Débora Yoshihara Caldeira Brandt

Diferenciação populacional em genes sob forte
seleção balanceadora:
um estudo de caso com genes HLA

Population differentiation at genes under strong
balancing selection:
a case study on the HLA genes

São Paulo
2015

Débora Yoshihara Caldeira Brandt

Diferenciação populacional em genes sob forte
seleção balanceadora:
um estudo de caso com genes HLA

Population differentiation at genes under strong
balancing selection:
a case study on the HLA genes

Dissertação apresentada ao Instituto de
Biociências da Universidade de São Paulo,
para a obtenção de Título de Mestre em
Ciências, na Área de Biologia (Genética).
Orientador: Prof. Dr. Diogo Meyer

São Paulo
2015

Brandt, Débora Yoshihara Caldeira

Diferenciação populacional em genes sob forte seleção
balanceadora: um estudo de caso com genes HLA

83 páginas

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

1. Diferenciação populacional 2. HLA 3. 1000 Genomas 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. Dr. Diogo Meyer

Para Célia e Fernando

Agradecimentos

Agradeço ao Diogo pela orientação atenciosa, pelo incentivo constante e pela confiança em me oferecer tantas oportunidades ao longo do caminho. Sempre será uma grande inspiração e referência para mim como cientista.

Ao professor Jérôme Goudet por ter me recebido em Lausanne e pelo inestimável avanço que esse trabalho teve com suas contribuições.

Aos professores Gabriel Marroig, Cristina Miyaki e Sergio Mاتيoli pelas contribuições dadas a esse trabalho em minha banca de qualificação.

Agradeço profundamente à Lúcia Armelin-Correa, pelo bom humor e entusiasmo com a pesquisa que me contagiaram para sempre.

A esse grupo de pessoas lindas que acompanhou de perto esse trabalho: Kelly, Vitor, Jonatas, Bárbara, Limão e Carol, devo muito a vocês. Companheiros de HLA, de R e de café com bolo.

Agradeço em particular a todos os coautores do primeiro capítulo desta dissertação, que o construíram lado a lado comigo, contribuindo com ideias e muita mão na massa.

Um obrigada muito especial à Bárbara, por me proteger e mimar desde o princípio. Pela amizade, pelas conversas, e por tudo que me ensinou, talvez até sem se dar conta.

Ao Binho pela força logo no início desse projeto, quando eu mal sabia onde isso tudo ia dar.

À Elisa e ao Ricardo que acompanharam a parte suíça dessa história e me fazem lembrar dela com muito mais carinho.

Agradeço também aos cafés e às trocas com todos os habitantes e agregados d'O Porão.

Aos amigos de Biologia, Pira e Toshiba, que estiveram tão presentes durante esses anos de mestrado. E especialmente ao Yoshi, que esteve sempre lá para dividir angústias e um monte de alegrias nos últimos tempos. Que venham mais anos assim! E aos outros biólogos queridos que estão por aí, Plank, Gê, Erro, Jow, Praize, Cassis, pelo carinho de sempre.

Às amigas Rê, Bibi, Freitas, Magá, Analú, Carol, Fê, Mari, que mesmo longe sempre estão perto. E que nunca vão largar a campanha pelo Nobel de Biologia.

À minha família pelo amor que nunca falhou e por sempre ter me oferecido um mundo de oportunidades. À minha mãe por ser a pessoa mais forte e decidida que eu conheço e pelo melhor colo do mundo. Ao meu pai por me mostrar o que é ser cientista, e por fazer eu gostar da ideia. E à minha irmã por aquelas conversas que são só com ela.

À minha tia Sandra, minha segunda mãe, meu tio Alfredo e meus primos Dê, Dani e Fê, porque "o vestibular tá chegando", e pelas aulas de tabuada.

Ao Daniel, minha nova família e meu melhor amigo. Obrigada por me entender como ninguém, e por tantos chororôs que viraram risadas. Obrigada por acreditar em mim mais que eu mesma, pela paciência, pelo incentivo, pela companhia "na alegria e na tristeza", no trabalho e no descanso.

Agradeço à FAPESP e ao CNPq pelo apoio financeiro a este projeto.

Índice

Lista de Figuras	xi
Lista de Tabelas	xiii
Introdução Geral	1
Definindo regimes seletivos	1
Assinaturas de seleção natural no genoma	3
Teste de neutralidade de Lewontin-Krakauer	4
Medindo diferenciação populacional	6
Seleção balanceadora e os genes HLA	7
Motivação: estudos de diferenciação em HLA	9
Estrutura da Dissertação	11
1 Mapping bias overestimates reference allele frequencies at the HLA genes in the 1000 Genomes Project phase I data	13
1.1 Introduction	14
1.2 Methods	16
1.3 Results	19
1.4 Discussion	27
2 How balancing selection shapes population differentiation at the HLA genes	29
2.1 Introduction	30
2.2 Methods	32
2.3 Results	35
2.4 Discussion	42
Discussão Geral e Conclusões	45
Controles Metodológicos	45
Diferenciação nos HLA	48
Resumo	51
Abstract	53
Referências Bibliográficas	55
Apêndices	61
I Supporting Information for Chapter 1	63
II Supporting Information for Chapter 2	81

Lista de Figuras

1	Distribuição de valores fenotípicos em uma população sujeita a diferentes modos de seleção	2
1.1	Genotype mismatches between the 1000G and PAG2014 datasets	19
1.2	Reference allele frequency per site in each HLA gene in the 1000 Genomes (1000G) and Sanger sequencing (PAG2014) datasets	20
1.3	Relationship between coverage and (a) genotype mismatch and (b) error in frequency estimate	22
1.4	Difference in reference allele frequency between 1000G and PAG2014 at each polymorphic site, in each population	23
1.5	Number of differences to the reference genome at 1860 51bp windows centered at sites <i>HLA-B</i> 132 and <i>HLA-DQB1</i> 244 with reference (REF) or alternative (ALT) allele at those sites	24
1.6	Number of differences to the reference genome at 51bp windows centered at each SNP in the <i>HLA-A</i> , <i>-B</i> and <i>-DQB1</i> genes	25
1.7	Heterozygosity of SNPs at HLA genes estimated from the PAG2014 dataset . .	26
1.8	Relationship between SNP heterozygosity (H) and (a) absolute value of deviation or (b) magnitude and direction of deviation	27
2.1	Distribution of F_{ST} values for SNP from the following groups: intergenic, exonic control group, HLA genes and at peri-HLA genes.	35
2.2	Maximum F_{ST} as a function of MAF for 10 populations	37
2.3	Population differentiation, measured by F_{ST} , as a function of minor allele frequency (MAF) for the exonic SNPs from the 1000 Genomes Project phase I data.	37
2.4	Distribution of minor allele frequency (MAF) for exonic control, peri-HLA genes and HLA genes	38
2.5	F_{ST} distributions as a function of minor allele frequency (MAF) intervals . . .	39
2.6	Average F_{ST} at the HLA and in the control for each pair of populations	41
2.7	Population differentiation as a function of minor allele frequency at the haplotype level	42

Lista de Tabelas

- 1.1 Number of sites with overestimation of reference (REF) or alternative (ALT) allele frequency in each HLA locus 24
- 2.1 Mean F_{ST} at the HLA and control SNPs using two averaging approaches. . . . 39
- 2.2 Resampling control SNPs to account for linkage disequilibrium among HLA SNPs 40
- 3 F_{ST} , frequência do alelo mais raro e paradoxo de Yule-Simpson 47

Introdução Geral

A diversidade genética dentro e entre populações é moldada pela ação da seleção natural, deriva genética e migração, que agem sobre a variação introduzida pelos processos de mutação e recombinação.

Genomas estão sujeitos às ações de mutação, recombinação, migração, deriva e seleção e, conseqüentemente, sua estrutura física também modela os efeitos dessas forças evolutivas. A ligação física de variantes presentes num mesmo cromossomo faz com que forças seletivas conflitantes muitas vezes atuem sobre uma mesma região.

Compreender os padrões gerados por esses processos no genoma e discernir os efeitos de cada um deles é um dos grandes desafios da genética de populações. Esse tipo de conhecimento abre as portas para compreendermos melhor a história de espécies, incluindo eventos demográficos e seletivos, e gera aplicações diretas em áreas como conservação da biodiversidade, melhoramento genético e medicina.

Definindo regimes seletivos

Diferentes áreas da Biologia Evolutiva usam diferentes terminologias para classificar formas de seleção natural (revisadas em Nielsen, 2005). Antes de revisar as assinaturas deixadas pela seleção natural no genoma, é fundamental definir essa terminologia.

Em Genética Quantitativa, a seleção natural é tradicionalmente classificada em três modos: direcional, estabilizadora e disruptiva. A seleção é **direcional** quando um fenótipo extremo tem a maior aptidão, **estabilizadora** quando um fenótipo intermediário é o mais apto, e **disruptiva** (ou diversificadora) quando mais de um fenótipo extremo tem a aptidão maior que os fenótipos intermediários (Figura 1) (Mather (1953) *apud* Kimura (1983), p. 119).

É comum encontrar a terminologia descrita acima para caracterizar regimes seletivos com base nas aptidões de genótipos. Nesse caso, **seleção direcional** é definida como o regime em que um dos homozigotos (e.g. A_1A_1) tem aptidão maior que o heterozigoto (A_1A_2), que por sua vez tem aptidão maior que o outro homozigoto (A_2A_2). **Seleção estabilizadora** ocorre quando o heterozigoto tem aptidão maior que ambos os homozigotos, e **seleção disruptiva** ocorre quando o heterozigoto tem a menor aptidão dentre todos os genótipos (revisão em Losos *et al.*, 2013).

Apesar de comum, o uso da terminologia original da genética quantitativa para descrever modos de seleção agindo sobre genótipos pode causar confusão. Nessa transição, há um

pressuposto implícito de que fenótipos extremos são determinados por genótipos homozigotos, o que pode nem sempre ser verdade (Futuyma, 2006).

Figura 1. Distribuição de valores fenotípicos em uma população sujeita a diferentes modos de seleção. Os painéis da esquerda mostram seleção direcional, *i.e.* um fenótipo extremo tem maior aptidão, levando a uma mudança na aptidão média da população. Os painéis centrais mostram seleção estabilizadora, *i.e.* fenótipos intermediários têm maior aptidão, levando a uma redução na variância dos valores fenotípicos da população. Painéis à direita mostram seleção disruptiva, *i.e.* fenótipos extremos têm maior aptidão que os intermediários, o que pode levar a uma distribuição bimodal dos valores fenotípicos. Fonte: Hamilton (2009).

No campo da Evolução Molecular, um outro conjunto de termos classifica a seleção natural sob a perspectiva dos valores adaptativos de variantes novas na população. Nesse cenário, seleção natural pode ser classificada como positiva, negativa ou balanceadora. Seleção **positiva** é o processo de aumento da frequência de novas variantes vantajosas, enquanto seleção **negativa** (ou purificadora) é o processo de redução da frequência de variantes deletérias da população. Os regimes de seleção positiva e negativa tendem à fixação ou perda de variantes, respectivamente, e portanto reduzem a diversidade genética das populações onde atuam. O terceiro tipo de seleção, seleção **balanceadora**, engloba todo regime seletivo que mantém a diversidade genética de populações em níveis mais altos que o esperado sob neutralidade. Regimes possíveis de seleção balanceadora incluem vantagem de heterozigotos (ou sobredominância), vantagem de alelo raro ou pressões seletivas que variam no tempo e/ou no espaço (Andrés, 2011).

Seleção positiva, frequentemente chamada de "seleção Darwiniana", é comumente usada como sinônimo de seleção direcional. Em um cenário em que uma nova variante de dominância incompleta determina um fenótipo vantajoso, tanto os conceitos fenotípico e genotípico de seleção direcional, quanto o conceito de seleção positiva se aplicam.

Por outro lado, a relação entre seleção negativa e seleção balanceadora e a terminologia da genética quantitativa é mais complicada. Gillespie (2004), por exemplo, trata de seleção balanceadora exclusivamente na forma de vantagem de heterozigotos. Dessa perspectiva, seleção balanceadora seria equivalente ao conceito genotípico de seleção estabilizadora. Contudo, um cenário no qual o fenótipo extremo é o mais vantajoso, e este é determinado por um genótipo heterozigoto constituiria seleção balanceadora (por vantagem de heterozigotos), seleção estabilizadora do ponto de vista do genótipo, e seleção direcional, do ponto de vista do fenótipo selecionado. Também é comum fazer um paralelo entre seleção balanceadora e seleção disruptiva, já que ambas aumentam a diversidade: a primeira aumenta a diversidade genética e pode também aumentar a diversidade fenotípica, enquanto a segunda por definição aumenta a diversidade fenotípica.

Portanto, não há uma correspondência única entre as classificações de tipos de seleção natural das áreas de Genética Quantitativa e de Evolução Molecular. É importante ter esses conceitos em mente ao se levantar as hipóteses sobre assinaturas de diferentes modos de seleção no genoma. Neste trabalho, será usada a classificação da literatura de Evolução Molecular (*i.e.* seleção positiva, negativa ou balanceadora).

Assinaturas de seleção natural no genoma

Cada regime de seleção natural modifica a diversidade das regiões genômicas onde atua de maneira diferente, gerando padrões característicos. A busca por padrões, ou assinaturas, permite identificar regiões do genoma associadas a adaptações e revelar aspectos da história evolutiva de genes e populações (Nielsen, 2005; Sabeti *et al.*, 2006).

Quando uma mutação vantajosa surge e aumenta de frequência por seleção positiva, algumas das assinaturas que podem ser identificadas na variante selecionada e nas regiões vizinhas são: redução da variação genética e um espectro de frequências alélicas* com excesso de variantes derivadas (Nielsen, 2005).

O efeito em regiões vizinhas ocorre pois a mutação vantajosa surge em um haplótipo específico, e quando aumenta de frequência carrega consigo esse haplótipo em detrimento dos demais. O processo de aumento de frequência das variantes ligadas ao alelo selecionado é conhecido como efeito carona (termo cunhado por Smith & Haigh (1974)), e o cenário resultante, de eliminação da variação na região adjacente à mutação, é chamado de varredura seletiva.

Seleção negativa sobre novas variantes também gera efeitos particulares em regiões vizinhas. Esse processo é conhecido como seleção de fundo, e resulta em redução de diversidade e um espectro de frequências alélicas com excesso de variantes raras. Por fim, seleção balanceadora resulta em aumento de diversidade e espectro de frequências alélicas com excesso de variantes em frequência intermediária.

É possível estimar o nível de diversidade genética esperado na ausência de seleção natural a partir da teoria neutra. Essa estimativa parte do pressuposto de que as populações estão

*Histograma de frequências alélicas de um conjunto de variantes.

em equilíbrio entre mutação e deriva, isto é, seus tamanhos populacionais são constantes, de modo que sua diversidade está em equilíbrio entre a perda de variação por deriva genética e a introdução de novas variantes por mutação.

Um desafio em se detectar seleção natural no genoma usando as assinaturas listadas acima está no fato de que muitas delas podem surgir quando populações desviam do equilíbrio, ainda que na ausência de seleção natural. Isso pode acontecer devido a fenômenos demográficos, como gargalos ou expansões populacionais, que causam diminuição e aumento da diversidade genética, respectivamente. Para lidar com esse problema, foram desenvolvidos testes que buscam desvios em relação a cenários esperados em situações nas quais as populações não estão em equilíbrio.

Teste de neutralidade de Lewontin-Krakauer

Um dos primeiros testes de neutralidade propostos, o teste de Lewontin-Krakauer, baseia-se em uma assinatura de seleção positiva local, isto é, seleção positiva agindo em uma população específica (Lewontin & Krakauer, 1973). Nesse cenário, espera-se que as variantes sob seleção e suas regiões vizinhas, levadas por carona, tenham frequências alélicas muito distintas entre populações. Em contrapartida, espera-se que pressões seletivas partilhadas entre populações (seleção global) façam com que as frequências alélicas tornem-se mais semelhantes entre populações.

Contudo, fenômenos demográficos, como gargalos populacionais, também podem aumentar a diferenciação genética entre populações. Lewontin e Krakauer, baseados nas ideias de Cavalli-Sforza (1966), desenvolveram um método para lidar com esse efeito confundidor da demografia sobre assinaturas de seleção natural. Eles chamaram a atenção para o fato de que a seleção natural opera de maneira diferente sobre cada *locus*, enquanto o efeito da demografia é uniforme em todo o genoma. Assim, eles propuseram que uma maneira de se identificar genes sob seleção local seria comparar medidas de diferenciação populacional entre *loci*. Aqueles com valores extremos em relação aos demais seriam candidatos a estarem sob seleção (Lewontin & Krakauer, 1973).

A crescente facilidade de obtenção de dados genômicos em muitas populações tornou viável a abordagem sugerida por Lewontin e Krakauer. Em 2002, Akey *et al.* realizaram a primeira varredura (*scan*) genômica que buscou extremos de diferenciação populacional. Esse estudo comparou 26.530 polimorfismos de nucleotídeo único (SNPs, do inglês, *Single Nucleotide Polymorphisms*) em três populações humanas em busca de *loci* para os quais o valor de diferenciação populacional, medida pela estatística $F_{ST}^{\dagger}$, estava em um dos extremos da distribuição para o conjunto total de SNPs analisados (*outliers*). Regiões contendo SNPs classificados como *outliers* foram consideradas candidatas a estarem sob seleção, e 174 genes candidatos foram encontrados. Dois desses já haviam apresentado indícios de seleção em outros estudos, e 18 haviam sido preditos por estudos computacionais.

O teste de Lewontin e Krakauer foi explicitamente proposto para detecção de seleção local (em uma população ou região geográfica) ou global (regimes seletivos partilhados entre

[†] F_{ST} será descrito em maior detalhe na próxima seção

populações) dependendo se os valores da estatística F_{ST} são extremos na direção de valores altos ou baixos, respectivamente:

"For the selection loci, the [variance in allele frequencies] and therefore the estimated [F_{ST}] will be too large if selection is different in different populations, while if there is a selection in common, the variance among populations will be too small."

Lewontin & Krakauer (1973)

Contudo, é comum encontrar na literatura associações de extremos de diferenciação populacional a regimes de seleção positiva, negativa ou balanceadora. Apesar de comum, ilustraremos a seguir os desafios de se associar um regime seletivo específico às assinaturas de diferenciação inusualmente alta ou baixa.

Muitos dos genes com diferenciação aumentada apresentam também outras assinaturas de seleção positiva (por exemplo, haplótipos longos e excesso de variantes derivadas em frequência alta) (Akey *et al.*, 2002; Hofer *et al.*, 2012; Barreiro *et al.*, 2008). Um dos exemplos mais didáticos de seleção positiva associada à assinatura de diferenciação populacional alta é a variante regulatória que gera a persistência da lactase em humanos (LCT -13910 C>T, rs4988234), que é muito mais frequente em europeus que em outras populações (Bersaglieri *et al.*, 2004). Variantes associadas a adaptação à altitude em tibetanos (no gene EPAS1), que têm frequências muito diferentes na população chinesa, também são um exemplo marcante (Xu *et al.*, 2011). Nesses casos, o aumento de diferenciação populacional é uma assinatura de seleção positiva robusta, uma vez que detecta variantes semelhantes às encontradas por estudos que usaram outros métodos de detecção de seleção positiva.

Enquanto é relativamente simples associar regiões de alta diferenciação a seleção positiva local, é mais complexo interpretar regiões de baixa diferenciação. Partindo do pressuposto de que seleção negativa atua da mesma maneira em diferentes populações, contribuindo para a conservação de regiões funcionalmente importantes do genoma, seria esperada uma associação entre baixa diferenciação populacional a seleção negativa (Nielsen, 2005; Barreiro *et al.*, 2008).

Contudo, também é comum na literatura a expectativa de que seleção balanceadora leve à redução de diferenciação populacional (Bowcock *et al.*, 1991; Excoffier & Lischer, 2010; Lotterhos & Whitlock, 2014), e essa expectativa é embasada por estudos teóricos com simulações (Schierup *et al.*, 2000; Muirhead, 2001). Consequentemente, varreduras genômicas buscando regiões de diferenciação populacional extrema (alta ou baixa) costumam atribuir regiões com valores de diferenciação extremamente baixos à atuação de seleção balanceadora (Akey *et al.*, 2002; Barreiro *et al.*, 2008; Hofer *et al.*, 2012).

Portanto, baixa diferenciação populacional pode ser uma assinatura de seleção positiva global, de seleção negativa, ou de seleção balanceadora. Para diferenciar esses regimes seletivos, é necessário considerar também as frequências alélicas: sob seleção negativa, espera-se que as variantes com baixa diferenciação populacional sejam variantes derivadas em baixa frequência, sob seleção balanceadora espera-se variantes com frequências intermediárias e sob seleção positiva global espera-se variantes derivadas em alta frequência (Barreiro *et al.*, 2008).

A relação entre seleção balanceadora e baixa diferenciação populacional é intuitiva em casos de seleção balanceadora sobre *loci* bialélicos, quando a manutenção de ambos os alelos invariavelmente levaria à redução da diferenciação populacional. Contudo, seleção balanceadora agindo sobre *loci* multialélicos pode manter alta diferenciação populacional se diferentes conjuntos de alelos forem vantajosos em diferentes populações.

Para examinar se há suporte para a expectativa de baixa diferenciação populacional em genes sob seleção balanceadora, comparamos os achados de varreduras que buscaram assinaturas de seleção balanceadora às regiões encontradas em varreduras que buscaram regiões com diferenciação populacional extrema (Akey *et al.*, 2002; Barreiro *et al.*, 2008; Hofer *et al.*, 2012).

Candidatos a alvos da seleção balanceadora foram identificados em duas varreduras genômicas que buscaram diferentes assinaturas desse tipo de seleção: Bubb *et al.* (2006) procuraram regiões com alta densidade de SNPs e Andrés *et al.* (2009) buscaram regiões com polimorfismo aumentado e espectro de frequências alélicas com excesso de variantes de frequência intermediária. Verificamos que nenhum dos genes encontrados nessas varreduras para seleção balanceadora aparecem entre os pouco diferenciados em varreduras buscando diferenciação populacional extrema (Akey *et al.*, 2002; Barreiro *et al.*, 2008; Hofer *et al.*, 2012). Entretanto, dois dos 16 genes encontrados por Bubb *et al.* (2006) (*CNTNAP2* e *CSMD1*) e um dos 60 genes encontrados por Andrés *et al.* (2009) (*CAMK2B*) estão entre os de diferenciação elevada encontrados por Barreiro *et al.* (2008). Essa revisão da literatura sugere uma ausência de correspondência entre a classe de genes sob seleção balanceadora e aqueles com baixa diferenciação, o que vai de encontro à frequente associação entre esses dois processos em estudos baseados no teste de Lewontin & Krakauer (1973).

Em resumo, o teste proposto por Lewontin & Krakauer (1973) tem a capacidade de discernir o grau de compartilhamento de regimes seletivos (local ou global). Regiões com alta diferenciação populacional são frequentemente associadas a eventos de seleção positiva local, potencialmente relacionadas a adaptações locais. Contudo, as expectativas sobre a relação entre diferenciação populacional e seleção balanceadora são menos claras.

Medindo diferenciação populacional

O parâmetro mais comumente usado para descrever o grau de diferenciação genética entre populações, F_{ST} , foi definido por Wright (1951) como uma medida da variância das frequências alélicas em um *locus* (σ_p) em relação a seu valor máximo possível, dada a frequência alélica na população total ($\bar{p}(1 - \bar{p})$):

$$F_{ST} = \frac{\sigma_p}{\bar{p}(1 - \bar{p})}$$

Desde que foi descrito pela primeira vez, diversas interpretações e métodos para estimar esse parâmetro foram desenvolvidos (revisados em Weir, 2012). F_{ST} pode ser interpretado como a proporção da variância de frequências alélicas explicada por subdivisão populacional (Cockerham, 1969), como uma medida da redução de heterozigose decorrente de subdivisão

populacional (Nei, 1973), ou como uma medida que compara os tempos de coalescência dentro de subpopulações ao tempo médio de coalescência na população como um todo (Slatkin, 1991).

A interpretação de F_{ST} baseada em redução da heterozigose, desenvolvida por Nei (1973) e trabalhos seguintes (Nei, 1977; Nei & Chesser, 1983), considera F_{ST} como uma estatística dos dados (e não um parâmetro do processo evolutivo), e fornece uma descrição da distribuição de frequências alélicas no tempo atual, para o conjunto de populações amostradas. Uma vantagem dessa concepção de F_{ST} é que ela não inclui pressupostos sobre a relação entre populações, enquanto F_{ST} interpretado em termos de componentes de variância (Cockerham, 1969) pressupõe que todas as populações são igualmente relacionadas umas às outras - o que pode ser pouco realista na maioria dos casos (Nei & Kumar, 2000, Cap. 12).

Neste trabalho utilizo principalmente a concepção de F_{ST} como proporção da variância de frequências alélicas explicada por subdivisão populacional. Essa concepção de F_{ST} foi escolhida porque foi demonstrado que seus estimadores são menos enviesados pelo conjunto particular de populações amostradas, além de terem fundamentação estatística mais precisa (Balding *et al.*, 2007, cap. 29). Nesse contexto, o estimador de F_{ST} mais usado é o de Weir & Cockerham (1984). Entre as vantagens desse estimador, está o fato de que ele leva em conta o efeito de deriva genética - o que o autor denomina de "amostragem evolutiva", em comparação à amostragem estatística que em geral é a única considerada ao elaborar estimadores de parâmetros populacionais (Holsinger & Weir, 2009). Além disso, foi demonstrado que quando os tamanhos amostrais são semelhantes, e quando médias de F_{ST} entre múltiplos marcadores são ponderadas de modo adequado, como no nosso caso, o estimador de Weir & Cockerham (1984) é apropriado (Bhatia *et al.*, 2013, veja também Capítulo 2 - Estimation of F_{ST}).

Apesar das diferenças conceituais entre os estimadores de diferenciação populacional de Weir & Cockerham (1984) e Nei (1973), ambos são frequentemente usados na literatura. Comparações de estimativas de F_{ST} usando essas duas abordagens mostram que os resultados são muito semelhantes quando o número de indivíduos e populações amostradas são grandes (Balding *et al.* (2007) Cap. 29, Nei & Kumar (2000), Cap. 12).

Seleção balanceadora e os genes HLA

O principal interesse no estudo de seleção balanceadora está no fato de que este tipo de seleção natural aumenta a diversidade genética. A compreensão dos processos envolvidos na manutenção da diversidade genética, e a importância relativa da seleção balanceadora nessa manutenção, remontam a um debate antigo dentro da genética de populações.

Até os anos 40, predominava na área de genética de populações a visão de que cada *locus* carrega um alelo muito comum que confere um fenótipo "selvagem", e alguns poucos alelos deletérios, chamados de "mutantes". Essa *escola clássica* predizia, portanto, que populações carregavam pouca variação genética. Os pesquisadores de maior destaque dentre os que sustentavam essa visão eram Thomas H. Morgan e Hermann J. Muller, através de seus estudos de herdabilidade de traços fenotípicos clássicos de *Drosophila*, como padrão de venação nas asas ou cor dos olhos, em populações altamente endocruzadas (Templeton, 2006).

Mais tarde, estudos de citogenética de Alfred Sturtevant e Theodosius Dobzhansky com populações naturais de *Drosophila* revelaram que existe grande diversidade genética de inversões cromossômicas. Essas observações subsidiaram o desenvolvimento de uma nova escola de pensamento em genética de populações, que ficou conhecida como a *escola balanceada*. Essa escola defendia que populações possuíam altos níveis de variação genética, e que esta era vantajosa e mantida por seleção natural (Templeton, 2006).

Nos anos 60, com o advento da técnica de eletroforese de proteínas, ficou claro que populações carregam uma grande quantidade de variação genética. Uma vez que esse conhecimento ficou estabelecido, o debate entre as escolas clássica e balanceada mudou: enquanto a escola balanceada permanecia defendendo que a variação genética era mantida por seleção natural, a escola clássica transformou-se na *escola neutra*, defendendo que a maior parte da variação genética não tinha efeito fenotípico, e seria invisível à seleção natural (Templeton, 2006).

O desenvolvimento da teoria neutra, marcado pelos trabalhos de Kimura em 1968 e de Ohta no início dos anos 70 (Ohta & Gillespie, 1996), fortaleceu a visão da *escola neutra*, mostrando que grande parte da variação genética em populações podia ser explicada pela combinação de mutação e deriva genética, sem a ação de seleção natural. Contudo, a importância relativa da seleção balanceadora na manutenção de polimorfismo em escala genômica ainda é assunto de debate (Charlesworth, 2015).

Ubíqua ou não, seleção balanceadora é inegavelmente importante na evolução de algumas regiões do genoma. Um exemplo clássico é o caso de vantagem de heterozigotos para o gene da β -globina em populações humanas que vivem em regiões afetadas pela malária. Indivíduos homozigotos para o alelo mutante, que confere fenótipo falciforme às hemácias, sofrem de grave anemia falciforme. Indivíduos homozigotos para o alelo selvagem são suscetíveis à malária. Já os heterozigotos não sofrem de anemia em condições ambientais normais e são resistentes à malária, portanto têm maior vantagem seletiva em relação aos demais genótipos neste ambiente (Pasvol *et al.*, 1978).

Outro exemplo notável de seleção balanceadora, que parece não depender de um ambiente tão particular quanto o anterior, é a família de genes do Complexo Principal de Histocompatibilidade (MHC, do inglês, *Major Histocompatibility Complex*). Essa família gênica está presente em todos os vertebrados mandibulados, e codifica proteínas responsáveis pela apresentação de peptídeos aos linfócitos T. Em humanos, esses genes são denominados Antígenos Leucocitários Humanos (HLA, do inglês, *Human Leukocyte Antigen*). Os genes HLA estão contidos numa região do cromossomo 6 (6p21.3) de aproximadamente 4 megabases que, confusamente, também é chamada de MHC (Hughes, 2008).

Os genes HLA clássicos são divididos em genes de classe I (HLA-A, -B e -C) e classe II (HLA-DRA, -DRB1, -DRB3, -DRB4, -DRB5, -DQA1, -DQB1, -DPA1, -DPB1). Os genes clássicos de classe I codificam proteínas expressas em todas as células somáticas nucleadas, que apresentam peptídeos citoplasmáticos a linfócitos T citotóxicos. Estes últimos, ao reconhecerem uma molécula de HLA de classe I ligada a um peptídeo não-próprio, matam a célula infectada. Os genes HLA clássicos de classe II são expressos em células apresentadoras de antígenos do sistema imune, e apresentam peptídeos processados em organelas celulares aos linfócitos T auxiliares. Caso um peptídeo não-próprio seja reconhecido pelos linfócitos T auxiliares,

estes desencadeiam uma resposta imune complexa, que pode incluir o recrutamento de linfócitos T citotóxicos e linfócitos B, responsáveis pela produção de anticorpos (Hughes, 2008).

Além dos genes HLA clássicos, a região do MHC apresenta outros genes de estrutura muito semelhante, mas que não efetuam a função de apresentação de antígenos. Estes são chamados de genes HLA não-clássicos.

A família dos genes HLA é a mais polimórfica do genoma humano, e possui diversas evidências de seleção balanceadora: distribuição de frequências alélicas mais uniforme que a esperada sob neutralidade, diversidade genética aumentada, alto desequilíbrio de ligação, taxa de substituição não-sinônima maior do que a sinônima e polimorfismo transespecífico (Meyer & Thomson, 2001). Essas evidências baseiam-se em pressupostos distintos, o que torna robusta a conclusão de que seleção balanceadora atua sobre esses genes.

À luz da compreensão da função exercida pelos produtos desses genes, diferentes ideias foram propostas para explicar como a seleção balanceadora atua sobre eles. Doherty & Zinkernagel (1975) propuseram que indivíduos heterozigotos nos genes HLA têm vantagem seletiva em relação aos homozigotos, uma vez que têm maior diversidade de moléculas de HLA, e portanto podem apresentar uma maior variedade de antígenos.

Um regime de seleção movido por patógenos, sem que necessariamente ocorra a vantagem de heterozigotos, é outro cenário plausível. Nesse caso, parte-se do pressuposto de que o conjunto de patógenos presente em cada população é diferente, favorecendo a manutenção de diferentes alelos em diferentes localidades. Esse modelo de seleção prevê que populações expostas a uma maior diversidade de patógenos serão mais variáveis para genes HLA, algo de fato observado (Prugnolle *et al.*, 2005; Qutob *et al.*, 2012). Outros resultados que apoiam esse modelo incluem um estudo de identidade por descendência no genoma, que mostrou através de simulações que o padrão observado no entorno de genes HLA é mais compatível com seleção dependente de frequência do que com vantagem de heterozigotos (Albrechtsen *et al.*, 2010). Estudos de simulação também mostraram que um regime que inclui adaptação a patógenos locais é capaz de explicar melhor a variabilidade extrema observada nos genes HLA, pois mantém um número muito maior de alelos numa população do que o de regime de vantagem de heterozigotos por si só (Borghans *et al.*, 2004).

Em conjunto, esses trabalhos mostram que regimes de seleção com adaptação a patógenos locais fazem previsões que se encaixam melhor aos dados observados do que o modelo de vantagem de heterozigoto. Entretanto, é importante lembrar que esses mecanismos não são mutuamente excludentes.

Motivação: estudos de diferenciação em HLA

Em oposição à expectativa teórica de que seleção balanceadora leva a redução de diferenciação populacional, os genes HLA raramente foram encontrados em varreduras genômicas como sendo pouco diferenciados. Entre os diversos *loci* clássicos com evidências de seleção balanceadora, apenas o *locus HLA-C* foi classificado como pertencente a uma região genômica de baixa diferenciação (Hofer *et al.*, 2012), enquanto duas outras varreduras encontraram

diferenciação elevada em genes HLA não clássicos (*HLA-DPB2* e *HLA-DPA2* em Barreiro *et al.* (2008) e Bhatia *et al.* (2011), respectivamente), ao passo que um terceiro trabalho não encontrou *loci* HLA em nenhum dos extremos (Akey *et al.*, 2002).

Além dessas varreduras, alguns estudos analisaram diretamente a diferenciação populacional nos genes HLA e em genes do MHC em outros animais, comparando-a a outras regiões do genoma, e encontraram resultados diversos.

Estudos com organismos não-modelo encontraram tanto níveis de diferenciação populacional similares entre os genes do MHC e marcadores em outras regiões do genoma (Miller *et al.*, 2010), quanto diferenciação maior (Loiseau *et al.*, 2009; Oliver *et al.*, 2009; Cammen *et al.*, 2011) ou menor (McCairns *et al.*, 2011) nos genes do MHC.

Em humanos, Sanchez-Mazas (2007) encontrou diferenciação menor nos genes HLA que em outras regiões, porém não realizou teste estatístico para avaliar a significância dessa diferença, enquanto Meyer *et al.* (2006) realizaram um teste de reamostragem e não encontraram diferença significativa entre os valores de diferenciação para os *loci* HLA e para outras regiões. Contudo, há limitações no tipo de dado usado nesses estudos: Sanchez-Mazas (2007) e Meyer *et al.* (2006) compararam diferenciação populacional entre genes HLA e outros marcadores (Sanchez-Mazas usou microssatélites e RFLPs[‡] e Meyer *et al.* usaram microssatélites e outras sequências em regiões não codificantes), e as diferenças entre os modelos mutacionais e a diversidade desses diferentes sistemas dificultam a interpretação dos resultados.

Em um estudo mais recente, Nunes (2011) comparou a diferenciação entre populações ameríndias em 16 microssatélites localizados na região do genes HLA à de 61 microssatélites localizados em outras regiões do genoma. Controlando para o tipo de marcador genético, Nunes (2011) encontrou diferenciação aumentada na região dos genes HLA, corroborando a hipótese de que pressões seletivas locais - possivelmente patógenos - moldam a distribuição da variação genética em genes HLA entre populações.

O trabalho de Nunes (2011) foi uma importante motivação para esse estudo, que visa explorar algumas questões que permaneceram em aberto. Primeiramente, Nunes (2011) estava interessada especificamente nas populações ameríndias, e portanto levantamos a questão do quão geral seria o padrão de diferenciação populacional aumentada em genes HLA na escala global. A disponibilidade de dados de resequenciamento de indivíduos de diferentes continentes, oriundos do Projeto 1000 Genomas (The 1000 Genomes Project Consortium, 2012a), permitiu que realizássemos essa análise em escala global. Além disso, esse conjunto de dados permite o uso de SNPs como marcadores genéticos, o que facilita as análises de diferenciação, uma vez que o mecanismo mutacional complexo dos microssatélites pode dificultar a interpretação das medidas de diferenciação populacional, como discutido no trabalho de Nunes (2011).

Nesse contexto, realizamos um estudo da diferenciação populacional em genes com fortes evidências a priori de seleção balanceadora: os genes HLA. A principal motivação deste trabalho foi compreender melhor o efeito da seleção balanceadora sobre a diferenciação populacional. Especificamente, investigamos duas questões principais: i) na escala global, o

[‡]Polimorfismo de comprimento de fragmento de restrição, do inglês *Restriction Fragment Length Polymorphism*

regime de seleção balanceadora atuante sobre os genes HLA consiste em pressões seletivas globais (partilhadas entre populações) ou locais?; ii) como reconciliar a expectativa teórica de baixa diferenciação populacional em regiões sob seleção balanceadora à observação de alta diferenciação populacional em genes do MHC em humanos e outras espécies?

Estrutura da Dissertação

Capítulo 1 Mapping bias overestimates reference allele frequencies at the HLA genes in the 1000 Genomes Project phase I data

No primeiro capítulo, exploro a qualidade dos dados do projeto 1000 Genomas na região dos genes HLA. A alta diversidade genética desses genes torna essa região um desafio às tecnologias de sequenciamento de DNA em massa usadas pelo projeto 1000 Genomas. Demonstro que a principal fonte de erros nesses dados é o viés de mapeamento a um único genoma referência, e defino uma lista de SNPs cuja frequência alélica é confiável nos dados do 1000 Genomas.

Capítulo 2 How balancing selection shapes population differentiation at the HLA genes

No segundo capítulo, apresento o estudo de diferenciação populacional com base em SNPs nos genes HLA, aplicando filtros definidos no Capítulo 1, e usando um controle genômico apropriado. Discuto os padrões de diferenciação populacional encontrados em diferentes escalas geográficas, e usando diferentes definições de alelos (no nível de SNPs ou haplótipos).

Mapping bias overestimates reference allele frequencies at the HLA genes in the 1000 Genomes Project phase I data

Brandt, D. Y. C., V. R. C. Aguiar, B. D. Bitarello, K. Nunes, J. Goudet, & D. Meyer, 2015 Mapping bias overestimates reference allele frequencies at the HLA genes in the 1000 Genomes Project phase I data. *G3:Genes|Genomes|Genetics* 5: 931–941

Abstract

Next Generation Sequencing (NGS) technologies have become the standard for data generation in studies of population genomics, as the 1000 Genomes Project (1000G). However, these techniques are known to be problematic when applied to highly polymorphic genomic regions, such as the Human Leukocyte Antigen (HLA) genes. Because accurate genotype calls and allele frequency estimations are crucial to population genomics analyses, it is important to assess the reliability of NGS data. Here, we evaluate the reliability of genotype calls and allele frequency estimates of the SNPs reported by 1000G (phase I) at five HLA genes (*HLA-A*, *-B*, *-C*, *-DRB1*, *-DQB1*). We take advantage of the availability of HLA Sanger sequencing of 930 of the 1,092 1000G samples, and use this as a gold standard to benchmark the 1000G data. We document that 18.6% of SNP genotype calls in HLA genes are incorrect, and that allele frequencies are estimated with an error higher than ± 0.1 at approximately 25% of the SNPs in HLA genes. We found a bias towards overestimation of reference allele frequency for the 1000G data, indicating mapping bias is an important cause of error in frequency estimation in this dataset. We provide a list of sites that have poor allele frequency estimates, and discuss the outcomes of including those sites in different kinds of analyses. Since the HLA region is the most polymorphic in the human genome, our results provide insights into the challenges of using of NGS data at other genomic regions of high diversity.

Resumo

Tecnologias de sequenciamento de nova geração (NGS, do inglês *Next Generation Sequencing*) tornaram-se o padrão para a geração de dados em estudos de genômica populacional, como no caso do Projeto 1000 Genomas (1000G). Contudo, essas técnicas são conhecidamente problemáticas quando aplicadas a regiões genômicas altamente polimórficas, como os genes HLA (do inglês, *Human Leukocyte Antigen*). Como genótipos e estimativas de frequências alélicas são cruciais para análises de genômica populacional, é importante investigar a confiabilidade dos dados de NGS. Neste trabalho, avaliamos a confiabilidade dos genótipos atribuídos e

frequências alélicas dos SNPs reportados pelo 1000G (fase I) em cinco genes HLA (*HLA-A*, *-B*, *-C*, *-DRB1*, *-DQB1*). Utilizamos como padrão ouro dados recém publicados de sequenciamento Sanger de 930 das 1.092 amostras do 1000G. Documentamos que 18.6% dos genótipos em genes HLA estão incorretos, e que frequências alélicas são estimadas com um erro maior que ± 0.1 em aproximadamente 25% dos SNPs em genes HLA. Encontramos um viés de superestimativa da frequência do alelo referência nos dados do 1000G, indicando que viés de mapeamento é uma causa importante do erro nas estimativas de frequência feitas a partir desses dados. Fornecemos uma lista de sítios que têm frequência alélica mal estimada, e discutimos as consequências da inclusão desses sítios em diferentes tipos de análises. Como a região dos genes HLA é a mais polimórfica do genoma humano, nossos resultados são relevantes para a compreensão dos desafios do uso de dados de NGS em outras regiões de alta diversidade.

1.1 Introduction

Whole genome resequencing data for large numbers of human individuals, as generated by the 1000 Genomes Project (www.1000genomes.org), provide unprecedented amounts of information about microevolutionary processes and demographic histories. Such inferences rely on either genotypic or allelic frequency information for each variable position, which constitute the data for downstream analyses and hypothesis testing.

The calling of SNPs and genotypes and the estimation of allele frequencies from Next Generation Sequencing (NGS) has undergone rapid development, along with likelihood-based and Bayesian methods created to deal with challenges associated to heterogeneity in read quality and coverage (Nielsen *et al.*, 2011). In Phase I of the 1000 Genomes Project, genotypes were called using a combination of different approaches: first, primary call sets were independently generated by different centers with different sequencing platforms, alignment and variant calling methods; then, a consensus SNP call set was generated and made publicly available (?).

The data generated by the 1000 Genomes Project have frequently been used to make inferences about evolutionary processes affecting our species, including the detection of targets of natural selection (Hernandez *et al.*, 2011; Ward & Kellis, 2012; Andersen *et al.*, 2012) and understanding the genetic basis of complex phenotypes (Lappalainen *et al.*, 2013). In addition, the detailed catalogue of genetic variation it provides across multiple human populations has been used to understand the processes affecting specific genes.

Among the well documented targets of selection is the Major Histocompatibility Complex (MHC) region of the human genome, which harbors the highly polymorphic classical Human Leukocyte Antigen (HLA) class I and II loci. The interest in these loci stems from their strong association to various autoimmune disorders (?), susceptibility and resistance to infection (Chapman & Hill, 2012), and striking signatures of genetic variation indicating strong balancing selection (Meyer & Thomson, 2001). Such types of investigations can naturally be extended to the analysis of the 1000 Genomes data, which provide a rich resource of population genetic variation within and around HLA genes.

In spite of this interest, the use of NGS data for HLA loci is hampered by a major technical hurdle, which is the mapping of short sequence reads to genes that are both highly polymorphic and which constitute a multi-gene family. The high polymorphism may decrease the probability that short reads will be successfully mapped to the reference genome, in the event that the sequenced individual carries a variant that is highly diverged from that used in the index (Nielsen *et al.*, 2011). In addition, the fact that many HLA genes have close paralogues increases the chance that a read will map to two or more genomic regions, leading it to be discarded from most sequencing analyses pipelines, and thus decreasing the amount of usable information for genotype calling (Treangen & Salzberg, 2012).

Previous studies explored the applicability of NGS to genotype the HLA alleles of an individual, where an allele is typically defined as the haplotype determined by a combination of SNPs within a given *HLA* gene (*e.g.* Erlich *et al.* (2011); Major *et al.* (2013)). To this end, Erlich *et al.* (2011) proposed NGS methodologies in which different steps - from sample preparation to haplotype level allele calling - were adapted to deal with the issues of high polymorphism and paralogy of HLA genes. In this way, they were able to successfully validate their methodology in a study of 270 samples that had been previously typed by sequence specific oligonucleotide (SSO) hybridization, which they treated as a gold standard dataset. The same gold standard dataset was used by Major *et al.* (2013), who also examined the reliability of calling HLA alleles using NGS, but using the 1000 Genomes alignment data, and showed that this publicly available dataset can be used for this purpose, after appropriate filters (*e.g.* coverage) are applied.

Both Erlich *et al.* (2011) and Major *et al.* (2013) were interested in using NGS data to determine HLA alleles. Information regarding HLA alleles is of biomedical relevance since HLA genotypes are often an important covariate to account for in association studies, and HLA typing is critical to hematopoietic transplantation. In this study, however, we evaluate the quality of SNP level genotype calls from the 1000 Genomes at the *HLA* genes.

The analysis of genotype and allele frequencies for SNPs contained within HLA genes has proven of great value in biomedical and evolutionary studies, and the 1000 Genomes dataset is a recurrently used resource in this context. Examples of the use of HLA SNP data from the 1000 Genomes Project include: (a) In GWAS studies SNPs in HLA genes are often associated with phenotypes of interest, and it is useful to understand the prevalence of these variants in additional populations; (b) GWAS studies benefit from knowledge of the haplotype structure surrounding HLA genes, which can be inferred from the dense SNP data of the 1000 Genomes for multiple populations (*e.g.* Hill-Burns *et al.* 2011); (c) When testing for selection, many studies have found strong evidence associated to the HLA region, using the 1000 Genomes as a source of polymorphism data (*e.g.* Leffler *et al.* 2013).

All the above applications of the 1000 Genomes Project SNP data in HLA genes are dependent on the reliability of genotype calls at each SNP. However, no study to date has provided a detailed survey of the reliability of individual genotype calls and allele frequency estimates at the SNPs in HLA genes, in spite of their frequent usage. We address this issue, discuss likely causes for cases of incorrect genotype calls and provide a list of reliable sites for the HLA loci in the 1000 Genomes data. As in previous studies (Erlich *et al.*, 2011; Major

et al., 2013), we used a dataset in which individuals had their *HLA* genes genotyped using Sanger sequencing as a gold standard to benchmark the genotypes called at the 1000 Genomes Project. However, differently from these other studies, which were interested in reconstructing the *HLA* haplotypes using NGS, here we have deconstructed the haplotypes determined from Sanger sequencing data into SNPs, and compared genotypes at the SNP level to the 1000 Genomes data. We took advantage of the recent availability of a dataset of Sanger sequencing based *HLA* genotyping of *HLA-A*, *-B*, *-C*, *-DQB1* and *-DRB1* for 930 of the samples from the 1000 Genomes Project (Gourraud *et al.*, 2014). Our results have implications for other studies that use SNP data from the 1000 Genomes in order to estimate allele frequencies. Because *HLA* loci are the most polymorphic in the human genome, they most likely represent the worst case scenario for mapping bias and, consequently, allele frequency estimation error.

1.2 Methods

In this study we compare NGS genotype calls and allele frequency estimates reported by the 1000 Genomes Project with those obtained in a study which used Sanger sequencing to genotype *HLA* genes. For the purpose of our analysis we assembled a dataset comprising the intersection of the 1000 Genomes and Sanger sequencing samples, resulting in 930 individuals from 12 populations. Figure S1.1 summarizes the pre-processing of both datasets, which preceded genotype and allele frequency comparisons.

1.2.1 1000 Genomes dataset (1000G)

SNP genotypes were acquired from the chromosome 6 integrated Variant Call Format (VCF) file from version 3 of the 1000 Genomes Project Phase I data, which is available at <ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/> (The 1000 Genomes Project Consortium, 2012a). We selected only SNPs in exons encoding the antigen recognition sites (ARS), which are exons 2 and 3 for *HLA-A*, *-B*, and *-C* (Bjorkman *et al.*, 1987) and exon 2 for *HLA-DQB1* and *-DRB1* (Brown *et al.*, 1993). Sites were selected based on the most inclusive coordinates of the RefSeq database in 22 July 2014 (see). Both SNP and sample selection were carried out using VCFtools v0.1.12b (Danecek *et al.*, 2011).

1.2.2 *HLA* reference panel by Gourraud *et al.* (2014) (PAG2014)

Gourraud *et al.* (2014) typed class I *HLA-A*, *-B* and *-C*, and class II *HLA-DRB1* and *-DQB1* genes of 1266 individuals from 14 different populations in Africa, Europe, Asia and America. The *HLA* sequence-based typing was performed with specific PCR amplification of ARS exons followed by Sanger sequencing. Data are available at the dbMHC website (<http://www.ncbi.nlm.nih.gov/gv/mhc/xslcgi.fcgi?cmd=cellsearch>) (Helmberg *et al.*, 2014).

Data from Gourraud *et al.* (2014) are available in the form of *HLA* allele names per individual. Allele naming for *HLA* genes follows specific rules (Marsh *et al.*, 2010). Briefly, allele names are composed of a letter indicating the locus, followed by 2 to 4 numeric fields separated by colons. Each numeric field indicates specific forms of variation: the 1st field distinguishes groups of alleles by serological type, and the following fields distinguish

nonsynonymous polymorphisms, synonymous polymorphisms, and non-coding differences, respectively. In order to obtain SNP genotypes and frequencies from the Sanger sequencing data, we converted all allele names to their associated sequences for ARS encoding exons. Sequences were acquired from the IMGT database (Robinson *et al.*, 2013), which keeps a well curated repository of all known HLA allele sequences.

Our analysis was restricted to ARS exons because the HLA typing method used by Gourraud *et al.* (2014) only probed genetic variation in these specific exons. As a consequence, multiple HLA alleles are compatible with the sequencing results, since the sites that differentiate them are in other exons. This results in what we refer to as an "ambiguous allele call" for an HLA allele (e.g., the allele is identified as B*35:03 but we cannot establish whether it is B*35:03:01 or B*35:03:02, or a group of alleles is attributed to an individual, such as B*35:02/B*35:03/B*35:04). Ambiguous allele calls may also happen when sequencing has low quality at bases that differentiate two alleles. In addition, there are also genotypic ambiguities, which occur when different pairs of alleles are compatible with the sequencing results. For individuals that bear ambiguous alleles, we created a consensus sequence in which ambiguous sites were reported with both possible alleles (e.g. A/T, see Figure S1.1). In this way, we incorporate the uncertainty associated to the sequence-based typing into downstream analyses.

Although we cannot rule out technical errors in the Sanger sequencing that generated the PAG2014 data (Gourraud *et al.*, 2014), we assume that this method provides the most reliable estimate of HLA alleles (and hence SNP genotypes), and will serve as a standard to estimate the reliability of genotype calls and allele frequencies for the 1000 Genomes data (De Santis *et al.*, 2013).

1.2.3 Genotype comparisons

We initially quantified how well the 1000G and PAG2014 data agreed with respect to genotype calls. Genotypes at each site in each individual were compared between the 1000G data and the PAG2014 data, here considered as a gold standard. In the case of sites with ambiguity (e.g. T/A) in the PAG2014 data, if one of the two possible alleles matched an allele present in the 1000G, we considered this an allele match and PAG2014 was corrected, by attributing the allele present in the 1000G data to the ambiguous site. After correcting the ambiguous sites in PAG2014, we only considered genotypes to be a match if both alleles in 1000G were present in the PAG2014 data, at that site.

Throughout this paper, sites are numbered according to their position in the ARS exons coding sequences (1-546 at the class I loci and 1-270 at the class II loci).

1.2.4 Allele frequency comparisons

After correcting all possible ambiguities in PAG2014 (as described above), we calculated allele frequencies for SNPs in both datasets. By comparing the frequency of the reference allele in 1000G to its value in PAG2014 we assessed the accuracy of allele frequency estimation. The reference allele was defined as the allele present in the hg19 build of the reference sequence

of the human genome. RefSeq IDs of the reference sequences used for each HLA gene are reported on .

We computed the error in 1000G frequency estimates per site i (FE_i) as:

$$FE_i = f_{i,1000G} - f_{i,PAG2014} \quad (1.1)$$

where $f_{i,1000G}$ and $f_{i,PAG2014}$ are the frequency of the reference allele at site i in 1000G and PAG2014, respectively. We also computed the mean absolute error in frequency estimates per gene as a mean of absolute FE_i for all sites within a gene (MAE):

$$MAE = \frac{1}{n} \sum_{i=1}^n |f_{i,1000G} - f_{i,PAG2014}| \quad (1.2)$$

where n is the number of SNPs in the gene.

1.2.5 Coverage in 1000G

Sequencing coverage per individual per site was calculated from the 1000 Genomes Project phase I BAM files for the low coverage experiments using the genomeCoverageBed program from BEDTools (Quinlan & Hall, 2010). BAM files are available on [ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/data/\[sampleID\]/alignment/](ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/data/[sampleID]/alignment/). Only low coverage BAM files were used to estimate coverage because genotype likelihoods for the data we analyzed (1000 Genomes Project Phase I integrated VCF files) were estimated from this source. Genotype likelihoods were estimated from high coverage exome BAM files only for a minority of sites that were exclusively discovered on the exome experiments, and were not used in the coverage analysis (See Table S1.1).

1.2.6 Testing for mapping bias

After demonstrating that there is an overestimation of reference allele frequency in the 1000G SNPs (see Results), we hypothesized that mapping bias was the underlying cause. To test this hypothesis we examined whether reads carrying the alternative allele at a SNP are less likely to map to the reference genome than reads carrying the reference allele. First, for each HLA allele present in the PAG2014 dataset, we defined windows of 51 basepairs that were centered on each SNP (25 basepairs upstream and 25 basepairs downstream of the SNP, including non-polymorphic sites). The set of windows centered on a specific SNP was then separated in two groups: i) those that carry the reference allele at the central site and ii) those that carry the alternative allele at the central site. Next, all windows were compared to the reference genome (hg19) sequence (the same sequence that was used as an index in the 1000 Genomes Project), and the number of mismatches was counted, excluding the mismatch at the central SNP. If mapping bias was influencing allele frequency estimates, we expected that, for SNP positions with overestimation of the reference allele frequency in the 1000G, the alternative alleles would be flanked by additional alternative alleles (and thus have a higher mismatch count against the reference sequence).

1.3 Results

1.3.1 Genotypic mismatch frequency

We found that, on average, 18.6% of genotypes were mismatched between 1000G and PAG2014 when individual genotypes for each site in the 5 classical HLA genes were compared, and exons with higher nucleotide diversity tend to have a higher proportion of genotype mismatches (Figure S1.2). We also observed that mismatches are specially concentrated on a few sites (Figure 1.1), with 18.7% of sites concentrating 50% of the mismatches over the 5 loci we analyzed.

Figure 1.1. Genotype mismatches between the 1000G and PAG2014 datasets. Results per polymorphic site ("Position") and per individual (930 in total). Individuals are ordered by number of mismatches (individuals with less mismatches on top). Sites are numbered according to their position in ARS exons coding sequence. Dark squares indicate mismatches between genotypes in the two datasets.

1.3.2 Reference allele frequency accuracy

Accuracy of estimation of allele frequencies in 1000G was assessed comparing the observed frequency of the reference allele in the 1000G data with that of PAG2014, for both the global dataset (consisting of a pooled set of all individuals) and for each population separately (see Figures S1.3-S1.7). We chose a difference of 0.1 between the frequencies on both datasets as a threshold that determines a "large frequency difference".

For the global dataset (Figure 1.2) we found that for *HLA-A* and *-C* most SNPs have similar frequency estimates for 1000G and PAG2014, with few large deviations (only 9/66 and 8/44 SNPs with absolute difference in frequencies ($|FE|$) larger than 0.1, respectively). The *HLA-DQB1* locus shows an intermediate proportion of SNPs with large deviations (10/42 SNPs with $|FE| > 0.1$), and *HLA-B* and *HLA-DRB1* show the greatest proportion of sites with large frequency differences between 1000G and PAG2014 (23/64 and 15/35 sites with $|FE| > 0.1$). Overall, the mean absolute difference in frequency between SNPs in the 1000G and PAG2014 data is 0.08, and it is higher at the HLA genes with the highest levels of nucleotide diversity (*HLA-B*, *-DQB1* and *-DRB1* all deviate by ± 0.1).

The proportion of genotype mismatches and allele frequency deviations per site are highly correlated (Pearson correlation = 0.86, p-value $< 10^{-16}$; Figure S1.8). However, some SNPs with a high proportion of genotype mismatches have well estimated allele frequencies. One example is site 465 at *HLA-B*, in which 44% of genotypes are mismatched, but $|FE|$ is only 0.007. Overall, 15 sites have more than 25% mismatched genotypes while showing $|FE| < 0.1$. (See Figure S1.8). This is possible when the frequency of genotype errors in which the reference allele is overrepresented is similar to the frequency of errors in which the alternative allele is overrepresented.

Allele frequency at the Axiom Exome Genotyping Array - Affymetrix

Because genotyping arrays constitute an additional frequently used resource to genotype SNPs within HLA genes, playing an important role in GWAS studies, we have also investigated the accuracy of allele frequency estimation from this genotyping technology. We estimated allele frequencies from Axiom Exome data, and we found that those allele frequency estimates are as reliable as the ones from the 1000 Genomes NGS data, at the same SNPs (see Figure S1.9).

Figure 1.2. (On next page) Reference allele frequency per site in each HLA gene in the 1000 Genomes (1000G) and Sanger sequencing (PAG2014) datasets. Continuous line indicates the expected relationship (i.e., no difference) between 1000G and PAG2014. Dashed lines indicate a ± 0.1 deviation from the expected frequency (as estimated from PAG2014 dataset). MAE (mean absolute error) defined in Methods. Numbers indicate site position in ARS exons sequence.

1.3.3 Relationship between sequencing coverage and genotypic mismatches

To investigate whether low sequencing coverage could explain genotype mismatches and deviations from expected allele frequencies, we compared sequencing coverage between mismatched and matched genotypes (Figure 1.3a) and assessed the relationship between coverage and frequency deviation (Figure 1.3b).

Figure 1.3. (a) Distribution of coverage (x-axis) at matched and mismatched genotypes; y-axis is the square root of the relative frequency (Mann-Whitney one-tailed test, p-value $< 10^{-16}$); (b) Relationship between mean coverage (x-axis) and absolute frequency difference ($|FE|$, y-axis) between 1000G and PAG2014 ($r = -0.11$, p-value = 0.09). All polymorphic sites from *HLA-A*, *-B*, *-C*, *-DRB1* and *-DQB1* genes are included in both *a* and *b*.

Sites with mismatched genotypes have on average lower sequencing coverage than sites with matched genotypes (Figure 1.3a; Mann-Whitney one-tailed test p-value $< 10^{-16}$). This is the expected relationship if low sequencing coverage explains genotype mismatches between datasets. However, the difference in sequencing coverage between sites with matched and mismatched genotypes is small (mean coverage in matching genotypes is 1.95, and 1.75 in non-matching genotypes, a difference of 6,2%), and has likely achieved very high significance only due to the large number of observations. Similarly, correlation between allele frequency deviation and sequencing coverage is weak and not significant (Figure 1.3b; $r = -0.11$, p-value = 0.09), although the direction of correlation is in agreement with what would be expected if lower coverage explained larger deviations in frequency estimation. We have also investigated the possible effect of the position of the SNPs relative to exon edges on the allele frequency deviations and found no correlation between those factors (Figure S1.10). We therefore investigated other factors that may account for errors in genotype calling.

1.3.4 Direction of frequency deviation

We found that most of the genotype mismatches are due to miscalling an alternative allele as a reference allele (Table S1.2). Furthermore, most deviations in allele frequency estimates are in the direction of an overestimation of reference allele frequencies in the 1000 Genomes

data (Figure 1.2). This information is summarized in Figure 1.4 which shows the location and magnitude of frequency deviations between the 1000G and PAG2014 data.

Figure 1.4. Difference in reference allele frequency between 1000G and PAG2014, measured by *FE* (see Methods), at each polymorphic site, in each population. Shades of red indicate overestimation of reference allele frequency and shades of blue indicate underestimation of reference allele frequency in 1000G. Full population names are given in Table S1.3.

The overall shift in the direction of overestimating reference alleles is summarized in Table 1.1, which shows the number of SNPs with more than 0.1 frequency difference in at least two populations, for each locus. For *HLA-A*, *-B* and *-DQB1* most sites with large frequency differences between 1000G and PAG2014 are skewed in the direction of overestimating the reference allele (p-value = 0.057 for *HLA-A* and p-value < 10^{-4} for *HLA-B* and *-DQB1*, binomial test for null hypothesis of equal numbers of deviations in direction of REF or ALT), whereas *HLA-C* and *HLA-DRB1* show no evidence for an excess of large deviations in the direction of reference alleles.

Table 1.1. Number of sites with overestimation of reference (REF) or alternative (ALT) allele frequency in each HLA locus ($|FE| > 0.1$ in 2 or more populations). Genomic coordinates of those sites are given in Table S1.4.

	<i>A</i>	<i>B</i>	<i>C</i>	<i>DQB1</i>	<i>DRB1</i>
REF	11	30	6	22	11
ALT	3	2	3	2	11

1.3.5 Testing for mapping bias

We hypothesized that the observed reference allele bias was caused by a lower efficiency in the mapping of reads containing the alternative allele. This is expected under the assumption that the reads carrying the alternative allele on average have more differences with respect to the reference genome (used by the 1000 Genomes Consortium as the index to align NGS reads) than reads carrying the reference allele. In this scenario, some sites would have a stronger bias than others if the alternative alleles in those sites are flanked by additional alternative alleles.

Figure 1.5. Number of differences to the reference genome at 1860 51bp windows centered at sites *HLA-B* 132 and *HLA-DQB1* 244 with reference (REF) or alternative (ALT) allele at those sites. Windows were defined from all HLA alleles present in the 930 samples from the PAG2014 dataset.

To test this hypothesis, we aligned sequences of all alleles present in PAG2014 to the *HLA* sequences present in the hg19 build of the reference human genome (the same sequences used for the alignment of reads in the 1000 Genomes Project) and defined windows of 51 basepairs around each SNP. We then quantified the number of differences with respect to the reference genome for windows surrounding i) reference (REF) and ii) alternative (ALT) alleles. If reference allele mapping bias is driving errors in frequency estimation, it is expected that sites with an overestimation of reference allele frequency would present the following pattern: windows carrying the reference allele (REF) with fewer differences to the reference

genome than sequences centered on the alternative allele (ALT). For sites with well estimated frequencies, on the other hand, we did not expect such a difference between REF and ALT windows.

To illustrate this effect, Figure 1.5 shows the results for the two most extreme cases of frequency deviation shown in Figure 1.4: site 244 of *HLA-DQB1* and site 132 of *HLA-B* (0.56 and 0.52 absolute increase in reference allele frequency in the 1000 Genomes data with respect to PAG2014). In both cases, ALT windows bear more differences to the reference sequence than REF windows.

These results support the hypothesis that these sites with poorly estimated allele frequencies have their alternative alleles residing in haplotypes with substantially more differences with respect to the reference genome than haplotypes centered on the reference allele, thus accounting for the observed bias.

To gain a broader perspective of this issue we classified SNPs from the HLA loci with reference allele bias (*HLA-A*, *-B*, and *-DQB1*) into three categories: i) sites where the reference allele frequency was overestimated, i.e. $FE > 0.1$ (“overestimated”), ii) sites where the reference allele frequency was underestimated, i.e. $FE < -0.1$ (“underestimated”) and iii) sites where allele frequencies were well estimated ($|FE| < 0.01$, here referred to as “well estimated”). We compared these three categories of sites with respect to the number of differences relative to the reference genome in REF and ALT windows (Figure 1.6).

Figure 1.6. Number of differences to the reference genome at 51bp windows centered at each SNP in the *HLA-A*, *-B* and *-DQB1* genes. Windows around each SNP were defined from the set of 1860 alleles present in the 930 samples from the PAG2014 dataset. Next, the set of windows was divided in three groups: those centered on SNPs with *overestimated*, *well estimated* and *underestimated* reference allele frequencies (red, yellow and blue boxplots, respectively). Then, each group was divided in two: windows in which the central site contains the reference allele (*REF*, dark boxplots) and windows centered on an alternative allele (*ALT*, light colored boxplots). Upper and lower hinges correspond to the 25th and 75th percentiles, horizontal lines represent the median, whiskers are 1.5 times the interquartile range, and outliers are represented by dots.

We found that the *overestimated* group has significant excess of differences at alternative allele bearing haplotypes. In this group of SNPs, ALT windows have on average 4.4 other differences relative to the reference genome, while those centered on the reference allele (REF) have 1.9 differences (excess of differences on windows centered on the alternative allele was tested with a one tailed Mann-Whitney test; $p\text{-value} < 10^{-16}$). Sites with *well estimated* or *underestimated* reference allele frequency, on the other hand, do not show a similar excess of differences in the haplotypes bearing the alternative allele, although the difference between REF and ALT windows is statistically significant due to the large sample size (*well estimated*: ALT mean=1.7; REF mean=1.8; one tailed Mann-Whitney test $p\text{-value} < 10^{-16}$; *underestimated*: ALT mean = 1.9 ; REF mean = 1.2; one tailed Mann-Whitney $p\text{-value} < 10^{-16}$).

1.3.6 Impact of biases in frequency estimation to population genetic statistics

Our analysis was able to identify a subset of SNPs in the HLA genes for which genotype calls and allele frequency estimates from the 1000G showed a high error rate with respect to the PAG2014 dataset. To evaluate the impact of the errors introduced by including these sites in population genetic analyses, we compared the distribution of sample heterozygosity between the sites with low and high error rates. Heterozygosity is defined as $H = 2p(1 - p)$ for biallelic loci, as is the case for the 1000 Genomes Phase I SNPs, since tri- or quad-allelic SNPs were not reported on Phase I.

The removal of sites with poor frequency estimates ($|FE| > 0.1$ in at least two populations) results in a marked change in the distribution of H , with a significant drop in the frequency of sites with large H and a shift in the distribution towards lower values (Figure 1.7).

Figure 1.7. Heterozygosity of SNPs at HLA genes estimated from the PAG2014 dataset. Orange bars show distribution of heterozygosity at sites with a high error rate in frequency estimation ($|FE| > 0.1$ in two or more populations). Blue bars show the distribution of heterozygosity after exclusion of SNPs with high error rate.

Note that the H values in Figure 1.7 are estimated from the PAG2014 data, implying that the high values of H among "excluded" sites are not due to the deviations in allele frequencies generated by NGS errors, but are the true heterozygosities at those sites. These results therefore document that because sites with high heterozygosity tend to have greater deviations from the "true" frequency (*i.e.*, based on the PAG2014 dataset), the removal of poorly estimated sites results in a reduction in H values.

1.3.7 The effect of heterozygosity on allele frequency estimation bias

We found an overall positive correlation between SNP heterozygosity and the magnitude of error in allele frequency estimates (Figure 1.8a; Pearson's correlation = 0.32; p-value = 1.938×10^{-7}). This result provides further evidence that sites with higher heterozygosity tend to have poorer estimates for allele frequencies in the 1000G. Also, heterozygosity is even more strongly correlated to the deviation in frequency, considering the direction of the deviation (Figure 1.8b; Pearson's correlation = 0.59; p-value $< 10^{-16}$). Together, these results show that HLA SNPs with higher heterozygosities not only have more errors in frequency estimation but also a stronger bias towards overestimation of reference allele frequency.

Figure 1.8. Relationship between SNP heterozygosity (H) and (a) absolute value of deviation ($|FE|$; Pearson's correlation = 0.32; p-value = 1.938×10^{-7}) or (b) magnitude and direction of deviation (FE ; Pearson's correlation = 0.59; p-value $< 10^{-16}$).

1.4 Discussion

The 1000 Genomes Project data were generated by various sequencing centers, which relied on different sequencing platforms, read lengths, aligners and variant and genotype calling algorithms (The 1000 Genomes Project Consortium, 2012a), creating challenges to an overall assessment of data reliability. In this study, we specifically examine the performance of NGS-based genotype calls and allele frequency estimates for the highly polymorphic and intensely studied classical HLA genes. We took advantage of the possibility of comparing downstream genotype calls from the 1000 Genomes and HLA typing based on Sanger sequencing for the same set of samples to assess data quality and test hypothesis about possible biases.

We show that the 1000 Genomes SNPs called in the HLA genes have many differences at the genotype level, when compared to results obtained using Sanger sequencing. However, considerably high genotype mismatching is possible with only modest deviations in allele frequencies, and we conclude that for the 1000 Genomes data allele frequency estimates for SNPs at HLA genes are considerably more reliable than the individual genotype calls.

Low coverage did not explain the errors in genotypes and allele frequencies in the 1000 Genomes dataset. Instead, we found evidence that read mapping bias was responsible for those errors. Mapping bias is well known for NGS, and highly polymorphic regions such as HLA genes are specially susceptible to its effects (Nielsen *et al.*, 2011), particularly when a single reference genome is used as an index for the alignment of NGS reads. In this situation, many true variants fail to be identified because they are present in haplotypes that differ from the genome used as index, and thus reads generated from these regions are not aligned and are lost. Together, these results suggest that increasing coverage would not improve allele frequency estimates at those sites if a single reference sequence is still used as index. By mapping to multiple genomes (*e.g.*, using strategies similar to Boegel *et al.* (2012) or Diltney *et al.* (2015)), it would be possible to improve genotype calling and allele frequency estimates.

In our study, *HLA-A*, *-B*, and *DQB1* show evidence of reference allele mapping bias. The *HLA-DRB1* locus, on the other hand, did not present reference allele frequency overestimation, a finding that can be explained by the existence of multiple copies of this gene (both pseudogenes and functional copies), which may result in biases/errors that make reference allele bias comparatively less visible (Degner *et al.*, 2009). The *HLA-C* locus also shows a weaker reference allele bias, a pattern that may be explained by its lower degree of polymorphism which leads to a decrease in the number of mismatches of reads with respect to the reference genome, thus decreasing the mapping bias.

We provide a list of unreliable SNPs within the HLA genes, defined by us as those with an absolute difference in frequency larger than 0.1 ($|FE| > 0.1$) in two or more populations (Table S1.4). We show that these unreliable SNPs on average have higher heterozygosities in our gold standard dataset. As a consequence, although filtering out those unreliable sites improves the overall accuracy in allele frequency estimation, it leads to an underestimation of the mean heterozygosity of SNPs in HLA genes, a bias that should be taken into account in downstream analyses. Analyses that require genotype calls at the individual level, including haplotype-based analyses, should be performed with caution when using the data from the 1000 Genomes at HLA genes.

Our results have implications to studies that use SNP data from the 1000 Genomes in other genomic regions with high variability, such as KIR and olfactory receptors. Because HLA loci are the most polymorphic in the human genome, they represent a worst case scenario for mapping bias and subsequent allele frequency estimation errors. We found a significant correlation between SNP heterozygosity and the absolute difference in frequency between 1000 Genomes data and our gold standard. This suggests that in genome-wide studies, SNPs with high heterozygosities, and contained within regions with additional SNPs, have an increased chance of presenting poor frequency estimates.

How balancing selection shapes population differentiation at the HLA genes

This chapter is presented in the form of manuscript, which is currently in preparation to be submitted for publication. This study was co-supervised by professor Diogo Meyer, from the University of São Paulo, and professor Jérôme Goudet, from the University of Lausanne.

Abstract

Balancing selection is defined as any kind of selective regime that increases genetic variability in populations relative to what is expected under neutrality. Theory predicts that balancing selection reduces population differentiation. However, balancing selection regimes in which different sets of alleles are maintained in different populations could increase population differentiation. To better understand the effects of balancing selection on the distribution of genetic variation among populations, we investigated population differentiation at the HLA genes, which constitute the most striking example of balancing selection in humans. We analysed the public dataset of the 1000 Genomes project, which sequenced 1092 individuals from different populations. We demonstrate the challenges of using a dataset rich in rare variants in population differentiation studies. Controlling for this effect, we found that population differentiation of single nucleotide polymorphisms (SNPs) at the HLA genes is lower than that of SNPs in other genomic regions. This suggests a predominant role of selective pressures shared among populations at a global scale. However, we also show evidence that population differentiation of HLA haplotypes may be higher than what we observe at the SNP level, suggesting that local selective pressures may influence the distribution of haplotypes among populations. Altogether, our results indicate that it is possible to reconcile low population differentiation at the SNP level - as predicted by theory - to higher differentiation at haplotypes, which are possibly subject to local selective pressures.

Resumo

Seleção balanceadora é definida como aquela que aumenta a variabilidade genética de uma população em relação ao esperado sob neutralidade. Uma expectativa teórica sobre seus efeitos é a redução da diferenciação populacional nos genes onde atua. Contudo, regimes de seleção balanceadora que mantêm conjuntos distintos de alelos em diferentes populações poderiam resultar em aumento de diferenciação populacional. Com o objetivo de compreender melhor os efeitos da seleção balanceadora sobre a distribuição da variação genética entre populações, investigamos a diferenciação populacional em genes HLA, uma vez que estes constituem o exemplo mais clássico de seleção balanceadora em humanos. Neste estudo, analisamos

dados públicos do projeto 1000 Genomas, que sequenciou os genomas de 1092 indivíduos de diferentes populações. Chamamos a atenção para os desafios de trabalhar com dados que apresentam uma abundância de variantes raras. Controlando para esse efeito, descobrimos que a diferenciação populacional de polimorfismos de nucleotídeo único (SNPs) nos genes HLA é menor que a diferenciação de SNPs em outras regiões do genoma. Esse resultado aponta para um papel dominante de pressões seletivas partilhadas entre populações globalmente distribuídas. Contudo, apresentamos também evidências de que a diferenciação populacional de haplótipos nos genes HLA pode ser maior do que a observada no nível dos SNPs, sugerindo que pressões locais podem influenciar a distribuição de haplótipos entre populações. Nossos achados indicam que é possível reconciliar baixa diferenciação populacional em SNPs - como predito por estudos teóricos - com maior diferenciação em haplótipos, que possivelmente estão sujeitos a pressões seletivas locais.

2.1 Introduction

Natural selection shapes the genetic variation within and the differentiation between populations. In the case of a locus where a variant is favored in one population but not in another (*i.e.*, in which selection drives local adaptation), we expect differentiation to exceed that expected under purely demographic processes (Lewontin & Krakauer, 1973). This is the case of well known examples, such as the regulatory variant that generates lactase persistence in adulthood, which is more frequent in Europeans (Bersaglieri *et al.*, 2004) and variants of the *EPAS* gene that provide adaptation to high altitude in Tibetans (Xu *et al.*, 2011). Purifying selection, on the other hand, is expected to remove functional variants which contribute to among population differences, and is expected to reduce genetic differentiation with respect to that expected under a strictly neutral scenario (*e.g.* Barreiro *et al.* (2008)).

A third regime, balancing selection, is related to diversity and differentiation in more complex ways. By definition balancing selection encompasses all selective regimes that result in increased genetic variation with respect to neutral expectations. There are several distinct modes of balancing selection: heterozygote advantage, frequency dependent selection and selection that varies over time and space (Andrés, 2011).

The loci with strongest evidence for balancing selection in the human genome are the classical HLA class I and II loci (especially the *HLA-A*, *-B*, *-C*, *-DRB1*, *-DQB1*, *-DQA1* and *-DRA1* loci). Those loci are located in the Major Histocompatibility Complex (MHC) region, and present homologues in all jawed vertebrates. The HLA genes are responsible for binding peptides to be presented to T-cell receptors, thus mediating a critical step of the adaptive immune response (Klein & Sato, 2000). Balancing selection at these loci has been strongly supported by a wide variety of methods, and evidence includes an excess of alleles at intermediate frequency with respect to neutral expectations (Hedrick & Thomson, 1983), high linkage disequilibrium (?), higher non-synonymous to synonymous substitution rate (Hughes & Nei, 1988) and trans-specific polymorphism (Lawlor *et al.*, 1988) (Meyer & Thomson, 2001, for a review).

Balancing selection at HLA genes is well documented, but the evidence presented above is compatible with regimes that predict both increased or reduced population differentiation. For example, the regime of frequency dependent selection, or rare allele advantage, could be acting due to the coevolution between HLA and pathogens. In this scenario, rare HLA alleles would be advantageous, since few pathogens would have evolved resistance to them (Meyer & Thomson, 2001). Rare allele advantage is expected to increase the effective migration rate: migrants will often be rare in the population they arrive to, and thus will be advantageous and increase in frequency in the new population (Schierup *et al.*, 2000; Muirhead, 2001). Therefore, this regime of balancing selection is expected to reduce population differentiation.

However, in the case of HLA genes, balancing selection through population-specific selective pressures, where the different sets of pathogens in each population select locally advantageous variants, is also plausible, and is expected to increase population differentiation. One line of evidence supporting this hypothesis is the finding that HLA and pathogen diversities are correlated (Prugnolle *et al.*, 2005). This regime also implies that specific HLA alleles are more effective in presenting antigens of certain pathogens. Evidence supporting this assumption come from studies reporting significant associations between disease susceptibility, resistance or progression with variation at HLA (reviewed in Chapman & Hill, 2012), and from theoretical studies documenting the ability of this kind of pathogen driven selection models to explain observed levels of HLA variation (Hedrick, 2002; Borghans *et al.*, 2004).

Previous studies which measured population differentiation at the HLA genes also found evidence for both increased or reduced population differentiation. For example, Meyer *et al.* (2006) found no significant difference between differentiation at HLA genes and a set of neutral markers used as a control (microsatellites), while Sanchez-Mazas (2007) found lower differentiation at HLA loci than in their genome-wide control (microsatellites and RFLPs). A limitation of these studies is that differences between the neutral genetic markers and the sequence data used for HLA genes introduce confounding variables, making it difficult to determine the roles of selection or characteristics inherent to the marker (*e.g.* mutation rate and diversity).

A more recent study compared differentiation on markers of the same type (microsatellites) located at HLA genes or near them and those located in other genomic regions, which serve as controls. In this study, Nunes (2011) found increased differentiation in regions near HLA genes. Nonetheless, some issues remain unresolved: Nunes (2011) was exclusively interested in native American populations, and used a limited number of markers. Furthermore, the complexity of the mutational mechanism of microsatellites complicates the interpretation of results.

For non-model organisms a similarly wide array of results have been found, with the MHC region showing either equal (Miller *et al.*, 2010), higher (Loiseau *et al.*, 2009; Oliver *et al.*, 2009; Cammen *et al.*, 2011) or lower (McCairns *et al.*, 2011) differentiation than genome-wide averages.

In summary, it remains unclear whether balancing selection on HLA genes drives increased differentiation due to selection favoring adaptation to locally occurring pathogens, or whether

it results in decreased genetic differentiation, due to the maintenance of shared polymorphisms among populations.

Here, we revisit the question of population differentiation at the HLA genes. Initially, we focus on SNP level analyses, which have the advantage of allowing the use of genomic controls. We used data from the 1000 Genomes Project phase I (The 1000 Genomes Project Consortium, 2012a), which sequenced 1092 individuals from 14 populations and identified almost 37 million SNPs. We compare differentiation at SNPs within and close to HLA genes to that of SNPs in other genomic regions, which we treated as a control group.

Differently from SNP-based studies such as scans that seek genome-wide significance for specific SNPs, we rely on the fact that we have an *a priori* grouping of SNPs we are querying (those within or around HLA genes). This allows us to test for significant differences between groups of SNPs, which is more powerful than seeking individually significant SNPs. Finally, we consider the fact that the entity that defines the fitness of an individual may be more complex than individual SNPs, and thus compare the findings regarding differentiation at the SNP level to those found at the haplotype level.

2.2 Methods

2.2.1 SNP Data

SNP genotypes were acquired from the integrated Variant Call Format (VCF) file from version 3 of the 1000 Genomes Project Phase I data (The 1000 Genomes Project Consortium, 2012a), which is available at <ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/>.

This dataset includes variants discovered via high coverage exome targeted resequencing and low coverage resequencing. Because of higher coverage in exonic regions, all comparisons between HLA SNPs and a group of non-HLA SNPs treated as a control were made within the same functional category (e.g. intronic or exonic). To this end SNPs were annotated using ANNOVAR, which classified each SNP in one of the following categories: exonic, splicing, UTR5, UTR3, intronic, upstream or downstream of a gene, intergenic, and non coding RNA (Wang *et al.*, 2010). Our findings for differences between HLA and the non-HLA control were qualitatively the same when using either exonic or intronic regions, so throughout the paper we focus on the results for exonic regions.

The 1000 Genomes Project Phase I data contains the genomes of 1092 individuals from 14 populations. We note that Phase I of the 1000 Genomes Project contains only bi-allelic variants. All cases of tri- or quad-allelic SNPs were reported as bi-allelic, reporting all alternative alleles as the most frequent alternative allele. Phase I data were used because the available analysis of ancestry allowed us to select only non-admixed populations, as described below (The 1000 Genomes Project Consortium, 2012b). Among the 14 populations, four are admixed: Americans of African Ancestry in SW USA (ASW), Mexican Ancestry from Los Angeles USA (MXL), Puerto Ricans from Puerto Rico (PUR) and Colombians from Mendellin, Colombia (CLM). Those populations were excluded from our analyses because they would decrease our overall levels of population differentiation due to an event in a different timescale (much more

recent) than the ones causing differentiation among the other native populations, and this could hamper the interpretation of our results. We have also excluded one of the Han Chinese populations - the Southern Han Chinese (CHS) - because keeping a pair of populations which is much more similar than any other pair would also disproportionally decrease the overall levels of population differentiation in our dataset.

Finally, we excluded the Iberian Population in Spain (IBS) because of its reduced sample size (14 individuals). After applying those filters, a total of 736 individuals in 8 populations were kept (see Table S2.1).

2.2.2 Minor allele frequency calculations

Allele frequency (AF) was calculated per SNP, per population using VCFtools v0.1.12b (Danecek *et al.*, 2011). Minor allele frequency (MAF) was calculated as $1 - AF$, when $AF > 0.5$. Global minor allele frequency (over all populations) was calculated as an average of MAFs per population, weighted by sample size.

2.2.3 Estimation of F_{ST}

Population differentiation was calculated as a ratio between the components of the variance in allele frequency among (a) and within (w) populations:

$$F_{ST} = \frac{a}{a + w} \quad (2.1)$$

It was estimated using the Weir & Cockerham (1984) F_{ST} estimator implemented in VCFtools v0.1.12b (Danecek *et al.*, 2011). The Weir & Cockerham (1984) F_{ST} estimator was chosen due to the fact that it is an unbiased estimator of F_{ST} when sample sizes are large and similar, as in the case of our dataset. F_{ST} was calculated per SNP i) over all populations and ii) between each pair of populations.

When summarizing F_{ST} over multiple loci, we used the approach suggested by Reynolds *et al.* (1983): we first computed the components of variance among (a) and within populations (w) for each locus, then we computed the average of a and $a + w$ for the desired set of loci, and finally computed the ratio of both averages. It was shown that this approach provides the least biased estimate of F_{ST} , while performing a simple average of the F_{ST} s of each SNP leads to an underestimation of differentiation, especially in datasets rich in rare variants (Reynolds *et al.*, 1983; Bhatia *et al.*, 2013). Variance components (a and $a + w$) were estimated using a minor modification of the VCFtools source code, to include those values in the output of the program.

2.2.4 Definition of HLA, peri-HLA and control regions

We define "HLA SNPs" as those contained within the classical HLA genes *HLA-A*, *-B*, *-C*, *-DRA*, *-DRB1*, *-DQA*, *-DQB1*. The classical *HLA-DPA1* and *-DPB1* genes were excluded from our analysis unless otherwise mentioned. This was done because previous studies have shown weak or no evidence of balancing selection in them (Solberg *et al.*, 2008; Begovich *et al.*, 2001), and even instances of directional selection (Hollenbach *et al.*, 2001), making them

inappropriate for our central question concerning the role of balancing selection on population differentiation at HLA loci.

Peri-HLA genes were defined as those that flank the HLA genes and have higher diversity relative to the average of chromosome 6 in the Yoruba (YRI) population (Mendes, 2013), indicating that their increased polymorphism is driven by hitchhiking due to the strongly selected HLA loci. The 25 peri-HLA genes are shown in Supporting Information (Appendix S2.2).

All SNPs contained outside the HLA and peri-HLA genes were included in the control group.

2.2.5 Filter out SNPs with unreliable allele frequency estimates

The high level of variability at the HLA genes makes this region very challenging for the alignment of short reads generated by Next Generation Sequencing (NGS) platforms as the ones used in the 1000 Genomes Project.

We have shown that the use of a single reference genome to map reads creates a strong mapping bias at some HLA SNPs in the 1000 Genomes Project dataset (Brandt *et al.*, 2015). This bias is especially strong at SNPs which are part of haplotypes that contain additional variants that differ from the reference sequence. We excluded SNPs in the HLA genes which were shown to have unreliable frequency estimates in the 1000 Genomes phase I dataset, due to mapping bias. After applying this filter, 101 out of 251 SNPs tested for mapping bias were excluded. Those SNPs are listed on Table S1.4.

In total, 2775 SNPs were kept in the HLA group, 6895 in the peri HLA group and 36820992 in the control group. Among those, the number of exonic SNPs are: 388 in the HLA, 688 in the peri HLA and 481610 in the control.

2.2.6 Accounting for linkage disequilibrium among HLA SNPs

The set of HLA SNPs are on average under much higher linkage disequilibrium (LD) than the SNPs in the control group, to which we compared them. To account for this effect we compare the F_{ST} values of HLA SNPs with a set control SNPs that was matched for similar levels of LD. To this end, we selected the control SNPs in the following way. We sampled a random exonic SNP from outside the HLA and peri-HLA genes and took all other SNPs that were in the same gene. We repeated this operation until we reached a total number of SNPs that was similar to that in the set of HLA SNPs, for each interval of MAF. We then compared the F_{ST} distribution of this set of linked control SNPs to that of the HLA SNPs, applying a χ^2 test (see Figure S2.1). This resampling algorithm was repeated 1,000 times and we quantified the number of resamplings in which the F_{ST} distribution at the HLA SNPs was not significantly different from that of the resampled SNPs.

2.2.7 Simulations for maximum F_{ST} as a function of MAF

In order to gain a better understanding on how F_{ST} is constrained by SNP allele frequencies, we simulated bi-allelic loci evolving neutrally in a metapopulation with low migration rate, allowing populations to achieve high differentiation.

Simulations were performed using the *sim.genot* function of the hierfstat R package (Goudet, 2005). We simulated the neutral evolution of 10,000 bi-allelic loci (SNPs) in 10 populations, each with population size 1,000, virtually no migration among them (proportion of migration = 10^{-5}) and very low mutation rate (10^{-8}). This allowed SNPs to achieve maximum F_{ST} values possible given their MAF.

2.2.8 Haplotype level analyses

In addition to SNP based analyses done using the 1000 Genomes dataset, we investigated population differentiation at the HLA genes at the level of alleles, defined by the entire coding sequence of the HLA genes. We will treat this set of analyses as "haplotype level" analyses.

HLA genotypes at the haplotype level for 1000 Genomes samples from YRI, CEU and CHB populations (see Table S2.1 for full names) were acquired from Gourraud *et al.* (2014). Data is available at the dbMHC website (<http://www.ncbi.nlm.nih.gov/gv/mhc/xslcgi.fcgi?cmd=cellsearch>). Note that, differently from previous SNP based analyses, here we used data from only 3 populations from the 1000 Genomes. Also, we have only performed the haplotype based analyses for class I loci (*HLA-A*, *-B* and *-C*).

To allow F_{ST} values at the HLA haplotypes to be compared to a null distribution, as was done for the SNPs, we recoded each HLA gene as a biallelic locus. This recoding was done by treating the most frequent allele at each gene as "allele 1", and all other alleles as "allele 2". Finally, F_{ST} was computed as described for SNPs.

2.3 Results

2.3.1 Higher overall F_{ST} in HLA

In an initial analysis, we compared the distribution of F_{ST} between SNPs of the HLA and control groups (Figure 2.1).

Figure 2.1. Distribution of F_{ST} values for SNP from the following groups: intergenic, exonic control group, HLA genes and at peri-HLA genes.

The distribution of F_{ST} values at SNPs within HLA genes is shifted towards higher values compared to the control SNPs and the difference is significant (p-value $< 10^{-16}$, one-tailed Mann-Whitney-Wilcoxon test).

Theory predicts that balancing selection affects only a narrow genomic region, the size of which is defined by the intensity of selection and the recombination rate (Charlesworth *et al.*, 1997). To evaluate if the increased differentiation at coding SNPs within HLA genes was also observed in loci that flank the HLA, we applied the same test to the peri-HLA genes. As is the case for the HLA SNPs, the peri-HLA SNPs have an F_{ST} distribution which is shifted to significantly higher values, for both introns and exons (p-value $< 10^{-16}$, one-tailed Mann-Whitney-Wilcoxon test; Figure 2.1).

Although we include the intergenic data in Figure 2.1 as a reference, we note that the difference in coverage between intergenic and exonic data makes a direct comparison between these inappropriate (see Methods). Therefore, in the remainder of the paper, we focus on contrasts between exonic SNPs within and outside HLA loci.

2.3.2 Constraints of Minor Allele Frequency on F_{ST}

Allele frequencies impose constraints on the F_{ST} value (Roesti *et al.*, 2012; Maruki *et al.*, 2012; Elhaik, 2012; Jakobsson *et al.*, 2013), with SNPs that have very low minor allele frequencies (MAF) being bounded to low F_{ST} values.

Based on the approach of Maruki *et al.* (2012), we have analytically investigated the behavior of maximum F_{ST} among multiple populations as a function of MAF.

Assuming all populations are the same size, we define p_i as the frequency of the minor allele in population i , and $p = \frac{\sum_{i=1}^n p_i}{n}$ as the minor allele frequency over all n populations. We considered the expected value of F_{ST} as defined by

$$F_{ST} = \frac{H_T - H_S}{H_T}, \quad (2.2)$$

where

$$H_T = 2p(1-p) \quad H_S = \frac{\sum_{i=1}^n 2p_i(1-p_i)}{n}. \quad (2.3)$$

In this scenario, maximum F_{ST} for a given p is achieved when all occurrences of the minor allele are concentrated in as few populations as possible, *i.e.* when $\lfloor pn \rfloor$ * populations are fixed for the minor allele, and the remainder of minor alleles are all in the same population, with frequency $p^* = pn - \lfloor pn \rfloor$.

All populations that are fixed for either the minor or major allele will not contribute to H_S , since either p or $1-p$ will be zero. So the H_S formula in the scenario of maximum F_{ST} can be simplified to

$$H_{S_{maxF_{ST}}} = \frac{2p^*(1-p^*)}{n}. \quad (2.4)$$

*The notation $\lfloor \cdot \rfloor$ represents the integer part of the number pn

For example, with $n = 10$ populations and MAF of $p = 0.15$, maximum F_{ST} will be achieved when $\lfloor 1.5 \rfloor = 1$ population is fixed for the minor allele, 1 population has $p^* = 1.5 - 1 = 0.5$ minor alleles, and the remainder 8 populations are fixed for the major allele. In this case, maximum F_{ST} is 0.8.

To illustrate this, we have simulated the neutral evolution of SNPs in 10 populations, with virtually no migration among them, allowing SNPs to achieve maximum differentiation among populations. Figure 2.2 shows the F_{ST} values of 10,000 simulated SNPs as a function of their MAFs, as well as the maximum F_{ST} values calculated as described above.

Figure 2.2. Maximum F_{ST} as a function of MAF for 10 populations. Maximum F_{ST} is achieved when all minor alleles are concentrated in as few populations as possible. Black line shows maximum F_{ST} as a function of MAF, calculated as described in main text. Gray points are simulations of biallelic SNPs evolving neutrally in 10 populations, with no migration among them, which allows them to achieve maximum F_{ST} .

Figure 2.3. Population differentiation, measured by F_{ST} , as a function of minor allele frequency (MAF) for the exonic SNPs from the 1000 Genomes Project phase I data.

This relationship is empirically illustrated for the 1000 Genomes exome data in Figure 2.3, which shows that the constraint of F_{ST} to low values is stronger for SNPs with lower MAFs, and only SNPs with $MAF > 0.2$ achieve $F_{ST} > 0.75$.

Our analysis shows that the constraint of MAF on F_{ST} values is particularly strong in the range of low MAF (up to 0.2). After this MAF value the constraint is no longer evident. This suggests that in datasets with large numbers of rare variants the site frequency spectrum (SFS) of the SNPs being studied will have a profound effect on the F_{ST} value which is estimated.

2.3.3 Lower F_{ST} in HLA when accounting for MAF

In view of the constraints imposed by MAF on F_{ST} values, we examined whether the difference in F_{ST} values between SNPs in HLA genes and in our control group were determined by the difference in abundance of low MAF SNPs in HLA and non-HLA loci.

Balancing selection shifts the site frequency spectrum of the selected locus toward an excess of intermediate frequency variants. This expectation was confirmed for the HLA SNPs (Figure 2.4), while most control SNPs have extremely low MAF and peri-HLA SNPs occupy intermediate values. This finding suggests that the difference in F_{ST} for the HLA and control SNPs may be simply accounted for by the relative paucity of low MAF SNPs in the HLA group.

Figure 2.4. Distribution of minor allele frequency (MAF) for exonic control, peri-HLA genes and HLA genes. There is a depletion of low-frequency SNPs at the HLA and peri-HLA genes compared to other genes.

To take into account the effects of differences in SFS between HLA and control SNPs when comparing population differentiation among those groups, we compared the F_{ST} of HLA and control SNPs within bins of MAF (Figure 2.5). For all bins with MAF higher than 0.01 the F_{ST} at HLA and peri-HLA SNPs was significantly lower than at the control SNPs. For the lowest MAF bin (<0.01), the distribution of F_{ST} at HLA SNPs was not significantly different from the control SNPs distribution, and it was significantly higher at peri-HLA SNPs.

Figure 2.5. F_{ST} distributions as a function of minor allele frequency (MAF) intervals. SNPs at the HLA and peri-HLA genes show lower F_{ST} than exonic control SNPs in all intervals of $MAF > 0.01$. * Mann-Whitney one-tailed test p-value $< 10^{-9}$. Outliers were removed from figure (but not from statistical test) for better visualization.

Another strategy to control for the constraint of MAF on F_{ST} was applied by Bhatia *et al.* (2013), in the context of averaging F_{ST} over loci. Reynolds *et al.* (1983) showed that the less biased estimate of F_{ST} for multiple loci is obtained by computing a "ratio of the averages" of the numerator and denominator of the AMOVA-based F_{ST} estimator (see Methods).

However, Bhatia *et al.* (2013) have shown that a rather naïve approach was chosen by the 1000 Genomes Project when they reported their results in 2010: they reported a simple average of F_{ST} s at individual SNPs (*i.e.* an "average of ratios").

Bhatia *et al.* (2013) discuss that the "average of ratios" approach creates a bias leading to underestimation of F_{ST} , and that this bias is stronger in datasets rich in rare variants. Instead, the desirable averaging approach (ratio of averages) gives less weight to variants with low MAF, which are less informative about population differentiation.

We compared the result of both averaging approaches on our data, and we found that the "ratio of averages" approach (which controls for MAF) reveals lower F_{ST} at the HLA SNPs than in the control SNPs. In contrast, the simple mean of individual F_{ST} values for the HLA SNPs is higher than genome-wide, as in our initial analysis that did not account for MAF (Table 2.1). This further emphasizes the need of taking the differences of the site frequency spectrum of HLA and control into account when assessing population differentiation.

Table 2.1. Mean F_{ST} at the HLA and control SNPs using two averaging approaches.

	HLA	control
Average of ratios	0.04	0.02
Ratio of averages	0.06	0.13

HLA-DP genes

The classical *HLA-DPA1* and *-DPB1* genes were excluded from the analysis reported above. This was done because previous studies have shown weak or no evidence of balancing selection in them (Solberg *et al.*, 2008; Begovich *et al.*, 2001), and even instances of directional selection (Hollenbach *et al.*, 2001), making them inappropriate for our central question concerning the role of balancing selection on population differentiation. Consistently with our expectations, our results revealed that *HLA-DPA1* and *-DPB1* show a pattern of population differentiation which is very different from the other classical HLA loci: population differentiation at those genes is higher than in the genomic control, even when minor allele frequency is accounted for (see Figure S2.2).

2.3.4 Accounting for linkage disequilibrium among HLA SNPs

The above approach used SNPs outside the HLA and peri-HLA genes to provide a control for the HLA SNPs. However, because the HLA SNPs are located in 5 loci, they are not independent, due to both intra and inter-locus associations. As a consequence, our p-values could be inflated by treating a set of correlated SNPs as independent, and comparing them to a set of control SNPs which are in their majority independent.

To control for this non-independence, we used a resampling approach, in which sets of linked SNPs were sampled from our control group, and their F_{ST} values were compared to those from the HLA SNPs, at each MAF bin (see Methods). We repeated this resampling approach 1000 times and each time the difference between F_{ST} distributions from the sampled SNPs and HLA SNPs at each MAF bin was tested.

At all bins of MAF higher than 0.01, F_{ST} at HLA SNPs was significantly lower than at the resampled SNPs, with the only exception of two samples of linked SNPs in which F_{ST} was not significantly different from HLA (χ^2 test p-values > 0.05) in one of the MAF bins (Table 2.2). Thus, we show that F_{ST} at the HLA SNPs is significantly lower than in our control when the non-independence of the HLA SNPs is accounted for.

Table 2.2. Number of samples of linked control SNPs whose F_{ST} is not significantly higher than the F_{ST} of HLA SNPs at each minor allele frequency (MAF) bin, as observed in the total dataset.

MAF	<0.01	0.01-0.1	0.1-0.2	0.2-0.3	0.3-0.4	>0.4
Samples whose F_{ST} distribution is not significantly higher than HLA	0.437	0	0	0	0.002	0

2.3.5 Divergence time is related to lower F_{ST} at HLA SNPs

Our previous analyses examined a global F_{ST} , which captures patterns of differentiation among all 8 populations in our sample. Next, we asked how specific populations were contributing to our findings. In order to investigate this question, and to evaluate how the geographical scale (within and among continents), influences differentiation at the HLA, we have analyzed F_{ST} between pairs of populations.

We found that the pattern of lower differentiation at HLA SNPs than at the control SNPs demonstrated in our previous results (Figure 2.5) arises only on a broader geographical scale (between continents), between pairs of more diverged populations (Figure 2.6). Population pairs that are within the same continent show similar or slightly higher differentiation at the HLA SNPs compared to control SNPs.

Figure 2.6. Average F_{ST} at the HLA and at the control SNPs for each pair of populations. Red squares represent African populations, blue triangles are European populations and green crosses are Asian populations. Points with a single color represent pairs of populations from the same continent.

In summary, we have shown that at the SNP level HLA genes show lower population differentiation when the appropriate control for minor allele frequency is applied. We explored the geographical scale of this global pattern and found that it is mostly driven by differences among distantly related populations.

2.3.6 Differentiation at HLA alleles (haplotypes) and SNPs

Next, we explored population differentiation at the haplotype level (*i.e.* the whole coding sequence of each HLA gene). Haplotype level analyses were motivated by the idea that the fitness of individuals might be defined by the combination of SNPs they carry in each gene, which will determine the peptides the HLA molecules will be able to present, rather than fitness being defined by individual SNPs.

Diversity is known to pose constraints on population differentiation measures (Jakobsson *et al.*, 2013). Since no other genomic region has levels of diversity comparable to that of the HLA genes, this complicates the evaluation of population differentiation at the HLA genes, when analyzing haplotypes of SNPs (which define the HLA alleles).

To be able to compare population differentiation at the HLA haplotypes to a null distribution, we recoded each HLA gene as a biallelic locus, based on the most frequent allele at each gene (see Methods). We focused on the most frequent alleles because their diversity is more likely to have been shaped by recent selection favoring a locally advantageous haplotype. Population

differentiation at the recoded HLA haplotypes was then compared to the distribution of differentiation at genome-wide distributed SNPs, while controlling for minor allele frequency, as we did for SNPs.

Differently from our finding with SNPs, variation at the HLA haplotype level does not show significantly lower differentiation than genome-wide. In the three cases analysed, the most common HLA haplotype shows levels of differentiation in the extreme high portion of the F_{ST} distribution for HLA SNPs within the same MAF interval (Figure 2.7). This indicates that while long term balancing selection has shaped the maintenance of shared polymorphisms among populations (leading to low F_{ST} at the SNP level), these are distributed into a wide diversity of haplotypes that are more differentiated among populations.

Figure 2.7. Population differentiation (F_{ST}) as a function of minor allele frequency (MAF) for the exonic control SNPs (black), for exonic HLA SNPs (red) and for the most common haplotype-level alleles at each of the *HLA-A*, *-B* and *-C* genes after recoding as biallelic (see Methods).

2.4 Discussion

Population differentiation at the HLA at the SNP level

We found significantly decreased genetic differentiation at HLA SNPs, contrary to the expectation of a model according to which adaptation to local pathogens would drive differentiation. Our findings extend to the non-coding regions and to genes flanking the HLA loci, indicating that selection on HLA loci influences variation and differentiation in a larger genomic region. This result will be further discussed together with the haplotype based results.

Our SNP based results also highlight the importance of accounting for minor allele frequency (MAF) when comparing F_{ST} values for different sets of SNPs, as previously shown by others (Roesti *et al.*, 2012; Maruki *et al.*, 2012; Elhaik, 2012; Jakobsson *et al.*, 2013). Not accounting for MAF leads to an underestimation of population differentiation for sets of SNPs rich in rare variants. When comparing sets of SNPs from genomic regions with different MAF distributions, this may lead to a misleading interpretation of the selective regime acting on each region.

In the case of a set of SNPs undergoing purifying selection, it is expected that they will be enriched in low frequency variants. It is also expected that purifying selection leads to lower population differentiation. Since population differentiation in low frequency variants is constrained to low values, the relationship between MAF and F_{ST} leads both signatures to the same direction. This effect has been demonstrated in *Drosophila melanogaster* (Jackson *et al.*, 2014).

However, in the case of balancing selection, the relationship between MAF and F_{ST} creates misleading results when not accounted for. We have shown that the depletion of rare variants on a region under balancing selection leads overall population differentiation to be higher than in other regions. However, when variants with similar MAF are compared, it is revealed that population differentiation at variants under balancing selection is actually lower than observed in variants in other regions.

Contrast to previous studies

Two classical HLA genes, *HLA-DPA1* and *HLA-DPB1*, were exceptions to our finding of lower population differentiation at HLA SNPs. SNPs in those genes showed higher population differentiation, even when MAF was accounted for.

A genomic scan performed by Bhatia *et al.* (2011) also found a SNP near a HLA-DP gene to have unusually high population differentiation (rs2179915, which is 30kb from *HLA-DPA2*). Similarly, Barreiro *et al.* (2008) found unusually high F_{ST} at the *HLA-DPB2* locus. Together, these results indicate that the HLA-DP genes are under a selective regime which is different from the other HLA genes. In addition, these studies were designed with an emphasis on the detection of extreme high differentiation, and thus did not capture the pattern of low differentiation which is characteristic of most HLA loci.

Hofer *et al.* (2012), on the other hand, used an approach where regions of adjacent SNPs with extreme F_{ST} were scanned for, and detected the *HLA-C* locus among these, with evidence of unusually low F_{ST} . These results show that when a test is designed to account for the possibility of unusually low F_{ST} values, and when a genomic region is queried (as opposed to individual SNPs), a signature of low differentiation which would otherwise not be detected can be found.

Divergence time effect

By taking advantage of multiple populations made available by the 1000 Genomes project, we were also able to examine if the excess of low differentiation at HLA SNPs holds at all timescales of differentiation. Interestingly, we find that for population pairs with low divergence (those from the same continent) some instances of higher F_{ST} at the HLA SNPs than in the genomic control are observed. For highly diverged population pairs (those from different continents), we consistently find lower differentiation among HLA SNPs. This implies that genetic differentiation at the HLA SNPs increases at a lower rate than genome-wide variation, over time.

We cannot specify if this effect has a biological basis, or simply reflects the increased power to detect low differentiation due to selection when comparing more highly diverged

populations (and symmetrically, more power to detect highly differentiated SNPs when comparing populations with low overall differentiation). However, this result shows that a specific set of SNPs may differ in how they deviate from the genomic background depending on the timescale of population divergence, implying that a single "overall" result may mask interesting patterns, and that analyses at different timescales are necessary.

Reconciling low F_{ST} at SNPs and local adaptation

Our finding of decreased population differentiation at HLA SNPs may be reconciled with the hypothesis of local adaptation when a haplotype-based analysis is considered. We have shown evidence that haplotypes at the HLA genes do not show unusually low differentiation, as SNPs do. Therefore, it is possible that pathogens vary among populations and select locally adaptive HLA haplotypes, that will be different in each population. The most adaptive haplotypes would be the ones that are more effective in presenting peptides from the most frequent pathogens in each population.

Intragenic recombination and gene conversion play an important role in generating HLA diversity at the haplotype level (Parham & Ohta, 1996). Those processes generate new haplotype-level alleles with the same SNP-level alleles, so different haplotypes at the HLA genes share many SNPs. Therefore, it is possible that the most frequent HLA haplotypes in each population are different (due to local selection), but that the allele frequencies at the SNP level are very similar.

Discussão Geral e Conclusões

A principal questão motivadora deste trabalho foi: qual o efeito da seleção balanceadora sobre a diferenciação populacional? Especificamente, investigamos duas questões principais: i) na escala global, o regime de seleção balanceadora atuante sobre os genes HLA consiste em pressões seletivas globais ou locais?; ii) como reconciliar a expectativa teórica de baixa diferenciação populacional em regiões sob seleção balanceadora à observação de alta diferenciação populacional em genes do MHC em humanos e outras espécies?

Para responder ambas questões, realizamos um estudo para determinar se a diferenciação populacional em genes HLA é inusualmente alta ou baixa em relação ao restante do genoma, comparando populações distribuídas em 3 continentes, com diferentes graus de divergência. Nesse estudo, foi essencial contar com um controle genômico contra o qual comparamos a diferenciação nos genes HLA.

Controles metodológicos

Ao realizar esse estudo, nos deparamos com importantes questões metodológicas. A primeira delas diz respeito à dificuldade em se aplicar métodos de sequenciamento de nova geração (NGS) a regiões extremamente polimórficas, como a do MHC. Usando uma abordagem empírica, comparamos os genótipos e frequências alélicas de variantes nos genes HLA obtidos a partir de NGS (do projeto 1000 Genomas) e sequenciamento Sanger, tratando esses últimos como padrão ouro. Como resultado, obtivemos uma lista de sítios cujas frequências alélicas eram confiáveis nos dados de NGS do projeto 1000 Genomas, e excluímos os sítios pouco confiáveis das análises de diferenciação populacional.

A segunda questão metodológica que abordamos foi o efeito da frequência do alelo mais raro (MAF) sobre medidas de diferenciação populacional (F_{ST}). Quando essa relação entre MAF e F_{ST} é ignorada, detectamos nos nossos dados um sinal de diferenciação aumentada nos SNPs dos genes HLA em relação ao controle genômico. Contudo, mostramos que esse sinal é gerado pelo fato de que o conjunto de marcadores que usamos como controle genômico tem predomínio de variantes extremamente raras (ou seja, com MAF muito reduzida). Ao compararmos os valores de F_{ST} dentro de cada classe de MAF, revelamos que na realidade SNPs nos genes HLA desviam na direção oposta, de baixa diferenciação, em relação ao controle genômico.

A abordagem dessas duas questões metodológicas por si só gerou resultados importantes desse trabalho, que serão discutidos em maior detalhe a seguir.

Qualidade dos dados

O efeito de viés de mapeamento de regiões altamente polimórficas a um genoma referência já era conhecido e esperado (Degner *et al.*, 2009; Nielsen *et al.*, 2011; Boegel *et al.*, 2012; Dilthey *et al.*, 2015). Contudo, nosso estudo quantificou os efeitos desse viés em um banco de dados amplamente usado (1000 Genomas), mostrando que ele explica grande parte dos erros de genótipos nesses dados. Essa quantificação poderá ser útil para pesquisadores interessados em estudar outras regiões altamente polimórficas usando os dados do projeto 1000 Genomas.

Também avaliamos a confiabilidade de sítios individuais nos genes HLA nos dados do 1000 Genomas e disponibilizamos essa informação, que pode ser útil para outros grupos interessados em HLA, e para pesquisadores que realizam estudos de associação em escala genômica (GWAS, do inglês *Genome-wide Association Study*). Nesse tipo de estudo, frequentemente são encontradas variantes nos genes HLA associadas a caracteres de interesse, mas muitas vezes estas são descartadas porque os dados nessa região são menos confiáveis. Nossos resultados poderão ser um recurso útil para pesquisadores que desejam avaliar a confiabilidade dessas variantes.

Nossas análises mostraram que os erros causados pelo viés de mapeamento não seriam atenuados com o aumento da cobertura em experimentos de NGS. A solução para esse problema parece estar no desenvolvimento de novos métodos de mapeamento de fragmentos sequenciados (*reads*). Boegel *et al.* (2012) e Dilthey *et al.* (2015), por exemplo, desenvolveram estratégias que evitam o problema do mapeamento a um único genoma referência usando um banco de dados de sequências de HLA como referência. Esses autores mostraram que esse tipo de método melhora muito a inferência de genótipos nos genes HLA. Nosso grupo também buscou uma solução para esse problema, e Aguiar *et al.* (2015, resultados não publicados) desenvolveram um método que usa um banco de dados de sequências de HLA para mapear fragmentos de sequenciamento de RNA (RNAseq), e demonstraram que esse método estima genótipos HLA com alta precisão.

Efeito da diversidade genética sobre a diferenciação populacional

A restrição imposta pela diversidade de marcadores genéticos sobre valores de F_{ST} é discutida há muito tempo, com foco principalmente em marcadores multialélicos de grande diversidade, como microssatélites. Foi demonstrado que marcadores muito polimórficos limitam F_{ST} a valores baixos (Beaumont & Nichols, 1996; Hedrick, 1999; Jost, 2008).

Nos últimos anos, passou a ser discutido o fato que em marcadores genéticos com baixa diversidade, F_{ST} também fica restrito a valores baixos (Roesti *et al.*, 2012; Maruki *et al.*, 2012; Elhaik, 2012; Jakobsson *et al.*, 2013). Contudo, até então havia pouco interesse em variantes raras, que em muitos estudos eram filtradas em etapas iniciais, sob a justificativa de serem pouco confiáveis.

Mais recentemente, a atenção às variantes raras tem crescido, movido pela observação de que a recente explosão de tamanho populacional em humanos aumenta a proporção de variantes raras e potencialmente deletérias no genoma (Keinan & Clark, 2012). Além do interesse biológico em variantes raras, o desenvolvimento tecnológico também teve papel

crucial no aumento do interesse em variantes raras. Até os anos 2000, o método predominante em genômica populacional era o uso de microarranjos de DNA. Esse método realiza a genotipagem de SNPs averiguados *a priori*, o que cria um viés para variantes mais comuns. A recente disseminação do uso de sequenciamento em escala populacional permitiu a detecção de muitas variantes raras, e a compreensão de sua abundância no genoma humano.

Nesse contexto, a relação entre a frequência do alelo mais raro (MAF) e diferenciação populacional demonstrada por Maruki *et al.* (2012) e Jakobsson *et al.* (2013) passa a ser cada vez mais relevante. Nosso estudo soma aos achados de Maruki *et al.* (2012) e Jakobsson *et al.* (2013), pois analisamos o comportamento de F_{ST} em função de MAF para múltiplas populações, enquanto aqueles estudos focaram em descrever essa relação para um par de populações.

Nosso trabalho também revelou uma outra faceta da restrição criada pela diversidade dos marcadores genéticos (MAF) sobre as medidas de diferenciação populacional (F_{ST}). Mostramos que, em regiões sob seleção balanceadora, padrões totalmente opostos emergem dos dados quando se considera ou não a relação entre MAF e F_{ST} . Esse tipo de fenômeno estatístico é conhecido como paradoxo de Yule-Simpson (Yule, 1903; Simpson, 1951), e pode ser definido como a observação de uma tendência em grupos de dados, que é invertida quando esses grupos são amalgamados. A tendência que observamos em grupos de SNPs em cada intervalo de MAF é de diferenciação menor nos genes HLA que no controle genômico. Quando amalgamamos SNPs de todos os intervalos de MAF, observamos a tendência oposta, ou seja, maior diferenciação nos genes HLA que no controle genômico (Tabela 3).

Tabela 3. Número de SNPs nos genes HLA com F_{ST} maior ou menor que a mediana dos F_{ST} s dos SNPs no grupo controle em cada intervalo de frequência do alelo mais raro (MAF), e para o conjunto total de dados, amalgamando todos os intervalos de MAF. Valores em negrito indicam a classe de F_{ST} mais frequente em cada intervalo de MAF. Em todos os intervalos de MAF, há mais SNPs nos HLA com valores de F_{ST} abaixo da mediana do grupo controle do que acima. Contudo, ao amalgamar todos os intervalos de MAF, observa-se o padrão contrário: mais SNPs nos HLA com F_{ST} acima da mediana do controle.

Intervalo de MAF	F_{ST} mediano no controle	F_{ST} HLA < mediana do controle	F_{ST} HLA > mediana do controle	Número total de SNPs nos HLA
<0.01	0.004	83	65	148
0.01 - 0.1	0.049	658	166	824
0.1 - 0.2	0.068	657	63	720
0.2 - 0.3	0.079	570	7	577
0.3 - 0.4	0.085	290	8	298
0.4 - 0.5	0.089	263	3	266
all	0.01	581	2252	2833

Questão original: diferenciação nos HLA

Retomando nossas questões iniciais,

- i) na escala global, o regime de seleção balanceadora atuante sobre os genes HLA consiste em pressões seletivas globais ou locais?*
- ii) como reconciliar a expectativa teórica de baixa diferenciação populacional em regiões sob seleção balanceadora à observação de alta diferenciação populacional em genes do MHC em humanos e outras espécies?*

Nossos resultados indicam que as respostas a ambas as perguntas dependem da entidade genética analisada. Em nossas análises baseadas em SNPs, detectamos diferenciação populacional menor nos genes HLA que no restante do genoma, o que indica que pressões seletivas globais atuam sobre os SNPs dos genes HLA. Por outro lado, é possível que análises no nível do haplótipo (definido como o conjunto de SNPs que forma a região apresentadora de antígenos dos genes HLA) revelem diferenciação aumentada em relação ao restante do genoma, corroborando a hipótese de que pressões seletivas locais atuam sobre os genes HLA. A interpretação desses resultados dependerá, portanto, de suposições sobre qual é o nível de organização biológica que define a aptidão dos indivíduos: SNPs ou haplótipos.

Essa diferença na escala do marcador usado pode responder também a segunda questão, explicando a expectativa de alta ou baixa diferenciação nos genes do MHC. Nossos resultados levantam uma questão motivadora para novos estudos: É possível que haplótipos de HLA sejam selecionados localmente, respondendo à pressão exercida por patógenos, e levando a alta diferenciação na distribuição desses haplótipos nas populações, e que no nível dos SNPs a expectativa teórica de diferenciação populacional reduzida seja atendida? Em caso afirmativo, a escala do marcador genético usado pode reconciliar os resultados diversos encontrados na literatura sobre diferenciação populacional em genes do MHC (humano e não humano), e em genes sob seleção balanceadora, em geral.

Resumo

Seleção balanceadora é definida como aquela que aumenta a variabilidade genética de populações em relação ao esperado sob neutralidade. Uma expectativa sobre seus efeitos é a redução da diferenciação populacional nos genes onde atua. Contudo, regimes que mantêm conjuntos distintos de alelos em diferentes populações poderiam resultar em aumento de diferenciação populacional. Com o objetivo de compreender melhor os efeitos da seleção balanceadora sobre a distribuição da variação genética entre populações, investigamos a diferenciação populacional em genes dos Antígenos Leucocitários Humanos (HLA, do inglês, *Human Leukocyte Antigen*), que são os genes mais polimórficos do genoma humano e o exemplo mais clássico de seleção balanceadora em humanos. As proteínas HLA são responsáveis pela apresentação de peptídeos aos linfócitos T, mediando uma etapa crítica da resposta imune. A vantagem da manutenção de variação nesses genes está possivelmente associada à capacidade de resposta a uma maior diversidade de patógenos.

Neste estudo, analisamos dados do projeto 1000 Genomas (1000G), que sequenciou indivíduos de diferentes populações usando sequenciamento de nova geração (NGS, do inglês, *Next Generation Sequencing*). Essa técnica de sequenciamento é conhecidamente problemática quando aplicada a regiões altamente polimórficas, como os genes HLA. Por isso, avaliamos a confiabilidade dos genótipos e frequências alélicas estimados a partir dos dados do 1000G nos genes HLA, utilizando como padrão ouro dados de sequenciamento Sanger de 930 das 1092 amostras do 1000G. Encontramos um viés de superestimativa da frequência do alelo referência em alguns polimorfismos de nucleotídeo único (SNPs, do inglês, *Single Nucleotide Polymorphisms*), indicando que viés de mapeamento é uma causa importante de erros nos dados do 1000G. Esses resultados são relevantes para a compreensão dos desafios do uso de dados de NGS em outras regiões de alta diversidade. Usando os resultados dessa análise, excluímos de nosso estudo de diferenciação populacional sítios com estimativas de frequências alélicas pouco confiáveis nos dados do 1000G. Em uma segunda etapa de controle metodológico, demonstramos o efeito do uso de dados ricos em variantes raras em estudos de diferenciação populacional. Controlando para esse efeito, e usando apenas sítios que demonstramos ser confiáveis em nossos dados, descobrimos que a diferenciação populacional de SNPs nos genes HLA é menor que a diferenciação de SNPs em outras regiões do genoma. Esse resultado aponta para um papel predominante de pressões seletivas globais na distribuição da variação genética de HLA entre populações. Contudo, apresentamos também evidências de que a diferenciação populacional de haplótipos nos genes HLA pode ser maior do que a observada no nível dos SNPs, sugerindo que pressões locais podem influenciar a distribuição de haplótipos entre populações. Nossos achados indicam que é possível reconciliar baixa diferenciação populacional em SNPs com maior diferenciação em haplótipos, possivelmente sujeitos a pressões seletivas locais.

Abstract

Balancing selection is defined as any kind of selective regime that increases genetic variability in populations relative to what is expected under neutrality. Theory predicts that balancing selection reduces population differentiation. However, balancing selection regimes in which different sets of alleles are maintained in different populations could increase population differentiation. To better understand the effects of balancing selection on the distribution of genetic variation among populations, we investigated population differentiation at the Human Leukocyte Antigen (HLA) genes, which are the most polymorphic genes in the human genome, and constitute the most striking example of balancing selection in humans. The HLA molecules are responsible for the presentation of peptides to T cells, thus mediating a critical step of the immune response. The advantage of maintaining variation in those genes through balancing selection is possibly related to the increased ability of the immune system to respond to a wider variety of pathogens.

In this study, we analysed the public dataset of the 1000 Genomes project (1000G), which sequenced 1092 individuals from different populations using Next Generation Sequencing (NGS) technologies. These sequencing techniques are known to be problematic when applied to highly polymorphic genomic regions, such as the HLA genes. Therefore, we evaluated the reliability of genotype calls and allele frequency estimates of the SNPs reported by 1000G at HLA genes, using Sanger sequencing data of 930 of the 1092 1000G samples as a gold standard. We found a bias towards overestimation of reference allele frequency for some single nucleotide polymorphisms (SNPs), indicating mapping bias is an important cause of error in frequency estimation in the 1000G data. These results provide insights into the challenges of using of NGS data at other genomic regions of high diversity. Using the results of this analysis, we selected a list of sites that have reliable allele frequency estimates in the 1000G data to be used in our population differentiation study. In another methodological control, we demonstrate the effect of using a dataset rich in rare variants in population differentiation studies. Controlling for this effect, and using only the sites which we demonstrated that were reliable, we found that population differentiation of single nucleotide polymorphisms (SNPs) at the HLA genes is lower than that of SNPs in other genomic regions. This suggests a predominant role of global selective pressures in shaping the distribution of variation at the HLA genes among populations. However, we also show evidence that population differentiation of HLA haplotypes may be higher than what we observe at the SNP level, suggesting that local selective pressures may influence the distribution of haplotypes among populations. Altogether, our results indicate that it is possible to reconcile low population differentiation at the SNP level - as predicted by theory - to higher differentiation at haplotypes, which are possibly under local selective pressures.

Referências Bibliográficas

- Aguiar, V. R. C., J. E. Cesar, E. T. Dermitzakis, & D. Meyer, 2015 Expression and eQTL mapping of HLA genes in large-scale RNAseq assays - Poster presented at the Biology of Genomes Meeting at the Cold Spring Harbor Laboratories.
- Akey, J. M., G. Zhang, K. Zhang, L. Jin, & M. D. Shriver, 2002 Interrogating a High-Density SNP Map for Signatures of Natural Selection : 1805–1814.
- Albrechtsen, A., I. Moltke, & R. Nielsen, 2010 Natural Selection and the Distribution of Identity-by-Descent in the Human Genome. *Genetics* **186**.
- Andersen, K. G., I. Shylakhter, S. Tabrizi, S. R. Grossman, C. T. Happi, & P. C. Sabeti, 2012 Genome-wide scans provide evidence for positive selection of genes implicated in Lassa fever. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **367**: 868–77.
- Andrés, A. M., 2011 Balancing Selection in the Human Genome. *Encyclopedia of Life Sciences* : 1–8.
- Andrés, A. M., M. J. Hubisz, A. Indap, D. G. Torgerson, J. D. Degenhardt, A. R. Boyko, *et al.*, 2009 Targets of balancing selection in the human genome. *Molecular biology and evolution* **26**: 2755–64.
- Balding, D. J., M. Bishop, & C. Cannings, editors, 2007 *Handbook of Statistical Genetics - Volume 1*. John Wiley & Sons, Ltd, 3rd edition.
- Barreiro, L. B., G. Laval, H. Quach, E. Patin, & L. Quintana-Murci, 2008 Natural selection has driven population differentiation in modern humans. *Nature genetics* **40**: 340–5.
- Beaumont, M. A., & R. A. Nichols, 1996 Evaluating Loci for Use in the Genetic Analysis of Population Structure. *Proceedings of the Royal Society B: Biological Sciences* **263**: 1619–1626.
- Begovich, A. B., P. V. Moonsamy, S. J. Mack, L. F. Barcellos, L. L. Steiner, S. Grams, *et al.*, 2001 Genetic variability and linkage disequilibrium within the HLA-DP region: analysis of 15 different populations. *Tissue antigens* **57**: 424–439.
- Bersaglieri, T., P. C. Sabeti, N. Patterson, T. Vanderploeg, S. F. Schaffner, J. A. Drake, *et al.*, 2004 Genetic Signatures of Strong Recent Positive Selection at the Lactase Gene. *American journal of human genetics* : 1111–1120.
- Bhatia, G., N. Patterson, B. Pasaniuc, N. Zaitlen, G. Genovese, S. Pollack, *et al.*, 2011 Genome-wide Comparison of African-Ancestry Populations from CARE and Other Cohorts Reveals Signals of Natural Selection. *The American Journal of Human Genetics* **89**: 368–381.
- Bhatia, G., N. Patterson, S. Sankararaman, & A. L. Price, 2013 Estimating and interpreting Fst: the impact of rare variants. *Genome research* **23**: 1514–1521.

- Bjorkman, P. J., M. A. Saper, B. Samraoui, W. S. Bennett, J. L. Strominger, & D. C. Wiley, 1987 Structure of the human class I histocompatibility antigen, HLA-A2. *Nature* **329**: 506–512.
- Boegel, S., M. Löwer, M. Schäfer, T. Bukur, J. de Graaf, V. Boisguérin, *et al.*, 2012 HLA typing from RNA-Seq sequence reads. *Genome medicine* **4**: 102.
- Borghans, J. a. M., J. B. Beltman, & R. J. De Boer, 2004 MHC polymorphism under host-pathogen coevolution. *Immunogenetics* **55**: 732–9.
- Bowcock, A. M., J. R. Kidd, J. L. Mountain, J. M. Hebert, L. Carotenuto, K. K. Kidd, *et al.*, 1991 Drift, admixture, and selection in human evolution: a study with DNA polymorphisms. *Proceedings of the National Academy of Sciences of the United States of America* **88**: 839–843.
- Brandt, D. Y. C., V. R. C. Aguiar, B. D. Bitarello, K. Nunes, J. Goudet, & D. Meyer, 2015 Mapping bias overestimates reference allele frequencies at the HLA genes in the 1000 Genomes Project phase I data. *G3:Genes|Genomes|Genetics* **5**: 931–941.
- Brown, J. H., T. S. Jardetzky, J. C. Gorga, L. J. Stern, R. G. Urban, J. L. Strominger, *et al.*, 1993 Three-dimensional structure of the human class II histocompatibility antigen HLA-DR1. *Nature* **364**: 33–39.
- Bubb, K. L., D. Bovee, D. Buckley, E. Haugen, M. Kibukawa, M. Paddock, *et al.*, 2006 Scan of human genome reveals no new Loci under ancient balancing selection. *Genetics* **173**: 2165–77.
- Cammen, K., J. I. Hoffman, L. a. Knapp, J. Harwood, & W. Amos, 2011 Geographic variation of the major histocompatibility complex in Eastern Atlantic grey seals (*Halichoerus grypus*). *Molecular Ecology* **20**: 740–752.
- Cavalli-Sforza, L. L., 1966 Population Structure and Human Evolution. *Proceedings of the Royal Society B: Biological Sciences* **164**: 362–379.
- Chapman, S. J., & A. V. S. Hill, 2012 Human genetic susceptibility to infectious disease. *Nature reviews. Genetics* **13**: 175–88.
- Charlesworth, B., 2015 Causes of natural variation in fitness: Evidence from studies of *Drosophila* populations. *Proceedings of the National Academy of Sciences* **112**: 201502053.
- Charlesworth, B., M. Nordborg, & D. Charlesworth, 1997 The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genetical research* **70**: 155–74.
- Cockerham, C. C., 1969 Variance of gene frequencies. *Evolution* **23**: 72–84.
- Danecek, P., A. Auton, G. R. Abecasis, C. a. Albers, E. Banks, M. a. DePristo, *et al.*, 2011 The variant call format and VCFtools. *Bioinformatics (Oxford, England)* **27**: 2156–8.
- De Santis, D., D. Dinuer, J. Duke, H. a. Erlich, C. L. Holcomb, C. Lind, *et al.*, 2013 16(th) IHIW : review of HLA typing by NGS. *International journal of immunogenetics* **40**: 72–6.
- Degner, J. F., J. C. Marioni, A. a. Pai, J. K. Pickrell, E. Nkadori, Y. Gilad, *et al.*, 2009 Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* **25**: 3207–12.
- Dilthey, A., C. Cox, Z. Iqbal, M. R. Nelson, & G. McVean, 2015 Improved genome inference in the MHC using a population reference graph. *Nature genetics* : 006973.
- Doherty, P. C., & R. M. Zinkernagel, 1975 Enhanced immunological surveillance in mice heterozygous at the H-2 gene complex. *Nature* **256**: 50–52.

- Elhaik, E., 2012 Empirical distributions of F_{ST} from large-scale human polymorphism data. *PloS one* **7**: e49837.
- Erlich, R. L., X. Jia, S. Anderson, E. Banks, X. Gao, M. Carrington, *et al.*, 2011 Next-generation sequencing for HLA typing of class I loci. *BMC genomics* **12**: 42.
- Excoffier, L., & H. E. L. Lischer, 2010 Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular ecology resources* **10**: 564–7.
- Futuyma, D. J., 2006 *Evolutionary Biology*. Sinauer Associates.
- Gillespie, J. H., 2004 *Population Genetics: A Concise Guide*. Johns Hopkins University Press, 2nd edition.
- Goudet, J., 2005 Hierfstat, a package for R to compute and test hierarchical F-statistics. *Molecular Ecology Notes* **5**: 184–186.
- Gourraud, P.-A., P. Khankhanian, N. Cereb, S. Y. Yang, M. Feolo, M. Maiers, *et al.*, 2014 HLA Diversity in the 1000 Genomes Dataset. *PLoS ONE* **9**: e97282.
- Hamilton, M. B., 2009 *Population Genetics*. John Wiley & Sons, Ltd.
- Hedrick, P. W., 1999 Highly variable loci and their interpretation in evolution and conservation. *Evolution* **53**: 313–318.
- Hedrick, P. W., 2002 Pathogen resistance and genetic variation at MHC loci. *Evolution; international journal of organic evolution* **56**: 1902–8.
- Hedrick, P. W., & G. Thomson, 1983 Evidence for balancing selection at HLA. *Genetics* **104**: 449–56.
- Helmberg, W., M. Feolo, R. Dunivin, & D. Hoffman, 2014 dbMHC.
- Hernandez, R. D., J. L. Kelley, E. Elyashiv, S. C. Melton, A. Auton, G. McVean, *et al.*, 2011 Classic selective sweeps were rare in recent human evolution. *Science (New York, N.Y.)* **331**: 920–4.
- Hill-Burns, E. M., S. a. Factor, C. P. Zabetian, G. Thomson, & H. Payami, 2011 Evidence for more than one Parkinson's disease-associated variant within the HLA region. *PloS one* **6**: e27109.
- Hofer, T., M. Foll, & L. Excoffier, 2012 Evolutionary forces shaping genomic islands of population differentiation in humans. *BMC genomics* **13**: 107.
- Hollenbach, J. A., G. Thomson, K. Cao, M. Fernandez-Vina, H. A. Erlich, T. L. Bugawan, *et al.*, 2001 HLA Diversity, Differentiation, and Haplotype Evolution in Mesoamerican Natives. *Human Immunology* **62**: 378–390.
- Holsinger, K. E., & B. S. Weir, 2009 Genetics in geographically structured populations: defining, estimating and interpreting F_{ST} . *Nature reviews. Genetics* **10**: 639–50.
- Hughes, A. L., 2008 Major Histocompatibility Complex (MHC) Genes: Evolution. *Encyclopedia of Life Sciences* : 1–4.
- Hughes, A. L., & M. Nei, 1988 Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* **335**: 167–170.
- Jackson, B. C., J. L. Campos, & K. Zeng, 2014 The effects of purifying selection on patterns of genetic differentiation between *Drosophila melanogaster* populations. *Heredity* **114**: 163–174.

- Jakobsson, M., M. D. Edge, & N. A. Rosenberg, 2013 The relationship between F_{ST} and the frequency of the most frequent allele. *Genetics* **193**: 515–28.
- Jost, L., 2008 G_{ST} and its relatives do not measure differentiation. *Molecular Ecology* **17**: 4015–4026.
- Keinan, a., & a. G. Clark, 2012 Recent Explosive Human Population Growth Has Resulted in an Excess of Rare Genetic Variants. *Science* **336**: 740–743.
- Kimura, M., 1968 Evolutionary rate at the molecular level. *Nature* **217**: 624–626.
- Kimura, M., 1983 *The Neutral Theory of Molecular Evolution*. Cambridge University Press.
- Klein, J., & A. Sato, 2000 The HLA System. *The New England Journal of Medicine* **343**: 702–709.
- Lappalainen, T., M. Sammeth, M. R. Friedländer, P. a. C. 't Hoen, J. Monlong, M. a. Rivas, *et al.*, 2013 Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**: 506–11.
- Lawlor, D. A., F. E. Ward, P. D. Ennis, A. P. Jackson, & P. Parham, 1988 HLA-A and B polymorphisms predate the divergence of humans and chimpanzees. *Nature* **335**: 268–271.
- Leffler, E. M., Z. Gao, S. Pfeifer, L. Séguérel, A. Auton, O. Venn, *et al.*, 2013 Multiple instances of ancient balancing selection shared between humans and chimpanzees. *Science* (New York, N.Y.) **339**: 1578–82.
- Lewontin, R. C., & J. Krakauer, 1973 Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* **74**: 175–95.
- Loiseau, C., M. Richard, S. Garnier, O. Chastel, R. Julliard, R. Zoorob, *et al.*, 2009 Diversifying selection on MHC class I in the house sparrow (*Passer domesticus*). *Molecular Ecology* **18**: 1331–1340.
- Losos, J. B., D. A. Baum, D. J. Futuyma, H. E. Hoekstra, R. E. Lenski, A. J. Moore, *et al.*, editors, 2013 *The Princeton Guide to Evolution*. Princeton reference. Princeton University Press.
- Lotterhos, K. E., & M. C. Whitlock, 2014 Evaluation of demographic history and neutral parameterization on the performance of F_{ST} outlier tests. *Molecular ecology* **23**: 2178–92.
- Major, E., K. Rigo, T. Hague, A. Bérces, S. Juhos, A. Be, *et al.*, 2013 HLA Typing from 1000 Genomes Whole Genome and Whole Exome Illumina Data. *PloS one* **8**: e78410.
- Marsh, S. G. E., E. D. Albert, W. F. Bodmer, R. E. Bontrop, B. Dupont, H. a. Erlich, *et al.*, 2010 Nomenclature for factors of the HLA system, 2010. *Tissue antigens* **75**: 291–455.
- Maruki, T., S. Kumar, & Y. Kim, 2012 Purifying selection modulates the estimates of population differentiation and confounds genome-wide comparisons across single-nucleotide polymorphisms. *Molecular biology and evolution* **29**: 3617–23.
- Mather, K., 1953 The genetical structure of populations. *Symposium of society for Experimental Biology* **7**: 66–95.
- McCairns, R. J. S., S. Bourget, & L. Bernatchez, 2011 Putative causes and consequences of MHC variation within and between locally adapted stickleback demes. *Molecular Ecology* **20**: 486–502.
- Mendes, F. H. K., 2013 *Seleção natural em genes HLA e seu efeito sobre regiões adjacentes do genoma*. Msc, Universidade de São Paulo.

- Meyer, D., R. M. Single, S. J. Mack, H. a. Erlich, & G. Thomson, 2006 Signatures of demographic history and natural selection in the human major histocompatibility complex Loci. *Genetics* **173**: 2121–42.
- Meyer, D., & G. Thomson, 2001 How selection shapes variation of the human major histocompatibility complex: a review. *Annals of human genetics* **65**: 1–26.
- Miller, H. C., F. Allendorf, & C. H. Daugherty, 2010 Genetic diversity and differentiation at MHC genes in island populations of tuatara (*Sphenodon* spp.). *Molecular Ecology* **19**: 3894–3908.
- Muirhead, C. a., 2001 Consequences of population structure on genes under balancing selection. *Evolution; international journal of organic evolution* **55**: 1532–41.
- Nei, M., 1973 Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences of the United States of America* **70**: 3321–3.
- Nei, M., 1977 F-statistics and analysis of gene diversity in subdivided populations. *Annals of human genetics* **41**: 225–233.
- Nei, M., & R. K. Chesser, 1983 Estimation of fixation indices and gene diversities. *Annals of human genetics* **47**: 253–259.
- Nei, M., & S. Kumar, 2000 *Molecular Evolution and Phylogenetics*. Oxford University Press.
- Nielsen, R., 2005 Molecular signatures of natural selection. *Annual review of genetics* **39**: 197–218.
- Nielsen, R., J. S. Paul, A. Albrechtsen, & Y. S. Song, 2011 Genotype and SNP calling from next-generation sequencing data. *Nature reviews. Genetics* **12**: 443–51.
- Nunes, K., 2011 *Populações ameríndias da América do Sul: estudo multi-locus e inferência histórico-demográfica e seletiva*. Doutorado, Universidade de São Paulo.
- Ohta, T., & J. Gillespie, 1996 Development of Neutral and Nearly Neutral Theories. *Theoretical population biology* **49**: 128–42.
- Oliver, M. K., X. Lambin, T. Cornulier, & S. B. Piertney, 2009 Spatio-temporal variation in the strength and mode of selection acting on major histocompatibility complex diversity in water vole (*Arvicola terrestris*) metapopulations. *Molecular Ecology* **18**: 80–92.
- Parham, P., & T. Ohta, 1996 Population biology of antigen presentation by MHC class I molecules. *Science* **272**: 67–74.
- Pasvol, G., D. J. Weatherall, & R. J. M. Wilson, 1978 Cellular mechanism for the protective effect of haemoglobin S against *P. falciparum* malaria.
- Prugnolle, F., A. Manica, M. Charpentier, V. Guernier, M. Population, G. Group, *et al.*, 2005 Pathogen-Driven Selection and Worldwide HLA Class I Diversity **15**: 1022–1027.
- Quinlan, A. R., & I. M. Hall, 2010 BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)* **26**: 841–2.
- Qutob, N., F. Balloux, T. Raj, H. Liu, S. Marion de Procé, J. Trowsdale, *et al.*, 2012 Signatures of historical demography and pathogen richness on MHC class I genes. *Immunogenetics* **64**: 165–75.
- Reynolds, J., B. S. Weir, & C. C. Cockerham, 1983 Estimation of the coancestry coefficient: basis for a short-term genetic distance. *Genetics* **105**: 767–79.
- Robinson, J., J. A. Halliwell, H. McWilliam, R. Lopez, P. Parham, & S. G. E. Marsh, 2013 The IMGT/HLA database. *Nucleic acids research* **41**: D1222–7.

- Roesti, M., W. Salzburger, & D. Berner, 2012 Uninformative polymorphisms bias genome scans for signatures of selection. *BMC evolutionary biology* **12**: 94.
- Sabeti, P. C., S. F. Schaffner, B. Fry, J. Lohmueller, P. Varilly, O. Shamovsky, *et al.*, 2006 Positive natural selection in the human lineage. *Science* (New York, N.Y.) **312**: 1614–20.
- Sanchez-Mazas, A., 2007 An apportionment of human HLA diversity. *Tissue antigens* **69 Suppl 1**: 198–202.
- Schierup, M. H., X. Vekemans, & D. Charlesworth, 2000 The effect of subdivision on variation at multi-allelic loci under balancing selection. *Genetical research* **76**: 51–62.
- Simpson, E. H., 1951 The Interpretation of Interaction in Contingency Tables. *Journal of the Royal Statistical Society. Series B (Methodological)* **13**: 238–241.
- Slatkin, M., 1991 Inbreeding coefficients and coalescence times. *Genetical Research* **58**: 167.
- Smith, J. M., & J. Haigh, 1974 The hitch-hiking effect of a favourable gene. *Genetics Research* **23**: 23–35.
- Solberg, O. D., S. J. Mack, A. K. Lancaster, R. M. Single, Y. Tsai, A. Sanchez-Mazas, *et al.*, 2008 Balancing selection and heterogeneity across the classical human leukocyte antigen loci: a meta-analytic review of 497 population studies. *Human immunology* **69**: 443–64.
- Templeton, A. R., 2006 *Population genetics and microevolutionary theory*. John Wiley & Sons, Inc.
- The 1000 Genomes Project Consortium, 2012a An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**: 56–65.
- The 1000 Genomes Project Consortium, 2012b An integrated map of genetic variation from 1,092 human genomes: Supplementary Material. *Nature* .
- Treangen, T. J., & S. L. Salzberg, 2012 Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature reviews. Genetics* **13**: 36–46.
- Wang, K., M. Li, & H. Hakonarson, 2010 ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research* **38**: 1–7.
- Ward, L. D., & M. Kellis, 2012 Evidence of Abundant Purifying Selection in Humans for Recently Acquired Regulatory Functions. *Science* **337**.
- Weir, B. S., 2012 Estimating F -Statistics : A Historical View. *Philosophy of Science* **79**: 637–643.
- Weir, B. S., & C. C. Cockerham, 1984 Estimating F-statistics for the analysis of population structure. *Evolution* **38**: 1358–1370.
- Wright, S., 1951 The genetical structure of populations. *Annals of Eugenics* .
- Xu, S., S. Li, Y. Yang, J. Tan, H. Lou, W. Jin, *et al.*, 2011 A genome-wide search for signals of high-altitude adaptation in Tibetans. *Molecular biology and evolution* **28**: 1003–11.
- Yule, G. U., 1903 Notes on the theory of association of attributes in statistics. *Biometrika* **2**: 121–134.

Apêndices

Supporting Information for Chapter 1

File S1

ARS_exons.bed

BED file giving the coordinates for ARS exons used in this study. Coordinates were acquired from UCSC Table Browser using the RefSeq Genes track on 22 July 2014. When more than one transcript was available in the database, the pair of coordinates including more positions was chosen. RefSeq IDs from which ARS exon coordinates were acquired are NM_002116 (HLA-A), NM_005514 (HLA-B), NM_002117 (HLA-C), NM_001243961 (HLA-DQB1) and NM_001243965 (HLA-DRB1). Coordinates in the BED file are given using one-based start and end coordinates.

File S1 is available for download at <http://www.g3journal.org/content/suppl/2015/03/17/g3.114.015784.DC1/FileS1.bed>

Figure S1.1. Workflow for preparation of next generation sequencing dataset from the 1000 Genomes Project (1000G) and Sanger sequencing dataset generated by Gourraud et al. (2014) (PAG2014) for comparisons of genotypes and allele frequencies (see main text).

Table S1.1. List of polymorphic sites at the HLA genes that were discovered in the 1000 Genomes project exclusively on the high-coverage exome experiments. Positions in coordinates relative to the human reference genome hg19 build and relative to the ARS exons are given.

Gene	hg19 position	ARS position
A	29910673	140
A	29910682	149
A	29910717	184
A	29910719	186
A	29910750	217
A	29910752	219
A	29910761	228
A	29910768	235
B	31324570	140
B	31324589	146
B	31324595	165
C	31238984	412
DQB1	32632598	128
DQB1	32632599	131
DQB1	32632601	244
DQB1	32632714	246
DQB1	32632717	247
DRB1	32552067	65
DRB1	32552072	69
DRB1	32552075	75
DRB1	32552079	77
DRB1	32552081	81
DRB1	32552087	84
DRB1	32552091	89

Figure S1.2. Relationship between the proportion of genotype mismatches and nucleotide diversity (P_i) per exon.

Figure S1.3. Continues on next page.

Figure S1.3. Reference allele frequency per population and per site in the HLA-A gene in the 1000 Genomes (1000G; y-axis) and Sanger sequencing (PAG2014; x-axis) datasets. Dashed lines indicate a ± 0.1 deviation from the expected frequency (as estimated from PAG2014 dataset). MAE (mean absolute error) defined in Methods. Numbers indicate site position in ARS exons sequence.

Figure S1.4. Continues on next page.

Figure S1.4. Reference allele frequency per population and per site in the HLA-B gene in the 1000 Genomes (1000G; y-axis) and Sanger sequencing (PAG2014; x-axis) datasets. Dashed lines indicate a ± 0.1 deviation from the expected frequency (as estimated from PAG2014 dataset). MAE (mean absolute error) defined in Methods. Numbers indicate site position in ARS exons sequence.

Figure S1.5. Continues on next page.

Figure S1.5. Reference allele frequency per population and per site in the HLA-C gene in the 1000 Genomes (1000G; y-axis) and Sanger sequencing (PAG2014; x-axis) datasets. Dashed lines indicate a ± 0.1 deviation from the expected frequency (as estimated from PAG2014 dataset). MAE (mean absolute error) defined in Methods. Numbers indicate site position in ARS exons sequence.

Figure S1.6. Continues on next page.

Figure S1.6. Reference allele frequency per population and per site in the HLA-DQB1 gene in the 1000 Genomes (1000G; y-axis) and Sanger sequencing (PAG2014; x-axis) datasets. Dashed lines indicate a ± 0.1 deviation from the expected frequency (as estimated from PAG2014 dataset). MAE (mean absolute error) defined in Methods. Numbers indicate site position in ARS exons sequence.

Figure S1.7. Continues on next page.

Figure S1.7. Reference allele frequency per population and per site in the HLA-DRB1 gene in the 1000 Genomes (1000G; y-axis) and Sanger sequencing (PAG2014; x-axis) datasets. Dashed lines indicate a ± 0.1 deviation from the expected frequency (as estimated from PAG2014 dataset). MAE (mean absolute error) defined in Methods. Numbers indicate site position in ARS exons sequence.

Figure S1.8. Continues on next page.

Figure S1.8. Relationship between proportion of mismatched genotypes per site (considering all individual genotypes) and mean difference in reference allele frequency estimated from the 1000 Genomes NGS data and Gourraud et al. (2014) Sanger sequencing data. Numbers indicate site position in ARS exons sequence.

Figure S1.9. Genotypes from the Axiom Exome Genotyping Array - Affymetrix for 1000 Genomes samples were acquired from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/axiom_genotypes/ALL.wex.axiom.20120206.snps_and_indels.genotypes.vcf.gz. For the first and second sets of points (ARS exons), Axiom Exome and 1000G datasets were filtered to keep only sites at exons 2 and 3 of *HLA-A*, *-B* and *-C* and exon 2 of *-DQB1* and *-DRB1* genes and only individuals present in the PAG2014 dataset. For the third set of points, the Axiom Exome dataset was filtered to keep only sites at the extended MHC region (positions 29570005 to 33377699 in the hg19 build of the human reference genome), and only individuals present in the 1000 Genomes phase I dataset (1000G). Both individual and site filters were applied using VCFtools v0.1.12b.

Allele frequencies were calculated from the Axiom Exome array genotypes and compared to frequencies estimated from PAG2014 genotypes or the 1000G in the same way that frequencies from 1000G were previously compared to the PAG2014 frequencies (described in Methods). A single SNP had a very discrepant reference allele frequency between PAG2014 and the Axiom array data: rs145937432, which is not present in the 1000G dataset. This SNP has "C" as its reference allele, and its frequency among the 930 individuals we analysed is 0.001 in the Axiom Exome dataset, and 0.969 in PAG2014. This site was excluded from this analysis.

The difference in frequency between Axiom and PAG2014 was smaller than the difference between 1000G and PAG2014 (p-value = 0.004 using a permutation approach). However, sites that were present in both datasets (shown in red) show that their frequency differences are small for both Axiom Exome and 1000G, relative to PAG2014. The overall divergence between 1000G and Axiom Exome is also small for SNPs surrounding the HLA genes. This indicates that 1) SNP allele frequencies estimated from this array are reliable; 2) allele frequencies of SNPs present in this array are similarly reliable when estimated from NGS.

Figure S1.10. Absence of relationship between absolute deviation in allele frequency estimation in the 1000 Genomes dataset relative to Sanger sequencing (PAG2014) and the distance of the SNP relative do the center of the exon.

Table S1.2. Proportion of each genotype in the PAG2014 dataset (Sanger sequencing) as called by the 1000 Genomes. The diagonal shows the proportion of correctly called genotypes. ALT = alternative allele; REF = reference allele.

		1000 Genomes		
		ALT/ALT	ALT/REF	REF/REF
PAG2014	ALT/ALT	0.699	0.227	0.074
	ALT/REF	0.029	0.681	0.290
	REF/REF	0.001	0.069	0.930

Table S1.3. Full names of 1000 Genomes Project populations.

Code	Population name
ASW	African Ancestry from Southwest, USA
CEU	Northern and Western European from Utah, USA
CHB+JPT	Han Chinese from Beijing, China + Japanese from Tokyo, Japan
CHS	Han from south, China
CLM	Colombian from Medellin, Colombians
FIN	Finnish, Finland
GBR	British from England and Scotland, UK
LWK	Luhya from Webuye, Kenya
MXL	Mexican Ancestry from Los Angeles-California, USA
PUR	Puerto Rican, Puerto Rico
TSI	Italian from Tuscany, Italy
YRI	Yoruba from Ibadan, Nigeria

Table S1.4. Genomic coordinates (hg19) of sites with poorly estimated frequency in 1000G in each HLA locus. Those sites have difference larger than 0.1 in the frequency estimated by 1000G relative to PAG2014 in 2 or more populations.

Gene	<i>HLA-A</i>	<i>HLA-B</i>	<i>HLA-C</i>	<i>HLA-DQB1</i>	<i>HLA-DRB1</i>
Number of sites	14/66	32/64	9/44	24/42	22/35
hg19 coordinates	29910558	31324711	31239050	32632795	32552147
	29910700	31324705	31238983	32632782	32552112
	29910717	31324702	31238957	32632770	32552091
	29910719	31324664	31238942	32632749	32552029
	29910759	31324603	31238930	32632703	32551999
	29911056	31324586	31238851	32632700	32551998
	29911115	31324549	31239060	32632688	32551995
	29911119	31324547	31239006	32632687	32551970
	29911228	31324536	31238992	32632660	32551935
	29911239	31324528		32632659	32551912
	29911240	31324526		32632638	32551899
	29910688	31324525		32632637	32552092
	29911296	31324516		32632635	32552087
	29911306	31324506		32632628	32552075
		31324491		32632627	32552048
		31324489		32632608	32552039
		31324210		32632605	32552017
		31324208		32632601	32552016
		31324194		32632599	32551939
		31324184		32632589	32551938
		31324176		32632581	32551928
		31324100		32632578	32551905
		31324086		32632598	
		31324077		32632587	
		31324036			
		31324024			
		31324004			
		31323960			
		31323958			
		31323953			
		31324154			
		31324104			

Supporting Information for Chapter 2

Table S2.1. 1000 Genomes Project populations and sample sizes. Total sample: 736 individuals.

Population	Code	Sample size	Continent	Sample Size
Utah Residents (CEPH) with Northern and Western European Ancestry	CEU	85	EUR	365
British in England and Scotland	GBR	89		
Finnish in Finland	FIN	93		
Toscani in Italia	TSI	98		
Japanese in Tokyo, Japan	JPT	89	ASN	186
Han Chinese in Beijing, China	CHB	97		
Luhya in Webuye, Kenya	LWK	97	AFR	185
Yoruba in Ibadan, Nigeria	YRI	88		

Table S2.2. List of peri HLA genes and their genomic coordinates in hg19

Start	End	Gene
29691117	29695073	HLA-F
29759683	29765584	LOC554223
29794756	29798899	HLA-G
30029017	30032686	ZNRD1
30951485	30957675	MUC21
30973729	31003179	MUC22
31079000	31080332	C6orf15
31082865	31088252	CDSN
31082608	31107869	PSORS1C1
31105311	31107127	PSORS1C2
31110216	31126015	CCHCR1
31126303	31131992	TCF19
31132114	31138451	POU5F1
31371371	31383090	MICA
31465855	31478901	MICB
32260475	32339656	C6orf10
32362513	32374900	BTNL2
32485154	32498006	HLA-DRB5
32709163	32714664	HLA-DQA2
32723875	32731330	HLA-DQB2
32780540	32784825	HLA-DOB
32793187	32806547	TAP2
32808494	32812712	PSMB8
32812986	32821748	TAP1
32821938	32827628	PSMB9

Figure S2.1. Resampling algorithm to account for LD among *HLA* SNPs

(a)

(b)

Figure S2.2. F_{ST} as a function of MAF for all intronic (a) and exonic (b) SNPs. HLA genes are coloured by gene, with *HLA-DPA1* shown in pink and *HLA-DPB1* in purple. Gray background contains all variants in the genomic control. Green: *HLA-A*; red: *HLA-B*; dark blue: *HLA-C*; orange: *-DQA*; dark green: *-DQB*; blue: *-DRA*; light green: *-DRB*.