNPs HighD e (2) genes sem SNPs HighD, mas que começam a no máximo uma distância A e terminam no máximo a uma distância B do SNP HighD. A cor vermelha indica genes que foram excluídos da análise: Genes sem SNPs HighD que começam a uma distância menor que A , mas terminam a uma distância maior que B . A e B são distâncias em kb partindo do SNPs HighD e variam nos continentes testados, sendo $A=125$ e $B=135$ nos europeus e $A=225$ e $B=235$ nos leste asiáticos. Sendo A , a janela definida pela heterozigosidade significativamente menor do que a média genômica. 73

Lista de tabelas

Tabela 1.1 – Quantificação de possíveis vieses apresentado pelos quatro métodos de reamostragem.	25
Tabela 2.1 – Conjunto de genes utilizados sob seleção positiva (i.e.compresença de SNPs HighD) em europeus (EUR) e leste asiáticos (EAS). Em negrito o único gene presente nos dois grupos continentais.	38
Tabela 3.1 – Número de SNPs nos genes HLA clássicos de classe I, com e sem os éxons ARS nas populações estudadas.	54
Tabela 3.2 – Sumário das estatísticas utilizadas no capítulo 3 separado por genes do HLA não-ARS.	56
Tabela S2.1–Sumário das estatísticas do capítulo 2 separado por genes do grupo positivo para o grupo continental europeu (EUR).	74
Tabela S2.1–Continuação do sumário das estatísticas do capítulo 2 separado por genes do grupo positivo para o grupo continental europeu (EUR).	75
Tabela S2.2–Sumário das estatísticas do capítulo 2 separado por genes do grupo positivo para o grupo continental leste asiático (EAS).	76
Tabela S2.2–Continuação do sumário das estatísticas do capítulo 2 separado por genes do grupo positivo para o grupo continental leste asiático (EAS).	77
Tabela S2.2–Continuação do sumário das estatísticas do capítulo 2 separado por genes do grupo positivo para o grupo continental leste asiático (EAS).	78

Tabela S3.1—Tabela com a comparação dos valores da estatística $P_{\text{del}}/P_{\text{neu}}$ encontrados neste trabalho para a região não-ARS do HLA e encontrados por César e colaboradores (em preparação) para os genes vizinhos do HLA para os diversos grupos continentais: europeu (EUR), leste asiático (EAS), sul asiático (SAS), africano (AFR) e todas as populações em conjunto (ALL).	79
---	----

Resumo

É o processo de mutação que, em última instância, introduz diversidade nas populações. Quando a mutação é pontual e segrega na população, recebe o nome de polimorfismo de nucleotídeo único (ou SNP, de *Single Nucleotide Polymorphism*). Caso esse SNP seja vantajoso, ele aumenta de frequência na população e estabelece um forte desequilíbrio de ligação com seus SNPs vizinhos, consequência do processo chamado de efeito de carona genética. Esse efeito pode ter duas consequências na proximidade dos sítios selecionados: (1) pode levar ao aumento de frequência de mutações neutras ligadas a mutação vantajosa ou (2) pode reduzir o tamanho efetivo da população (N_e) nessas regiões, e, consequentemente aumentar a importância da deriva genética na variação genética dessas regiões. Como consequência, é comum que a região próxima a sítios vantajosos apresente acúmulo de mutações neutras com frequências acima do esperado sob neutralidade. No entanto, a maior parte de novas mutações são deletérias e seria esperado que fossem removidas da população por seleção purificadora. Entretanto, como os padrões de variabilidade observados no genoma resultam de processos seletivos atuando em conjunto com a deriva genética, não é incomum SNPs fracamente deletérios atingirem frequências elevadas nos diferentes contextos de história demográfica e seletiva das populações. O acúmulo desses SNPs fracamente deletérios no genoma reduz progressivamente a aptidão média da população, o que representa um aumento de sua carga genética. Neste estudo investigamos se na vizinhança de sítios vantajosos selecionados há um acúmulo de SNPs deletérios, contribuindo para a carga genética da população. Para responder essa questão, dividimos nosso estudo em três passos. Primeiro, desenvolvemos uma nova metodologia para

estimar a diferença na carga genética entre região vizinha a sítios selecionados e o resto do genoma. Nossa metodologia é mais robusta do que a tradicionalmente utilizada pois garante que a região influenciada pela seleção natural e o controle apresentem o mesmo número de SNPs, e possibilita controlar para variáveis confundidoras, como a frequência dos polimorfismos, permitindo explorar regiões alvo com diferentes características e reduzir resultados espúrios. Em seguida, utilizamos nossa metodologia para testar se a carga genética na região vizinha a sítios sob seleção positiva é maior do que no restante do genoma. Identificamos um acúmulo de SNPs deletérios em europeus e leste asiáticos para amostras do projeto 1000 genomas. Mostramos também que as mesmas regiões exploradas em populações sob efeito de seleção positiva não apresentaram carga genética aumentada em populações nas quais essas regiões não experimentaram seleção positiva. Por último, utilizamos essa mesma metodologia para avaliar a carga genética em genes sob seleção balanceadora, os genes clássicos do HLA de classe I. Esses genes compõe uma das famílias gênicas mais estudadas em humanos e para os quais existem evidências indicando os sítios específicos que estão sob seleção balanceadora. Nosso teste mostrou que as regiões vizinhas aos sítios sob seleção balanceadora dentro dos genes do HLA apresentam acúmulo de SNPs deletérios para pelo menos dois agrupamentos populacionais. Assim, com os três passos aqui descritos, conseguimos apontar o aumento da carga genética em regiões próximas a sítios sob dois regimes de seleção distintos, utilizando uma metodologia capaz de considerar características específicas das regiões estudadas.

Palavras-chave: Carga genética; Reamostragem; Efeito Carona; SNPs deletérios

Abstract

It is the mutation process that ultimately introduces diversity into populations. When a mutation is punctual and segregates in the population, it is called a single nucleotide polymorphism (SNP). If this mutation is advantageous, it increases in frequency in the population and creates strong linkage disequilibrium with its neighboring *loci*, as a consequence of the hitchhiking effect. This effect can have two consequences in the vicinity of the selected sites: (1) it can increase the frequency of neutral mutations linked to this advantageous mutation or (2) can reduce the effective population size (N_e) in the region around the selected site, and consequently increase the importance of genetic drift in the genetic variation of these regions. As a consequence, it is common that the region around sites under advantageous selection to accumulate more neutral mutations at high frequencies than expected under neutrality. However, most new mutation are deleterious and it is expected that these mutation will be removed from the population by purifying selection. However, since the patterns of variability observed in the genome result from selective processes interacting with genetic drift, it is common for weakly deleterious SNPs to reach high frequencies in different population demographic histories. The accumulation of these weakly deleterious SNPs in the genome progressively reduces the population's fitness, increasing their genetic load. In this study we investigated whether in the neighborhood of selected sites there is an accumulation of deleterious SNPs due to hitchhiking, contributing to the genetic load of the population. To answer this question, we divided our study into three steps. First, we developed a new methodology to estimate the genetic load difference between the region around selected sites and the rest of the genome.

Our methodology is more robust than those traditionally used one because it ensures that both the region influenced by natural selection and control have the same number of SNPs, reducing the probability of spurious results, and is flexible to different genome peculiarities, making it possible to explore target regions with different characteristics. We then used our methodology to test whether genetic load in the region around sites under positive selection is greater than in the rest of the genome. We found consistent evidence of an accumulation of deleterious SNPs in European and East Asian continental groups. We also showed that these same region do not present the same deleterious SNP accumulation in populations in which these region did not experience positive selection. Finally, we used this same methodology to evaluate the genetic load in genes under the balancing selection, the classic HLA class I genes. These genes are one of the most studied in humans and there is evidence that indicates the specific sites within these genes that are under balancing selection. We found that the regions around these sites under balancing selection within the HLA genes had deleterious SNPs accumulating for at least two continental groups. Thus, with the three steps described here, we were able to point out the increase of genetic load in regions around selected sites for two distinct selective regimes using a methodology that takes into account specific characteristics of the studied regions.

Key words: Genetic load; Bootstrap; Hitchhiking effect; deleterious SNPs

Introdução Geral

As populações humanas estão hoje mundialmente distribuídas, porém esse é um padrão recente de nossa história evolutiva. A migração para fora da África aconteceu há cerca de 150 mil anos e atingiu todos os continentes habitados há cerca de 15 mil anos. Dessa forma, nossos ancestrais foram expostos a uma grande variedade de ambientes, deparando-se com diferentes temperaturas, altitudes, níveis de radiação ultravioleta, alimentos e patógenos (SCHEINFELDT; TISHKOFF, 2013; COLONNA et al., 2014). Ao longo desse período, também passaram por uma dinâmica demográfica complexa, caracterizada por migrações e expansões populacionais. Esses diferentes contextos da história humana moldam complexos padrões de diversidade genética nas populações modernas com muitas camadas de eventos demográficos e evolutivos agindo em diferentes escalas de tempo (BALARESQUE; BALLEREAU; JOBLING, 2007; SCHRAIBER; AKEY, 2015). Compreender os processos responsáveis por esses padrões é uma questão em genética evolutiva.

Regimes Seletivos

Em última instância, é o processo de mutação que introduz diversidade nas populações (SABETI et al., 2006). Quando novas variantes aparecem através de mutação pontual e segregam na população, recebem o nome de polimorfismo de nucleotídeo único (ou SNP, ou *Single Nucleotide Polymorphisms*) e podem ser classificados de acordo com o impacto que causam ao fenótipo do indivíduo. Denomina-se SNPs sinônimo aqueles presentes em regiões codificadoras que não resultam em trocas de aminoácidos. De maneira geral,

aceita-se que a frequência desses SNP oscila na população apenas por ação da deriva genética, e sua contagem pode ser expressa pelo número total de sítios dessa categoria (PS). Por outro lado, os SNPs que resultam na troca de aminoácidos são denominados não sinônimos, e podem segregar por períodos que dependem do balanço entre deriva genética e intensidade da seleção natural, sendo sua contagem expressa pelo número total de sítios dessa categoria numa amostra populacional (PN, HENN et al., 2015).

Os diversos campos da biologia evolutiva utilizam distintas terminologias para descrever formas de seleção natural (NIELSEN, 2005). Para responder as perguntas em evolução molecular é necessário descrever como são classificados os regimes seletivos de acordo com os geneticistas de população. De maneira geral, a genética de populações distingue a seleção natural como sendo positiva, balanceadora ou negativa. Um novo SNP pode aumentar a aptidão do indivíduo, e por essa razão, aumentar de frequência na população ao longo das gerações. Esse processo é denominado seleção positiva. Os trabalhos que buscam identificar a ocorrência desse processo são em grande parte guiados pelo modelo de varredura seletiva (HERNANDEZ et al., 2011), que prevê a redução da diversidade genética na região ao redor do sítio selecionado (SABETI et al., 2006; AKEY, 2009; HERNANDEZ et al., 2011). O capítulo 2 deste trabalho discutirá com mais detalhes os padrões deixados no genoma pela assinatura de seleção positiva. Outro regime de seleção natural que aumenta a aptidão do indivíduo que carrega o sítio vantajoso é a seleção balanceadora. Ao contrário da seleção positiva, a seleção balanceadora engloba os regimes que resultam na manutenção da diversidade genética nas populações em nível maior do que esperado sob neutralidade. Pode-se observar seleção balanceadora em ambientes onde o genótipo com maior aptidão é o heterozigoto, mantendo os dois alelos na população; em ambientes onde o alelo menos frequente é vantajoso; ou quando há pressões seletivas que variam no tempo e espaço (ANDRÉS, 2011). O Capítulo 3 deste trabalho abordará as consequências da ação do regime de seleção balanceadora para a região vizinha ao sítio selecionado e discutirá características específicas e detalhadas de seleção balanceadora.

A maior parte das mutações não sinônimas afetam negativamente a aptidão do indivíduo (EYRE-WALKER; KEIGHTLEY, 2007; LOHMUELLER, 2014a). Assim, o regime

seletivo mais comum no genoma dos organismos é a seleção negativa, também chamada de seleção purificadora. Esse regime seletivo atua reduzindo a frequência das variantes deletérias na população (TENNESSEN et al., 2012; HENN et al., 2015).

Entretanto, como os padrões de variabilidade observados no genoma resultam de processos seletivos atuando em conjunto com a deriva genética, não é incomum SNPs fracamente deletérios atingirem frequências elevadas dependendo da história demográfica da população (LOHMUELLER, 2014b; DO et al., 2015; HENN et al., 2015). Ainda assim, resultados clássicos mostram que a manutenção de SNPs sinônimos em altas frequências é mais comum que SNPs não sinônimos, apontando a importância da seleção purificadora em moldar as frequências dos polimorfismos nas populações (TENNESSEN et al., 2012; HENN et al., 2015). O fato da variação genética ser moldada pelo balanço entre seleção natural e deriva genética implica que esses dois fatores precisam ser contemplados. Para tanto, na próxima seção examinamos aspectos da história demográfica humana que podem ter influenciado a variação genética através da deriva.

História Demográfica Humana

Na última década, o avanço da tecnologia através do Next Generation Sequencing (NGS) permitiu sequenciar amostras vindo de inúmeras populações ao redor do mundo e testar previsões teóricas que dominaram discussões entre evolucionistas no século XX (COVENTRY et al., 2010; MARTH et al., 2011; KEINAN; CLARK, 2012; TENNESSEN et al., 2012; FU et al., 2012). Para a história evolutiva humana, esse avanço possibilitou compreender de maneira mais precisa a expansão dos seres humanos modernos para fora do continente africano, em especial, explorar dois processos de nossa história demográfica que impactaram a distribuição de SNPs deletérios em nosso genoma (LOHMUELLER, 2014b; LOHMUELLER, 2014a; HENN et al., 2015; KOCH; NOVEMBRE, 2017). O primeiro processo é a passagem por um, ou mais, gargalos populacionais (VOIGHT et al., 2006; GAZAVE et al., 2013), levando a uma redução do tamanho efetivo populacional (N_e) das populações que se estabeleceram fora da África (MARTH et al., 2004; LAVAL et al., 2010; NELSON et al., 2012; KEINAN; CLARK, 2012). É esperado que essa redução do N_e , por

representar um aumento na influência da deriva genética, explique a manutenção de SNPs deletérios que esperaríamos que fossem removidos pela seleção purificadora (TENNESSEN et al., 2012; SCHRAIBER; AKEY, 2015). O segundo processo é a expansão populacional que ocorreu nos últimos 5000 anos em todo o mundo, em especial nas populações não-africanas (COVENTRY et al., 2010; TENNESSEN et al., 2012; GAZAVE et al., 2013; LOHMUELLER, 2014b). Com maior número de indivíduos nas populações, há também o maior influxo de novas mutações (GAZAVE et al., 2013). O resultado é uma distribuição das frequências alélicas (conhecida como espectro de frequência alélica ou SFS, de *Site Frequency Spectrum*) com a maior parte de seus alelos ou, neste caso, SNPs concentrados em baixas frequências (i.e. há muitas mutações restritas a poucos indivíduos da amostra). Grande parte desses SNPs são fracamente deletérios e, por serem extremamente recentes, ainda não passaram pelo crivo seleção purificadora (TENNESSEN et al., 2012; GAZAVE et al., 2013).

Carga genética

O acúmulo de SNPs fracamente deletérios no genoma reduz progressivamente a aptidão média da população em relação a uma hipotética aptidão média dessa mesma população sem essas mutações. A diferença entre essas aptidões é classicamente definida como carga genética (HALDANE, 1957; KIMURA; MARUYAMA; CROW, 1963; LOHMUELLER, 2014a). No entanto, o cálculo da carga genética em populações humanas seguindo sua definição clássica não é uma tarefa trivial (LOHMUELLER, 2014a; HENN et al., 2015; KOCH; NOVEMBRE, 2017). Diversas métricas foram desenvolvidas para calcular a carga genética e permitir compreender os processos responsáveis pela sua origem dentro de populações, assim como, comparar sua magnitude em populações humanas diferentes (GRAVEL, 2016; DO et al., 2015; HENN et al., 2015). Pode-se, por exemplo, estimar a carga genética como o número médio de alelos deletérios por indivíduo na população (SIMONS et al., 2014; DO et al., 2015). Outro método é calcular a frequência média de alelos deletérios segregando na população (NELSON et al., 2012; GIBSON, 2012). Nos capítulos que se seguem adotamos um terceiro método baseado no cálculo da razão entre

os números de SNPs deletérios e SNPs neutros ($P_{\text{del}}/P_{\text{neu}}$) (CHUN; FAY, 2011; LENZ et al., 2016; ZHANG et al., 2016). Esta abordagem foi inicialmente proposta como a divisão simples entre o número de SNPs não sinônimos (PN) e sinônimo (PS), o PN/PS, assumindo que a probabilidade de um SNP não sinônimo reduzir a aptidão do indivíduo é maior do que chance de ser vantajoso ou neutro. No entanto, nem todo SNP não sinônimo é deletério e recentemente algoritmos para estimar o efeito dos SNPs não sinônimos sobre aptidão do indivíduo têm sido desenvolvidos buscando tornar essas proporções de $P_{\text{del}}/P_{\text{neu}}$ mais precisas. Embora os processos demográficos citados anteriormente afetem todas as abordagens sobre carga genética, estudos intrapopulacionais (i.e. aqueles que comparam variações de carga genética dentro do genoma de uma mesma população) e estudos interpopulacionais (i.e. os que comparam a carga genética de diferentes populações) tratam os efeitos demográficos de formas distintas. De maneira geral, investigando populações com diferentes histórias demográficas, os estudos interpopulacionais esbarram na dificuldade de entender o que os modelos demográficos e regimes seletivos prevêm sobre a carga genética (KOCH; NOVEMBRE, 2017). Já em estudos intrapopulacionais, a história demográfica é compartilhada entre todos os SNPs do genoma e, portanto, as diferenças encontradas (e.g. acúmulo de SNPs deletérios) entre as regiões estudadas e o resto do genoma podem ser atribuídas a processos de seleção natural que atuaram naquelas regiões (CHUN; FAY, 2011; LENZ et al., 2016; ZHANG et al., 2016). Nos dois últimos capítulos desta dissertação iremos explorar distribuição de carga genética em um contexto intrapopulacional sob o efeito de distintos regimes de seleção vantajosa. No próximo capítulo apresentaremos uma metodologia desenvolvida para testar essa distribuição desigual dentro de um mesmo genoma.

Estrutura da dissertação

A presente dissertação está organizada em três capítulos escritos em português. No primeiro, intitulado "*Metodologia de reamostragem para comparar carga genética em populações humanas*", apresentamos a metodologia que permite a comparação da carga genética entre diferentes regiões de um mesmo genoma, que será utilizado nos capítulos subsequentes. Consideramos em quais condições cada tipo particular da metodologia de reamostragem deve ser utilizada, assim como, as melhorias que essa metodologia traz para a discussão de carga genética. Por se tratar de um capítulo metodológico, ele não tem como objetivo testar uma hipótese, e sim descrever o método e facilitar a compreensão dos capítulos seguintes.

No segundo capítulo, redigido em formato de artigo e intitulado "*Carga genética por efeito de carona na adjacência de sítios sob seleção positiva*", testamos a hipótese de que regiões adjacentes a SNPs com assinaturas de seleção positiva apresentam enriquecimento para SNPs deletérios. Investigamos a manutenção desses SNPs deletérios em regiões sob seleção positiva em populações europeias e asiáticas. Discutimos que há evidência de acúmulo de SNPs deletérios apenas nas populações sob seleção positiva, apresentando o comportamento dessas mesmas regiões em populações onde essas regiões não apresentam assinatura de seleção.

No terceiro capítulo, também redigido em formato de artigo e intitulado "*Carga genética em genes do HLA por efeito carona de sítios sob seleção balanceadora*", exploramos o acúmulo de SNPs deletérios em éxons não-ARS de genes de HLA-Clássicos. Os genes HLA são exemplos clássicos de genes sob Seleção balanceadora e são um dos poucos genes para os quais há evidências apontando os éxons selecionados. Excluímos os éxons sob seleção balanceadora (ARS) e testamos acúmulo de SNPs deletérios nos éxons adjacentes. Os resultados indicam que a seleção balanceadora nos éxons ARS é responsável pela manutenção de SNPs deletérios nos éxons adjacentes.

Metodologia de reamostragem para comparar carga genética em populações humanas

1.1 Contextualização

Como visto na introdução geral, a diversidade genética é moldada pelos processos evolutivos e demográficos que atuam sobre as mutações que surgem ao acaso nas populações. Um SNP ou genótipo vantajoso aumentando de frequência na população por efeito de seleção natural quando em desequilíbrio de ligação com seus SNPs vizinhos, irá gerar um efeito de carona genética. Esse efeito pode ter duas consequências na proximidade dos sítios selecionados. Primeiro, pode levar ao aumento de frequência de mutações neutras ligadas à mutação vantajosa (KIMURA; MARUYAMA; CROW, 1963). Segundo, pode reduzir o tamanho efetivo da população (N_e) nessas regiões e, consequentemente, aumentar a importância da deriva genética como força evolutiva que molda a variação genética dessas regiões (CHARLESWORTH, 2013). Como consequência, é comum que a região próxima a sítios sob seleção vantajosa apresente acúmulo de SNPs neutros acima do esperado sob neutralidade (KIMURA; MARUYAMA; CROW, 1963). Testamos nos capítulos seguintes desta dissertação se o processo de carona genética também influencia a variação observada para SNPs deletérios, sob dois regimes distintos de seleção vantajosa. O acúmulo desses SNPs deletérios é uma forma de identificar que há um aumento da carga genética (i.e. perda de aptidão) como consequência da seleção vantajosa (HENN et al., 2015).

Para testar se há diferenças no número de SNPs deletérios em regiões próximas a genes selecionados é necessário estabelecer um teste estatístico robusto, que leve em consideração diversos fatores, como por exemplo grande diferença no número de SNPs entre as regiões testadas. Seguindo uma estratégia comum em estudos de genética evolutiva, desenvolvemos um teste no qual as expectativas em relação às quais os dados de interesse são comparados são geradas por um processo de reamostragem. Este é um capítulo metodológico com o objetivo de apresentar essa metodologia que desenvolvemos e que será utilizada nos capítulos seguintes.

1.2 Métodos tradicional para comparar carga genética intrapopulacional

Para testar se existe distribuição desigual de carga genética em diferentes regiões de um genoma devido a fatores como, por exemplo, proximidade a genes selecionados ou a taxa de recombinação local, muitos estudos partem para comparações diretas, ou seja, dividem o genoma em duas categorias: a região alvo, que será testada para a hipótese alternativa de uma maior carga genética; e a controle, formado pelo resto do genoma. Essas abordagens tipicamente testam a hipótese nula de igual carga genética entre essas regiões (CHUN; FAY, 2011; ZHANG et al., 2016; LENZ et al., 2016), utilizando, por exemplo, métricas de proporções como PN/PS (PIGANEAU; EYRE-WALKER, 2009; CHARLESWORTH; CHARLESWORTH, 2010) ou $P_{\text{del}}/P_{\text{neu}}$ (CHUN; FAY, 2011; LENZ et al., 2016) para calcular a carga genética e testes estatísticos como teste exato de Fisher, teste x^2 , teste de Wilcoxon, dentre outros (BETANCOURT; PRESGRAVES, 2002; CHUN; FAY, 2011; ZHANG et al., 2016; LENZ et al., 2016). Um problema com essas abordagens é o fato das regiões alvo e as regiões controle possuírem tamanhos muito diferentes, com as regiões controle tipicamente sendo muito maiores do que as regiões alvo (BETANCOURT; PRESGRAVES, 2002; CHUN; FAY, 2011; ZHANG et al., 2016; LENZ et al., 2016). Desta forma, com a região alvo muito menor, aumenta a chance de haver proporções diferentes de SNPs deletérios simplesmente por efeitos amostrais (KOCH; NOVEMBRE, 2017).

1.3 Metodologia proposta

Para delimitar um controle com o mesmo número de SNPs que a região alvo, seguimos uma abordagem de reamostragem de SNPs vindos do genoma inteiro, sendo esses, SNPs codificantes independente de sua anotação (i.e. sinônimos e não sinônimos). De maneira geral, a reamostragem permite formar uma distribuição de qualquer estatística proveniente do genoma e comparar com o valor da mesma estatística para a região alvo, gerando um p-valor empírico que permite detectar possíveis *outliers*. Neste contexto, *outliers* referem-se às regiões que apresentam uma contagem ou proporção de polimorfismos deletérios que ocupam uma posição extrema na distribuição gerada para o controle, superando assim um limiar (por exemplo, os 5% mais extremos) que permite identificá-los como estatisticamente significantes (condicional ao valor do limiar adotado). Nessa abordagem, cada amostra que compõe a reamostragem deve apresentar o mesmo número de SNPs da região alvo. Isso garante que efeitos da diferença de tamanhos entre as regiões comparadas sejam eliminados. No entanto, cada região do genoma apresenta características específicas e controlar somente para o número de SNPs não é a maneira mais precisa para montar a reamostragem, pois estas características específicas podem gerar diferenças espúrias.

Além da diferença do número de SNPs, outro fator que deve ser considerado é o SFS (de *Site Frequency Spectrum*) da região alvo (HENN et al., 2015). Considerando-se o genoma humano inteiro para a maior parte das populações humanas, o SFS é tipicamente em formato de L e apresenta mais SNPs deletérios em baixa frequência (TENNESSEN et al., 2012; HENN et al., 2015; BITARELLO et al., 2018), pois as mutações surgem em baixa frequência e devido à seleção purificadora as variantes deletérias não atingem frequências mais elevadas, permanecendo em frequências baixas, até que venham a ser removidas (TENNESSEN et al., 2012). Consequentemente, ao sortear um SNP do genoma com tal SFS, é provável que ele seja deletério e esteja em baixa frequência. No entanto, esse padrão geral genômico não necessariamente se reflete em regiões menores do genoma, uma vez que a força de seleção não é a mesma ao longo do genoma, de maneira que ela molda a variação genética de forma diferente em regiões diferentes. Neste caso, uma região que

esteja sob fraca força de seleção terá um SFS diferente de uma região sobre forte força de seleção. Como consequência desses processos, regiões específicas podem ter SFS diferentes do genoma inteiro e, consequentemente, probabilidades de reamostrar um SNP deletério diferentes (ANDRÉS et al., 2009; MEYER et al.,). Por exemplo, se a região alvo tem um padrão de SFS inflado para alelos comuns, ela tem, também, menos SNPs deletérios (comuns em baixas frequências) (ANDRÉS et al., 2009; MEYER et al., ; BITARELLO et al., 2018). Ao realizar a reamostragem do genoma sem considerar essa diferença, o controle terá maior quantidade de SNPs deletérios do que a região alvo devido a diferença em seus SFS, o que dificulta a detecção de carga genética na região alvo.

Para que a reamostragem elimine os efeitos dessa contribuição dada pelo SFS, é importante que a região alvo e a controle tenham SFS semelhantes. Somente assim podemos testar se há efeito de seleção (i.e. carga genética) para além dos efeitos que a própria seleção provoca ao SFS. Assim, montamos dois tipos de reamostragem: 1) STD-Boot (*Standard Bootstrap*), que deve ser utilizado quando previamente se tem informação de que os grupos que serão comparados apresentam SFS semelhantes. Nesse caso os SNPs são amostrados sem nenhuma restrição ao longo das regiões codificadoras do genoma e 2) SFS-Boot (*SFS Bootstrap*), quando sabe-se que os SFS são diferentes, ou não se tem essa informação e deseja-se controlar para uma possível diferença. Nesse caso, para garantir que o conjunto de SNPs reamostrados para constituir os grupos controle tenham um SFS igual ao da região alvo, utilizamos o seguinte protocolo. Todos os SNPs do genoma são separados em 12 categorias de frequências do alelo menos frequente (MAF - *Minor Allele Frequency*), e sorteamos SNPs seguindo o número de SNPs de cada categoria de frequência da região alvo. Por exemplo, se a região alvo tem 10 SNPs com frequência entre 0 e 4% e 5 SNPs com frequência entre 4 e 8%, amostramos do genoma 10 SNPs entre 0 e 4% e 5 SNPs entre 4 e 8% a cada amostra da reamostragem.

Veremos com detalhes nos próximos capítulos desta dissertação os critérios que estabelecemos para definir as regiões alvo. Em ambos, entretanto, um elemento em comum é que a região alvo é formada por SNPs contíguos dentro de um conjunto de genes. Assim, deve haver um grande desequilíbrio de ligação entre esses SNPs (REICH et al., 2001).

Ao reamostrar do genoma controlando apenas pelo número de SNPs e/ou o SFS da região alvo, é provável que os SNPs sorteados estejam distribuídos num número maior de genes do que aqueles que compõem a região alvo, e, por consequência, apresentam um desequilíbrio de ligação menor (REICH et al., 2001). Esperamos que essa diferença no desequilíbrio de ligação entre os SNPs da região alvo e do grupo controle influencie nossas análises. Frente a esse problema, buscamos um critério de reamostragem que permita estabelecer um desequilíbrio de ligação semelhante entre os SNPs que compõem a região alvo e os SNPs amostrados que irão compôr o controle. Isso foi feito através de dois métodos de reamostragem 1) STD-Gene-Boot (*Standard Gene Bootstrap*) e 2) SFS-Gene-Boot (*SFS Gene Bootstrap*), que diferem dos dois primeiros ao controlarem também para que a região alvo e a controle apresentem o número mais parecido possível de genes seguindo as demais regras impostas para STD-Boot e SFS-Boot, respectivamente. Assim, o controle também será formado por SNPs contíguos e consequentemente, um desequilíbrio de ligação mais próximos da região alvo.

É necessário ressaltar que os critérios que estabelecem cada amostra da reamostragem seguem um conjunto de operações organizadas de forma hierárquica: para formar o controle, todas as amostras devem ter o mesmo número de SNPs que a região alvo; os SFS devem seguir as categorias da região alvo (quando esse critério está presente); e, por último, as amostras devem ter um número de genes próximo ao da região alvo (quando esse critério está presente). Então, seguindo essa hierarquia, ao usar o SFS-Gene-Boot, por exemplo, quando uma amostra alcança o mesmo número de genes que a região alvo, porém, ainda não ao mesmo número de SNPs, relaxa-se o critério de número de genes e sorteia-se mais genes até o número de SNPs necessário ser alcançado.

1.4 Avaliando a metodologia proposta

Seguindo as considerações teóricas discutidas na seção anterior, elaboramos quatro diferentes critérios de reamostragem para realizar testes estatísticos que contemplem diferenças em tamanho amostral e SFS entre a região sob estudo e os SNPs que servem de

controle: STD-Boot, SFS-Boot, STD-Gene-Boot e SFS-Gene-Boot. É importante verificar se no contexto explorado nos capítulos 2 e 3 desta dissertação (i.e. região alvo formada por SNPs com grande desequilíbrio de ligação), esses critérios de reamostragem apresentam algum viés. Para tanto, elaboramos uma simulação que segue o seguinte protocolo: amostramos 50 genes ao acaso do genoma, para os quais mediu-se cinco estatísticas que medem a carga genética (as estatísticas utilizadas serão descritas nos capítulos 2 e 3 deste trabalho). Essas foram comparadas com o controle formado utilizando os quatro critérios de reamostragem, gerando um p-valor para cada critério de reamostragem. Repetiu-se essa metodologia 1000 vezes e montou-se um histograma com os p-valores (Figura 1.1). Sob a hipótese de ausência de viés, esperaríamos uma distribuição uniforme, pois os 50 genes testados não possuem qualquer propriedade que deva levá-los a possuir uma posição extrema em relação à distribuição gerada por reamostragem.

Figura 1.1 – Viés da distribuição de p valores para cada estatística avaliada (serão explicadas nos capítulos 2 e 3) e métodos de reamostragens (SFS-Gene-Boot, STD-Gene-Boot, SFS-Boot e STD-Boot).

Assim como esperado, ao construir uma distribuição de estatística utilizando STD-Boot e SFS-Boot e compará-la com a mesma estatística em uma região específica do genoma, estamos comparando o mesmo número de SNPs, porém, em todos os casos as distribuições apresentaram, a cada reamostragem, um número maior de genes. Tanto STD-Boot como SFS-Boot, que não controlaram para o desequilíbrio de ligação, apresentam um forte viés nos p-valores (Figura 1.1)), com uma grande proporção de réplicas em que os 50 genes testados possuíam estatísticas mais extremas do que os SNPs usados para gerar a distribuição nula, evidenciando a ocorrência de falsos positivos (Tabela 1.1).

Considerando que o valor esperado para qualquer estatística é o mesmo entre os 50 genes e a distribuição nula, e que apresentar viés significa alterar as distribuições geradas pelas reamostragens, analisamos as distribuições formadas por cada critério de reamostragem para uma das estatística que será utilizada nesta dissertação (Figura 1.1) a fim de entender qual aspecto da distribuição diferia entre as reamostragem e possivelmente causava o viés.

Notamos que dessas distribuições, STD-Boot e SFS-Boot apresentam variâncias que correspondem aproximadamente a metade daquelas apresentadas por STD-Gene-Boot e SFS-Gene-Boot para a mesma estatística (Figura 1.2a). Além disso, notamos que os métodos que controlam para o desequilíbrio de ligação (STD-Gene-Boot e SFS-Gene-Boot) e os que não controlam para o desequilíbrio de ligação (STD-Boot e SFS-Boot) apresentam variâncias semelhantes entre si.

Como os padrões de variância das distribuições formaram dois grupos que coincidem com a utilização do desequilíbrio de ligação como critério para formação da reamostragem, analisamos também as variâncias dessas distribuições separando em dois grupos de acordo com o desequilíbrio de ligação dos SNPs sorteados na reamostragem. Para tanto, combinamos as quatro distribuições em somente duas (ClasseGene e ClasseSNP): as controlavam para desequilíbrio de ligação (STD-Gene-Boot e SFS-Gene-Boot) e as que controlam apenas pelo número de SNPs (STD-Boot e SFS-Boot), respectivamente. Essas classes também apresentaram diferença na variância, com a ClasseSNP tendo aproximadamente metade da variância da ClasseGene (Figura 1.2b). De fato, em uma análise de *outlier*, a diferença na variância das distribuições nula interfere na possibilidade de co-

meter erro tipo I (EFRON; TIBSHIRANI, 1994). O mesmo valor da estatística pode ser considerado como *outlier* em uma distribuição nula com menor variância ou como pertencente à distribuição caso a variância seja maior. Isso indica que a variância pode ser o aspecto da distribuição que determina se há ou não um erro que leva ao falso positivo.

Figura 1.2 – Distribuições de SNPs deletérios separadas (a) pelos quatro métodos de reamostragem: STD-Boot, SFS-Boot, STD-Gene-Boot e SFS-Gene-Boot e (b) por classes de reamostragens que controlam para SNPs e genes (ClasseGene: STD-Gene-Boot e SFS-Gene-Boot) ou apenas para SNP (ClasseSNP: STD-Boot, SFS-Boot) pelo número aproximado de genes.

Apesar de não ser nosso objetivo detalhar como o desequilíbrio de ligação afeta a variância dessas distribuições, nossos resultados mostram a influência da forma como os SNPs são sorteados sobre o teste estatístico. Claramente, amostras de SNPs aleatórios resultam em distribuições nulas mais estreitas, favorecendo resultados significativos. Isso resulta da independência maior de SNPs, que reduz a variância em relação às reamostragens de SNPs vindos do mesmo gene (REICH et al., 2001). Os dois capítulos que seguem deste trabalho exploram regiões específicas do genoma, com SNPs contíguos dentro dos genes escolhidos por estarem sob regimes de seleção específicos. Dessa forma, todas as reamostragens realizadas ao longo deste estudo controlarão para o número de SNPs e também para o números aproximados de genes, visando comparar blocos de SNPs com desequilíbrios de ligação semelhantes, reduzindo a taxa de falso positivo nos resultados.

Em nenhum dos casos o viés foi completamente eliminado (Tabela 1.1), uma vez que, seguindo os critérios já apresentados, não é incomum os blocos sorteados e regiões alvo

apresentarem pequenas distorções no número de genes (tanto em STD-Gene-Boot como SFS-Gene-Boot) e consequentemente no desequilíbrio de ligação. Ainda assim, a redução do viés foi importante com as reamostragens que controlam para o desequilíbrio de ligação. Dentre as reamostragens que controlam para o desequilíbrio de ligação (STD-Gene-Boot e SFS-Gene-Boot), a diferença de viés entre é resultado da ordem de prioridades estabelecidas ao se montar as reamostragens. O SFS-Gene-Boot, por exemplo, controla por número de SNP, SFS e gene, nessa ordem de prioridade. Assim, é mais comum que o número de genes entre grupo controle e região alvo sejam diferentes se comparados com o STD-Gene-Boot, que controla por número de SNPs e genes, apenas.

1.5 Quando utilizar cada critério de reamostragem

Os possíveis vieses causados pelos critérios de reamostragem debatido na seção anterior tem relação direta com as questões desenvolvidas nos próximos capítulos desta dissertação. Como em ambos os capítulos abordamos regiões alvo formadas por SNPs com grande desequilíbrio de ligação, implementamos uma simulação que testa o efeito da dependência (i.e. desequilíbrio de ligação) entre os SNPs sobre as taxas de falso positivo e do poder estatístico dos métodos de reamostragem propostos.

Em outros cenários, os critérios de reamostragem que não controlam para o desequilíbrio de ligação podem gerar resultados satisfatórios, sem falsos positivos. Por exemplo, se nossa intenção fosse testar se há diferença entre conjuntos de SNPs que ocorrem dispersos pelo genoma, sem desequilíbrio de ligação entre eles e outros SNPs desse mesmo genoma, também dispersos, é provável que as reamostragens STD-Boot e SFS-Boot fossem as mais adequadas, pois a região alvo e controle apresentam desequilíbrios de ligação semelhantes (REICH et al., 2001).

1.6 Conclusão

Neste capítulo apresentamos a metodologia de reamostragem desenvolvida para comparar a carga genética de diferentes regiões do genoma de forma mais robusta do que as metodologias usualmente utilizadas. Mostramos que para responder essas questões em regiões alvo formadas a partir de SNPs contíguos a reamostragem precisa controlar, não só o número de SNPs (tamanho da amostra), mas também o desequilíbrio de ligação. Mostramos também que caso regiões diferentes do genoma apresentem diferentes SFS, esse também precisam ser considerados na reamostragem. Por fim, para esta dissertação utilizamos apenas as metodologias STD-Gene-Boot e SFS-Gene-Boot por essas apresentarem baixo viés de reamostragem. A decisão sobre qual deverá ser utilizada em cada análise depende do SFS da região alvo explorada nos trabalhos apresentados em cada capítulo subsequente.

Tabela 1.1 – Quantificação de possíveis vieses apresentado pelos quatro métodos de reamostragem.

		SFS-Gene-Boot	STD-Gene-Boot	SFS-Boot	STD-Boot
< 5%	$P_{\text{del}}/P_{\text{neu}}$	0,06	0,05	0,08	0,09
	P_{del}	0,06	0,05	0,08	0,09
	H	0,06	0,05	0,09	0,11
	PN	0,07	0,05	0,09	0,10
	PS	0,07	0,05	0,09	0,10
> 95%	$P_{\text{del}}/P_{\text{neu}}$	0,07	0,05	0,09	0,09
	P_{del}	0,07	0,06	0,09	0,09
	H	0,05	0,06	0,09	0,11
	PN	0,06	0,05	0,10	0,10
	PS	0,06	0,05	0,10	0,10

Carga genética por efeito de carona na adjacência de sítios sob seleção positiva

Resumo

Chun e Fay (2011) exploraram a relação entre seleção natural e carga genética, mostrando que, em populações humanas, regiões vizinhas a sítios sob seleção positiva apresentam mais polimorfismos de nucleotídeo único (SNPs) deletérios do que regiões distantes desses sítios por carona genética. Entretanto, a abordagem deles não implementou um teste estatístico que contemplava diferenças no número de SNPs entre regiões próximas a genes selecionados e as regiões controle, e também não corrigia para efeitos de uma possível diferenças no SFS. A partir de um método robusto, que é capaz de lidar com diferenças no número de SNPs entre as regiões testadas, temos como objetivo revisar essa questão e investigar se, para humanos, há acúmulo de SNPs deletérios por carona genética e avaliar se as mesmas regiões em populações que experimentaram e não experimentaram seleção positiva contribuem de forma diferente para a carga genética dessas populações. Nossos resultados mostram que há evidência de acúmulo de SNPs deletérios na vizinhança a sítios sob seleção positiva em europeus e leste asiáticos, e mostramos também o comportamento dessas mesmas regiões em populações onde elas não apresentam assinatura de seleção. Esses resultados indicam que, pelo menos nas regiões estudadas, a seleção natural é fundamental para aumentar a carga genética no genoma humano.

Palavras-chave: carona genética; $P_{\text{del}}/P_{\text{neu}}$; seleção positiva; reamostragem

Abstract

The relationship between natural selection and genetic load has been established by Chun and Fay (2011) that demonstrated, although not properly, regions near *loci* under selection had more deleterious single nucleotides polymorphisms (SNPs) than other regions in human genome. Based on a more robust methodology, capable of dealing with different number of SNPs between studied regions, we aim to revisit this question and investigate whether there are more deleterious SNPs due to hitchhiking in regions near *loci* under selection and evaluate if these same genomic regions carry more deleterious SNPs in other human population that were not under natural selection. We found evidence of more deleterious SNPs near *loci* under selection in european and east asian populations. We also showed that this pattern was only present in these population that experienced natural selection, indicating that, at least for the genome regions studied, natural selection was important to increase genetic load in human genome.

Key words: bootstrap methodology; $P_{\text{del}}/P_{\text{neu}}$; genetic load; positive selection

2.1 Introdução

O gargalo experimentado pelas populações durante a saída da África (ou OOA, de *Out of Africa*) (RAMACHANDRAN et al., 2005; ALVES et al., 2012; VEERAMAH; HAMMER, 2014). e a recente expansão populacional nos últimos 5000 anos (COVENTRY et al., 2010; TENNESSEN et al., 2012; GAZAVE et al., 2013) são eventos que impactaram decisivamente a distribuição dos polimorfismos de nucleotídeo único (SNPs) em humanos. Concomitante a esses efeitos demográficos, ao ser espalhado pelo mundo, o genoma humano passou por mudanças resultantes de processos adaptativos nos diferentes ambientes (ALVES et al., 2012; VEERAMAH; HAMMER, 2014). Como resultado, a variação genética em populações humanas é reflexo tanto de processos seletivos como de eventos demográficos que ocorreram em diferentes escalas de tempo (COLONNA et al., 2014).

Dentro desse contexto, estudos da variação genética deletéria têm estimulado as discussões em genética de populações humanas (LOHMUELLER, 2014a; HENN et al., 2015; KOCH; NOVEMBRE, 2017). Mesmo quando deletérios, os SNPs podem segregar por algum tempo, dependendo da intensidade da deriva e da seleção purificadora (HENN et al., 2015). O acúmulo desses SNPs deletérios pode ser utilizado como proxy de carga genética, uma vez que reduz progressivamente a aptidão média da população (HALDANE, 1957; KIMURA; MARUYAMA; CROW, 1963; LOHMUELLER, 2014a; HENN et al., 2015). Recentemente, diversos trabalhos têm investigado a carga genética humana sob perspectivas intrapopulacionais ou interpopulacionais. Os estudos interpopulacionais estão interessados em entender se diferenças nas histórias demográficas e seletivas de distintas populações humanas resultam em níveis variáveis de carga genética em todo genoma (LOHMUELLER, 2014a; SIMONS et al., 2014; HENN et al., 2015; KOCH; NOVEMBRE, 2017). Já os estudos intrapopulacionais investigam a dimensão intragenômica da variação demográfica, buscando entender de que forma fatores como proximidade a genes selecionados ou a taxa de recombinação local influenciam a eficácia da seleção natural em diferentes regiões do genoma (BETANCOURT; PRESGRAVES, 2002; CHUN; FAY, 2011; LENZ et al., 2016; ZHANG et al., 2016). Para os estudos intrapopulacionais que trabalham com a questão do efeito da proximidade a genes selecionados, o primeiro desafio é localizar com precisão regiões do genoma sob seleção natural.

Muitos trabalhos se dedicaram a entender quais padrões na variação genética humana indicam assinaturas de seleção natural (AKEY et al., 2002; NIELSEN, 2005; SABETI et al., 2006). Antes da era genômica, essas buscas por assinaturas de seleção eram guiadas quase exclusivamente por meio de genes candidatos (SABETI et al., 2006; AKEY, 2009). A capacidade de sequenciar genomas humanos inteiros em escala populacional permitiu testar processos adaptativos de maneira menos enviesada por expectativas prévias (CONSORTIUM et al., 2012; COLONNA et al., 2014), e as buscas por evidências de seleção positiva passaram a ser feitas em nível genômico, guiadas em maior parte pelo modelo de varredura seletiva. Esse modelo prevê que um SNP sob forte seleção positiva tende a aumentar rapidamente de frequência na população e, por efeito carona, aumentar também a frequência das variantes neutras ou mesmo fracamente deletérias fisicamente ligadas à

variante selecionada (SMITH; HAIGH, 1974), de modo que sua ascensão varre (i.e. reduz) a diversidade na região vizinha (BERRY; AJIOKA; KREITMAN, 1991; HERNANDEZ et al., 2011; MESSER; PETROV, 2013; CORBETT-DETIG; HARTL; SACKTON, 2015).

Um primeiro estudo a explorar a relação entre seleção natural e o acúmulo de mutações deletérias mostrou que, em populações humanas, que regiões vizinhas a sítios sob seleção positiva apresentam mais SNPs deletérios do que regiões distantes desses sítios (CHUN; FAY, 2011), e, por consequência, essas regiões devem contribuir proporcionalmente mais para carga genética dessas populações do que o resto de seus genomas. Embora o estudo tenha contribuído ao trazer para o âmbito da espécie humana a investigação empírica da carga genética, acreditamos que a pergunta proposta por esses autores pode ser explorada de forma mais precisa, conforme detalhado a seguir.

Em primeiro lugar, Chun e Fay (2011) utilizaram apenas uma estatística para quantificar a intensidade da carga genética. Há entretanto, importantes debates sobre qual a melhor métrica para medir a magnitude da carga genética numa região genômica. A comparação de diferentes estatísticas contribui tanto para a compreensão de como as estatísticas se comportam, como para avaliar a robustez dos achados prévios, indicando que há uma carga genética aumentada na região próxima a genes selecionados. Em segundo lugar, os testes estatísticos comparando regiões sob efeito de carona genética com o restante do genoma envolviam o contraste entre conjuntos de SNPs com grandes diferenças de tamanho (havendo muito mais SNPs no conjunto controle). Essa diferença de tamanho entre as regiões comparadas (i.e. a com carona genética e o resto do genoma) pode inflar os resultados com falsos positivos, na medida em que diferenças na proporção de SNPs deletérios entre os grupos podem ocorrer como simples decorrência do menor número amostral da região sob carona genética (KOCH; NOVEMBRE, 2017). Uma metodologia baseada em amostragem, que compare regiões de tamanhos semelhantes dentro da mesma população, é mais adequada para responder às questões propostas por Chun e Fay (2011). Outro ganho, decorrente dos avanços dos dados, é a possibilidade de estudar o efeito da carona genética em escala continental, o que, até onde sabemos, ainda não foi feito. Por escala continental nos referimos a casos em que uma mutação vantajosa tornou-se comum

num continente e não em outro. Mutações com esse perfil de distribuição tem as diferenças de frequência apoiadas por dados em múltiplas populações, de modo que a robustez da inferência de seleção é maior.

Este trabalho tem como objetivos: 1) investigar se, para humanos, há acúmulo de SNPs deletérios por efeito de carona genética utilizando mais de uma estatística e uma metodologia para testar diferenças entre regiões genômicas que seja baseada em reamostragens, capaz de lidar com diferenças no número de SNPs entre as regiões testadas, e, 2) também, avaliar se as mesmas regiões em populações que experimentaram e não experimentaram seleção positiva contribuem de forma diferente para a carga genética dessas populações.

2.2 Dados e Métodos

2.2.1 Banco de dados

As informações dos SNPs codificantes foram obtidas a partir de arquivos VCF da *phase3* do projeto 1000 genomas anotado para a versão hg19 do genoma humano (CONSORTIUM et al., 2015). Para testar se há efeito de carona genética em escala continental, estabelecemos grupos continentais a partir das populações que compõem o continente estudado, de maneira que o grupo europeu (EUR) foi formado a partir de 503 indivíduos das populações CEU (descendentes da Europa setentrional e ocidental vivendo em Utah, EUA), TSI (Toscana, Itália), FIN (Finlândia), GBR (britânicos da Inglaterra e Escócia) e IBS (Ibéricos da Espanha), o leste asiático (EAS) formado a partir de 503 indivíduos das populações CHB (grupo étnico Han de Pequim, China), JPT (Tóquio, Japão), CHS (grupo étnico Han do sul, China), CDX (grupo étnico Dai de Xishuangbanna, China) e KHV (grupo étnico Kinh de Ho Chi Minh, Vietnã), o sul asiático (SAS) formado a partir de 489 indivíduos das populações GIH (grupo étnico indiano Gujarati vivendo no Texas, EUA), PJI (indianos de Punjabi vivendo em Lahore, Paquistão), BEB (grupo étnico Bengali, Bangladesh), STU (grupo étnico Tamil do Sri Lanka vivendo no Reino Unido) e ITU (grupo étnico indiano Telugu vivendo no Reino Unido) e o africano (AFR)

formado a partir de 504 indivíduos das populações YRI (grupo étnico Iorubá de Ibadan, Nigéria), LWK (grupo étnico Luhya de Webuye, Kenya), GWD (Gâmbia), MSL (Mende, Serra Leoa) e ESN (grupo étnico Esan, Nigéria). Excluimos as populações miscigenadas da América AMR (i.e. MXL (descendentes de mexicanos vivendo em Los Angeles, EUA), PUR (Porto Rico), CLM (Medellín, Colômbia), PEL (Lima, Peru)) e outras também miscigenadas segundo material suplementar do 1000 genomas (i.e. ASW (afro descendentes norte americanos, EUA, ACB (afro descendentes caribenhos, Barbados)). Calculamos a frequência alélica (AF - *Allele Frequency*) para cada SNP nos continentes utilizando o programa *VCFtools* v0.1.12b (DANECEK et al., 2011). Mantivemos em nossas análises apenas sítios bialélicos, sendo MAF a frequência do alelo menos frequente calculada com $1 - AF$ quando $AF > 0,5$.

2.2.2 Identificando seleção positiva

Estudar os efeitos da carona genética tem como primeiro desafio definir com precisão as regiões adequadas para se testar a hipótese (i.e. regiões sob seleção positiva). Os alelos, ao experimentarem seleção positiva, deixam diversas assinaturas no genoma. Essas assinaturas podem ser identificadas ao se observar desvios em regiões ou sítios específicos do genoma em relação ao esperado sob neutralidade (SABETI et al., 2006). Uma ferramenta clássica utilizada para encontrar essas assinaturas é a diferenciação genética entre duas populações (WEIR et al., 2005; NIELSEN et al., 2007; AKEY, 2009). Assume-se que a frequência do alelo que carrega o SNP vantajoso irá aumentar apenas na população selecionada, aumentando a diferenciação entre as populações (SABETI et al., 2006). No entanto, um desafio central dessa abordagem é definir qual é o grau de diferenciação que supera o esperado apenas em decorrência de eventos demográficos (i.e. efeitos de deriva genética, HOLSINGER, WEIR, 2009; COLONNA et al., 2014). Dessa forma, mesmo as regiões mais diferenciadas do genoma humano podem resultar da combinação de seleção natural e deriva genética, e o uso apenas de uma metodologia para encontrar assinatura de seleção pode apresentar falsos positivos (AKEY, 2009). Frente a esse desafio, duas abordagens são classicamente utilizadas para identificar casos de seleção e controlar os

efeitos demográficos. Primeiro, pode-se usar testes de "*outlier*", em que um gene é classificado como selecionado se tiver um nível de diferenciação extremo em relação ao resto do genoma. Em segundo lugar, podem ser usadas simulações de cenários demográficos realistas, de modo a identificar diferenciação que supera as expectativas sob neutralidade e demografia realista. No contexto deste trabalho, para minimizar a taxa de falsos positivos no teste da hipótese e para controlar para efeitos demográficos, adotamos como critério para identificar genes que apresentam assinaturas de seleção positiva a ocorrência de sinais de seleção com base em diversos métodos de buscas que individualmente controlam para demografia com abordagens de *outlier* e/ou simulação, reduzindo o efeito da variação de fundo causada pela deriva genética (VOIGHT et al., 2006; SABETI et al., 2006; AKEY, 2009).

Um desses testes que estabelece evidência para seleção positiva é baseado em medidas de F_{st} . O F_{st} , uma clássica medida utilizada para estimar a diferenciação genética, está diretamente relacionado à variação na frequência do alelo entre as populações e inversamente relacionado ao grau de semelhança entre indivíduos dentro das populações (HOLSINGER; WEIR, 2009). Recentemente, pesquisadores calcularam as diferenças de frequência do alelos derivados (ΔDAF - *Delta Derived Allele Frequency*), estatística altamente correlacionada ao F_{st} (*Spearman* $r = 0,93$ para $0,95$, todos os p -valores $< 10^{-16}$), para acessar a diferenciação genética (Colonna et al. 2014). Usando essa métrica, identificaram um conjunto de genes com os SNPs que tinham $\Delta DAF > 0,7$ (conjunto 1), chamados de "HighD" (para *High Differentiation*), definindo assim o conjunto de SNPs mais altamente diferenciados do genoma humano (COLONNA et al., 2014). Através de simulações os autores demonstraram que, assim como o F_{st} , ΔDAF tem mais 98% de poder para detectar varreduras seletivas parciais e completas e tem um desempenho melhor do que *XP-EHH* (um teste baseado em diversidade haplotípica) para varreduras parciais. O mesmo estudo delimitou, a partir de revisão da literatura, outro conjunto de genes com múltiplas evidências de estarem sob seleção positiva (i.e. determinados como genes sob seleção positiva a partir de diferentes métodos, além do ΔDAF ; conjunto 2) (COLONNA et al., 2014). Comparando os dois conjuntos, o estudo mostrou que se considerados somente os genes com HighD em suas regiões codificadoras e entre continentes, existe 63%

de sobreposição do conjunto 2 no conjunto 1 (COLONNA et al., 2014), concluindo que esses SNPs são os prováveis responsáveis pelo aumento da aptidão (ou fortemente ligados aos que conferem esse fenótipo) e, conseqüentemente, os sítios alvos de seleção positiva. Para aumentar a probabilidade das regiões que utilizamos de fato flanquearem sítios selecionados, o conjunto de genes utilizado para testar nossa hipótese foi formado pelos genes do conjunto 1, que continham SNPs HighD calculados entre continentes em suas regiões codificadoras.

2.2.3 Identificando sinais de seleção positiva em escala continental

Dentre os grupos continentais formados por populações não miscigenadas presentes no 1000 genomas (AFR, EAS, EUR, SAS), Colonna e colaboradores (2014) identificaram SNPs HighD a partir de comparações entre AFR, EAS e EUR. Considerando esses três grupos continentais, investigamos os pares EAS-AFR e EUR-AFR, pois esses apresentaram os SNPs HighD com as maiores diferenciações ($\Delta DAF > 0,92$ e $\Delta DAF > 0,83$, respectivamente; Colonna et al. 2014). Quando há um sítio altamente diferenciado entre duas populações, é possível que ambas ou apenas uma delas esteja sob seleção positiva (NIELSEN, 2005; SABETI et al., 2006). Para descobrir em qual continente o SNP HighD estava sob seleção positiva, olhamos a frequência do alelo derivado (DAF) dos SNPs HighD, uma vez que alelos derivados em frequências elevadas são uma evidência adicional de seleção positiva (SABETI et al., 2006). No caso do par EAS-AFR, foram encontrados 32 e 0 SNPs com maior DAF em EAS e AFR, respectivamente. Já no par EUR-AFR foram encontrados 32 e 10 SNPs com maior DAF em EUR e AFR, respectivamente. Portanto, neste trabalho avaliamos somente os grupos continentais EAS e EUR por serem os que contêm o maior número de SNPs HighD com evidência de terem sido selecionados naquela região.

2.2.4 Diversidade ao redor de sítios sob seleção positiva

Inicialmente, determinamos se as regiões adjacentes aos SNPs HighD seguem as expectativas do modelo de varredura seletiva, com redução de diversidade. Para tanto, definimos janelas de diversos tamanhos flanqueando os SNPs HighD e analisamos o comportamento da diversidade à medida que a janela tornava-se maior, distanciando-se dos SNPs HighD. Para quantificar a variabilidade estimamos a taxa da heterozigosidade como $2MAF(1 - MAF)$, para cada SNP da janela, e em seguida calculamos a heterozigose média (H) desses SNPs.

2.2.5 Definição da janelas de estudo

Após determinar que as regiões adjacentes aos SNPs HighD seguem o modelo de varredura (ver Resultados), determinamos o tamanho de janela a ser utilizado nas etapas seguintes do trabalho, em que se quantifica a quantidade de SNPs deletérios na região vizinha ao SNP selecionado. Quanto maior a janela, maior o número de SNPs que são testados e consequentemente mais confiáveis são as estimativas. Porém, devido ao processo de recombinação, o efeito da seleção natural é mais forte na vizinhança imediata dos SNPs selecionados, e diminui com o aumento da distância a partir deles (CHARLESWORTH, 2006). Nesse sentido, o tamanho de janela ideal é aquele que representa um balanço entre essas duas condições: grande o suficiente para resultar em um número mínimo de SNPs para realizar testes, mas não grande demais, para que ela compreenda uma região sob influência da seleção natural. Portanto, o primeiro critério para determinar o tamanho de janela foi a busca por aquela que resulta no maior conjunto de SNPs possível cuja a diversidade média seja significativamente menor do que a média genômica, utilizando um critério de *outlier* ($p < 0,05$) em uma distribuição gerada pelo método de reamostragem Gene-Std-Boot (Figura 2.1 dos resultados). O tamanho de janela escolhida usando esses critérios foi de 250kb e 450kb em europeus e leste asiáticos, respectivamente. Porém esse tamanho de janela ainda pode sofrer pequenas oscilações em decorrência de um segundo critério: que elas incluam apenas genes inteiros. Nos casos em que a parte final da janela

se sobrepunha a alguma região de um gene, 1) excluímos a parte da janela que caía sobre o gene caso uma parte muito grande (maior que 5kb) dele estivesse fora da janela e 2) incluímos o restante do gene caso contrário, pois incluir todos os genes indiscriminadamente pode aumentar a diversidade para valores próximos à média genômica (Figura S2.1). Após a inclusão ou exclusão dos genes de acordo com esse critério, obtivemos um total de 766 SNPs na região ao redor de SNPs highD distribuídos em 57 genes para os europeus (EUR) e 1336 SNPs na região ao redor de SNPs highD distribuídos em 91 genes para os leste asiáticos (EAS). Todos os genes dentro dessas janelas foram classificados como grupo positivo (por serem genes vizinhos a uma região sob seleção positiva), e SNPs de todo o resto do genoma foram classificados como grupo controle. A Tabela S2.1 e Tabela S2.2 apresentam um sumário dos genes que formam o grupo positivo em europeus (EUR) e leste asiáticos (EAS).

2.2.6 Quantificação de carga genética

As estatísticas que utilizamos para quantificar SNPs deletérios são aplicáveis apenas a sítios que codificam proteína. Assim, restringimos nossas análises apenas a essa classe de SNPs. Para classificá-los, utilizamos as anotações do *Polyphen* (KIRCHER et al., 2014) que aloca os SNPs não sinônimos em três categorias funcionais: *Probably Damaging*, *Possibly Damaging* ou *Benign*, sendo o número total de SNPs em cada uma dessas classes, abreviada por P_{d1} , P_{d2} e P_b , respectivamente. Optamos por essa ferramenta de anotações pelo fato do *Polyphen* ser baseado em um naïve bayes machine learning que infere a probabilidade de um SNP ser deletério considerando em conjunto conservação filogenética e medidas empíricas de substituições que afetam a estrutura tridimensional da proteína, enquanto anotações análogas (*Sift* ou *GERP*) consideram apenas os padrões de conservação filogenética. Segundo a classificação do *Polyphen*, SNPs classificados como *Probably Damaging* apresentam maior probabilidade de serem deletérios, e portanto utilizamos P_{d1} como estatística para quantificar o número de SNPs deletérios (abreviado por P_{del}). Outra estatística utilizada foi a divisão do número de SNPs deletérios pelo o número de SNPs neutros (razão abreviada aqui como P_{del}/P_{neu}). Para isso, calculamos o

P_{del} nos grupos positivo e controle e dividimos pela contagem de todas as outras classes de SNPs ($P_{\text{neu}} = P_{\text{b}} + P_{\text{d2}} + PS$). Nem todos os SNPs não sinônimos estão classificados pelo *Polyphen*. Para garantirmos uma análise precisa, utilizamos apenas genes com uma proporção alta de SNPs classificados pelo Polyphen, especificamente genes que satisfaçam a condição $(P_{\text{b}} + P_{\text{d1}} + P_{\text{d2}})/PN > 0,95$, implicando que 95% de seus SNPs classificados pelo *Polyphen*.

2.2.7 Reamostragem em janelas sob seleção positiva

Para comparar o número de SNPs deletérios entre a região alvo e a controle, utilizamos uma metodologia de reamostragem. Uma vez que diferenças entre os SFS (*Site Frequency Spectrum*) das regiões podem interferir na detecção de diferenças na carga genética entre elas, caso o SFS seja diferente, é necessário controlar para esse fator. Portanto, para determinar qual critério de reamostragem descrito no capítulo 1 deve ser utilizado, realizamos um teste de premissa, no qual, avaliamos a semelhança entre o SFS dos grupos positivo e controle (ver Resultados).

2.2.8 Análises Interpopulacionais

Toda metodologia apresentada até este momento visa testar a hipótese do aumento da carga genética devido ao efeito carona dentro de um mesmo continente (i.e análise intrapopulacional). Para os casos em que o teste revelou aumento da carga genética no grupo positivo, efetuamos uma segunda análise, que consiste na comparação de achados entre continentes. Especificamente, nós realizamos uma análise interpopulacional com o objetivo de avaliar se o aumento de SNPs deletérios na região adjacente ao SNP HighD é um efeito da carona genética ou, pelo contrário, seria resultado de alguma característica das regiões estudadas, não relacionada ao seu status de ser vizinha de um SNP com forte evidência de seleção natural. Para tanto, submetemos os mesmos genes às mesmas análises descritas até aqui (i.e. contagem de SNPs deletérios por meio de P_{del} e $P_{\text{del}}/P_{\text{neu}}$, seguida de uma análise de *outlier* comparando com restante do genoma), porém, para os

grupos continentais que não experimentaram seleção positiva. Alinhado a nossa hipótese, esperamos que para as populações sem evidência de seleção positiva na vizinhança dos genes sob análise, não exista acúmulo de SNPs deletérios.

Em seguida, avaliamos a variação na quantidade de SNPs deletérios em função da distância das regiões selecionadas, semelhante ao que fizemos com a medida de diversidade, nos mesmos genes em todas os continentes. Nossa expectativa é que apenas os grupos continentais sob seleção positiva apresentem aumento no número de SNPs deletérios próximo aos sítios selecionados e redução na quantidade desses SNPs com o aumento da distância desses sítios. Para os demais grupos continentais, o número de SNPs deletérios deve oscilar ao redor da média genômica, independente da distância até o sítio selecionado. Para isso, montamos janelas com distâncias cumulativas flanqueando os SNPs HighD e medimos $P_{\text{del}}/P_{\text{neu}}$ para cada janela. A fim de controlar o efeito demográfico, subtraímos do $P_{\text{del}}/P_{\text{neu}}$ de cada janela a média de $P_{\text{del}}/P_{\text{neu}}$ genômico do grupo continental analisado. Desse modo, cada grupo continental teve o valor de $P_{\text{del}}/P_{\text{neu}}$ controlado por sua própria média, o que permite a comparação direta entre eles.

2.3 Resultados

Neste capítulo trabalhamos com populações europeias e leste asiáticas, e podemos notar que alguns genes classicamente reportados como estando sob efeito de seleção positiva nessas regiões não estão no conjunto de genes analisados (Tabela 2.1). São exemplos de variantes ausentes, porém com evidências prévias de seleção, o SNP rs4988235 na região promotora do gene LCT, associado à persistência da lactase na vida adulta (BERSAGLI-ERI et al., 2004), e o SNP rs12913832 em HERC2 que implica na coloração azul nos olhos (EIBERG et al., 2008; STURM et al., 2008). Ambos estão presentes na lista de Colonna e colaboradores (2014), mas foram excluídos de nossas análises, pois estão restritos à população britânica e não apresentam evidências forte para a seleção positiva em todo continente europeu. No entanto, nossa lista também apresenta alguns alelos previamente identificados em estudos de seleção positiva, como é o caso do rs1426654 em SLC24A5

e do rs16891982 em SLC45A2, ambos associados a coloração da pele em europeus (LAMASON et al., 2005; NORTON et al., 2007). Obtivemos apenas um gene presente em ambas os continentes estudados, o gene HIP1, que apresenta SNPs HighD em tanto em europeus como em asiáticos. Todos os outros genes estão sob seleção em somente um dos continentes (Tabela 2.1).

Tabela 2.1 – Conjunto de genes utilizados sob seleção positiva (i.e.compresença de SNPs HighD) em europeus (EUR) e leste asiáticos (EAS). Em negrito o único gene presente nos dois grupos continentais.

Grupos Continentais			
EUR		EAS	
HIP1	ATP1A1	HIP1	DOK5
RCBTB2	TLR1	HINT3	ACSS1
SMYD3	ABCA12	ITGB1BP1	CUX1
TPRG1	CNTNAP4	STAG1	GRB2
TPRG1	AVEN	BRF1	IKZF2
DNAJB6	CELSR1	TBC1D4	FBXO31
NAV2	CYP26B1	HMCN1	BCAS3
CNTNAP2	ZEB1	MFAP3	EXOC6B
SF1	PDE7B	TENM4	VAV2
ATP6V1H	DACH1	SLC35F3	L3MBTL4
GALNT5	CD226	PRTG	ADCY2
LRRC4B	ARHGAP10	CDH7	MAGI1
GRB2	NECAB2	SNTG1	SLC30A9
FOXP1	MTHFD2L	RABGGTB	VWF
AGAP1	ALK	RILPL1	BACE2
EPHA6	SLC24A5	AP5M1	XKR

Em relação à verificação de que os dados seguem o modelo de varredura, os resultados mostram que tanto europeus quanto a leste asiáticos apresentam uma redução da diversidade genética na adjacência aos SNPs HighD (Figura 2.1). Devido à recombinação, a diversidade genética retorna a valores mais altos, com algumas oscilações, conforme aumenta a distância em relação aos sítios selecionados (CHARLESWORTH, 2006; CUTTER; PAYSEUR, 2013; CORBETT-DETIG; HARTL; SACKTON, 2015) e alcança a

média genômica em distâncias superiores a 2Mb (Figura 2.1). Porém, submetido a uma análise de *outlier*, o desvio é significativamente menor que a média genômica somente em distâncias menores (para leste asiática, até 450 kb, para europeia, até 250 kb; Figura 2.1).

Figura 2.1 – *Baixa diversidade (medida pela heterozigosidade média dos SNPs na região) na região ligada aos SNPs HighD formada com janelas cumulativas em (a) leste asiáticos (EAS) e (b) europeus (EUR). Os retângulos preenchidos indicam o tamanho da janela para a qual há uma diversidade significativamente menor do que a média genômica, conforme teste feito utilizando a reamostragem STD-Gene-Boot.*

No teste de premissa, no qual avaliamos a semelhança entre o SFS dos grupos positivo e controle, não encontramos diferença entre o SFS do grupo positivo e controle (Figura 2.2a e b). Assim, utilizamos como metodologia de reamostragem, para este capítulo, o STD-Gene-Boot, que gera uma distribuição baseada em reamostragens nas quais o grupo controle e positivo apresentam o mesmo número de SNPs e de genes, para cada amostra, mas não é feito nenhum procedimento para parear o SFS entre grupos de SNPs.

Figura 2.2 – *Espectro de Frequência Alélica (SFS) dos grupos positivo e controle, para europeus e leste asiáticos. SFS de todos os SNPs codificantes para o grupo positivo e controle em (a) europeus (EUR) e (b) leste asiáticos (EAS); SFS separando por grupos de SNPs codificantes (não sinônimo, N; sinônimos, S) para o grupo positivo e controle para (c) europeus (EUR) e (d) leste asiáticos (EAS).*

Para duas estatísticas de carga genética ($P_{\text{del}}/P_{\text{neu}}$ e P_{del}), o grupo positivo apresenta valores que são extremos em relação àqueles obtidos por reamostragens de SNPs controle, usando o critério de reamostragem STD-Gene-Boot, tanto para europeus quanto leste asiáticos. Essa evidência de efeito da carona genética é observada apenas para os SNPs definidos como pertencentes ao grupo positivo por serem adjacentes ao um SNP highD daquela região: a análise mostra que não há um excesso de variantes deletérias, por qualquer estatística, quando se quantifica os mesmos genes em outro grupo continental (Figura 2.3). Repare que para os casos em que SNPs HighD são selecionados em EAS (duas fileiras superiores), há P_{del} e $P_{\text{del}}/P_{\text{neu}}$ significativos em EAS (terceira coluna), mas não nos demais grupos continentais (i.e. AFR, EUR e SAS), mesmo quando exatamente os mesmos genes são analisados. O mesmo acontece para SNPs HighD selecionados em EUR. Repare também que apesar dos sul asiáticos e africanos não terem sido incluídos na análise de efeito carona pelas razões expostas anteriormente, na comparação interpopu-

lacional eles não tiveram valores de P_{del} e P_{del}/P_{neu} significativos quando analisados para genes adjacentes aos SNPs HighD sob seleção em europeus ou leste asiáticos (Figura 2.3, primeira e quarta coluna), o que reforça a ideia de que o efeito carona está presente apenas na vizinhança de sítios sob seleção positiva naquela região específica.

Figura 2.3 – Análise de outlier para as estatísticas P_{del} e P_{del}/P_{neu} . Os violin plots indicam as regiões sob seleção positiva para (a) leste asiáticos (EAS) e (b) europeus (EUR), utilizando o genoma de grupo continental como controle. Os valores acima de cada violin plot são os p-valores da respectiva análise de outlier. A cor verde indica os casos em que há significância seguindo o critério de valor crítico de 0,05, e vermelho indica não significância, pelo mesmo critério.

O progressivo decréscimo de P_{del}/P_{neu} para janelas que abrangem regiões maiores a partir dos SNPs HighD também indica o papel determinante da carona genética no acúmulo de SNPs deletérios. Janelas pequenas no entorno dos SNPs HighD apresentam grandes desvios no P_{del}/P_{neu} , enquanto janelas grandes, que incluem maiores distâncias, apresentam menores desvios de P_{del}/P_{neu} em consequência da redução do efeito da carona genética (Figura 2.4). As mesmas regiões em grupos continentais que não experimentaram

seleção natural não tiveram grandes oscilações de P_{del}/P_{neu} (Figura 2.4).

Figura 2.4 – Diferença entre a carga genética (estatística P_{del}/P_{neu}) ao redor dos SNPs HighD e sua média genômica, para diferentes distâncias em relação ao SNP HighD. As linhas indicam os diferentes grupos continentais utilizando os mesmos genes que estão na adjacência de SNPs HighD em (a) leste asiáticos (EAS) e (b) europeus (EUR).

2.4 Discussão

Neste estudo testamos a hipótese de que forte seleção positiva aumenta a carga genética em genes vizinhos através do efeito de carona genética. Para tanto, utilizamos um critério conservador para determinação de genes sob seleção em escalas continentais, assim como um teste mais robusto do que tem sido utilizado na literatura, que controla para o número de SNPs e desequilíbrio de ligação entre os SNPs.

2.4.1 Maior carga genética em regiões em que há efeito carona

Ambas estatísticas, $P_{\text{del}}/P_{\text{neu}}$ e P_{del} , indicam que as regiões vizinhas a SNPs highD estão estatisticamente enriquecidas para SNPs deletérios tanto no leste da Ásia quanto na Europa (Figura 2.3). O enriquecimento de SNPs deletérios nas regiões vizinhas aos sítios selecionados pode ser interpretado como uma redução no tamanho efetivo populacional (N_e) nessa região (CHARLESWORTH, 2013). Assim, a carona genética promove o aumento da deriva genética nessas regiões e, como consequência, provoca a redução da eficácia da seleção purificadora em retirar os SNPs deletérios. Portanto, independente da estatística utilizada, nossos resultados apontam a seleção positiva como o principal mecanismo atuando no aumento de SNP deletérios no grupo positivo em todo continente europeu e em todo leste asiático, indicando que os efeitos da carona genética apresentam em humanos sinais mesmo em escalas continentais.

O fato de apenas um gene estar ao mesmo tempo sob seleção na Europa e no leste da Ásia (Tabela 2.1; COLONNA et al., 2014) indica que os genes responsáveis pelo acúmulo dos SNPs deletérios são exclusivos das populações e, portanto, que o acúmulo desses SNPs em escala continental aconteceu em lugares diferentes do genoma para cada continente.

2.4.2 Carga genética somente nos sítios sob seleção

De acordo com grande parte dos estudos recentes, populações não-africanas carregam maiores carga genéticas do que populações africanas (LOHMUELLER et al., 2008; LOHMUELLER, 2014b). Quando as mutações deletérias têm pequeno impacto no fenótipo, o tamanho da população tem grande importância na quantidade de variação deletéria segregante (KOCH; NOVENBRE, 2017). Dessa forma, a diferença na carga genética entre africanos e não-africanos foi atribuída à história demográfica dessas populações, que reduziu a capacidade da seleção purificadora em retirar as variações deletérias em populações não-africanas devido aos sucessivos gargalos durante o OOA (LOHMUELLER et al., 2008; LOHMUELLER, 2014b). Recentemente, esses resultados foram contestados por estudos que, utilizando novas métricas para medir a carga genética (o número de mutações deleté-

rias por indivíduo), apontaram que populações africanas e não-africanas carregam cargas genéticas muito semelhantes, indicando que as histórias demográfica dessas populações tiveram pouco impacto sobre a variação deletéria segregante (SIMONS et al., 2014; DO et al., 2015), e que provavelmente os padrões observados não refletem mudanças na eficácia da seleção purificadora em não-africanos. No entanto, é consenso que o gargalo enfrentado por não-africanos durante o OOA e o posterior crescimento dessas populações não foram processos fortes o suficiente para provocar grandes alterações nas cargas genéticas entre as populações humanas (KOCH; NOVEMBRE, 2017). Portanto, a discordância desses estudos pode ser atribuída as diferentes métricas utilizadas para definir carga genética e análise de diferentes conjuntos de dados (LOHMUELLER, 2014b).

Nossos resultados dizem respeito a um fator que também contribui para o aumento da carga genética, ainda que não em escala genômica. Eles mostram que a seleção purificadora foi mais eficiente ao retirar SNPs deletérios em regiões do genoma em que não há evidência de seleção positiva forte para aquele grupo populacional, (Figura 2.3 e Figura 2.4). Com isso, os resultados indicam que o acúmulo de SNPs deletérios em regiões vizinhas a sítios sob seleção positiva é um efeito direto de carona genética característica daquela região geográfica, e não o resultado de particularidades daquelas regiões genômicas. O efeito de carona genética é uma função inversa da distância ao sítio selecionado (TENNESSEN et al., 2012; HENN et al., 2015) e, ao aumentar a distâncias dos SNPs HighD, é maior a eficiência da seleção purificadora em retirar os SNPs deletérios (Figura 2.4). Em grupos continentais que não experimentaram seleção positiva nessas regiões, a seleção purificadora retira SNPs deletérios com maior eficiência em todos os tamanhos de janelas testados (i.e. em todas as distâncias dos SNPs HighD).

Os resultados indicam que a varredura seletiva experimentada no leste da Ásia e na Europa foi fundamental para o acúmulo de SNPs deletérios em consequência de carona genética, contribuindo para a carga genética desses grupos. Indicam também que, ao trabalharmos com genes sob seleção exclusivamente na Europa e outros genes exclusivamente no leste da Ásia (Tabela 2.1), conseguimos apontar que a seleção natural nessas localidades reduziu a eficiência da seleção purificadora em retirar SNPs deletérios nesses

sítios, efeito não observado em grupos continentais que não experimentaram seleção positiva. Essas conclusões só foram possíveis pela aplicação da análise intrapopulacional em diversas grupos continentais (que aqui chamamos de interpopulacionais) e, portanto, sugerimos que próximos trabalhos intrapopulacionais também o façam.

2.4.3 Padrão de SFS encontrado

O padrão de SFS encontrado para as regiões adjacentes a SNPs HighD (Figura 2.2) já foi reportado pela literatura (TENNESSEN et al., 2012; HENN et al., 2015). O fato de haver grande concentração de SNPs tanto sinônimos como não sinônimos em baixa frequência, pode ser explicado pela recente expansão no tamanho das populações humanas que contribuiu efetivamente para influxo de novas mutações, que aparecem em baixa frequência (TENNESSEN et al., 2012; HENN et al., 2015). Também, o fato de SNPs não sinônimos atingirem frequências elevadas em menor proporção do que SNPs sinônimos (Figura 2.2c e d) indica o efeito da seleção purificadora agindo contra alelos deletérios, resultantes de SNPs não sinônimos (FAY; WYCKOFF; WU, 2001; TENNESSEN et al., 2012).

2.4.4 Escolha conservadora dos genes sob seleção

O número de genes incluídos nesta análise é menor do que tem sido reportado na literatura de seleção natural como prováveis genes que experimentaram seleção positiva (AKEY, 2009; GROSSMAN et al., 2010; GROSSMAN et al., 2013). Uma limitação comum desses estudos é estabelecer de forma exata quais padrões da variação genética se referem a processos seletivos e quais se referem a processos demográficos. Simulações mostram que os métodos utilizados por esses estudos inflam os resultados para falsos positivos e o cálculo da taxa desse erro depende de parâmetros difíceis de estimar (AKEY, 2009). Não foi de nosso interesse conseguir o maior número de genes que experimentaram seleção positiva e sim buscar um número suficiente de genes com fortes evidências para testarmos nossa hipótese (Tabela 2.1). Visando um critério conservador, usamos filtros para reduzir o

número de falsos positivos entre os genes identificados com assinatura de seleção positiva (SIMONSEN; CHURCHILL; AQUADRO, 1995; PRZEWORSKI; HUDSON; RIENZO, 2000; AKEY et al., 2004; STAJICH; HAHN, 2004).

2.4.5 Contribuições do uso de reamostragem

Nas análises apresentadas neste capítulo, o grupo positivo contém 754 SNPs no continente europeu e 1304 no continente asiático, números muito inferiores aos do grupo controle (223.656 e 233.925, respectivamente). É comum estudos de carga genética intrapopulacionais realizarem seus testes em conjuntos de dados com grande diferença no número de SNPs entre a região alvo e o controle (CHUN; FAY, 2011). O aumento da frequência de SNPs deletérios por efeito carona foi observado em arroz domesticados utilizando um teste χ^2 (LU et al., 2006). Resultado semelhante foi encontrado em populações de milho, a partir de um teste de Mann-Whitney (MEZMOUK; ROSS-IBARRA, 2014). Em populações humanas, Chun e Fay (2011) também encontraram aumento de SNPs deletérios em regiões próximas a sítios sob seleção positiva, utilizando um teste exato de Fisher. Essa comparação direta, sem considerar a diferença de tamanho em relação ao controle, é perigosa, pois pode afetar a conclusão dos resultados. Uma pequena perturbação (ex. presença de um SNP deletério) provoca um efeito proporcionalmente maior em uma amostra pequena do que em uma amostra grande. Dessa forma, ao se comparar regiões pequenas com o resto do genoma, pode-se inflar os resultados para falsos positivos. Ao utilizar uma abordagem baseada em reamostragem, nossos métodos garantem que as regiões comparadas têm o mesmo tamanho, e assim asseguram que os resultados encontrados são mais robustos, sendo reflexos da seleção natural e não de diferenças provocadas pelo tamanho da amostra.

2.5 Conclusão

A partir de uma metodologia atualizada onde geramos o controle a partir de uma abordagem de reamostragem e da utilização de diferentes estatísticas, encontramos um

aumento da carga genética por efeito de carona genética em escalas continentais em europeus e leste asiáticos. Esses resultados, em conjunto com uma avaliação interpopulacional, indicam que, ao menos em regiões próximas a sítios selecionados, a seleção natural é fundamental para a manutenção da variação deletéria dentro e entre populações humanas. Dada a robustez do método utilizado, que além de controlar para o tamanho da amostra também controla para o desequilíbrio de ligação das regiões testadas, permitindo atribuir a diferença encontrada somente à ação da seleção positiva e, também, o ganho interpretativo em avaliar a mesma região em diversas populações e não somente uma, entendemos que os resultados encontrados sejam definitivos.

Carga genética em genes do HLA por efeito carona de sítios sob seleção balanceadora

Resumo

Em humanos, inúmeras evidências apontam os genes do HLA como alvos de seleção balanceadora, que mantém a diversidade genética nos sítios selecionados acima do esperado por neutralidade. Os genes do HLA codificam proteínas responsáveis por apresentar peptídeos antigênicos na superfície de células de defesa. Dentro dos genes HLA, os sítios de reconhecimento do antígeno (ou ARS, de *Antigen Recognition Site*) são os alvos da seleção balanceadora por codificarem a região da molécula à qual o antígeno é ligado. Usando o método de reamostragem descrito no capítulo 1, investigamos se há acúmulo de SNPs deletérios em éxons não-ARS nos genes clássicos do HLA de classe I por efeito carona provocado pelos éxons ARS, contribuindo para a carga genética do genoma humano. Apontamos o acúmulo de SNPs deletérios nos éxons não-ARS para ao menos dois agrupamentos populacionais, convergindo com o que tem sido reportado na literatura de carga genética, para os genes vizinhos dos genes do HLA.

Palavras-chave: Seleção Balanceadora, genes do HLA, carga genética, SNPs deletérios, não-ARS.

Abstract

In humans, there are many evidence pointing to HLA genes being under balancing selection, a selection regime that maintains genetic diversity at selected sites above expected by neutrality. HLA genes encode proteins responsible for antigenic peptides on the surface of defense cells. Within HLA genes, antigen recognition sites (ARS) are the targets of balancing selection because they encode the region of the molecule to which the antigen is linked. Using the resampling method described in chapter one, we investigated whether there is increasing load in non-ARS exons in classic HLA class I genes as a consequence of hitchhiking effect caused by the ARS exons. We point out increased load in non-ARS exons for at least two continental groups, converging with what has been reported in the genetic load literature for the genes near to HLA genes.

Key words: Balancing Selection, HLA genes, genetic load, deleterious SNPs, non-ARS.

3.1 Introdução

No segundo capítulo desta dissertação, investigamos o acúmulo de SNPs deletérios nas regiões vizinhas a sítios sob seleção positiva. Nele, vimos que a seleção natural foi o mecanismo fundamental para esse acúmulo ao estabelecer um efeito de carona genética em sítios próximos aos sítios selecionados. O efeito de carona genética reduz o tamanho efetivo (N_e) desses sítios, aumentando a deriva genética aleatória e por consequência reduzindo o poder da seleção purificadora em retirar os SNPs deletérios. No entanto, como adiantado na introdução geral, a seleção positiva não é o único mecanismo evolutivo que aumenta a aptidão do indivíduo que carrega uma variante benéfica. A seleção balanceadora, por exemplo, é outro regime seletivo que promove vantagens ao indivíduo e que seguindo argumentos semelhantes ao apresentados no capítulo 2, pode contribuir para a carga genética em humanos.

A seleção balanceadora é um termo abrangente que engloba todo processo evolutivo que, de forma adaptativa, mantém a diversidade genética nos sítios selecionados acima do esperado por neutralidade (CHARLESWORTH, 2006). Pode-se observar seleção balanceadora em ambientes onde o genótipo com maior aptidão é o heterozigoto; em ambientes onde o alelo menos frequente é o alelo vantajoso; ou quando há pressões seletivas que variam no tempo e no espaço (ANDRÉS, 2011). Em todos esses cenários, a seleção balanceadora mantém os diferentes alelos na população em grandes escalas de tempo, reduzindo a probabilidade de que haja fixação de algum deles (KAPLAN; DARDEN; HUDSON, 1988). Um sinal comum deixado pela seleção balanceadora no genoma é o excesso de polimorfismos segregando em frequências intermediárias (ANDRÉS et al., 2009; MEYER et al., ; BITARELLO et al., 2018). Como visto no capítulo 1, em humanos, o SFS de todo genoma tipicamente tem o formato de L, com a maior parte dos SNPs distribuídos em baixas frequências (TENNESSEN et al., 2012; HENN et al., 2015; BITARELLO et al., 2018). Já em regiões sob seleção balanceadora, é comum o SFS apresentar o padrão diferente onde a frequência dos alelos é desviada para níveis intermediários (ANDRÉS et al., 2009; MEYER et al., ; BITARELLO et al., 2018).

Nos vertebrados mandibulados, o Complexo Principal de Histocompatibilidade (ou MHC, de *Major Histocompatibility Complex*) é um dos conjuntos gênicos mais documentado devido a importância de suas proteínas na defesa frente a patógenos. Em humanos, esse complexo gênico é denominado sistema Antígeno Leucocitário Humano (ou HLA, de *Human Leukocyte Antigen*) e apresenta diversas evidências que indicam que seus genes estão sob seleção balanceadora. São alguma dessas evidências: a diversidade anormalmente aumentada se comparada com o esperado sob neutralidade, taxa de substituição não-sinônima maior do que a sinônima, alto desequilíbrio de ligação (MEYER; THOMSON, 2001; GARRIGAN; HEDRICK, 2003; SPURGIN; RICHARDSON, 2010), SFS com excesso de variantes em frequências intermediárias (GARRIGAN; HEDRICK, 2003) e correlação positiva entre o número de polimorfismo em HLA e a diversidade de patógenos (PRUGNOLLE et al., 2005).

Os genes HLA continuam um subconjunto de uma região maior que confusamente

também é denominada MHC (HUGHES, 2008). Para evitar imprecisão, neste capítulo, utilizamos o nome HLA para nos referirmos ao conjunto gênico humano e MHC ao conjunto dos vertebrados. Os genes HLA estão localizados no braço curto do cromossomo 6 (A LICHTMAN A, 2000), e, são divididos em genes de classe I, que se subdivide em clássicos (HLA-A, -B e -C) e não clássicos (HLA-E, HLA-F e HLA-G), e classe II (HLA-DRA, -DRB1, -DRB3, -DRB4, -DRB5, -DQA1, -DQB1, -DPA1, -DPB1). Os genes de classe I clássicos codificam proteínas expressas em todas as células somáticas nucleadas, sendo responsáveis por apresentar peptídeos antigênicos na superfície celular a linfócitos T citotóxicos (KLEIN; SATO, 2000; NEEFJES et al., 2011). Dentro dos genes HLA, os sítios que codificam a região da molécula à qual o antígeno é ligado são conhecidos como sítios de reconhecimento do antígeno (ARS - *Antigen Recognition Site*). Estudos apontam os sítios que codificam para o ARS como o principal alvo da seleção balanceadora dentro dos genes HLA de classe I clássicos (HEDRICK; WHITTAM; PARHAM, 1991; HUGHES, 1988; BITARELLO; FRANCISCO; MEYER, 2016). Sendo os outros sítios desses genes (aqui chamados de "não-ARS") codificadores para proteínas da região transmembrana e/ou da cauda citoplasmática (KLEIN; FIGUEROA, 1986; KLEIN et al., 1986; MEL-LOR, 1987) não tendo contato com o antígeno e não apresentando evidências de serem mantidos por seleção balanceadora (BITARELLO; FRANCISCO; MEYER, 2016).

Devido aos fatos dos genes clássicos do HLA de classe I 1) formarem um conjunto gênico para os quais há forte evidência de seleção balanceadora, 2) representarem genes para os quais temos informação sobre os sítios específicos que estão sob seleção e, 3) comporem a família gênica mais enriquecida para polimorfismos no genoma humano, eles são bons candidatos para estudar o acúmulo de SNPs deletérios sob o efeito de seleção balanceadora. Portanto, este capítulo tem como objetivo investigar se os éxons não-ARS de genes HLA de classe I apresentam acúmulo de SNPs deletérios por efeito de carona genética aos éxons sob seleção balanceadora em genes clássicos do HLA de classe I.

3.2 Dados e Métodos

3.2.1 Banco de dados

Para testar nossa hipótese de que a seleção balanceadora nos éxons ARS dos genes de HLA de classe I contribuiu para o acúmulo de SNPs deletérios nos éxons não-ARS, consideramos todas as posições de SNPs codificadores de proteínas (não-sinônimos, N e sinônimo, S) a partir dos arquivos VCF da *phase3* do projeto 1000 anotado para a versão hg19 do genoma humano (CONSORTIUM et al., 2015). Para testar se há efeito de carona genética em escala continental, estabelecemos grupos continentais a partir das populações que compõem o continente estudado, de maneira que o grupo europeu (EUR) foi formado a partir de 503 indivíduos das populações CEU (descendentes da Europa setentrional e ocidental vivendo em Utah, EUA), TSI (Toscana, Itália), FIN (Finlândia), GBR (britânicos da Inglaterra e Escócia) e IBS (Ibéricos da Espanha), o leste asiático (EAS) formado a partir de 503 indivíduos das populações CHB (grupo étnico Han de Pequim, China), JPT (Tóquio, Japão), CHS (grupo étnico Han do sul, China), CDX (grupo étnico Dai de Xishuangbanna, China) e KHV (grupo étnico Kinh de Ho Chi Minh, Vietnã), o sul asiático (SAS) formado a partir de 489 indivíduos das populações GIH (grupo étnico indiano Gujarati vivendo no Texas, EUA), PJL (indianos de Punjabi vivendo em Lahore, Paquistão), BEB (grupo étnico Bengali, Bangladesh), STU (grupo étnico Tamil do Sri Lanka vivendo no Reino Unido) e ITU (grupo étnico indiano Telugu vivendo no Reino Unido) e o africano (AFR) formado a partir de 504 indivíduos das populações YRI (grupo étnico Iorubá de Ibadan, Nigéria), LWK (grupo étnico Luhya de Webuye, Kenya), GWD (Gâmbia), MSL (Mende, Serra Leoa) e ESN (grupo étnico Esan, Nigéria). Excluímos as populações miscigenadas da América AMR (i.e. MXL (descendentes de mexicanos vivendo em Los Angeles, EUA), PUR (Porto Rico), CLM (Medellín, Colômbia), PEL (Lima, Peru)) e outras também miscigenadas segundo material suplementar do 1000 genomas (i.e. ASW (afro descendentes norte americanos, EUA), ACB (afro descendentes caribenhos, Barbados). Calculamos a frequência alélica (AF - *Allele Frequency*) para cada SNP nos continentes utilizando o programa *VCFtools* v0.1.12b (DANECEK et al.,

2011). Mantivemos em nossas análises apenas sítios bialélicos, sendo MAF a frequência do alelo menos frequente calculada com $1 - AF$ quando $AF > 0,5$.

3.2.2 Identificando os sítios não-ARS

Para delimitar a região alvo retiramos dos genes de HLA de classe I os éxons 2 e 3 dos genes HLA-(A, B e C) (Tabela 3.1), por serem esses que contém posições que codificam sítios ARS, responsáveis pela ligação de antígenos e alvo de seleção balanceadora. Para os sítios restantes, aqui denominados de não-ARS, não há evidência de função na apresentação de peptídeos ou seleção balanceadora. As coordenadas dos éxons ARS utilizadas nas análises deste capítulo foram adquiridas no UCSC *Table Browser* usando o *RefSeq Genes* em janeiro de 2017. Quando mais de um transcrito estava disponível no banco de dados, escolhemos o par de coordenadas mais inclusivo (i.e. menor começo de leitura e maior fim de leitura, gerando o maior transcrito possível). As coordenadas utilizadas referem-se aos transcritos com os seguintes IDs: RefSeq NM_002116 (HLA-A), NM_005514 (HLA-B), NM_002117 (HLA-C). A região alvo delimitada (i.e. HLA-A, B e C sem os éxons ARS) foi denominada de grupo não-ARS e todo o resto do genoma, de grupo controle.

3.2.3 Enriquecimento de polimorfismo nos sítios não-ARS

É bem documentado o enriquecimento de variabilidade nos genes do HLA acima do esperado sob neutralidade (MEYER; THOMSON, 2001; GARRIGAN; HEDRICK, 2003; SPURGIN; RICHARDSON, 2010). Devido à função biológica dos éxons ARS, que codificam a região da molécula responsável pela interação com o antígeno, há também um enriquecimento de SNPs com evidência de seleção balanceadora concentrado nesses éxons (Tabela 3.1). Por outro lado, um possível enriquecimento de SNPs também nos sítios não-ARS, que têm uma função biológica distinta daquela dos sítios ARS, deve ser interpretado como um sinal de que a carona genética está afetando a manutenção de SNPs nessa região. Assim, para verificar se o enriquecimento de polimorfismos nos genes do HLA está somente concentrado nos éxons ARS ou também nos não-ARS, realizamos uma análise

preliminar: elaboramos uma reamostragem de três genes sorteados ao acaso no genoma e contamos o número de SNPs presentes nesses genes. Repetimos essa reamostragem 10 mil vezes e verificamos em qual quantil se encontram os genes HLA não-ARS.

Tabela 3.1 – Número de SNPs nos genes HLA clássicos de classe I, com e sem os éxons ARS nas populações estudadas.

POP	Genes	n. de SNPs		
		Total	não-ARS	ARS
EUR	HLA-A	97	33	64
	HLA-B	71	23	48
	HLA-C	86	48	38
EAS	HLA-A	94	29	65
	HLA-B	77	23	54
	HLA-C	77	37	40
SAS	HLA-A	93	30	63
	HLA-B	70	23	47
	HLA-C	88	49	39
AFR	HLA-A	99	33	66
	HLA-B	81	27	54
	HLA-C	87	48	39
ALL	HLA-A	117	43	74
	HLA-B	96	27	59
	HLA-C	102	55	47

3.2.4 Quantificação de carga genética

Neste capítulo seguimos a mesma metodologia para quantificação de SNPs deletérios que aquela aplicada no capítulo 2, onde utilizamos as anotações do *Polyphen* (KIRCHER et al., 2014) que alocam os SNPs não sinônimos em três categorias funcionais: *Probably Damaging*, *Possibly Damaging* ou *Benign*, sendo o número total de SNPs em cada uma dessas classes, abreviada por P_{d1} , P_{d2} e P_b , respectivamente. Optamos por essa ferramenta de anotações pelo fato do *Polyphen* ser baseado em um naïve bayes machine learning que infere a probabilidade de um SNP ser deletério considerando em conjunto conservação filogenética e medidas empíricas de substituições que afetam a estrutura tridimensional

da proteína, enquanto anotações análogas (*Sift* ou *GERP*) consideram apenas os padrões de conservação filogenética. Segundo a classificação do *Polyphen*, SNPs classificados como *Probably Damaging* apresentam maior probabilidade de serem deletérios, e portanto utilizamos P_{d1} como estatística para quantificar o número de SNPs deletérios (abreviado por P_{del}). Outra estatística utilizada foi a divisão do número de SNPs deletérios pelo o número de SNPs neutros (razão abreviada aqui como P_{del}/P_{neu}). Para isso, calculamos o P_{del} nos grupos positivo e controle e dividimos pela contagem de todas as outras classes de SNPs ($P_{neu} = P_b + P_{d2} + PS$). Nem todos os SNPs não sinônimos estão classificados pelo *Polyphen*. Para garantirmos uma análise precisa, utilizamos apenas genes com uma proporção alta de SNPs classificados pelo *Polyphen*, especificamente genes que satisfaçam a condição $(P_b + P_{d1} + P_{d2})/PN > 0,95$, implicando que 95% de seus SNPs classificados pelo *Polyphen*.

3.2.5 Reamostragem em janelas sob seleção balanceadora

Como discutido no capítulo 1, há diferenças na proporção de mutações deletérias entre conjuntos de SNPs com distribuições de frequências diferentes, havendo menos deletérias entre os SNPs com alto MAF (FAY; WYCKOFF; WU, 2001; TENNESSEN et al., 2012; HENN et al., 2015). Como SNPs ARS são deslocados para MAFs altos, a comparação deles com uma amostra aleatória de SNPs do genoma (que seria enriquecida para SNPs com MAF baixo, mais comuns) consistiria numa comparação entre SFS diferentes. No entanto, como excluímos os sítios alvos de seleção balanceadora dentro dos genes clássicos do HLA (i.e. os éxons ARS), essa característica pode não ser mantida. Assim, como feito no capítulo anterior, realizamos uma análise do SFS da região sendo investigada, para avaliar se o seu SFS é suficientemente próximo do restante do genoma para permitir que a reamostragem seja feita ao acaso, ou se é necessário que haja um controle explícito para diferenças no SFS. Essa análise consistiu em verificar qual o padrão do SFS da região não-ARS dentro dos genes HLA sob estudo, permitindo definir qual método de reamostragem utilizar em nossas análises subsequentes, seguindo as orientações do primeiro capítulo.

3.3 Resultados

Na região não-ARS, o continente africano (AFR) apresentou o maior número de SNPs, seguido pelo continente europeu (EUR); na Ásia, vemos maior número de SNPs no sul (SAS) seguido pelo leste (EAS) (108, 104, 92 e 90 SNPs, respectivamente) (Tabela 3.2). O grupo formado por todas as populações amalgamadas (ALL) apresentou um número maior, 125 SNPs. Em relação aos genes HLA não-ARS estudados, se os considerarmos individualmente, o HLA-C apresentou em média o maior número de SNPs e o HLA-B em média o menor número de SNPs (Tabela 3.2).

Tabela 3.2 – Sumário das estatísticas utilizadas no capítulo 3 separado por genes do HLA não-ARS.

	Grupo Continental	nº de SNPs	P _{del}	P _{del} /P _{neu}
HLA-A	EUR	33	0	0,00
	EAS	29	2	0,07
	SAS	30	1	0,03
	AFR	33	1	0,03
	ALL	43	4	0,10
HLA-B	EUR	23	3	0,15
	EAS	24	3	0,14
	SAS	23	3	0,15
	AFR	27	4	0,17
	ALL	27	4	0,17
HLA-C	EUR	48	5	0,12
	EAS	37	1	0,03
	SAS	49	5	0,11
	AFR	48	5	0,12
	ALL	55	6	0,12

Ainda em relação ao número de SNPs, podemos notar que, em nossa análise preliminar, os três genes do HLA mostram enriquecimento de SNPs se comparados com uma distribuição do número de SNPs proveniente do sorteio de três genes quaisquer do genoma

(todos os grupos continentais apresentaram contagens de SNPs superior em pelo menos 99,3% dos sorteios; Figura 3.1).

Figura 3.1 – Maior número de SNPs da região não-ARS do gene HLA se comparada com uma distribuição do número de SNPs de três genes sorteados do genoma, para todos os grupos continentais: europeu (EUR), leste asiático (EAS), sul asiático (SAS), africano (AFR) e todos amalgamados (ALL). O grupo continental EAS apresenta o menor enriquecimento de SNPs, apresentando contagem de SNPs superior a 99,3% dos sorteios.

O enriquecimento de SNPs na região não-ARS se refletiu também no SFS, com um marcante desvio na direção de sítios em frequência intermediária, uma grande diferença em relação ao SFS de SNPs controle (Figura 3.2). Para controlar para efeitos de diferenças no SFS entre as regiões sob estudo e as controle, é necessário estabelecer uma metodologia que permita o controle desse fator. Nesse caso, como apresentado no primeiro capítulo, o método de reamostragem que controla para o padrão de SFS da região alvo (SFS-Gene-Boot) é o mais adequado. Esse método reamostra SNPs do genoma de modo a obter conjuntos de SNPs com SFS semelhante aos da região não-ARS dos genes do HLA, ao mesmo tempo que controla para que cada amostra tenha o mesmo número de SNPs e um números semelhantes de genes aos da região alvo.

Figura 3.2 – Espectro de Frequência Alélica (SFS) de SNPs codificantes dos grupos não-ARS e controle para os continentes africano (AFR), europeu (EUR), leste asiático (EAS), sul asiático (SAS) e todos amalgamados (ALL).

Definido o método de reamostragem pela análise do SFS, realizamos um teste de *outlier*, no qual o critério de significância estatística é obtido identificando o quantil da distribuição das reamostragens que possui medida de carga genética mais extrema que a esperada sob neutralidade. Os resultados apontam que os SNPs não-ARS desviaram significativamente do controle tanto para a estatística P_{del} como $P_{\text{del}}/P_{\text{neu}}$ no grupo composto por todas as populações (ALL) e para a estatística P_{del} para o continente africano (AFR) (Figura 3.3). Se considerarmos os grupos continentais separadamente, nossos resultados não indicam uma concentração de SNPs deletérios significativamente maior no grupo não-ARS (Figura 3.3). Embora não tenhamos feito testes, pode-se notar que, se considerarmos os HLA não-ARS separados, a carga genética, tanto por P_{del} quanto por $P_{\text{del}}/P_{\text{neu}}$, é maior nos genes HLA-B e HLA-C do que no gene HLA-A (Tabela 3.2).

Figura 3.3 – Análise de outlier para as estatísticas P_{del} e P_{del}/P_{neu} para o grupo de SNPs não-ARS de genes HLA clássico de classe I não-ARS para o continente africano (AFR), europeu (EUR) e asiático (EAS e SAS) e com as populações amalgamadas (ALL), utilizando o genoma de cada grupo como controle. Os valores acima de cada violin plot são os p-valores da respectiva análise de outlier. A cor verde indica os casos em que há significância seguindo o critério de valor crítico de 0,05, e vermelho indica não significância, pelo mesmo critério.

3.4 Discussão

Neste capítulo testamos a hipótese de que a seleção balanceadora nos éxons ARS de genes de HLA de classe I clássicos é responsável pelo aumento da carga genética nos sítios não-ARS desses mesmos genes. Para tanto, utilizamos um teste mais robusto do que tem sido utilizado na literatura, que garante que os sítios não-ARS e controle tenham o mesmo número de SNPs, e também SFS e desequilíbrio de ligação semelhantes.

3.4.1 Enriquecimento de SNPs por efeito carona

Nossos resultados indicam que os sítios não-ARS dos genes clássicos do HLA de classe I estão enriquecidos para SNPs se comparados com conjuntos aleatórios de três genes

reamostrados do genoma (Figura 3.1), sugerindo que a seleção balanceadora afeta o enriquecimento de SNPs nos sítios vizinhos a sítios selecionados, uma consequência da carona genética. Nessa análise não consideramos o tamanho dos genes, que poderia influenciar os resultados. Porém, um resultado enviesado seria esperado em genes grandes (i.e. genes possivelmente com mais SNPs) e os genes aqui estudados, mesmo inteiros, têm tamanhos muito menores (HLA-A: 4.625 bp, HLA-B: 3.316 bp, HLA-C: 3.338 bp) (HARROW et al., 2012) se comparados a média ou mediana genômica (67.008 e 26.775 bp, respectivamente) (HARROW et al., 2012; PIOVESAN et al., 2016). Assim, podemos observar que mesmo retirando os dois éxons que compõem os sítios ARS, nossos resultados sugerem um forte enriquecimento de SNPs nos genes do HLA.

3.4.2 Maior carga genética por efeito carona

Nossos resultados apontam aumento significativo de SNPs deletérios para as duas estatísticas testadas ($P_{\text{del}}/P_{\text{neu}}$ e P_{del}) no grupo formado por todas as populações amalgamadas (ALL) e para a estatística P_{del} na população AFR. Esses achados indicam que, para ao menos dois agrupamentos populacionais, há evidência de um enriquecimento de SNPs deletérios na região não-ARS por efeito de carona genética nos sítios ARS em genes do HLA classe I clássicos. O acúmulo desses SNPs nessas regiões indica a seleção balanceadora como força que contribuiu para o aumento da carga genética nessas populações humanas por meio do efeito de carona genética (HENN et al., 2015). Uma interpretação alternativa para esse resultado é que, na realidade, estamos diante de um excesso de SNPs funcionais sob seleção balanceadora e não sob efeito de carona genética provocada pela seleção balanceadora. Entretanto, não há, até o momento, evidências de que sítios não-ARS estão sob seleção balanceadora, o que indica que a primeira interpretação é a mais provável (HUGHES, 1988; HEDRICK; WHITTAM; PARHAM, 1991; BITARELLO; FRANCISCO; MEYER, 2016).

Um estudo recente, utilizando o método estatístico clássico que compara diretamente a região alvo com o restante do genoma, encontrou aumento na frequência de SNPs deletérios em mais de 100 genes vizinhos ao genes do HLA (LENZ et al., 2016). Outro

estudo (CÉSAR et al., em preparação), utilizando os mesmos grupos continentais e métodos deste trabalho, também apresentou resultados congruentes aos apresentados aqui. Por um lado, esse estudo indicou um aumento de carga genética em genes vizinhos aos HLA em africanos (AFR) e em populações amalgamadas (ALL) (CÉSAR et al., em preparação), o que reforça nossas conclusões discutidas anteriormente para esses dois grupos. Por outro lado, o estudo de César e colaboradores indicou que europeus (EUR) e leste asiáticos (EAS) não apresentam evidências do aumento de carga genética por efeito de carona genética em genes situados ao redor dos genes HLA. Sabendo que a força do efeito carona é uma função inversa da distância genética (CHARLESWORTH, 2006), a ausência de carga genética em genes vizinhos não implica em ausência de carga genética dentro do gene, onde os efeitos deveriam ser mais fortes na mesma direção. O fato de não encontrarmos evidências de carga genética mesmo na região mais próxima aos sítios selecionados sugere que em europeus (EUR) e leste asiáticos (EAS) a seleção balanceadora não é forte o suficiente para estabelecer um efeito de carona genética capaz de levar ao acúmulo SNPs deletérios de modo que seja estatisticamente significativo com os métodos e dados aqui usados.

No entanto, diferentemente deste trabalho, César e colaboradores (em preparação) encontraram acúmulo de SNPs deletérios nos genes vizinhos ao HLA em sul asiáticos (SAS), tornando surpreendente o fato de não termos encontrado assinatura similar nas nossas análises. A diferença entre os resultados é inesperada pois quanto mais próximo ao sítio selecionado, mais forte é o efeito de carona genética. Entretanto, há possíveis explicações para as diferenças entre os achados. Devido à importância funcional dos genes do HLA, é possível que a restrição para SNPs deletérios seja maior dentro desses genes. De acordo com essa hipótese, a força de seleção balanceadora nos sítios ARS estaria gerando um acúmulo maior de SNPs deletérios em genes fora do HLA do que dentro dos próprios genes do HLA em sua porção não-ARS, onde a seleção purificadora é mais intensa. Um resultado que aponta para essa direção é a comparação entre os valores de P_{del}/P_{neu} encontrados por este estudo e os encontrados por César e colaboradores (em preparação; Tabela S3.1). Podemos notar que em todos os grupos continentais, os valores de P_{neu}/P_{neu} são menores nas regiões não-ARS dos genes HLA do que nos genes vizinhos ao HLA, sugerindo

que embora tenhamos trabalhado com sítios mais próximos aos selecionados, também trabalhamos com genes com maiores restrições funcionais. Nesse caso, dos três grupos continentais que apresentaram acúmulo de SNPs deletérios em César e colaboradores (em preparação), somente o sul asiático (SAS) não apresentou tal acúmulo. É possível que a restrição sobre o acúmulo de SNPs deletérios seja mais forte no grupo continental SAS.

Uma interpretação alternativa para esse resultado pode ser que a ausência de apoio para uma carga genética aumentada em exons não-ARS seja consequência do baixo número de SNPs que estamos investigando nesta análise. Ao trabalharmos apenas com três genes, mesmo enriquecidos por polimorfismos, nos limitamos a um baixo número de SNPs e, neste caso, todos os grupos continentais estão sujeitos a baixo poder estatístico em decorrência desse reduzido número de SNPs. De fato, os grupos continentais asiáticos são os que apresentam menor quantidade de SNPs em nossos dados, sendo portanto, os com menor poder para detectar um eventual excesso de polimorfismos deletérios. Porém, o fato da carga genética dos sítios não-ARS ser congruente com a carga genética encontrada na adjacência de genes do HLA para todos os grupos continentais, com exceção do SAS, aponta para a possibilidade da ausência de significância dos desvios que encontramos ser uma particularidade do grupo continental SAS. Neste trabalho não foi possível verificar se alguma dessas alternativas levantadas explica adequadamente o resultado que encontramos para o grupo continental SAS.

3.4.3 Genes sob seleção balanceadora

Neste terceiro capítulo nossa estratégia foi explorar os genes de HLA classe I clássicos, exemplos bem documentados de seleção balanceadora, altamente polimórficos e com informações prévias sobre os alvos de seleção balanceadora dentro dos genes, assim, podemos testar nossas previsões utilizando um número reduzido de genes.

Mesmo considerando o enriquecimento de polimorfismos nessas regiões, conseguimos apontar o efeito da carona genética apenas nos grupos continentais com maiores números de SNPs (AFR e ALL). Estudos futuros com objetivos semelhantes podem utilizar a

mesma abordagem apresentada neste trabalho com um número maior de genes sob seleção balanceadora, aumentando o poder estatístico da análise.

Embora a seleção balanceadora tenha sido amplamente explorada a partir de uma perspectiva teórica, ainda se tem pouco conhecimento sobre sua prevalência em estudos empíricos. Isso decorre em parte da dificuldade de detecção de suas assinaturas, e em parte por se tratar de um evento seletivo comparativamente raro (ASTHANA; SCHMIDT; SUNYAEV, 2005; BUBB et al., 2006; ALONSO et al., 2008; ANDRÉS et al., 2009; BITARELLO et al., 2018). Porém, estudos recentes têm apresentado listas contendo centenas de genes candidatos a seleção balanceadora (DEGIORGIO; LOHMUELLER; NIELSEN, 2014; SIEWERT; VOIGHT, 2017; BITARELLO et al., 2018). Como esses estudos recentes ainda não apresentam informações sobre os alvos de seleção balanceadora dentro desses genes, optamos por trabalhar apenas com os genes HLA de classe I clássicos. É bem provável que estudos futuros, que tenham essa informação, possam apresentar resultados mais precisos na mesma direção que o nosso.

3.4.4 Conclusão

Neste capítulo, mostramos o acúmulo de SNPs deletérios nos éxons não-ARS de genes clássicos do HLA de classe I por efeito carona aos éxons ARS em populações continentais amalgamadas e em africanos. O acúmulo desses SNPs contribui para o aumento da carga genética nessas populações. Apontamos que para se obter resultados mais robustos em populações separadas ainda há a necessidade de novos estudos que visem encontrar mais genes sob efeito de seleção balanceadora e os alvos de seleção dentro desses genes.

Considerações Finais

Nesta dissertação exploramos o aumento da carga genética em populações humanas como produto de dois processos adaptativos distintos. Nossos resultados mostram que tanto a seleção positiva como a seleção balanceadora foram fundamentais para reduzir a eficiência da seleção purificadora em retirar SNPs deletérios, aumentando a carga genética em humanos. Diferentemente de processos demográficos, todos os processos seletivos afetam regiões específicas do genoma e a força de seleção sobre a dinâmica de regiões vizinhas é uma função inversa da distância entre elas e os sítios selecionados. Sendo assim, o aumento da carga genética por efeito da seleção sob sítios vantajosos é restrito às regiões selecionadas e suas adjacências. Nossos resultados mostram que tanto a seleção positiva como a seleção balanceadora foram fundamentais para aumentar a carga genética em regiões específicas do genoma. No entanto, como esses regimes seletivos não são eventos comuns no genoma, não é possível com esse trabalho, e, tampouco foi o objetivo, dizer que a seleção natural vantajosa é o processo evolutivo mais impactante para a diferença entre as cargas genéticas em todo genoma humano.

Diferente desequilíbrio de ligação entre regiões genômicas

Como indicado no capítulo 1, tanto o capítulo 2 quanto o capítulo 3 exploraram regiões específicas do genoma, com SNPs contíguos dentro dos genes escolhidos por estarem sob regimes de seleção específicos. Dessa forma, todas as reamostragens realizadas ao longo deste estudo controlaram para o números aproximados de genes, visando comparar

blocos de SNPs com desequilíbrios de ligação semelhantes. A comparação de regiões com diferentes desequilíbrios de ligação se mostrou equivocada, uma vez que, reduziu a variância da distribuição reamostrada resultando em falsos positivos.

Diferentes assinaturas de seleção e diferentes reamostragens

Para responder às questões sobre carga genética dentro dos grupos continentais optamos por uma análise baseada em reamostragem. O primeiro desafio de se trabalhar com dois processos seletivos distintos, mantendo a metodologia de reamostragem para ambos, é determinar as particularidades de cada processo que devem ser consideradas durante a reamostragem.

A seleção positiva e seleção balanceadora são dois processos que aumentam a aptidão do indivíduo portador do SNP ou genótipo vantajoso, mas apresentam inúmeras diferenças nos sinais deixados no genoma. Em especial, como desvios no SFS por si só trazem mudanças na quantidade de mutações deletérias (havendo menos deletérias entre conjuntos de polimorfismos comuns), o padrão do SFS dessas regiões foi fundamental para a escolha do método de reamostragem utilizado no capítulo 2 (STD-Gene-Boot) e do método utilizado no capítulo 3 (SFS-Gene-Boot). Controlamos só para o que potencialmente difere entre região alvo e controle. No capítulo 2, não se mostrou necessário reamostrar controlando para que as regiões apresentassem o mesmo SFS. Já no capítulo 3, a região alvo sob efeito de seleção balanceadora apresentou o SFS desviado para frequências intermediárias, muito diferente do padrão encontrado no genoma. Para evitar que nosso controle tenham um excesso de SNPs deletérios simplesmente por ter mais variantes raras, controlamos para que a região alvo e controle apresentassem o mesmo SFS.

A partir dessas aplicações da metodologia proposta no capítulo 1 nos capítulos 2 e 3, demonstramos sua robustez e flexibilidade para responder perguntas relacionadas à carga genética em uma perspectiva populacional, e ainda, apontamos a importância de considerar as particularidades deixadas por diferentes regimes de seleção natural na implementação da reamostragem, reduzindo a probabilidade de falsos positivos.

Referências

A LICHTMAN A, P. J. A. *Cellular And Molecular Immunology*. 4th editio. ed. Philadelphia: [s.n.], 2000.

AKEY, J. M. Constructing genomic maps of positive selection in humans: Where do we go from here? **Genome research**, v. 19, n. 206, p. 711–722, 2009.

AKEY, J. M. et al. Population history and natural selection shape patterns of genetic variation in 132 genes. **PLoS biology**, v. 2, n. 10, p. e286, 10 2004. ISSN 1545-7885. Disponível em: <<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=515367&tool=pmcentrez&rendertype=ab>>.

_____. Interrogating a high-desity SNP map for signatures of natural selection. **Genome Research**, v. 12, p. 1805–1814, 2002. ISSN 10889051.

ALONSO, S. et al. Overdominance in the human genome and olfactory receptor activity. **Molecular biology and evolution**, Oxford University Press, doi = 10.1093/molbev/msn049, v. 25, n. 5, p. 997–1001, 2008.

ALVES, I. et al. Genomic data reveal a complex making of humans. **PLoS Genetics**, v. 8, n. 7, 2012. ISSN 15537390.

ANDRÉS, A. M. Balancing selection in the human genome. **Encyclopedia of life sciences**, Wiley Online Library, 2011.

ANDRÉS, A. M. et al. Targets of balancing selection in the human genome. **Molecular biology and evolution**, Oxford University Press, v. 26, n. 12, p. 2755–2764, 2009.

ASTHANA, S.; SCHMIDT, S.; SUNYAEV, S. A limited role for balancing selection. **Trends in Genetics**, Elsevier, v. 21, n. 1, p. 30–32, 2005.

BALARESQUE, P. L.; BALLEREAU, S. J.; JOBLING, M. a. Challenges in human genetic diversity: demographic history and adaptation. **Human molecular genetics**, v. 16 Spec No, n. 2, p. 134–9, 10 2007. ISSN 0964-6906. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/17911157>>.

BERRY, A. J.; AJIOKA, J.; KREITMAN, M. Lack of polymorphism on the drosophila fourth chromosome resulting from selection. **Genetics**, Genetics Soc America, v. 129, n. 4, p. 1111–1117, 1991.

- BERSAGLIERI, T. et al. Genetic signatures of strong recent positive selection at the lactase gene. **The American Journal of Human Genetics**, Elsevier, v. 74, n. 6, p. 1111–1120, 2004.
- BETANCOURT, A. J.; PRESGRAVES, D. C. Linkage limits the power of natural selection in drosophila. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 99, n. 21, p. 13616–13620, 2002.
- BITARELLO, B. D. et al. Signatures of long-term balancing selection in human genomes. **Genome biology and evolution**, Oxford University Press, v. 10, n. 3, p. 939–955, 2018.
- BITARELLO, B. D.; FRANCISCO, R. dos S.; MEYER, D. Heterogeneity of dn/ds ratios at the classical hla class i genes over divergence time and across the allelic phylogeny. **Journal of molecular evolution**, Springer, v. 82, n. 1, p. 38–50, 2016.
- BUBB, K. L. et al. Scan of human genome reveals no new loci under ancient balancing selection. **Genetics**, Genetics Soc America, v. 173, n. 4, p. 2165–2177, 2006.
- CÉSAR, J. et al. The burden of selection at hla genes over their genetic neighborhood. em preparação.
- CHARLESWORTH, B. Background selection 20 years on: the wilhelmine e. key 2012 invitational lecture. **Journal of Heredity**, Oxford University Press US, v. 104, n. 2, p. 161–171, 2013.
- CHARLESWORTH, B.; CHARLESWORTH, D. **Elements of evolutionary genetics**. [S.l.]: Roberts and Company Publishers Greenwood Village, CO, 2010. v. 42.
- CHARLESWORTH, D. Balancing selection and its effects on sequences in nearby genome regions. **PLoS genetics**, Public Library of Science, v. 2, n. 4, p. e64, 2006.
- CHUN, S.; FAY, J. C. Evidence for hitchhiking of deleterious mutations within the human genome. **PLoS Genetics**, v. 7, n. 8, 2011. ISSN 15537390.
- COLONNA, V. et al. Human genomic regions with exceptionally high levels of population differentiation identified from 911 whole-genome sequences. **Genome biology**, BioMed Central, v. 15, n. 6, p. 1, 2014.
- CONSORTIUM, . G. P. et al. An integrated map of genetic variation from 1,092 human genomes. **Nature**, Nature Publishing Group, v. 491, n. 7422, p. 56, 2012.
- _____. A global reference for human genetic variation. **Nature**, Nature Publishing Group, v. 526, n. 7571, p. 68, 2015.
- CORBETT-DETIG, R. B.; HARTL, D. L.; SACKTON, T. B. Natural Selection Constrains Neutral Diversity across A Wide Range of Species. **PLOS Biology**, v. 13, n. 4, p. e1002112, 2015. ISSN 1545-7885. Disponível em: <<http://dx.plos.org/10.1371/journal.pbio.1002112>>.
- COVENTRY, A. et al. Deep resequencing reveals excess rare recent variants consistent with explosive population growth. **Nature communications**, Nature Publishing Group, v. 1, p. 131–136, 2010.

- CUTTER, A. D.; PAYSEUR, B. A. Genomic signatures of selection at linked sites: unifying the disparity among species. **Nature Reviews Genetics**, Nature Publishing Group, v. 14, n. 4, p. 262, 2013.
- DANECEK, P. et al. The variant call format and vcftools. **Bioinformatics**, Oxford University Press, v. 27, n. 15, p. 2156–2158, 2011.
- DEGIORGIO, M.; LOHMUELLER, K. E.; NIELSEN, R. A model-based approach for identifying signatures of ancient balancing selection in genetic data. **PLoS genetics**, Public Library of Science, v. 10, n. 8, p. e1004561, 2014.
- DO, R. et al. No evidence that selection has been less effective at removing deleterious mutations in europeans than in africans. **Nature genetics**, Nature Publishing Group, v. 47, n. 2, p. 126–131, 2015.
- EFRON, B.; TIBSHIRANI, R. J. **An introduction to the bootstrap**. [S.l.]: CRC press, 1994.
- EIBERG, H. et al. Blue eye color in humans may be caused by a perfectly associated founder mutation in a regulatory element located within the herc2 gene inhibiting oca2 expression. **Human genetics**, Springer, v. 123, n. 2, p. 177–187, 2008.
- EYRE-WALKER, A.; KEIGHTLEY, P. D. The distribution of fitness effects of new mutations. **Nature Reviews Genetics**, Nature Publishing Group, v. 8, n. 8, p. 610–618, 2007.
- FAY, J. C.; WYCKOFF, G. J.; WU, C.-I. Positive and negative selection on the human genome. **Genetics**, Genetics Soc America, v. 158, n. 3, p. 1227–1234, 2001.
- FU, W. et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. **Nature**, v. 493, n. 7431, p. 216–220, 2012. ISSN 0028-0836. Disponível em: <<http://www.nature.com/doifinder/10.1038/nature11690>>.
- GARRIGAN, D.; HEDRICK, P. W. Perspective: detecting adaptive molecular polymorphism: lessons from the mhc. **Evolution**, BioOne, v. 57, n. 8, p. 1707–1722, 2003.
- GAZAVE, E. et al. Population growth inflates the per-individual number of deleterious mutations and reduces their mean effect. **Genetics**, v. 195, n. 3, p. 969–978, 2013. ISSN 00166731.
- GIBSON, G. Rare and common variants: twenty arguments. **Nature Reviews Genetics**, Nature Publishing Group, v. 13, n. 2, p. 135–145, 2012. ISSN 1471-0056. Disponível em: <<http://www.nature.com/doifinder/10.1038/nrg3118>>.
- GRAVEL, S. When is selection effective? **Genetics**, Genetics Soc America, v. 203, n. 1, p. 451–462, 2016.
- GROSSMAN, S. R. et al. Identifying recent adaptations in large-scale genomic data. **Cell**, Elsevier, v. 152, n. 4, p. 703–713, 2013.
- _____. A composite of multiple signals distinguishes causal variants in regions of positive selection. **Science**, American Association for the Advancement of Science, v. 327, n. 5967, p. 883–886, 2010.

- HALDANE, J. B. S. The cost of natural selection. **Journal of Genetics**, Springer, v. 55, n. 3, p. 511–524, 1957.
- HARROW, J. et al. Gencode: the reference human genome annotation for the encode project. **Genome research**, Cold Spring Harbor Lab, v. 22, n. 9, p. 1760–1774, 2012.
- HEDRICK, P. W.; WHITTAM, T. S.; PARHAM, P. Heterozygosity at individual amino acid sites: extremely high levels for hla-a and-b genes. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 88, n. 13, p. 5897–5901, 1991.
- HENN, B. M. et al. Estimating the mutation load in human genomes. **Nature Reviews Genetics**, Nature Publishing Group, v. 16, n. 6, p. 333–343, 2015.
- HERNANDEZ, R. D. et al. Classic selective sweeps were rare in recent human evolution. **science**, American Association for the Advancement of Science, v. 331, n. 6019, p. 920–924, 2011.
- HOLSINGER, K. E.; WEIR, B. S. Genetics in geographically structured populations: defining, estimating and interpreting F_{ST} . **Nature reviews. Genetics**, v. 10, n. 9, p. 639–650, 2009. ISSN 1471-0056.
- HUGHES, A. L. Major histocompatibility complex (mhc) genes: evolution. **eLS**, Wiley Online Library, 2008.
- HUGHES, M. N. A. L. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. **Nature**, v. 335, p. 167 – 170, 1988.
- KAPLAN, N. L.; DARDEN, T.; HUDSON, R. R. The coalescent process in models with selection. **Genetics**, Genetics Soc America, v. 120, n. 3, p. 819–829, 1988.
- KEINAN, A.; CLARK, A. G. Recent explosive human population growth has resulted in an excess of rare genetic variants. **Science**, American Association for the Advancement of Science, v. 336, n. 6082, p. 740–743, 2012.
- KIMURA, M.; MARUYAMA, T.; CROW, J. F. The mutation load in small populations. **Genetics**, Genetics Society of America, v. 48, n. 10, p. 1303–1312, 1963.
- KIRCHER, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. **Nature genetics**, Nature Research, v. 46, n. 3, p. 310–315, 2014.
- KLEIN, J.; FIGUEROA, F. Evolution of the major histocompatibility complex. **Critical reviews in immunology**, v. 6, n. 4, p. 295–386, 1986.
- KLEIN, J. et al. **Natural history of the major histocompatibility complex**. [S.l.]: Wiley, 1986.
- KLEIN, J.; SATO, A. The hla system. **New England Journal of Medicine**, Mass Medical Soc, v. 343, n. 10, p. 702–709, 2000.
- KOCH, E.; NOVEMBRE, J. A temporal perspective on the interplay of demography and selection on deleterious variation in humans. **G3: Genes, Genomes, Genetics**, G3: Genes, Genomes, Genetics, v. 7, n. 3, p. 1027–1037, 2017.

- LAMASON, R. L. et al. Slc24a5, a putative cation exchanger, affects pigmentation in zebrafish and humans. **Science**, American Association for the Advancement of Science, v. 310, n. 5755, p. 1782–1786, 2005.
- LAVAL, G. et al. Formulating a historical and demographic model of recent human evolution based on resequencing data from noncoding regions. **PloS one**, Public Library of Science, v. 5, n. 4, p. e10284, 2010.
- LENZ, T. L. et al. Excess of Deleterious Mutations around HLA Genes Reveals Evolutionary Cost of Balancing Selection. p. 1–30, 2016.
- LOHMUELLER, K. E. The distribution of deleterious genetic variation in human populations. **Current opinion in genetics & development**, Elsevier, v. 29, p. 139–146, 2014.
- _____. The impact of population demography and selection on the genetic architecture of complex traits. **PLoS genetics**, Public Library of Science, v. 10, n. 5, p. e1004379, 2014.
- LOHMUELLER, K. E. et al. Proportionally more deleterious genetic variation in european than in african populations. **Nature**, Nature Publishing Group, v. 451, n. 7181, p. 994–997, 2008.
- LU, J. et al. The accumulation of deleterious mutations in rice genomes: a hypothesis on the cost of domestication. **Trends in Genetics**, Elsevier, v. 22, n. 3, p. 126–131, 2006.
- MARTH, G. T. et al. The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. **Genetics**, Genetics Soc America, v. 166, n. 1, p. 351–372, 2004.
- _____. The functional spectrum of low-frequency coding variation. **Genome biology**, BioMed Central, v. 12, n. 9, p. R84, 2011.
- MELLOR, A. L. Analysis of major histocompatibility complex class i gene transcription using oligonucleotide probes. **Molecular and cellular probes**, Elsevier, v. 1, n. 3, p. 207–220, 1987.
- MESSER, P. W.; PETROV, D. A. Population genomics of rapid adaptation by soft selective sweeps. **Trends in Ecology & Evolution**, Elsevier Ltd, v. 28, n. 11, p. 659–669, 2013. ISSN 01695347. Disponível em: <<http://linkinghub.elsevier.com/retrieve/pii/S0169534713002073>>.
- MEYER, D. et al. A genomic perspective on hla evolution. **Immunogenetics**, v. 70, n. 1.
- MEYER, D.; THOMSON, G. How selection shapes variation of the human major histocompatibility complex: a review. **Annals of human genetics**, Cambridge University Press, v. 65, n. 1, p. 1–26, 2001.
- MEZMOUK, S.; ROSS-IBARRA, J. The pattern and distribution of deleterious mutations in maize. **G3: Genes, Genomes, Genetics**, G3: Genes, Genomes, Genetics, v. 4, n. 1, p. 163–171, 2014.

- NEEFJES, J. et al. Towards a systems understanding of mhc class i and mhc class ii antigen presentation. **Nature Reviews Immunology**, Nature Publishing Group, v. 11, n. 12, p. 823, 2011.
- NELSON, M. R. et al. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. **Science**, American Association for the Advancement of Science, v. 337, p. 100–104, 2012.
- NIELSEN, R. Molecular signatures of natural selection. **Annual review of genetics**, v. 39, p. 197–218, 1 2005. ISSN 0066-4197. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/16285858>>.
- NIELSEN, R. et al. Recent and ongoing selection in the human genome. **Nature Reviews Genetics**, Nature Publishing Group, v. 8, n. 11, p. 857, 2007.
- NORTON, H. L. et al. Genetic evidence for the convergent evolution of light skin in europeans and east asians. **Molecular biology and evolution**, Oxford University Press, v. 24, n. 3, p. 710–722, 2007.
- PIGANEAU, G.; EYRE-WALKER, A. Evidence for variation in the effective population size of animal mitochondrial DNA. **PloS one**, v. 4, n. 2, p. e4396, 1 2009. ISSN 1932-6203. Disponível em: <<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2635931&tool=pmcentrez&rendertype=ab>>.
- PIOVESAN, A. et al. Genebase 1.1: a tool to summarize data from ncbi gene datasets and its application to an update of human gene statistics. **Database**, Oxford University Press, v. 2016, 2016.
- PRUGNOLLE, F. et al. Pathogen-driven selection and worldwide HLA class I diversity. **Current Biology**, v. 15, n. 11, p. 1022–1027, 2005. ISSN 09609822.
- PRZEWORSKI, M.; HUDSON, R. R.; RIENZO, A. D. Adjusting the focus on human variation. **Trends in Genetics**, Elsevier, v. 16, n. 7, p. 296–302, 2000.
- RAMACHANDRAN, S. et al. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in africa. **Proceedings of the National Academy of Sciences of the United States of America**, National Acad Sciences, v. 102, n. 44, p. 15942–15947, 2005.
- REICH, D. E. et al. Linkage disequilibrium in the human genome. **Nature**, Nature Publishing Group, v. 411, n. 6834, p. 199–204, 2001.
- SABETI, P. C. et al. Positive natural selection in the human lineage. **science**, American Association for the Advancement of Science, v. 312, n. 5780, p. 1614–1620, 2006.
- SCHEINFELDT, L. B.; TISHKOFF, S. A. Recent human adaptation: genomic approaches, interpretation and insights. **Nature reviews Genetics**, v. 14, n. 10, p. 692–702, 2013. ISSN 1471-0064. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/24052086>>.
- SCHRAIBER, J. G.; AKEY, J. M. Methods and models for unravelling human evolutionary history. **Nature Reviews Genetics**, Nature Publishing Group, v. 16, n. 12, p. 727, 2015.

- SIEWERT, K. M.; VOIGHT, B. F. Detecting long-term balancing selection using allele frequency correlation. **Molecular biology and evolution**, Oxford University Press, v. 34, n. 11, p. 2996–3005, 2017.
- SIMONS, Y. B. et al. The deleterious mutation load is insensitive to recent population history. **Nature genetics**, NIH Public Access, v. 46, n. 3, p. 220, 2014.
- SIMONSEN, K. L.; CHURCHILL, G. A.; AQUADRO, C. F. Properties of statistical tests of neutrality for dna polymorphism data. **Genetics**, Genetics Soc America, v. 141, n. 1, p. 413–429, 1995.
- SMITH, J. M.; HAIGH, J. The hitch-hiking effect of a favourable gene. **Genetics Research**, Cambridge University Press, v. 23, n. 1, p. 23–35, 1974.
- SPURGIN, L. G.; RICHARDSON, D. S. How pathogens drive genetic diversity: Mhc, mechanisms and misunderstandings. **Proceedings of the Royal Society of London B: Biological Sciences**, The Royal Society, p. rspb20092084, 2010.
- STAJICH, J. E.; HAHN, M. W. Disentangling the effects of demography and selection in human history. **Molecular Biology and Evolution**, Oxford University Press, v. 22, n. 1, p. 63–73, 2004.
- STURM, R. A. et al. A single snp in an evolutionary conserved region within intron 86 of the herc2 gene determines human blue-brown eye color. **The American Journal of Human Genetics**, Elsevier, v. 82, n. 2, p. 424–431, 2008.
- TENNESSEN, J. A. et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. **Science (New York, N.Y.)**, v. 337, n. 6090, p. 64–9, 2012. ISSN 1095-9203. Disponível em: <<http://www.sciencemag.org/content/337/6090/64.long>>. <http://www.sciencemag.org/content/337/6090/64.long>.
- VEERAMAH, K. R.; HAMMER, M. F. The impact of whole-genome sequencing on the reconstruction of human population history. **Nature reviews Genetics**, Nature Publishing Group, v. 15, n. 3, p. 149–62, 2014. ISSN 1471-0064. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/24492235>>.
- VOIGHT, B. F. et al. A Map of Recent Positive Selection in the Human Genome. **PLoS biology**, Public Library of Science, v. 4, n. 3, p. e72, 2006.
- WEIR, B. S. et al. Measures of human population structure show heterogeneity among genomic regions. **Genome research**, Cold Spring Harbor Lab, v. 15, n. 11, p. 1468–1476, 2005.
- ZHANG, M. et al. Recombination rate variation, hitchhiking, and demographic history shape deleterious load in poplar. **Molecular Biology and Evolution**, SBE, v. 33, n. 11, p. 2899–2910, 2016.

Material Suplementar do capítulo 2

Figura S2.1 – *Representação esquemática dos critérios de inclusão de genes na análise. A cor verde indica genes incluídos na análises: (1) genes que continham SNPs HighD e (2) genes sem SNPs HighD, mas que começam a no máximo uma distância A e terminam no máximo a uma distância B do SNP HighD. A cor vermelha indica genes que foram excluídos da análise: Genes sem SNPs HighD que começam a uma distância menor que A, mas terminam a uma distância maior que B. A e B são distâncias em kb partindo do SNPs HighD e variam nos continentes testados, sendo $A = 125$ e $B = 135$ nos europeus e $A = 225$ e $B = 235$ nos leste asiáticos. Sendo A, a janela definida pela heterozigosidade significativamente menor do que a média genômica. .*

Tabela S2.1 – Sumário das estatísticas do capítulo 2 separado por genes do grupo positivo para o grupo continental europeu (EUR).

	nº de SNPs	P _{del}	P _{del} /P _{neu}
NAV2	55	15	0,38
ABCA12	45	4	0,10
ALK	36	6	0,20
CDC42BPG	34	8	0,31
CNTNAP2	29	5	0,21
CNTNAP4	28	4	0,17
TLR10	26	3	0,13
GTSE1	26	3	0,13
TLR6	22	1	0,05
PYGM	21	6	0,40
NECAB2	21	6	0,40
TLR1	20	6	0,43
ARHGAP10	19	5	0,36
SLC38A8	19	4	0,27
GALNT5	18	5	0,38
MBTPS1	18	1	0,06
GGA3	18	6	0,50
OSGIN1	17	4	0,31
TTC38	17	0	0,00
LRRC4B	14	0	0,00
RASGRP2	13	0	0,00
MAP4K2	13	4	0,44
ZEB1	12	3	0,33
HIP1	11	3	0,38
SMYD3	10	2	0,25
CYP26B1	10	0	0,00
AGAP1	10	1	0,11
SLC45A2	10	4	0,67
MTFR2	10	2	0,25
MEN1	10	0	0,00
MLYCD	10	0	0,00
ASPDH	10	0	0,00
TRMU	9	3	0,50
FOXP1	8	2	0,33
SYT3	8	0	0,00
ATP1A1	7	0	0,00
MTHFD2L	7	1	0,17

Tabela S2.1 – Continuação do sumário das estatísticas do capítulo 2 separado por genes do grupo positivo para o grupo continental europeu (EUR).

	nº de SNPs	P _{del}	P _{del} /P _{neu}
RCBTB2	7	0	0,00
AVEN	7	2	0,40
MRPS7	7	2	0,40
PDE7B	6	1	0,20
ATP6V1H	6	1	0,20
SLC25A19	6	1	0,20
CD226	6	1	0,20
FAM71E1	6	0	0,00
C19orf81	6	2	0,50
LPAR6	5	3	1,50
CHRM5	5	1	0,25
TPRG1	4	0	0,00
EHD1	4	0	0,00
MIF4GD	4	0	0,00
JOSD2	4	0	0,00
ERMN	3	1	0,50
KLF3	3	0	0,00
CXCL2	3	0	0,00
SLC24A5	2	0	0,00
AC006372.1	1	0	0,00

Tabela S2.2 – Sumário das estatísticas do capítulo 2 separado por genes do grupo positivo para o grupo continental leste asiático (EAS).

	nº de SNPs	P _{del}	P _{del} /P _{neu}
HMCN1	120	33	0,38
VWF	83	12	0,17
TENM4	57	5	0,10
PKDREJ	50	2	0,04
CASKIN2	37	6	0,19
MAGI1	31	7	0,30
POM121C	30	4	0,15
DBH	29	8	0,38
LLGL2	29	3	0,12
SARDH	27	3	0,13
VAV2	26	2	0,08
GTSE1	26	3	0,13
TSEN54	25	4	0,20
MTRR	24	4	0,20
JAG2	24	3	0,14
PACS2	24	2	0,09
KIAA0195	24	3	0,14
ADCY2	23	2	0,10
SBNO1	19	0	0,00
MYO15B	19	0	0,00
ASAP2	18	1	0,06
GGA3	18	6	0,50
BRF1	17	1	0,06
CDH7	17	1	0,06
TTC38	17	0	0,00
ADAM17	16	1	0,07
ENTPD6	16	5	0,45
MX2	16	2	0,14
DDX55	15	3	0,25
TBC1D4	15	2	0,15
MSH4	14	1	0,08
FASTKD3	13	2	0,18
TMEM200C	13	0	0,00
BEND4	12	2	0,20
GPR132	12	0	0,00
BCAS3	12	3	0,33
NUP85	12	0	0,00

Tabela S2.2 – Continuação do sumário das estatísticas do capítulo 2 separado por genes do grupo positivo para o grupo continental leste asiático (EAS).

	nº de SNPs	P _{del}	P _{del} /P _{neu}
VSX1	12	3	0,33
EXOC6B	11	2	0,22
IKZF2	11	3	0,38
SLC30A9	11	0	0,00
HIP1	11	3	0,38
C14orf105	11	4	0,57
APMAP	11	1	0,10
ACADM	10	0	0,00
TRMT11	10	2	0,25
XKR6	10	3	0,43
SNRNP35	10	0	0,00
TBX2	10	1	0,11
PPARA	10	0	0,00
C5orf49	9	2	0,29
NCOA7	9	1	0,13
ADAMTSL2	9	0	0,00
C14orf79	9	1	0,13
SLC35F3	8	1	0,14
STAG1	8	0	0,00
SLC35G2	8	2	0,33
RILPL1	8	2	0,33
CDCA4	8	0	0,00
HEY2	7	0	0,00
NUDT14	7	0	0,00
DYX1C1	7	4	1,33
C17orf82	7	0	0,00
MRPS7	7	2	0,40
L3MBTL4	7	1	0,17
NCK1	6	0	0,00
FAM114A2	6	1	0,20
HINT3	6	0	0,00
SLC25A19	6	1	0,20
SETD8	5	0	0,00
GTF2H3	5	0	0,00
AP5M1	5	0	0,00
RP11-178L8.4	5	3	1,50
FBXO31	5	1	0,25

Tabela S2.2 – Continuação do sumário das estatísticas do capítulo 2 separado por genes do grupo positivo para o grupo continental leste asiático (EAS).

	nº de SNPs	P _{del}	P _{del} /P _{neu}
FAM3B	5	1	0,25
CDPF1	5	0	0,00
ASB17	4	0	0,00
MFAP3	4	0	0,00
TRIM73	4	0	0,00
MIF4GD	4	0	0,00
RABGGTB	3	1	0,50
ITGB1BP1	3	1	0,50
TMEM33	3	0	0,00
EXOC5	3	0	0,00
NAA30	3	0	0,00
AF131215.5	2	0	0,00
EIF2B1	2	0	0,00
UCHL3	2	0	0,00
DOK5	2	0	0,00
AL391152.1	1	0	0,00
MAP1LC3B	1	0	0,00

Material Suplementar do capítulo 3

Tabela S3.1 – Tabela com a comparação dos valores da estatística $P_{\text{del}}/P_{\text{neu}}$ encontrados neste trabalho para a região não-ARS do HLA e encontrados por César e colaboradores (em preparação) para os genes vizinhos do HLA para os diversos grupos continentais: europeu (EUR), leste asiático (EAS), sul asiático (SAS), africano (AFR) e todas as populações em conjunto (ALL).

	não-ARS	César e colaboradores (em preparação)
EUR	0,083	0,14
EAS	0,071	0,13
SAS	0,097	0,16
AFR	0,102	0,14
ALL	0,126	0,19