Rodrigo Fernandes Ramalho

Análise do polimorfismo da região cis-reguladora do
gene *CCR5* em populações ameríndias

São Paulo
2007

Rodrigo Fernandes Ramalho

Análise do polimorfismo da região cis-reguladora do gene *CCR5* em populações ameríndias

Dissertação apresentada ao Instituto de Biociências da Universidade de São Paulo, para a obtenção de Título de Mestre em Ciências, na Área de Genética Evolutiva.

Orientador(a): Dr. Diogo Meyer

São Paulo
2007

Ramalho, Rodrigo Fernandes
Análise do polimorfismo na região cis-
reguladora do gene *CCR5* em populações
ameríndias

Número de páginas 103

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

1. seleção natural 2. *CCR5* 3.
ameríndios I. Universidade de São Paulo.
Instituto de Biociências. Departamento de
Genética

Comissão Julgadora:

Prof(a). Dr(a).

Prof(a). Dr(a).

Prof(a). Dr(a).

Prof(a). Dr(a).

Prof(a). Dr.(a).

Orientador(a)

Aos meus pais

Agradecimentos

Ao professor Diogo Meyer pela orientação e confiança.

Aos professores Eduardo Gorab, Regina Célia Mingroni Netto e João Stenghel Morgante por permitirem o início e boa parte do desenvolvimento deste projeto.

Aos meus amigos companheiros do Laboratório de Biologia Evolutiva e Conservação (LABEC) do IB-USP, Nádia, Gisele, Felipe, Juliana, Aninha, Carol, Fernanda, Rodrigo, Márcia, Kelly e Gustavo.

Aos professores Eduardo José Melo dos Santos da Universidade Federal do Pará e Nelson Jurandi Fagundes Rosa da Pontifícia Universidade Católica do Rio Grande do Sul por suas orientações, sugestões e colaborações no decorrer deste projeto.

Ao Dr. Michael Bamshad da Universidade de Utah (E.U.A.) pela colaboração.

À todos os alunos da professora Regina Célia Mingroni Netto, principalmente Dra. Cláudia Blanes Angeli, Dra. Karina Lezirovitz e Lilian Kimura pela grande ajuda nas etapas de biologia molecular.

Às técnicas responsáveis pelo sequenciador automático do Laboratório Biologia Molecular de Plantas (IB-USP) e do Centro de Estudos do Genoma Humano (IB-USP), Silvia Blanco e Camila respectivamente.

À todos que de alguma forma colaboraram direta ou indiretamente para a realização deste trabalho.

Índice

Introdução.....	7
Objetivo e Justificativa	20
Materiais e Métodos.....	21
Resultados	43
Discussão.....	69
Conclusão e Perspectivas Futuras.....	86
Resumo.....	88
Abstract	90
Bibliografia.....	92

Introdução

Fatores seletivos e não seletivos influenciam a variabilidade genética

Nas últimas décadas, avanços teóricos, especialmente vindos da teoria neutra de evolução molecular (Kimura, 1991) e da teoria da coalescência (Kingman, 2000), ao lado dos imensos avanços na geração de dados sobre a variação genética em populações, permitiram o desenvolvimento de uma linha de pesquisa voltada para a detecção de assinaturas de seleção natural.

Uma das formas de se estudar a possível ação de seleção natural com dados genéticos é através de testes de neutralidade, que têm como hipótese nula o modelo neutro de evolução molecular. Esse modelo é analiticamente testável e faz previsões sobre as frequências dos alelos, a proporção esperada de polimorfismos e divergência para diferentes genes ou classes de mutação.

Essa linha de pesquisa inclui tanto abordagens que testam a hipótese de ação de seleção natural em genes previamente definidos, geralmente com importância funcional conhecida (Bersaglieri, Sabeti *et al.*, 2004), assim como outras que buscam regiões do genoma com evidências de seleção natural, independentemente de conhecimento funcional prévio (*scans* genômicos) (Nielsen, 2005). Esses estudos foram estimulados pelos resultados dos projetos de seqüenciamento do genoma humano e por iniciativas de caracterização dos polimorfismos da espécie humana como, por exemplo, o projeto HapMap (www.hapmap.org).

Os testes aplicados para o estudo de seleção natural podem ser similares entre as abordagens de *scans* genômicos e de genes candidatos. Duas classes de testes muito utilizados nesses estudos são os baseados na frequência de polimorfismos de seqüências de DNA (por exemplo, D de Tajima) e os baseados na diferenciação genética intra e

interpopulacional (por exemplo, testes baseados em valores de F_{st}) (Harris e Meyer, 2006).

Testes de neutralidade baseados na frequência de polimorfismos de seqüências de DNA

A seleção natural e a demografia causam impactos diversos e às vezes semelhantes nas frequências alélicas de uma população. A seleção positiva ou direcional resulta no aumento de frequência de uma determinada variante genética (alelo) resultando na redução de diversidade genética no loco selecionado e nos locos ligados a ele. Esse processo é chamado *selective sweep* (Smith e Haigh, 1974; Braverman, Hudson *et al.*, 1995). À medida que novas mutações vão acumulando em um loco que foi alvo do *selective sweep*, espera-se que este apresente uma grande proporção de alelos em frequências baixas. A mesma assinatura é esperada em caso de expansão populacional, pois o aumento no tamanho efetivo da população permite o maior acúmulo de mutações (Rogers e Harpending, 1992). Seleção negativa ou purificadora é aquela que atua removendo alelos novos e deletérios e também se espera que ela cause um excesso de alelos raros na amostra populacional. Isso porque os alelos deletérios ou fracamente deletérios que surgem por mutação são constantemente removidos por seleção e não atingem frequências relativamente altas.

A seleção balanceadora mantém dois ou mais alelos em frequências similares em um determinado loco genético na população amostrada e, conseqüentemente, as medidas de polimorfismo desses locos são relativamente altas. Entre os mecanismos seletivos propostos para explicar o efeito balanceador da seleção natural estão: vantagem de heterozigotos, seleção dependente de frequência e mudanças no regime de seleção em diferentes períodos da história evolutiva (Charlesworth, 2006; Hedrick, 2006). Entretanto, uma assinatura genética semelhante à da seleção balanceadora é

esperada para cenários demográficos complexos que envolvem gargalos populacionais que removem preferencialmente alelos raros, miscigenação e subdivisão populacional (Przeworski, Hudson *et al.*, 2000; Richman, 2000).

A assinatura de seleção natural em dados genéticos formados por seqüências de DNA é geralmente analisada através de testes de neutralidade que usam como hipótese nula a teoria neutra de evolução molecular. Tipicamente, esses testes adotam como premissa a aderência da população à um regime neutro e às condições de equilíbrio populacional. Entre as premissas de equilíbrio estão a de que a população sob estudo é panmítica, apresenta ausência de estruturação populacional e de variações no tamanho da população ao longo do tempo (Bamshad e Wooding, 2003).

Entretanto, populações naturais dificilmente encontram-se em equilíbrio e portanto, eventos demográficos tais como variações no tamanho e subdivisões da população devem ser levados em conta nas análises, pois geram efeitos no padrão de variabilidade genética que podem deixar assinaturas similares àsquelas deixadas por determinados tipos de eventos seletivos.

Os testes de neutralidade baseados na freqüência de polimorfismos de seqüências de DNA utilizam basicamente estatísticas-sumário de polimorfismo como parâmetros (Bamshad e Wooding, 2003). Essas medidas, como por exemplo, o número de sítios polimórficos e a taxa de heterozigose, são descrições resumidas da quantidade de variação genética presente numa amostra populacional. Tanto os aspectos demográficos como os seletivos podem alterar a distribuição das freqüências dos alelos de uma amostra populacional, e conseqüentemente influenciar as estatísticas-sumário e os resultados dos testes de neutralidade baseados na freqüência de polimorfismo de seqüência de DNA.

Testes de neutralidade baseados na diferenciação genética intra e interpopulacional

O modelo neutro propõe que grande parte da diferenciação genética entre populações é devida ao mecanismo de deriva genética (Kimura, 1991). Entretanto, a seleção natural também pode contribuir para a diferenciação devido à adaptação diferencial das populações aos seus respectivos ambientes. A medida mais comum de diferenciação genética entre populações é o índice de fixação (F_{st}) proposto por (Wright, 1950). Portanto, um loco que apresente um valor de F_{st} maior que o esperado por um modelo de evolução neutro é um candidato a ser alvo de seleção natural (Vallender e Lahn, 2004).

Abordagens utilizadas na distinção entre seleção natural e neutralidade

Atualmente, destacam-se duas estratégias para se distinguir os efeitos da demografia dos efeitos seletivos. A primeira baseia-se na comparação de estatísticas-sumário de polimorfismo, resultados de testes de neutralidade ou valores de F_{st} de loco(s) candidato(s), com a distribuição dessas mesmas estatísticas obtidas através da análise de diversos marcadores considerados neutros, dispersos pelo genoma e analisados nos mesmos indivíduos ou populações (Bamshad, Mummidi *et al.*, 2002). A segunda estratégia é similar, porém a comparação é feita em relação à uma distribuição de valores simulados sob a hipótese nula. Isto é, a comparação do valor da estatística-sumário, do resultado do teste de neutralidade ou do F_{st} observado para o loco candidato à seleção, se dá em relação à uma distribuição da mesma estatística calculada para milhares de amostras simuladas sob um modelo neutro que pode assumir a existência de eventos demográficos (Akey, Eberle *et al.*, 2004).

Como ferramenta para análise de dados, o modelo coalescente pode ser implementado em programas de computador permitindo testes de hipótese com rigor estatístico (Hudson, 1990). Uma grande variedade de cenários evolutivos podem ser modelados usando o coalescente. Eventos históricos como aumento e diminuição do tamanho populacional em um determinado momento, migração e crescimento constante ao longo do tempo e a combinação de todos estes eventos em diferentes momentos da genealogia podem ser simulados milhares de vezes com o intuito de se obter a distribuição da estatística de interesse como por exemplo, o número de sítios polimórficos. A esta distribuição, gerada sob uma hipótese nula (nesse caso ausência de seleção natural), é comparado o valor observado para o loco candidato à seleção natural e verifica-se sua significância estatística, geralmente assumindo um intervalo de confiança de 95%.

Evidências genéticas de desvio do equilíbrio populacional em humanos

Os estudos genéticos realizados na década de 90 com o DNA mitocondrial (mtDNA) indicaram uma forte expansão da população humana com início entre 100 e 50 mil anos atrás (Rogers e Harpending, 1992; Underhill, Jin *et al.*, 1997; Wall e Przeworski, 2000). Resultados similares foram encontrados para o cromossomo Y (Underhill, Jin *et al.*, 1997; Wall e Przeworski, 2000). A evidência genética original para esse crescimento foi um excesso de alelos raros no mtDNA em relação ao esperado pelo modelo neutro (Di Rienzo e Wilson, 1991). Estudo teórico mostrou que esta característica poderia ser decorrente de expansão populacional (Slatkin e Hudson, 1991). Assinatura similar também poderia ser causada por seleção positiva causando efeito de carona em locos neutros ligados (Braverman, Hudson *et al.*, 1995) ou mesmo por seleção purificadora atuando no mtDNA (Nachman, Brown *et al.*, 1996). Efeitos da

seleção natural podem ser melhor avaliados quando o loco candidato é comparado com a variação de vários locos independentes distribuídos no genoma, pois esses servem como um controle da variação genética resultante de processos demográficos que atuam em todo o genoma. A análise de diversos microsatélites em populações humanas também corrobora a hipótese de expansão populacional, porém a estimativa da data de início da expansão foi controversa devido à imprecisão na estimativa da taxa de mutação dos microsatélites (Di Rienzo, Donnelly *et al.*, 1998; Kimmel, Chakraborty *et al.*, 1998; Reich e Goldstein, 1998).

Em contraste com os marcadores com altas taxas de mutação, há diversos locos nucleares (estudados por re-seqüenciamento) que não apóiam o modelo de expansão populacional recente. Esse padrão é ainda mais acentuado quando são analisadas separadamente amostras de africanos e não-africanos. Essa característica é atualmente foco de intenso debate sendo que as hipóteses contemplam seleção balanceadora, cenários demográficos complexos com sucessivos gargalos populacionais, eventos de subdivisão e migração e também miscigenação dos humanos modernos com formas arcaicas (Wall e Przeworski, 2000; Templeton, 2005; Garrigan e Hammer, 2006).

Populações nativas da América vivenciaram processos demográficos complexos e relativamente recentes

De modo geral, haplótipos encontrados fora do continente africano são subconjuntos da grande variabilidade haplotípica encontrada em africanos, sugerindo reduções sucessivas no tamanho das populações que emigraram da África e se estabeleceram na Eurásia e a América (Ramachandran, Deshpande *et al.*, 2005). Além disso, a distância geográfica a partir da África está significativamente associada com a

redução de diversidade genética das populações não-africanas (Prugnolle, Manica *et al.*, 2005).

De acordo com o modelo de origem africana recente, as populações humanas se expandiram a partir da África há aproximadamente 100 mil anos atrás e substituíram completamente formas arcaicas do gênero *Homo* (Stringer e Andrews, 1988; Gibbons, 2001; Tishkoff e Kidd, 2004). Por esse modelo, a América representa um dos continentes mais recentemente ocupados por populações humanas, tendo portanto passado por sucessivos eventos demográficos, incluindo gargalos populacionais e expansões.

Marcadores autossômicos (Salzano e Callegari-Jacques, 2006), mitocondriais (Bonatto e Salzano, 1997a; b; Silva, Bonatto *et al.*, 2002) e do cromossomo Y (Santos, Rodriguez-Delfin *et al.*, 1996; Bianchi, Bailliet *et al.*, 1997; Bianchi, Catanesi *et al.*, 1998) estudados em ameríndios, mostram uma reduzida variabilidade (diversidade genética) em relação a outras populações do mundo, sugerindo um efeito de sucessivos gargalos populacionais.

Diversos estudos genéticos foram feitos com ameríndios e mostraram que as populações nativas da América do Norte, Central e Sul apresentam nível de diferenciação interpopulacional mais alto que de outras regiões do mundo. Esta característica é atribuída ao processo de deriva genética que seria uma das principais forças evolutivas que moldaram o padrão de evolução das populações nativas americanas (Salzano, 2002; Salzano e Callegari-Jacques, 2006).

Várias hipóteses procuram explicar o povoamento das Américas baseando-se em dados lingüísticos, arqueológicos ou genéticos. Atualmente existe um consenso entre os geneticistas que o local de origem do influxo migratório originário dos nativos americanos foi o nordeste asiático e que a rota utilizada nesse movimento foi uma

porção de terra localizada na região da Beríngia que atualmente se encontra coberta por água oceânica, mas que há aproximadamente 20 mil anos atrás (durante o último período glacial) interligava os continentes asiático e americano. Entretanto, existem controvérsias a respeito do número de influxos migratórios, a data de ocupação do continente americano e da região geográfica de origem.

Afinidades genéticas entre nativos americanos e povos da Mongólia (Neel, Biggar *et al.*, 1994; Kolman, Sambuughin *et al.*, 1996; Merriwether, Hall *et al.*, 1996) ou Sibéria (Torroni, Sukernik *et al.*, 1993; Grahovac, Sukernik *et al.*, 1998; Karafet, Zegura *et al.*, 1999; Santos, Pandya *et al.*, 1999; Lell, Sukernik *et al.*, 2002) foram associadas com o local de origem ancestral das populações nativas da América.

Por outro lado, Brown, Hosseini *et al.*, (Brown, Hosseini *et al.*, 1998) analisando o haplogrupo X do DNA mitocondrial, levantaram a hipótese de que povos europeus poderiam ter sido os primeiros a chegarem à América e estudos arqueológicos e paleoantropológicos, sugerem que povos pré-Mongolóides (Prous, 1995; Munford *et al.*, 1995; Lahr, 1995; Neves *et al.*, 1996) foram os primeiros migrantes do continente americano.

Greenberg *et al.* (1986) sugeriram um modelo de três influxos migratórios, originados de populações distintas, para justificar a existência de três principais grupos lingüísticos dos nativos americanos: Ameríndio, Na-Dene (do noroeste da América do Norte) e Eskimó-Aleuta (da região sub-ártica).

Entretanto, essa hipótese foi recentemente desafiada pela descoberta de um alelo do microssatélite D9S1120, em alta frequência, exclusivo de indígenas da América e compartilhado por diversas populações pertencentes a grupos lingüísticos distintos. Esse resultado sugere uma origem única e exclusiva para todos os grupos populacionais de nativos americanos (Schroeder, Schurr *et al.*, 2007).

Seleção natural em nativos americanos é um campo pouco explorado

Na última década diversos estudos genéticos de populações nativas das Américas foram feitos usando DNA mitocondrial, cromossomos sexuais e autossomos com o objetivo de caracterizar a diversidade genética desses grupos e inferir aspectos como região geográfica de origem, tempo da chegada ao continente americano, número de ondas migratórias e intensidade do gargalo populacional. Entretanto, apenas recentemente, estudos envolvendo re-seqüenciamento de locos autossômicos vem sendo feitos aplicando metodologias específicas para se avaliar a importância da seleção natural na dinâmica evolutiva destas populações e vice-versa, ou seja, utilizar as populações ameríndias, de história demográfica complexa, para compreender conceitos teóricos básicos relacionados aos mecanismos de ação de seleção natural (Fagundes, Salzano *et al.*, 2005; Tarazona-Santos e Tishkoff, 2005; Battilana, Cardoso-Silva *et al.*, 2007). Aspectos micro-evolutivos em escala intra ou intercontinental são campos abertos à exploração científica usando marcadores possivelmente sob seleção natural.

Os estudos em escala genômica com populações africanas, européias e asiáticas encontraram uma série de genes candidatos a serem alvos de seleção natural. Esses genes incluem tanto aqueles que evidenciam adaptação regional a ambientes distintos (Sabeti, Varilly *et al.*, 2007), como genes com padrão de variação que indica seleção balanceadora (Akey, Eberle *et al.*, 2004). Entretanto, populações ameríndias não foram incluídas nesses estudos. Dado que elas ocuparam uma nova região do mundo num período relativamente recente, encontraram novos ambientes à medida que realizaram essa ocupação, e apresentam tamanhos demográficos reduzidos, o que em teoria diminui a eficácia da seleção natural, há considerável interesse em compreender de que modo a seleção natural atuou sobre essas populações.

A importância evolutiva das regiões cis-reguladoras de genes

Recentemente, um número crescente de estudos evolutivos indica que regiões cis-reguladoras de genes estão sob seleção natural (Bamshad, Mummidi *et al.*, 2002; Meyer, May *et al.*, 2002; Wray, 2007). As regiões cis-reguladoras são segmentos de DNA que regulam e promovem o processo de transcrição genética através da ligação a fatores de transcrição e influenciam a expressão gênica. Tais segmentos tipicamente localizam-se à montante (5') do sítio de início de transcrição do gene e, portanto, não estão presentes no transcrito primário. Porém para alguns genes, como no caso do *CCR5*, a região cis-reguladora pode estar localizada na região não-traduzida (*UTR – untranslated region*) do transcrito primário (Mummidi, Bamshad *et al.*, 2000). As *UTRs* são transcritas, podem apresentar *exons* e *introns* e desempenham importante papel na regulação da expressão gênica em uma etapa pós-transcricional (Mignone, Gissi *et al.*, 2002). Sendo assim, certas regiões podem ter dupla função (pré e pós-transcricional) no controle da expressão gênica. Hughes *et al.*, (2005) analisando 481 genes humanos, encontrou alta taxa de heterozigose nas regiões 5' *UTRs* de genes de função imunológica, sugerindo uma assinatura de seleção balanceadora nessas regiões.

ESEs e ISEs: Uma importante categoria de elementos reguladores da expressão gênica está associada ao mecanismo de processamento (*splicing*) do RNA primário. Essas regiões denominadas *ativadores de splicing exônico (exonic splicing enhancers – ESEs)* e *ativadores de splicing intrônico (intronic splicing enhancers – ISEs)* (Blencowe, 2000; Fairbrother, Yeh *et al.*, 2002; Yeo, Hoon *et al.*, 2004), podem estar em regiões codificantes ou dentro de *UTRs*. Mutações em tais regiões podem induzir a maquinaria de *splicing (spliceosomo)* a reconhecer bordas *exons/introns* alternativas durante o processamento do *pré-RNA* e conseqüentemente alterar a expressão gênica (Cartegni, Chew *et al.*, 2002; Wang e Cooper, 2007). O trabalho de Carlini e Genut

(2006), através da análise de aproximadamente 7,5 milhões de polimorfismos de sítio único do genoma humano, sugeriu que através da análise de mutações sinônimas é possível detectar forte assinatura de seleção purificadora em *ESEs* corroborando a importância funcional dessas regiões.

CCR5: padrão de expressão e importância imunológica

O receptor 5 de quimiocina CC (CCR5) é receptor das β quimiocinas RANTES, MIP-1 α , MIP-1 β e desempenha papel fundamental para a resposta imunológica inata e para o processo inflamatório (Premack e Schall, 1996).

Rottman, Ganley *et al.*, (1997), usando um painel de anticorpos monoclonais específicos para o CCR5 humano, demonstraram através de imunohistoquímica e citometria de fluxo que o CCR5 é expresso em células provenientes da medula óssea, incluindo uma subpopulação de linfócitos e monócitos/macrófagos do sangue, órgãos linfóides primários e secundários e tecidos não inflamados. No sistema nervoso central o CCR5 é expresso nos neurônios, astrócitos e microglia. Em outros tecidos o CCR5 é expresso no epitélio, endotélio, músculo liso vascular e fibroblastos.

A região cis-reguladora do CCR5 apresenta uma assinatura de seleção balanceadora

O CCR5 ganhou grande destaque no meio científico a partir do ano de 1995 devido à descoberta de uma deleção de 32 pares de bases (chamada de $\Delta 32$) na região codificante do gene que, em homozigose, confere aos portadores, resistência à infecção pelo vírus da imunodeficiência humana (*HIV-Human Immunodeficiency Virus*) e que, em heterozigose, confere aos portadores progressão mais lenta à síndrome de imunodeficiência adquirida (*AIDS - acquired immunodeficiency syndrome*) (Benkirane, Jin *et al.*, 1997).

Entretanto, estudos feitos por Cohen *et al.*, (1997) e Wu *et al.*, (1997) mostraram que o polimorfismo na região codificante do *CCR5* não era o único fator determinante da progressão mais lenta à *AIDS* e da variabilidade de expressão do *CCR5* entre indivíduos.

Estes resultados estimularam o estudo detalhado da região cis-reguladora do *CCR5* e de seus efeitos na progressão à *AIDS* e na expressão desse receptor.

Martin *et al.*, (1998) mostraram através de estudo de associação que indivíduos homozigotos para um haplótipo contendo o alelo *CCR5P1* na região promotora, quando infectados pelo vírus da imunodeficiência humana (*HIV – human immunodeficiency virus*), apresentavam progressão à *AIDS* mais rápida do que os indivíduos que possuíam outros genótipos na região cis-reguladora do *CCR5*.

Em 2000, Mummidi e colaboradores sequenciaram e redescreveram a organização estrutural do gene e diversos sítios polimórficos na região cis-reguladora do *CCR5* foram encontrados. Os haplótipos humanos para a região cis-reguladora do *CCR5* foram filogeneticamente classificados em sete haplogrupos (HHA-HHG) usando como grupo externo seqüências da região ortóloga de primatas não-humanos. Ensaaios com vetores de expressão contendo o gene repórter da luciferase caracterizaram variações específicas aos diferentes haplogrupos. Os haplogrupos HHA e HHC apresentaram atividade promotora reduzida em relação aos haplogrupos HHE, HHF e HHG. Esse estudo também mostrou que o processamento (*splicing*) alternativo do transcrito primário do *CCR5* gera diversas isoformas de RNA mensageiro (*RNA_m*) que diferem entre si em suas regiões não-traduzidas localizadas a montante do gene (5'-UTR) e que os polimorfismos na região cis-reguladora do *CCR5* influem no conjunto de fatores de transcrição que se ligam a essa região.

A análise populacional da região cis-reguladora do *CCR5* decorreu dos resultados obtidos pelo estudo de associação interessado nos efeitos dos haplogrupos da região controladora do *CCR5* na progressão à *AIDS* em diferentes grupos étnicos da população humana. Esses autores encontraram freqüências relativamente similares dos haplogrupos da região controladora do *CCR5* nos grupos controle de origem africana, afro-americana, asiática e caucasiana.

Em 2002, um estudo populacional usando amostras de europeus, africanos, asiáticos não-indianos e indianos, mostrou que a região cis-reguladora do *CCR5* apresenta um polimorfismo maior que o esperado em um regime de evolução sob neutralidade (diversidade nucleotídica aproximadamente 4 vezes maior que a observada para locos não-codificantes do genoma humano (Zhao, Yu *et al.*, 2006)), baixa diferenciação genética entre as populações estudadas ($F_{st}=0,025$), um alto compartilhamento de haplótipos e um excesso de sítios polimórficos em freqüências intermediárias. Os haplogrupos mais freqüentes em escala mundial (HHE e HHC) são separados por um número relativamente alto de mutações. Em conjunto, essas características são fortes indícios de ação da seleção balanceadora. O valor da estimativa de diferenciação interpopulacional F_{st} encontrado para a região cis-reguladora do *CCR5* foi praticamente três vezes menor do que os valores encontrados para 100 locos neutros (elementos Alu) ($F_{st} = 0,073$) tipados na mesma amostra, sugerindo que baixa diferenciação inter-populacional observada para a região cis-reguladora do *CCR5*, não era explicável exclusivamente por eventos demográficos ou pela estrutura populacional (Bamshad, Mummidi *et al.*, 2002).

Até o momento nenhum estudo procurou avaliar a possível correlação entre isoformas de *RNA*m do *CCR5* com sua expressão nos diferentes tecidos humanos. Entretanto, sabendo-se que o *CCR5* é expresso em diferentes tecidos, a hipótese de que

diferentes isoformas de *RNA*m possam ser funcionalmente mais adaptadas à determinados tecidos representa uma alternativa, para justificar a seleção balanceadora observada na região cis-reguladora do *CCR5*.

Objetivo e Justificativa

Nosso objetivo foi descrever a variação da região cis-reguladora do *CCR5* em populações nativas da América do Sul e fazer inferências sobre os processos micro-evolutivos que operaram nesse loco. Os processos que podem estar moldando a variação nesse gene incluem tanto eventos demográficos (por exemplo, gargalos e expansões populacionais) como seletivos.

As pressões seletivas sobre a região cis-reguladora do *CCR5* geram uma assinatura de diversidade oposta àquela esperada com base na história demográfica dos ameríndios, reduzindo a variação interpopulacional e contribuindo para uma elevação das taxas de polimorfismo. Até o momento, nenhuma população indígena da América do Sul foi estudada para a região cis-reguladora do *CCR5*, e de um modo geral há poucos estudos sobre os efeitos dos gargalos populacionais sofridos pelas populações que ocuparam as Américas sobre um gene intensamente selecionado.

No presente estudo investigamos a interação entre processos demográficos e seletivos que atuaram sobre a região cis-reguladora do gene *CCR5* de ameríndios da região amazônica brasileira. Os resultados obtidos foram comparados com aqueles de estudos de outras populações do mundo.

O polimorfismo da região cis-reguladora foi testado em relação ao esperado sob a hipótese de neutralidade e equilíbrio populacional, que constitui a hipótese nula mais geral usada em genética de populações.

Investigamos de que modo a história demográfica das populações interfere com os resultados dos testes de neutralidade. Essas análises foram feitas usando abordagens de simulação e comparações entre resultados obtidos para a região cis-reguladora do *CCR5* e marcadores neutros.

O estudo populacional foi complementado por uma análise *in silico*, que testou se havia associação entre o polimorfismo na região cis-reguladora do *CCR5* e regiões funcionais que controlam o mecanismo de *splicing* do RNA.

Materiais e Métodos

Amostras populacionais

Nós investigamos o polimorfismo na região cis-reguladora do gene *CCR5* em uma amostra de população indígena da região Amazônica composta por 62 indivíduos distribuídos entre sete grupos distintos: 13 indivíduos da tribo Arara do Laranjal, 13 Arara do Iriri, 12 Kayapó, 13 Xikrin, 2 Mapuera-Katuena, 3 Parakanã e 6 Waiãpi.

Os grupos Arara do Laranjal, Arara do Iriri e Mapuera-Katuena pertencem ao grupo lingüístico Karib, os Kayapó e Xikin falam língua do grupo Gê e os Parakanã e Waiãpi são do grupo Tupi (Callegari-Jacques, Salzano *et al.*, 1994); Mousinho-Ribeiro, *et al.*, 2003). A tribo Arara do Iriri provavelmente foi fundada por um casal de irmãos (Ribeiro-Dos-Santos, Guerreiro *et al.*, 2001) e portanto foi excluída da maioria das análises.

Os DNAs foram extraídos a partir de sangue coletado em expedições de saúde organizadas em 1990-1992 por agentes da FUNAI e pesquisadores do Laboratório de Genética Humana e Médica (LGHM) da Universidade Federal do Pará (UFPA). Foi obtido o consentimento das tribos para o uso do DNA em estudos genéticos e

antropológicos. A localização geográfica aproximada destes grupos encontra-se no mapa da Figura 1.

Figura 1: Mapa com a localização aproximada dos grupos estudados. Cores representam tribos de acordo com a legenda abaixo. Modificado de Callegari-Jacques, Salzano *et al.*, 1994.

Legenda			
Cor	Grupo	Afiliação Lingüística	Coordenadas geográficas
Vermelho	Mapuera-Katuena	Karib	0°40'S; 57°55'W
Preto	Waiãpi	Tupi	1°N, 53°W
Verde	Arara do Iriri e Laranjal	Karib	3°30'S; 53°21'W
Azul	Kayapó, Xikrin	Gê	7°28'S; 52°44'W
Cinza	Parakanã	Tupi	5°45'S; 51°52' W

Para as análises comparativas, usamos três conjuntos de dados distintos, obtidos da literatura. O primeiro conjunto de dados nos foi fornecido pelo Dr. Michael Bamshad da Universidade de Utah (E.U.A.) e continha 13 sítios polimórficos (*SNPs* - *single nucleotide polymorphism*) da região cis-reguladora do *CCR5*, sem fase resolvida, de amostras de africanos, asiáticos não-indianos, indianos e europeus (Bamshad, Mummidi

et al., 2002). O segundo consistiu de 783 microssatélites distribuídos ao longo dos autossomos humanos em uma amostra de asiáticos não-indianos e nativos sul-americanos (Rosenberg, Mahajan *et al.*, 2005). O terceiro consistiu de 18 locos neutros de populações nativas da América do Sul. Esses últimos, foram analisados por Fagundes, Ray *et al.*, (2007) em recente trabalho usado para se estimar um modelo demográfico para a população humana.

A amostra africana contendo a região cis-reguladora do gene *CCR5* foi composta por 3 pigmeus Biaka, 7 pigmeus Mbuti, 3 Alur, 3 Nande, 3 Hema, 1 keniano, 2 nigeriano, 3 Sotho, 3 Nguni, e 3 San, totalizando 31 indivíduos. A amostra de asiáticos não-indianos foi composta por 3 indivíduos do Cambódia, 5 chineses (Han), 6 japoneses, 3 aborígenes malaio, 3 vietnamitas, 2 indivíduos da Mongólia, 1 das Ilhas Nicobar, 3 dagestaneses e 1 descendente de asiático, totalizando 27 indivíduos. A amostra europeia foi composta por 4 indivíduos do norte da Europa, 5 franceses, 4 italianos, 5 poloneses, e 6 finlandeses, totalizando 24 indivíduos. A amostra de indianos foi composta por 3 Brahmin, 3 Kshatriya, 2 Kapu, 3 Yadava, 3 Mala, 3 Madiga, 3 Relli, 3 Irula, 3 Khonda Dora, 1 Santal, e 3 Chenchu, totalizando 30 indivíduos.

O segundo conjunto de dados continha 783 microssatélites genotipados nas seguintes populações: chineses (44), cambodianos (11), japoneses (29), Adygei (17), indígenas colombianos (13), Karitianas (24) e Suruis (21). Entre parênteses estão os números de indivíduos amostrados por população. Esses dados estão disponíveis no sítio virtual do Dr. Noah Rosenberg no seguinte endereço eletrônico: <http://rosenberglab.bioinformatics.med.umich.edu/diversity.html>. Essas amostras representativas da Ásia e América foram escolhidas devido à maior proximidade geográfica com os grupos analisados para a região cis-reguladora do *CCR5*.

O terceiro conjunto de dados foi formado pelos locos *T2557*, *T1412*, *T2064*, *T2558*, *T2568*, *T1469*, *T2924*, *T15*, *T2020*, *T1505*, *T2659*, *T2988*, *T787*, *T813*, *T2294*, *T2019*, *T2021*, *T292* seqüenciados em uma amostra de 14 nativos sul-americanos composta por 2 Kuben-Kran-Keng (Amazônia), 2 Lengua (Paraguai), 2 Quechua (Peru), 2 Tiryo (Amazônia), 2 WaiWai (Amazônia), 2 Xavante (Amazônia), 2 Zoro (Amazônia). Esses 18 locos foram originalmente analisados por Chen e Li, (2001) e foram selecionados dos autossomos humanos seguindo os seguintes critérios: pertencerem a *contigs* do genoma humano com ausência de genes conhecidos e validados *in vitro*; estarem a pelo menos 5 kb de distância, tanto a 5' como a 3', de potenciais genes preditos através de Bioinformática (*in silico*); não possuir elementos repetitivos e terem comprimentos maiores que 800 pares de bases.

Estratégia de re-amostragem

No presente estudo realizamos a análise comparativa entre o polimorfismo observado para a região cis-reguladora do *CCR5* e o de outros locos. Entretanto, diferenças no tamanho amostral influenciam as medidas de diversidade genética e as estatísticas-sumário usadas nos testes de neutralidade. Para descartar tais influências, realizamos sorteios para gerar bases de dados com tamanhos amostrais iguais para os diferentes marcadores. Além disso, também foram escolhidas populações que pertenciam a regiões geográficas similares àquelas analisadas para a região cis-reguladora do *CCR5*.

Estimando a significância estatística do F_{st} observado para a região cis-reguladora do CCR5

Partindo da base de dados de microssatélites foram realizados sorteios de 2000 amostras, sendo 1000 de asiáticos e 1000 de ameríndios. Cada amostra asiática foi composta por 5 indivíduos chineses, 6 japoneses, 8 adygeis e 8 cambodianos, totalizando 27 indivíduos. Cada amostra ameríndia foi composta por 6 indígenas colombianos (Piapoco e Curripaco), 13 Karitianas e 12 Suruis, sendo as duas últimas tribos da região amazônica brasileira. O tamanho da amostra asiática, 27 indivíduos, foi escolhido por ser o mesmo da amostra contendo os dados da região cis-reguladora do CCR5. O tamanho da amostra ameríndia, 31 indivíduos, foi escolhido por ser o mesmo de uma sub-amostra contendo os dados da região cis-reguladora do CCR5 das tribos Kayapó (12), Arara do Laranjal (13), Waiãpi (6). Escolhemos essas três tribos principalmente por que elas são de grupos linguísticos distintos, o que teoricamente, resultaria em valores de F_{st} mais elevados, tornando o teste mais conservador. Além disso, outros dois critérios usados na escolha foram: maior tamanho amostral e menor desvio das proporções genotípicas esperadas pelo princípio de Hardy-Weinberg (ver Tabela 11).

Para os 1000 pares de amostras foram calculados 1000 valores de F_{st} . Em seguida, esses valores foram utilizados para geração da distribuição de F_{st} . À esta distribuição foi comparado o valor de F_{st} observado para a região cis-reguladora do CCR5 de asiáticos e ameríndios. Nesse caso, a amostra de ameríndios foi formada apenas pelas populações Kayapó, Arara do Laranjal e Waiãpi.

Estimando a significância estatística do D de Tajima observado para a região cis-reguladora do CCR5

Estratégia similar foi utilizada para se comparar o polimorfismo da região cis-reguladora do *CCR5* de ameríndios com os 18 locos neutros mencionados. Porém, nesse caso, re-amostamos os dados da região cis-reguladora do *CCR5* para gerar amostras de 14 indivíduos, distribuídos entre populações de modo semelhante ao usado por Fagundes *et al.*, (2007). Partindo de uma amostra contendo todos os indivíduos ameríndios analisados para a região cis-reguladora do *CCR5*, exceto os Arara do Iriri, foram geradas 1000 amostras através de sorteios. Cada amostra foi composta por 3 indivíduos Arara do Laranjal, 3 Kayapó, 3 Xikrin, 3 Waiãpi e 2 Parakanã totalizando 14 indivíduos, o mesmo tamanho amostral da base de dados dos locos neutros. Para cada uma das 1000 amostras, foi calculado o valor de D de Tajima e em seguida os 18 valores de D de Tajima observados para os locos neutros foram comparados com a distribuição dos 1000 valores de D obtidos a partir das amostras simuladas.

Para a realização dos sorteios usamos o banco de dados *MySQL* através de um *stored procedure* (*script* de repetição em linguagem *SQL*) contendo a função de aleatorização *RAND*. A conversão de formato dos arquivos saídos do banco de dados para o formato Arlequin usados para posteriores análises populacionais foi feita através dos programas de edição de texto do sistema *Linux* (*sed*, *vi*, *cut*, *split*), de um *script* em linguagem *Awk* escrito por Diogo Meyer e do programa *Convert* (Glaubitz, 2004).

Técnicas de Biologia Molecular

Amplificação genômica

Inicialmente, todas as amostras de DNA usadas neste trabalho foram submetidas à amplificação genômica através da metodologia de amplificação através de múltiplos

deslocamentos (*multiple displacement amplification* ou *MDA*) implementada no kit *GenomiPhi* v.2 comercializado pela *GE HealthCare*. Na amplificação de genoma inteiro por *MDA*, a DNA polimerase do bacteriófago *Phi29* amplifica exponencialmente o DNA molde através de uma reação enzimática em que uma fita de DNA é deslocada à medida que a outra é replicada, tendo oligonucleotídeos aleatórios com o tamanho de seis pares de bases como iniciadores. A polimerase do bacteriófago *Phi29* gera produtos relativamente grandes (de 12 a 100 kpb) e possui atividade exonucleolítica 3'→5', apresentando baixa taxa de erro no processo replicativo (Short, Kennedy *et al.*, 2005). Este procedimento permitiu um aumento de aproximadamente 50 vezes na massa de DNA das amostras. Para cada amostra foi utilizado 10ng de DNA original seguindo o protocolo do fabricante.

Amplificação da região cis-reguladora do CCR5

A região cis-reguladora do gene *CCR5* foi amplificada através da reação em cadeia da polimerase (*polymerase chain reaction* – *PCR*) utilizando o par de iniciadores descritos na Tabela 1.

Tabela 1: Sequências de iniciadores usados nas reações de amplificação da região cis-reguladora do *CCR5*.

nome	seqüência (5' – 3')
<i>CCR5L2</i>	CCAAACTGTGACCCTTTCC
<i>CCR5R2b</i>	CTCATAGCTTCAGATAGATTAT

Este par de iniciadores amplifica uma região de 1009 pares de base da região controladora do gene *CCR5* que possui 1108 pares de base. Os iniciadores da PCR estão representados na Figura 2.

Inicialmente utilizamos o par de iniciadores L2 e R2 descritos em Bamshad, Mummidi *et al.*, (2002), entretanto o iniciador R2 foi substituído pelo R2b por razão que será descrita na seção “Resultados e Discussão”.

A reação da PCR foi realizada usando o protocolo presente na Tabela 2.

Tabela 2: Protocolo da PCR usado na amplificação da região controladora do gene *CCR5*.

Reagentes	Volume (µl)	Concentração Final
10x PCR Buffer*	5	1x
Cloreto de Magnésio (50mM)*	1,5	1,5mM
DNTP Mix (10mM)*	1	0,2mM
Iniciador R2b (10 µM)	1	0,2µM
Iniciador L2 (10 µM)	1	0,2µM
Platinum TAQ (5 U/µl)*	1	-
DNA	1	50ng
H ₂ O MilliQ	38,5	-

* Invitrogen

As reações foram submetidas ao programa de termociclagem descrito na Tabela 3.

Tabela 3: Ciclos de temperatura usados na PCR para a região controladora do gene *CCR5*. Programa usado no termociclador modelo PTC-200 da marca MJ.

Temperatura (°C)	Tempo (seg.)	Repetições
96	180	0
96	25	
54	60	39
72	120	
72	480	0

Foram aplicados 2 µl dos produtos da PCR em gel de agarose 2% para análise de tamanho e massa através de comparação com o marcador DNA *Low Mass Ladder* (Invitrogen).

Purificação

A purificação do produto da PCR foi feita com as enzimas Exonuclease I e *Shrimp Alkaline Phosphatase* comercializadas pela *GE Healthcare* seguindo o protocolo recomendado pelo fabricante.

Sequenciamento automático

A PCR feita a partir de DNAs resultantes de amplificação de genoma inteiro (implementado pelo GenomiPhi) gerou, na maioria das amostras, bandas não específicas adicionais ao produto de interesse. Sendo assim os iniciadores de amplificação (R2b, L2) não foram usados nas reações de sequenciamento, pois eles geraram seqüências de baixa qualidade devido ao anelamento destes iniciadores tanto no produto de interesse como nos produtos inespecíficos. Portanto, para fazer a leitura da região de interesse, foram usados os iniciadores internos B, I, E, A, D, L2b (Figura 2).

As seqüências dos iniciadores usados nas reações de sequenciamento estão listadas abaixo (Tabela 4).

Tabela 4: Sequências de iniciadores usados nas reações de seqüenciamento da região cis-reguladora do *CCR5*.

nome	seqüência (5'-3')
L2b	TATGAGCACTTGGTGTTCGCC
CCR5D	CCGTAAATAAACCTTCAGACCAG
CCR5E	CAGGGCTTTTCAACAGTAAGG
CCR5A	CGGGCTTTTCTCACTGG
CCR5B	AACTATGGGCTCACGGG
CCR5I	GGTTAATGTGAAGTCCAGG

Figura 2: Esquema do gene *CCR5*. Os retângulos representam exons, quadrados pretos representam introns. As posições de início e término dos exons estão indicadas acima dos retângulos. A flecha grande indica a região analisada neste estudo. Os iniciadores de amplificação e seqüenciamento estão representados por setas pequenas de cores cinza e branca respectivamente. As setas para a direita simbolizam os iniciadores *forward* e para esquerda os iniciadores *reverse*. Os nomes dos iniciadores estão indicados abaixo das setas. Modificado de Mummidi *et al.*, (2000).

As reações de seqüenciamento foram realizadas usando o protocolo descrito na Tabela 5.

Tabela 5: Protocolo da reação de seqüenciamento.

Reagentes	Volume (µl)	Concentração Final
BigDye* ou ET Terminator**	2	-
Iniciador (10µM)	1	1µM
DNA molde	-	30ng
H ₂ O MilliQ	q.s.p 10 µl	-

*Applied Biosystems, ** Ge HealthCare

As reações foram submetidas ao programa de termociclagem presente na Tabela 6.

Tabela 6: Ciclos de temperatura usados na reação de seqüenciamento para a região controladora do gene *CCR5*. Programa usado no termociclador modelo PTC-200 da marca MJ.

Temperatura (°C)	Tempo (seg.)	Repetições
95	20	40
56,5	15	
60	60	

Precipitação

As reações de seqüenciamento foram precipitadas usando o seguinte protocolo:

Adicionar 8 µl de isopropanol 75%, deixar 20 minutos em local escuro, centrifugar a 13000 r.p.m. a temperatura ambiente durante 25 minutos. Descartar o sobrenadante, adicionar 200 µl de etanol 70%, centrifugar a 13000 r.p.m., a temperatura ambiente durante 10 minutos. Secar os tubos com a tampa aberta em local escuro durante 3 horas.

As amostras foram submetidas à leitura no analisador genético MegaBACE 1000, disponível no Centro de Estudos do Genoma Humano (IB-USP), ou no seqüenciador ABI3100 do Laboratório de Biologia Molecular de Plantas, do Departamento de Botânica do IB-USP.

Análises das seqüências de DNA

A qualidade das seqüências foi avaliada através do programa *Phred* (Ewing e Green, 1998) e montadas através dos programas *Phrap* (http://www.phrap.org/phredphrapconsed.html#block_phrap).

Para detectar os indivíduos heterozigotos foi usado o programa *PolyPhred* (Nickerson, Tobe *et al.*, 1997). Esse programa identifica sítios heterozigotos em seqüências produzidas através do seqüenciamento direto de produtos da PCR. *PolyPhred* usa as informações geradas pelo *Phred* e os alinhamentos gerados pelo *Phrap* para identificar e classificar sítios heterozigotos. Cada sítio recebe um valor que varia de 1 a 6. Sítios com valor igual a 1 tem chance de 97% de serem verdadeiros e os sítios com valor igual a 6 tem 1% de chance de serem verdadeiros. Os resultados dessas análises são visualizados através do programa *Consed* (Gordon, Abajian *et al.*, 1998) (Figura 3).

Figura 3: a) Janela parcial do programa *Consed* mostrando alinhamento de seqüências criado pelo programa *PHRAP*. Os sítios polimórficos foram marcados com três cores distintas que significam: (1) sítios polimórficos contendo alelos minoritários* em homozigose, (2) sítios polimórficos com alelos majoritários** em homozigose e (3) sítios polimórficos com os dois alelos presentes (i.e. em heterozigose). b) Cromatogramas de três indivíduos distintos. O cromatograma superior apresenta na posição 936/284 o alelo minoritário em homozigose (T/T). O cromatograma do meio apresenta, nesta mesma posição, o alelo majoritário em homozigose (G/G). O cromatograma inferior apresenta, nesta mesma posição, os dois alelos (G/T) sendo, portanto, um heterozigoto.

*Alelo minoritário: alelo que aparece em menor freqüência em um sítio polimórfico.

**Alelo majoritário: alelo que aparece em maior freqüência em um sítio polimórfico.

A inspeção visual de cada sítio polimórfico deve ser feita para complementar as análises do *Polyphred*. O *Consed* facilita esta inspeção, pois permite a visualização de todos os cromatogramas alinhados na posição polimórfica escolhida.

Inferência da fase genotípica e nomenclatura dos haplótipos

Caso um indivíduo seja heterozigoto em mais de um sítio da seqüência estudada não é possível especificar diretamente quais alelos destes sítios ocorrem em um mesmo cromossomo e dizemos que o genótipo tem fase desconhecida. Nesses casos, para a definição da fase, foi utilizada a técnica de inferência estatística dos haplótipos implementada no programa *Phase* (Stephens, Smith *et al.*, 2001).

Este programa usa a estatística Bayesiana associada aos conceitos da teoria da coalescência para reconstruir os haplótipos mais prováveis para cada genótipo da amostra. Assim como outros métodos de estimativa de freqüências haplotípicas, o algoritmo do *Phase* usa as freqüências dos homozigotos presentes na amostra. Entretanto, baseado na teoria da coalescência o programa também considera a

semelhança ao nível de sequência de DNA de toda a amostra ao calcular as probabilidades dos haplótipos serem reais.

A inferência de fase foi realizada separadamente para as amostras de africanos, asiáticos, europeus, indianos e ameríndios. Os detalhes sobre a inferência de fase e caracterização dos haplótipos da amostra do Velho Mundo estão na seção “Resultados”.

Análises populacionais

Inicialmente foi feito para cada população e para o conjunto das populações o cálculo das frequências haplotípicas e dos respectivos desvios padrão. A variância da frequência alélica (haplotípica) é dada pela fórmula abaixo e o desvio é calculado como raiz quadrada da variância,

$$V(p_i) = \frac{\hat{p}_i(1 - \hat{p}_i)}{n - 1}$$

onde p_i é a frequência do alelo i , n é o número total de alelos da amostra

Para verificar se as frequências genotípicas observadas estavam de acordo com o esperado pelo princípio de Hardy-Weinberg (HW), foram calculadas as significâncias dos desvios através do método de Guo e Thompson (1992), implementado no programa *Arlequin* (Schneider, Roessli *et al.*, 2000). Esse teste foi aplicado para a amostra total (representando o conjunto total de amostras de populações ameríndias) e também para cada população individual. Desvios significativos das proporções genotípicas esperadas pelo equilíbrio de (HW) são compatíveis com vários processos micro-evolutivos, incluindo a ocorrência de endocruzamento, casamentos preferenciais, seleção natural, e estrutura populacional. Além disso, um excesso de homozigotos também pode resultar

de erros técnicos na execução das etapas de amplificação, seqüenciamento ou detecção de heterozigotos.

As seguintes estatísticas-sumário de diversidade genética foram calculadas para cada população e para a amostra total através do programa *Arlequin*.

Número de cromossomos estudados (2n): Representa o número total de cromossomos presentes na amostra, e é igual ao dobro do número de indivíduos diplóides presentes na amostra (para o caso de genes autossômicos); é o denominador da fração no cálculo das freqüências haplotípicas.

Número de sítios polimórficos (S): Representa o total de sítios polimórficos observados no conjunto de seqüências alinhadas, é usado para estimativas de diversidade nucleotídica.

Número médio de diferenças entre pares de seqüências (π): É uma média de diversidade genética de um loco baseada nas freqüências alélicas por sítio. Representa a média aritmética das diferenças por sítio entre os pares de seqüências da amostra. Quando este valor é dividido pelo comprimento da seqüência passa a ser chamado *diversidade nucleotídica* e geralmente é representado em porcentagem (Nei, 1987).

Número de haplótipos (k): Representa o total de haplótipos distintos encontrados na amostra.

Diversidade gênica (H): É uma medida de diversidade genética baseada nas freqüências haplotípicas encontradas na amostra. É igual à taxa de heterozigose esperada sob Hardy-Weinberg corrigida pelo tamanho amostral,

$$\hat{H} = \frac{n}{n-1} \left(1 - \sum_{i=1}^k p_i^2 \right)$$

onde n é o número de cromossomos da amostra, k é o número de alelos (haplótipos) e p_i é a freqüência do alelo (haplótipo) i .

As estatísticas π e θ podem ser usadas para estimar θ , que é uma das medidas mais importantes da diversidade genética intrapopulacional pois expressa a seguinte relação: $\theta = 4N\mu$, onde, N é o tamanho efetivo da população e μ é a taxa de mutação do loco (Kimura, 1991).

A teoria da coalescência demonstrou que as medidas empíricas de diversidade genética π e θ_w são estimadores de θ em um cenário em que a população evolui de acordo com o modelo neutro de infinitos sítios. Nesse modelo mutacional cada nova mutação gera um novo alelo ou haplótipo, ou seja, homoplasias não são admitidas.

Theta W ou Theta S (θ_w ou θ_s): É uma medida de diversidade genética baseada no número de sítios polimórficos (S) corrigida pelo tamanho amostral (Nei, 1987).

$$\theta_s = \frac{S}{a_1}$$

Onde S é o número de sítios polimórficos da amostra total e a_1 é dado por

$$a_1 = \sum_{i=1}^{n-1} \frac{1}{i}$$

onde n é número de seqüências da amostra total.

Testes de neutralidade

D de Tajima: Este teste realiza uma comparação entre os estimadores de *Theta* (θ_π e θ_s). Para uma população ideal que evolui de modo estritamente neutro, seguindo o modelo de infinitos sítios, e está em condições de equilíbrio, as esperanças para as duas estimativas são iguais, de modo que o valor esperado para a diferença entre elas, expressa por D , é igual a zero (Tajima, 1989). A significância estatística do valor de D é geralmente obtida através de comparação do D observado com a distribuição de valores de D esperados sob a hipótese nula de neutralidade, modelo de infinitos sítios e

equilíbrio populacional. Essa distribuição pode ser gerada através de simulação coalescente. As amostras geradas por simulação devem ter o mesmo tamanho da amostra de interesse enquanto que o loco gerado por simulação deve ter o mesmo tamanho (em pares de bases) e a mesma taxa de mutação do loco de interesse. Valores de D estatisticamente significativos podem indicar não só a ação de pressões seletivas como também a existência de desvios do equilíbrio populacional, causado por fatores demográficos como expansão, redução e subdivisão populacional (Bamshad e Wooding, 2003).

Teste de homozigose de Ewens-Watterson: Este teste é baseado na teoria de amostragem de alelos neutros de Ewens (Ewens, 1972). Este teste compara a taxa de homozigose esperada sob Hardy-Weinberg com a distribuição de valores de Fe esperada para amostras simuladas sob o modelo estritamente neutro de infinitos alelos que apresentam o mesmo tamanho e o mesmo número de haplótipos distintos (k) da amostra de interesse (Fe , soma dos quadrados das frequências alélicas).

$$F_E = \sum_{i=1}^k p_i^2$$

Onde k : número de alelos (haplótipos), p_i : frequência do alelo (haplótipo) i .

A significância do teste é dada pela probabilidade de se observar amostras aleatórias com valores de Fe iguais ou menores do que o valor obtido para a amostra de interesse. Esse tipo de desvio significativo, isto é, quando mais de 95% dos valores simulados de Fe são maiores que o valor observado, é indicativo de excesso na taxa de heterozigose da amostra de interesse que pode ser um efeito da seleção balanceadora. O teste poderia também ser aplicado para o outro extremo, isto é, quando o Fe observado

situa-se na cauda superior da distribuição, indicando pouca diversidade. Isso seria observado após uma varredura seletiva.

Teste de neutralidade para amostras populacionais que não estão em condições de equilíbrio através de simulações coalescentes

Nós usamos o programa *Simcoal* v.2.1.2 (Excoffier, Novembre *et al.*, 2000) para realizar simulações coalescentes e estimar a diversidade nucleotídica (π) e valores de D de Tajima para amostras populacionais representativas de ameríndios e assim comparar os valores observados empiricamente para a região cis-reguladora do *CCR5* com os valores simulados em cenário demográfico sem seleção natural.

A taxa de mutação da região cis-reguladora do *CCR5* usada na simulação foi de $1,03 \times 10^{-8}$ por sítio por geração, estimada a partir da fórmula $\mu = K/2t$ (Li e Graur, 1991) onde μ é a taxa de mutação no modelo neutro, K é o número de substituições ocorridas entre espécies e t é o tempo em anos da divergência entre as espécies. O tempo de geração usado para conversão da unidade da taxa de mutação de “loco/ano” para “sítio/geração” foi de 25 anos. A região cis-reguladora do *CCR5* apresentou 7 diferenças fixas entre humanos e chimpanzés (Bamshad, Mummidi *et al.*, 2002). O tempo de divergência entre humanos e chimpanzés utilizado foi de 6 milhões de anos atrás.

O modelo demográfico utilizado na simulação foi baseado no modelo teórico de substituição recente a partir da África com crescimento exponencial (*AFREG- African recent with exponential growth*) proposto por Fagundes *et al.*, (2007). Esse modelo foi escolhido como mais provável quando comparado com outros dois modelos básicos: de Assimilação, que postula que os humanos modernos que emigraram da África sofreram miscigenação com *Homo erectus* (Templeton, 2005) e de Evolução Multiregional que

postula origens independentes da anatomia humana moderna na Ásia e África porém relativamente sincronizadas (Eswaran, Harpending *et al.*, 2005). As amostras contendo sequências de DNA, geradas através de simulação coalescente usando o modelo AFREG, apresentaram valores de diversidade genética, compatíveis com os valores observados para 50 locos neutros sequenciados em africanos, asiáticos e ameríndios. O modelo *AFREG* sugere que uma população arcaica africana de tamanho efetivo igual a 12.772 indivíduos deu origem aos humanos modernos há 141 mil anos após um evento de gargalo populacional associado à especiação que reduziu o tamanho efetivo da população para 600 indivíduos. Crescendo exponencialmente numa taxa de $r=0,0010$, uma parcela da população humana moderna com tamanho efetivo de 462 indivíduos migrou da África para a Ásia há 51.102 anos. Segundo esse modelo, a América teria sido ocupada à 10.500 anos atrás por 450 indivíduos. Essa estimativa da idade da ocupação da América é mais recente do que as usualmente consideradas, porém o limite superior do intervalo de confiança estimado inclui as datas dos sítios arqueológicos mais antigos, de aproximadamente 14 mil anos atrás (Meltzer, 1997). O tamanho efetivo da população fundadora que chegou à América estimado pelo modelo *AFREG* foi seis vezes maior que o estimado por Hey (2005) e está de acordo com resultados obtidos para locos nucleares (Battilana, Fagundes *et al.*, 2006) e DNA mitocondrial (Bonatto e Salzano, 1997a).

Rede de haplótipos

Este método de representação das relações filogenéticas entre os táxons é mais adequado que a representação de árvores bifurcantes para detecção de eventos de recombinação, homoplasias e a ocorrência de mutações paralelas ou reversas que atuam no nível de espécie. Neste trabalho a rede de haplótipos foi construída através do

algoritmo *Median Join* implementado no programa *Network* (Bandelt, Forster *et al.*, 1999). Uma distinção fundamental entre métodos de rede e de árvores é que os primeiros, aceitam a hipótese de haver haplótipos ancestrais presentes na amostra e, não impõem, como no caso das árvores bifurcantes, que os haplótipos atuais estejam apenas nos ramos terminais da árvore.

Diferenciação genética interpopulacional - Fst

A diferenciação genética entre as amostras populacionais foi estimada através do índice de fixação *Fst* que mede a proporção da diversidade genética total que está distribuída entre as populações. Estimativas de *Fst* baseadas na frequência haplotípica (*Fst* padrão) e na diversidade nucleotídica foram calculadas. A estimativa baseada na diversidade nucleotídica é similar ao *Fst* padrão, porém a medida de diversidade genética básica utilizada é o π , em lugar da diversidade gênica (*H*), utilizada no método padrão. A significância estatística destas estimativas foi calculada através de 10000 permutações. Essas permutações consistem no sorteio e troca de indivíduos entre populações, permitindo que se gere a distribuição de valores da estatística de diferenciação sob a hipótese nula de não haver diferenciação entre elas.

Análises funcionais: Identificação de ativadores e inativadores de splicing exônico

(ESEs – exonic splicing enhancer, ISEs – intronic splicing enhancers)

Os *ESEs* e *ISEs* são pequenas seqüências de nucleotídeos, com aproximadamente 6 pares de bases de tamanho que estão presentes nos exons e introns respectivamente e favorecem o *splicing* do pre-*RNA*m. Os *ESEs* foram inicialmente identificados através de abordagem computacional e tais regiões foram selecionadas por atenderem aos seguintes critérios: I) Serem significativamente mais frequentes em

exons que em introns. II) Serem significativamente mais frequentes em exons com sítios de *splicing* fracos (não-canônicos) que em exons com sítios de *splicing* fortes (canônicos). Posteriormente, a importância funcional dos *ESEs* foi comprovada em experimentos *in vitro* (Fairbrother *et al.*, 2002).

Usando uma estratégia de busca similar àquela usada para identificação de *ESEs*, Yeo, *et al.*, (2004) identificaram os *ISEs*.

Para a identificação de possíveis ativadores de *splicing* nos exons (*ESEs*) e introns (*ISEs*) na região cis-reguladora do *CCR5* usamos a seguinte metodologia básica. Foram selecionadas, em torno de cada *SNP* (descritos na Figura 5), seqüências de 6 pares de bases. Essas seqüências foram escolhidas através do método de janela deslizante, de modo, que o *SNP* ocupou cada uma das 6 possíveis posições nos micro-haplótipos (hexâmeros-candidatos). Como os *SNP* da região cis-reguladora do *CCR5* apresentaram apenas dois alelos, obtivemos 12 micro-haplótipos para cada *SNP* (Figura 4). As exceções à essa regra devem-se aos *SNPs* que estavam localizados a menos de 6 pares de distância entre si (Figura 5). Nesse caso, mais de um *SNP* foi incluído no micro-haplótipo e, as combinações de alelos já descritas na literatura (haplótipos conhecidos), foram exploradas na busca por *ESEs* e *ISEs*.

Figura 4: Estratégia de identificação de *ESEs* e *ISEs*. Em torno de cada *SNP*, descrito para a região cis-reguladora do CCR5, foram escolhidos hexâmeros-candidatos de modo que o *SNP* ocupou cada uma das 6 possíveis posições nesses micro-haplótipos. Como cada *SNP* apresentou 2 alelos, foram obtidos 12 hexâmeros-candidatos para cada *SNP*.

A identificação das bordas exon/intron da região cis-reguladora do CCR5 foi obtida da anotação descrita em Mummidi *et al.*, (1997) e a identificação dos *SNPs* presentes em exons e introns foi baseada no trabalho de Mummidi *et al.*, (2000) e Bamshad *et al.*, (2002).

Por fim, cada micro-haplótipo exônico foi alinhado aos 238 *ESEs* definidos por Fairbrother *et al.*, (2002) através da ferramenta de busca *Rescue-ESE*. Estratégia similar foi aplicada aos micro-haplótipos intrônicos, só que nesse caso a ferramenta de busca utilizada foi o *Acescan 2*. Esses programas são baseados simplesmente no alinhamento entre a sequência de interesse e os 238 possíveis *ESEs* descritos por Fairbrother *et al.*, (2002) ou nos *ISEs* descritos por Yeo *et al.*, (2004) respectivamente. Os programas estão disponíveis na internet nos seguintes endereços eletrônicos:

Rescue-ESE - <http://genes.mit.edu/burgelab/rescue-ese/>

Acscan2 - <http://genes.mit.edu/acscan2/index.html>

Para testar uma possível associação entre a frequência alélica dos *SNPs* e sua localização em regiões funcionais ou não-funcionais, usamos o teste de qui-quadrado e o teste exato de Fisher a partir de tabelas de contingência 2x2.

Resultados

Definição dos haplótipos e nomenclatura utilizada

Dos 13 sítios polimórficos da região cis-reguladora do *CCR5* encontrados por Bamshad, Mummidi *et al.*, (2002) nas amostras de africanos, asiáticos não-indianos, europeus e indianos (-2852, -2838, -2733, -2554, -2459, -2454, -2256, -2150, -2135, -2132, -2086, -1915 e -1835) sete (-2733, -2554, -2459, -2135, -2132, -2086 e -1835) já haviam sido descritos no trabalho de Mummidi *et al.*, (2000). Os seis novos sítios, foram encontrados em apenas uma cópia na amostra total (*singletons*). Os sítios -2852, -2838, localizados na extremidade 5' da região não foram amplificados pelo par de iniciadores utilizados neste trabalho e, portanto foram removidos do conjunto de dados analisados.

Em 2002, Bamshad, Mummidi *et al.*, utilizaram a mesma nomenclatura de Mummidi *et al.*, (2000) para identificar os haplogrupos encontrados na amostra mundial, porém, as seqüências dos haplótipos encontrados nessa amostra não foram descritas e a nomenclatura usada foi exclusiva.

Com o intuito de identificar, com maior resolução, os haplótipos pertencentes aos haplogrupos encontrados nas populações do Velho Mundo e utilizar uma nomenclatura haplotípica comum a todas as populações analisadas neste trabalho usamos uma definição haplotípica baseada no trabalho de Mummidi *et al.*, (2000). Sendo assim, inferimos os haplótipos das amostras africana, asiática não-indiana, européia, indiana e ameríndia utilizando um conjunto de dados com 37 sítios, composto por 33 descritos em Mummidi *et al.*, (2000) e mais os quatro novos sítios descritos em Bamshad, Mummidi *et al.*, (2002) (-2454, -2256, -2150, -1915) (Figura 5). Dos 33 sítios descritos por Mummidi *et al.*, (2000), 26 não foram polimórficos na amostra mundial analisada por Bamshad, Mummidi *et al.*, (2002) e, portanto apresentaram os alelos majoritários fixados (Bamshad M. comunicação pessoal).

Os sítios -2733, -2554, -2459, -2135, -2086 e -1835 estão presentes no banco de dados de polimorfismos genéticos *dbSNP* (<http://www.ncbi.nlm.nih.gov/projects/SNP/>) sob os seguintes números de acesso: RS2856758, RS2734648, RS1799987, RS1799988, RS1800023, RS1800024 respectivamente.

A sequência de todos os haplótipos da região cis-reguladora do *CCR5* já encontrados para a espécie humana estão descritos na Figura 5.

A inclusão de todos os haplótipos na Figura 5, incluindo os raros encontrados por Mummidi *et al.*, (2000), se justifica devido às análises de possível associação entre polimorfismos raros e não-raros com regiões funcionais envolvidas em *splicing* que serão descritas posteriormente.

Figura 5: Haplótipos (1-28, rec1, rec2, rec4) e haplogrupos (HHA-HHG) da região cis-reguladora do CCR5. Os haplogrupos foram definidos filogeneticamente, em Mummid *et al.*, (2000), usando a sequência homóloga de chimpanzé (*Pan troglodytes*) como grupo externo. Os números na parte inferior da figura correspondem aos sítios polimórficos encontrados em humanos. A sequência referência (sublinhada) foi escolhida por ser a mais similar à sequência de chimpanzé. Pontos indicam identidade de bases em relação à sequência referência. *SNPs* comuns à vários haplótipos estão agrupados por caixas coloridas e caracterizam os haplogrupos. Demais *SNPs* caracterizam cada haplótipo. Sítios polimórficos cujos alelos foram encontrados em frequências intermediárias na população humana estão indicados por um traço vertical. Haplótipos rec1, rec2 e rec4 são recombinantes encontrados em frequências inferiores a 1% em populações do Velho Mundo. Haplótipos 16.1, 16.2 e 16.3 foram adicionados à descrição de Mummid *et al.*, (2000) pois apresentam sítios polimórficos descritos por Bamshad *et al.*, (2002). Tais sítios estão representados por *.

Frequências haplotípicas em amostras da África, Ásia, Europa e Índia.

Na Tabela 7 são comparadas as frequências dos haplótipos inferidos neste estudo, a partir das amostras das populações do Velho Mundo, com as frequências haplotípicas encontradas por Bamshad, Mummidi *et al.*, (2002), que usou um conjunto de dados com apenas 13 sítios polimórficos. As frequências dos haplótipos inferidos são muito similares às aquelas publicadas, e dessa forma, foi possível atribuir aos haplótipos inferidos neste estudo, uma nomenclatura baseada no trabalho de Mummidi *et al.*, (2000). Todos os indivíduos da amostra do Velho Mundo tiveram seus genótipos inferidos com probabilidades médias de serem reais superiores a 99% com exceção de 2 indivíduos africanos com probabilidade de 85%, um indiano com probabilidade de 87% e um europeu com probabilidade de 95%

Tabela 7: Quadro comparativo contendo as frequências haplotípicas para as amostras de africanos, asiáticos não-indianos, europeus e indianos. Colunas não sombreadas contém as frequências publicadas em Bamshad, Mummidi *et al.*, (2002). Colunas sombreadas contém as frequências dos haplótipos inferidos neste trabalho. Entre parênteses está o número de cromossomos analisados por população.

[illegible]

Inferência de fase e frequências haplotípicas em ameríndios

Na amostra total de 124 cromossomos, foram identificados 6 sítios polimórficos e 5 haplótipos diferentes. Nenhum haplótipo novo foi detectado, em relação aos previamente descritos em amostras de outras regiões do mundo (Bamshad, Mummidi *et al.*, 2002). As frequências dos haplótipos encontrados na amostra completa e na amostra sem a população Arara do Irii estão na Tabela 8 e a sequência dos haplótipos caracterizados por 37 sítios podem ser verificadas na Figura 5. As frequências dos haplótipos encontrados para cada população ameríndia analisada estão na Tabela 9. Todos os indivíduos da amostra de ameríndios tiveram seus genótipos inferidos com probabilidade de 100% de serem reais de acordo com critérios Bayesianos implementados no programa *Phase*.

Tabela 8: Frequências haplotípicas para a amostra total de ameríndios e para a amostra sem a tribo Arara do Irii. * número de cromossomos

** Clusters filogenéticos descritos em Mummidi *et al.*, (2000) (ver Figura 7). Entre parênteses está o desvio padrão.

Haplogrupo/Haplótipo		Ameríndios (124)*	Ameríndios sem Arara do Irii (98)*
Cluster 1**	HHA / 2	0,048 (0,019)	0,061 (0,024)
	HHC / 6	0,306 (0,042)	0,388 (0,049)
Cluster 2**	HHE / 13	0,113 (0,029)	0,143 (0,036)
	HHF / 21	0,306 (0,042)	0,163 (0,038)
	HHG / 27	0,226 (0,038)	0,245 (0,044)

Tabela 9: Frequências haplotípicas para cada grupo ameríndio analisado. * número de cromossomos

** Clusters filogenéticos descritos em Mummidi *et al.*, (2000) (ver Figura 7).

Entre parênteses está o desvio padrão.

	Haplogrupo/ Haplótipo	Arara do Iriri 26*	Arara do Laranjal 26*	Kayapó 24*	Xikrin 26*	Mapuera Katuenta 4*	Parakanã 6*	Waiãpi 12*
Cluster 1**	HHA / 2	0	0	0,042 (0,042)	0	0	0,167 (0,167)	0,333 (0,142)
	HHC / 6	0	0,308 (0,092)	0,458 (0,104)	0,5 (0,1)	0,25 (0,250)	0,167 (0,167)	0,333 (0,142)
	HHE / 13	0	0,115 (0,064)	0,25 (0,090)	0,038 (0,038)	0,25 (0,250)	0,167 (0,167)	0,167 (0,112)
Cluster 2**	HHF / 21	0,846 (0,072)	0,077 (0,053)	0,125 (0,069)	0,192 (0,079)	0,5 (0,289)	0,5 (0,224)	0,083 (0,083)
	HHG / 27	0,154 (0,072)	0,5 (0,1)	0,125 (0,069)	0,269 (0,089)	0	0	0,083 (0,083)

A importância do teste do equilíbrio de Hardy-Weinberg na detecção de falha metodológica

No primeiro ano do projeto (06/2005-06/2006) a região controladora do gene *CCR5* foi amplificada e sequenciada usando o DNA genômico de 45 indivíduos indígenas da região amazônica brasileira utilizando na PCR os iniciadores descritos por Bamshad, Mummidi *et al.*, (2002).

Entretanto, as frequências genotípicas encontradas apresentaram um significativo desvio das proporções esperadas pelo princípio de equilíbrio de Hardy-Weinberg, causado por um marcante excesso de genótipos homozigotos, havendo 39 homozigotos e apenas 6 heterozigotos ($p < 0.0001$). Desvios significativos também ocorreram quando cada população foi analisada. Na análise por sítio polimórfico também foram encontrados desvios significativos na amostra total e por população. Como parte da mesma amostra já havia sido analisada com outros marcadores genéticos (micro e mini satélites) e estes não indicaram um desvio significativo do equilíbrio de

Hardy-Weinberg (Santos E. comunicação pessoal) consideramos improvável que um desvio de tal magnitude fosse atribuível apenas ao efeito de estrutura populacional ou endogamia. O excesso de homozigotos sugeriu que os iniciadores utilizados na reação em cadeia da polimerase (*PCR*) da região controladora do *CCR5* não estavam amplificando os dois alelos presentes em cada indivíduo. Quando isso ocorre, um indivíduo heterozigoto é incorretamente classificado como homozigoto. Uma explicação para a não amplificação de alguns alelos é a presença de mutações de ponto na região de hibridação dos iniciadores, impedindo a ligação destes com a fita molde. Condições da *PCR* (temperatura de termociclagem e concentrações de reagentes) também poderiam gerar tais efeitos.

Para testar a hipótese de que polimorfismos no sítio de anelamento do iniciador resultaram na não amplificação de certos alelos, o iniciador R2 foi substituído pelo R2b, desenhado para anelar em uma região a 80 pares de base de distância da extremidade 3' do iniciador R2. Nenhuma alteração foi feita no iniciador L2.

Geração de dados usando os iniciadores R2b e L2

O novo par de iniciadores apresentou resultados satisfatórios e o desvio do equilíbrio de *Hardy-Weinberg* observado inicialmente foi solucionado para a maioria das populações analisadas. Sendo assim, o projeto prosseguiu, e dos 45 indivíduos que tiveram seus genótipos analisados, 21 foram re-analisados com os novos iniciadores, permitindo a comparação dos genótipos obtidos com o antigo par de iniciadores. Além desses indivíduos, mais 41 foram acrescentados ao estudo.

Através da comparação entre as seqüências de DNA obtidas a partir do par de iniciadores original (R2, L2) e do novo par (R2b, L2), foi possível verificar que os novos iniciadores detectaram mais de um haplótipo em vários dos indivíduos

anteriormente classificados como homozigotos. As únicas mudanças genótípicas observadas decorreram do acréscimo de um novo haplótipo em genótipos anteriormente classificados como homozigotos (Tabela 10).

Tabela 10: Tabela comparativa entre os genótipos obtidos anteriormente com os iniciadores R2 e L2 e os genótipos obtidos através da amplificação com iniciadores novos R2b, L2. Células em branco representam indivíduos que só foram analisados com os iniciadores R2b, L2. Os haplótipos sombreados são aqueles que foram detectados, com o novo par de iniciadores, em indivíduos classificados como homozigotos pelo par de iniciadores antigos.

População	Indivíduo	Iniciadores novos (R2b/L2)	Iniciadores antigos (R2/L2)
Arara do Iriri	ari03	hp27/hp21	
	ari05	hp27/hp21	
	ari09	hp21/hp21	
	ari17	hp21/hp21	
	ari22	hp21/hp21	
	ari24	hp21/hp21	
	ari26	hp21/hp21	
	ari27	hp21/hp21	
	ari28	hp27/hp21	
	ari32	hp21/hp27	
	ari35	hp21/ hp21	
	ari36	hp21/ hp21	
	ari37	hp21/ hp21	
Arara do Laranjal	ara08	hp27/ hp27	
	arl03	hp27/ hp27	
	arl06	hp21/hp27	
	arl14	hp6/hp27	
	arl19	hp06/hp06	
	arl20	hp13/hp27	
	arl21	hp06/ hp27	hp27/hp27
	arl22	hp27/ hp27	hp27/hp27
	arl25	hp06/hp21	
	arl26	hp27/hp27	
	arl28	hp06/hp13	
	arl29	hp06/hp27	
	arl30	hp06/hp13	
Kayapó	k01	hp06/hp13	hp13/hp13
	k05	hp06/hp13	
	k06	hp06/ hp27	hp27/ hp27
	k07	hp06/hp13	
	k09	hp06/ hp06	hp06/ hp06

	k17	hp02/hp06	
	k18	hp06/hp13	hp13/hp13
	k37	hp06/hp27	hp27/hp27
	k43	hp06/hp13	
	k50	hp13/ hp21	hp13/hp13
	K51	hp21/hp21	
	k70	hp06/ hp27	hp27/hp27
Xikrin	xik04	hp06/ hp06	hp06/ hp06
	xik10	hp06/hp21	
	xik15	hp06/hp13	hp13/ hp13
	xik29	hp06/hp21	
	xik30	hp06/hp06	
	xik32	hp06/ hp06	hp06/ hp06
	xik35	hp21/ hp21	hp21/ hp21
	xik36	hp06/ hp06	hp06/ hp06
	xik41	hp06/hp27	
	xik68	hp27/hp27	
	xik70	hp06/hp21	
	xik95	hp27/ hp27	hp27/ hp27
	xik98	hp27/ hp27	hp27/ hp27
Waiãpi	wpi08	hp13/hp21	hp13/hp13
	wpi11	hp02/hp06	
	wpi17	hp13/hp27	hp13/hp27
	wpi25	hp02/hp06	
	wpi40	hp02/hp06	
	wpi47	hp02/hp06	
Mapuera-Katuena	mapktn28	hp06/hp13	hp13/hp13
	mapktn209	hp21/hp21	
Parakanã	prk02	hp06/hp13	hp13/hp13
	prk09b	hp02/hp21	
	prk27	hp02/hp21	hp21/hp21

Teste do equilíbrio de Hardy-Weinberg para a amostra definitiva

O teste do equilíbrio de Hardy-Weinberg, baseado nas frequências haplotípicas da amostra total e para a amostra sem a população Arara do Iriri, indicou um desvio significativo das frequências genotípicas observadas em direção a um excesso de homozigotos (Tabela 11). O excesso de homozigotos observado neste caso é um exemplo empírico do Efeito de Wahlund (Hartl e Clark, 1997).

O teste aplicado para cada população, não indicou desvios significativos das frequências genotípicas esperadas excetuando-se as populações Xikrin e Waiãpi. Estas populações mostraram desvios significativos tendendo a um excesso de homozigotos e de heterozigotos respectivamente (Tabela 11). Entretanto esses desvios observados

deixam de ser significativos após a aplicação da correção de Bonferroni para múltiplos testes.

Tabela 11: Resultados dos testes do equilíbrio de Hardy-Weinberg, baseados nas frequências haplotípicas, de cada população, da amostra total e da amostra total sem a população Arara do Iriri. * Número de genótipos. ** desvios significativos ($p < 0,05$) após correção de Bonferroni.

População	Tamanho da amostra*	Taxa de heterozigose Observada	Taxa de heterozigose esperada	Valor de P
Arara do Iriri	13	0,307	0,270	1
Arara do Laranjal	13	0,615	0,661	0,868
Parakanã	3	0,666	0,800	0,596
Mapuera-	2	0,500	0,833	0,335
Katuena				
Kayapó	12	0,833	0,724	0,130
Xikrin	13	0,384	0,664	0,031
Waiãpi	6	1,000	0,803	0,021
ameríndios	62	0,580	0,752	0,000**
ameríndios sem Arara do Iriri	49	0,653	0,746	0,005**

Diferenciação genética interpopulacional (F_{st})

Comparações intercontinentais

Nós calculamos os valores de F_{st} entre todos os pares de grandes grupos geográficos. A média do valor de F_{st} encontrada para comparações envolvendo ameríndios, asiáticos e indianos foi igual a 0,05 sendo que o valor médio de p foi menor do que 0,01 em todos os casos.

O valor de F_{st} entre os ameríndios e a amostra africana foi significativo ($p < 0,01$) e igual a 0,07. Por outro lado, o valor de F_{st} entre a amostra de indígenas

amazônicos e a amostra européia não foi significativamente diferente de zero (Tabela 12a).

Tabela 12a: Valores de F_{st} haplotípicos. Os respectivos valores de p estão entre parênteses. Entre colchetes estão os tamanhos amostrais em número de cromossomos.

	ameríndios sem Arara do Iri [98]	asiáticos	européus	indianos
asiáticos [54]	0,05 (0,00)	0		
européus [48]	0,01 (0,14)	0,02 (0,12)	0	
indianos [60]	0,04 (0,00)	0,00 (0,58)	0,00 (0,42)	0
africanos [62]	0,07 (0,00)	0,08 (0,00)	0,02 (0,03)	0,03 (0,00)

Usando a amostra total de ameríndios (contendo os 13 indivíduos Arara do Iri) todas as comparações entre pares de amostras continentais resultaram em valores de F_{st} statisticamente significativas ($p < 0,05$) e com média igual a 0,06 (Tabela 12b).

Tabela 12b: Valores de F_{st} haplotípicos. Os respectivos valores de p estão entre parênteses. Entre colchetes estão os tamanhos amostrais em número de cromossomos.

ameríndios [124]	
asiáticos [54]	0,08 (0,00)
européus [48]	0,03 (0,01)
indianos [60]	0,07 (0,00)
africanos [62]	0,06 (0,00)

A comparação entre ameríndios e asiáticos é particularmente interessante porque vários estudos genéticos apontam para uma ancestralidade asiática para os nativos americanos. O valor de diferenciação genética encontrado para comparação desses dois grupos populacionais foi igual a 0,05 ($F_{st} = 0,05$).

Uma vez que esse valor é aparentemente baixo em relação a valores publicados para outras bases de dados, decidimos investigar de modo mais sistemático se os valores de F_{st} entre ameríndios e asiáticos tipicamente são maiores do que o valor por nós encontrado para a região cis-reguladora do *CCR5*. Para isso, utilizamos uma base de dados de 783 microssatélites descrita na seção “Materiais e Métodos”. Uma vez que o valor do F_{st} é sensível ao número de indivíduos da amostra populacional e de sua procedência geográfica, realizamos 1000 re-amostragens dos indivíduos da base de dados de microssatélites para gerar amostras de mesmo tamanho daquelas analisadas para a região cis-reguladora do *CCR5*. Além disso, incluímos nessas amostras apenas indivíduos de populações indígenas da América do Sul. Dessa forma geramos uma distribuição de 1000 valores de F_{st} calculados entre asiáticos e nativos sul-americanos usando marcadores neutros.

Figura 6: Distribuição de 1000 valores de F_{st} médios entre amostras de Asiáticos e nativos amazônicos.

O valor de F_{st} observado para a região cis-reguladora do *CCR5* entre asiáticos e ameríndios ($F_{st}=0,05$) foi menor que todos os valores da distribuição que variou de aproximadamente 0,081 a 0,096.

Comparações intracontinentais

A seguir, analisamos a diferenciação entre populações ameríndias (Tabela 13). De todas as comparações realizadas apenas a minoria apresentou valores de F_{st} estatisticamente significativos ($p < 0,05$). A população Arara do Iriri apresentou diferenciação genética significativa em todas as comparações. Considerando apenas as amostras contendo mais de 20 cromossomos, as populações Arara do Laranjal e Kayapó mostraram valores de F_{st} significativos.

Tabela 13: Valores de F_{st} haplotípicos. Os respectivos valores de p estão entre parênteses. * Valores de p não significativos.

População	Arara do laranjal (26)	Kayapó	Xikrin	Mapuera Katuena	Parakanã	Waiãpi
Kayapó (24)	0,085 (0,032)	0				
Xikrin (26)	0,042 (0,109)*	0,012 (0,251)*	0			
Mapuera Katuena (4)	0,157 (0,102)*	-0,009 (0,493)*	0,053 (0,226)*	0		
Parakanã (6)	0,198 (0,013)	0,069 (0,117)*	0,116 (0,083)*	-0,223 (0,999)*	0	
Waiãpi (12)	0,125 (0,022)	0,011 (0,348)*	0,075 (0,066)*	0,021 (0,428)*	0,021 (0,371)*	0
Arara do Iriri (26)	0,456 (0,000)	0,436 (0,000)	0,412 (0,000)	0,207 (0,086)*	0,175 (0,025)	0,474 (0,000)

Rede de haplótipos

A rede de haplótipos permite a fácil visualização das mutações que separam os haplótipos e do nível de compartilhamento de haplótipos entre os grupos analisados.

Construímos uma rede de haplótipos para a amostra mundial e outra somente para a

amostra ameríndia (Figura 7). Nessas figuras podemos verificar que os alelos mais freqüentes encontrados em ameríndios pertencem a “clusters” filogenéticos distintos, assim como o observado nas demais populações estudadas (africanos, asiáticos não-indianos, europeus e indianos). Uma característica marcante de ameríndios é a total ausência das mutações raras encontradas nas demais populações, principalmente nas africanas, indicando um provável efeito de sucessivos gargalos populacionais associados aos processos de ocupação dos continentes até a chegada à América.

Figura 7: a) Rede de haplótipos construída através do algoritmo Median-Join (Bandelt, Forster *et al.*, 1999) contendo os haplótipos encontrados nas populações africanas, asiáticas não-indianas, européias, indianas (Bamshad, Mummidi *et al.*, 2002) e ameríndias. O tamanho de cada nó é proporcional à frequência do haplótipo na população. Cada cor representa uma população estudada de acordo com a legenda da figura. Os números em vermelho representam os *SNPs* que diferenciam os haplótipos. Cluster 1 e 2 contêm haplótipos associados à expressão fraca e forte do gene *CCR5*, respectivamente. b) Rede de haplótipos construída através do algoritmo Median-Join (Bandelt, Forster *et al.*, 1999) contendo os haplótipos encontrados nas populações ameríndias. Cada cor representa uma população estudada de acordo com a legenda da figura

Análise de polimorfismo e testes de neutralidade

Para testar hipóteses sobre os efeitos da seleção natural e da história demográfica sobre populações ameríndias nós calculamos algumas estatísticas-sumário de polimorfismo e o D de Tajima das populações do Velho Mundo (Tabela 14) assim como para a amostra de indígenas da região amazônica brasileira (Tabela 15).

Tabela 14: Estatísticas-sumário de polimorfismo para as populações do Velho Mundo e para nativos americanos da região amazônica

** Não-Africanos incluindo ameríndios. n: número de cromossomos; *H*: diversidade gênica; K: número de haplótipos; S: número de sítios polimórficos; D.P.:desvio padrão; entre parênteses: significância estatística do D de Tajima estimada por 1000 simulações coalescentes para uma população neutra em condições de equilíbrio.

População	n	H ± D.P	K	S	θ_s	$\pi \pm D.P., \%$	D deTajima
Africanos	62	0,8503 +/- 0,0187	11	11	2,34 +/- 0,91	0,26 ± 0,015	0,16 (0,40)
Não-Africanos**	286	0,7648 +/- 0,0125	8	6	0,96 +/- 0,43	0,26 ± 0,004	3,05 (<0,01)
Não-Africanos	162	0,7322 +/- 0,0210	8	6	1,06 +/- 0,48	0,24 ± 0,005	2,39 (0,02)
Asiáticos	54	0,6625 +/- 0,0458	5	5	1,09 +/- 0,55	0,24 ± 0,012	2,46 (0,01)
Europeus	48	0,8014 +/- 0,0299	7	6	1,35 +/- 0,64	0,25 ± 0,012	1,96 (0,03)
Indianos	60	0,7401 +/- 0,0327	6	5	1,07 +/- 0,54	0,23 ± 0,010	2,45 (0,04)
Ameríndios	124	0,7522 +/- 0,0149	5	6	1,11 +/- 0,51	0,27 ± 0,009	2,82 (<0,01)
Ameríndios sem Arara do Iriri	98	0,7465 +/- 0,0232	5	6	1,16 +/- 0,53	0,28 ± 0,006	2,80 (<0,01)

Tabela 15: Estatísticas-sumário de polimorfismo para as populações ameríndias analisadas

(ver legenda da Tabela 14 para descrição das siglas).

População	n	H± D.P	K	S	θ_s	$\pi \pm D.P., \%$	D de Tajima
Arara do Iriri	26	0,270 +/- 0,099	2	2	0,52 +/- 0,38	0,058 ± 0,021	0,07 (0,36)
Arara do Laranjal	26	0,661 +/- 0,064	4	6	1,57 +/- 0,78	0,263 ± 0,032	1,63 (0,05)
Kayapó	24	0,724 +/- 0,066	5	6	1,60 +/- 0,80	0,274 ± 0,018	1,74 (0,04)
Xikrin	26	0,664 +/- 0,062	4	6	1,57 +/- 0,78	0,303 ± 0,018	2,33 (0,01)
Kayapó+Xikrin	50	0,698 +/- 0,047	5	6	1,34 +/- 0,64	0,28 ± 0,011	2,47 (0,01)
Mapuera Katuena	4	0,833 +/- 0,222	3	5	2,72 +/- 1,78	0,287 ± 0,121	-0,21 (0,43)
Parakanã	6	0,800 +/- 0,172	4	5	2,18 +/- 1,34	0,251 ± 0,077	0,36 (0,33)
Waiãpi	12	0,803 +/- 0,077	5	6	1,98 +/- 1,06	0,245 ± 0,037	0,55 (0,28)

O valor de D encontrado para a região cis-reguladora do *CCR5* em populações indígenas da região amazônica ($D = 2,82$) foi positivo e estatisticamente significativo. Esse valor é compatível com efeitos da história demográfica ou pode indicar a ação de seleção balanceadora. Para levar em consideração os efeitos da história demográfica, nós comparamos o valor de D observado para a região cis-reguladora do *CCR5* com uma distribuição de valores de D calculados a partir de dados empíricos e simulados sob a hipótese nula de ausência de seleção natural. Para a primeira abordagem, utilizamos valores de D calculados a partir de 18 locos neutros seqüenciados em uma amostra de 14 indivíduos nativos sul-americanos oriundos de populações do Paraguai, Peru, e Amazônia Brasileira.

A Figura 8 contém um histograma dos valores de D de Tajima encontrados para esses locos (barras sombreadas). Nenhum dos 18 valores de D encontrados para os locos neutros foi maior que o observado para a região cis-reguladora do *CCR5* ($D=2,82$) nem atingiram significância estatística.

Para verificarmos se este resultado poderia ser causado pela grande diferença no tamanho das amostras (62 indivíduos para a região cis-reguladora do *CCR5* e 14 para os

locos neutros), calculamos 1000 valores de D de Tajima para a região cis-reguladora do *CCR5*, obtidos de 1000 amostras sorteadas, cujos tamanhos foram iguais a 14 indivíduos cada. A média dos 1000 valores de D de Tajima foi superior a todos os valores observados para os locos neutros ($D_{\text{médio ccr5}} = 1,98$). Na Figura 8 a linha tracejada representa a distribuição desses 1000 valores de D de Tajima.

Figura 8: Valores de D de Tajima para 18 locos neutros de uma amostra de nativos da América do Sul (Fagundes, Ray *et al.*, 2007). Linha tracejada representa a distribuição dos valores de D de Tajima para a região cis-reguladora do *CCR5* de oriundos de 1000 amostras de 14 indivíduos indígenas da região amazônica.

Para a realização da segunda abordagem, usamos simulação coalescente contendo parâmetros histórico-demográficos que foram escolhidos, através de estatística Bayesiana, por apresentarem maior nível de encaixe com a diversidade genética observada para 50 locos neutros seqüenciados em africanos, asiáticos e ameríndios.

O histograma contendo os valores de diversidade nucleotídica (π) em porcentagem, obtidos para as 2000 amostras simuladas usando o modelo demográfico está na Figura 9. O valor de π observado para a região cis-reguladora do *CCR5* ($\pi=0,0027$ ou 0,27%) para as populações indígenas da região amazônica foi maior que 99,8% dos valores obtidos por simulação, ou seja, a probabilidade de se observar um valor de π igual ao observado para a região cis-reguladora do *CCR5* em populações nativas da região amazônica assumindo o modelo demográfico neutro utilizado é igual a 0,002.

Figura 9: Distribuição dos 2000 valores de diversidade nucleotídica (π) em porcentagem para amostras de nativos americanos simuladas usando o modelo demográfico *AFREG* (Fagundes, Ray *et al.*, 2007).

Nós utilizamos as seqüências simuladas sob o modelo demográfico *AFREG* para calcular a distribuição de uma segunda estatística, o D de Tajima (Figura 10). Para populações evoluindo sem seleção natural e sem variações em seu tamanho efetivo é esperado que o D de Tajima igual seja igual a zero. A distribuição obtida apresenta um pequeno deslocamento em direção aos valores positivos. Esta característica pode estar associada aos gargalos populacionais simulados no cenário demográfico.

Figura 10: Distribuição dos 2000 valores de D de Tajima para amostras de nativos americanos simuladas usando o modelo demográfico *AFREG*.

O resultado indica que o valor de D de Tajima observado para a região cis-reguladora do *CCR5* em populações ameríndias da Amazônia ($D=2,82$) foi maior que 99,6% dos valores simulados. Esse resultado indica que o modelo neutro utilizado não é

capaz de explicar o valor de D de Tajima observado para a amostra ameríndia, mesmo levando em conta um cenário demográfico de não-equilíbrio.

Teste de neutralidade Ewens-Watterson

O teste de Ewens-Watterson aplicado à amostra ameríndia total indicou um desvio significativo em direção a um excesso na taxa heterozigose da amostra em relação ao esperado para uma população neutra e em condições de equilíbrio. Esse resultado corrobora a assinatura de seleção balanceadora na região cis-reguladora do *CCR5*. A Tabela 16 contém os resultados dos testes aplicados às amostras total e sem a população Arara do Iriri.

Tabela 16: Resultados do teste de neutralidade de Ewens-Watterson aplicado à amostra total de ameríndios e à amostra sem o grupo Arara do Iriri. n : número de cromossomos, K : número de haplótipos, Fe : soma do quadrado das frequências alélicas.

	Ameríndios	Ameríndios sem Arara do Iriri
n	124	98
K	5	5
Fe observado	0.25390	0.26114
Fe esperado	0.50512	0.48870
p	0.00970	0.02300

O resultado do teste de Ewens-Watterson aplicado a cada população ameríndia analisada apontou valores de Fe não significativos, ou seja, a taxa de homozigose observada não foi significativamente diferente dos valores encontrados para 1000 amostras simuladas evoluindo de acordo com o modelo neutro de infinitos alelos.

Entretanto, a tendência do desvio foi para a maioria das populações (Arara do Laranjal, Kayapó, Xikrin, Waiãpi), de um excesso na taxa de heterozigose observada (Tabela 17).

Tabela 17: Resultados do teste de neutralidade de Ewens-Watterson aplicado às populações ameríndias analisadas. n: número de cromossomos, K: número de haplótipos, F: soma do quadrado das frequências alélicas.

	Arara do Iri	Arara do Laranjal*	Kayapó*	Xikrin*	Waiãpi*	Mapuera- Katuenta	Parakanã
n	26	26	24	26	12	4	6
K	2	4	5	4	5	3	4
Fe	0,73964	0,36391	0,30556	0,36095	0,26389	0,37500	0,33333
observado							
Fe esperado	0,73550	0,46146	0,37865	0,46440	0,30110	0,37500	0,30300
p	0,52100	0,26600	0,27700	0,23900	0,37000	1,00000	1,00000

* Tribos com tendência ao excesso na taxa de heterozigose observada em relação ao esperado pelo modelo neutro de infinitos alelos

Sítios polimórficos em ESEs e ISEs

Nós investigamos se havia alguma correlação entre as frequências alélicas dos *SNPs* da região cis-reguladora do *CCR5* e suas possíveis funções regulatórias. Através do programa *Rescue-ESE*, observamos que dentre os nove sítios polimórficos exônicos (localizados em exons da região cis-reguladora do *CCR5*) descritos em Mummidi *et al.*, (2000) (-2733, -2717, -2135, -2132, -2086, -2076, -2044, -2006, -1990) (Figura 5) todos aqueles que apresentaram alelos em frequências intermediárias, no estudo populacional em escala mundial, apresentaram micro-haplótipos que correspondiam às sequências

caracterizadas como ativadores de processamento exônico (*ESEs*) (Blencowe, 2000). Dos sítios polimórficos exônicos que apresentaram um dos dois alelos em baixa frequência ($<0,01$) nenhum apresentou micro-haplótipos localizados em *ESE* (Tabela 18).

Tabela 18: Quantidade de *ESEs* encontrados nos hexâmeros candidatos selecionados a partir de cada SNP da região exônica do promotor do gene CCR5. Os SNPs estão agrupados em duas classes de frequência alelica: raros e não-raros segundo dados de Bamshad *et al.*, (2002)

	SNP exônicos	ESEs	Hexâmeros candidatos
não –raros	-2733	6	12
(freq. > 0.05)	-2135, -2132*	3	22
	-2086	9	12
raros	-2757	0	12
(freq. < 0,01)	-2076	0	12
	-2044	0	12
	-2006	0	12
	-1990	0	12

* SNP analisados em conjunto por não distarem mais de 7 p.b. entre si.

Para confirmar a significância estatística destas observações foi construída uma tabela de contingência (Tabela 19) a partir dos dados da Tabela 18. Em seguida, foi realizado o teste do qui-quadrado com um grau de liberdade e o teste exato de Fisher.

Tabela 19: Tabela de contingência 2x2 contendo a quantidade de alelos raros e não-raros presentes nas classes funcionais *ESE* e *não-ESE*.

	ESEs	não-ESEs
raros (freq. <0,01)	0	60
não -raros (freq. > 0,01)	18	28

Os resultados dos testes do qui-quadrado e o do teste exato de Fisher rejeitam a hipótese nula de independência destas classes ($p < 0,0001$), ou seja, existe uma forte associação entre *ESEs* e SNPs de frequência intermediária.

Os *introns* também possuem regiões envolvidas no processamento alternativo ou constitutivo do transcrito primário (pré-*RNA*m). Yeo *et al.*, (2004) identificaram através de Bioinformática que, em humanos e camundongos, as trincas de guanina (GGG) e hexâmeros ricos em citosina (C) funcionam como prováveis ativadores de processamento intrônicos (*ISEs-Intronic splicing enhancers*). Utilizando o programa *Acscan 2*, verificamos quais dos *SNPs* da região cis-reguladora do *CCR5* estavam em prováveis regiões *ISEs*. A Tabela 20 apresenta a quantidade de *ISEs* encontrados nos hexâmeros-candidatos selecionados a partir dos sítios polimórficos intrônicos, raros e não-raros da região cis-reguladora *CCR5*.

Tabela 20: Quantidade de *ISEs* encontrados nos hexâmeros candidatos selecionados a partir de cada SNP da região intrônica do promotor do gene *CCR5*. Os SNPs estão agrupados em duas classes de frequência alelica: raros e não-raros segundo dados de Bamshad *et al.*, (2002).

	SNP intrônicos	ISEs	Hexâmeros candidatos
não -raros	-2554,-2553*	1	19
(freq. > 0,01)	-2459	6	12
	-1840,1837,1835*	0	16
raros	-2668	1	12
(freq. <0,01)	-2585	3	12
	-2562	1	12
	-2523	5	12
	-2470	4	12
	-2401	0	12
	-2388	0	12
	-2381,2377	0	22
	-2352	2	12
	-2328	1	12
	-2268	0	12
	-2238	1	12
	-2216	0	12
	-1882	0	12
	-1872,-1871,-1867* [@]	7	30

* SNP analisados em conjunto por não distarem mais de 7 p.b. entre sí. @ Não foi possível explorar todos os hexâmeros candidatos por falta de sequência a 3' do último SNP (-1867).

A partir dos dados da Tabela 21, foram realizados os testes do qui-quadrado com um grau de liberdade e o teste exato de Fisher.

Tabela 21: Tabela de contingência 2x2 contendo a quantidade de alelos raros e não-raros presentes nas classes funcionais *ISE* e *não-ISE*.

	ISE	não-ISE
raro (freq. <0,01)	25	183
não raro (freq. > 0,01)	7	40

Diferentemente do que foi observado para os *SNPs* de *exons*, os resultados dos testes do qui-quadrado e exato de Fisher, sugerem independência entre a frequência alélica dos *SNPs* e suas localizações funcionais definidas por *ISEs* ou *não-ISEs* ($p > 0,6$). Portanto, nesse caso não descartamos a hipótese nula.

Discussão

Nesta discussão serão abordados três temas: aspectos metodológicos, inferências sobre efeitos demográficos na região cis-reguladora do *CCR5* de ameríndios e inferências sobre seleção natural na região cis-reguladora do *CCR5*.

Aspectos metodológicos

Desvios do equilíbrio de Hardy-Weinberg (HW) como ferramenta para detecção de problemas técnicos

O teste do equilíbrio de *HW* foi importante para identificarmos uma falha metodológica durante o primeiro ano de execução deste trabalho. O marcante excesso de homozigotos encontrado na amostra foi detectado através do método de Guo e Thompson

(1992) e nos sugeriu uma possível inadequação dos iniciadores de PCR descritos em Bamshad, Mummidi *et al.*, (2002) para a utilização em amostra de ameríndios. Esse problema foi resolvido com o desenho de um novo iniciador de amplificação e dessa forma foi possível prosseguir com o projeto e realizar a inferência dos haplótipos da amostra de ameríndios.

Caracterização dos haplótipos da região cis-reguladora do CCR5 seguindo a nomenclatura proposta por Mummidi et al., (2000).

A inferência de haplótipos da região cis-reguladora do *CCR5* para as amostras africana, asiática não-indiana, européia e indiana, utilizando os 33 sítios polimórficos encontrados por Mummidi *et al.*, (2000) conjuntamente com os quatro sítios polimórficos adicionais descritos em Bamshad, Mummidi *et al.*, (2002) nos permitiu a utilização de uma nomenclatura comum para os haplótipos em todas as amostras de populações analisadas, incluindo as ameríndias aqui descritas. O uso de uma nomenclatura comum permitiu a comparação dos nossos dados com as seqüências de haplótipos descritas por Mummidi *et al.*, (2000) e disponíveis no *GenBank* (<http://www.ncbi.nlm.nih.gov/>). O alto valor de probabilidade associado a cada genótipo inferido através do programa *Phase* é decorrência do alto nível de desequilíbrio de ligação existente na região cis-reguladora do *CCR5* (Clark e Dean, 2004) e fornece bom suporte para as análises baseadas em dados haplotípicos que foram aplicadas neste trabalho.

Escolha de amostras e marcadores para as análises comparativas

Uma limitação metodológica enfrentada neste projeto foi a ausência de marcadores neutros tipados nas mesmas amostras que foram analisadas para a região cis-reguladora do

CCR5. Porém, para controlar os efeitos da influência da demografia sobre a diferenciação genética entre ameríndios e asiáticos, escolhemos amostras e marcadores seguindo critérios para maximizar a legitimidade da comparação entre amostras sequenciadas para a região cis-reguladora do *CCR5* e aquelas estudadas para outros locos. Em primeiro lugar, reamostramos os dados genéticos das populações usadas nas meta-análises, para que as amostras tivessem tamanhos iguais. Em segundo lugar, escolhemos uma quantidade de populações-controle, representativas de cada região geográfica (Ásia e América), similares às utilizadas para a região cis-reguladora do *CCR5*. Finalmente, escolhemos populações-controle geograficamente próximas às usadas no estudo da região cis-reguladora do *CCR5*.

Uma outra diferença importante entre os dados gerados neste estudo e aqueles usados como controle, na análise envolvendo os valores de F_{st} , diz respeito à diferença no tipo de marcador usado. Enquanto nossas análises são baseadas em re-sequenciamento, os dados de Rosenberg *et al.*, (2002) são genótipos de microssatélites, que possuem um mecanismo mutacional distinto.

Entretanto, Barbujani *et al.*, (1997) analisando 30 microssatélites e 70 fragmentos de restrição polimórficos (*Restriction Fragment Length Polymorphisms – RFLPs*) em 16 populações humanas, encontraram valores de diferenciação genética interpopulacional muito semelhantes para esses dois tipos de marcadores e que ainda corroboravam os resultados obtidos pelos estudos clássicos de polimorfismos de proteínas.

Além disso, o próprio estudo de Rosenberg *et al.*, (2002) mostrou que os resultados obtidos por eles, analisando a diferenciação genética interpopulacional, através de microssatélites são similares às aqueles obtidos por RFLPs. Portanto, esses achados indicam que, em escala comparativa intercontinental, os padrões de diferenciação medidos por F_{st}

não apresentam diferenças marcantes em função do tipo de marcador utilizado, legitimando nossa comparação.

Inferências sobre efeitos demográficos na região cis-reguladora do CCR5 de ameríndios

Efeito de Wahlund

Para a amostra total de ameríndios, foi observado desvio significativo nas frequências genotípicas esperadas pelo princípio de *HW* (Tabela 11). O excesso de homozigotos observado para a amostra total é um exemplo empírico do efeito de Wahlund (Nei, 1987). Para as populações com tamanhos amostrais maiores que 20 cromossomos não foram encontrados desvios significativos, com exceção de Xikrin, que apresentou desvio marginalmente significativo ($p=0,03$) e que após correção de Bonferroni não apresentou significância.

Fst significativo entre Arara e Kayapó

Embora a maioria das comparações entre as populações ameríndias analisadas resultarem em valores não significativos de *Fst*, a comparação entre as populações Arara do Laranjal e Kayapó resultou em valor de *Fst* significativo ($Fst = 0,08$, $p=0,03$). Esse resultado está de acordo com o trabalho de Ribeiro-dos-Santos *et al.*, (2001) que, analisando a região hipervariável da região controladora do genoma mitocondrial (*D loop*) em sete tribos amazônicas, encontraram entre Arara e Kayapó, uma das mais altas diferenciações genéticas em relação à todos os pares de populações comparados. Resultado similar foi encontrado por Zago *et al.*, (1996), analisando quatro *VNTRs* (*variable number of tandem repeats*) autossômicos, um ligado ao cromossomo X e um ligado ao Y, em cinco tribos

amazônicas. Além disso, Zago *et al.*, (1996), assim como neste estudo (Tabela 15), encontraram o menor valor de diversidade genética para o grupo Arara quando comparado com outras tribos amazônicas.

O mesmo foi encontrado por Santos (1999) analisando 12 microssatélites e 3 minissatélites em sete populações nativas da região amazônica. Esses resultados sugerem que processos demográficos que atuaram na população Arara do Laranjal foram fortes o suficiente para influenciar a diversidade de um loco com forte assinatura de seleção balanceadora, como é o caso da região cis-reguladora do *CCR5*.

Possível miscigenação entre Waiãpi e afro-descendentes

A população Waiãpi apresentou o maior número de haplótipos, o maior valor de diversidade genética (H), o maior valor de θ_s e o menor valor de D de Tajima dentre as populações ameríndias com amostras com 12 ou mais cromossomos (Tabela 15). Todas essas características se assemelham às apresentadas pelos Africanos quando comparados com as demais populações do mundo (Tabela 14). Além disso, a população Waiãpi apresentou o haplótipo 2 em frequência relativamente alta (aproximadamente 33%, Tabela 9). Este haplótipo foi encontrado em todas as populações do Velho Mundo, entretanto os africanos foram os que apresentaram esse haplótipo em frequência mais alta (aproximadamente 25%, Tabela 7). Uma possível hipótese para explicar esses resultados seria que a população Waiãpi teria um certo grau de miscigenação com africanos ou afro-descendentes. O trabalho de Franco *et al.*, (1997), estudando o gene do *Fator IX*, e de Mousinho-Ribeiro, Pante-De-Souza *et al.*, (2003), analisando o gene da β -globina em populações de nativos da Amazônia, encontraram um número relativamente alto de

haplótipos nas tribos Waiãpi e Urubu-Kaapor respectivamente. A população Urubu-Kaapor e a população Waiãpi são do grupo lingüístico Tupi e apresentam similaridade genética relativamente alta (Black, Santos *et al.*, 1988). Os autores dos trabalhos citados acima, atribuíram essa elevada diversidade genética dos Waiãpi, à possíveis eventos de miscigenação entre este grupo ameríndio e afro-descendentes da Guiana Francesa durante o século 19. Por outro lado, marcadores dos cromossomos mitocondrial e Y não sugerem essa possível miscigenação entre os Waiãpi e os afro-descendentes (Mendes-Junior, C. T comunicação pessoal). Como os marcadores dos cromossomos mitocondrial e Y têm tamanhos efetivos menores que os dos cromossomos autossômicos eles estão mais sujeitos aos efeitos da deriva genética e apresentam tempos de coalescência menores que os de marcadores dos cromossomos autossômicos. Sendo assim, uma possível explicação para a ausência de evidência de miscigenação entre os Waiãpi e os afro-descendentes, apontada por marcadores dos cromossomos mitocondrial e Y, seria que o evento de miscigenação não foi suficientemente recente para ser detectado por marcadores de tamanhos efetivos relativamente pequenos.

Arara do Iriri apresentou baixo nível de polimorfismo

O grupo Arara do Iriri apresentou os menores valores para as estatísticas-sumário de polimorfismo, refletindo o alto nível de endogamia desse grupo, que foi fundado por dois irmãos. Essa é a primeira descrição de uma população que apresentou apenas haplótipos de um mesmo cluster filogenético (cluster 2). É evidente a alta discrepância do valor de D de Tajima deste grupo em relação aos outros (Tabela 15).

Efeitos demográficos são detectáveis na região cis-reguladora do CCR5 de ameríndios

Os resultados discutidos acima sugerem que ao se analisar as populações individualmente, é possível detectar uma correlação entre o padrão de diversidade apresentado por vários marcadores genéticos e a região cis-reguladora do *CCR5*.

Portanto, efeitos demográficos colaboraram para o padrão de diversidade genética observado na região cis-reguladora do *CCR5* em populações ameríndias apesar do forte sinal de seleção balanceadora observado para este loco.

Inferências sobre seleção natural na região cis-reguladora do CCR5

Evidências de seleção balanceadora na região cis-reguladora do CCR5 em humanos.

A diversidade nucleotídica (π) da região cis-reguladora do *CCR5* em humanos (0.0026) é aproximadamente duas vezes mais alta do que a média encontrada para os sítios 4 vezes degenerados do genoma humano (considerados neutros pois qualquer mudança de nucleotídeo nesses sítios não causa alteração de aminoácidos dos códons ao quais eles pertencem) (Li e Sadler, 1991) e duas vezes mais alto do que o encontrado para a região codificante do *CCR5* (0.0011) (Chakravarti, 1999).

O alto valor de π encontrado para a região cis-reguladora do *CCR5* não pode ser explicado por uma alta taxa de mutação pois a comparação desse loco entre humanos e chimpanzés revelou 0,7% de divergência enquanto que a média feita a partir de outros locos é 1,4% (Takahata e Satta, 1997; Bamshad, Mummidi *et al.*, 2002). Dentre todos os haplótipos da região cis-reguladora do *CCR5* que segregam na população humana, os mais

freqüentes são os que estão separados por um número relativamente alto de mutações (Figura 7). Este padrão de genealogia não é esperado pelo modelo estritamente neutro (Bamshad e Wooding, 2003), e é interpretado como manutenção de linhagens antigas. A manutenção de haplótipos ou haplogrupos, separados por um número relativamente alto de mutações e associados a padrões fenotípicos distintos em nível molecular, é uma característica de locos sob seleção balanceadora ou estabilizadora. A seleção balanceadora é capaz de preservar duas ou mais linhagens durante um período de tempo maior do que o esperado pelo modelo neutro e conseqüentemente, a genealogia de locos que estiveram sob este regime de seleção apresentam duas ou mais linhagens que são separadas por ramos relativamente longos (Richman, 2000; Bamshad e Wooding, 2003).

Sendo assim, foram encontrados valores de F_{st} muito baixos para todas as comparações entre amostras de africanos, asiáticos não-indianos, europeus e indianos (F_{st} médio = 0,03) e entre as populações não-africanas (F_{st} médio = 0,007). Outra característica que evidencia a ação de seleção natural na região cis-reguladora do *CCR5* é o valor aproximadamente igual de π entre africanos e asiáticos (padrão pouco comum para a maioria dos locos analisados) (Jorde, Watkins *et al.*, 2001).

Evidências de seleção balanceadora na região cis-reguladora do CCR5 em ameríndios

De modo semelhante ao o que foi previamente descrito para as populações asiática não-indiana, européia e indiana, os haplótipos mais freqüentes em ameríndios pertencem aos dois *clusters* filogenéticos associados à padrões de expressão gênica opostos (fraco/forte). Entretanto, enquanto que para a Ásia, Europa e Índia os dois haplogrupos

mais freqüentes foram HHC e HHE (Tabela 7), para a amostra de indígenas amazônicos os dois haplogrupos mais comuns foram HHC e HHF (Tabela 8). Considerando a amostra de ameríndios, sem Arara do Iriri, temos HHC e HHG como os haplogrupos mais comuns (Tabela 8, Figura 7). Ainda que a identidade dos alelos mais comuns varie entre os ameríndios e as demais populações, os nossos resultados indicam que para todas as populações do mundo os alelos mais comuns incluem um de cada padrão de expressão.

O nível de heterozigose na amostra de ameríndios, medido através da diversidade nucleotídica (π), foi o mais alto dentre todas as populações já estudadas para este loco ($\pi=0,0027$). Este valor de π foi aproximadamente 4 vezes mais alto do que o valor encontrado para a região hipervariável do cromossomo 16 analisado em populações indígenas da América do Sul e Central (Battilana, Cardoso-Silva *et al.*, 2007) e aproximadamente 3 vezes maior que o valor de π encontrado para o gene da interleucina-13 (*IL-13*) em amostra de nativos sul-americanos (Tarazona-Santos e Tishkoff, 2005).

A medida da taxa de heterozigose, baseada no número de sítios polimórficos (θ_s) para ameríndios, foi similar aos valores encontrados para as populações não-africanas, sendo que em africanos o valor foi praticamente o dobro.

Uma possível interpretação para esses resultados é que o alto valor de π reflete a manutenção, em ameríndios, de haplótipos com divergência relativamente antiga (Figura 7), provavelmente por seleção balanceadora. O valor de θ_s para ameríndios, por ser similar aos valores encontrados para as demais populações não africanas, sugere que o efeito de gargalo populacional associado à ocupação do continente americano não foi tão forte quando comparado ao efeito de gargalo populacional associado à saída da África.

Resultados similares foram encontrados para outros locos nucleares (Battilana, Fagundes *et al.*, 2006) e mitocondrial (Bonatto e Salzano, 1997b).

O valor de π mais alto do que o de θ_s , resulta num valor de D de Tajima significativamente positivo, e nesse caso significativo ($D=2,82$, $p<0,01$), consistente com um regime de seleção balanceadora.

Um estudo que analisou 313 genes em uma amostra de 82 indivíduos contendo em iguais proporções caucasianos, afro-americanos, asiáticos e “hispânicos”, verificou que mais de 90% deles apresentavam valores negativos de D de Tajima e este resultado foi interpretado como sendo uma evidência de expansão da população humana (Stephens, Schneider *et al.*, 2001). Estudo feito para um loco do cromossomo 16 em amostra contendo 90 indivíduos de diversas populações nativas da América do Sul encontrou valor do D de Tajima igual a $-2,08$ e sugeriu um cenário de expansão da população ameríndia iniciada no mínimo à 18 mil anos atrás (Battilana, Cardoso-Silva *et al.*, 2007).

Para avaliar a importância de eventos demográficos sobre os valores de D na região controladora do *CCR5* nós comparamos nossos resultados com aqueles obtidos para uma análise de 18 locos nucleares não-codificantes estudados por Fagundes *et al.*, (2007), em amostra de 14 indivíduos indígenas da América do Sul (Figura 8).

Os locos nucleares neutros apresentaram um valor médio de $D=0,51$, sendo que nenhum loco apresentou valores de D que desviavam significativamente dos valores de D esperados sob a hipótese de neutralidade e equilíbrio populacional. Os 18 locos apresentaram valores de D inferiores ao valor de D de Tajima observado para a região cis-reguladora da amostra total de ameríndios.

Para evitar que efeitos amostrais influenciassem no resultado, realizamos 1000 sorteios de amostras de populações ameríndias contendo os dados da região cis-reguladora do *CCR5*. Essas amostras foram de mesmo tamanho (14 indivíduos) que aquelas utilizadas por Fagundes *et al.*, (2007). Em seguida, calculamos o valor de D de Tajima para cada uma das 1000 réplicas.

A média do D de Tajima calculada para as 1000 amostras foi igual a 1,98. Esse valor foi maior do que todos os valores de D encontrados para os locos neutros. Dezesete dos 18 locos neutros apresentaram valores de D menores do que 99% dos valores encontrados para as 1000 réplicas. Dessa forma descartamos a hipótese de que o alto D de Tajima encontrado para a região cis-reguladora do *CCR5* de ameríndios, tenha sido maior que os valores de D observados para os locos neutros, apenas devido aos diferentes tamanhos amostrais.

As populações não altamente endogâmicas e com maiores amostras (Arara do Laranjal (2N=26), Kayapó (2N=24) e Xikrin (2N=26), também apresentaram valores de D de Tajima superiores aos valores dos 18 locos neutros. Portanto, o padrão de variação do *CCR5*, em ameríndios, apresenta um espectro de frequências alélicas (medido pelo D de Tajima) e níveis de polimorfismo (medido pelo π) consistentes com a ação de seleção balanceadora.

As comparações dessas estatísticas-sumário com valores obtidos por outros estudos que re-sequenciaram genes autossômicos em ameríndios, mostram que a região cis-reguladora do *CCR5* apresenta valores mais extremos. Esse resultado sugere que o padrão de variação observado para a região cis-reguladora do *CCR5* não é atribuível à sua história demográfica recente.

Apesar dos pequenos tamanhos amostrais, as populações Mapuera-Katuena e Parakanã, apresentaram haplótipos dos dois clusters filogenéticos (Figura 7). Esse fato, associado aos pequenos tamanhos amostrais, podem ter contribuído para o aumento do θ_s em relação ao valor de π e conseqüentemente causar uma redução dos valores de D de Tajima observados para esses grupos.

Além das abordagens empíricas descritas acima, nós também utilizamos simulações, para avaliar o efeito da história demográfica sobre a diversidade observada para a região cis-reguladora do *CCR5* em grupos indígenas da Amazônia.

As simulações foram feitas usando um cenário demográfico que continha gargalos associados à ocupação dos continentes e as posteriores expansões populacionais. Através delas, foram geradas as distribuições de estatísticas de diversidade genética (D e π) sob a hipótese nula de ausência de seleção natural. Essa análise revelou que a diversidade nucleotídica (π) da região cis-reguladora do *CCR5* foi maior que 99,8% dos valores simulados e, portanto, que a probabilidade de se observar tal valor no cenário demográfico escolhido é 0,002. O valor de D de Tajima observado também foi significativamente mais alto que a maioria dos valores simulados ($p=0,004$).

Portanto, em comparação com os valores esperados sob um cenário demográfico que leva em conta gargalos populacionais associados às saídas da África para a Ásia e da Ásia para a América, assim como expansões populacionais, o *CCR5* apresenta valores de π e D maiores do que o esperado.

Também realizamos um teste de neutralidade de Ewens-Watterson cuja hipótese nula é o modelo de infinitos alelos. Desvio significativo nas frequências dos alelos esperadas pelo modelo neutro de infinitos alelos, medido através do teste de Ewens-

Watterson indicaram um déficit na taxa de homozigose esperada, ou seja, um excesso na taxa de heterozigose esperada tanto para a amostra total de ameríndios como para a amostra sem Arara do Iriri (Tabela 16).

Garrigan e Hedrick, (2003), calcularam através de simulação, que uma população de tamanho efetivo de 1000 alelos, com 5 alelos em iguais frequências e evoluindo apenas por deriva genética, atingiria uma taxa de homozigose, que seria detectada pelo teste de *Ewens-Watterson* como consistente com modelo de evolução neutro, em aproximadamente 10.530 anos. Este resultado sugere que o tempo decorrido desde a ocupação do continente americano seria suficiente para resultar na perda de um sinal de seleção, caso esta tivesse deixado de atuar após a ocupação das Américas, e que portanto, a região cis-reguladora do *CCR5* das populações ameríndias estudadas, sofreu ou sofre ação da seleção balanceadora.

Por ser baseado na frequência de sítios polimórficos, acreditamos que o *D* de Tajima seja mais sensível à seleção natural ocorrida em um passado mais remoto. Partindo dessa premissa, espera-se que esse teste detecte desvios da neutralidade na população de estudo em consequência de padrões de variação herdados de uma população ancestral que sofreu seleção natural. Por outro lado, o teste de *EW*, por ser baseado em frequências haplotípicas, seria menos sensível a este tipo de evento. Dessa forma o resultado significativo do teste de *EW*, indicando que os alelos ocorrem em frequências intermediárias, corrobora a hipótese de seleção mais recente.

A rede de haplótipos construída para a região cis-reguladora do *CCR5* para todas as populações analisadas (Figura 7a) e somente para os nativos americanos (Figura 7b) mostra um alto grau de compartilhamento de haplótipos entre regiões e populações. Essa observação é também refletida em valores de *Fst* relativamente baixos. De modo similar

aos resultados de Bamshad, Mummidi *et al.*, (2002), encontramos valores muito baixos de F_{st} para as comparações entre as amostras de populações africanas, asiáticas não-indianas, européias e indianas, sendo que apenas as comparações envolvendo africanos resultaram em valores de F_{st} significativamente diferentes de zero (Tabela 12a). De todas as comparações feitas, as comparações entre ameríndios e as demais populações do Velho Mundo foram as que apresentaram valores de F_{st} mais altos, com exceção da comparação entre asiáticos e africanos. O valor de F_{st} entre asiáticos e ameríndios foi o mais alto entre populações não-africanas ($F_{st} = 0,05$). Para interpretar esse valor levando em consideração os efeitos da história demográfica experimentada por asiáticos e ameríndios, comparamos o valor de F_{st} observado, com a distribuição de 1000 valores de F_{st} calculados a partir 783 microssatélites em amostras de asiáticos e ameríndios de mesmo tamanho da utilizada para a amostra da região cis-reguladora do *CCR5*. O resultado indicou que o valor de F_{st} observado para a região cis-reguladora do *CCR5* foi mais baixo que todos os valores obtidos em todas as 1000 comparações realizadas (Figura 6). Tal resultado é consistente com a hipótese de seleção balanceadora na região cis-reguladora do *CCR5* em populações indígenas da Amazônia.

Num estudo de 377 marcadores microssatélites de 52 populações globalmente distribuídas, Rosenberg *et al.*, (2002) encontraram o maior valor de diferenciação genética para a região contendo 5 populações de indígenas americanos ($F_{st} = 0,116$), marcadamente maior do que F_{st} encontrado entre 18 populações do leste-asiático ($F_{st}=0,013$); 8 populações européias ($F_{st}=0,007$) e 6 populações africanas ($F_{st}=0,03$). A estimativa de F_{st} para ameríndios feita por Rosenberg *et al.*, (2002) foi baseada num estudo de cinco populações, sendo que duas apresentam evidências de conterem indivíduos aparentados (Surui, Karitiana). Esse fator causa um viés no sentido de aumentar o F_{st} entre populações

indígenas da América. Entretanto, Wang *et al.*, (*in press*) analisando 678 microssatélites em amostra de aproximadamente 20 indivíduos de 24 populações indígenas das Américas, corroborou esse resultado, pois o valor de F_{st} encontrado para as populações indígenas americanas continuou sendo o mais alto ($F_{st}=0,08$) quando comparado com as demais populações do mundo. Wang *et al.* encontraram uma heterogeneidade nos valores de F_{st} entre regiões da América do Sul. As sete populações amostradas localizadas na região não-andina (quatro tribos da região amazônica, duas da região centro-sul e uma do Paraguai), apresentaram valor de $F_{st} = 0,147$ aproximadamente três vezes maiores do que daquelas localizadas na região da Cordilheira dos Andes (nove tribos) ($F_{st} = 0,057$). Diante desse resultado, o baixo valor de F_{st} encontrado para a região cis-reguladora do *CCR5* de indígenas da Amazônia torna-se surpreendente e corrobora a existência de seleção balanceadora nesse grupo populacional.

O F_{st} médio para a região cis-reguladora do *CCR5* entre os pares de populações ameríndias com tamanhos amostrais maiores que 20 cromossomos foi igual a 0,05 e o F_{st} médio medido partir da diversidade nucleotídica para estas mesmas populações foi igual a 0,017.

Nenhum dos valores atingiu significância estatística em relação à hipótese nula de não haver estruturação genética entre as populações ($p_{médio} = 0,06$). Em média, o valor de F_{st} observado para a região cis-reguladora do *CCR5* é provavelmente o mais baixo já descrito para populações nativas da América do Sul.

É interessante notar que, até o momento, além deste trabalho, os outros dois trabalhos que encontraram valores de D de Tajima positivos e significativos para populações de indígenas sul americanos também analisaram *UTRs*, sendo elas pertencentes aos genes *IL-13* (Tarazona-Santos e Tishkoff, 2005) e *LDLR* (Fagundes, Salzano *et al.*,

2005; Tarazona-Santos e Tishkoff, 2005). Esses resultados são evidências da importância funcional regulatória das *UTRs* e apóiam a hipótese de que essas regiões são alvos da seleção natural. Hughes *et al.*, (2005) encontraram forte assinatura de seleção balanceadora em *UTRs*, analisando 4911 *SNPs* em 481 genes humanos.

Toda a argumentação apresentada até o momento sugere a existência de seleção balanceadora relativamente recente na região cis-reguladora do *CCR5* de populações indígenas da região amazônica. Entretanto, tal assinatura não exclui um cenário em que a população ancestral àquelas amostradas por nós, foi alvo de seleção natural e que as populações ameríndias apenas retiveram essa assinatura, ainda que não sujeitas à seleção natural. Portanto, podemos concluir que há seleção na história da região cis-reguladora do *CCR5*, e que essa seleção é detectável em ameríndios, mas se ela atuou na história recente dessas populações é uma questão que permanece em aberto. Acreditamos que através de simulações coalescentes ou simulações que considerem a ação da seleção natural é possível abordar sistematicamente essa questão.

Aspectos funcionais do polimorfismo na região cis-reguladora do CCR5, e sua relevância para hipóteses seletivas

Os resultados que indicam que a região cis-reguladora do *CCR5* está sob seleção natural são de natureza genética, baseados nos padrões de variação populacional. Entretanto, há pouca compreensão das bases funcionais de processo seletivo que explica a manutenção, de duas linhagens de haplótipos (haplogrupos) que conferem expressão gênica forte e fraca. As hipóteses mais exploradas envolvem a interação entre a espécie humana e certos patógenos ao longo da história evolutiva (Lim, Glass *et al.*, 2006) e de aspectos relacionados a doenças auto-imunes (Pralhad, Bohnsack *et al.*, 2006). Neste trabalho

demonstramos a existência de uma associação estatisticamente significativa ($p < 0,01$) entre os alelos mantidos em frequências intermediárias na população humana e regiões ativadoras de *splicing* localizadas em exons (*ESEs*). Todos os *ESEs* presentes na região cis-reguladora do CCR5 foram encontrados nos hexâmeros-candidatos gerados a partir dos SNPs balanceados, isso é, aqueles cujos alelos apresentam frequências intermediárias. Conseqüentemente, não foram encontrados *ESEs* nos hexâmeros-candidatos gerados a partir de SNPs com alelos raros. Sendo assim, uma forte associação ($p < 0,0001$) entre SNPs balanceados e regiões *ESEs* da região cis-reguladora do CCR5, foi descoberta.

Quando o mesmo tipo de análise foi aplicado aos introns da região cis-reguladora do CCR5 não foi observada associação estatisticamente significativa entre os *SNPs* balanceados e regiões ativadoras de *splicing* intrônicas (*ISEs*) ($p = 0,6$). Esses resultados corroboram a importância dos *ESEs* na regulação do *splicing* e sugerem haver uma importância funcional para o *splicing* alternativo que já foi observado na região cis-reguladora do CCR5. O resultado observado, de forte associação entre regiões *ESEs* e polimorfismos balanceados da região cis-reguladora do CCR5 pode estar relacionado com possíveis efeitos de variantes de *splicing* na manifestação dos fenótipos de expressão forte e fraca observados em escala global e também ao fato de que o CCR5 é um receptor expresso em diversos tecidos humanos (Rottman, Ganley *et al.*, 1997). Em recente trabalho, Dupanloup *et al.* (*artigo submetido à publicação*) explicita através de modelo matemático que variantes de *splicing* são mantidos por seleção natural em genes que são expressos em vários tecidos resultando em efeito balanceador intraespecífico.

Conclusão e Perspectivas Futuras

Neste trabalho estudamos o polimorfismo da região cis-reguladora do *CCR5* em populações nativas da região amazônica brasileira. Assim como para as amostras das populações asiática não-indiana, européia e indiana, os haplótipos mais frequentes em ameríndios são de clusters filogenéticos distintos e que estão associados a um padrão de expressão gênica oposto (fraco/forte).

Os testes de neutralidade (D de Tajima e Ewens Watterson) detectaram na região cis-reguladora do *CCR5* de ameríndios um desvio significativo no polimorfismo quando comparado aos valores esperados sob a hipótese nula de neutralidade e equilíbrio populacional. O valor positivo do D de Tajima pode ser resultado de seleção balanceadora que atuou na região cis-reguladora do *CCR5* de ameríndios. Este loco apresentou nas populações indígenas da região amazônica, um alto nível de compartilhamento de haplótipos, refletido também nos baixos valores de diferenciação interpopulacional medidos através do *Fst*. Estas características parecem não ser explicáveis apenas por eventos demográficos, uma vez que os resultados permaneceram significativos em dois tipos de análise adicionais: (a) comparações com marcadores neutros de populações de regiões geográficas próximas às analisadas para a região cis-reguladora do *CCR5*, (b) comparação com dados simulados sob modelo realista da história demográfica, incluindo a ocorrência de gargalos e expansões.

O modelo demográfico utilizado não leva em consideração eventos de estruturação populacional ocorridos nas populações indígenas da América. Seria interessante acrescentar tais eventos à esse modelo, para que o impacto dessas subdivisões na diferenciação genética (*Fst*) entre as tribos fosse analisada.

A maior limitação da presente análise, deve-se ao fato de termos utilizado como controle das condições de neutralidade, amostras de populações distintas daquelas analisadas para a região cis-reguladora do *CCR5*. A base de dados ideal para corrigir os efeitos demográficos, deve conter resultados da análise de locos neutros feita nos mesmos indivíduos analisados para o loco candidato à seleção. Um projeto para gerar esses dados está atualmente em andamento em nosso laboratório.

Apesar da forte assinatura de seleção balanceadora encontrada para a região cis-reguladora do *CCR5* em populações nativas da região amazônica as populações Arara e Waiãpi parecem ainda evidenciar a ação de fortes eventos demográficos.

A assinatura de seleção balanceadora encontrada para a região cis-reguladora do *CCR5* pode ser resultado da manutenção de um sinal de seleção que atuou nas populações ancestrais. Uma possível maneira de se testar essa hipótese é através de uma análise de simulação de um cenário em que a seleção tenha operado no passado, e deixado de operar mais recentemente, após a ocupação da América. Apesar de simulações coalescentes serem baseadas em modelo neutro, é possível simular a ação da seleção natural de forma indireta ajustando parâmetros, como por exemplo, o tamanho efetivo populacional. Outra maneira seria usar simulações não-neutras que permitem introduzir na simulação a ação de seleção natural. Para corroborar a hipótese alternativa, de que a seleção natural atuou ou atua na região cis-reguladora do *CCR5* de populações nativas da região amazônica, uma simulação representando um evento de forte seleção balanceadora em uma população ancestral de ameríndios deve ser realizada. A rejeição da hipótese nula (seleção balanceadora em população ancestral e ausência de seleção na América) corroboraria a hipótese de seleção natural nas populações de ameríndios.

Por fim, demonstramos que para a região cis-reguladora do *CCR5*, existe uma associação estatisticamente significativa entre regiões ativadoras de *splicing* (*ESEs*) e os sítios polimórficos com alelos em frequências intermediárias na espécie humana. Esses resultados corroboram a importância dos *ESEs* na regulação do *splicing* e sugerem uma importância funcional para o *splicing* alternativo já observado na região cis-reguladora do *CCR5*.

Resumo

Populações nativas da América do Sul apresentam diversidade genética reduzida em relação às demais populações do mundo e alta diferenciação interpopulacional dentro do continente. As pressões seletivas sobre a região cis-reguladora do *CCR5* geram uma assinatura de diversidade oposta àquela esperada com base na história demográfica, reduzindo a variação interpopulacional e contribuindo para uma elevação das taxas de polimorfismo. No presente estudo investigamos a interação entre esses processos micro-evolutivos, analisando o polimorfismo da região cis-reguladora de ameríndios da América do Sul e comparando os resultados com aqueles obtidos em outras regiões do mundo.

Sequenciamos 927 pares de base (pb) da região controladora do *CCR5* em 7 tribos indígenas da região amazônica. Nenhum haplótipo exclusivo foi encontrado em ameríndios e os dois haplogrupos mais comuns foram de diferentes clusters filogenéticos. A diversidade nucleotídica (π) para a amostra total foi a mais alta dentre todas as regiões do mundo já estudadas ($\pi=0,0027$). A estimativa da diversidade genética populacional para ameríndios, baseada no número de sítios polimórficos (θ_s), foi similar aos valores encontrados para as populações não-africanas, sendo que em africanos o valor foi

praticamente o dobro. Conseqüentemente, foi observado um valor de D de Tajima positivo e estatisticamente significativo ($D=2,82$, $p<0,01$), maior do que aquele visto para outras regiões do mundo.

Através de comparações empíricas e de simulações coalescentes verificamos que o alto valor de D de Tajima encontrado para ameríndios não é explicável por um recente modelo demográfico de evolução humana. O teste Ewens-Watterson indicou para a amostra ameríndia uma taxa de heterozigose maior do que esperada para uma população neutra.

O valor de F_{st} observado para comparação envolvendo asiáticos e ameríndios foi mais baixo que 1000 valores de F_{st} calculados a partir de 783 microssatélites genotipados em amostras de mesmo tamanho de regiões geográficas similares.

Essas características corroboram a hipótese de seleção balanceadora na região cis-reguladora do CCR5 de ameríndios da região amazônica brasileira.

No presente estudo demonstramos também a existência de uma associação estatisticamente significativa ($p<0,01$) entre os alelos mantidos em frequências intermediárias na população humana e regiões ativadoras de *splicing* localizadas em exons (ESEs).

Finalmente, nós acreditamos que um modelo demográfico de evolução humana, complementar ao utilizado neste trabalho, deve ser testado para se refutar a hipótese de que a forte assinatura de seleção balanceadora observada em ameríndios não foi causada por seleção natural que atuou em população ancestral à de indígenas americanos.

Abstract

Native american populations show lower genetic diversity and higher interpopulational genetic differences than populations from other continents. Within South America groups from the non-andean geographic region shows extremely high genetic differentiation.

The strong signature of balancing selection observed for 5' cis-regulatory region of the *CCR5* gene at the worldwide scale represents an opposite pattern of genetic variation to that expected by demographic process which acted in Amerindians.

To evaluate the impact of complex demographic history on the 5' cis-regulatory region of *CCR5* we resequenced 927 bp of this locus in 62 individuals from different native groups located in the Brazilian Amazonian region. No new haplotype was detected and the two most common haplogroups observed were from different phylogenetic clusters, according with the pattern observed for other human populations. The level of heterozygosity of the total sample, measured by nucleotide diversity (π) was the highest yet described for human populations ($\pi=0,0027$) while values based on polymorphic sites (θ_s) were similar among Amerindians and non-Africans populations. Consequently we observed a positive and significant Tajima's D value ($D=2,82$, $p < 0,01$). This value was higher than all D values observed for the other human populations.

Observed summary statistics (π and D) were significantly higher than same statistics estimated from 2000 simulated samples assuming a recently proposed demographic model of human evolution. We also found significant deviations in Ewens-Watterson homozygosity test toward an excess of heterozygosity for Amerindian sample. Another

striking feature of *CCR5* cis-regulatory region was the low F_{st} value among populations. The F_{st} among Asians and Amerindians was an outlier among 1000 F_{st} values estimated by 783 microsatellites of samples from similar geographic regions and with the same sample sizes as those of the *CCR5* 5' cis-regulatory data. These features corroborates the strong signature of balancing selection described for *CCR5* 5' cis regulatory region.

We also contributed to discussion about functional aspects of *CCR5* cis-regulatory region demonstrating a significant association between intermediate frequency SNPs and exonic splicing enhancers (*ESEs*) regions ($p < 0,01$). Finally, we believe that improved models of demographic human evolution must be tested to refute the hypothesis that the strong signature of balancing selection observed in Amerindians was not caused by a selection pressure which occurred in an ancestral population.

Bibliografia

- Akey, J., M. Eberle, *et al.* Population history and natural selection shape patterns of genetic variation in 132 genes. PLoS Biol, v.2, n.10, 2004 Oct, p.e286. 2004.
- Bamshad, M. e S. P. Wooding. Signatures of natural selection in the human genome. Nat Rev Genet, v.4, n.2, Feb, p.99-111. 2003.
- Bamshad, M. J., S. Mummidi, *et al.* A strong signature of balancing selection in the 5' cis-regulatory region of CCR5. Proc Natl Acad Sci U S A, v.99, n.16, Aug 6, p.10539-44. 2002.
- Bandelt, H. J., P. Forster, *et al.* Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol, v.16, n.1, Jan, p.37-48. 1999.
- Barbujani, G., A. Magagni, *et al.* An apportionment of human DNA diversity. Proc Natl Acad Sci U S A, v.94, n.9, 1997 Apr, p.4516-9. 1997.
- Battilana, J., L. Cardoso-Silva, *et al.* Molecular variability of the 16p13.3 region in Amerindians and its anthropological significance. Ann Hum Genet, v.71, n.Pt 1, Jan, p.64-76. 2007.
- Battilana, J., N. J. Fagundes, *et al.* Alu insertion polymorphisms in Native Americans and related Asian populations. Ann Hum Biol, v.33, n.2, Mar-Apr, p.142-60. 2006.
- Benkirane, M., D. Y. Jin, *et al.* Mechanism of transdominant inhibition of CCR5-mediated HIV-1 infection by ccr5delta32. J Biol Chem, v.272, n.49, Dec 5, p.30603-6. 1997.
- Bersaglieri, T., P. Sabeti, *et al.* Genetic signatures of strong recent positive selection at the lactase gene. Am J Hum Genet, v.74, n.6, 2004 Jun, p.1111-20. 2004.

Bianchi, N. O., G. Bailliet, *et al.* Origin of Amerindian Y-chromosomes as inferred by the analysis of six polymorphic markers. Am J Phys Anthropol, v.102, n.1, Jan, p.79-89. 1997.

Bianchi, N. O., C. I. Catanesi, *et al.* Characterization of ancestral and derived Y-chromosome haplotypes of New World native populations. Am J Hum Genet, v.63, n.6, Dec, p.1862-71. 1998.

Black, F. L., S. E. B. Santos, *et al.* Genetic variation within the Tupi linguistic group: new data on three Amazonian tribes. Annals of Human Biology, v.15, n.5, p.337 - 351. 1988.

Blencowe, B. Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. Trends Biochem Sci, v.25, n.3, 2000 Mar, p.106-10. 2000.

Bonatto, S. L. e F. M. Salzano. A single and early migration for the peopling of the Americas supported by mitochondrial DNA sequence data. Proc Natl Acad Sci U S A, v.94, n.5, Mar 4, p.1866-71. 1997a.

_____. Diversity and age of the four major mtDNA haplogroups, and their implications for the peopling of the New World. Am J Hum Genet, v.61, n.6, Dec, p.1413-23. 1997b.

Braverman, J., R. Hudson, *et al.* The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics, v.140, n.2, 1995 Jun, p.783-96. 1995.

Brown, M., S. Hosseini, *et al.* mtDNA haplogroup X: An ancient link between Europe/Western Asia and North America? Am J Hum Genet, v.63, n.6, 1998 Dec, p.1852-61. 1998.

Callegari-Jacques, S. M., F. M. Salzano, *et al.* Further blood genetic studies on Amazonian diversity--data from four Indian groups. Ann Hum Biol, v.21, n.5, Sep-Oct, p.465-81. 1994.

Cartegni, L., S. Chew, *et al.* Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nat Rev Genet, v.3, n.4, 2002 Apr, p.285-98. 2002.

Chakravarti, A. Population genetics--making sense out of sequence. Nat Genet, v.21, n.1 Suppl, 1999 Jan, p.56-60. 1999.

Charlesworth, D. Balancing selection and its effects on sequences in nearby genome regions. PLoS Genet, v.2, n.4, 2006 Apr, p.e64. 2006.

Chen, F. e W. Li. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. Am J Hum Genet, v.68, n.2, 2001 Feb, p.444-56. 2001.

Clark, V. e M. Dean. Characterisation of SNP haplotype structure in chemokine and chemokine receptor genes using CEPH pedigrees and statistical estimation. Hum Genomics, v.1, n.3, 2004 Mar, p.195-207. 2004.

Cohen, O. J., M. Vaccarezza, *et al.* Heterozygosity for a defective gene for CC chemokine receptor 5 is not the sole determinant for the immunologic and virologic phenotype of HIV-infected long-term nonprogressors. J Clin Invest, v.100, n.6, Sep 15, p.1581-9. 1997.

Di Rienzo, A., P. Donnelly, *et al.* Heterogeneity of microsatellite mutations within and between loci, and implications for human demographic histories. Genetics, v.148, n.3, 1998 Mar, p.1269-84. 1998.

Di Rienzo, A. e A. Wilson. Branching pattern in the evolutionary tree for human mitochondrial DNA. Proc Natl Acad Sci U S A, v.88, n.5, 1991 Mar, p.1597-601. 1991.

Eswaran, V., H. Harpending, *et al.* Genomics refutes an exclusively African origin of humans. J Hum Evol, v.49, n.1, 2005 Jul, p.1-18. 2005.

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

Excoffier, L., J. Novembre, *et al.* SIMCOAL: a general coalescent program for the simulation of molecular data in interconnected populations with arbitrary demography. J Hered, v.91, n.6, 2000 Nov-Dec, p.506-9. 2000.

Fagundes, N., N. Ray, *et al.* Statistical evaluation of alternative models of human evolution. Proc Natl Acad Sci U S A, v.104, n.45, 2007 Nov, p.17614-9. 2007.

Fagundes, N. J., F. M. Salzano, *et al.* Worldwide genetic variation at the 3'-UTR region of the LDLR gene: possible influence of natural selection. Ann Hum Genet, v.69, n.Pt 4, Jul, p.389-400. 2005.

Fairbrother, W., R. Yeh, *et al.* Predictive identification of exonic splicing enhancers in human genes. Science, v.297, n.5583, 2002 Aug, p.1007-13. 2002.

Franco, R. F., A. G. Araujo, *et al.* Factor IX gene haplotypes in Amerindians. Hum Biol, v.69, n.1, Feb, p.1-9. 1997.

Garrigan, D. e M. Hammer. Reconstructing human origins in the genomic era. Nat Rev Genet, v.7, n.9, 2006 Sep, p.669-80. 2006.

Garrigan, D. e P. W. Hedrick. Perspective: detecting adaptive molecular polymorphism: lessons from the MHC. Evolution Int J Org Evolution, v.57, n.8, Aug, p.1707-22. 2003.

Gibbons, A. Human anthropology. Modern men trace ancestry to African migrants. Science, v.292, n.5519, May 11, p.1051-2. 2001.

Gordon, D., C. Abajian, *et al.* Consed: a graphical tool for sequence finishing. Genome Res, v.8, n.3, Mar, p.195-202. 1998.

- Grahovac, B., R. Sukernik, *et al.* Polymorphism of the HLA class II loci in Siberian populations. Hum Genet, v.102, n.1, 1998 Jan, p.27-43. 1998.
- Guo, S. W. e E. A. Thompson. Performing the exact test of Hardy-Weinberg proportion for multiple alleles. Biometrics, v.48, n.2, Jun, p.361-72. 1992.
- Harris, E. E. e D. Meyer. The molecular signature of selection underlying human adaptations. Am J Phys Anthropol, v.Suppl 43, p.89-130. 2006.
- Hartl, D. L. e A. G. Clark. Principles of Population Genetics: Sinauer Associates, Inc. 1997.
- Hedrick, P. W. Genetic Polymorphism in Heterogeneous Environments: The Age of Genomics. Annual Review of Ecology, Evolution, and Systematics, v.37, n.1, p.67-93. 2006.
- Hey, J. On the number of New World founders: a population genetic portrait of the peopling of the Americas. PLoS Biol, v.3, n.6, 2005 Jun, p.e193. 2005.
- Hudson, R. R. Gene genealogies and the coalescent process. Oxford Surveys in Evolutionary Biology: Oxford University Press 1990.
- Hughes, A., B. Packer, *et al.* High level of functional polymorphism indicates a unique role of natural selection at human immune system loci. Immunogenetics, v.57, n.11, 2005 Dec, p.821-7. 2005.
- Jorde, L. B., W. S. Watkins, *et al.* Population genomics: a bridge from evolutionary history to genetic medicine. Hum Mol Genet, v.10, n.20, Oct 1, p.2199-207. 2001.
- Karafet, T., S. Zegura, *et al.* Ancestral Asian source(s) of new world Y-chromosome founder haplotypes. Am J Hum Genet, v.64, n.3, 1999 Mar, p.817-31. 1999.

Kimmel, M., R. Chakraborty, *et al.* Signatures of population expansion in microsatellite repeat data. Genetics, v.148, n.4, 1998 Apr, p.1921-30. 1998.

Kimura, M. Recent development of the neutral theory viewed from the Wrightian tradition of theoretical population genetics. Proc Natl Acad Sci U S A, v.88, n.14, Jul 15, p.5969-73. 1991.

Kingman, J. F. Origins of the coalescent. 1974-1982. Genetics, v.156, n.4, Dec, p.1461-3. 2000.

Kolman, C., N. Sambuughin, *et al.* Mitochondrial DNA analysis of Mongolian populations and implications for the origin of New World founders. Genetics, v.142, n.4, 1996 Apr, p.1321-34. 1996.

Lell, J., R. Sukernik, *et al.* The dual origin and Siberian affinities of Native American Y chromosomes. Am J Hum Genet, v.70, n.1, 2002 Jan, p.192-206. 2002.

Li, W. e L. Sadler. Low nucleotide diversity in man. Genetics, v.129, n.2, 1991 Oct, p.513-23. 1991.

Lim, J., W. Glass, *et al.* CCR5: no longer a "good for nothing" gene--chemokine control of West Nile virus infection. Trends Immunol, v.27, n.7, 2006 Jul, p.308-12. 2006.

Martin, M. P., M. Dean, *et al.* Genetic acceleration of AIDS progression by a promoter variant of CCR5. Science, v.282, n.5395, Dec 4, p.1907-11. 1998.

Meltzer, D. J. Anthropology: Monte Verde and the Pleistocene Peopling of the Americas. Science, v.276, n.5313, May 2, 1997, p.754-755. 1997.

Merriwether, D., W. Hall, *et al.* mtDNA variation indicates Mongolia may have been the source for the founding population for the New World. Am J Hum Genet, v.59, n.1, 1996 Jul, p.204-12. 1996.

Meyer, C. G., J. May, *et al.* TNFalpha-308A associated with shorter intervals of Plasmodium falciparum reinfections. Tissue Antigens, v.59, n.4, Apr, p.287-92. 2002.

Mignone, F., C. Gissi, *et al.* Untranslated regions of mRNAs. Genome Biol, v.3, n.3, p.REVIEWS0004. 2002.

Mousinho-Ribeiro, R. C., G. Pante-De-Souza, *et al.* Genetic relationships among native americans based on β -globin gene. Genetics and Molecular Biology, v.26, n.3, p.229-234. 2003.

Mummidi, S., S. Ahuja, *et al.* The human CC chemokine receptor 5 (CCR5) gene. Multiple transcripts with 5'-end heterogeneity, dual promoter usage, and evidence for polymorphisms within the regulatory regions and noncoding exons. J Biol Chem, v.272, n.49, 1997 Dec, p.30662-71. 1997.

Mummidi, S., M. Bamshad, *et al.* Evolution of human and non-human primate CC chemokine receptor 5 gene and mRNA. Potential roles for haplotype and mRNA diversity, differential haplotype-specific transcriptional activity, and altered transcription factor binding to polymorphic nucleotides in the pathogenesis of HIV-1 and simian immunodeficiency virus. J Biol Chem, v.275, n.25, Jun 23, p.18946-61. 2000.

Nachman, M., W. Brown, *et al.* Nonneutral mitochondrial DNA variation in humans and chimpanzees. Genetics, v.142, n.3, 1996 Mar, p.953-63. 1996.

Neel, J., R. Biggar, *et al.* Virologic and genetic studies relate Amerind origins to the indigenous people of the Mongolia/Manchuria/southeastern Siberia region. Proc Natl Acad Sci U S A, v.91, n.22, 1994 Oct, p.10737-41. 1994.

Nei, M. Molecular Evolutionary Genetics: Columbia University Press, New York 1987.

Nickerson, D. A., V. O. Tobe, *et al.* PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. Nucleic Acids Res, v.25, n.14, Jul 15, p.2745-51. 1997.

Nielsen, R. Molecular signatures of natural selection. Annu Rev Genet, v.39, p.197-218. 2005.

Prahalad, S., J. Bohnsack, *et al.* Association of two functional polymorphisms in the CCR5 gene with juvenile rheumatoid arthritis. Genes Immun, v.7, n.6, 2006 Sep, p.468-75. 2006.

Premack, B. A. e T. J. Schall. Chemokine receptors: gateways to inflammation and infection. Nat Med, v.2, n.11, Nov, p.1174-8. 1996.

Prugnolle, F., A. Manica, *et al.* Geography predicts neutral genetic diversity of human populations. Curr Biol, v.15, n.5, Mar 8, p.R159-60. 2005.

Przeworski, M., R. R. Hudson, *et al.* Adjusting the focus on human variation. Trends Genet, v.16, n.7, Jul, p.296-302. 2000.

Ramachandran, S., O. Deshpande, *et al.* Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proc Natl Acad Sci U S A, v.102, n.44, 2005 Nov, p.15942-7. 2005.

Reich, D. e D. Goldstein. Genetic evidence for a Paleolithic human population expansion in Africa. Proc Natl Acad Sci U S A, v.95, n.14, 1998 Jul, p.8119-23. 1998.

Ribeiro-Dos-Santos, A. K., J. F. Guerreiro, *et al.* The split of the Arara population: comparison of genetic drift and founder effect. Hum Hered, v.51, n.1-2, p.79-84. 2001.

Richman, A. Evolution of balanced genetic polymorphism. Mol Ecol, v.9, n.12, p.1953-63. 2000.

Rogers, A. e H. Harpending. Population growth makes waves in the distribution of pairwise genetic differences. Mol Biol Evol, v.9, n.3, 1992 May, p.552-69. 1992.

Rosenberg, N., S. Mahajan, *et al.* Clines, clusters, and the effect of study design on the inference of human population structure. PLoS Genet, v.1, n.6, 2005 Dec, p.e70. 2005.

Rosenberg, N. A., J. K. Pritchard, *et al.* Genetic structure of human populations. Science, v.298, n.5602, Dec 20, p.2381-5. 2002.

Rottman, J. B., K. P. Ganley, *et al.* Cellular localization of the chemokine receptor CCR5. Correlation to cellular targets of HIV-1 infection. Am J Pathol, v.151, n.5, Nov, p.1341-51. 1997.

Sabeti, P., P. Varilly, *et al.* Genome-wide detection and characterization of positive selection in human populations. Nature, v.449, n.7164, 2007 Oct, p.913-8. 2007.

Salzano, F. e S. Callegari-Jacques. Amerindian and NonAmerindian autosome molecular variability--a test analysis. Genetica, v.126, n.1-2, 2006 Jan, p.237-42. 2006.

Salzano, F. M. Molecular variability in Amerindians: widespread but uneven information. An Acad Bras Cienc, v.74, n.2, Jun, p.223-63. 2002.

Santos, F., A. Pandya, *et al.* The central Siberian origin for native American Y chromosomes. Am J Hum Genet, v.64, n.2, 1999 Feb, p.619-28. 1999.

Santos, F., L. Rodriguez-Delfin, *et al.* North and South Amerindians may have the same major founder Y chromosome haplotype. Am J Hum Genet, v.58, n.6, 1996 Jun, p.1369-70. 1996.

Schneider, S., D. Roessli, *et al.* Arlequin: A software for population genetics data analysis. Ver 2.000. 2000. 2000.

Schroeder, K., T. Schurr, *et al.* A private allele ubiquitous in the Americas. Biol Lett, v.3, n.2, 2007 Apr, p.218-23. 2007.

Short, A. D., L. J. Kennedy, *et al.* Canine DNA subjected to whole genome amplification is suitable for a wide range of molecular applications. J Hered, v.96, n.7, p.829-35. 2005.

Silva, W. A., Jr., S. L. Bonatto, *et al.* Mitochondrial genome diversity of Native Americans supports a single early entry of founder populations into America. Am J Hum Genet, v.71, n.1, Jul, p.187-92. 2002.

Slatkin, M. e R. Hudson. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics, v.129, n.2, 1991 Oct, p.555-62. 1991.

Smith, J. e J. Haigh. The hitch-hiking effect of a favourable gene. Genet Res, v.23, n.1, 1974 Feb, p.23-35. 1974.

Stephens, J. C., J. A. Schneider, *et al.* Haplotype variation and linkage disequilibrium in 313 human genes. Science, v.293, n.5529, Jul 20, p.489-93. 2001.

Stephens, M., N. J. Smith, *et al.* A new statistical method for haplotype reconstruction from population data. Am J Hum Genet, v.68, n.4, Apr, p.978-89. 2001.

Stringer, C. B. e P. Andrews. Genetic and fossil evidence for the origin of modern humans. Science, v.239, n.4845, Mar 11, p.1263-8. 1988.

Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics, v.123, n.3, p.585-95. 1989.

Takahata, N. e Y. Satta. Evolution of the primate lineage leading to modern humans: phylogenetic and demographic inferences from DNA sequences. Proc Natl Acad Sci U S A, v.94, n.9, 1997 Apr, p.4811-5. 1997.

Tarazona-Santos, E. e S. A. Tishkoff. Divergent patterns of linkage disequilibrium and haplotype structure across global populations at the interleukin-13 (IL13) locus. Genes Immun, v.6, n.1, Feb, p.53-65. 2005.

Templeton, A. Haplotype trees and modern human origins. Am J Phys Anthropol, v.Suppl 41, p.33-59. 2005.

Tishkoff, S. A. e K. K. Kidd. Implications of biogeography of human populations for 'race' and medicine. Nat Genet, v.36, n.11 Suppl, Nov, p.S21-7. 2004.

Torrioni, A., R. Sukernik, *et al.* mtDNA variation of aboriginal Siberians reveals distinct genetic affinities with Native Americans. Am J Hum Genet, v.53, n.3, 1993 Sep, p.591-608. 1993.

Underhill, P., L. Jin, *et al.* Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. Genome Res, v.7, n.10, 1997 Oct, p.996-1005. 1997.

Vallender, E. J. e B. T. Lahn. Positive selection on the human genome. Hum Mol Genet, v.13 Spec No 2, Oct 1, p.R245-54. 2004.

Wall, J. e M. Przeworski. When did the human population size start increasing? Genetics, v.155, n.4, 2000 Aug, p.1865-74. 2000.

Wang, G. e T. Cooper. Splicing in disease: disruption of the splicing code and the decoding machinery. Nat Rev Genet, v.8, n.10, 2007 Oct, p.749-61. 2007.

Wray, G. A. The evolutionary significance of cis-regulatory mutations. Nat Rev Genet, v.8, n.3, Mar, p.206-16. 2007.

Wright, S. Genetical structure of populations. Nature, v.166, n.4215, Aug 12, p.247-9. 1950.

Wu, L., W. A. Paxton, *et al.* CCR5 levels and expression pattern correlate with infectability by macrophage-tropic HIV-1, in vitro. J Exp Med, v.185, n.9, May 5, p.1681-91. 1997.

Yeo, G., S. Hoon, *et al.* Variation in sequence and organization of splicing regulatory elements in vertebrate genes. Proc Natl Acad Sci U S A, v.101, n.44, 2004 Nov, p.15700-5. 2004.

Zhao, Z., N. Yu, *et al.* Nucleotide variation and haplotype diversity in a 10-kb noncoding region in three continental human populations. Genetics, v.174, n.1, 2006 Sep, p.399-409. 2006.