UNIVERSIDADE CATÓLICA DE BRASÍLIA

TULIO CESAR DE LIMA LINS

Impacto da Miscigenação na Aplicação do HapMap para a População Brasileira Avaliado nos Genes *PTPN22* e *VDR*

BRASÍLIA
2007

TULIO CESAR DE LIMA LINS

Impacto da Miscigenação na Aplicação do HapMap para a População Brasileira Avaliado nos Genes *PTPN22* e *VDR*

Dissertação apresentada ao programa de Pós-Graduação *Stricto Sensu* em Ciências Genômicas e Biotecnologia da Universidade Católica de Brasília, como requisito para obtenção do Título de Mestre em Ciências Genômicas e Biotecnologia.

Orientador: Rinaldo Wellerson Pereira

BRASÍLIA
2007

L759i Lins, Tulio Cesar de Lima.
Impacto da miscigenação na aplicação do HapMap para a população brasileira avaliados nos genes PTPN22 e VDR / Tulio Cesar de Lima Lins. – 2007.
194 f. : il. ; 30 cm

Dissertação (mestrado) – Universidade Católica de Brasília, 2007.
Orientação: Rinaldo Wellerson Pereira

1. Genética. 2. Genes. 3. Miscigenação. 4. População brasileira. I. Pereira, Rinaldo Wellerson., orient. II. Título

CDU 575

Ficha elaborada pela Coordenação de Processamento do Acervo do SIBI – UCB.

Dissertação defendida e apresentada como requisito parcial para obtenção do Título de Mestre em Ciências Genômicas e Biotecnologia, defendida e aprovada em 02 de Abril de 2007, pela banca examinadora constituída por:

_____

Orientador: Prof. Dr Rinaldo Wellerson Pereira

Universidade Católica de Brasília

_____

Examinadora: Profª. Drª. Rosane Garcia Collevatti

Universidade Católica de Brasília

_____

Examinadora externa: Profª. Drª. Maria Luiza Petzl-Erler

Universidade Federal do Paraná

AGRADECIMENTOS

Ao Orientador Prof. Dr. Rinaldo Wellerson Pereira, pela parceria, amizade e incentivo, por idealizar, encorajar e acreditar na realização deste projeto;

Aos Profs. Drs. Márcio Elias Ferreira e Dario Grattapaglia, por sempre acreditar e incentivar minha trajetória acadêmica e pela grande colaboração neste trabalho;

Às Prof^{as} Dr^{as} Rosane Garcia Collevatti e Maria Luiza Petzl-Erler, membros da banca avaliadora, agradeço a oportunidade de avaliar e colaborar com este documento;

Aos amigos e colegas do Laboratório de Genética Vegetal da EMBRAPA-Cenargen, em especial, Marco Pessoa, Eva Mamani, Danielle Faria, Juliano Pádua e Marília Pappas, pelo imenso auxílio nas genotipagens e nas análises estatísticas;

Aos amigos e colegas do Laboratório Heréditas: Polyanna Diener, Nathália Bueno, André Santos, Mônica Ribeiro, Cláudia Guimarães, Camila Oliveira, Sandra de Andrade, Militize Sanches e Tânia Grattapaglia, agradeço o apoio e o auxílio na organização das amostras;

Aos amigos e colegas do Laboratório de Ciências Genômicas, Alessandra Reis, Aline Braga, Breno Abreu e Erika Grandi pelos grandes momentos de produção "científica e cultural";

Aos alunos de Iniciação Cinetífica Luciana Rollemberg, Rodrigo Vieira e Meiriele Silva que contribuíram com muito esforço e aprendizagem para o desenvolvimento deste trabalho;

Aos colegas Paulo Gentil e Ricardo Moreno, e ao Prof Dr Ricardo Jacó de Oliveira do Programa de Pós-Graduação em Educação Física da UCB pelo incentivo e colaboração em projetos paralelos;

A todos os professores, funcionários e demais alunos do programa de Ciências Genômicas da UCB que durante esses dois anos me apoiaram, incentivaram, ajudaram e ensinaram;

Aos grandes amigos Fabiano, Adriane, Mariana (Mari), Mariana (Mamá), Wagner, Isabel, Xandão, Marco, Bianca, Gá, Rato, Glycon, Carol, Iuri, Paulinho, Vanessa,

Marcelo e Juliana por simplesmente serem meus amigos e me apoiarem sempre que necessário;

Aos meus pais Tales e Graça, meus irmãos Tales e Taís, e demais familiares queridos pelo carinho e incentivo que nunca faltaram;

À amada Cecília por me oferecer os melhores momentos que tive durante esses anos, por apoiar, debater e incentivar este projeto;

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior do Ministério da Educação (CAPES/MEC) pelo fornecimento da bolsa de Mestrado;

Aos demais não mencionados, porém não esquecidos, que em algum dia na vida cruzaram meu caminho, sem vocês eu não estaria aqui;

Agradeço a Deus por ter chegado até aqui com mente sã e corpo são.

Meu muito obrigado a todos!

I.

Whether we write or speak or do but look
We are ever unapparent. What we are
Cannot be transfused into word or book.
Our soul from us is infinitely far.
However much we give our thoughts the will
To be our soul and gesture it abroad,
Our hearts are incommunicable still.
In what we show ourselves we are ignored.
The abyss from soul to soul cannot be bridged
By any skill of thought or trick of seeming.
Unto our very selves we are abridged
When we would utter to our thought our being.
We are our dreams of ourselves, souls by gleams,
And each to each other dreams of others' dreams.

Fernando Pessoa, 1918 – 35 Sonnets

RESUMO

Os genes *PTPN22* e *VDR* são associados a várias doenças auto-imunes e a fenótipos complexos das vias do metabolismo ósseo. Além disso, a associação de alguns polimorfismos desses genes é específica para alguns grupos populacionais e, portanto estudos em populações miscigenadas devem ser criteriosamente avaliados para esses fatores. A população brasileira é considerada uma das mais heterogêneas do mundo e estudos de associação genética nessa população devem ser conduzidos com controle genético para estratificação em termos dos níveis de ancestralidade. O desequilíbrio de ligação é uma ferramenta de estatística genética utilizada para detectar associações por blocos haplotípicos em genes candidatos. O presente trabalho teve por objetivo a investigação da contribuição de miscigenação na população brasileira para investigar os padrões de desequilíbrio de ligação nos genes *PTPN22* e *VDR* e avaliar os efeitos da ancestralidade genética nos mesmos. Para isso, 200 indivíduos brasileiros, não relacionados e separados por região geográfica, foram genotipados para 34 marcadores PTPN22

ABSTRACT

The *PTPN22* and *VDR* genes are commonly associated to several autoimmune diseases and complex phenotypes related to bone metabolism. The association of some of those polymorphisms are ethnic-specific, and therefore, admixture population studies must be conditionally evaluated to these factors. The Brazilian population is considered the most heterogeneous in the world and their genetic association studies must be controlled to the stratified levels of ancestry proportions. Linkage disequilibrium is a genetic statistic method to detect association with candidate genes by haplotype blocks. The aim of the present work was to estimate the admixture contribution levels of the Brazilian population, to evaluate the linkage disequilibrium patterns of *PTPN22* and *VDR* genes and to investigate the effects of ancestry proportions on these genes. For that, 200 non-related Brazilian individuals, separated by geographic region were genotyped for 34 ancestry informative markers, six SNPs in the *PTPN22* gene and 21 in the *VDR* gene by single base extension method (SNaPshot™ – Applied Biosystems). The results showed that the markers were able to assign 19,5% of African ancestry proportion and 80,5% of European to the Brazilian population, however they were unable to assign correctly the Amerindian proportion. The *PTPN22* gene showed high genetic ancestry influence on the distribution of haplotypes, while the *VDR* gene showed a similar haplotype block pattern to those described to European population. The transferability of tagSNPs among HapMap populations to Brazilian population was divergent and uncertain for the two genes, with high loss of variability in the *PTPN22* gene using tagSNPs of the European derived population and, to the *VDR* gene, there was higher loss of variability when using the tagSNPs selected for African or Asian derived populations. This indicates that the use of tagSNPs should not be generalized in the Brazilian population. The information generated at this work may contribute to several prospective genetic association studies involving the *PTPN22* and *VDR* genes, and also as a base for population stratification studies in admixture populations.

Keywords: Brazilian Population, SNP, Linkage Disequilibrium, Haplotypes, PTPN22, VDR.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

SUMÁRIO

INTRODUÇÃO

Doenças Complexas

Há muitas décadas, avanços na genética permitiram à medicina a identificação de genes que causam doenças e identificação dos fatores de risco que podem modular a severidade das mesmas. Neste contexto, a hereditariedade é um fator etiológico importante para determinação do caráter das enfermidades. Várias doenças de evidências genéticas afligem a humanidade, e algumas delas são doenças raras causadas pela ação de um único, ou de poucos genes (doenças Mendelianas), e que, pela simples análise de heredogramas é definido o modo de herança, tipicamente familiar. Como alguns exemplos temos a doença de Huntington, hemofilia, fibrose cística e fenilcetonúria (BADANO; KATSANIS, 2002). Outras, por sua vez, não estão associadas diretamente a registros de ocorrência familiar (as chamadas doenças complexas) e geralmente são classificadas como doenças genéticas de alta freqüência, como, por exemplo, cânceres de vários tipos, doenças coronarianas, hipertensão, diabetes, obesidade, e alcoolismo (DE BOER; VAN DEN BERG; VAN VELDHUISEN, 2006; LANDER; SCHORK, 1994; POCIOT; MCDERMOTT, 2002; THOMSON; ESPOSITO, 1999).

As doenças complexas ocorrem quando o fenótipo é resultado da ação de alelos em diversos locos ou genes que agem em conjunto em um número na maioria das vezes desconhecido, com efeitos aditivos ou sinergéticos (BOTSTEIN; RISCH, 2003; LANDER; SCHORK, 1994; RISCH; MERIKANGAS, 1996). Os indivíduos afetados podem ter diferentes mutações ou polimorfismos genéticos que conduzem ao mesmo fenótipo, e, geralmente, existem fatores ambientais que podem atuar conjuntamente na determinação da doença (BOTSTEIN; RISCH, 2003; LANDER; SCHORK, 1994; RISCH; MERIKANGAS, 1996). Por outro lado, as doenças mendelianas estão relacionadas a um ou a poucos genes e são dependentes dos seus efeitos de penetrância, herdabilidade e haploinsuficiência,

além de fatores populacionais, como consangüinidade e endogamia (BADANO; KATSANIS, 2002; BUCHANAN; WEISS; FULLERTON, 2006).

No último século as pesquisas se voltaram para a identificação dos vários fatores genéticos associados a pré-disposição ao desenvolvimento de doenças complexas. A busca por fatores que tornam uma característica biológica distinta de outra se origina de estudos remotos em que se procuravam alterações protéicas e enzimáticas que pudessem evidenciar causas ou motivos que levassem a um fenótipo distinto (SMITHIES; WALKER, 1955), até os tempos atuais que o intuito das pesquisas está voltado para polimorfismos pontuais e estruturais do ácido desoxirribonucléico, o DNA (GUSELLA et al., 1983).

Com o foco voltado para o DNA, o conhecimento da seqüência completa do genoma humano por meio de investimentos em consórcios de laboratórios de centros de pesquisa e Universidades trouxe a expectativa de que as pesquisas no campo de genética médica pudessem traçar novos caminhos para o entendimento das doenças humanas. Nas vertentes históricas, dois grupos se propuseram a trabalhar no que seria chamado de projeto genoma humano (LANDER et al., 2001; VENTER et al., 2001), que atualmente está considerado completo (COLLINS et al., 2003) e depositado em bancos de dados eletrônicos para consulta pública (LETOVSKY et al., 1998). No decorrer do processo de seqüenciamento do genoma humano, foram identificados milhões de polimorfismos de uma única base, os SNPs (do inglês, *Single Nucleotide Polymorphisms*) (SACHIDANANDAM et al., 2001). Porém, o grande avanço para a medicina genômica se encontra em determinar e compreender como a variabilidade alélica, a arquitetura genômica e a funcionalidade dos polimorfismos no DNA podem atuar sobre as características de predisposição genética a doenças complexas (PELTONEN; MCKUSICK, 2001).

Os estudos de epidemiologia genética em doenças complexas podem ser realizados basicamente de duas maneiras: por análise de ligação ou por associação genética. Análise de ligação tem por finalidade procurar por regiões cromossômicas co-herdadas com doenças em famílias, ou seja, análise da doença baseada no pedigree genético familiar. Estudos de associação genética procuram diferenças entre freqüências de alelos em uma amostra populacional constituída de

um grupo de pessoas com a doença e outro grupo sem a doença (BURTON; TOBIN; HOPPER, 2005; DAWN TEARE; BARRETT, 2005; KRUGLYAK, 1997).

Estudos de associação têm sido usados para testar a participação de genes candidatos nas doenças e refinar a posição dos genes nas regiões identificadas por análise de ligação e análises funcionais. Tais estudos podem ser conduzidos pelas estratégias direta ou indireta. Ambas consistem na hipótese de que variáveis genéticas comuns causam susceptibilidade a doenças comuns (CARDON; BELL, 2001; KRUGLYAK, 1999; LOHMUELLER et al., 2006; STEPHENS, J. C. et al., 2001).

A estratégia direta consiste em catalogar todos os locos comuns em regiões codificadoras e regulatórias dos genes, na esperança de que esta coleção contenha os polimorfismos que influenciam a susceptibilidade a doença. As freqüências destes polimorfismos seriam comparadas entre o grupo de estudo e o grupo controle, com a expectativa que um alelo de risco apresente-se mais freqüente nos indivíduos afetados (CARDON; BELL, 2001; KRUGLYAK, 1999).

A estratégia indireta evita a necessidade de catalogar potenciais locos de susceptibilidade, confiando preferivelmente na associação entre a doença e polimorfismos localizados perto de um alelo de risco (KRUGLYAK, 1999). Tais associações podem surgir em conseqüência do desequilíbrio de ligação (DL) entre o loco de risco e os polimorfismos próximos. Assim, a estratégia indireta emprega um mapa denso de marcadores polimórficos para fazer a varredura do genoma nas regiões associadas com a doença (KRUGLYAK, 1999). Os polimorfismos de base única são marcadores utilizados extensivamente nesses estudos por causa de sua alta freqüência, baixas taxas de mutação e da crescente facilidade de automatização na tipagem (BURTON; TOBIN; HOPPER, 2005; CARDON; ABECASIS, 2003; CARDON; BELL, 2001; DAWN TEARE; BARRETT, 2005; KRUGLYAK, 1997, 1999).

Polimorfismos de Base Única

Os polimorfismos de base única estão distribuídos de forma não aleatória por todo o genoma e ocorrem a uma freqüência notável de aproximadamente um em cada 1200 pares de bases (aproximadamente 1%) (LANDER et al., 2001; SACHIDANANDAM et al., 2001; SHERRY et al., 2001; VENTER et al., 2001), e, portanto, representam as variáveis mais comuns no genoma humano.

Em teoria, os SNPs podem ser de natureza polimórfica bi, tri, ou tetra alélica. No entanto, eles existem praticamente como marcadores variáveis bi-alélicos, sendo os tri e tetra alélicos mais raros ao ponto de quase não existirem (BROWN, 2002). Isso ocorre pela forma em como o SNP surge numa população. Um SNP se origina quando uma mutação pontual ocorre no genoma, convertendo um determinado nucleotídeo em outro qualquer. As fontes desse tipo de mutação englobam erros de replicação do DNA pela enzima DNA polimerase, transições (purina→purina ou pirimidina→pirimidina) e transversões (purina→pirimidina ou pirimidina→purina) geralmente causadas por alterações químicas (metilação, desaminação oxidativa e tautomeria) de origem espontânea ou induzida (BROWN, 2002). Forças evolutivas como seleção, deriva e migração modulam a fixação ou desaparecimento dessa mutação ao longo de gerações em uma população. Para um terceiro alelo surgir é preciso que uma outra mutação diferente da primeira ocorra na mesma posição do genoma, o que torna isso um evento raro. Ainda mais, é necessário que essa segunda mutação seja transmitida para as futuras gerações e que fique bem estabelecida na população. Mutações nas quais surgem três ou quatro alelos ocorrem em freqüência muito menor do que somente para dois e ainda assim a taxa de fixação de alelos na população é muito baixa (BROWN, 2002).

Devido a sua freqüência ao longo do genoma humano estima-se a existência de cerca de 3 milhões de SNPs (GABRIEL et al., 2002; LANDER et al., 2001; SACHIDANANDAM et al., 2001; SHERRY et al., 2001; VENTER et al., 2001). Porém, baseado na análise funcional de SNPs presentes em regiões codificantes e devido a degeneração do código genético, apenas uma pequena porção de todos os

SNPs (<1%) tem potencial impacto sobre as alterações físicas, estruturais e funcionais de proteínas (VENTER et al., 2001). Portanto, a grande maioria está presente em regiões não codificantes ou representam mutações silenciosas. Isto significa que é estimado em apenas milhares, e não em milhões as variações genéticas que devam contribuir com a diversidade estrutural nas proteínas humanas (VENTER et al., 2001). Ainda, as variações fenotípicas englobam SNPs em regiões importantes para o controle da expressão gênica ainda não claramente identificadas (STRANGER et al., 2005; STRANGER et al., 2007).

Por outro lado, a descoberta de que alguns desses SNPs estão diretamente ou indiretamente associados a doenças tem abastecido o interesse de pesquisas neste campo. O mapeamento de diversos SNPs é uma ferramenta reconhecidamente útil para estudos de associação genética a fim de identificar genes candidatos a doenças por estudos de ligação em famílias e em estudos de desequilíbrio de ligação em populações por análises de associação em estudos do tipo caso-controle (KRUGLYAK, 1997, 1999; RISCH; MERIKANGAS, 1996; WANG, D. G. et al., 1998). Neste sentido, os estudos de desequilíbrio de ligação estão em voga pelo fato de que associações genéticas podem ser detectadas se o loco de risco para determinada doença estiver em alto desequilíbrio de ligação com os polimorfismos à sua volta (CARDON; ABECASIS, 2003; CARDON; BELL, 2001; KRUGLYAK, 1999; WANG, D. G. et al., 1998; WANG, N. et al., 2002).

Padrões de Desequilíbrio de Ligação

Estudos de associação genética indireta utilizando polimorfismos comuns, como os SNPs, são possíveis pelo fato de que é freqüentemente predito que indivíduos que carreguem determinado alelo em um loco também carreguem alelos específicos em locos próximos. A correlação estatística deste fato é chamada de desequilíbrio de ligação (DL). A combinação particular de alelos em DL ao longo

de um segmento cromossômico de baixa recombinação, herdado de forma independente, é determinada como haplótipo (ALTSHULER et al., 2005).

Por se tratar de uma medida estatística, o DL pode assumir valores que determinam o quão freqüentemente um alelo está associado a outro, e estes são calculados de acordo com métodos estatísticos empregados para tal finalidade (DEVLIN; RISCH, 1995; HARTL; CLARK, 1997; LEWONTIN, 1964, 1995; MANIATIS et al., 2005). Os métodos estatísticos mais utilizados são baseados nas freqüências alélicas dos locos estudados e têm como componente básico, a diferença entre as freqüências de haplótipos observada e esperada pelo parâmetro de desequilíbrio de ligação, ou D, segundo equação 1 (LEWONTIN, 1964). Considerando um par de locos com alelos A/a no loco 1, e B/b no loco 2, as respectivas freqüências alélicas serão iguais a πA, πa, πB, e πb, e as freqüências haplotípicas resultantes são πAB, πAb, πaB, e πab.

$$D = \pi_{AB} \cdot \pi_{ab} - \pi_{aB} \cdot \pi_{Ab} \qquad (1A)$$

Ou simplificando a equação 1a temos:

$$D = \pi_{AB} - \pi_{A} \cdot \pi_{B} \qquad (1B)$$

Assim, as estatísticas empregadas em estudos de DL se diferenciam no modo em que empregam essa diferença. O primeiro modelo é o de desequilíbrio gamético simples, conhecido por D', que constitui na medida normalizada de D que estabelece que os alelos dos locos sejam idênticos por estado (DALY et al., 2001). O desequilíbrio gamético simples é baseado nas freqüências alélicas observadas e pode ser descrito conforme as equações 2a e 2b (LEWONTIN, 1995).

$$D' = \frac{D}{\min(\pi_A \pi_b ; \pi_B \pi_a)} \qquad \text{para D > 0} \qquad (2A)$$

$$D' = \frac{D}{\min(\pi_A \pi_B ; \pi_a \pi_b)} \qquad \text{para D < 0} \qquad (2B)$$

Já o segundo modelo, conhecido como r^2, ou ainda Δ^2, constitui a probabilidade de que dado dois alelos em um determinado loco sejam idênticos por descendência, então dois alelos em um segundo loco serão idênticos por descendência pela mesma via (SVED, 1971). A equação matemática que descreve essa relação utiliza o quadrado da equação 1 em relação ao quociente dos produtos das freqüências alélicas observadas, conforme equação 3 (HILL, 1974).

desequilíbrio de ligação, do inglês “*Solid Spine of linkage disequilibrium*” (BARRETT et al., 2005).

Este método utiliza um algoritmo que procura por uma coluna de forte DL de um marcador a outro ao longo do segmento, o que representa que o primeiro e o último marcador estão em forte DL com todos os marcadores intermediários, mas estes não necessariamente se encontram em DL entre si (BARRETT et al., 2005). Todos estes métodos são utilizados estatisticamente para determinar um valor para a associação entre os alelos e, a partir desse valor, definir um conjunto que forme o bloco haplotípico.

O estudo de desequilíbrio de ligação pelo polimorfismo de base única é, também, uma estratégia utilizada para identificar genes de susceptibilidade em doenças complexas (RISCH; MERIKANGAS, 1996; WANG, D. G. et al., 1998). A adequação desse tipo de estudo depende criticamente da existência de desequilíbrio de ligação entre alelos funcionais e marcadores nos arredores. Nas regiões de alto desequilíbrio de ligação é estatisticamente provável que um marcador qualquer forneça a mesma informação de outros marcadores ao redor e, assim, exista uma redundância de informação nos polimorfismos. Os SNPs que possuem essa características são denominados de tagSNPs, ou seja, SNPs que representam outros com grande eficiência estatística em uma região genômica de alto DL (GABRIEL et al., 2002). Desta forma as chances de se detectar uma associação com a utilização de apenas uma pequena parte desses marcadores podem ser as mesmas de quando se utilizam vários (GABRIEL et al., 2002). Ao contrário, nas regiões de baixo desequilíbrio de ligação, a baixa correlação entre marcadores significa que a região só pode ser caracterizada adequadamente pela tipagem de muitos, senão todos, os marcadores, a fim de localizar locos ligados a doença, o que torna dispendioso os estudos de associação genética (ALTSHULER et al., 2005; CARDON; ABECASIS, 2003; GABRIEL et al., 2002).

Outra vantagem dos blocos de haplótipos é que regiões de alto desequilíbrio de ligação exibem diversidade limitada (ALTSHULER et al., 2005; CARDON; ABECASIS, 2003). Assim, um pequeno número distinto de haplótipos é predito em cada bloco para a maior parte dos cromossomos nas populações em comparação ao número predito pela permutação de combinações entre eles, caso

os alelos dos diferentes locos estudados se associem ao acaso, ou seja, 2^n, em que n é o número de marcadores (CARDON; ABECASIS, 2003).

Portanto, o uso de blocos haplotípicos em estudo de associação por DL tem sido aplicado extensamente e revelado grande importância na associação de genes candidatos em doenças complexas. O avanço de tecnologias mais simples, rápidas e acuradas permite que projetos sejam feitos tanto em regiões gênicas de pequena extensão em estudos de associação indireta (CARLTON et al., 2005; FANG et al., 2005), como em estudos de associação em genoma completo (SLADEK et al., 2007; STRANGER et al., 2005).

Muito embora a recombinação tenha papel de destaque na determinação do padrão de blocos haplotípicos, outras forças como a seleção natural, a taxa de mutação e os efeitos demográficos também influenciam na extensão dos blocos haplotípicos (JORDE, 2000; PRITCHARD; PRZEWORSKI, 2001; WAKELEY; LESSARD, 2003; WANG, N. et al., 2002). Assim, a compreensão das relações entre estrutura genética definida pelos efeitos demográficos e os padrões de DL é de suma importância para o desenvolvimento de estudos de associação.

Estrutura Genética de Populações

A estruturação genética é um termo freqüentemente usado em genética de populações para designar diferenças entre duas ou mais subpopulações distintas em termos de freqüências alélicas e genotípicas. O estudo de estruturação populacional é baseado na distribuição das freqüências genotípicas e alélicas sob influência de quatro forças evolutivas: seleção natural, deriva gênica, mutação e recombinação. Além dessas forças existem outros eventos que influenciam a estrutura genética de populações, tais como migração, endocruzamento, tamanho efetivo populacional, efeito Walund, efeito fundador e efeito gargalo, entre outros.

Outro fator que pode ser citado é a distribuição espacial da população e sua relação com essas forças evolutivas a fim de explicar fenômenos como adaptação e especiação entre populações (HARTL; CLARK, 1997).

A base da genética de populações é feita na observação da variação dos polimorfismos genéticos de indivíduos, aplicados a métodos estatísticos que medem o grau de diferenciação entre e dentro de populações (HARTL; CLARK, 1997). O conceito de variação genética levou o matemático Inglês Godfrey Hardy (HARDY, 1908) e o médico Alemão Wilhelm Weinberg (WEINBERG, 1908) a desenvolverem independentemente o princípio matemático que é conhecido como Equilíbrio de Hardy-Weinberg. O princípio indica que, sob determinadas condições, após uma geração de acasalamentos aleatórios, as freqüências genotípicas em um único loco de um gene tornar-se-á fixa em um valor particular no equilíbrio. Especifica também que as freqüências de equilíbrio podem ser representadas como uma função simples das freqüências alélicas nesse loco, ou seja, o equilíbrio é a conseqüência direta da segregação de alelos na meiose dos heterozigotos (HARDY, 1908; HARTL; CLARK, 1997). O princípio de Hardy-Weinberg é uma demonstração da noção de uma população "no equilíbrio genético" e é um princípio básico da genética de populações.

No entanto, o princípio do equilíbrio de Hardy-Weinberg refere-se somente a populações mendelianas em equilíbrio, ou seja, populações infinitamente grandes, em que os cruzamentos ocorrem ao acaso (panmixia) e com a ausência de forças evolutivas (HARDY, 1908; HARTL; CLARK, 1997), o que raramente é encontrado em populações reais.

Os conceitos de estrutura populacional hierárquica foram desenvolvidos por Sewall Wright para quantificar diferenças genéticas entre subgrupos a partir dos níveis de heterozigose média e sua relação entre e dentro de populações. Essas relações foram matematicamente categorizadas e são conhecidas por Estatísticas F de Wright ou por índice de fixação (WRIGHT, 1965).

As estatísticas F em genética de populações são usadas para estabelecer uma relação do nível de heterozigose em uma população e as causas de sua redução quando comparadas ao esperado segundo o equilíbrio de Hardy-

Weinberg. Tais mudanças podem ser causadas pela deriva genética, pelo endocruzamento, pela seleção natural, mutação ou pela combinação destes. As estatísticas F, descritas e desenvolvidas por Wright, originalmente foram desenvolvidas para determinar o nível de endocruzamento (incluindo amostragem em populações naturais) e acasalamento aleatório, mas sua aplicação foi estendida para diferenciação seletiva de populações em busca dos padrões de estrutura genética entre e dentro de populações (WRIGHT, 1965).

A estratificação populacional pode ocorrer em grupos que possuem freqüências alélicas diferentes entre e dentro dos subgrupos ou em populações miscigenadas com diferentes frações de ancestralidade. A maior parte da variabilidade genética entre os humanos ocorre dentro de populações (80-90%), enquanto a menor parte da variação total é devido a diferenças entre populações (PRITCHARD; ROSENBERG, 1999; ROSENBERG et al., 2002). Portanto, a maioria

casos, algumas dessas populações, como as populações Européias (nórdicas e mediterrâneas), da África sub-Saarianas e das Américas, ainda podem ser subdivididas (ROSENBERG et al., 2002; SELDIN et al., 2006; TIAN et al., 2006; YU et al., 2002), indicando que, na pequena porção da variação entre populações existem locos gênicos com grande diferencial entre as que podem detectar estruturação populacional.

Em muitos estudos foi verificado que o desequilíbrio de ligação é extremamente variável entre e dentro de locos e também entre e dentro de populações (GABRIEL et al., 2002; REICH et al., 2001), portanto, a estruturação de populações é um fator que deve ser levado em consideração quando feito um estudo de associação genética (CARDON; BELL, 2001). Com base nisso, instituições públicas e privadas se juntaram para fundar o Consórcio Internacional de Mapas de Haplótipos, ou simplesmente HapMap.

Consórcio Internacional de Mapas de Haplótipos

O HapMap, anunciado em outubro de 2002 (WEISS; CLARK, 2002), é um ambicioso projeto criado a partir do seqüenciamento do genoma humano e teve como objetivo inicial guiar o desenho e análise de estudos de associação genética em medicina pela organização de um banco de dados público para os polimorfismos mais comuns do genoma humano (SNPs), e fornecer informações necessárias para estudos genéticos de doenças complexas (ALTSHULER et al., 2005). Além disso, o advento dessas fontes de dados no genoma abre uma nova era em genética de populações, oferecendo uma oportunidade de investigar as forças evolutivas que moldaram as variações genéticas em populações naturais (ALTSHULER et al., 2005).

O projeto tornou-se viável pela confluência dos seguintes fatos: (1) a disponibilidade da seqüência do genoma humano; (2) bancos de dados de SNPs

comuns, dos quais ensaios de genotipagem puderam ser desenvolvidos; (3) informações sobre o DL em humanos; (4) desenvolvimento de tecnologias acuradas e relativamente baratas para a genotipagem; (5) ferramentas de internet para armazenar e compartilhar dados; e (6) estruturas para direcionar tópicos associados à questões étnicas e culturais (ALTSHULER et al., 2005).

Com a iniciativa de construir e disponibilizar na internet um mapa de haplótipos baseado em um banco de genótipos de SNPs ao longo de todo genoma humano, o HapMap utilizou indivíduos de quatro populações humanas específicas, compreendendo os três grupos populacionais definidos por Rosenberg e colaboradores (2002), as quais: 90 indivíduos (30 trios) da população Iorubá de Ibadan, Nigéria (YRI); 45 japoneses de Tókio (JPT); 45 chineses de Han em Pequim (CHB); e 90 indivíduos (30 trios) de norte-americanos residentes de Utah com ancestralidade da Europa nórdica e ocidental (CEU) (ALTSHULER et al., 2005). Após três anos de intensa pesquisa, o consórcio conseguiu atingir a meta de aproximadamente um milhão de SNPs genotipados nas quatro populações. Com o uso de trios (pai, mãe e filho) e/ou métodos estatísticos computacionais foi possível estimar haplótipos em fase para cada grupo populacional (ALTSHULER et al., 2005).

O projeto HapMap, no entanto, baseia-se na premissa de que a recombinação é o fator principal na determinação do padrão de DL e atua de maneira similar em todas as populações. Porém, eventos demográficos, como efeito fundador, miscigenação e estratificação populacional possuem um importante papel na determinação do padrão de DL (CARDON; ABECASIS, 2003; CRAWFORD; CARLSON et al., 2004; PHILLIPS et al., 2003; WANG, N. et al., 2002; ZHANG et al., 2003). Estudos mostram que a transferibilidade de tagSNPs das populações do HapMap para outras é eficiente quando estas são étnicamente similares (DE BAKKER et al., 2006; GONZALEZ-NEIRA et al., 2006; GU et al., 2007; MUELLER et al., 2005; NEJENTSEV et al., 2004; WILLER et al., 2005). No entanto, a transferibilidade pode ter algum grau de comprometimento em populações miscigenadas (DE BAKKER et al., 2006) ou em populações geograficamente diferentes das do HapMap (GONZALEZ-NEIRA et al., 2006) por estas apresentarem diferentes padrões de DL. Assim, apesar da transferibilidade de tagSNPs definidos em populações de referência ser razoavelmente alta em populações similares, a

variabilidade pode não ser tão informativa em outras (GONZALEZ-NEIRA et al., 2006). Portanto, a simples transposição dos dados do HapMap para a população brasileira pode ser comprometida devido ao alto grau de miscigenação e heterogeneidade genética.

Ancestralidade Biogeográfica

Baseado na premissa de que a estrutura geográfica da população humana é compatível com dados de estrutura genética (ROSENBERG et al., 2002), existe a possibilidade de encontrar locos com diferenças de freqüência alélica entre populações continentais que possam atribuir a ancestralidade de um indivíduo, ou ainda estimar proporções de populações parentais em populações miscigenadas. Com isso, marcadores informativos de ancestralidade (AIM, do inglês *Ancestry Informative Markers*) estão sendo explorados cada vez mais com esse intuito. Um marcador informativo de ancestralidade pode ser compreendido por qualquer marcador genético que apresente diferenças de freqüências alélicas em duas ou mais populações distintas (CHAKRABORTY et al., 1992; SHRIVER et al., 1997).

Qualquer tipo de marcador molecular pode ser considerado um AIM, contanto que satisfaça a condição de diferença de freqüência alélica entre populações. Os microssatélites são potencialmente os marcadores moleculares mais informativos, devido ao seu caráter multi-alélico providenciar informações para mais de duas populações parentais ao mesmo tempo, enquanto os marcadores bi-alélicos, como os SNPs e Inserções-deleções (indels), podem fornecer informação para somente duas populações (PFAFF et al., 2004; ROSENBERG et al., 2002). A Informação de microssatélites com motivos di-nucleotídeos (a classe mais informativa de microssatélites) escolhidos aleatoriamente é em torno de cinco a oito vezes maior que SNPs selecionados ao acaso (ROSENBERG et al., 2002). No entanto, os marcadores bi-alélicos podem ter maior quantidade de informação sobre

ancestralidade que a média dos microssatélites (PFAFF et al., 2004; ROSENBERG et al., 2003).

Neste sentido, esses locos possuem a característica de apresentar grandes diferenciais de freqüências alélicas entre populações geograficamente distintas e, conseqüentemente, possuem baixos níveis de heterozigose e elevados valores de Fst (ROSENBERG et al., 2002; SHRIVER et al., 2003; SHRIVER et al., 1997). Portanto, considerando um loco bi-alélico, com os alelos A/a, dado as freqüências alélicas de duas populações (1 e 2) temos que a diferença de freqüências alélicas (δ) pode ser dada por:

$$\delta_{A12} = \left|\pi_{a1} - \pi_{a2}\right| \tag{4}$$

Assim, locos selecionados com esta base, podem permitir a identificação e classificação de populações distintas e ainda determinar percentuais de ancestralidade em populações miscigenadas (ROSENBERG et al., 2002; SHRIVER et al., 2003; SHRIVER et al., 1997).

O marcador bi-alélico ideal para distinguir uma população de outras, ou estimar proporções de ancestralidade em amostras miscigenadas, seria aquele em que um alelo se encontre totalmente fixado para uma população enquanto o outro alelo esteja fixado para as demais, fazendo com que δ=1,00 (PFAFF et al., 2004). No entanto, tais locos são relativamente raros quando se trata de duas populações parentais, e ainda mais raros quando o número de populações parentais é maior que dois (PFAFF et al., 2004). Portanto, a diferença de freqüência alélica (δ) indicada para um AIM ter conteúdo de informação para estimativas de ancestralidade é tal que $\delta \geq 0{,}60$ (HOGGART et al., 2003; ROSENBERG et al., 2003).

Um dos maiores problemas dos estudos de associação é a estratificação genética que pode ocorrer devido à utilização de indivíduos miscigenados ou de diferentes grupos continentais (PRITCHARD; ROSENBERG, 1999). A conseqüência seria a obtenção de resultados espúrios em estudos de associação genética. Em situações como esta, tem-se uma maior representação de indivíduos com ancestralidade para um determinado grupo entre os casos e, assim, alelos mais freqüentes neste grupo populacional aparecerão espuriamente

associados à doença em questão (KNOWLER et al., 1988). Na população Brasileira este fato deve ser levado em consideração para os estudos de associação genética a fim de se evitar esse erro.

Natureza Genética da População Brasileira

A população brasileira é considerada altamente miscigenada por um processo relativamente recente e recorrente (CALLEGARI-JACQUES et al., 2003; FAUCZ; PROBST; PETZL-ERLER, 2000; MARRERO et al., 2005; PARRA, F. C. et al., 2003). Apesar do povoamento imigratório do Brasil ter trazido grandes colônias de diversas partes do mundo, basicamente européias, africanas, árabes e orientais, a grande maioria se concentra em imigrações de grupos europeus de diferentes países (por exemplo: Portugal, Espanha, Itália e Alemanha) (SEYFERTH, 2000, 2002). Ademais, o processo de miscigenação ocorreu conforme um artifício sócio-antropológico de “europeização” da população brasileira no início do século XIX, condicionada à diminuição do intercruzamento das duas populações não européias mais comuns no país, as de origem Africana e Indígena. Acreditava-se, na época, que esses cruzamentos e o aumento dos cruzamentos dessas duas populações com indivíduos de origem européia, principalmente de europeus latinos, como os portugueses, espanhóis e italianos, resultariam no desaparecimento natural de africanos e índios (SEYFERTH, 2000, 2002).

Acreditava-se desta forma que a população resultante desses cruzamentos teria características fenotípicas predominantemente européias, devido à crença errônea da existência de uma raça superior (SEYFERTH, 2000, 2002). Mas, no entanto foi estruturada uma nova população altamente miscigenada, que manteve um histórico de miscigenação aleatória e recorrente durante dois séculos conseguintes.

Atualmente no Brasil, a classificação de grupo étnico ancestral é baseada no questionário de Censo Demográfico, realizado pelo Instituto Brasileiro de Geografia e Estatística (IBGE), que utiliza o critério de auto-declaração de cor de pele. No Censo Demográfico, o IBGE reconhece 5 grupos étnicos no Brasil, baseados na raça ou na cor da pele, são eles: Branco, Amarelo, Pardo (se incluem nesta categoria as pessoas que se auto-relataram como mulata, cabocla, cafuza, mameluca ou a mistura do Preto com outra cor ou a raça), Indígena, e Preto (IBGE, 2006). A distribuição desses grupos é completamente heterogênea ao longo do território brasileiro, e geralmente segue um padrão das condições histórico-sociais da população brasileira (Tabela 1).

Tabela 1 : Percentual de grupos étnicos da população brasileira por auto-declaração de cor de pele, de acordo com a nomenclatura utilizada na pesquisa nacional por amostra domiciliar de 2005.

População regional	Grupo Étnico / Cor de pele			
	Branca	Preta	Parda	Amarela ou Indígena
Brasil	49,90	6,30	43,20	0,70
Norte	24,00	3,80	71,50	0,60
Nordeste	29,50	7,00	63,10	0,30
Sudeste	58,50	7,20	33,40	0,90
Sul	80,80	3,60	15,00	0,60
Centro-Oeste	43,50	5,70	49,90	0,90

Fonte: IBGE, 2006

No contexto genético, este legado da história contribuiu para o aumento de heterogeneidade e um desbalanço nas freqüências alélicas e genotípicas entre a população resultante e as principais populações fundadoras (ALVES-SILVA et al., 2000; CARVALHO-SILVA et al., 2001; SEYFERTH, 2000, 2002). Os níveis de ancestralidade genômica na população brasileira atual têm sido investigados extensamente em pesquisas que envolvem marcadores moleculares de diversas classes (ABE-SANDES; SILVA; ZAGO, 2004; ALVES-SILVA et al., 2000; CALLEGARI-JACQUES et al., 2003; CARVALHO-SILVA et al., 2001; FERREIRA et

al., 2005; MARRERO et al., 2005; PARRA, F. C. et al., 2003; PIMENTA et al., 2006). Esses estudos mostram que a população brasileira é geneticamente heterogênea, porém com predominância européia em seus marcadores autossômicos e, ainda corroboram os dados históricos com a observação de linhagem patriarcal tipicamente Européia e matriarcal tri-parental, com grande influência de Indígenas e Africanos.

Teoricamente sabe-se que o número de demes em uma população é importante para determinar o nível de associação entre alelos, e é esperado que populações locais tenham maiores níveis de desequilíbrio de ligação na medida em que recebem novos imigrantes a cada geração (WAKELEY; LESSARD, 2003). Portanto, em teoria, acredita-se que os níveis de desequilíbrio de ligação entre alelos na população brasileira atual deva ser um produto resultante dos efeitos de migração e miscigenação principalmente, mas não exclusivamente, das populações européias, africanas e indígenas nativas (MARRERO et al., 2005; PARRA, F. C. et al., 2003; PROBST et al., 2000).

Com isso torna-se difícil a comparação de dados genéticos tanto entre a população brasileira com a européia, quanto com a africana ou indígena, pois a população brasileira possui diferentes frações de ancestralidade e assim, um padrão de desequilíbrio de ligação e diversidade haplotípica diferente das populações base (MORAES et al., 2003). Este é um dos maiores fatores que geram a necessidade do estudo comparativo com o HapMap e, por conseqüência, aprimorar os estudos de associação genética em doenças complexas na população brasileira. O trabalho desta dissertação utilizou os genes *PTPN22* e *VDR* para uma análise comparativa dos dados do HapMap com dados obtidos em uma amostra da população brasileira.

O Gene *PTPN22*

O gene da Fosfatase de Tirosina Protéica do tipo não–receptor 22 (*PTPN22*, do inglês *Protein Tyrosine Phosphatase, Nonreceptor-Type, 22*) faz parte da família das fosfatases de tirosina protéica (PTP), e também é conhecido pelos nomes Fosfatase Linfóide (LYP), Fosfatase da Tirosina Protéica do tipo não–receptor 8 (PTPN8) e Fosfatase de domínio PEST (prolina, ácido glutâmico, serina e treonina) (PEP). Sua localização se dá no cromossomo 1p13.3–p13.1 na orientação 3'-5' do DNA e sua proteína é uma fosfatase citoplasmática específica do tecido linfóide, com localização intracelular conferida pela seqüência de aminoácidos fora do sítio catalítico (PEST) (GREGERSEN et al., 2006; SIMINOVITCH, 2004).

Seu caráter específico do tecido linfóide é chave para associações de variáveis locais no gene e disfunções relacionadas à auto–imunidade, atenuando vias de sinalização integradas às respostas celulares imunológicas (SIMINOVITCH, 2004). Em 1996, Cloutier e Veillette (CLOUTIER; VEILLETTE, 1996) mostraram alta especificidade entre a quinase da tirosina citoplasmática (CSK) e PTP, especulando que as mesmas podem atuar como efetores e/ou reguladores da CSK em células T e em outras células hematopoéticas. Com isso, portas foram abertas para que outros estudos de associação genética envolvessem os genes da família das PTP com doenças auto–imunes, como diabetes tipo 1 ou artrite reumatóide.

Em 2004, Bottini e colaboradores identificaram no gene *PTPN22* um polimorfismo de base única na posição 1858, uma transição C/T, resultando na substituição do aminoácido arginina 620 para triptofano (R620W). Neste trabalho eles demonstraram uma associação do alelo de menor freqüência, 1858T, com diabetes tipo 1 em um estudo de caso-controle em coortes de Sardenha, Itália, e de Norte-americanos com descendência européia (BOTTINI et al., 2004). Em seguida, no mesmo ano, Begovich e colaboradores descreveram uma associação do mesmo alelo de menor freqüência com artrite reumatóide, outra doença de caráter auto-imune (BEGOVICH et al., 2004). O polimorfismo foi então definido com o número de referência rs2476601 no banco de dados de SNP, o dbSNP, do Centro Nacional para Informação de Biotecnologia (*National Center for Biotechnology Information* –

NCBI). Ambos os autores que primeiro descreveram a associação do alelo de menor freqüência com doenças auto-imunes, além de outros, descreveram a mutação como um potencial mecanismo que pode envolver a regulação negativa da ativação de células T (BEGOVICH et al., 2004; BOTTINI et al., 2004; LADNER et al., 2005; SIMINOVITCH, 2004; SMYTH et al., 2004).

O estudo de desequilíbrio de ligação (DL) pelo polimorfismo de base única é uma estratégia utilizada para identificar genes de susceptibilidade em doenças complexas assim como a análise de haplótipos em testes de transmissão em famílias (WANG, D. G. et al., 1998). Desta maneira, grupos de pesquisa se esforçaram em caracterizar a extensão do desequilíbrio de ligação ao longo desse gene e definir se os haplótipos mais comuns podem predispor indivíduos a doenças auto–imunes (CARLTON et al., 2005; ONENGUT-GUMUSCU; BUCKNER; CONCANNON, 2006). Um dos importantes achados destes grupos, mesmo utilizando um conjunto diferente de polimorfismos de única base, foi o fato de que o alelo de menor freqüência do SNP rs2476601, o alelo T, possui haplótipo único e significativamente associado com artrite reumatóide (CARLTON et al., 2005) e diabetes tipo 1 (ONENGUT-GUMUSCU; BUCKNER; CONCANNON, 2006), enquanto o mesmo haplótipo diferenciado apenas pelo alelo C não mostrou associação com as doenças. Além disso, e talvez o mais importante achado destas pesquisas, é o fato de outros haplótipos também apresentarem associação com essas doenças, mesmo na ausência do alelo T em rs2476601. Esse fato reflete a existência de outras possíveis variáveis no gene *PTPN22* detectadas diretamente (ONENGUT-GUMUSCU; BUCKNER; CONCANNON, 2006) ou por desequilíbrio de ligação entre as variáveis causais com os polimorfismos genotipados (CARLTON et al., 2005).

Um fato pertinente a respeito deste polimorfismo em questão e os padrões de desequilíbrio de ligação ao longo do gene *PTPN22* é a diferença da freqüência alélica entre populações de diferentes etnias. Dados do dbSNP/HapMap, reportam a presença do alelo T somente em populações caucasianas, e além disso, diversos trabalhos confirmam a ausência do alelo T nas populações Asiáticas e Africanas (BEGOVICH et al., 2004; KAWASAKI et al., 2006; MORI et al., 2005; RAY et al., 2006). Ainda nesse sentido, Gregersen e colaboradores mostraram em um

estudo de revisão que existe um gradiente biogeográfico da freqüência do alelo T em populações européias caucasianas, crescente no sentido sul–norte da Europa (GREGERSEN et al., 2006). Assim, a alta associação do alelo T do rs2476601 em populações caucasianas e a ausência, ou baixa freqüência, do polimorfismo em outras populações levanta a possibilidade de alelos de risco adicionais estarem presentes no gene *PTPN22*, ou ainda que a associação do loco rs2476601 seja reflexo direto do DL com outra variação genética que de fato é o polimorfismo associado com doenças auto-imunes (CARLTON et al., 2005; KAWASAKI et al., 2006; ONENGUT-GUMUSCU; BUCKNER; CONCANNON, 2006).

Além do loco rs2476601, outros polimorfismos do tipo SNP apresentam distribuição alélica e genotípica diferenciada entre populações de diferentes etnias, o que implica na distribuição de haplótipos diferenciada e suas respectivas conseqüências nos estudos de associação genética em populações miscigenadas.

O Gene *VDR*

O gene que codifica o receptor de vitamina D3 (VDR, do inglês *Vitamin D Receptor*) é um importante gene para o metabolismo humano. O *VDR* é um receptor hormonal intracelular que se liga especificamente à forma ativa da vitamina D3 (a 1,25-dihidroxivitamina D3 ou calcitriol) e interage com núcleo de células-alvo para produzir uma variedade de efeitos biológicos relacionados com fisiologia e expressão gênica (BAKER et al., 1988; WALTERS, 1992).

O gene *VDR* está localizado no cromossomo 12.q12-q14 (TAYMANS et al., 1999), contém 14 exons e se estende por aproximadamente 75 kilobases de DNA genômico (MIYAMOTO et al., 1997). Os Exons IA até IF codificam a região 5' não traduzida, os exons II e III codificam o domínio de ligação do DNA, e os exons FV-IX codificam a região de ligação do ligante (CROFTS et al., 1998; MIYAMOTO et al., 1997). A expressão do *VDR* humano está sobre complexo controle de

transcrição por promotores de diversos tecidos específicos (CROFTS et al., 1998). A presença de transcritos do gene já foi encontrada em mais de 30 tecidos diferentes, no entanto, as maiores concentrações estão no intestino, onde a vitamina D ajuda a aumentar a absorção de cálcio e fósfor

códon de iniciação de tradução, e por conseqüência leva a formação de uma proteína mais curta em três aminoácidos. Outro polimorfismo no promotor 1e do *VDR* foi encontrado por análise de seqüenciamento de um sítio de ligação de um fator de transcrição específico do intestino chamado Cdx-2 (FANG et al., 2003; YAMAMOTO et al., 1999). Estes cinco polimorfismos são considerados os mais importantes em estudos de associação entre variadas características fenotípicas relacionadas com metabolismo ósseo e o gene *VDR* (ARAI et al., 2001; FANG et al., 2005; FARACO et al., 1989; MOTOHASHI et al., 2003; NEJENTSEV et al., 2004; UITTERLINDEN et al., 2004; YAMAMOTO et al., 1999; ZMUDA; CAULEY; FERRELL, 2000; ZMUDA; SHEU; MOFFETT, 2005).

Entretanto, grande parte das associações observadas é inconsistente entre diversos estudos. Por exemplo, estudos que analisam a interação entre atividade física, os polimorfismos do *VDR* na região 3'UTR e densidade mineral óssea têm apresentado resultados conflitantes. Alguns autores sugerem que uma resposta mais favorável para os genótipos do ApaI G/G e TaqI A/A (KITAGAWA et al., 2001), outros não encontraram associação entre os genótipos dos polimorfismos ApaI (JARVINEN et al., 1998) ou BsmI (RABON-STITH et al., 2005), enquanto outros descreveram respostas atenuadas para o genótipo C/C do BsmI (BELL et al., 2001). Ainda, outros estudos mostram resultados divergentes para outras características fenotípicas (ECCLESHALL et al., 1998; HAUACHE et al., 1998; SPOTILA et al., 1996; UITTERLINDEN et al., 1996; ZMUDA et al., 1999). As razões usadas para explicar essas diferenças podem incluir fatores ambientais não controláveis, estratificação populacional, heterogeneidade genética e padrões de desequilíbrio de ligação em estudos que muitas vezes são conduzidos com baixo poder estatístico (FANG et al., 2005; IOANNIDIS, 2003; SHEN et al., 2005). No entanto, essas divergências sugerem que esses SNPs podem conferir deferentes respostas por serem meros marcadores em alto DL com a verdadeira variável causal (FANG et al., 2005; NEJENTSEV et al., 2004).

Atualmente é possível localizar vários polimorfismos *in silico* com o uso e auxílio de ferramentas de bioinformática e bancos de dados como o dbSNP e o HapMap, que contém diversas informações acerca de genótipos em diversas populações. A avaliação de blocos haplotípicos no gene contribui com esses

estudos devido as maiores chances de captar uma associação verdadeira pela maior densidade de SNPs e um mapeamento mais fino do gene (FANG et al., 2005; NEJENTSEV et al., 2004; UITTERLINDEN et al., 2004).

Em estudos de larga escala, com genotipagem de diversos SNPs, pesquisadores demonstraram que o gene *VDR* contém entre três e oito blocos de haplótipos, dependendo dos marcadores e das populações utilizadas (FANG et al., 2005; NEJENTSEV et al., 2004). As principais regiões onde os blocos estão localizados são a região intergênica do *VDR* e *COL2A1*, a região promotora do *VDR* (onde se localiza o polimorfismo Cdx-2), a região codificadora (onde se localizam os polimorfismos FokI, BsmI, ApaI, e TaqI) e a região 3' não traduzida do gene (3'UTR). Um ponto em comum comprovado nesses estudos em diversas populações, é o fato do polimorfismo FokI não estar em DL com nenhum outro SNP e por isso ser considerado um ponto quente de recombinação. Os polimorfismos da região 3'UTR, o BsmI, ApaI, e TaqI, apresentam grande DL entre eles e geralmente, em estudos de associação genética, são utilizados ou separadamente, ou em conjunto, ou em pares, (FANG et al., 2005; MORRISON et al., 1994; NEJENTSEV et al., 2004; UITTERLINDEN et al., 1996). Nestes casos, o bloco haplotípico formado por esses três SNPs apresentam maior resposta à associação do que os SNPs propriamente ditos e o bloco haplotípico que concerne a região codificadora e a região 3'UTR possui maior resposta (FANG et al., 2005; UITTERLINDEN et al., 2001).

Em populações de diferentes etnias, os fenótipos relacionados com metabolismo ósseo, como osteoporose e densidade mineral óssea possuem correlação com a proporção de ancestralidade genética (BARRETT-CONNOR et al., 2005; BONILLA et al., 2004; GONG et al., 2006; LEI et al., 2006; MARICIC, 2005). No gene *VDR* essas diferenças são enfatizadas nos polimorfismos associados principalmente em populações européias e asiáticas (FANG et al., 2005; FANG et al., 2007; GONG; HAYNATZKI, 2003; LEI et al., 2006; UITTERLINDEN et al., 1996; UITTERLINDEN et al., 2001), mas também em populações de origem africana (GONG et al., 2006; UITTERLINDEN et al., 2004; ZMUDA et al., 1999) e outras populações com histórico de miscigenação, incluindo a brasileira (GENTIL, 2006; GROSS et al., 1996; HAUACHE et al., 1998; LAZARETTI-CASTRO et al., 1997; LIMA, 2006; MAISTRO et al., 2004; RAMALHO et al., 1998).

OBJETIVOS

Objetivo Geral

Estudar os padrões de desequilíbrio de ligação nos genes PTPN22 e VDR em uma amostra da população brasileira caracterizada quanto a sua ancestralidade biogeográfica e comparar com os dados do HapMap.

Objetivos Específicos

Identificar a estrutura genética da população Brasileira em indivíduos não-relacionados utilizando marcadores autossômicos informativos de ancestralidade;

Avaliar o poder de informação dos marcadores autossômicos informativos de ancestralidade;

Selecionar SNPs ao longo dos genes *PTPN22* e *VDR* e desenvolver ensaios de genotipagem automatizada baseados em extensão de única base;

Verificar os padrões de desequilíbrio de ligação dos genes *PTPN22* e *VDR* nas populações alocadas no HapMap;

Genotipar SNPs em uma amostra da população brasileira;

Estimar blocos haplotípicos e avaliar o desequilíbrio de ligação para a população brasileira nos SNPs genotipados;

Avaliar a influência da ancestralidade biogeográfica nos blocos haplotípicos;

Comparar a transferibilidade de tagSNPs entre as populações brasileira e as do HapMap.

APRESENTAÇÃO DOS CAPÍTULOS

O presente trabalho foi dividido em três capítulos. No capítulo 1 é apresentada a estrutura genética da população Brasileira em indivíduos não-relacionados utilizando marcadores informativos de ancestralidade. Neste capítulo é apresentada a maior parte das informações sobre as amostras e parte das técnicas laboratoriais e ferramentas de informática utilizadas no estudo. Portanto, os dados descritos no capítulo 1 serão extensivamente usados nos demais capítulos. Nos capítulos 2 e 3 são apresentados os estudos do padrão de desequilíbrio de ligação dos genes *PTPN22* e *VDR* respectivamente, da mesma população retratada no capítulo 1.

1 Estrutura Genética da População Brasileira Avaliada por SNPs Informativos de Ancestralidade

1.1 MATERIAIS E MÉTODOS

1.1.1 Amostra Populacional

A amostragem populacional utilizada no presente estudo foi constituída de 200 indivíduos brasileiros não relacionados, que foram submetidos a exames de paternidade. As amostras foram cedidas em colaboração com os professores Dr Dario Grattapaglia e Dr Márcio Elias Ferreira. Esta amostragem inclui 40 indivíduos da população urbana (capitais e principais cidades) de cada uma das cinco regiões geográficas da Federação Brasileira. O sexo de cada indivíduo foi identificado no momento da coleta e estabelecido como: Feminino = Mãe (M) e Masculino = Suposto Pai (SP). Somente na região Sudeste foram utilizadas 17 indivíduos do sexo feminino e 23 do sexo masculino. Para as demais regiões foram utilizadas 20 amostras de cada sexo.

Uma vez que o estudo trata de uma caracterização de variabilidade genética em termos populacionais, os indivíduos que compuseram a amostragem não passaram por critérios de seleção para inclusão e exclusão de acordo com características fenotípicas. Essas amostras foram utilizadas nos demais capítulos desta dissertação.

Além da população brasileira, dados de genótipos das quatro populações alocadas no HapMap foram utilizados em todas as análises dos três capítulos desta dissertação. Como sugerido pelo consórcio (ALTSHULER et al., 2005) as populações CHB e JPT foram agrupadas para formar uma população Asiática, aqui denominada por ASN. Neste primeiro capítulo, por se tratar de um estudo de genética de populações, as análises foram realizadas com indivíduos não relacionados, ou seja, foram excluídos todos os filhos presentes nos trios das populações CEU e YRI do HapMap.

Ademais, para análise de estruturação genética foram usados dados de mais três populações depositadas pela Perlegen Sciences Inc, as quais: AFD_AFR_PANEL (AFM) – Amostra composta de 23 indivíduos Afro-Americanos selecionados a partir de um painel de variação humana de 50 Afro-Americanos (HD50AA) do Repositório de Células Coriell; AFD_CHN_PANEL (CHN) – Amostra composta de 24 indivíduos Chineses selecionados a partir de um painel de variação humana de 100 Chineses Han de Los Angeles TDo3HI(A) do 1(posit7-5.ó)1-5.(ório 5( )

Quadro 1.1 : Identificação dos indivíduos segundo o número do repositório de Células Coriell de acordo com cada amostra populacional e número de indivíduos utilizados.

C E U		Y R I		ASN-CHB	ASN-JPT	EUR	AFM	CHN
NA12144	NA11994	NA18501	NA19127	NA18524	NA18940	NA12560	NA17102	NA17733
NA12145	NA11995	NA18502	NA19128	NA18526	NA18942	NA06990	NA17103	NA17734
NA12146	NA12264	NA18504	NA19130	NA18529	NA18943	NA10848	NA17104	NA17735
NA12239	NA12234	NA18505	NA19131	NA18532	NA18944	NA07019	NA17105	NA17736
NA06994	NA12154	NA18507	NA19137	NA18537	NA18945	NA10851	NA17106	NA17737
NA07000	NA12236	NA18508	NA19138	NA18540	NA18947	NA10850	NA17107	NA17738
NA07022	NA12155	NA18516	NA19140	NA18542	NA18948	NA07349	NA17108	NA17739
NA07056	NA12156	NA18517	NA19141	NA18545	NA18949	NA07348	NA17109	NA17740
NA07034	NA12248	NA18522	NA19143	NA18547	NA18951	NA10857	NA17110	NA17741
NA07055	NA12249	NA18523	NA19144	NA18550	NA18952	NA10852	NA17111	NA17742
NA06993	NA12003	NA18852	NA19152	NA18552	NA18953	NA10858	NA17112	NA17743
NA06985	NA12004	NA18853	NA19153	NA18555	NA18956	NA10853	NA17113	NA17744
NA12056	NA12005	NA18855	NA19159	NA18558	NA18959	NA10854	NA17114	NA17745
NA12057	NA12006	NA18856	NA19160	NA18561	NA18960	NA10860	NA17115	NA17746
NA07357	NA12750	NA18858	NA19171	NA18562	NA18961	NA10861	NA17116	NA17747
NA07345	NA12751	NA18859	NA19172	NA18563	NA18964	NA10830	NA17133	NA17749
NA12043	NA12760	NA18861	NA19192	NA18564	NA18965	NA10831	NA17134	NA17752
NA12044	NA12761	NA18862	NA19193	NA18566	NA18966	NA10842	NA17135	NA17753
NA11881	NA12762	NA18870	NA19200	NA18570	NA18967	NA10843	NA17136	NA17754
NA11882	NA12763	NA18871	NA19201	NA18571	NA18968	NA10845	NA17137	NA17755
NA11839	NA12812	NA18912	NA19203	NA18572	NA18969	NA10844	NA17138	NA17756
NA11840	NA12813	NA18913	NA19204	NA18573	NA18970	NA17201	NA17139	NA17757
NA11829	NA12814	NA19092	NA19206	NA18576	NA18971	NA12547	NA17140	NA17759
NA11830	NA12815	NA19093	NA19207	NA18577	NA18972	NA12548		NA17761
NA11831	NA12872	NA19098	NA19209	NA18579	NA18973			N
NA11832	NA12873	NA19099	NA19210	NA18582	NA18974			N
NA12716	NA12874	NA19101	NA19222	NA18592	NA18975			N
NA12717	NA12875	NA19102	NA19223	NA18593	NA18976			N
NA11992	NA12891	NA19116	NA19238	NA18594	NA18978			N
NA11993	NA12892	NA19119	NA19239	NA18603	NA18980			N
				NA18605	NA18981			
				NA18608	NA18987			
				NA18609	NA18990			
				NA18611	NA18991			
				NA18612	NA18992			
				NA18620	NA18994			
				NA18621	NA18995			
				NA18622	NA18997			
				NA18623	NA18998			
				NA18624	NA18999			
				NA18632	NA19000			
				NA18633	NA19003			
				NA18635	NA19005			
				NA18636	NA19007			
				NA18637	NA19012			

Fonte: O Autor

1.1.2 Extração de DNA

O DNA genômico das amostras da população brasileira foi extraído dos leucócitos periféricos com modificações do método *salting out* (MILLER; DYKES; POLESKY, 1988). Primeiramente a quebra das membranas celulares foi realizada por meio da adição do Tampão A, composto por sacarose a 0,32 M, Tris-HCl pH 7,6 a 10 mM, $MgCl_2$ a 5 mM e detergente não iônico Triton X 100 (1%). Após a homogeneização do sangue, um volume inicial de 750 µL foi depositado em um microtubo de 1,5 mL. Adicionou-se ao sangue 750 µL do Tampão A, e o material em seguida foi centrifugado a 2.500 rpm por 20 minutos para condensação do pellet e posterior descarte do sobrenadante.

Para remoção das proteínas celulares e histonas ligadas ao DNA, o pellet foi suspenso em um Tampão B composto de 25mM de EDTA pH 8,0 e 75mM de NaCl, dodecil sulfato de sódio a 10% e proteinase K (10 mg/mL). Os tubos foram então incubados a 37ºC durante uma hora para ação da enzima. Em seguida foi adicionado 6,0 M de NaCl para precipitar impurezas protéicas.

Os tubos foram centrifugados e em seguida o sobrenadante foi transferido para um novo tubo de 2,0 mL e a ele foi adicionado o dobro de volume de etanol absoluto. Os tubos foram fechados e invertidos para precipitar o DNA. Outra centrifugação foi realizada com a finalidade de aderir o DNA ao fundo dos tubos, seguida de descarte do sobrenadante. Os tubos foram incubados a 37ºC durante uma hora para a evaporação do etanol e depois foram adicionados 300 µL de Tris-EDTA para eluição e estocagem do DNA.

Para utilização do material nas reações de PCR as amostras foram quantificadas em gel de agarose 1,0% corado com brometo de etídio (1 mg/mL) utilizando padrões λ de DNA de concentrações conhecidas e visualizado em luz ultravioleta. A partir da concentração original, as amostras foram diluídas para a concentração de trabalho de 10 ng/µL.

1.1.3 PCR e Genotipagem

Os marcadores utilizados neste estudo foram previamente selecionados na literatura a partir das diferenças em freqüências alélicas (δ) das populações parentais Européia, Africana e Ameríndia (SHRIVER et al., 2003; SMITH et al., 2004). Foram selecionados 33 SNPs informativos de ancestralidade que foram agrupados em três sistemas, para a amplificação em multiplex por meio de PCR e ensaio de extensão de única base, além de um marcador informativo de ancestralidade do tipo inserção-deleção (AT3) que pôde ser amplificado por PCR e visualizado em gel de agarose. A Tabela 1.1 descreve as freqüências alélicas para cada um dos 34 locos nas três populações parentais. Os iniciadores para amplificação dos locos (Tabela 1.2) foram desenhados previamente (PEDROSA, 2006) de acordo com as indicações do manual de SNaPshot™ Reaction (Applied Biosystems).

A amplificação da PCR para os locos individuais foi realizada em um volume de reação final de 12,5 µL contendo de 10 a 50 ng do DNA molde; 0,25 µM de cada iniciador (direto e reverso) em multiplex; 0,25 µM de dNTP mix (ATGC); 1,5 µM de MgCl2; 0,16 mg/mL de albumina sérica bovina (BSA); 1 unidade (U) de Platinum Taq DNA Polimerase (Invitrogen); 1 X de Tampão Platinum Taq (Invitrogen) e H_2O Milli-Q qsp.

Para a amplificação dos locos (inclusive AT3) foi usada a condição de termociclagem do programa *Two-step Touchdown PCR*, que consiste na fusão dos métodos *Two-step Gradient PCR* (LOPEZ; PREZIOSO, 2001) e *Touchdown PCR* (DON et al., 1991) da seguinte maneira: 3 minutos de desnaturação a 94ºC, 20 ciclos de 15 segundos a 94ºC seguidos de 30 segundos a 65ºC decrescendo até 55ºC (-0.5ºC por ciclo). Em seguida, mais 10 ciclos a 94ºC por 15 segundos e 55ºC por 30 segundos e 5 minutos de extensão final a 72ºC.

Tabela 1.1 : Descrição dos 34 locos de acordo com suas posições gênicas e físicas, seus alelos e as freqüências alélicas do alelo 1 nas respectivas populações parentais, Européia (EUR), Africana (AFR) e Ameríndia (AMR).

	Posição	Posição	Alelos		Freqüência do Alelo 1		
Loco	gênica	física	1	2	EUR	AFR	AMR
WI11153 (rs17203)	3p12	77070188	G	C	0,171	0,785	0,805
MID93 (rs16383)	22q13	40378198	A	T	0,220	0,739	0,895
rs1426654	15q21	46213776	C	T	0,000	0,980	0,950
TSC1102055 (rs2065160)	1q32	203057600	C	T	0,078	0,512	0,850
rs4305737	6q24	145093287	A	G	0,250	0,929	1,000
rs727563	22q13	40197323	C	T	0,260	0,820	0,950
rs734780	15q26	87365962	C	T	0,070	0,710	0,854
rs730570	14q32	100212643	A	G	0,860	0,185	0,100
rs1129038	15q13	26030454	C	T	0,224	0,995	0,983
rs1240709	1p36.3	1327700	A	G	0,794	0,036	0,103
rs3796384	3p14	64543452	C	G	0,154	0,783	0,875
rs2278354	5p15.2	10487111	G	T	0,120	0,704	0,839
AT3 (rs3138521)	1q25	172153366	i	d	0,282	0,858	0,061
CRH (rs3176921)	8q13	67253933	G	A	0,073	0,682	0,017
CYP3A4 (rs2740574)	7q22	99220032	G	A	0,042	0,802	0,041
FYNULL (rs2814778)	1q23	157441307	C	T	0,998	0,001	1,000
LPL (rs285)	8p21	19859469	G	A	0,508	0,029	0,558
OCA2 (rs1800404)	15q13	25909368	G	A	0,254	0,885	0,552
RB (rs2252544)	13q14	47776293	C	T	0,315	0,926	0,175
rs1480642	6q23	136541221	C	T	0,994	0,106	0,621
rs6034866	20p12	17551728	G	A	0,917	0,051	0,857
rs7349	10p11.2	31857911	G	A	0,939	0,016	1,000
rs803733	9q33	124872700	C	T	0,880	0,015	0,411
rs1871534	8q24.3	145610489	C	G	0,981	0,040	1,000
rs222541	6q16	95287884	G	C	0,743	0,036	0,982
rs267071	5q22	116976402	C	T	0,654	0,082	1,000
rs310612	20q13.3	61651952	G	A	0,237	1,000	0,034
rs3768641	2p13	72221698	G	C	0,077	0,990	0,000
rs3780293	9q21	79266067	G	A	0,286	0,985	0,017
rs3791896	2q35	218520326	G	A	0,234	1,000	0,086
rs4280128	13q22	74876840	G	A	0,643	0,041	0,966
rs4766807	12q24.2	115785834	A	T	0,622	0,030	0,948
rs730086	17q21	37525283	C	T	0,661	0,067	1,000
rs736556	7p15	31023593	C	T	0,244	0,939	0,018

Fontes: SHRIVER, 2003; SMITH, 2004

Tabela 1.2 : Seqüência e tamanho dos iniciadores para PCR e para a extensão de base única e seus respectivos arranjos em multiplex.

Primer	Seqüência	Direção do primer	Tamanho do primer	Multiplex
AT3-F	GCCTGAAGGTAGCAGCTTGT	direto	20	–
AT3-R	CCCACACTCCCTCACTCTTC	reverso	20	
TSC1102055-F	CTGCTGTGCTAGCTGCTGAT	direto	20	
TSC1102055-R	GCTGTGAGGACGTCAAACCT	reverso	20	1
TSC1102055-S	gactCCTCTCGATGAGTAAATATGGG	reverso	26	
WI11153-F	CGTTGGGAATATTTCTATCTCACCT	direto	25	
WI11153-R	CATCTTGCTGACAACTTAAATATGC	reverso	25	1
WI11153-S	ATCCAACAGTCAAGGTCTTC	reverso	20	
MID93-F	TGACATGACCCCAGTTACTAATGT	direto	24	
MID93-R	CGTTGTGTTTATTTGTGCAGTC	reverso	22	1
MID93-S	ATGAGCCATAGCACAAAAGA	direto	20	
rs3796384-F	GCCAATGTCGGAAGGATTAC	direto	20	
rs3796384-R	GCTAGCCAATGTGCAAGACA	reverso	20	1
rs3796384-S	$(gact)_6$CGTTCTTCTCTCCATTCAGA	direto	44	
rs2278354-F	GCTCCGTGACCCACTTTCTA	direto	20	
rs2278354-R	GCCTCTGCCTTCTCTGTCTG	reverso	20	1
rs2278354-S	$(gact)_7$AGAAGCGCAAGGCAGGAAGG	direto	48	
rs4305737-F	TGGTGAACACGTGAGGTTACA	direto	21	
rs4305737-R	TGGAGAAACCAGTCTCACCTG	reverso	21	1
rs4305737-S	$(gact)_3$AATTGAGGCCCCTGAAGA	direto	30	
rs1426654-F	TTCAGCCCTTGGATTGTCTC	direto	20	
rs1426654-R	AATTGCAGATCCAAGGATGG	reverso	20	1
rs1426654-S	gactGACCGCTGCCATGAAAGTTG	reverso	24	
rs730570-F	GCCTTCCATGGTTTCTCTGA	direto	20	
rs730570-R	AGATTGTGGGGACTGTGAGC	reverso	20	1
rs730570-S	$(gact)_4$TCACCTGCATCTCACACTGC	direto	36	
rs727563-F	CACGGTATCCAGAACAAGCA	direto	20	
rs727563-R	ACACTGCCTCCCAATAACCA	reverso	20	1
rs727563-S	$(gact)_3$ACCAGGCTGTCTCAAATAAC	reverso	32	
rs734780-F	GATGGCACTGACCTTCCTTC	direto	20	
rs734780-R	AGGTTGCAGTGAGCCAAGAT	reverso	20	1
rs734780-S	$(gact)_4$CCCAGCAGTGGGTATCAC	direto	34	
rs1129038-F	CAGCAGCGACGATTCAGATA	direto	20	
rs1129038-R	ATCACGGCCAGTCAGTCTCT	reverso	20	1
rs1129038-S	(gact)5ACAGTCTACACAGCAGCGAG	reverso	40	

Tabela 1.2 Continuação

Primer	Seqüência	Direção do primer	Tamanho do primer	Multiplex
rs1240709-F	ATCCTATCTGGGTGGCACAG	direto	20	
rs1240709-R	CAGCAGTCAGCTCAGTCAGG	reverso	20	1
rs1240709-S	$(gact)_5$ATGTGGACACGGGTGAGGGA	direto	40	
CYP3A4-F	CTGGGTTTGGAAGGATGTGT	direto	20	
CYP3A4-R	TGTTACTGGGGAGTCCAAGG	reverso	20	2
CYP3A4-S	CAGCCATAGAGACAAGGGCA	direto	20	
FYNULL-F	TCACCCTGTGCAGACAGTTC	direto	20	
FYNULL-R	GTGGGGTAAGGCTTCCTGAT	reverso	20	2
FYNULL-s	$(gact)_2$gacCTCATTAGTCCTTGGCTCTTA	reverso	32	
CRH-F	TTTGTGCCCCTTCACTATGG	direto	20	
CRH-R	CCATCTTTCTGCCTGGAAAA	reverso	20	2
CRH-S	$(gact)_3$TGCAGAAGCAAGGCCAATAA	reverso	32	
LPL-F	CAGTGGGTTCAAGGCTCTGT	direto	20	
LPL-R	AACAACAACAAAACCCCACA	reverso	20	2
LPL-S	$(gact)_4$gACAACAAAACCCCACAGCT	reverso	36	
OCA2-F	CAGGCTTTCGTGTGTGCTAA	direto	20	
OCA2-R	TGAGCTGACATCCCACTGAG	reverso	20	2
OCA2-S	$(gact)_2$gGTGCACAGAACTCTGGC	direto	26	
RB-F	GTCAAGTTGAAGCCGAGACC	direto	20	
RB-R	GTCAGGTGAGCGAGCAGAG	reverso	19	2
RB-S	CCCGCCTCCTCCCGCCGGGA	direto	20	
rs1480642-F	TTCTTGACCTGAGTGGTGGTT	direto	21	
rs1480642-R	CAAACCAGTGGGCAAGAGAT	reverso	20	2
rs1480642-S	$(gact)_4$TTATATGTGAGGGAAAGCTC	direto	36	
rs7349-F	GCAATTGGTTCTCCTGCATT	direto	20	
rs7349-R	GAAATGAGAGTTGTATGGTTAGGC	reverso	24	2
rs7349-S	$(gact)_5$AAATGAGAGTTGTATGGTTAGGCT	reverso	44	
rs803733-F	TCCCCAAGAGTTCAACCAAC	direto	20	
rs803733-R	AACCTTAGGCTTGAGCATGG	reverso	20	2
rs803733-S	$(gact)_7$ATGTCATTGTGGAGGAGATA	reverso	48	
rs6034866-F	TTGTGAGTCAAGGCAAGCTG	direto	20	
rs6034866-R	TAGCTAGGGCAGGAGGTGAA	reverso	20	2
rs6034866-S	$(gact)_7$TGTGAGTCAAGGCAAGCTGG	direto	48	

Tabela 1.2 Continuação

Primer	Seqüência	Direção do primer	Tamanho do primer	Multiplex
rs3768641-F	GGGGTGTATCTGATGGATGG	direto	20	
rs3768641-R	GTAGAGCTGGGCACAGGAAG	reverso	20	3
rs3768641-S	TGGAGAAGGAGCTAGAGAAT	direto	20	
rs4766807-F	ACTCGGGCCCTAATGGATAC	direto	20	
rs4766807-R	CAGAGCGAGACCCAGTTTCT	reverso	20	3
rs4766807-S	CTTGTGACTGTCAGCTGACAT	reverso	21	
rs3791896-F	GCACTGCCATCTCCTTTCTC	direto	20	
rs3791896-R	TGCTCCTCTCCTTGTCCCTA	reverso	20	3
rs3791896-S	gactAGGGAGAAAAGAGGTGAGGAGA	reverso	26	
rs267071-F	TGGGACTTGAGCATTATAGGG	direto	21	
rs267071-R	TGCTAATGCAGCCTTCTCAA	reverso	20	3
rs267071-S	gactCTTTAATGCTCTGAAAACCTTA	direto	26	
rs730086-F	GAGAACACTGGGGAGGTTCA	direto	20	
rs730086-R	CAAAGTTCAGCACACCCTGA	reverso	20	3
rs730086-S	$(gact)_3$CACCTCATTCCTGGTTTTAT	direto	32	
rs310612-F	CACCCCTTCCTCACCTCAC	direto	19	
rs310612-R	TTTCCTCACTCCCCTCTGTG	reverso	20	3
rs310612-S	$(gact)_4$CAGGCGCAACACAGCCCAGC	direto	36	
rs3780293-F	GAGAACACTGGGGAGGTTCA	direto	20	
rs3780293-R	CAAAGTTCAGCACACCCTGA	reverso	20	3
rs3780293-S	$(gact)_4$CTGCGCCTTAGCAAGATCCC	direto	36	
rs736556-F	CTCTGCTTCTGGTTCCTTGC	direto	20	
rs736556-R	AGAACCAGCCGAAACTGAAA	reverso	20	3
rs736556-S	$(gact)_3$TGGTTCCTTGCAGATTACCAAATT	direto	36	
rs222541-F	TGCTAAAGCACTTCAGTTTTGA	direto	22	
rs222541-R	AATCCCCTAAGGCAGCTTTC	reverso	20	3
rs222541-S	$(gact)_4$CATTCATAACTGACTGATAGCATA	direto	40	
rs1871534-F	GGGCTGAGTCTGGAAGAAAA	direto	20	
rs1871534-R	GACCTTGGGCGTCAGATG	reverso	18	3
rs1871534-S	$(gact)_6$CAGGACAGCGTCCCCAGTGA	reverso	44	
rs4280128-F	TATTTGGGTAGCGAGGGACT	direto	20	
rs4280128-R	ATTTCTCTGCATGGGTGGAG	reverso	20	3
rs4280128-S	$(gact)_7$ATTATATTTGGGTAGCGAGGGACT	direto	52	

Fonte: PEDROSA, 2006

Após a amplificação, foi realizada uma purificação enzimática para degradação do excesso de iniciadores e de dNTPs em 3,0 µL do produto de PCR com uma mistura de 1,0 U da enzima exonuclease 1 (EXO1, 10 U/µL), 0,90 U de fosfatase alcalina de camarão (SAP, 1 U/µL) e 0,5 µL de tampão 10X da enzima EXO1, durante 90 minutos a 37ºC seguido de 20 minutos a 80ºC para desnaturação das enzimas.

Posteriormente, a genotipagem dos SNPs foi realizada utilizando o sistema SNaPshot™ Multiplex System (Applied Biosystems), com a reação em volume final de 5,0 µL contendo 1,0 µL do amplicon purificado, 1,0 µL de SNaPshot™ Kit Reaction Mix, multiplex de iniciadores SNP em concentrações variadas (Tabela 1.3) e H_2O Milli-Q qsp. A termociclagem foi feita com alteração no protocolo sugerido pelo fabricante. A mudança foi feita com a adição de ciclos em *touchdown* a partir de uma temperatura de anelamento mais elevada que a sugerida inicialmente. O programa de termociclagem, denominado *snapdown* adotou o seguinte protocolo: 2 minutos de desnaturação a 96ºC, seguido de 30 ciclos de 96ºC por 15 segundos, anelamento a 55ºC durante 20 segundos com decaimento de 0,5ºC por ciclo e extensão a 60ºC durante 30 segundos; em seguida 15 ciclos de desnaturação a 95ºC durante 15 segundos, anelamento a 40ºC por 20 segundos e extensão a 60ºC por 30 segundos.

Finalmente, uma nova purificação foi realizada para remoção dos grupos fosforil das extremidades 5’ dos ddNTPs fluorescentes. Esta foi feita com 0,5 U de SAP e 0,5 µL de tampão 10X da SAP adicionados diretamente às reações, que em seguida foram incubadas durante 60 minutos a 37ºC e 15 minutos a 75ºC para desnaturação da enzima.

A eletroforese foi realizada no seqüenciador automático ABI 3100 (Applied Biosystems) com 1,0 µL do amplicon final, 8,9 µL de formamida Hi-Di e 0,1 µL de Liz120 *Size Standard* (Applied Biosystems). Os parâmetros de corrida foram: Polímero ABI3700 POP 6, com tensão de 3 volts e 40 segundos de injeção e 15 volts de corrida por 1000 segundos. As corridas foram analisados com os programas GeneScan Analysis 3.7 e os eletroferogramas foram analisadas por meio de macros automatizadas para cada multiplex de reação no programa Genotyper 3.7 (Applied Biosystems).

1.1.4 Análises Estatísticas

Para estudos de análises genéticas, o cálculo das freqüências alélicas foi realizado a partir dos dados genotípicos gerados, utilizando o programa GenAlEx (PEAKALL; SMOUSE, 2006). O teste exato de Fischer para as freqüências genotípicas quanto ao Equilíbrio de Hardy-Weinberg, a análise de variância para as estatísticas F de Wright e coeficientes de distância e similaridade par a par entre as populações foram realizadas pelo programa GDA - Genetic Data Analysis (LEWIS; ZAYKIN, 2001).

O programa Structure 2.1 (PRITCHARD; STEPHENS; DONNELLY, 2000) foi usado para avaliar a existência de estruturação populacional. Este programa modela a estrutura populacional, miscigenação populacional e miscigenação individual utilizando inferência Bayesiana a partir dos dados de genótipos, portanto, o programa não requer que o usuário especifique o número de populações parentais, tampouco as freqüências alélicas das mesmas. Assim, o número de populações (K) parentais presentes em uma amostragem em função do seu grau de estruturação genética ou populações ancestrais são estimados a partir dos dados fornecidos. Deste modo, o programa foi rodado com quatro tipos de análises: (A1) utilizando os dados de genótipos das seis populações (três do HapMap e três do Perlegen) recuperados no dbSNP; (A2) utilizando os dados de genotipagem da população brasileira; (3) utilizando os dados da população brasileira juntamente com as populações do HapMap e do Perlegen; e (4) utilizando os dados da população brasileira juntamente com as populações de origem Européia e Africana do HapMap e do Perlegen, somente para os marcadores que possuíssem genótipos nessas populações e na população brasileira, ou seja, foram excluídos os locos que possuíam dados somente na população brasileira. Os locos que possuíam dados somente para uma das populações não foram retirados da análise porque o programa trata dados faltantes de forma sistemática por estatística bayesiana.

O programa ainda oferece uma opção para designar populações previamente conhecidas como forma de melhorar as estimativas de análise. Desta forma, cada região brasileira foi definida como uma subpopulação *a priori*, assim

como as subpopulações EUR, AFM, CHN, CEU, YRI e ASN. O programa foi rodado com 50.000 interações para o *burn-in period* e 10.000 repetições extras, com K variando de 1 a 10 repetidas por 10 vezes. Corridas adicionais com diferentes interações (35.000/35.000 e 50.000/35.000) foram realizadas para checar a consistência dos resultados. Em todos os cálculos utilizou-se o modelo de miscigenação com freqüências alélicas relacionadas entre populações. Para inferir o melhor K foi utilizada uma estatística *ad hoc*, DeltaK, baseada na taxa de mudança do logaritmo da probabilidade dos resultados entre sucessivos valores de K (EVANNO; REGNAUT; GOUDET, 2005).

Primeiramente a estatística DeltaK utiliza a média dos valores de ln P(D), denominado L(K), gerados no Structure para fazer uma primeira inferência. Em seguida, esses valores são corrigidos para a diferença entre K e K-1, chamado de L'(K). Posteriormente é tomado o módulo da diferença de L'(K) para K+1 e K, chamado de L"(K). Dividindo L"(K) pelo produto do desvio padrão das médias de L(K) e o número de repetições realizadas obtém-se DeltaK. Por levar em consideração K-1 e K+1 nos cálculos, DeltaK não é capaz de inferir ausência de estrutura e nem identificar K, se o mesmo for maior que o número máximo proposto (EVANNO; REGNAUT; GOUDET, 2005).

Os marcadores que possuíam genótipos para as populações do HapMap e do Perlegen foram testados para o conteúdo de informação para inferência de ancestralidade pelas estatísticas de quantidade de informação para atribuição de ancestralidade (In), quantidade de informação para coeficiente de ancestralidade (Ia) e taxa ótima de atribuição correta (ORCA) utilizando um script escrito em Perl, gentilmente cedido pelo Dr. Noah Rosenberg (ROSENBERG et al., 2003). Estas estatísticas são aplicáveis a um potencial número de populações parentais para determinar a quantidade e o conteúdo de informação que os marcadores podem fornecer sobre a ancestralidade individual, e são baseadas na diferença de freqüências alélicas (δ) e na quantidade de populações parentais (K) (ROSENBERG et al., 2003).

1.2 RESULTADOS

1.2.1 Recuperação de Genótipos no dbSNP

Foi possível acessar os genótipos das seis amostras populacionais utilizadas (CEU, YRI, ASN, AFM, CHN e EUR) por meio do portal eletrônico do dbSNP para a maioria dos locos estudados. De todos os 34 locos selecionados para o estudo apenas 26 foram recuperados, e destes, apenas oito para todas as populações (Quadro 1.2). Um indivíduo da amostra ASN-JPT foi excluído das análises por apresentar genótipo para somente quatro locos (amostra NA19012).

Quadro 1.2 : Genótipos recuperados em cada população do dbSNP.

	População						
Loco	EUR	AFM	CHN	CEU	ASN-CHB	ASN-JPT	YRI
rs222541	V	V	V	V	V	V	V
rs3176921	V	V	V	V	V	V	V
rs3780293	V	V	V	V	V	V	V
rs3791896	V	V	V	V	V	V	V
rs4280128	V	V	V	V	V	V	V
rs4766807	V	V	V	V	V	V	V
rs730570	V	V	V	V	V	V	V
rs803733	V	V	V	V	V	V	V
rs1800404	V	V	V	V	V		V
rs1240709	V	V	V				
rs2065160	V	V	V				
rs3796384	V	V	V				
rs736556	V	V	V				
rs310612							V
rs1426654				V	V	V	V
rs1480642				V	V	V	V
rs1871534				V	V	V	V
rs2278354				V	V	V	V
rs267071				V	V	V	V
rs2740574				V	V	V	V
rs3768641				V	V	V	V
rs734780				V	V	V	V
rs7349				V	V	V	V
rs285				V	V		V
rs727563				V	V		V
rs4305737				V	V		

Fonte: O Autor

1.2.2 Reação de PCR e Genotipagem

Os sistemas multiplex permitiram uma tipagem rápida e eficiente dos locos SNP em apenas uma reação de PCR seguida de tratamento enzimático e reação de SNaPshot™ para cada sistema. A concentração dos iniciadores utilizados na reação SNaPshot™ foi diferente para cada loco (Tabela 1.3) a fim de manter equivalente o balanço de intensidade entre os picos em unidades relativa de fluorescência (rfu, do inglês, *relative fluorescence unit*) (Figura 1.1). A Tabela 1.4 indica o tamanho esperado e observado dos picos de cada alelo amplificado para todos os locos. A utilização de termociclagem *touchdown* tanto na PCR quanto na reação de SNaPshot™ foi necessária para evitar o comprometimento da qualidade da amplificação com o aparecimento de produtos inespecíficos que dificultariam a análise. No entanto, o aparecimento de artefatos e picos não esperados foi observado em algumas reações (Figura 1.2).

Tabela 1.3 : Concentração dos iniciadores de extensão de base única na reação SNaPshot™ para cada sistema multiplex.

Multiplex AIM1		Multiplex AIM2		Multiplex AIM3	
Loco	Concentração (μM)	Loco	Concentração (μM)	Loco	Concentração (μM)
WI11153	0,60	CRH	0,40	rs1871534	0,15
MID93	0,80	CYP3A4	0,60	rs222541	0,60
rs1426654	0,50	FYNULL	0,05	rs267071	0,50
TSC1102055	0,50	LPL	0,20	rs310612	0,90
rs4305737	0,50	OCA2	0,70	rs3768641	0,15
rs727563	0,50	RB	0,60	rs3780293	0,15
rs734780	0,25	rs1480642	0,35	rs3791896	0,90
rs730570	0,50	rs6034866	0,05	rs4280128	0,15
rs1129038	0,50	rs7349	0,15	rs4766807	0,15
rs1240709	0,25	rs803733	0,10	rs730086	0,35
rs3796384	0,25			rs736556	0,15
rs2278354	0,25				

Fonte: O Autor

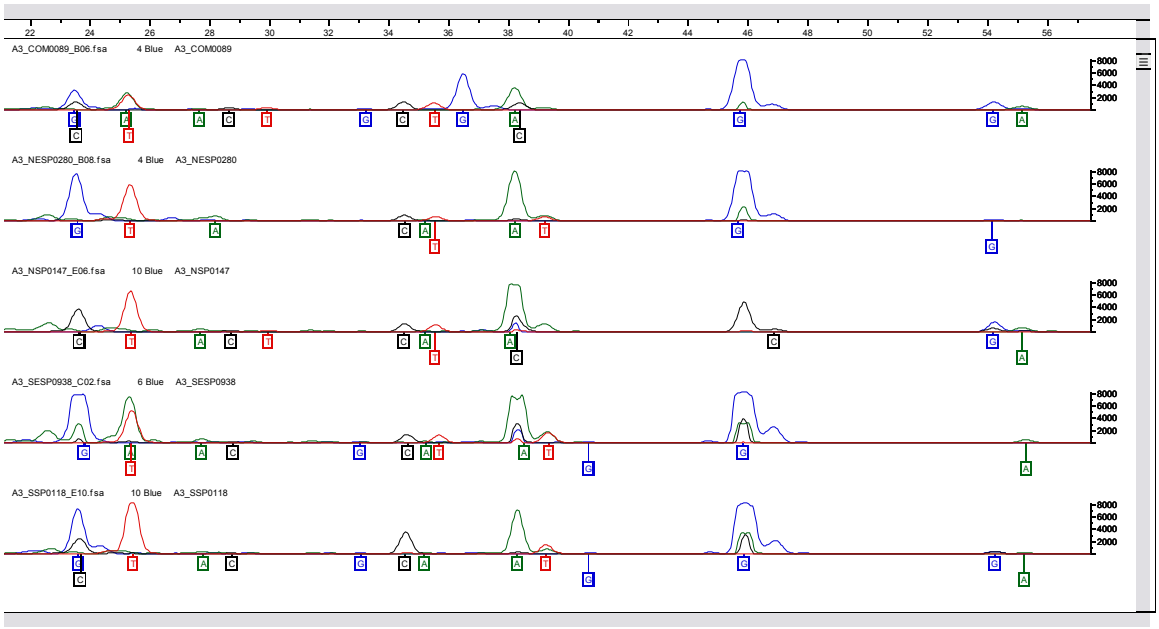


Figura 1.1 : Eletroferograma da amplificação dos alelos por SNaPshot™ para o sistema multiplex AIM3, contendo amostras de cada região geográfica: COM0089, NESP0280, NSP0147, SESP0938, SSP0118.
Fonte: O Autor

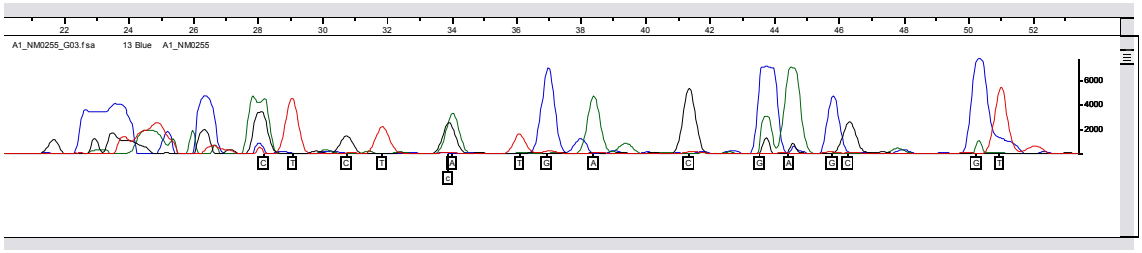


Figura 1.2 : Eletroferograma do Multiplex AIM1 indicando problemas na amplificação dos alelos por SNaPshot™, com ruídos e picos inespecíficos entre 20 e 28 pares de base.
Fonte: O Autor

Tabela 1.4 : Tamanho observado (OBS) e esperado (ESP) em pares de base (pb) dos picos na reação SNaPshot™ para cada sistema multiplex.

Multiplex AIM1				Multiplex AIM2				Multiplex AIM3			
		Tamanho do Alelo (pb)				Tamanho do Alelo (pb)				Tamanho do Alelo (pb)	
Loco	Alelos	OBS	ESP	Loco	Alelos	OBS	ESP	Loco	Alelos	OBS	ESP
WI11154	G	22,40	21	CYP3A4	G	24,00	21	rs3768641	G	23,50	21
	C	22,70	21		A	25,00	21		C	24,00	21
MID93	A	25,00	21	RB	C	26,50	21	rs4766807	A	25,20	22
	T	25,00	21		T	27,50	21		T	25,20	22
rs1426654	C	28,40	25	OCA2	G	29,15	27	rs3791896	G	27,40	27
	T	29,40	25		A	30,40	27		A	27,80	27
TSC1102055	C	31,10	27	CRH	G	33,60	33	rs267071	C	29,00	27
	T	32,20	27		A	35,00	33		T	30,20	27
rs4305737	G	32,50	31	FYNULL	C	34,90	33	rs310612	G	33,00	33
	A	33,90	31		T	36,10	33		A	35,00	33
rs727563	C	33,79	33	LPL	G	36,40	37	rs730086	C	34,70	37
	T	34,80	33		A	37,70	37		T	35,80	37
rs734780	C	34,90	35	rs1480642	C	38,20	37	rs3780293	G	36,70	37
	T	36,00	35		T	39,00	37		A	38,20	37
rs730570	G	36,80	37	rs7349	G	46,60	45	rs736556	C	38,50	37
	A	38,30	37		A	47,50	45		T	39,50	37
rs1129038	C	41,40	41	rs6034866	G	50,70	49	rs222541	G	40,50	41
	T	42,71	41		A	51,90	49		C	41,50	41
rs1240709	G	44,00	41	rs803733	C	50,50	49	rs1871534	G	45,60	45
	A	44,80	41		T	51,20	49		C	46,40	45
rs3796384	G	46,00	45					rs4280128	G	54,20	53
	C	46,50	45						A	55,20	53
rs2278354	G	50,00	49								
	T	50,80	49								

Fonte: O Autor

As estimativas de freqüências alélicas dos locos SNP foram tomadas a partir dos dados genotípicos gerados (Tabela 1.5) e foram utilizadas para análises de equilíbrio de Hardy-Weinberg (EHW) e análise de variância para as populações. Para a maioria das populações a maior parte dos locos se encontrou em EHW (Tabela 1.6). A figura 1.3 mostra a distribuição das freqüências alélicas nas subpopulações brasileiras.

Tabela 1.5 : Distribuição das freqüências alélicas dos 34 locos para a população brasileira e as seis populações de diferentes etnias

		População						
Loco	Alelo	EUR	AFM	CHN	CEU	YRI	ASN	BRA
rs1240709	A	0,771	0,227	0,021				0,570
	G	0,229	0,773	0,979				0,430
rs3796384	C	0,813	0,478	0,354				0,364
	G	0,188	0,522	0,646				0,636
TSC1102055	C	0,109	0,457	0,813				0,235
	T	0,891	0,543	0,188				0,765
rs730570	A	0,917	0,304	0,167	0,841	0,191	0,145	0,610
	G	0,083	0,696	0,833	0,159	0,809	0,855	0,390
rs4305737	A				0,206		0,825	0,458
	G				0,794		0,175	0,542
rs1426654	C				0,000	0,978	0,989	0,245
	T				1,000	0,022	0,011	0,755
rs734780	C				0,028	0,744	0,573	0,321
	T				0,972	0,256	0,427	0,679
rs727563	C				0,222	0,893	0,511	0,424
	T				0,778	0,107	0,489	0,576
rs2278354	G				0,094	0,739	0,697	0,271
	T				0,906	0,261	0,303	0,729
rs1480642	C				0,994	0,089	0,944	0,771
	T				0,006	0,911	0,056	0,229
rs7349	A				0,050	0,994	0,202	0,147
	G				0,950	0,006	0,798	0,853
OCA2	A	0,792	0,217	0,292	0,864	0,089	0,411	0,511
	G	0,208	0,783	0,708	0,136	0,911	0,589	0,489
CYP3A4	A				0,978	0,253	1,000	0,368
	G				0,022	0,747	0,000	0,632
rs1871534	C				1,000	0,017	1,000	0,767
	G				0,000	0,983	0,000	0,233
LPL	A				0,517	0,994	0,311	0,364
	G				0,483	0,006	0,689	0,636
CRH	A	0,938	0,432	1,000	0,922	0,383	1,000	0,439
	G	0,063	0,568	0,000	0,078	0,617	0,000	0,561
rs803733	C	0,729	0,261	0,370	0,822	0,022	0,208	0,702
	T	0,271	0,739	0,630	0,178	0,978	0,792	0,298
rs4766807	A	0,750	0,130	0,522	0,618	0,011	0,506	0,561
	T	0,250	0,870	0,478	0,382	0,989	0,494	0,439
rs222541	C	0,750	0,217	0,417	0,694	0,044	0,483	0,004
	G	0,250	0,783	0,583	0,306	0,956	0,517	0,996
rs736556	C	0,208	0,804	0,063				0,232
	T	0,792	0,196	0,938				0,768
rs4280128	A	0,375	0,935	0,000	0,381	0,989	0,039	0,448
	G	0,625	0,065	1,000	0,619	0,011	0,961	0,552
rs3791896	A	0,792	0,152	0,667	0,725	0,000	0,646	0,707
	G	0,208	0,848	0,333	0,275	1,000	0,354	0,293
rs3768641	C				0,083	1,000	0,062	0,173
	G				0,917	0,000	0,938	0,827
rs267071	C				0,661	0,017	0,972	0,545
	T				0,339	0,983	0,028	0,455

Tabela 1.5 continuação

		População						
Loco	Alelo	EUR	AFM	CHN	CEU	YRI	ASN	BRA
rs3780293	A	0,354	0,870	0,125	0,306	1,000	0,169	0,604
	G	0,646	0,130	0,875	0,694	0,000	0,831	0,396
rs310612	A					0,000		0,370
	G					1,000		0,630
WI11153	C							0,578
	G							0,422
MID93	A							0,371
	T							0,629
rs1129038	C							0,685
	T							0,315
FYNULL	C							0,118
	T							0,882
RB	C							0,579
	T							0,421
rs6034866	A							0,134
	G							0,866
rs730086	C							0,591
	T							0,409
AT3	i							0,340
	d							0,660

Fonte: O Autor

Tabela 1.6 : Análise do EHW teste exato de Fischer. Valores em negrito indicam P<0,05 e espaços em branco indicam população sem genótipo para o loco.

	População						
LOCO	EUR	AFM	CHN	CEU	YRI	ASN	BRA
rs1240709	0,119	0,541	**<0,001**				0,966
rs3796384	0,800	**0,045**	0,561				0,486
TSC1102055	0,445	0,126	0,782				0,212
rs730570	0,239	**0,001**	0,186	0,226	0,076	0,507	0,264
rs4305737				0,421		**0,014**	0,506
rs1426654				**<0,001**	0,060	0,011	**0,047**
rs734780				0,103	0,139	0,279	0,424
rs727563				0,861	0,472	0,481	0,483
rs2278354				0,332	0,778	0,213	0,371
rs1480642				**<0,001**	0,940	0,224	**0,003**
rs7349				0,387	**<0,001**	0,511	0,161
OCA2	0,631	0,496	0,540	0,383	0,288	0,529	**0,046**
CYP3A4				0,057	0,217	**<0,001**	**<0,001**
rs1871534				**<0,001**	**0,030**	**<0,001**	0,334
LPL				**0,019**	**<0,001**	0,836	**0,031**
CRH	0,115	0,630	**<0,001**	0,843	0,491	**<0,001**	0,475
rs803733	0,394	**0,030**	0,184	0,561	0,057	0,256	**<0,001**
rs4766807	0,879	0,598	0,238	0,863	**0,007**	0,112	0,354
rs222541	0,873	0,075	0,606	0,152	**0,013**	0,466	**<0,001**
rs736556	0,596	0,335	0,125				**0,002**
rs4280128	0,120	**<0,001**	**<0,001**	0,351	**0,014**	0,221	0,457
rs3791896	0,628	0,824	0,364	0,632	**<0,001**	0,797	**0,016**
rs3768641				0,208	**<0,001**	0,555	**0,003**
rs267071				0,738	**0,037**	0,123	**0,038**
rs3780293	**0,015**	0,053	**0,039**	0,112	**<0,001**	0,399	**0,036**
rs310612						**<0,001**	**<0,001**
WI11153							0,489
MID93							**0,049**
rs1129038							0,394
FYNULL							0,918
RB							**<0,001**
rs6034866							**0,041**
rs730086							0,621
AT3							0,311

Fonte: O Autor

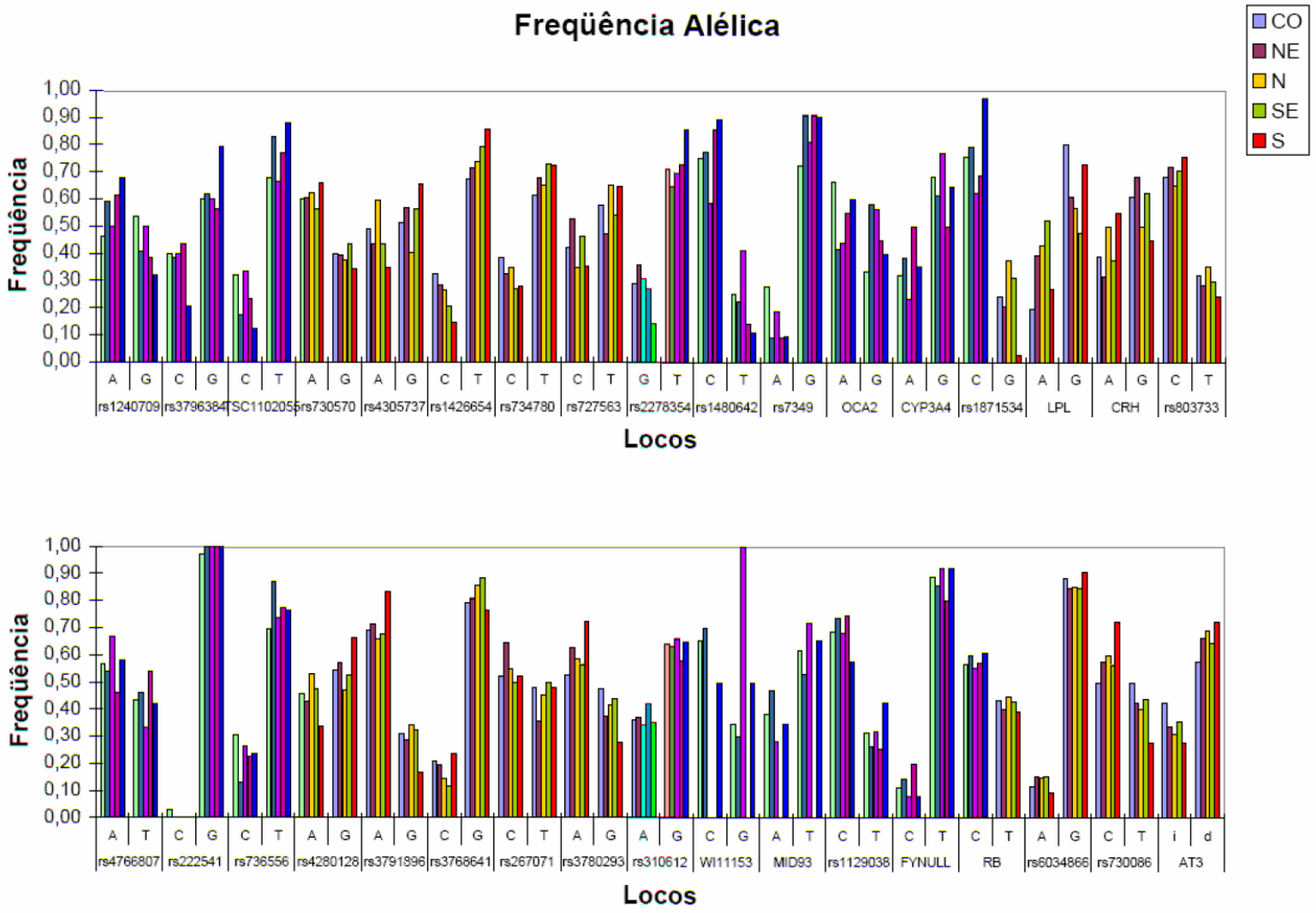


Figura 1.3 : Distribuição das freqüências alélicas dos 34 locos para as subpopulações brasileiras.
Fonte: O Autor

1.2.3 Análise de Estrutura Genética

A partir dos dados genotípicos foram estimadas a heterozigose média esperada e observada, e os índices de fixação populacionais (Fis, Fit e Fst) para a população brasileira (Tabela 1.7) e para as demais populações (Tabela 1.8)

Tabela 1.7 : Heterozigose média esperada (He) e observada (Ho), índices de fixação populacionais (Fis, Fit e Fst) e seus respectivos testes de significância para as amostras regionais brasileiras.

População	He	Ho	Fis	Fit	Fst
BRA_CO	0,431	0,413	0,040	-	-
BRA_NE	0,410	0,420	-0,027	-	-
BRA_N	0,411	0,417	-0,016	-	-
BRA_SE	0,412	0,439	-0,065	-	-
BRA_S	0,364	0,391	-0,074	-	-
Total	0,405	0,416	-0,036	-0,021	0,014
Teste de Significância					
Limite Superior			0,026	0,042	0,021
Limite Inferior			-0,108	-0,092	0,008
Nº. Repetições			1000	1000	1000
Intervalo de Confiança			95%	95%	95%

Fonte: O Autor

Tabela 1.8 : Heterozigose média esperada (He) e observada (Ho), índices de fixação populacionais (Fis, Fit e Fst) e seus respectivos testes de significância para as populações CEU, YRI, ASN, AFM, CHN e EUR.

População	He	Ho	Fis	Fit	Fst
EUR	0,329	0,353	-0,077	-	-
AFM	0,352	0,287	0,188	-	-
CHN	0,293	0,281	0,039	-	-
CEU	0,253	0,250	0,009	-	-
YRI	0,131	0,132	-0,006	-	-
ASN	0,270	0,254	0,058	-	-
Total	0,271	0,260	0,033	0,600	0,586
Teste de Significância					
Limite Superior			0,063	0,678	0,668
Limite Inferior			0,00001	0,529	0,509
Nº. Repetições			1000	1000	1000
Intervalo de Confiança			95%	95%	95%

Fonte: O Autor

Os valores de Fst par a par para a população brasileira contra as demais foram estimados (Tabela 1.9) e utilizados para computar uma matriz de

distância genética. Os dados de Fst par a par foram utilizados para agrupar as populações pelo método UPGMA (do inglês, *Unweighted Pair Group Method with Arithmetic mean*) para construção de árvore de distância genética (Figura 1.4).

Tabela 1.9 : Índice de fixação (Fst) da população brasileira par a par com as demais populações e seus respectivos testes de significância.

		Teste de Significância (95%)	
BRA x	Fst	Limite Superior	Limite Inferior
EUR	0,084	0,167	0,019
AFM	0,106	0,166	0,053
CHN	0,119	0,185	0,053
CEU	0,121	0,200	0,052
ASN	0,185	0,269	0,100
YRI	0,350	0,444	0,242

Fonte: O Autor

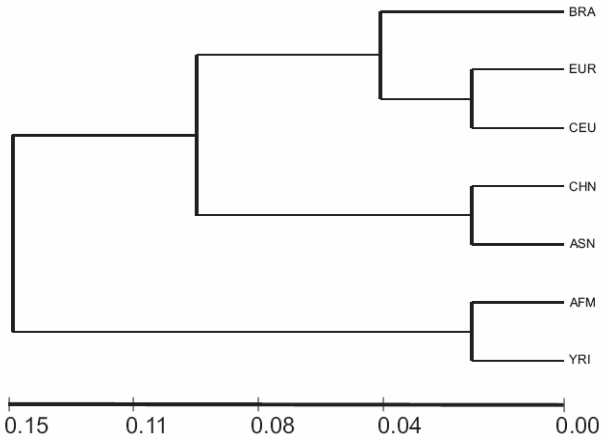


Figura 1.4 : Árvore de distância genética agrupada por UPGMA baseada nos valores de Fst par a par entre populações.
Fonte: O Autor

A análise de estruturação populacional utilizando o programa Structure 2.1 mostrou evidência de estruturação genética tanto na população brasileira isolada como em conjunto com as populações Européias, Africanas e Asiáticas.

Primeiramente, a análise de estruturação populacional foi realizada para as populações alocadas no dbSNP. A estatística DeltaK identificou nessas populações maior probabilidade em K=2 (Figura 1.5). Assim, foi possível a visualização da distribuição das populações conforme estrutura genética para K=2 (Figura 1.6) e K=3 (Figura 1.7).

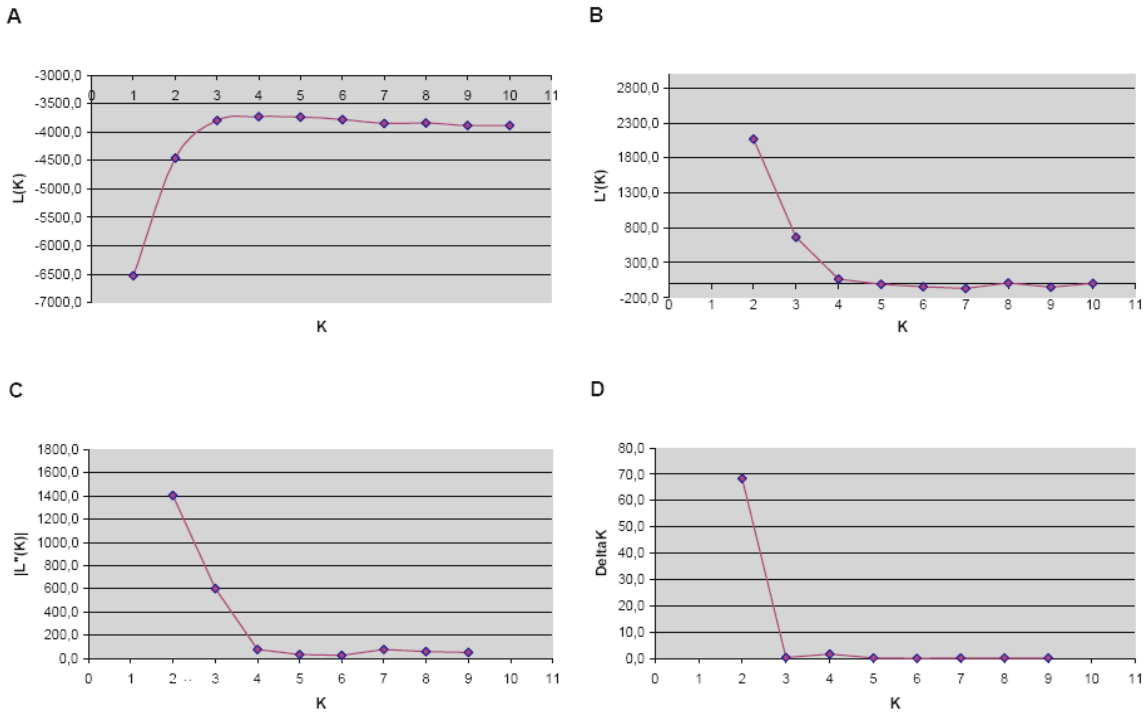


Figura 1.5 : Estimativa de melhor K para as populações alocadas no dbSNP de acordo com parâmetros de estimativa de DeltaK: (A) L(K), ou média dos valores de ln P(D); (B) L'(K), ou diferença entre L(K) e L(K-1); (C) |L"(K)|, ou módulo da diferença de L'(K) e L'(K+1) e (D) DeltaK, ou divisão do |L"(K)| pelo produto do desvio padrão das médias de L(K) e o número de repetições realizadas.
Fonte: O Autor

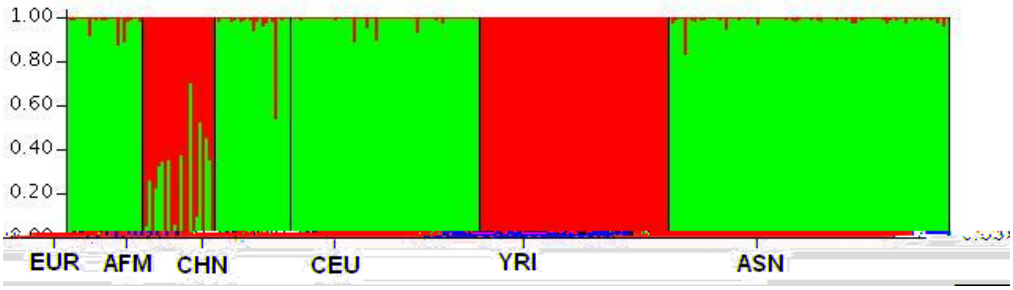


Figura 1.6 : Distribuição de estrutura genética das populações de origem Européia, Africana e Asiática para K=2.
Fonte: O Autor

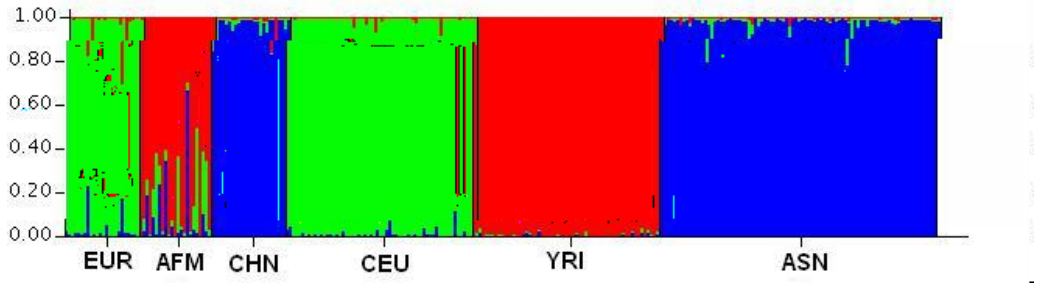


Figura 1.7 : Distribuição de estrutura genética das populações de origem Européia, Africana e Asiática para K=3.
Fonte: O Autor

Em seguida, a análise de estruturação populacional foi realizada somente com as amostras regionais brasileiras e identificou maior probabilidade em K=2 (Figura 1.8). Com isso foi possível a visualização da distribuição das proporções individuais de miscigenação na população brasileira e suas divisões regionais para K=2 (Figura 1.9) e K=3 (Figura 1.10).

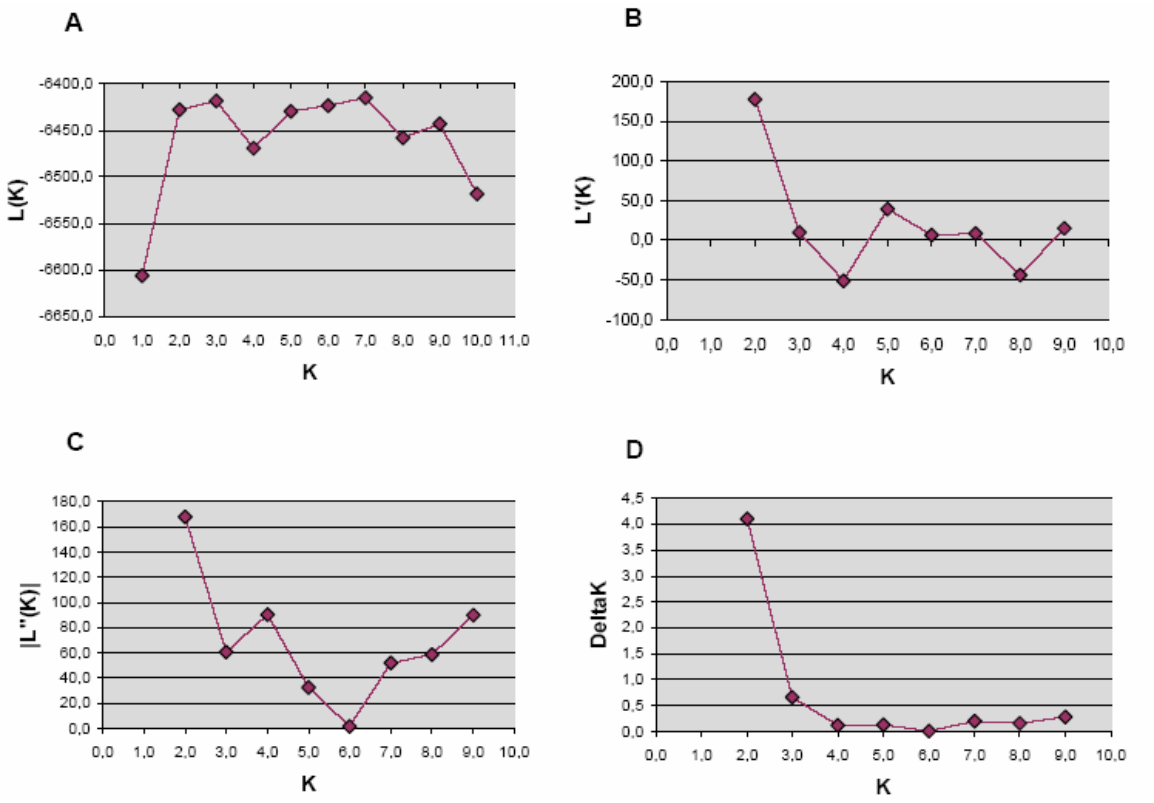


Figura 1.8 : Estimativa de melhor K para a população brasileira de acordo com parâmetros de estimativa de DeltaK: (A) L(K), ou média dos valores de ln P(D); (B) L'(K), ou diferença entre L(K) e L(K-1); (C) |L"(K)|, ou módulo da diferença de L'(K) e L'(K+1) e (D) DeltaK, ou divisão do |L"(K)| pelo produto do desvio padrão das médias de L(K) e o número de repetições realizadas.
Fonte: O Autor

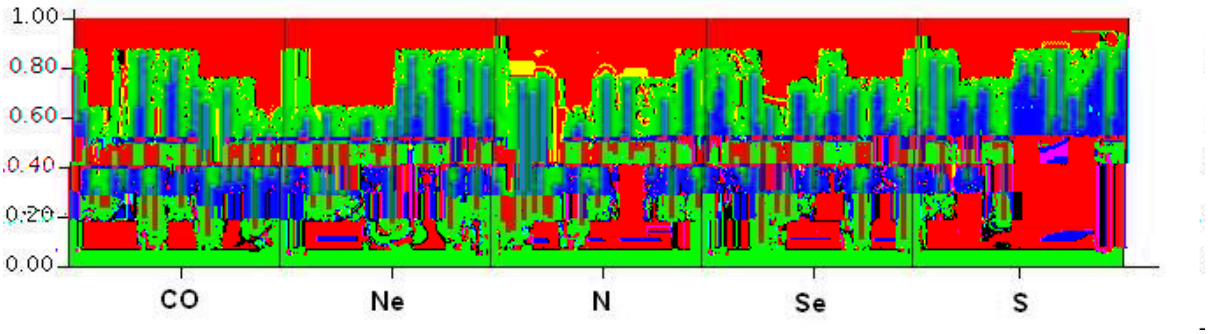


Figura 1.9 : Distribuição das proporções individuais de miscigenação nas divisões regionais da população brasileira para K=2.
Fonte: O Autor

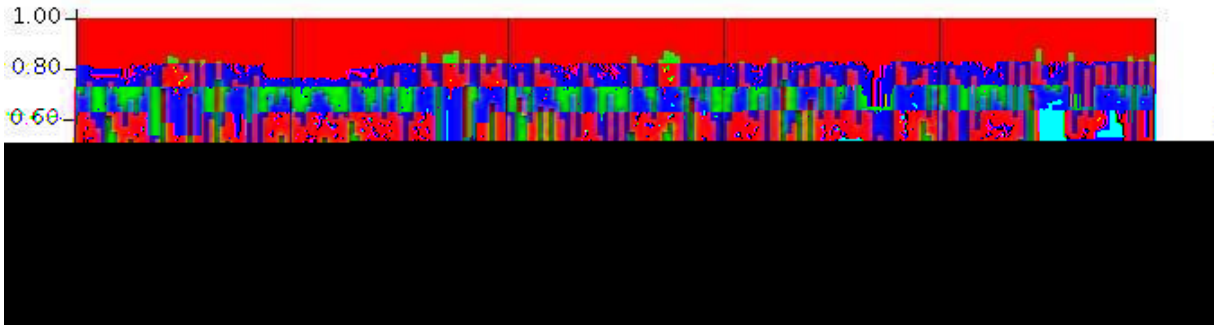


Figura 1.10 : Distribuição das proporções individuais de miscigenação nas divisões regionais da população brasileira para K=3.
Fonte: O Autor

Em seguida, a análise foi realizada com todas as 11 subpopulações. Utilizando todas as populações conjuntamente, a análise de estruturação populacional identificou maior probabilidade em K=2 (Figura 1.11). Com isso foi possível a visualização da estrutura genética das populações para K=2 (Figura 1.12) e K=3 (Figura 1.13).

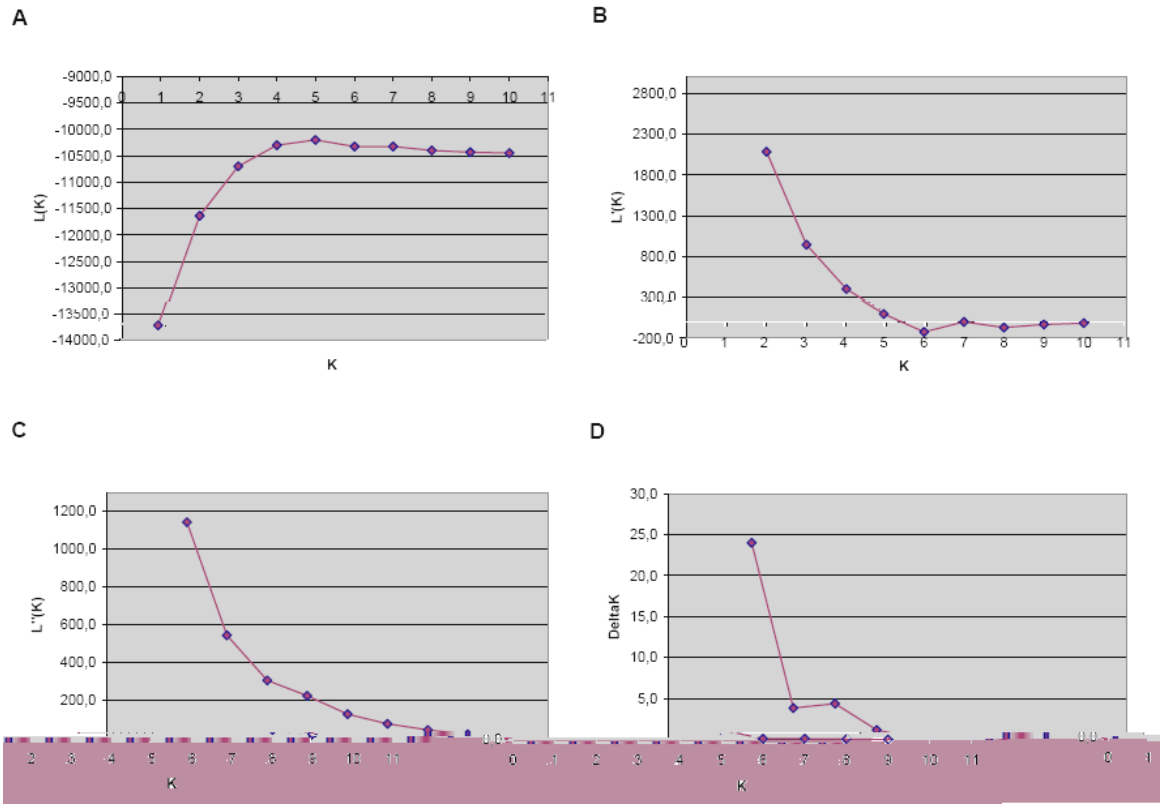


Figura 1.11 : Estimativa de melhor K para as populações alocadas no dbSNP de acordo com parâmetros de estimativa de DeltaK: (A) L(K), ou média dos valores de ln P(D); (B) L'(K), ou diferença entre L(K) e L(K-1); (C) |L"(K)|, ou módulo da diferença de L'(K) e L'(K+1) e (D) DeltaK, ou divisão do |L"(K)| pelo produto do desvio padrão das médias de L(K) e o número de repetições realizadas.
Fonte: O Autor

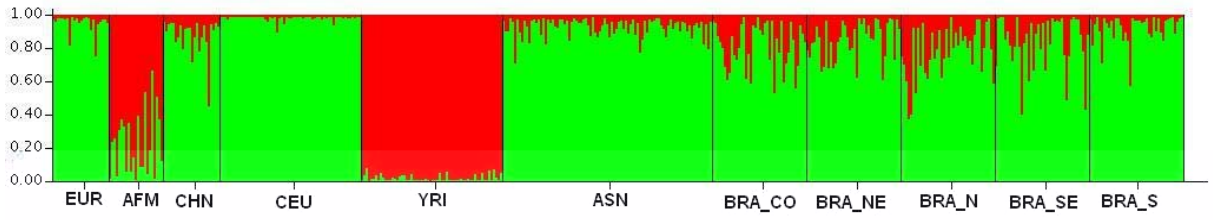


Figura 1.12 : Distribuição de estrutura genética de todas as populações utilizadas no estudo para K=2.
Fonte: O Autor

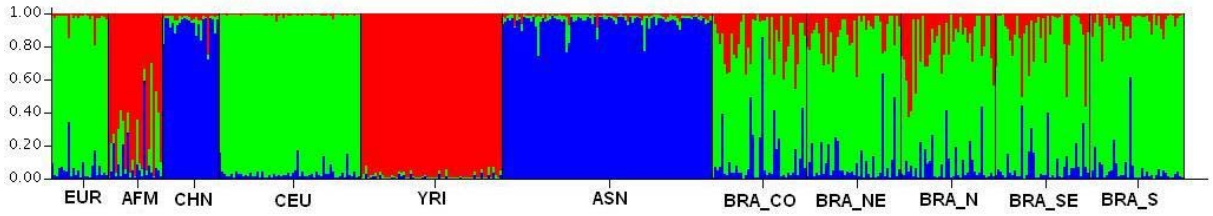


Figura 1.13 : Distribuição de estrutura genética de todas as populações utilizadas no estudo para K=3.
Fonte: O Autor

A quarta análise foi realizada com a exclusão das populações de origens asiáticas e somente com os marcadores presentes nas populações do dbSNP. Desta forma a análise de estruturação populacional identificou maior probabilidade em K=2 (Figura 1.14). Com isso foi possível a visualização da estrutura genética das populações para K=2 (Figura 1.15).

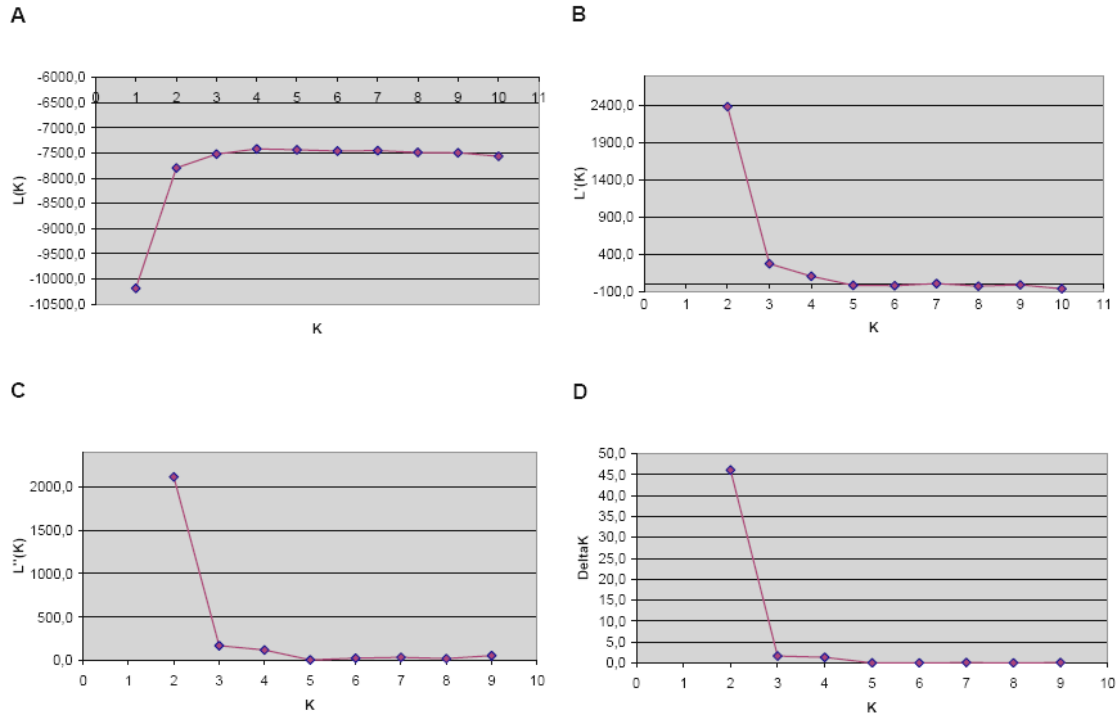


Figura 1.14 : Estimativa de melhor K para as populações de origens Européias e Africanas e a população brasileira de acordo com parâmetros de estimativa de DeltaK: (A) L(K), ou média dos valores de ln P(D); (B) L'(K), ou diferença entre L(K) e L(K-1); (C) |L"(K)|, ou módulo da diferença de L'(K) e L'(K+1) e (D) DeltaK, ou divisão do |L"(K)| pelo produto do desvio padrão das médias de L(K) e o número de repetições realizadas.
Fonte: O Autor

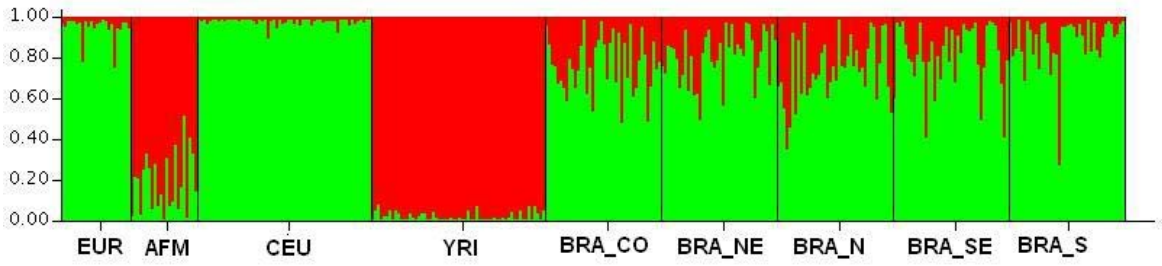


Figura 1.15 : Distribuição de estrutura genética das populações de origens Européias e Africanas e a população brasileira utilizadas no estudo para K=2.
Fonte: O Autor

A tabela 1.10 descreve os valores médios populacionais para as estimativas de ancestralidade de acordo com cada grupo ancestral em cada análise realizada para K = 2, enquanto a tabela 1.11 descreve os valores para K = 3.

Tabela 1.10 : Índices de ancestralidade populacionais de acordo com cada análise para K=2.

		População / n											
		EUR	AFM	CHN	CEU	YRI	ASN	BRA	CO	NE	N	SE	S
Análise (A)	Grupo Ancestral	24	23	24	60	60	89	200	40	40	40	40	40
A1	Eur	0,98	0,184	0,967	0,991	0,005	0,989	-	-	-	-	-	-
	Afr	0,02	0,816	0,033	0,009	0,995	0,011	-	-	-	-	-	-
A2	Eur	-	-	-	-	-	-	0,507	0,438	0,510	0,435	0,502	0,651
	Afr	-	-	-	-	-	-	0,493	0,562	0,490	0,565	0,498	0,349
A3	Eur	0,955	0,229	0,879	0,983	0,022	0,930	0,843	0,816	0,851	0,783	0,842	0,924
	Afr	0,045	0,771	0,121	0,017	0,978	0,070	0,157	0,184	0,149	0,217	0,158	0,076
A4	Eur	0,951	0,192	-	0,98	0,025	-	0,805	0,770	0,810	0,743	0,818	0,885
	Afr	0,049	0,808	-	0,02	0,975	-	0,195	0,230	0,190	0,257	0,182	0,115

Fonte: O Autor

Tabela 1.11 : Índices de ancestralidade populacionais de acordo com cada análise para K=3.

		População / n											
		EUR	AFM	CHN	CEU	YRI	ASN	BRA	CO	NE	N	SE	S
Análise (A)	Grupo Ancestral	24	23	24	60	60	89	200	40	40	40	40	40
A1	Eur	0,943	0,108	0,022	0,980	0,004	0,016	-	-	-	-	-	-
	Afr	0,028	0,810	0,014	0,007	0,991	0,006	-	-	-	-	-	-
	Asn	0,029	0,082	0,963	0,013	0,004	0,978	-	-	-	-	-	-
A2	Grupo 1	-	-	-	-	-	-	0,351	0,294	0,356	0,299	0,341	0,466
	Grupo 2	-	-	-	-	-	-	0,319	0,331	0,303	0,327	0,331	0,302
	Grupo 3	-	-	-	-	-	-	0,330	0,376	0,341	0,374	0,328	0,232
A3	Eur	0,908	0,148	0,043	0,956	0,012	0,031	0,779	0,716	0,782	0,721	0,788	0,889
	Afr	0,029	0,763	0,028	0,008	0,977	0,016	0,131	0,161	0,122	0,192	0,129	0,052
	Asn	0,062	0,089	0,928	0,036	0,011	0,954	0,089	0,123	0,096	0,086	0,083	0,059

Fonte: O Autor

1.2.4 Conteúdo de Informação para Inferência de Ancestralidade

O conteúdo de informação para inferência de ancestralidade foi testado em cada loco cujos os genótipos foram recuperados, exceto para o loco rs310612, o qual possui genótipo somente para uma população. Pelo critério de similaridade genética as populações mais próximas foram agrupadas em uma para simplificar as análises. Com isso as medidas de informação para atribuição de ancestralidade (In), informação para coeficientes de ancestralidade (Ia) e taxa ótima de atribuição correta (ORCA) foram tomadas a partir dos genótipos das três populações parentais. Para avaliar o conteúdo da informação de acordo com as populações foram testadas

todas elas juntas e em combinações par a par. A figura 1.16 ilustra o poder de informação de cada loco gerado para cada uma das estatísticas In (Figura 1.16A), Ia (Figura 1.16B) e ORCA (Figura 1.16C).

Pela sua natureza estatística, que relaciona a freqüência de um alelo entre duas ou mais populações, a estatística do conteúdo de informação para atribuição de ancestralidade é considerada altamente correlacionada com os valores de delta e de Fst. Portanto, testes de correlação foram realizados para In, delta e Fst entre os pares de populações (Figura 1.17). Os valores de δ foram tomados a partir das freqüências alélicas das populações e os de Fst (Theta-P neste caso) a partir dos resultados em cada loco para cada par de população.

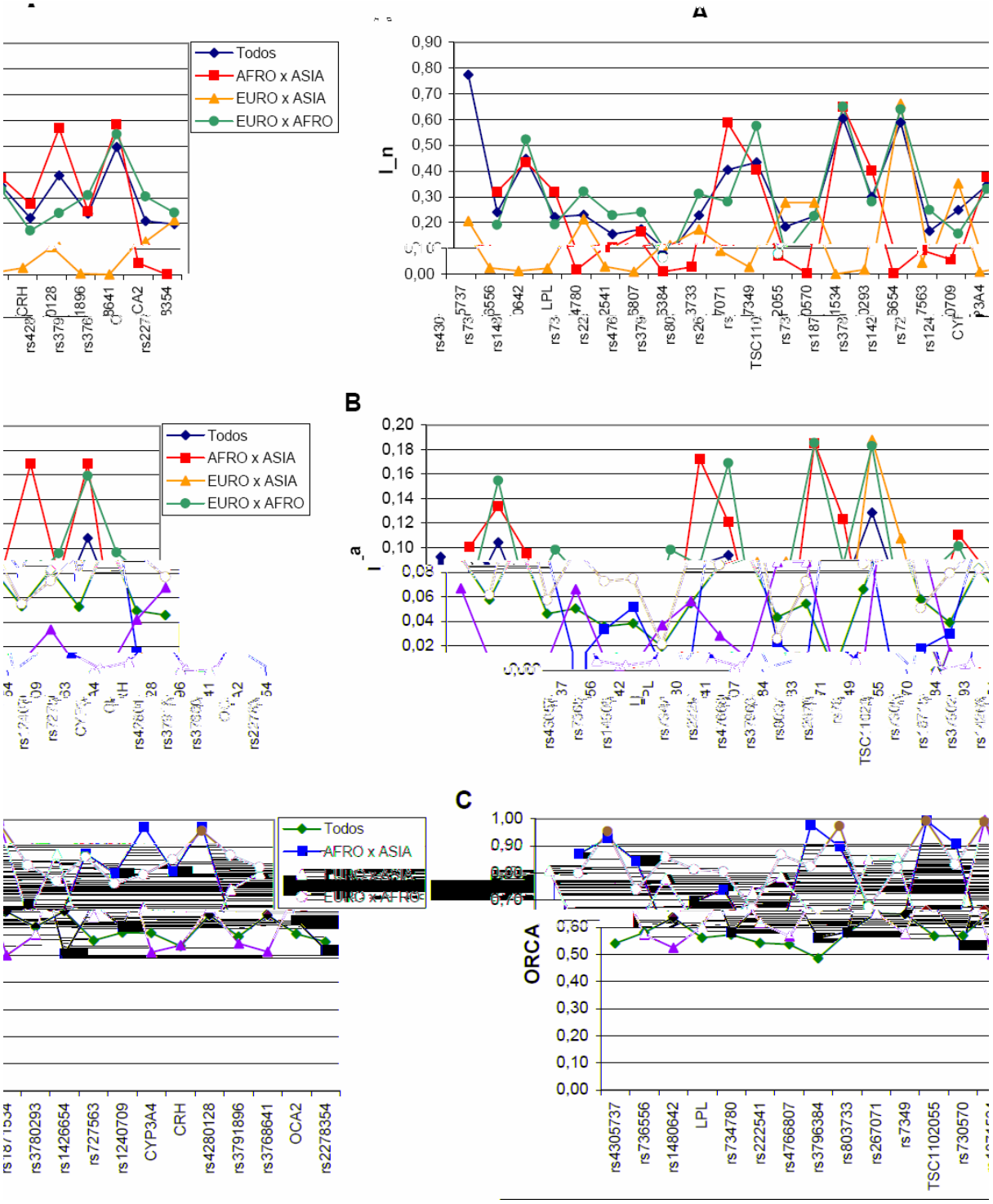


Figura 1.16 : Análise do conteúdo de informação de ancestralidade segundo as estatísticas de In(A), Ia (B) e ORCA (C).
Fonte: O Autor

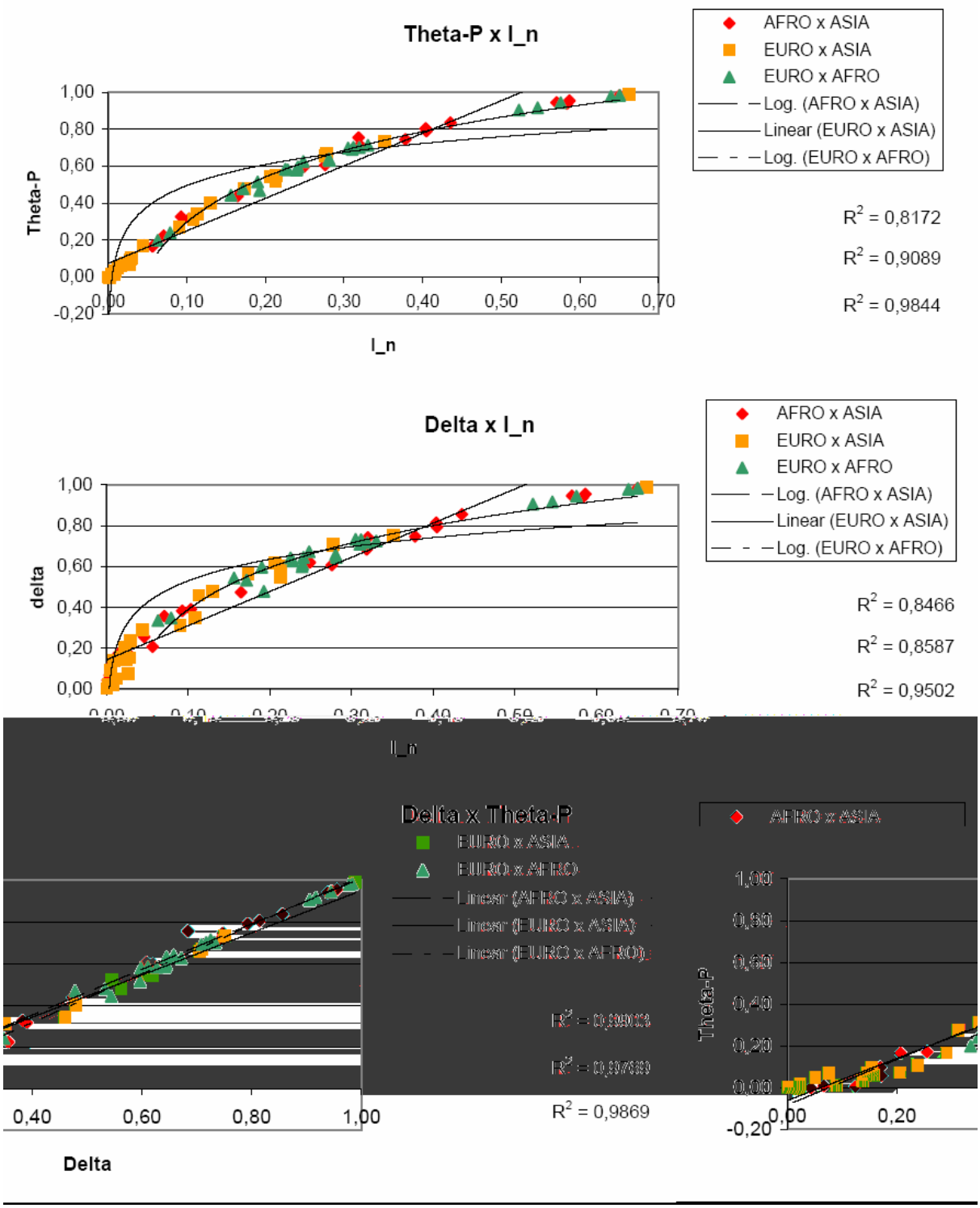


Figura 1.17 : Testes de correlação entre δ, Theta-P e ln para os pares de populações.
Fonte: O Autor

1.3 DISCUSSÃO

A história bioantropológica da população brasileira é extensamente ramificada quando se trata dos efeitos primórdios e recentes de miscigenação entre as populações nativas e seus imigrantes. Esse legado histórico-social repercute hoje em uma grande heterogeneidade genética na população brasileira. O interesse em caracterizar o grau de miscigenação e contribuição de frações genéticas ancestrais por meio de marcadores moleculares tem sido extensamente investigado em pesquisas que envolvem principalmente os marcadores moleculares de linhagem paterna – microssatélites do cromossomo Y (ABE-SANDES; SILVA; ZAGO, 2004; CARVALHO-SILVA et al., 2001) e materna – mtDNA (ALVES-SILVA et al., 2000; MARRERO et al., 2005), além de marcadores autossômicos do tipo microssatélite, Indel e SNP (CALLEGARI-JACQUES et al., 2003; FERREIRA et al., 2005; PARRA, F. C. et al., 2003; PIMENTA et al., 2006).

Neste estudo foi empregado um amplo conjunto de marcadores do tipo SNP com potencial informativo para atribuição de ancestralidade e seus respectivos coeficientes, para avaliar o grau de miscigenação da população brasileira em suas regiões geográficas.

As genotipagens foram realizadas satisfatoriamente utilizando o protocolo desenvolvido para extensão de base única, SNapShot™ Multiplex System (Applied Biosystems). O sucesso de genotipagem para a população brasileira chegou a aproximadamente 92% de todos os genótipos esperados. No entanto a média percentual de genótipo por loco chegou a 81%, com locos variando de 97% do total de genótipos, como o rs6034866, até 16% (WI11153) ou 29% (MID93). Esses números refletem o grau de dificuldade encontrado na técnica de extensão de base única. Dentre os problemas que dificultaram a análise pode-se citar a amplificação de picos inespecíficos na região de 20 a 26 pares de bases (Figura 1.2), e também a não amplificação ou amplificação de picos abaixo de 100 rfu's que geraram insegurança na genotipagem. A fonte desses problemas pode ser explicada, respectivamente, pelo excesso de iniciadores de PCR na reação SNaPshot™ que não foram devidamente degradados na purificação enzimática, por

locos com problema de amplificação na PCR e/ou desbalanço na concentração de iniciadores na reação de extensão de base única. O trabalho para resolução destes problemas está sendo conduzido em ensaios paralelos.

O padrão da distribuição de freqüências alélicas revelou somente um loco com freqüências similares à das populações de origem africana (rs222541). Nos demais locos o padrão da distribuição de freqüência dos alelos na população brasileira se mostrou similar tanto para Europeus quanto para Asiáticos.

O padrão da distribuição de freqüências alélicas entre as regiões geográficas brasileira se mostrou similar entre todas as regiões. Como conseqüência, os valores das estatísticas F (Fis, Fst e Fit) revelaram que nas subpopulações, tanto as variações dentro como as variações entre amostras regionais, a diminuição da heterozigose devido ao endocruzamento e deriva não são significativas em relação a população total. Por outro lado, o índice de fixação (Fst) foi significativo. Isto indica que a população amostrada possui diferenciação genética significativamente pequena, e que não há decréscimo na heterozigose média devido ao isolamento geográfico das subpopulações relativa à uma população total hipotética. De acordo com a tabela de Wright, o grau de diferenciação populacional estimado pelo Fst é significativamente baixo, indicando que a diferença genética entre os grupos geográficos da população brasileira pode ser atribuída em 0,014.

Este resultado levanta algumas questões e hipóteses referentes aos processos bioantropológicos do povoamento e miscigenação brasileiros: (1) A população brasileira sofreu um processo de colonização e miscigenação homogêneos quanto aos percentuais de ancestralidade genômica em todas as regiões; (2) Eventos estocásticos como taxa de migração entre regiões aliados à deriva gênica revelaram uma população homogênea em todas as regiões; (3) Os marcadores utilizados não forneceram informação suficiente para diferenciar devidamente as regiões geográficas.

Em vista disso, dados de indicadores sociais (Tabela 1) sugerem que a declaração de cor de pele é consideravelmente diferente entre as regiões, porém com predominância de brancos e pardos em todas elas. No entanto, estudos mostram que em determinadas amostras regionais da população brasileira, como a

do Centro-Oeste, não existe correlação entre auto declaração de cor de pele e ancestralidade genômica, e, por conseqüência, os níveis de ancestralidade genômica não divergem entre os grupos de brancos e pardos (ABREU, 2007; VIEIRA et al., 2006). Os resultados desses estudos mostram que, além da autodenominação de pele não ser uma ferramenta adequada para separar grupos homogêneos em estudos de associação, os principais grupos populacionais geográficos são homogêneos quanto à sua ancestralidade genômica.

No entanto, quando avaliadas todas as regiões Brasileiras, essas considerações podem não proceder ao mesmo resultado, devido aos particulares processos históricos e demográficos de povoamento e colonização. No presente estudo não foi possível coletar dados sobre a autodenominação de cor de pele, porém, revelou que grupos regionais são homogêneos quanto à sua ancestralidade genômica, fato também observado nas análises de estrutura populacional.

As análises de estrutura populacional feitas pelo programa Structure para a população brasileira, com atribuição *a priori* de subpopulação por região geográfica e sem a participação das demais populações, atribuiu a maior probabilidade do número de populações igual a dois (K=2), no entanto com baixo sinal. Esta análise gerou valores de contribuição de ancestralidade equipartido entre os grupos, exceto para a região sul (Tabela 1.9). No entanto, nesta análise especificamente, o efeito platô não foi observado, e o menor valor de *log Pr(X|K)* definido pelo Structure foi para K = 3. Da mesma forma, para o modelo tri-parental, os valores de contribuição de ancestralidade também foram equipartidos, exceto para a região sul (Tabela 1.10), não sendo possível, conseqüentemente, distinguir a contribuição africana da indígena. Essa dualidade de grupos parentais é extremamente conflitante, ainda mais quando os valores atribuídos para cada grupo são equivalentes, como visto. Em compensação, estimativas de proporção de miscigenação podem ser desafiadoras se existir pouca ou nenhuma amostra da população parental, porque na ausência de indivíduos não miscigenados pode haver, na estimativa estatística, uma atribuição inadequada das freqüências alélicas aos grupos parentais e, assim, desviar a atribuição de um alelo para um ou outro grupo (PRITCHARD; WEN, 2003). Portanto, dificilmente esta análise representa os níveis mais próximos da ancestralidade genômica nas regiões brasileiras.

Quando são utilizadas as outras populações de três origens distintas nas análises, o cenário muda completamente. A análise A3, (figura 1.11) levou em consideração os 34 locos e todas as populações juntas. Desta forma, a contribuição de genoma ancestral africano mudou substancialmente tanto para K = 2 quanto para K = 3. O resultado quando assumido modelo bi-parental, no entanto, deu indícios de desvio superestimado da ancestralidade européia. Apesar do programa tratar de dados faltantes com cautela (PRITCHARD; WEN, 2003), as estimativas estatísticas produzidas por dados incompletos podem reduzir significativamente o tamanho da amostra e assim diminuir o poder estatístico (BADZIOCH; THOMAS; JARVIK, 2003; HINRICHS; SUAREZ, 2005). Portanto, a análise A4 (figura 1.12) foi feita com a exclusão das populações asiáticas. Também foram excluídos todos os locos que não possuíam genótipos para as populações européias ou africanas e os locos rs803733 e rs310612, que possuíam genótipos somente para uma das populações, ou européia ou africana respectivamente, restando somente 25 locos.

As análises A3 (com K=3) e A4 (com K=2) geraram resultados mais condizentes com os apontados na literatura, cujos a população brasileira apresentou em torno de 70% a 80% de contribuição européia, 10 a 20% de contribuição Africana e entre 8 a 18% de contribuição Indígena (CALLEGARI-JACQUES et al., 2003). Como no presente estudo não foi utilizado populações indígenas, o estabelecimento de sua proporção fica a desejar, mesmo com a utilização da população asiática como um indicativo de população mais próxima. Os resultados de Fst par a par indicaram que as populações Asiáticas estão mais próximas da Brasileira que as populações de origem Africana.

A análise do Structure para as populações do dbSNP (Análise A1) atribuiu a maior probabilidade do número de populações igual a dois, e, portanto, não conseguiu separar as populações de origem Asiáticas das de origem Européias.

Apesar do fato de as três populações serem conhecidas e terem sido geneticamente distintas, quando o modelo de três populações foi aplicado, o efeito platô foi observado nas médias de L(K), e DeltaK providenciou indícios para K=2. Este fato pode ser particularmente explicado quando analisados os dados de ORCA (Figura 1.16C). A média dos valores de ORCA para todos os locos analisados revelou grande poder de atribuição para o par de populações Européias e Africanas

(0,84), maior que para os outros dois pares (Europeu/Asiático = 0,76 e Africano/Asiático = 0,67). No entanto, quando as três populações são analisadas juntas, a taxa cai drasticamente para 0,59. Portanto, os marcadores selecionados para esta análise não têm precisão suficiente para distinguir indivíduos Europeus de Asiáticos quando na presença de indivíduos Africanos.

Um fator que corrobora essa conclusão é a correlação entre In, δ e Fst. Rosenberg e colaboradores (2003) definiram In como uma função logarítmica dos parâmetros $\delta(|p_{11}-p_{21}|)$ e $\sigma(p_{11}+p_{21})$, portanto, seria esperado que os testes de regressão entre In e δ, e In e Fst fossem logarítmicas, enquanto entre δ e Fst seriam lineares. Os gráficos da figura 1.17 foram plotados para as correlações medidas de acordo com o máximo do coeficiente de regressão R^2, seja a correlação linear ou logarítmica. O maior coeficiente foi obtido para as relações de In entre as populações Européias e Africanas, enquanto os demais pares apresentam ou valores baixos de correlação, ou a mesma é linear e não logarítmica. Isto indica uma fraca correlação quando são analisadas as populações Africana com Asiática ou Européia com Asiática.

Foi interessante verificar que a população definida como Afro-americana apresentou indícios de contribuição de ancestralidade européia em cerca de 19% quando o modelo bi-parental é definido. Este valor é comparável a estimativa média de outras populações afro-americanas (PARRA, E. J. et al., 1998; SHRIVER et al., 1997), dando uma clara indicação de que os marcadores conseguem quantificar mistura genética a partir dos dados genotípicos em populações miscigenadas com contribuição Européia e Africana. Por outro lado, quando assumido o modelo tri-parental, é atribuído cerca de 8% de contribuição Asiática à população definida como Afro-americana, o que não condiz com as estimativas de trabalhos anteriores (PARRA, E. J. et al., 1998; SHRIVER et al., 1997). Nestes trabalhos citados, não são relatadas contribuições nem de miscigenação com Asiáticos, nem com Ameríndios na população Afro-americana, indicando um desvio na atribuição realizada nesta análise devido, provavelmente, aos marcadores utilizados.

O controle de estratificação populacional depende criticamente do tipo e da informação dos marcadores utilizados no estudo. Em muitos casos, o uso de microssatélites providencia informação genética suficiente para identificar e separar grupos de diferentes etnias e grupos miscigenados de acordo com sua proporção de ancestralidade (ROSENBERG et al., 2003; ROSENBERG et al., 2002; SMITH et al., 2001). No entanto, os marcadores bi-alélicos (SNP e Indel) têm sido usados com bastante freqüência para este propósito (CHOUDHRY et al., 2006; HOGGART et al., 2003; PARRA, E. J. et al., 2001; PARRA, F. C. et al., 2003; PFAFF; KITTLES; SHRIVER, 2002; PFAFF et al., 2001; SHRIVER et al., 2003; SHRIVER et al., 1997; SMITH et al., 2001). Um dos motivos da sua aplicabilidade é o fato de que a análise de múltiplos locos bi-alélicos que possuem alta diferenciação populacional (δ e Fst), tem maior conteúdo de informação para atribuição e, quando bem empregados, estes marcadores podem levar a uma melhor inferência de estrutura populacional do que os microssatélites (LIU, N. et al., 2005; PFAFF et al., 2004; ROSENBERG et al., 2003; SHRIVER et al., 1997). No entanto, o número e a quantidade de informação para atribuição destes marcadores podem alterar substancialmente a precisão e acurácia na estimativa de ancestralidade. O número de marcadores bi-alélicos independentes necessários em um modelo de miscigenação de duas populações parentais é consideravelmente maior que o número exigido em um modelo de populações não miscigenadas (ROSENBERG et al., 2003) e, ainda, a inclusão de marcadores menos informativos pode gerar ruídos e piorar a análise (LIU, N. et al., 2005). No presente estudo, foi observado que esses efeitos realmente ocorrem e são exclusivamente dependentes dos marcadores utilizados.

A escolha dos 34 marcadores utilizados seguiu critérios de seleção baseado na diferença das freqüências alélicas nas três populações parentais tomadas par a par. Essa estratégia acabou por selecionar locos com maior informação para atribuição de ancestralidade, informação para coeficientes de ancestralidade e taxa ótima de atribuição correta preferencialmente para o par Europeu-Africano. De certa maneira, os marcadores utilizados tiveram poder suficiente para diferenciar as populações Asiáticas das Européias quando na presença de indivíduos Africanos, porém geraram distorções quando populações miscigenadas foram analisadas. Embora eles não tenham sido escolhidos para essa finalidade, a identificação da população asiática seria útil como sugestão de

potencial indicativo de percentual Indígena na população brasileira, devido à semelhança das freqüências alélicas na maioria dos locos entre as duas populações (Tabela 1.1 para freqüências em Ameríndios e Tabela 1.5 para freqüências em Asiáticos). Dados que seriam extremamente úteis para avaliar o real poder de inferência desses marcadores para populações tri-parentais, como a Brasileira, seria a disponibilidade de todos esses genótipos nas populações-base do tripé populacional brasileiro, ou seja, Europeus de origem latina (Portugueses, Espanhóis e Italianos), Africanos de origens sub-Saariana e populações indígenas nativas do Brasil.

Para a população Brasileira, basicamente composta da miscigenação de três populações, a escolha de marcadores informativos de ancestralidade deve ser realizada baseada no conhecimento prévio das freqüências alélicas nas três populações parentais e podem ser escolhidos baseados nas medidas de delta somente, como descrito em outros trabalhos (HOGGART et al., 2003; PFAFF et al., 2004; SHRIVER et al., 2003; SHRIVER et al., 1997). Contudo, essa escolha deve ser tomada trio-a-trio, ou seja, o delta deve ser maximizado para uma população de forma que, dadas três populações parentais (1, 2 e 3), o delta de um loco, por exemplo um loco A, deverá ser tal que: $\delta_{A12} \geq 0{,}60$; $\delta_{A13} \geq 0{,}60$ e $\delta_{A23} \cong 0{,}00$. Da mesma maneira, locos diferentes devem ser selecionados priorizando as demais populações de forma equivalente. Hoggart e colaboradores sugeriram que um número mínimo ideal para estudo de populações tri-parentais seria de 40 marcadores, no entanto, não deixam claro o modo como esses marcadores devem ser escolhidos (HOGGART et al., 2003). Como discutido anteriormente, o número de marcadores bi-alélicos necessários em um modelo de miscigenação de duas populações parentais é consideravelmente maior que o número exigido em um modelo de populações não miscigenadas. Logo, o mesmo se aplica a modelos de miscigenação em maior grau, de modo que, quanto maior o grau de miscigenação, tanto em termos de número de populações parentais como em termos de recorrência do evento, maior deverá ser o número de marcadores usados (ROSENBERG et al., 2003; TIAN et al., 2006).

Na sugestão de maximizar o conteúdo de informação para atribuição de ancestralidade, na bateria de 34 marcadores utilizados, 11 foram priorizados para

a população Européia, enquanto 20 para a população Africana e os dois locos restantes (OCA2 e rs803733) possuem freqüências alélicas divergentes entre Europeus e Africanos, porém a freqüência em Ameríndios é intermediária (próxima de 0,5), o que torna $\delta \geq 0,60$ somente entre as populações Européias e Africanas. Nenhum marcador foi priorizado para a população Ameríndia ou Indígena. (Locos priorizados para Europeus: WI11153, MID93, rs1426654, rs4305737, rs727563, rs734780, rs730570, rs1129038, rs1240709, rs3796384 e rs2278354; Locos priorizados para Africanos: AT3, CRH, CYP3A4, FYNULL, LPL, RB, rs1480642, rs6034866, rs7349, rs1871534, rs222541, rs267071, rs310612, rs3768641, rs3780293, rs3791896, rs4280128, rs4766807, rs730086 e rs736556). A seleção de novos marcadores favorecendo a população Ameríndia está sendo conduzida em estudos paralelos para corrigir essa distorção e melhorar as estimativas populacionais nessas amostras.

Portanto, faz-se saber que ainda são necessários estudos mais aprofundados da aplicabilidade de marcadores informativos de ancestralidade a fim de utilizar essas estimativas populacionais em estudos de associação genética. Vários fatores influenciam essas inferências, como o poder de atribuição de ancestralidade, o número de locos necessários para estudos em populações com elevado grau de miscigenação e como esses locos devem ser distribuídos de forma a maximizar o poder de atribuição de ancestralidade e minimizar o número de locos utilizados. Como descrito na literatura, tanto o número de marcadores quanto a quantidade e a qualidade de informação dos mesmos podem providenciar melhor entendimento dos níveis de miscigenação nessa amostra da população brasileira em estudos futuros (HOGGART et al., 2003; PFAFF et al., 2004; ROSENBERG et al., 2003), possibilitando o uso mais coerente de classificação de indivíduos segundo sua ancestralidade e grau de miscigenação em estudos de associação genética.

Contudo, os dados de contribuição genômica na miscigenação da população brasileira gerados por essas análises foram utilizados nas análises em seguida para avaliar a influência de ancestralidade africana sobre os genótipos e haplótipos dos genes *PTPN22* e *VDR*.

2 Padrões de Desequilíbrio de Ligação no Gene da Fosfatase da Tirosina Protéica do Tipo Não–Receptor 22 (*PTPN22*) na População Brasileira

2.1 MATERIAIS E MÉTODOS

2.1.1 Amostra Populacional

A amostragem populacional utilizada no presente estudo foi constituída de 200 indivíduos brasileiros não relacionados como descrito no Capítulo 1, assim como a extração do DNA (seções 1.1.1 e 1.1.2 respectivamente). Foram utilizadas as amostras do HapMap como referência de pesquisa e comparação dos dados genotípicos do gene *PTPN22*.

2.1.2 Seleção de SNPs e Desenho dos Iniciadores

A seleção dos SNPs no gene *PTPN22* foi feita tomando como base o SNP rs2476601, conhecido como o principal SNP associado às doenças auto-imunes nesse gene, e a partir dele foram escolhidos SNPs que pudessem conferir estrutura de blocos haplotípicos nas populações do HapMap empregando uma consulta ao banco de dados Data Rel#16c.1 phaseI june05. Os critérios de seleção para os SNPs que flanqueiam o rs2476601 foram, preferencialmente, SNPs espaçados em uma média de 5 kb e com freqüência alélica mínima de 5% no HapMap e no dbSNP no *build* 124. Foi avaliado também o padrão de desequilíbrio de ligação nessas populações como critério de seleção dos marcadores.

Definidos os SNPs utilizados no estudo, o passo seguinte foi construir iniciadores para amplificação dessas regiões específicas via PCR. Para isso, o dbSNP foi utilizado para recuperar as seqüências flanqueadoras dessas regiões. O dbSNP tem como função disponibilizar seqüências com polimorfismo de base única para a comunidade científica com detalhamento de validação, alelo, alelo ancestral

(em alguns casos), seqüência flanqueadora da região, seqüência fasta, posição no cromossomo, dados de genótipos, freqüência alélica e heterozigose (SHERRY et al., 2001).

No entanto, a construção, ou desenho, dos iniciadores deve obedecer a critérios rigorosos quanto às seqüências flanqueadoras. A primeira consideração foi verificar se as seqüências selecionadas possuem seqüências de elementos repetitivos ou de baixa complexidade. Para isso, a ferramenta Repeat Masker, disponível na internet (SMIT; HUBLEY; GREEN, 1996), foi utilizado para verificar a existência de regiões mascaradas no genoma humano a partir da seqüência original fornecida e assim prover uma nova seqüência com as regiões mascaradas identificadas.

Em seguida, a partir da seqüência retornada pelo Repeat Masker os iniciadores foram desenhados utilizando o programa Primer3 (ROZEN; SKALETSKY, 2000), disponível na versão *online*. Foram utilizados os seguintes parâmetros: número de nucleotídeos = 20; Temperatura de fusão (Tm, do inglês, *temperature of melting*) = 60°C; tamanho do produto = 100 a 400 pares de base (pb). As seqüências pequenas ou impossibilitadas de aceitar iniciadores foram completadas utilizando seqüências do genoma completo a partir do banco de dados de seqüência genômica da Universidade Santa Cruz da Califórnia (KENT et al., 2002). Os iniciadores de extensão de base única foram desenvolvidos a partir dos produtos de PCR virtuais gerados no Primer3, e foram desenhados com o propósito de que a seqüência terminasse exatamente um nucleotídeo antes do SNP em questão. A adição de polinucleotídeos não homólogos (poli T) na extremidade 5’ de cada iniciador foi necessária para ajustar o tamanho dos produtos para a reação de extensão de base única.

Para aperfeiçoar o sucesso na genotipagem do conjunto de SNPs simultaneamente, a possível formação de grampos e auto complementaridade de todos os iniciadores de PCR e todos os iniciadores SNPs foram virtualmente testados com o programa Autodimer (VALLONE; BUTLER, 2004).

2.1.3 PCR e Genotipagem

Os iniciadores foram todos testados utilizando a reação em cadeia da polimerase (PCR). A amplificação do DNA via PCR foi realizada em um volume de reação final de 12,5 µL contendo de 10 a 25 ng do DNA molde; 0,25 µM de cada iniciador (direto e reverso); 0,25 µM de dNTP; 1,5 µM de $MgCl_2$; 0,16 mg/mL de albumina sérica de bovino (BSA); 1 unidade (U) de Taq Platinum (Invitrogen); 1 X de Tampão Taq e H_2O Milli–Q estéril qsp.

O ciclo de temperaturas utilizado para amplificar o DNA foi realizado no termociclador PE GeneAmp PCR System 9700 (Applied Biosystems) baseado no processo de PCR *touchdown* (DON et al., 1991), que é constituído dos seguintes passos: desnaturação a 95ºC por 3 minutos; seguido por 20 ciclos de 30 segundos de desnaturação a 95ºC, anelamento a 63ºC por 40 segundos com decaimento de 0,5ºC a cada ciclo e extensão a 72ºC por 50 segundos; em seguida 15 ciclos de 30 segundos de desnaturação a 95ºC, anelamento a 53ºC por 40 segundos e extensão a 72ºC por 50 segundos; e por fim, 5 minutos extensão final a 72ºC.

Em seguida, o processo de purificação enzimática e reação de extensão de base única via SNaPshot™ Multiplex System (Applied Biosystems, USA) foram realizados conforme descrito no capítulo 1 seção 1.1.3, levando em consideração a concentração dos iniciadores da reação de extensão de base única.

2.1.4 Análises Estatísticas

2.1.4.1 Estatística Descritiva e Análise de Variância Molecular

Para estudos de análises genéticas o cálculo das freqüências alélicas e teste exato de Fisher das freqüências genotípicas quanto ao Equilíbrio de Hardy-Weinberg foram realizadas pelo programa PowerMarker V 3.25 (LIU, K.; MUSE, 2005).

2.1.4.2 Análise de DL e de Haplótipos

As análises do desequilíbrio de ligação entre os marcadores utilizados nas populações do HapMap e na população brasileira foram feitas de acordo com os parâmetros estatísticos D' e r^2 e os blocos de haplótipos foram definidos por Coluna sólida de DL no programa Haploview versão 3.32 (BARRETT et al., 2005). O programa Haploview possui uma interface gráfica para visualização do desequilíbrio de ligação baseado nos valores de D', r^2 e intervalo de confiança de D'.

Os haplótipos na população brasileira foram inferidos por estatística Bayesiana utilizando o programa Phase (STEPHENS, M.; DONNELLY, 2003). Neste caso foi considerada a orientação 5'-3' do gene, e não a orientação da seqüência no cromossomo. Primeiramente foi realizada a estimativa de haplótipos para os trios do HapMap nas populações CEU e YRI a fim de estabelecer as fases dos haplótipos para os indivíduos não relacionados, ou seja mãe e pai (opção -P1 recomendada pelo manual do programa). Em seguida a fase dos haplótipos desses indivíduos foi especificada como conhecida usando a opção –k no programa e foi utilizada para

aprimorar a inferência dos haplótipos das demais populações (BRA e ASN), com número de interações = 100, intervalo de remoção = 1 e *burn-in* = 100.

As estimativas de ancestralidade genômica das amostras geradas na análise 4, descritas no capítulo 1 (Tabela 1.10), foram utilizadas como fator de correlação nas análises genéticas de freqüências alélicas, genotípicas e haplotípicas. O software Whap versão 2.9 (PURCELL; DALY; SHAM, 2007) foi utilizado para correlacionar as estimativas de ancestralidade em testes de regressão com permutação condicional, em que o genótipo é condicionado a característica quantitativa (ancestralidade), enquanto o teste padrão seria associar a característica fenotípica ao genótipo. O teste de permutação *omnibus* verifica a existência de associação para todos os haplótipos e, quando significativo, é feito um teste de correlação haplótipo-específica para identificar qual dos haplótipos confere a associação. Após esse teste, foi realizado um teste de permutação local para avaliar se a ancestralidade está condicionada a algum SNP em particular. Quando identificado algum SNP, os haplótipos foram testados para o modelo nulo, ou seja, testando a associação do haplótipo sem o efeito desse SNP e sem o efeito dos outros. A média e a variância da característica quantitativa (no caso, o percentual de ancestralidade africana) foram fixadas de acordo com os valores da população para aumentar o poder de detecção da análise haplotípica. Todos esses testes foram feitos no modelo de análise condicional.

2.1.4.3 Seleção Natural

A fim de investigar se as regiões genômicas sofreram algum efeito de recente seleção natural o programa Haplotter (VOIGHT et al., 2006) foi utilizado como ferramenta eletrônica para inferências estatísticas da pontuação integral de haplótipos (iHS, do inglês *Integrated Haplotype Score*). O programa SNPBrowser (Applied Biosystems, USA) foi utilizado para identificar a extensão de DL nos genes vizinhos ao *PTPN22* nas populações do HapMap.

2.1.4.4 Transferibilidade de tagSNPS

O estudo de transferibilidade de tagSNPs foi realizado com a predição de cobertura da variabilidade usando o algoritmo Stampa (HALPERIN; KIMMEL; SHAMIR, 2005) do programa Gevalt (DAVIDOVICH; KIMMEL; SHAMIR, 2007), baseada no código aberto do Haploview. Para isso, o algoritmo foi aplicado em cada população para os SNPs selecionados nesse estudo em conjuntos de 2 a 5 marcadores e, em seguida, o cálculo da perda de variabilidade foi feito na população brasileira utilizando-se os tagSNPs selecionados para cada uma das populações do HapMap, também em conjuntos de 2 a 5 marcadores. Utilizando os dados da fase II do HapMap, foi possível verificar qual o percentual de variabilidade captado pelos marcadores selecionados em cada uma das três populações.

2.2 RESULTADOS

2.2.1 Seleção de SNPs e Desenho dos Iniciadores

A busca para polimorfismos no gene *PTPN22* retornou 23 SNPs, os quais quatro deles não possuíam dados de genotipagem em pelo menos uma das quatro populações do banco. Os 19 SNPs restantes foram submetidos à análise de regiões mascaradas no Repeat Masker e a partir dos resultados foram selecionados os SNPs que flanqueiam o rs2476601 (Figura 2.1 e Tabela 2.1). Pela dificuldade de desenhar iniciadores em regiões mascaradas foram selecionados locos que não respeitavam o critério inicial de freqüência alélica mínima de 5%, mas que possuíssem freqüências alélicas diferenciadas em outras populações além das populações do HapMap e que mesmo assim conferissem extenso bloco de desequilíbrio de ligação (Figura 2.2). Pelo mesmo motivo e pela grande distância entre o rs2476601 e o próximo marcador foi selecionado um loco que não estava genotipado nas populações do HapMap, mas no entanto possuía freqüências alélicas e genotípicas em outras populações no banco de dados do dbSNP. A tabela 2.1 descreve as características dos marcadores com relação ao cromossomo, enquanto a tabela 2.2 relata as freqüências alélicas e genotípicas nas respectivas populações para os locos selecionados. Os valores médios do desequilíbrio de ligação entre os locos foi extremamente alto para D' em todas as populações, mas o mesmo não ocorreu para os valores de r^2 (Tabela 2.3).

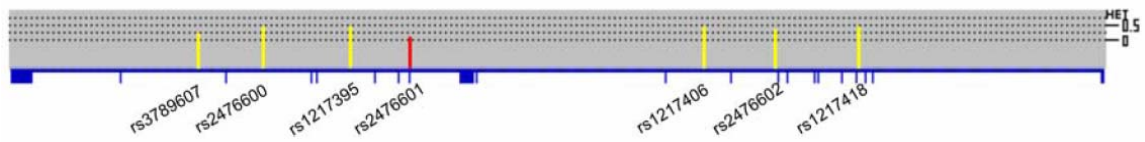


Figura 2.1 : Posição e heterozigose média dos locos no gene PTPN22 na direção 5'-3' da fita molde de DNA.
Fonte: O Autor

Tabela 2.1 : Características dos marcadores do gene PTPN22 com relação ao cromossomo e contig.

rs	Heterozigose média	Acesso do Contig	Versão do Contig	Posição no Contig	Posição no Cromossomo	Build Original	Build Atual	Alelos *	
3789607	0,23	NT_019273.18	Hs1_19429_36	10274118	114167956	107	123	Y	[C/T]
2476600	0,49	NT_019273.18	Hs1_19429_36	10277418	114171256	100	123	R	[A/G]
1217395	0,45	NT_019273.18	Hs1_19429_36	10282119	114175957	87	123	R	[A/G]
2476601	0,08	NT_019273.18	Hs1_19429_36	10285252	114179090	100	126	R	[A/G]
1217406	0,50	NT_019273.18	Hs1_19429_36	10300837	114194675	87	121	M	[A/C]
2476602	0,41	NT_019273.18	Hs1_19429_36	10304639	114198477	100	123	R	[A/G]
1217418	0,49	NT_019273.18	Hs1_19429_36	10308915	114202753	87	123	R	[A/G]

* Alelos referentes à orientação positiva (5'-3') do DNA
Fonte: O Autor

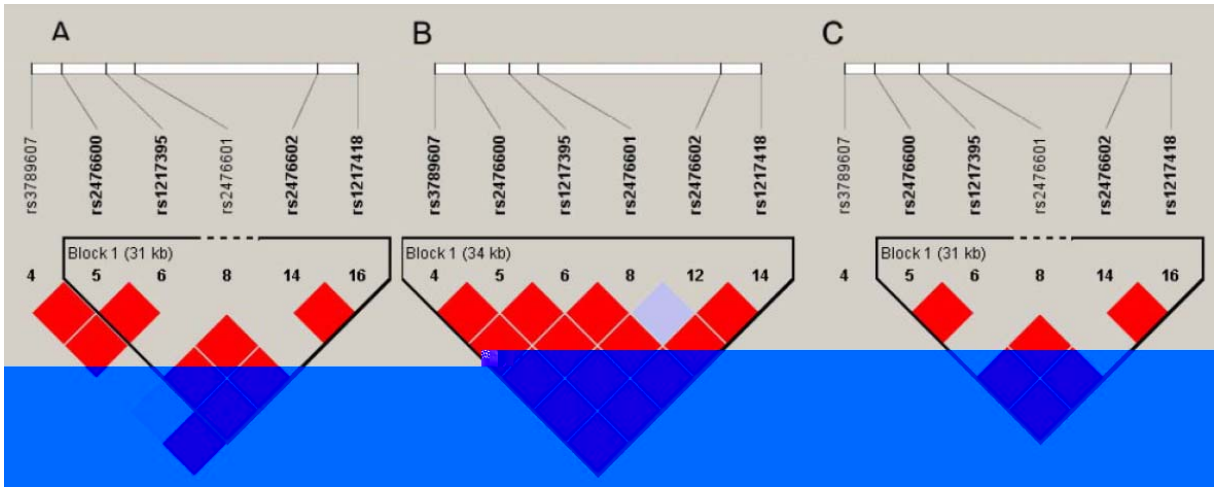


Figura 2.2 : Desequilíbrio de ligação entre os marcadores selecionados nas populações do HapMap: (A) ASN, (B) CEU e (C) YRI.
Os quadrados vermelhos indicam D' = 1 e LOD ≥ 2; e os azuis D' = 1 e LOD < 2. O loco rs2476601 é monomórfico nas populações ASN e YRI e o loco rs3789607 monomórfico na população YRI.
Fonte: O Autor

Tabela 2.2 : Freqüências alélicas e genotípicas das populações do HapMap e Sanger para os locos selecionados no gene PTPN22.
As populações SC_12_A, SC_12_AA e SC_12_C se referem respectivamente a populações Asiática, Afro-americana e Caucasiana da América do Norte, submetidas pelo grupo TSC-CSHL-Sanger. Alelos descritos como 1 e 2 são referentes à tabela 2.1.

SNP	População	Número de Amostras	Freqüência alélica		Freqüência genotípica		
			1	2	11	12	22
rs3789607	CEU	60	0,283	0,717	0,067	0,433	0,500
	CHB	90	0,044	0,956	-	0,089	0,911
	JPT	88	0,080	0,920	-	0,159	0,841
	YRI		-	1,000	-	-	1,000
rs2476600	CEU	60	0,500	0,500	0,217	0,567	0,217
	CHB	90	0,878	0,122	0,756	0,244	-
	JPT	88	0,773	0,227	0,568	0,409	0,023
	YRI	118	0,246	0,754	0,085	0,322	0,593
rs1217395	CEU	60	0,725	0,275	0,533	0,383	0,083
	CHB	90	0,378	0,622	0,156	0,444	0,400
	JPT	88	0,455	0,545	0,205	0,500	0,295
	YRI	118	0,942	0,058	0,883	0,117	-
rs2476601	CEU	60	0,142	0,858	-	0,283	0,717
	CHB	90	-	1,000	-	-	1,000
	JPT	88	-	1,000	-	-	1,000
	YRI	60	-	1,000	-	-	1,000
rs1217406	SC_12_A	20	0,800	0,200	0,700	0,200	0,100
	SC_12_AA	20	0,300	0,700	0,100	0,400	0,500
	SC_12_C	14	0,500	0,500	0,286	0,429	0,286
rs2476602	CEU	60	0,217	0,783	0,017	0,400	0,583
	CHB	90	0,078	0,922	-	0,156	0,844
	JPT	88	0,148	0,852	0,023	0,250	0,727
	YRI	60	0,717	0,283	0,533	0,367	0,100
rs1217418	CEU	60	0,500	0,500	0,217	0,567	0,217
	CHB	90	0,878	0,122	0,756	0,244	-
	JPT	88	0,761	0,239	0,545	0,432	0,023
	YRI	118	0,254	0,746	0,085	0,339	0,576

Fonte: O Autor

Tabela 2.3 : Valores do desequilíbrio de ligação nas populações do HapMap para os locos escolhidos no gene PTPN22.

Loco 1	Loco 2	Distância (pb)	CEU					YRI					ASN				
			D'	LOD	r²	IC inf	IC sup	D'	LOD	r²	IC inf	IC sup	D'	LOD	r²	IC inf	IC sup
rs3789607	rs2476600	3300	1	12,930	0,395	0,840	1						1	5,860	0,312	0,700	1
rs3789607	rs1217395	8001	1	5,480	0,150	0,710	1						1	2,610	0,093	0,460	1
rs3789607	rs2476601	11134	1	2,690	0,065	0,490	1										
rs3789607	rs2476602	30521	1	4,050	0,109	0,630	1						1	0,590	0,008	0,090	0,990
rs3789607	rs1217418	34797	1	12,930	0,395	0,840	1						1	5,670	0,301	0,690	1
rs2476600	rs1217395	4701	1	10,610	0,379	0,820	1	1	3,970	0,194	0,570	1	1	8,180	0,296	0,800	1
rs2476600	rs2476601	7834	1	4,530	0,165	0,650	1										
rs2476600	rs2476602	27221	1	6,400	0,277	0,740	1	1	20,600	0,805	0,880	1	1	12,990	0,600	0,840	1
rs2476600	rs1217418	31497	1	33,110	1,000	0,950	1	1	24,870	0,956	0,910	1	1	25,080	0,962	0,920	1
rs1217395	rs2476601	3133	1	10,730	0,435	0,800	1										
rs1217395	rs2476602	22520	1	2,390	0,105	0,440	1	1	3,480	0,157	0,530	1	1	4,500	0,178	0,660	1
rs1217395	rs1217418	26796	1	10,610	0,379	0,820	1	1	3,860	0,185	0,560	1	1	8,410	0,308	0,800	1
rs2476601	rs2476602	19387	1	0,460	0,046	0,080	0,98										
rs2476601	rs1217418	23663	1	4,530	0,165	0,650	1										
rs2476602	rs1217418	4276	1	6,400	0,277	0,740	1	1	21,900	0,842	0,890	1	1	12,550	0,578	0,840	1
Média		17252	1	8,523	0,289	0,680	0,999	1	13,113	0,523	0,723	1	1	8,644	0,364	0,680	0,999

Fonte: O Autor

Pelo banco de SNP disponível no NCBI e com auxílio de ferramentas de programação foram desenhados iniciadores para as regiões descritas. O desenho dos iniciadores de PCR e de extensão de base única, assim como os detalhes de tamanho de fragmento e alelos amplificados pode ser observado na tabela 2.4.

Tabela 2.4 : Seqüência e detalhes dos iniciadores para o gene PTPN22.

SNP	Primer	Seqüência	Direção do primer	Tamanho do primer (nts)	Tm (°C)
rs3789607	F	TTCTTTGCACTTGGCTGTTTT	direto	21	60
	R	CACCAACTCATGATGGCTGA	reverso	20	61
	S	$(T)_{10}$CATACCTTTTCTCCTAAAAAGAGTC	reverso	35	56
rs2476600	F	TCTTCAAGGAACCTACCCAAA	direto	21	59
	R	TCCATGGAAGAGCACATTTCT	reverso	21	60
	S	$(T)_{9}$CTCAAAGGTGTTCTGTTCCA	direto	29	60
rs1217395	F	TGTGCACCTTACACAGGGTTA	direto	21	59
	R	TCTTCCACTCAGCGAAACCT	reverso	20	60
	S	$(T)_{17}$CACAGATTATCACTTAGTGGTTCCA	reverso	42	57
rs2476601	F	CCAGCTTCCTCAACCACAAT	direto	20	60
	R	TTCCTTGAATGAACAAGTGTCAA	reverso	23	60
	S	$(T)_{21}$CACAATAAATGATTCAGGTGTCC	direto	44	58
rs1217406	F	TGAAGGCTTTTTCAGCGTCT	direto	20	60
	R	CAACCTGTGAGGGAGAAAGC	reverso	20	60
	S	$(T)_{9}$ATAACTAGGTTCCATACTTCTG	direto	31	50
rs2476602	F	ACCTGGGAACTAATTTAATC	direto	20	50
	R	GCTGTGGAAGGACTGGTGTT	reverso	20	60
	S	$(T)_{14}$ACTTGTAGACCCACTCTGTCAGA	reverso	37	55
rs1217418	F	GACCCTGGGTGGCAATATAA	direto	20	60
	R	AGCGGAGGACTAGGTGAGAA	reverso	20	59
	S	$(T)_{4}$GAAATTACACGGGGTGACTACA	reverso	26	59

Fonte: O Autor

2.2.2 Genotipagem e Condições de PCR

O sistema multiplex permitiu uma tipagem rápida e eficiente dos sete locos em apenas uma reação de PCR seguida de tratamento enzimático e reação de SNaPshot™ em multiplex. A concentração dos iniciadores (Tabela 2.5) utilizados na reação SNaPshot™ teve que ser diferente entre os locos a fim de manter equivalente o balanço de intensidade entre os picos em unidades relativa de fluorescência (rfu). A utilização de termociclagem *touchdown* tanto na PCR quanto na reação de extensão de base única foi utilizado para evitar o comprometimento da qualidade da amplificação com o aparecimento de produtos inespecíficos que dificultariam a análise.

Foi possível visualizar todos os alelos dos locos bi-alélicos (Figura 2.3), exceto para o loco rs1217406, cujo o único alelo amplificado (G) era diferente dos alelos esperados (A/C). Uma análise minuciosa revelou a ausência do nucleotídeo G na extremidade 3' do primer S deste loco, levando à exclusão do mesmo nas análises que seguem.

Tabela 2.5 : Concentração dos Iniciadores S na reação SNaPshot™.

Loco	Concentração (μM)
rs1217418	0,300
rs1217395	0,450
rs3789607	0,450
rs2476600	0,525
rs2476601	0,525
rs2476602	1,000

Fonte: O Autor

2.2.3 Análises Genéticas

Para finalidade de análise genética, os locos foram organizados no sentido da fita reversa, uma vez que o gene se posiciona no sentido 3'-5'. Os bancos de dados de SNPs disponíveis no NCBI e no HapMap quase sempre relatam os SNPs de acordo com a posição no cromossomo, ou no contig, em relação a orientação positiva do DNA, no entanto, em genes que seguem a orientação negativa do DNA, os haplótipos são relatados na mesma orientação do gene. Portanto os alelos genotipados foram convertidos para seus respectivos alelos na orientação 5'-3' do gene (Quadro 2.1).

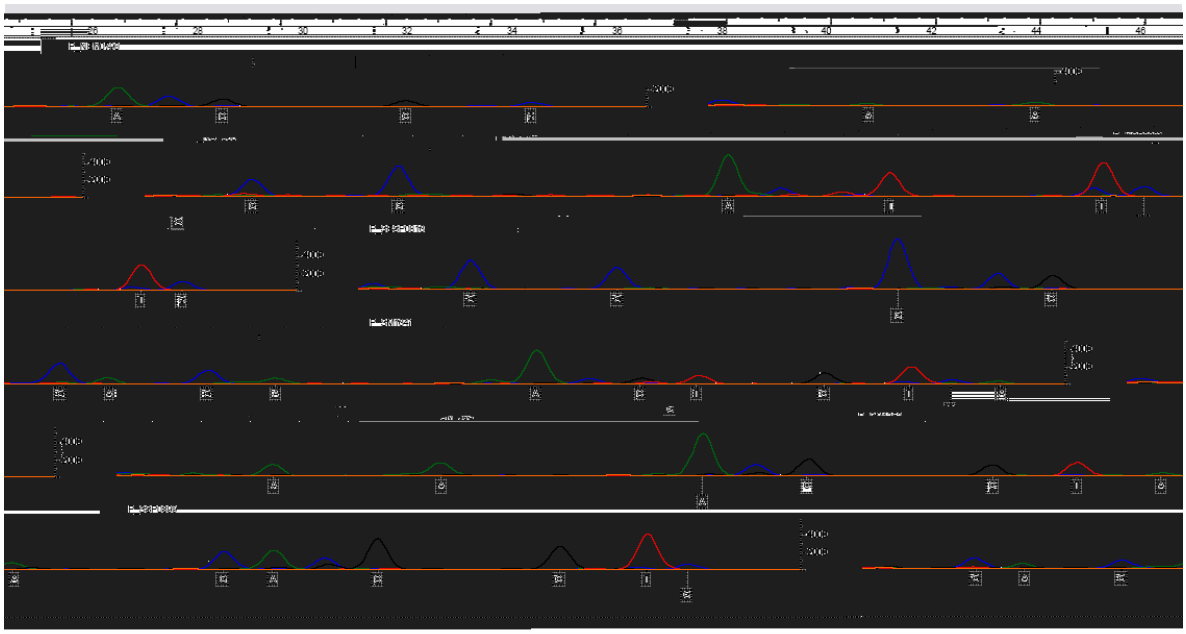


Figura 2.3 : Eletroferograma do sistema SNaPshot™ em multiplex para o gene PTPN22 com todos os 12 alelos e todos os 18 genótipos possíveis dos seis locos estudados. Amostras: NEM0790, NSP0301, SESP0813, SM1341, SM1353 e SSP0399.
Fonte: O Autor

Quadro 2.1 : Conversão dos alelos para estimativa dos haplótipos no gene *PTPN22*.

	Locos											
	rs1217418		rs2476602		rs2476601		rs1217395		rs2476600		rs3789607	
Alelo Genotipado	G	A	C	T	G	A	C	T	G	A	G	A
Alelo 3'-5'	C	T	C	T	C	T	C	T	C	T	G	A

Fonte: O Autor

A análise de dados de freqüências alélicas para a população Brasileira revelou freqüências alélicas mínimas maiores que 0,05 em todos os seis locos avaliados neste estudo, mesmo levando em consideração a divisão em subpopulações por região geográfica (Tabela 2.6). A distribuição das freqüências alélicas se mostrou homogênea entre as amostras regionais, com os alelos de menor freqüência mantidos entre as subpopulações.

Tabela 2.6 : Distribuição das freqüências alélicas dos seis locos na população brasileira e respectivas regiões.

População	rs1217418		rs2476602		rs2476601		rs1217395		rs2476600		rs3789607	
	C	T	C	T	C	T	C	T	C	T	G	A
BRA	0,551	0,449	0,697	0,303	0,927	0,073	0,301	0,699	0,569	0,431	0,229	0,771
BRA CO	0,500	0,500	0,795	0,205	0,934	0,066	0,313	0,688	0,550	0,450	0,262	0,738
BRA Ne	0,587	0,412	0,703	0,297	0,936	0,064	0,262	0,738	0,615	0,385	0,275	0,725
BRA N	0,526	0,474	0,641	0,359	0,923	0,077	0,397	0,603	0,513	0,487	0,162	0,837
BRA Se	0,566	0,434	0,684	0,316	0,919	0,081	0,225	0,775	0,551	0,449	0,225	0,775
BRA S	0,577	0,423	0,662	0,338	0,921	0,079	0,308	0,692	0,615	0,385	0,218	0,782

Fonte: O Autor

A análise de freqüência genotípica revelou uma heterogeneidade maior tanto na população total como nas amostras regionais principalmente nos locos rs2476600, rs1217395, rs2476602 e rs1217418 (Tabela 2.7), os quais as freqüências alélicas são diferenciadas nas populações do HapMap (Tabela 2.2).

Tabela 2.7 : Distribuição das [illegible]s dos seis locos na população brasileira e respectivas regiões.

	Freqüências genotípicas																
	Locos / Genótipos																
	rs1217418			rs2476602			rs2476601			rs1217395			rs2476600			rs3789607		
População	C/C	C/T	T/T	C/C	C/T	T/T	C/C	C/T	T/T	C/C	C/T	T/T	C/C	C/T	T/T	G/G	G/A	A/A
BRA	0,318	0,467	0,215	0,518	0,328	0,154	0,862	0,092	0,026	0,092	0,426	0,497	0,359	0,421	0,221	0,070	0,317	0,613
BRA CO	0,250	0,500	0,250	0,641	0,308	0,051	0,895	0,079	0,026	0,100	0,425	0,475	0,375	0,350	0,275	0,075	0,375	0,550
BRA Ne	0,375	0,425	0,200	0,568	0,270	0,162	0,872	0,128	0,000	0,075	0,375	0,550	0,436	0,359	0,205	0,100	0,350	0,550
BRA N	0,289	0,474	0,237	0,462	0,359	0,179	0,872	0,103	0,026	0,179	0,436	0,385	0,289	0,447	0,263	0,000	0,325	0,675
BRA Se	0,342	0,447	0,211	0,474	0,421	0,105	0,892	0,054	0,054	0,025	0,400	0,575	0,308	0,487	0,205	0,100	0,250	0,650
BRA S	0,333	0,487	0,179	0,514	0,297	0,189	0,868	0,105	0,026	0,077	0,462	0,462	0,385	0,462	0,154	0,077	0,282	0,641

Fonte: O Autor

Tabela 2.7 : Distribuição das

A partir das freqüências genotípicas foi estimado o equilíbrio de Hardy-Weinberg (EHW) com teste exato de Fisher. Dos seis locos apenas o rs2476601 e o rs2476602 não se encontram em EHW para a população brasileira como um todo (Tabela 2.8). Quando separados em regiões, as amostras regionais Norte, Sul e Centro-Oeste apresentaram todos os locos em equilíbrio de Hardy-Weinberg (Tabela 2.8). A amostra regional do Nordeste apresentou o loco rs2476602 fora do equilíbrio esperado segundo Hardy-Weinberg e a do Sudeste apresentou o loco rs2476601 na mesma situação (Tabela 2.8).

O índice de fixação (Fst) foi medido na população brasileira dividida por amostras regionais e não revelou diferença genética significativa (Fst = 0,003 e intervalos de confiança a 95%: superior = 0,002 e inferior = -0,008). O loco rs2476601 não apresentou diferença genética significativa nas amostras regionais em relação a população total brasileira (Theta-P = -0,0166). A análise par a par para a população brasileira e as demais populações revelou pequena diferença genética entre a população brasileira e CEU, e de moderada a alta para as populações YRI e ASN (Tabela 2.9). Quando comparadas entre si, as populações do HapMap possuem alta diferença genética, especialmente entre as populações YRI e ASN (Tabela 2.9).

Tabela 2.8 : Distribuição das populações segundo o loco, o número de amostras genotipadas (N), heterozigose observada (Ho), heterozigose esperada (He) e valor de p para o teste exato do equilíbrio de Hardy-Weinberg (p-val). Números em negrito indicam p < 0,05.

População	Loco	N	Ho	He	p-val
Centro-Oeste	rs1217418	40	0,500	0,500	1,0000
	rs2476602	39	0,308	0,326	0,6430
	rs2476601	38	0,079	0,123	0,1390
	rs1217395	40	0,425	0,430	1,0000
	rs2476600	40	0,350	0,495	0,0640
	rs3789607	40	0,375	0,387	1,0000
Nordeste	rs1217418	40	0,425	0,485	0,5190
	rs2476602	37	0,270	0,418	**0,0450**
	rs2476601	39	0,128	0,120	1,0000
	rs1217395	40	0,375	0,387	1,0000
	rs2476600	39	0,359	0,473	0,1720
	rs3789607	40	0,350	0,399	0,3910
Norte	rs1217418	38	0,474	0,499	0,7670
	rs2476602	39	0,359	0,460	0,1730
	rs2476601	39	0,103	0,142	0,1860
	rs1217395	39	0,436	0,479	0,7410
	rs2476600	38	0,447	0,500	0,5220
	rs3789607	40	0,325	0,272	0,5840
Sudeste	rs1217418	38	0,447	0,491	0,7360
	rs2476602	38	0,421	0,432	1,0000
	rs2476601	37	0,054	0,149	**0,0070**
	rs1217395	40	0,400	0,349	0,6660
	rs2476600	39	0,487	0,495	1,0000
	rs3789607	40	0,250	0,349	0,0780
Sul	rs1217418	39	0,487	0,488	1,0000
	rs2476602	37	0,297	0,447	0,0580
	rs2476601	38	0,105	0,145	0,1720
	rs1217395	39	0,462	0,426	0,7160
	rs2476600	39	0,462	0,473	1,0000
	rs3789607	39	0,282	0,341	0,3280
Total	rs1217418	195	0,467	0,495	0,4621
	rs2476602	190	0,332	0,422	**0,0034**
	rs2476601	191	0,094	0,136	**< 0,0001**
	rs1217395	198	0,419	0,420	0,8651
	rs2476600	195	0,421	0,490	0,0543
	rs3789607	199	0,317	0,353	0,1566

Fonte: O Autor

Tabela 2.9 : Matriz do índice de fixação (Fst) par a par entre as populações. Diagonal inferior representa os valores de Fst e diagonal superior seus respectivos testes de significância com intervalos de confiança superior (ic Sup) e inferior (ic Inf) a 95%.

	BRA	YRI	ASN	CEU
BRA		0,234 ic Sup 0,083 ic Inf	0,230 ic Sup 0,102 ic Inf	0,014 ic Sup 0,002 ic Inf
YRI	0,151		0,530 ic Sup 0,438 ic Inf	0,344 ic Sup 0,132 ic Inf
ASN	0,176	0,495		0,190 ic Sup 0,105 ic Inf
CEU	0,007	0,228	0,161	

Fonte: O Autor

2.2.4 Padrões de Desequilíbrio de Ligação

A análise de desequilíbrio de ligação mostrou que, na população brasileira, todos os locos estão em DL entre si, com valores de D' entre 0,777 e 1,0 e valores de r^2 variando de 0,014 a 0,818 (Tabela 2.10). Quando as amostras são separadas por região, o desequilíbrio de ligação cai e é quebrado entre os marcadores para as regiões Centro-Oeste, Sudeste e Sul (Figura 2.4). Não foi possível visualizar a definição de blocos haplotípicos utilizando os parâmetros de intervalo de confiança propostos por Gabriel (2002), porque os valores dos intervalos de confiança superior e inferior são menores do que os necessários. No entanto, foi possível identificar blocos a partir dos valores de D' por coluna sólida de DL, nos quais a coluna é estendida para cada D'>0,8.

A estimativa bayesiana de haplótipos para o gene *PTPN22* foi realizada com os genótipos em fase das populações CEU e YRI do HapMap, e os genótipos com fase desconhecida da população ASN do HapMap e a população brasileira, com finalidade de incrementar o poder de inferência estatística. A estimativa revelou 17 haplótipos para a população brasileira com freqüências variando de 0,286 à 0,003 (Tabela 2.11). Quando analisada por amostras regionais

separadas, a distribuição de haplótipos se mostrou diferenciada. A região Sul e Centro-Oeste apresentaram 13 haplótipos cada, com 10 haplótipos em comum diferenciados em freqüências. Em seguida, a região Sudeste apresentou 11 haplótipos, a região Norte 7 e a nordeste 6.

Tabela 2.10 : Testes de desequilíbrio de ligação par a par baseados em D' e r² para a população brasileira. O logaritmo de odd score (LOD) para D' e intervalos de confiança (IC) inferior (inf) e superior (sup) para r²

Loco 1	Loco 2	D'	LOD	r²	IC inf	IC sup	Distância
rs3789607	rs2476600	1,000	15,250	0,217	0,890	1,000	3300
rs3789607	rs1217395	0,935	7,050	0,109	0,710	0,990	8001
rs3789607	rs2476601	0,777	1,180	0,014	0,200	0,940	11134
rs3789607	rs2476602	0,899	8,320	0,107	0,700	0,970	30521
rs3789607	rs1217418	1,000	15,430	0,234	0,890	1,000	34797
rs2476600	rs1217395	0,808	22,050	0,370	0,700	0,880	4701
rs2476600	rs2476601	0,917	5,760	0,090	0,650	0,980	7834
rs2476600	rs2476602	0,926	19,600	0,306	0,810	0,980	27221
rs2476600	rs1217418	0,933	61,120	0,818	0,880	0,970	31497
rs1217395	rs2476601	0,813	6,400	0,123	0,570	0,930	3133
rs1217395	rs2476602	0,826	7,690	0,130	0,620	0,920	22520
rs1217395	rs1217418	0,815	20,320	0,350	0,700	0,890	26796
rs2476601	rs2476602	1,000	2,360	0,037	0,430	1,000	19387
rs2476601	rs1217418	0,909	5,030	0,083	0,620	0,980	23663
rs2476602	rs1217418	0,928	20,370	0,323	0,820	0,980	4276
Média		0,899	14,529	0,221	0,679	0,961	-

Fonte: O Autor

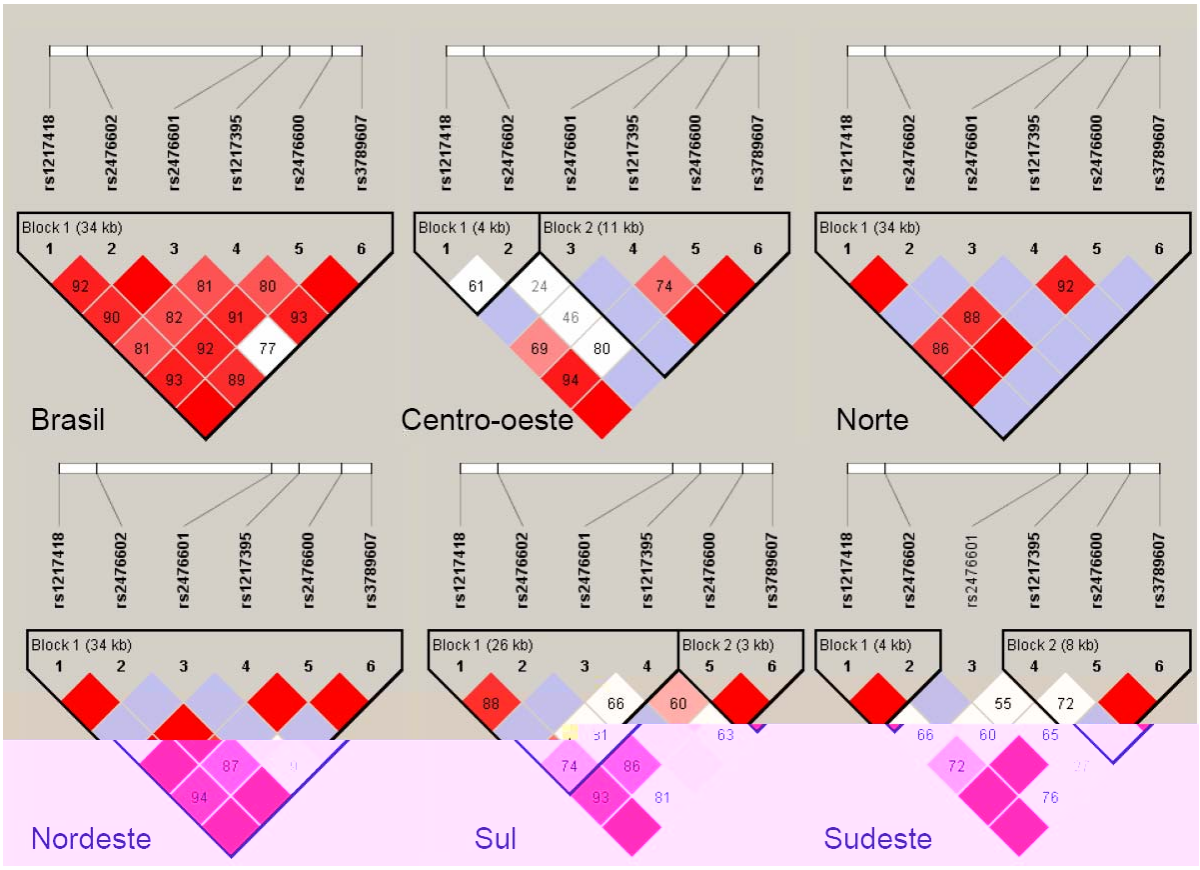


Figura 2.4 : Desequilíbrio de ligação entre os marcadores selecionados na população brasileira e nas amostras regionais. O número nos quadrados indica o valor percentual de D'. Quadrados entre rosa e vermelho indicam D' < 1 e LOD ≥ 2 e quadrados brancos D' < 1 e LOD < 2.
Fonte: O Autor

Tabela 2.11 : Identificação dos haplótipos (Hap ID) e distribuição das freqüências haplotípicas nas populações.

Hap ID	Haplótipo	CEU	YRI	ASN	BRA	CO	NE	N	SE	S
H01	TCCTTA	0,225	0,183	0,242	0,151	0,175	0,113	0,113	0,238	0,114
H02	CCCTTA	-	-	-	0,005	-	-	-	0,013	0,013
H03	TCTTTA	-	-	-	0,005	-	-	-	0,013	0,013
H04	TCCCTA	0,133	0,058	0,579	0,196	0,188	0,200	0,288	0,138	0,165
H05	CCCCTA	-	-	0,006	0,008	0,013	-	0,013	0,013	-
H06	TTCCTA	-	-	-	0,005	0,013	-	-	-	0,013
H07	TCTCTA	0,142	-	-	0,063	0,063	0,063	0,075	0,050	0,063
H08	TCCTCA	-	0,008	-	0,020	0,038	0,038	-	-	0,025
H09	CCCTCA	-	0,033	-	0,008	0,038	-	-	-	-
H10	TTCTCA	-	-	-	0,003	0,013	-	-	-	-
H11	CTCTCA	0,217	0,717	0,112	0,286	0,175	0,313	0,338	0,313	0,316
H12	TCCCCA	-	-	-	0,008	0,013	-	-	-	0,025
H13	CTCCCA	-	-	-	0,015	0,013	-	0,013	0,025	0,025
H14	CCCTCG	0,283	-	0,062	0,216	0,250	0,275	0,163	0,200	0,190
H15	CTCTCG	-	-	-	0,005	-	-	-	0,013	0,013
H16	CCTTCG	-	-	-	0,003	-	-	-	0,013	-
H17	CCCCCG	-	-	-	0,005	0,013	-	-	-	0,013

Fonte: O Autor

Para avaliar o efeito da ancestralidade genômica, foram feitos testes de correlação condicional em todos os haplótipos e em haplótipos específicos. O teste de regressão *omnibus* levou em consideração somente os haplótipos mais comuns, excluindo os haplótipos com freqüência abaixo de 0,01. Com isso 6,2% de todos os haplótipos foram retirados da análise e somente 180 dos 200 indivíduos foram analisados. De todos os 17 haplótipos revelados pelo Phase, o Whap retornou apenas sete (Tabela 2.12), dos quais três são comuns às populações CEU, YRI e ASN, um exclusivo da população YRI, um exclusivo da população CEU e um comum às populações CEU e ASN (Tabela 2.11) e foi detectada uma associação da ancestralidade condicionada aos haplótipos (p = 0,006). Em seguida, a análise haplótipo-específica indicou que a associação ocorre em dois desses sete haplótipos, o H14 e o H04, com coeficientes de regressão β não padronizado igual a -0,036 e 0,054, respectivamente.

Tabela 2.12 : Teste de regressão condicional de haplótipo-específico, seus coeficientes de regressão (β) e valor de p para o teste de χ^2. Números em negrito indicam p<0,05 .

Haplótipo	Freqüência	β	χ^2	p-val
H11	0,304	0,014	0,730	0,393
H14	0,231	-0,036	3,904	**0,048**
H04	0,204	0,054	9,000	**0,003**
H01	0,158	-0,034	2,432	0,119
H07	0,068	-0,014	0,179	0,672
H08	0,022	0,023	0,206	0,650
H13	0,014	-0,116	3,157	0,076

Fonte: O Autor

O teste de permutação local avaliou o efeito de cada SNP e identificou dois locos que significativamente exercem influência na associação da ancestralidade com os haplótipos (Tabela 2.13) e a influência de cada alelo (Tabela 2.14).

Tabela 2.13: Teste de permutação local e valor de p para o teste da razão de verossimilhança (LRT). Números em negrito indicam p<0,05.

Loco	LRT	p-val
rs1217418	1,964	0,161
rs2476602	0,251	0,617
rs2476601	0,982	0,322
rs1217395	4,402	**0,036**
rs2476600	1,180	0,277
rs3789607	3,992	**0,046**

Fonte: O Autor

Tabela 2.14 : Teste de permutação local com seus coeficientes de regressão (β) e valor de p para a razão de verossimilhança (LRT).

Loco	Alelo	Freqüência	β	LRT	p-val
rs1217395	T	0,699	0,000	4,402	0,036
	C	0,301	0,034		
rs3789607	A	0,771	0,000	3,992	0,046
	G	0,229	-0,035		

Fonte: O Autor

Os haplótipos foram testados para o modelo nulo dos SNPs identificados. No modelo nulo, primeiramente foram excluídos os dois locos juntos e, em seguida, foram realizados os testes excluindo cada loco de uma vez. Para os dois locos, a regressão *omnibus* continuou significativa (p = 0,017), assim como a regressão haplótipo-específica (p_{H14} = 0,048 e p_{H04} = 0,003). Com o modelo nulo para os locos rs1217395 e rs3789607, as regressões *omnibus* e haplótipo-específica apresentaram valores significativos (p = 0,017, p_{H14} = 0,048 e p_{H04} = 0,003; p =0,015, p_{H14} = 0,048 e p_{H04} = 0,003, respectivamente).

O alelo T do SNP rs2476601 foi testado para correlação com ancestralidade Européia pelo teste de permutação local e não revelou significância (p = 0,318). A análise caso a caso dos portadores do alelo T revelou que a média de ancestralidade Africana é de 0,190, com máximo de 0,501, mínimo de 0,027 e mediana em 0,174.

2.2.5 Seleção Positiva no Genoma

O programa Haplotter revelou indícios de que a região cromossômica de aproximadamente 1Mb perto do gene *PTPN22* possui recente seleção positiva no genoma humano (Figura 2.5). Não houve indícios de recente seleção positiva para o gene *PTPN22* propriamente dito (p_{CEU}=0,075, p_{YRI}=0,022 e p_{ASN}=0,044; significância a 1%), mas na população YRI, esta seleção é significativa para os genes que flanqueiam o *PTPN22* na região 5', o *MAGI3* (p=0,009), *PHTF1* (p=0,008) e *RSBN1* (p=0,008). Nas demais populações, apesar dos valores de iHS serem altos, eles não são significativos (0,015 < p < 0,191).

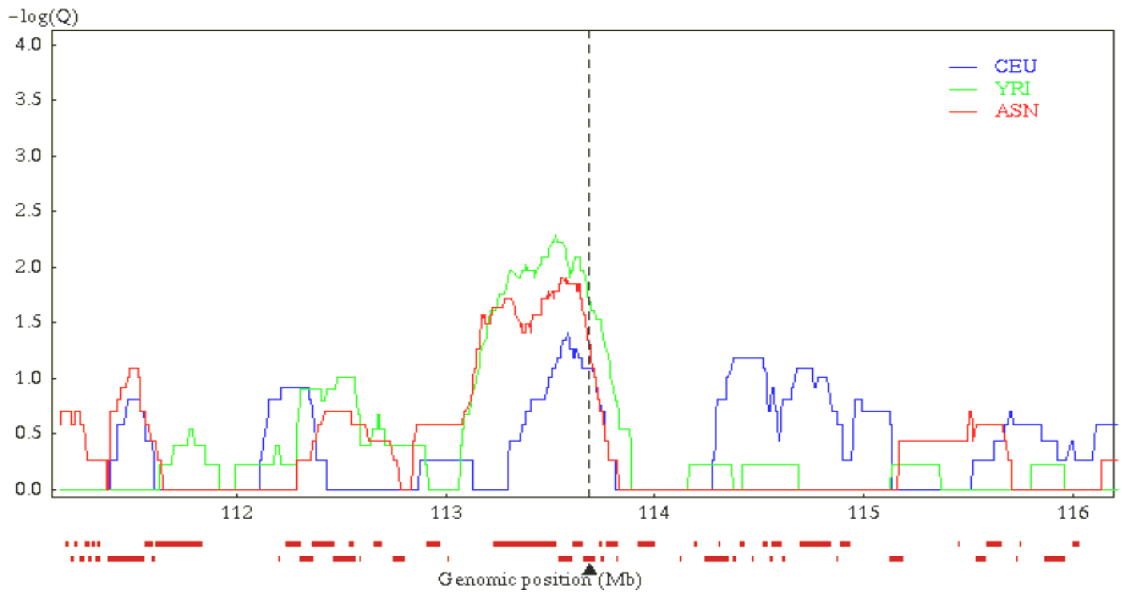
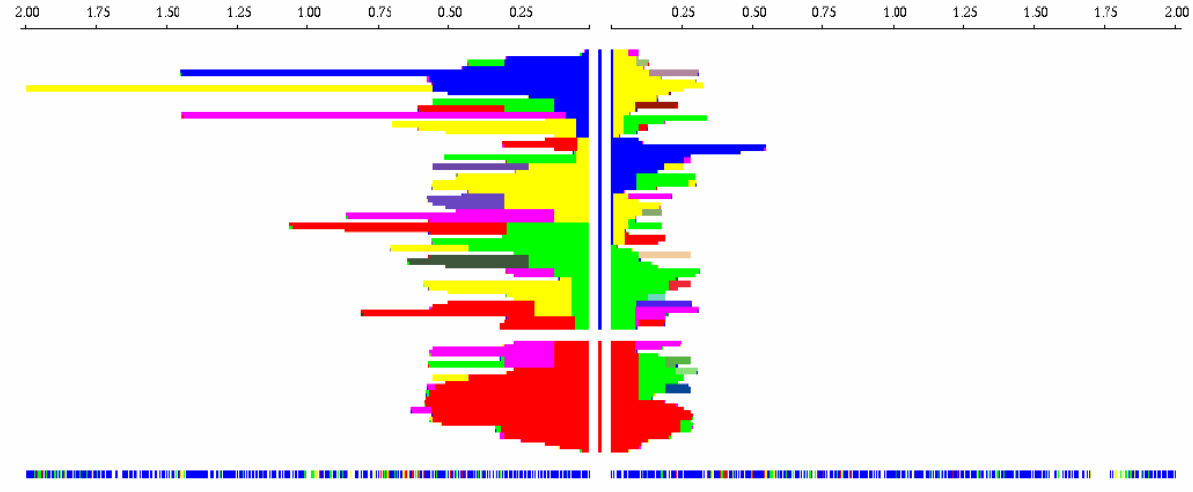


Figura 2.5 : Gráfico da pontuação integral de haplótipos na região genômica do *PTPN22*.
A linha pontilhada representa a posição cromossômica do gene. O eixo das ordenadas representa o log negativo do rank da estatística |iHS| observada para um determinado SNP[–log(Q)]. Para cada SNP, 25 SNPs de cada lado são avaliados para |iHS|>2. A proporção de SNPs nesse painel de 51 SNPs com |iHS|>2 é computado e dividido pelo número total de SNPs (Q).
Fonte: programa Haplotter, VOIGHT, 2006

Para cada população foi possível identificar SNPs que possuíssem valores altos e significativos de iHS. Dos marcadores avaliados nesse estudo, foram encontrados quatro SNPs com significativa pontuação integral de haplótipos. Na população CEU, o rs3789607 apresentou o alelo G sobre efeito de seleção com iHS igual a -2,928, enquanto na população YRI, o rs2476600 apresentou o alelo T sobre efeito de seleção com iHS igual a 2,538. A distância em megabases nas quais os haplótipos estão distribuídos a partir do SNP em questão está representada no topo dos gráficos da figura 2.6.

Com a identificação de associação de ancestralidade com o loco rs1217395 na população brasileira, foi avaliado o potencial indicativo de seleção nas populações parentais. Foi encontrado indicativo de seleção significativa para o alelo C somente na população ASN, com iHS de -2,074 (Figura 2.7A). Os demais SNPs também foram avaliados e somente o rs2476601 foi sobre efeito de seleção com iHS de -2,130 (Figura 2.7B). A figura 2.8 mostra a extensão do DL nos genes vizinhos ao *PTPN22* nas populações do HapMap.

A



B

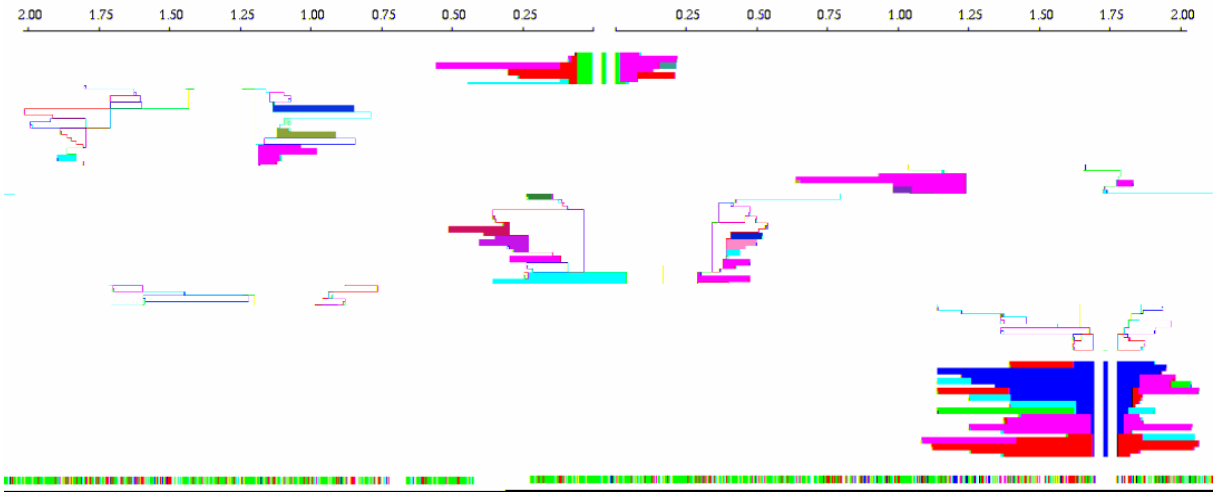


Figura 2.6 : Distribuição de haplótipos na região do SNP rs3789607 da população CEU (A) e na região do SNP rs2476600 da população YRI (B).
No primeiro caso (A) o alelo derivado (coluna vermelha central) está sobre efeito de seleção sobre o alelo ancestral, enquanto no segundo caso (B) o alelo ancestral está sobre efeito de seleção do outro alelo. O tamanho das colunas representa a freqüência alélica do respectivo SNP. As linhas horizontais ao lado das colunas representam os haplótipos. Uma linha de uma cor representa um bloco haplotípico e uma nova cor representa um novo haplótipo daquele ponto em diante. Os blocos são interrompidos quando os haplótipos se tornam únicos em toda a população. A barra inferior representa a posição dos SNPs adjacentes, onde os SNPs com freqüência alélica intermediária estão em azul (alelo de menor freqüência > 0,2). Os lados direito e esquerdo são distribuídos independentemente.
Fonte: programa Haplotter, VOIGHT, 2006

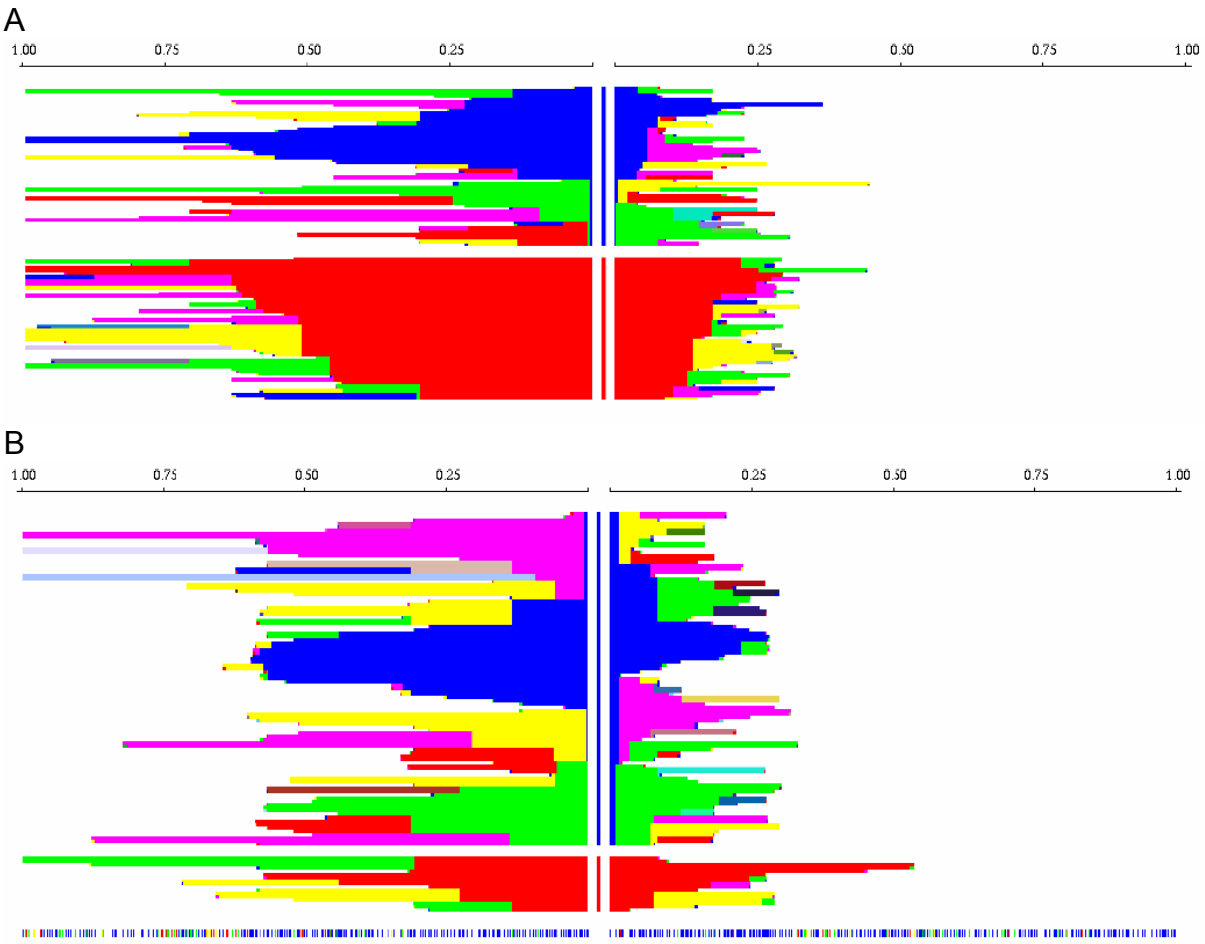


Figura 2.7 :Distribuição de haplótipos na região do SNP rs1217395 da população ASN (A) e do rs2476601 para a população CEU (B)
Fonte: programa Haplotter, VOIGHT, 2006

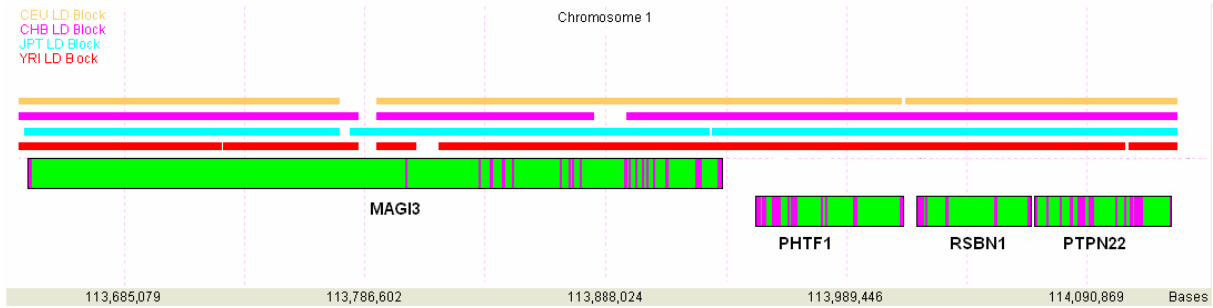


Figura 2.8 : Extensão do desequilíbrio de ligação e blocos haplotípicos nos genes vizinhos ao *PTPN22* nas populações do HapMap. De laranja a população CEU, de rosa CHB, de azul JPT e de vermelho YRI.
Fonte: O Autor

2.2.6 Transferibilidade de tagSNPs

A análise de transferibilidade de tagSNPs entre a população brasileira e as populações do HapMap mostrou que os níveis de ancestralidade influenciam a predição de cobertura da variabilidade genética na população brasileira. A tabela 2.15 descreve os tagSNPs para cada população e a predição de variabilidade no gene para cada número de marcadores usados.

Tabela 2.15 : Percentual de variabilidade captado por cada conjunto de marcadores em cada população.

	Populações			
	CEU	ASN	YRI	BRA
Número de marcadores	Percentual de variabilidade			
2	83,75	89,04	95,00	84,17
rs1217418			V	
rs2476602				V
rs2476601				
rs1217395				
rs2476600	V	V		
rs3789607	V	V	V	V
3	78,33	95,51	97,78	81,74
rs1217418	V			V
rs2476602	V		V	V
rs2476601				
rs1217395		V	V	
rs2476600	V	V	V	V
rs3789607		V		
4	67,5	93,82	100	77,39
rs1217418	V	V	V	V
rs2476602	V	V	V	V
rs2476601	V			
rs1217395		V	V	V
rs2476600	V	V	V	
rs3789607				V
5	76,67	100	100	88,44
rs1217418	V	V	V	V
rs2476602	V	V	V	V
rs2476601	V			
rs1217395		V	V	V
rs2476600	V	V	V	V
rs3789607	V	V	V	V

Fonte: O Autor

O cálculo de perda relativa de variabilidade genética pela transferibilidade de tagSNPs das populações do HapMap para a população brasileira mostrou-se maior quando são usados os marcadores escolhidos para a população CEU conforme aumenta-se o número de marcadores utilizados (Figura 2.9). O teste do percentual de variabilidade captada pelos seis SNPs utilizados no estudo em relação a todos os marcadores disponíveis na fase II do banco de dados do HapMap foi realizado para verificar o potencial do conjunto de SNPs selecionados (Tabela 2.15).

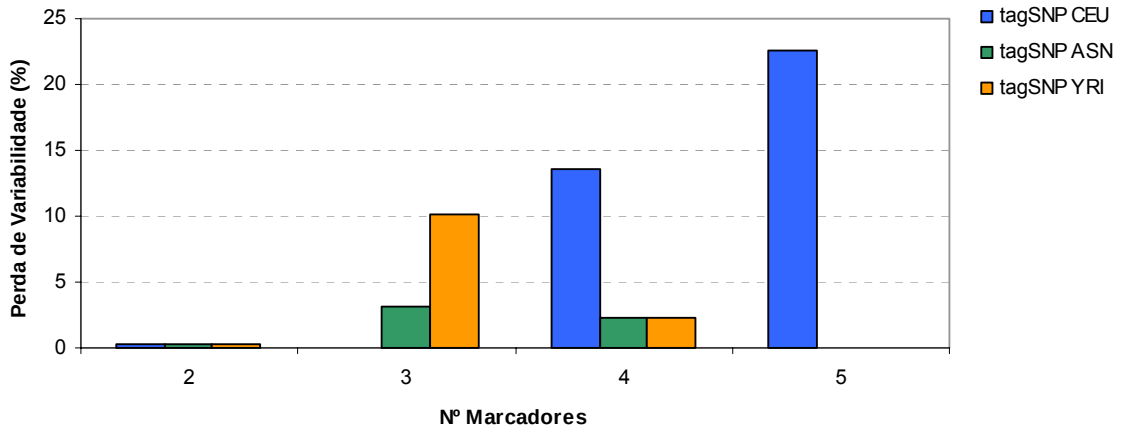


Figura 2.9 : Percentual de perda relativa de variabilidade na população brasileira captada por cada conjunto de tagSNPs definido em cada população do HapMap.
Fonte: O Autor

Tabela 2.16 : Percentual da variabilidade nas populações do HapMap captada pelos conjunto de tagSNPs em relação a todos os SNPs disponíveis no banco.

Populações e número total de marcadores	Número de marcadores				
	2	3	4	5	6
	Percentual de variabilidade				
CEU (32)	79,78	79,08	85,77	82,47	76,54
rs1217418	V	V	V	V	V
rs2476602		V	V	V	V
rs2476601				V	V
rs1217395			V	V	V
rs2476600		V	V	V	V
rs3789607	V				V
YRI (30)	94,17	93,89	93,97	92,53	89,93
rs1217418		V	V	V	V
rs2476602	V	V	V	V	V
rs2476601					V
rs1217395	V	V	V	V	V
rs2476600			V	V	V
rs3789607				V	V
ASN (32)	94,46	87,76	90,77	83,44	75,89
rs1217418			V	V	V
rs2476602			V	V	V
rs2476601					V
rs1217395	V	V	V	V	V
rs2476600	V	V	V	V	V
rs3789607		V		V	V

Fonte: O Autor

2.3 DISCUSSÃO

O gene *PTPN22* está na lista dos diversos genes que possuem SNPs associados a doenças auto imunes (GREGERSEN et al., 2006; IKARI et al., 2006; KAWASAKI et al., 2006; MORI et al., 2005). Neste caso, o polimorfismo freqüentemente usado em estudos de associação (rs2476601) possui freqüência alélica diferenciada em populações de diferentes etnias (MORI et al., 2005), e a associação alélica encontrada em uma população não é encontrada em outra, ou é revelada pela análise haplotípica (KAWASAKI et al., 2006; ONENGUT-GUMUSCU; BUCKNER; CONCANNON, 2006). Contudo, poucos trabalhos foram realizados com esse gene em populações miscigenadas em estudos de associação (BACA et al., 2006; GOMEZ et al., 2005; KAUFMAN et al., 2006), em que o efeito de contribuição de ancestralidade pode afetar nos resultados de associação.

No presente estudo foi abordada a diferença entre freqüências alélicas nas populações do HapMap para verificar o padrão de desequilíbrio de ligação na população brasileira de acordo com as proporções de miscigenação. A análise de freqüências alélicas revelou que a população brasileira possui padrões de distribuição similares aos da população CEU. Esta tendência pode ser comprovada pelas análises de variância entre as populações, na qual a população brasileira teve menor diferenciação genética com a população CEU do que com qualquer outra. Ainda, a diferença entre a população YRI foi menor do que YRI em relação à CEU, e houve maior diferenciação entre a população ASN com a brasileira, do que ASN em relação à CEU (Tabela 2.9). Isso mostra que, assim como na análise por marcadores autossômicos independentes, a amostra da população brasileira se assemelha à amostra da população de origem Européia. No entanto, para o gene *PTPN22*, a menor diferença genética observada entre a população de origem Africana e a Brasileira, em relação aos marcadores de ancestralidade, indica que a contribuição de ancestralidade Africana na miscigenação pode ter uma maior influência nos marcadores analisados nesse gene, o que pode criar desvios em análises de associação genética direta ou indireta (CHOUDHRY et al., 2006).

No caso do SNP rs2476601, estudos de associação genética já foram conduzidos e publicados em populações miscigenadas, porém, sem o controle de estratificação populacional (BACA et al., 2006; GOMEZ et al., 2005). No estudo realizado por Baca e colaboradores, pessoas afetadas com lúpus eritematoso sistêmico (LES), uma doença de caráter auto-imune, foram investigadas para o polimorfismo do rs2476601 em um estudo de caso-controle utilizando a população Mexicana. O resultado deste estudo mostrou associação com LES devido a maior freqüência do alelo T e do genótipo C/T no grupo de casos (BACA et al., 2006). Embora o estudo tenha descrito o grupo controle como “etnicamente correspondentes”, e como o termo “Hispânico” é por vezes generalizado nos países da América do Norte e Europa, é possível portanto, que os autores estivessem se referindo simplesmente à nacionalidade dos indivíduos, e não aos percentuais de ancestralidade genômica, uma vez que não foi realizado controle de estratificação genética com marcadores informativos de ancestralidade, ou qualquer outro marcador que identificasse estruturação. Da mesma forma, um estudo conduzido com a população Colombiana reportou associação do alelo T com síndrome primária de Sjogren, diabetes tipo 1 e LES, utilizando somente um grupo controle para as três doenças, sem um estudo prévio de estratificação genética entre os grupos de casos

A freqüência do alelo T do rs2476601 encontrada na população brasileira foi menor que na população CEU. Além disso, quando as regiões geográficas brasileiras são separadas, não é detectada diferença significativa entre a sua freqüência, ou seja, o alelo T está distribuído igualmente nas amostras regionais do Brasil. Por outro lado, a freqüência deste alelo já foi descrita como diferenciada entre populações européias em um gradiente crescente no sentido sul-norte, e sua freqüência na população brasileira é comparável à de outras populações de origem européia que deram origem à população brasileira, como a Espanhola (GREGERSEN et al., 2006; OROZCO et al., 2005). Além de estar distribuído igualmente, a ancestralidade européia não foi correlacionada com o alelo T, o que ratifica o fato que este alelo foi encontrado em pessoas com percentuais maiores de ancestralidade africana.

Um exemplo no qual este tipo de caso pode ocorrer é o estudo realizado por Suarez-Kurtz e colaboradores na população brasileira com o gene *CYP2C9*. Esse gene, muito importante em estudos de farmacogenômica, codifica a proteína CYP2C9, responsável pela hidrólise de uma vasta amplitude de medicamentos clínicos, e um de seus polimorfismos, o *CYP2C9*5*, é relatado somente em populações de origem Africana (KIRCHHEINER; BROCKMOLLER, 2005; SUAREZ-KURTZ et al., 2005). Apesar disso, esse polimorfismo foi observado em um indivíduo Brasileiro que se auto identificou como branco. A análise de ancestralidade genômica revelou contribuição de 92,0% de genoma Europeu, 7,5% Africano, e 0,5% Indígena. Por outro lado, a análise familiar (pais e irmãos) revelou que a contribuição africana neste indivíduo tem origem materna e, conseqüentemente, o alelo *CYP2C9*5* também (SUAREZ-KURTZ et al., 2005).

O estudo de Suarez-Kurtz e colaboradores (2005) é um forte indicativo de que alelos exclusivos de algumas populações em genes importantes para estudos de associação genética, como no caso do alelo T do rs2476601 no gene *PTPN22*, exclusivo de Europeus, podem ser encontrados em populações miscigenadas com predominância de outras populações. O presente estudo verificou esse fato com o achado que o alelo T do rs2476601 pode estar presente em indivíduos que possuem de 3 até 50% de contribuição Africana. Apesar de não ser predominante, a contribuição Africana neste caso é alta o suficiente para ser um

fator de conflito em estudos de associação que utilizam grupos separados por outras variáveis, como auto denominação de cor de pele (ABREU, 2007; PARRA, F. C. et al., 2003; VIEIRA et al., 2006) ou auto declaração de grupo étnico (BARNHOLTZ-SLOAN et al., 2005). Assim, o controle de estratificação populacional utilizando marcadores autossômicos independentes e não ligados ao gene em questão é uma estratégia prática que deve sempre ser utilizada quando forem conduzidos estudos do tipo caso-controle e estudos transversais em populações miscigenadas a fim de se evitar associações espúrias ou falhas em identificar associação (BARNHOLTZ-SLOAN et al., 2005; CHOUDHRY et al., 2006; HOGGART et al., 2003; TSAI et al., 2006).

O padrão de desequilíbrio de ligação em populações miscigenadas deveria ser maior conforme a miscigenação entre duas ou mais populações parentais (ALTSHULER et al., 2005; CARDON; ABECASIS, 2003; GABRIEL et al., 2002). Em estudos prévios na população brasileira, foram identificados maiores DL em determinados locos em relação a outras populações (ALLEBRANDT; SOUZA; CHAUTARD-FREIRE-MAIA, 2002; MORAES et al., 2003; SHI et al., 2003). No entanto, o DL médio no gene *PTPN22* observado na população brasileira foi menor do que nas populações do HapMap, tanto para os valores de D' quanto para os de r^2. Sawyer e colaboradores descreveram que existem algumas regiões genômicas onde o padrão de DL é bastante variado em diferentes populações, e ainda, que populações pertencentes às mesmas regiões geográficas particulares não mostram consistência no padrão de DL para os locos pesquisados (SAWYER et al., 2005). Portanto, o DL resultante na população brasileira pode ser um artifício da combinação de fatores demográficos e forças evolutivas, como deriva e seleção.

Outra conseqüência observada nos resultados foi um efeito não esperado da quebra do desequilíbrio de ligação nas amostras regionais brasileiras, com possíveis pontos de recombinação. No momento do estudo *in silico* e desenho dos iniciadores, o HapMap dispunha de aproximadamente 20 marcadores ao longo do gene (HapMap - Data Rel#16c.1 phase I june 2005) que conferiam alto DL nas três populações. Com o avanço recente da proposta do consórcio (HapMap - Data Rel#21a phase II january 2007), o número de marcadores ultrapassou o dobro do que havia disponível, deixando a região com maior densidade de marcadores e, por

conseqüência, o padrão de DL, principalmente na população ASN, foi alterado substancialmente. Por outro lado, o número de amostras nas populações regionais brasileiras pode contrabalancear a informação gerada pelos locos, favorecendo a captura de informação dos SNPs de menor freqüência e assim revelar diferentes padrões de blocos haplotípicos (ZEGGINI et al., 2005).

Em relação ao SNP rs2476601, associado às doenças auto-imunes, a literatura descreve que, nas populações que possuem o alelo T, independente de quais locos são utilizados, somente um haplótipo é encontrado (CARLTON et al., 2005; ONENGUT-GUMUSCU; BUCKNER; CONCANNON, 2006). A população brasileira apresentou três haplótipos com o alelo T no rs2476601, são eles o haplótipo H07, o qual foi observado também na população CEU, e os haplótipos H03 e H16, que foram considerados raros, com freqüências menores que 0,01, e, por isso, foram descartados das análises de correlação com ancestralidade. Desses dois haplótipos raros, um deles, o H03, difere do H07 somente no polimorfismo do loco rs1217395, onde o H07 possui o alelo C e o H03 o alelo T, portanto, descarta-se a hipótese de baixa probabilidade na estimativa de fase do haplótipo. Da mesma forma, o haplótipo H16 difere do H14 somente no polimorfismo do loco rs2476601 onde o H14 possui o alelo C e o H16 o alelo T. Tendo em vista que houve quebra de DL e formação de blocos nas amostras regionais, é possível insinuar que uma possível e rara recombinação tenha criado esses haplótipos.

Apesar de manter a discussão sobre o haplótipo carregador do alelo associado à doenças auto-imunes, os resultados mostraram a influência das proporções de ancestralidade sobre a distribuição dos outros haplótipos na população brasileira. Considerando somente os haplótipos com freqüência acima de 0,01, a população brasileira possui sete haplótipos, dos quais três são comuns às populações CEU, YRI e ASN, um é exclusivo da população YRI, um exclusivo da população CEU e um comum às populações CEU e ASN. As demais populações apresentam, na ordem, cinco (CEU) e quatro haplótipos (ASN e YRI). Dos três haplótipos mais freqüentes na população brasileira, o primeiro é o mais freqüente em YRI (H11), o segundo é o mais freqüente em CEU (H14) e o terceiro é o mais freqüente em ASN (H04). Esses três haplótipos somam 73,9% da diversidade haplotípica na população brasileira. Quando comparado com a população CEU os

mesmos haplótipos representam 63,3% da diversidade. O quarto haplótipo mais freqüente (H01) é comum nas quatro populações estudadas e, somando-o aos demais, representam 89,7% da diversidade haplotípica brasileira. Dos demais haplótipos, dois são exclusivos em YRI, porém de baixa freqüência, e, por último, o haplótipo exclusivo da população CEU que contém o alelo T no rs2476601 (H07).

Esses números levam a crer que a população brasileira possui uma diversidade haplotípica diretamente influenciada pelos níveis de miscigenação, o que corrobora os achados de outros estudos realizados na população brasileira. Moraes e colaboradores (MORAES et al., 2003) descreveram no promotor da Interleucina-10 haplótipos exclusivos de populações Européias presentes em Afro-Brasileiros e, ainda, um padrão de distribuição dos haplótipos similares entre Afro-Brasileiros e Euro-Brasileiros. Em outro estudo, Boldt e colaboradores (BOLDT et al., 2006) estimaram haplótipos do gene *MBL2* para uma diversidade de grupos da população brasileira composta por Afro-Brasileiros, Euro-Brasileiros, Brasileiros com origem asiática, populações indígenas e indivíduos miscigenados, e foi observado a semelhança no padrão de haplótipos entre estes grupos brasileiros com os respectivos grupos de origem, enquanto no grupo de indivíduos miscigenados a distribuição dos haplótipos se mostrou mais diversificada, provavelmente influenciada pelos níveis de miscigenação.

Para avaliar os efeitos da miscigenação no gene *PTPN22*, o teste de correlação da ancestralidade genômica condicionado aos haplótipos mostrou efeito significativo em dois haplótipos, segundo o teste de correlação haplótipo-específica. Os haplótipos H14 e H04 obtiveram coeficientes de correlação negativo e positivo, respectivamente, para o componente de ancestralidade africana. A análise de permutação local revelou que a ancestralidade africana está negativamente correlacionada com o alelo G do rs3789607. Isto fica evidente quando avaliada as freqüências alélicas deste loco, onde o alelo G é quase exclusivo da população CEU (f = 0,283), ocorrendo a baixas freqüências na população ASN (f = 0,062), enquanto o alelo A está fixado na população YRI. Entre os haplótipos H14 e o H04, somente o H14 possui este alelo G. Como mencionado anteriormente, o H14 é também o mais comum na população CEU, assim, ele ocorre mais freqüentemente em indivíduos

com menor contribuição de ancestralidade africana comparado aos outros haplótipos.

Por outro lado, a ancestralidade africana também se mostrou correlacionada positivamente com o alelo C do rs1217395. Neste caso, esta correlação não é tão evidente como a comentada anteriormente. Este alelo é de baixa freqüência na população YRI (f = 0,058), portanto, seria esperado o efeito contrário ao que foi observado. No entanto, sua freqüência é relativamente alta na população ASN (f = 0,584) e moderada na população CEU (f = 0,275). Ainda, esse alelo está presente no H04 (também correlacionado positivamente com ancestralidade africana), cuja freqüência é moderadamente elevada em ASN. Uma explicação plausível seria o fato da população brasileira com maior ancestralidade africana ter herdado esse alelo de duas maneiras: (1) da contribuição Européia, ou (2) da contribuição da população indígena pelo histórico de miscigenação atuando juntamente com forças evolutivas, como deriva genética e seleção natural, caso esse alelo também fosse muito freqüente nos Índios nativos do Brasil. Mesmo sem dados em outras populações ameríndias ou indígenas, é provável que essas duas hipóteses possam ter atuado conjuntamente, principalmente pelos dados observados de forte seleção natural na população asiática para este loco.

O conceito de mosaico de haplótipos foi introduzido primeiramente por Erhart e colaboradores para designar pares de haplótipos do complexo *t* de *Mus musculus* em que fossem observados tanto o haplótipo *t* quanto o seu homólogo selvagem em uma mesma amostra (ERHART et al., 1989). Neste trabalho esse conceito foi estendido no sentido em que mosaico de haplótipos fosse compreendido por um indivíduo miscigenado, que possuísse um par de haplótipos determinado por cada haplótipo exclusivo de duas populações distintas. Na população brasileira analisada, apenas um indivíduo pertencente à região Norte foi encontrado com essa característica, cujo par de haplótipos é formado pelo H05, um haplótipo raro exclusivo da população ASN, e o H07, o haplótipo portador do alelo T do rs2476601, exclusivo da população CEU. Esse fato, apesar de único, é mais um indicativo da miscigenação como forte evento demográfico atuando na distribuição dos blocos de haplótipos na população brasileira.

A seleção natural é um dos conceitos centrais na teoria evolutiva e está diretamente ligada aos fenótipos e ao processo de adaptabilidade. Se as diferenças em genótipos individuais afetarem a adaptabilidade, então as freqüências dos genótipos mudarão nas gerações seguintes, ou seja, os genótipos com adaptabilidade mais elevada tornar-se-ão mais comuns (HARTL; CLARK, 1997). A varredura seletiva (*selective sweep*), na qual uma mutação benéfica é levada para fixação em uma população por seleção natural também pode resultar na seleção de alelos adjacentes pelo efeito chamado de carona genética. Uma forte varredura seletiva tem grande impacto nos padrões de DL, pois resulta numa região genômica em que existe um haplótipo positivamente selecionado que é essencialmente o mais freqüente e de grande extensão em uma população (ALTSHULER et al., 2005; KIM; NIELSEN, 2004; MCVEAN, 2006; VOIGHT et al., 2006). Esse fato ocorreu na população ASN, com freqüência haplotípica de 0,58 para o H04, e na população YRI, com freqüência haplotípica de 0,72 para o H11. Se ponderado pelo número de marcadores utilizados na análise, os resultados do percentual de variabilidade genética na população YRI cai para aproximadamente 94% utilizando somente os 4 marcadores polimórficos num total de 30 marcadores da fase II do HapMap. Para a população ASN, o percentual de variabilidade haplotípica cai para aproximadamente 83% utilizando os 5 marcadores polimórficos em um total de 32 marcadores da fase II, no entanto, essa perda é menor quando somente 4 marcadores são utilizados, aproximadamente 90% de variabilidade é capturada. Portanto, tais resultados demonstram que parece haver uma seleção positiva na região do gene *PTPN22*.

A ferramenta eletrônica Haplotter deu indicativos de que a seleção positiva não ocorre no *PTPN22*, mas sim em genes vizinhos a ele com valores significativos para a população YRI nos genes *MAGI3*, *PHTF1* e *RSBN1*. A seleção natural atua sobre características fenotípicas, mas acaba por alterar as freqüências genotípicas para fixação de alelos importantes na adaptabilidade e sobrevivência de uma população (VOIGHT et al., 2006). Dos três genes, o *PHTF1* é um suposto gene que codifica para um fator de transcrição de homeodomínio. O gene *MAGI3* é uma guanilato quinase associada à membrana (do inglês, *membrane associated guanylate kinase, WW and PDZ domain containing 3*) com função de molécula *scaffolding* que liga a fosfatase receptora de tirosina com seus substratos na membrana plasmática (ADAMSKY et al., 2003). O gene *RSBN1* (*round spermatid*

basic protein 1) codifica para uma proteína com importante papel na regulação de transcrição em células haplóides espermáticas (TAKAHASHI et al., 2004). É possível que, algumas gerações atrás, um desses três genes tenha passado por forte seleção positiva, e, por possuir extenso desequilíbrio de ligação (Figura 2.9), tenha favorecido a alta freqüência haplotípica no gene *PTPN22* pelo efeito de carona genética na população YRI (KIM; NIELSEN, 2004; MCVEAN, 2006; VOIGHT et al., 2006).

No modelo proposto por Voight e colaboradores (2006), são esperados que alelos derivados da seleção surjam com excessivo DL em relação à sua origem. No entanto, um iHS positivo também é considerado candidato por que o alelo ancestral pode ter sofrido efeito carona ou ele mesmo pode ter sido alvo de seleção. Um iHS extremamente positivo (iHS > 2) significa que os haplótipos com origem no alelo ancestral são mais longos que os com origem no alelo derivado, e o contrário ocorre quando um iHS é extremamente negativo (iHS < -2), ou seja, os haplótipos com origem no alelo derivado são mais longos que os com origem no alelo ancestral.

Dos seis SNPs investigados neste estudo, quatro foram identificados para efeito de seleção com iHS significativo, sendo dois na população CEU e um em cada uma das outras três populações do HapMap. Destes quatro SNPs, apenas um deles, o rs2476600 na população YRI, apresentou o iHS positivo indicando que os haplótipos com origem no alelo ancestral são mais longos que os com origem no alelo derivado. Isso corrobora a idéia do alelo ancestral, e os haplótipos originados a partir dele, terem sofrido efeito carona na seleção positiva de um dos genes ao redor. Por outro lado, os demais locos identificados possuem iHS negativo e, portanto, os alelos derivados foram sujeitos a seleção positiva recente.

Apesar de o HapMap ser um projeto grande o suficiente para gerar informações de caráter genético em diferentes populações, o próprio consórcio questiona a maneira em como painéis de análise podem ser transferidos destas e/ou comparados com outras populações (ALTSHULER et al., 2005). Uma vez que a população brasileira estudada possui níveis de ancestralidade predominantemente europeus, seriam esperados padrões de desequilíbrio de ligação e de distribuição de freqüências haplotípicas semelhante aos da população CEU. No entanto foi

observada nas freqüências haplotípicas uma forte influência de ancestralidade, principalmente na contribuição africana, no que diz respeito às correlações aqui observadas. A utilização direta de tagSNPs das populações do HapMap se mostrou eficaz, mas não tão eficiente, em avaliar a variabilidade haplotípica e os padrões de DL na população brasileira. Pela maior similaridade com a população de origem européia, era esperado que tagSNPs referentes a essa população gerassem maior variabilidade, no entanto, foi observado o contrário, indicando que as demais populações podem ter grande influência nos padrões do DL da população brasileira.

A perda relativa de variabilidade na transferibilidade de tagSNP entre as populações do HapMap e a população Brasileira apóia o fato de que, apesar da população brasileira ter maior proximidade genética com a população CEU, a variabilidade haplotípica está condicionada aos tagSNPs das outras populações demonstrando a influência do efeito de miscigenação e de seleção na diversidade haplotípica do gene *PTPN22* na população brasileira.

No entanto, foi observada baixa perda de variabilidade genética nas populações do HapMap quando utilizados os marcadores selecionados contra todos disponíveis. Isto indica que em genes com forte DL entre todos os marcadores e com grandes extensões de haplótipos, as informações de DL puderam ser captadas mesmo por um número pequeno de marcadores em um número relativamente pequeno de amostras. É evidente, no entanto, que o padrão de blocos haplotípicos pode mudar substancialmente com a ampliação da densidade de marcadores, aumentando a extensão dos blocos ou revelando pontos de recombinação (DALY et al., 2001; GABRIEL et al., 2002; PHILLIPS et al., 2003; WANG, N. et al., 2002; ZEGGINI et al., 2005; ZHANG et al., 2002).

Dados sobre a distribuição das freqüências alélicas e genotípicas, estimativas de haplótipos e suas correlações com níveis de ancestralidade são extremamente importantes para estudos de associação genética em populações miscigenadas. A população brasileira está entre as populações com maior heterogeneidade genética do mundo, e, portanto estudos dessa natureza devem ser conduzidos sobre controle de estratificação. Antes de tudo, o estudo de genética de populações é fundamental para avaliar como essa heterogeneidade é modulada de acordo com as forças evolutivas. Neste capítulo foi mostrado que o gene *PTPN22*

possui uma grande diversidade genética na população brasileira, e que, o mesmo está intrinsecamente ligado aos níveis de miscigenação. Para estudos de associação que envolvam a população brasileira e o loco rs2476601 devem ser tomados cuidados com estratificação genética, a fim de se evitar possíveis falhas de associação espúria. Ainda, um campo inexplorado foi aberto para investigação do desequilíbrio de ligação em genes que sofreram recente seleção positiva em populações parentais e como se comporta esse gene na população brasileira de acordo com forças evolutivas.

3 Padrões de Desequilíbrio de Ligação no Gene do Receptor de Vitamina D (*VDR*) na População Brasileira

3.1 MATERIAIS E MÉTODOS

3.1.1 Amostra Populacional

A amostragem populacional utilizada no presente estudo foi constituída de 200 indivíduos brasileiros não relacionados como descrito no Capítulo 1, assim como o método de extração do DNA (seções 1.1.1 e 1.1.2 respectivamente). Foram utilizadas as amostras do HapMap como referência de pesquisa e comparação dos dados genotípicos do gene *VDR*.

3.1.2 Seleção de SNPs e Desenho dos Iniciadores

A seleção dos SNPs foi feita mediante consulta ao banco de dados disponível no *International HapMap Consortium* para o gene *VDR* e foram escolhidos SNPs que pudessem conferir estrutura de blocos haplotípicos nas populações do HapMap. Os critérios de seleção para os locos foram, preferencialmente, SNPs espaçados a uma média de 5 kb entre eles e com freqüência alélica mínima de 5%, empregando uma consulta ao banco de dados Data Rel#16c.1 phaseI june05 do HapMap e também ao dbSNP no *build* 124. Foi avaliado o padrão de desequilíbrio de ligação nessas populações como critério de seleção dos marcadores.

Além disso, foram utilizados conjuntamente na análise cinco marcadores adicionais (BsmI, TaqI, ApaI, FokI e Cdx-2) amplamente utilizados em estudos de associação, principalmente com fenótipos relacionados a densidade mineral óssea (ARAI et al., 2001; FARACO et al., 1989; GROSS et al., 1996; MORRISON et al., 1992). Sistemas para genotipagem em multiplex utilizando

extensão de base única e SNaPshot™ Multiplex System (Applied Biosystems, EUA) para esses cinco polimorfismos foram previamente desenvolvidos e utilizados em outros estudos no laboratório de Ciências Genômicas da Universidade Católica de Brasília (GENTIL, 2006; LIMA, 2006; ver também APÊNDICE A). Os critérios de desenho dos iniciadores e as ferramentas utilizadas foram os mesmos descritos no Capítulo 2, seção 2.1.2.

3.1.3 PCR e Genotipagem

A genotipagem dos marcadores selecionados para o gene *VDR* foi realizada em três conjuntos de multiplex com sete locos cada, a partir dos métodos descritos no capítulo 2, seção 2.1.3.

3.1.4 Análise Estatística

3.1.4.1 Estatística descritiva e Análise de variância molecular

As estimativas de freqüências alélicas dos locos SNPs foram tomadas a partir dos dados genotípicos e foram usadas para calcular o equilíbrio esperado segundo Hardy-Weinberg e as análises de variância molecular baseadas nas estatísticas F de Wright utilizando respectivamente os programas GenAlEx (PEAKALL; SMOUSE, 2006) e GDA - Genetic Data Analysis (LEWIS; ZAYKIN, 2001).

3.1.4.2 Análise de DL e de Haplótipos

As análises do desequilíbrio de ligação entre os marcadores utilizados nas populações do HapMap e na população brasileira foram feitas de acordo com os parâmetros estatísticos D' e r^2 e os blocos de haplótipos foram definidos por Coluna sólida de DL no programa Haploview versão 3.32 (BARRETT et al., 2005). Neste caso foi considerada a orientação 5'-3' do gene, e não do DNA. O programa Haploview possui uma interface gráfica para visualização do desequilíbrio de ligação baseado nos valores de D', r^2 e intervalo de confiança de D'. O programa Haploview foi utilizado para estimar os haplótipos das populações do HapMap.

Os haplótipos na população brasileira foram inferidos por estatística Bayesiana utilizando o programa Phase (STEPHENS, M.; DONNELLY, 2003). Primeiramente os dados foram utilizados para estimar as taxas de recombinação e estimar pontos quentes de recombinação no gene (CRAWFORD; BHANGALE et al., 2004). Em seguida, os blocos de haplótipos foram separados e tiveram suas freqüências estimadas.

As estimativas de ancestralidade genômica das amostras geradas na análise 4, descritas no capítulo 1 (Tabela 1.9), foram utilizadas como fator de correlação nas análises genéticas de freqüências alélicas, genotípicas e haplotípicas. O software Whap versão 2.9 (PURCELL; DALY; SHAM, 2007) foi utilizado para estimativas de ancestralidade em testes de regressão com permutação condicional.

3.1.4.3 Transferibilidade de tagSNPs

O estudo de transferibilidade de tagSNPs foi realizado com a predição de cobertura da variabilidade usando o algoritmo Stampa (HALPERIN; KIMMEL; SHAMIR, 2005) do programa Gevalt (DAVIDOVICH; KIMMEL; SHAMIR, 2007),

baseados no código aberto do Haploview. Para isso, o algoritmo foi aplicado em cada população para os SNPs selecionados nesse estudo, e em seguida, o cálculo da perda relativa de variabilidade foi feito na população brasileira pela diferença entre a variabilidade captada utilizando-se os tagSNPs selecionados para cada uma das populações do HapMap, relativo à variabilidade captada pelos tagSNPs da população brasileira. Utilizando os dados da fase II do HapMap, foi possível verificar qual o percentual de variabilidade captado pelos marcadores selecionados em cada uma das três populações. No caso do gene VDR nem todos os marcadores possuíam genótipos para todas as populações, portanto, estas análises foram feitas par a par de acordo com os marcadores presentes nas populações do HapMap.

3.2 RESULTADOS

3.2.1 Seleção de SNPs e Desenho dos Iniciadores

Os critérios de seleção para os SNPs que flanqueiam os SNPs disponíveis e validados no dbSNP foram, preferencialmente, SNPs espaçados em uma média de 5 kb e com freqüência alélica mínima de 5% empregando uma consulta ao banco de dados Data Rel#16c.1 phase I june 05 do HapMap e também ao dbSNP no build 124. Para o gene *VDR* foram escolhidos 18 marcadores, no entanto, em uma análise posterior verificou-se que em um deles existia no mesmo SNP dois números de referência nos bancos de dados. A consulta nos bancos de dados revelou que o SNP FokI possui dois números de referência, o rs10735810 empregado pelo dbSNP, e desenhado iniciadores para este estudo, e o rs2228570, empregado pelo HapMap, e com iniciadores previamente desenhados e utilizados em outros estudos (GENTIL, 2006; LIMA, 2006). Em outro, o SNP rs1544410 é referente ao BsmI, e, portanto outro SNP foi repetido na escolha e desenho dos iniciadores. Assim, dos cinco SNPs descritos anteriormente foram utilizados somente três, o ApaI, TaqI e Cdx-2. Ao todo foram utilizados 21 SNPs (Tabela 3.1) distribuídos ao longo do gene (Figura 3.1) para que fossem desenhados iniciadores (Tabela 3.2) e desenvolvidos sistemas multiplex baseados em extensão de base única.

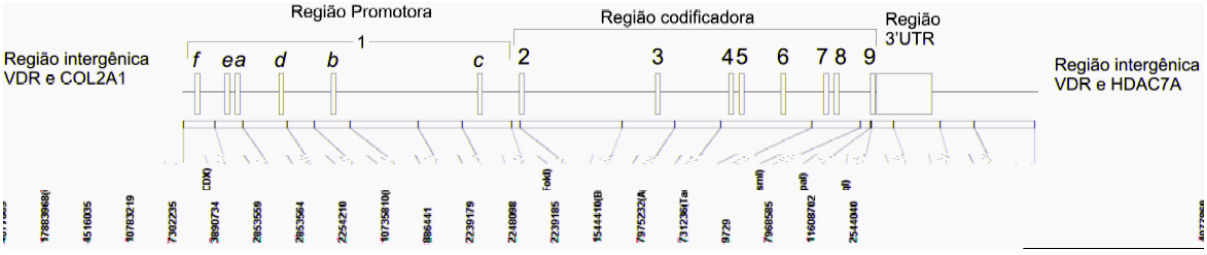


Figura 3.1 : Distribuição esquemática dos polimorfismos ao longo do gene *VDR*
Fonte: O Autor

Tabela 3.1 : Características dos marcadores do gene *VDR* com relação ao cromossomo e contig.

Loco	Heterozigose média	Acesso do Contig	Versão do Contig	Posição no Contig	Posição no Cromossomo	Build Original	Alelos *
rs2544040	0,18	NT_029419.10	Hs12_29578_35	10366251	46509213	100	G/A
rs11608702	0,26	NT_029419.10	Hs12_29578_35	10372073	46515035	121	A/T
rs7968585	0,26	NT_029419.10	Hs12_29578_35	10375398	46518360	121	C/T
rs9729	0,05	NT_029419.10	Hs12_29578_35	10379928	46522890	125	A/C
rs731236 (TaqI)	0,18	NT_029419.10	Hs12_29578_35	10382062	46525024	123	C/T
rs7975232 (ApaI)	0,03	NT_029419.10	Hs12_29578_35	10382142	46525104	121	C/A
rs1544410 (BsmI)	0,19	NT_029419.10	Hs12_29578_35	10383140	46526102	121	G/A
rs2239185	0,08	NT_029419.10	Hs12_29578_35	10387864	46530826	123	C/T
rs2248098	0,06	NT_029419.10	Hs12_29578_35	10396661	46539623	123	C/T
rs2239179	0,18	NT_029419.10	Hs12_29578_35	10401071	46544033	121	G/A
rs886441	0,24	NT_029419.10	Hs12_29578_35	10406269	46549231	121	C/T
rs10735810 (FokI)	0,17	NT_029419.10	Hs12_29578_35	10416200	46559162	121	G/A
rs2254210	0,16	NT_029419.10	Hs12_29578_35	10417019	46559981	121	C/T
rs2853564	0,18	NT_029419.10	Hs12_29578_35	10421792	46564754	121	C/T
rs2853559	0,27	NT_029419.10	Hs12_29578_35	10426110	46569072	120	C/T
rs3890734	0,27	NT_029419.10	Hs12_29578_35	10432660	46575622	123	G/A
rs7302235	0,27	NT_029419.10	Hs12_29578_35	10436143	46579105	123	C/T
rs10783219	0,26	NT_029419.10	Hs12_29578_35	10438793	46581755	123	A/T
rs4516035	0,27	NT_029419.10	Hs12_29578_35	10443131	46586093	123	C/T
rs17883968 (Cdx-2)	0,02	NT_029419.10	Hs12_29578_35	10445850	46588812	124	G/A
rs4077869	0,27	NT_029419.10	Hs12_29578_35	10448949	46591911	123	C/T

* Alelos referentes à orientação positiva (5'-3') do DNA
Fonte: O Autor

3.2.2 Recuperação de Genótipos do HapMap

A priori, foram utilizados dados da fase I do HapMap para recuperar os genótipos, mas, no entanto, com o lançamento da fase II (Rel 21a, Jan 2006) foi possível recuperar os genótipos das três amostras populacionais (CEU, YRI e ASN) para quase todos os locos (rs4077869 e rs2239185 ausentes em CEU; rs7302235 ausente em ASN). O DL nas populações do HapMap foi avaliado pelo software haploview (Figura o d7 Tt ferença nos padrões entre .32s.

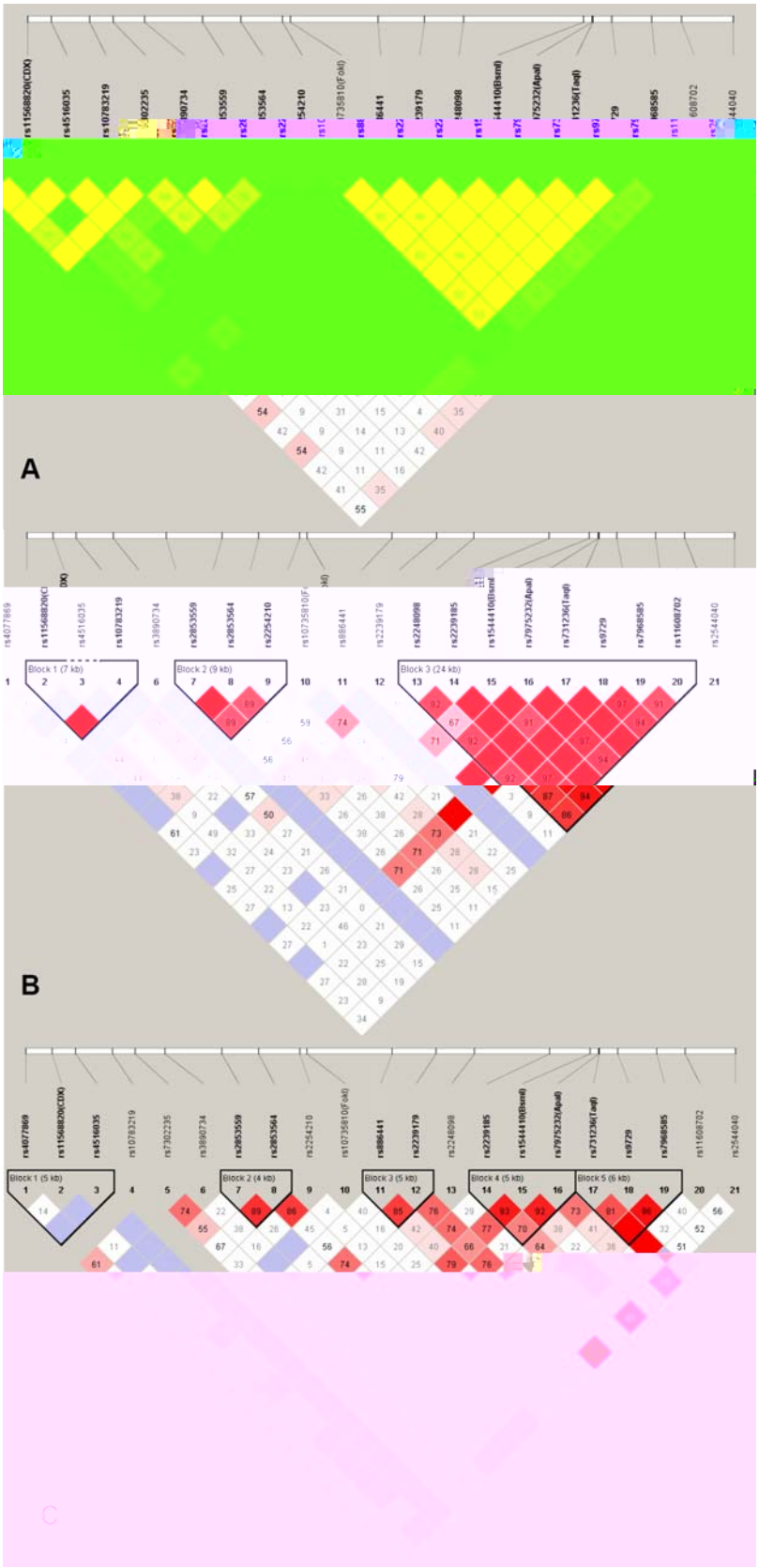


Figura 3.2 : Desequilíbrio de ligação entre os marcadores selecionados nas populações do HapMap: (A) CEU, (B) ASN e (C) YRI. O número nos quadrados indica o valor de D'. Quadrados entre rosa e vermelho indicam D' < 1 e LOD ≥ 2 e quadrados brancos D' < 1 e LOD < 2. Quadrados vermelhos sem número indicam D' = 1 e LOD ≥ 2; e os azuis D' = 1 e LOD < 2.
Fonte: O Autor

3.2.3 Genotipagem e Condições de PCR

O desenho dos iniciadores (Tabela 3.2) permitiu a montagem de três sistemas multiplex utilizados para uma tipagem rápida e eficiente dos 21 locos em reações de PCR seguidas de tratamento enzimático e reação de extensão de base única em multiplex. A concentração dos iniciadores (Tabela 3.3) utilizados na reação SNaPshot™ teve que ser diferente entre os locos a fim de manter equivalente o balanço de intensidade entre os picos em unidade relativa de fluorescência (Figuras 3.3 a 3.5). A utilização de termociclagem *touchdown* tanto na PCR quanto na reação de SNaPshot™ foi necessária para evitar o comprometimento da qualidade da amplificação com o aparecimento de produtos inespecíficos que dificultariam a análise.

Tabela 3.2 : Seqüência e detalhes dos iniciadores para o gene *VDR*. Iniciadores indicados são F – Direto e R – reverso para reações de PCR e S – primer de extensão de base única.

SNP	Primer	Direção	Seqüência dos oligos	Tamanho (pb)	TM (ºC)	Produto (pb) ou Alelo amplificado
rs2544040	F	direto	CTGGGAGAGGTGGAGTCATC	20	60	208
	R	reverso	GGCACATTTTGAGCTCCATT	20	60	
	S	direto	$T_{(4)}$GGGAGGGGCAGACCCACGCA	24	60	G/A
rs4077869	F	direto	CAGGGGCCATGAGAGTTTTA	20	60	284
	R	reverso	TGGTAAAGAGCTTGGGCTTG	20	60	
	S	reverso	$T_{(5)}$CAGTTTTGCCATCAGCAAA	24	51	G/A
rs11608702	F	direto	CCTGGGTAGGAGAGGGAAAG	20	60	233
	R	reverso	GTGCCCCAACTTTGTCAACT	20	60	
	S	direto	$T_{(8)}$ACATGTAAATCAGTGGGGCT	28	50	A/T
rs10783219	F	direto	TAGTGTGGTCCCCAGAGGTC	20	60	229
	R	reverso	TTCCCTTCACCCTTGACTTG	20	60	
	S	direto	$T_{(12)}$GAAGCTGGATGAGCAAAATG	32	51	A/T
rs2239179	F	direto	CCTAGCTGTGGGTCTGAGGA	20	60	234
	R	reverso	CCTCAGCTCTGCCTCTGTGT	20	61	
	S	reverso	$T_{(19)}$GTTACCTGACCTCTCCCCA	38	52	C/T
rs2248098	F	direto	GAATTGAAAACTGTGCTAAC	20	50	180
	R	reverso	CCAAACCTGTATGAGTTCTA	20	50	
	S	direto	$T_{(19)}$GGCAATCAAGAAATGCTTCA	38	50	C/T
rs10735810 (FokI)	F	direto	TCAAAGTCTCCAGGGTCAGG	20	60	249
	R	reverso	AGGGCGAATCATGTATGAGG	20	60	
	S	reverso	$T_{(24)}$GCTTGCTGTTCTTACAGGGA	44	52	C/T

Tabela 3.2 Continuação

SNP	Primer	Direção	Seqüência dos oligos	Tamanho (pb)	TM (ºC)	Produto (pb) ou Alelo amplificado
rs7302235	F	direto	TGTATGTGATATATCGCGGTTTT	23	58	186
	R	reverso	CTTGTGAGAGGCCCGTTTAG	20	60	
	S	reverso	$T_{(24)}$ATTCTTTTTTGTTCATTTTT	44	52	G/A
rs3890734	F	direto	CATTCTGGGAGGGGTACCTT	20	60	158
	R	reverso	GGCAAACAGCAAACATCACA	20	61	
	S	direto	$T_{(30)}$AAGGAATTCTACAGGAACAG	50	47	G/A
rs1544410 (BsmI)	F	direto	CCTCACTGCCCTTAGCTCTG	20	60	209
	R	reverso	CCATCTCTCAGGCTCCAAAG	20	60	
	S	reverso	$T_{(30)}$GAGCCTGAGTATTGGGAATG	50	50	C/T
rs7968585	F	direto	CCTGGGAGACAGCAGTTTACTT	22	60	275
	R	reverso	GAGCATAGTCCGTGCTGACA	20	60	
	S	direto	$T_{(9)}$CTTCTCCCACTGCACAGTG	28	54	C/T
rs2853559	F	direto	TCTCCTCTCTTGTTTTTATTGACTT	25	57	451
	R	reverso	ATAGGGGACTGGCCCATTTA	20	61	
	S	direto	$T_{(12)}$ATAATAATTGTACATA	28	42	C/T
rs2254210	F	direto	TGAATGTCCTTCTGAGTCATGT	22	60	201
	R	reverso	GCCCAGTTTTGGCTAATGAA	20	60	
	S	direto	$T_{(12)}$GGAGAAGAGACAGACGGAAC	32	52	C/T
rs886441	F	direto	CATTCCTGTTGTCAGGCAAA	20	60	188
	R	reverso	CTGCCTTTGACCCTCACTTG	20	61	
	S	direto	$T_{(18)}$GGCAATCTCCAACCCTCCTG	38	54	C/T
rs2853564	F	direto	ATCTATTGAATGGGATTTTC	20	51	234
	R	reverso	AATCTAGGTAGCTTAGCTCTG	21	50	
	S	direto	$T_{(24)}$GTGGAAGTGAAAGGTGTCCA	44	50	C/T
rs9729	F	direto	AGGGAGAGACCCTTGTTTGA	20	59	173
	R	reverso	GTCCTGGCCCACTTCTAGC	19	60	
	S	direto	$T_{(4)}$AACGAGTCAATCCCCTCATT	24	50	A/C
rs4516035	F	direto	GAATAGAAATGCTCACAAAA	20	51	268
	R	reverso	CTGGTTATGATGTAGTCCAG	20	50	
	S	direto	$T_{(30)}$CCTCCTTTAGCCAGGGAAGA	50	53	C/T
rs2239185	F	direto	GGGAAAGACGAAACAGCAAC	20	60	190
	R	reverso	GTGGTGGGAGTAGAGGTGGA	20	60	
	S	direto	$T_{(5)}$GCCCAGAAGCGGCTGCAGG	24	61	C/T
rs11568820 (Cdx-2)	F	direto	CATTGTAGAACATCTTTTGTATCAGGA	27	60	224
	R	reverso	GACAAAAAGGATCAGGGATGA	21	60	
	S	direto	CCTGAGTAAACTAGGTCACA	20	50	G/A
rs731236 (TaqI)	F	direto	CTGCCGTTGAGTGTCTGTGT	20	60	242
	R	reverso	TCGGCTAGCTTCTGGATCAT	20	60	
	S	reverso	$T_{(9)}$GCGGTCCTGGATGGCCTC	27	50	G/A
rs7975232 (ApaI)	F	direto	*			
	R	reverso	*			
	S	reverso	$T_{(12)}$GTGGTGGGATTGAGCAGTGAGG	34	50	G/T

* Iniciadores F e R para ApaI e TaqI são os mesmos pelo fato de serem separados em apenas 80 pb.
Fonte: O Autor

Tabela 3.3 : Sistemas multiplex de PCR e concentração dos iniciadores na reação de extensão de base única para os polimorfismos do gene *VDR*.

Multiplex VDR1		Multiplex VDR2		Multiplex VDR3	
Loco	Concentração (μM)	Loco	Concentração (μM)	Loco	Concentração (μM)
rs17883968 (CDX)	0,10	rs7975232 (ApaI)	0,08	rs10735810 (FokI)	0,60
rs2544040	0,04	rs731236 (TaqI)	0,02	rs1544410 (BsmI)	0,60
rs2853559	1,20	rs4077869	0,80	rs2239185	1,10
rs2254210	0,04	rs9729	0,10	rs7968585	0,40
rs2248098	0,40	rs2239179	0,60	rs11608702	0,05
rs2853564	0,60	rs7302235	0,60	rs10783219	0,60
rs3890734	0,40	rs4516035	1,20	rs886441	0,60

Fonte: O Autor

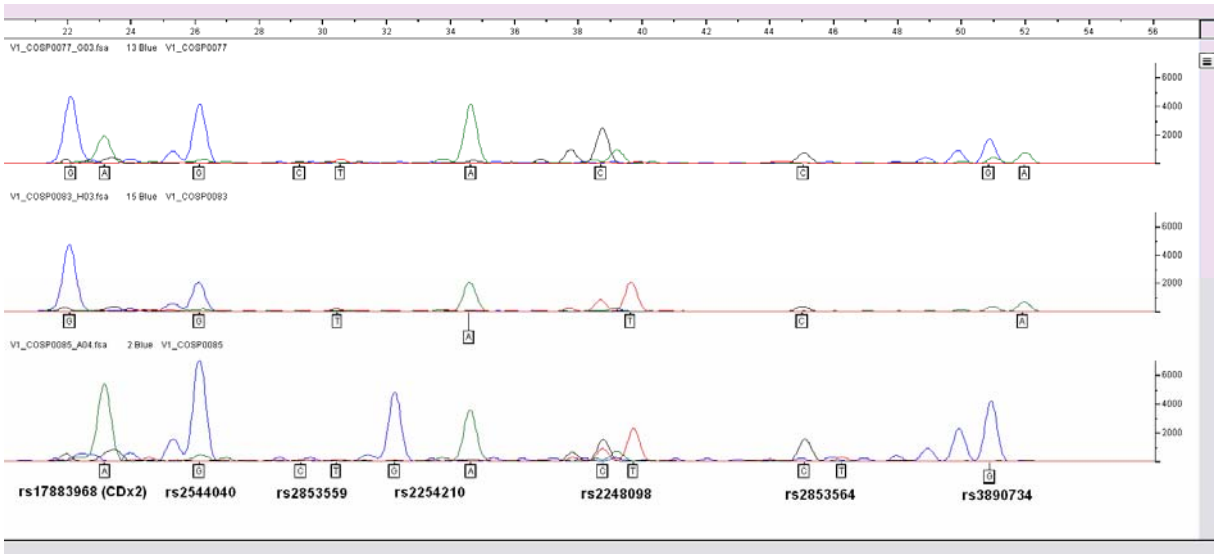


Figura 3.3 : Eletroferograma da reação SNaPshot™ para o multiplex 1 do gene *VDR*. Indivíduos: COSP0077, COSP0083 e COSP0085.
Fonte: O Autor

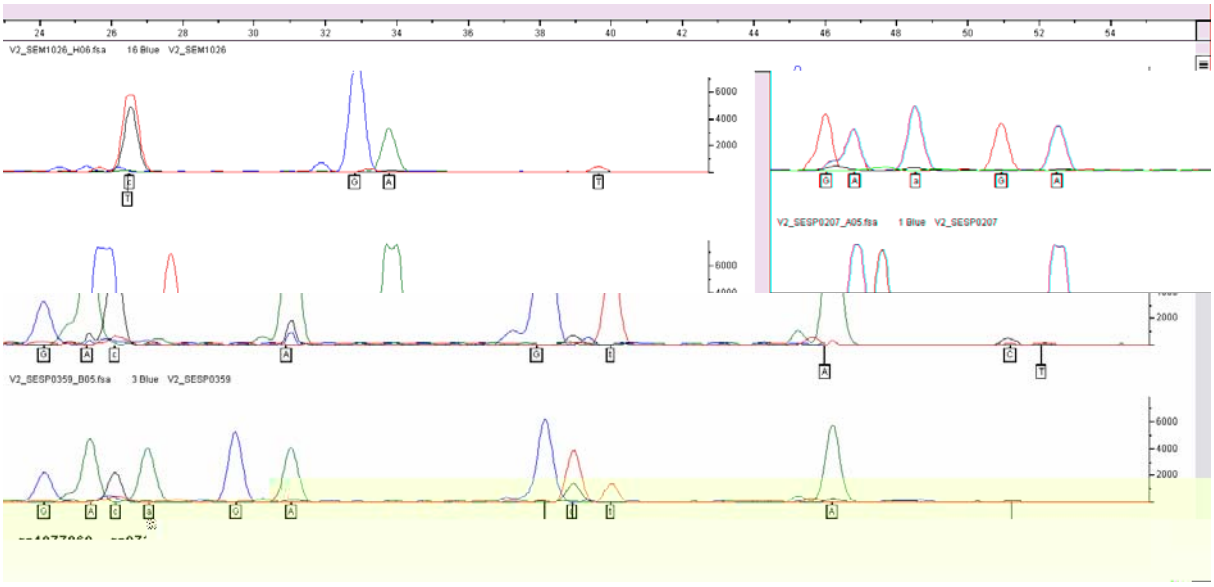


Figura 3.4 : Eletroferograma da reação SNaPshot™ para o multiplex 2 do gene *VDR*. Indivíduos: SEM1026, SESP0207 e SESP0359.
Fonte: O Autor

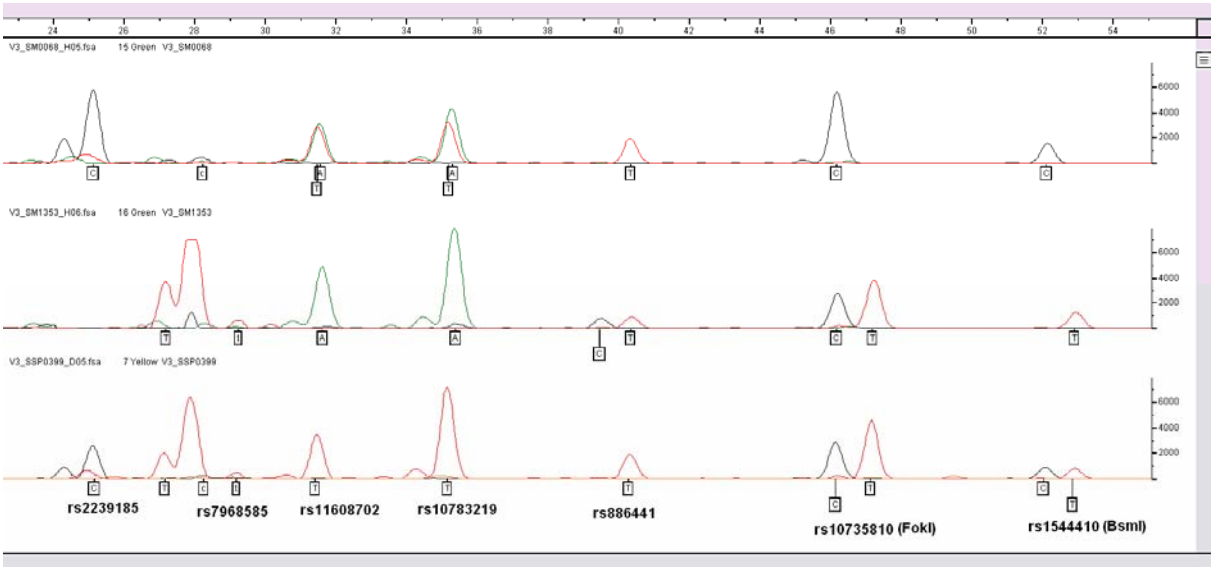


Figura 3.5 : Eletroferograma da reação SNaPshot™ para o multiplex 3 do gene *VDR*. Indivíduos: SM0068, SM1353 e SSP0399.
Fonte: O Autor

3.2.4 Análises Genéticas

Para análise genética, os alelos foram organizados no sentido reverso da fita de DNA, uma vez que o gene se posiciona no sentido 3'-5'. Os bancos de dados de SNPs disponíveis no NCBI e no HapMap sempre relatam os SNPs de acordo com a posição no cromossomo, ou no contig, em relação a orientação positiva do DNA, no entanto, em genes que seguem a orientação negativa do DNA os haplótipos são relatados na mesma orientação do gene. Portanto os alelos genotipados foram convertidos para seus respectivos alelos na orientação 5'-3' do gene (Quadro 3.2).

Quadro 3.1 : Conversão dos alelos para estimativa dos haplótipos no gene *VDR*.

Loco	rs4077869		rs11568820 (Cdx-2)		rs4516035		rs10783219		rs7302235		rs3890734		rs2853559	
Alelo Genotipado	G	A	G	A	C	T	A	T	G	A	G	A	C	T
Alelo Gene	G	A	C	T	G	A	T	A	G	A	C	T	G	A

Loco	rs2853564		rs2254210		rs10735810 (FokI)		rs886441		rs2239179		rs2248098		rs2239185	
Alelo Genotipado	C	T	G	A	C	T	C	T	C	T	C	T	C	T
Alelo Gene	G	A	C	T	C	T	G	A	C	T	G	A	G	A

Loco	rs1544410 (BsmI)		rs7975232 (ApaI)		rs731236 (TaqI)		rs9729		rs7968585		rs11608702		rs2544040	
Alelo Genotipado	C	T	G	T	G	A	A	C	C	T	A	T	G	A
Alelo Gene	C	T	G	T	G	A	T	G	G	A	T	A	C	T

Fonte: O Autor

As estimativas de freqüências alélicas dos locos SNPs foram tomadas a partir dos dados genotípicos gerados para cada subpopulação regional (Tabela 3.4) e para as populações brasileira e as do HapMap (Tabela 3.5) e foram utilizadas para a análise de equilíbrio de Hardy-Weinberg (EHW) e análise de variância para as populações. Para a maioria das populações a maior parte dos locos se encontrou em EHW (Tabela 3.6 e 3.7).

O índice de fixação (Fst) foi medido na população brasileira dividida por amostras regionais e revelou pequena e não significativa diferença genética (Fst = 0,003 e intervalos de confiança a 95% superior = 0,008 e inferior = -0,0006). Entre as populações, o índice de fixação revelou pequena, porém significativa, diferença genética entre a população brasileira e as populações do HapMap (Fst = 0,076720 e intervalos de confiança a 95%: superior = 0,107865 e inferior = 0,046214). O teste par a par revelou diferenciação de pequena a moderada, segundo a tabela de Wright, para os pares da população brasileira com as populações do HapMap. Porém, a diferença entre BRA e CEU não foi significativa. Entre as populações do HapMap, houve alta diferença entre YRI e ASN e nas demais a diferença foi moderada (Tabela 3.8).

Tabela 3.4 : Distribuição das freqüências alélicas dos polimorfismos do gene VDR nas regiões brasileiras.

Locos	Alelos	Populações Regionais				
		BRA_CO	BRA_NE	BRA_N	BRA_SE	BRA_S
rs4077869	A	0,650	0,650	0,538	0,731	0,769
	G	0,350	0,350	0,463	0,269	0,231
rs11568820 (Cdx-2)	C	0,637	0,705	0,663	0,688	0,731
	T	0,363	0,295	0,338	0,313	0,269
rs4516035	A	0,789	0,786	0,778	0,684	0,958
	G	0,211	0,214	0,222	0,316	0,042
rs10783219	A	0,333	0,325	0,244	0,282	0,346
	T	0,667	0,675	0,756	0,718	0,654
rs7302235	A	0,724	0,776	0,637	0,643	0,731
	G	0,276	0,224	0,363	0,357	0,269
rs3890734	C	0,700	0,688	0,679	0,675	0,676
	T	0,300	0,313	0,321	0,325	0,324
rs2853559	A	0,545	0,544	0,414	0,395	0,361
	G	0,455	0,456	0,586	0,605	0,639
rs2853564	A	0,410	0,338	0,262	0,488	0,363
	G	0,590	0,663	0,738	0,512	0,637
rs2254210	C	0,675	0,712	0,700	0,613	0,738
	T	0,325	0,287	0,300	0,387	0,262
rs10735810 (Fok)	C	0,667	0,744	0,675	0,628	0,658
	T	0,333	0,256	0,325	0,372	0,342
rs886441	A	0,868	0,718	0,800	0,705	0,744
	G	0,132	0,282	0,200	0,295	0,256
rs2239179	C	0,500	0,408	0,462	0,463	0,371
	T	0,500	0,592	0,538	0,537	0,629
rs2248098	A	0,425	0,551	0,387	0,450	0,512
	G	0,575	0,449	0,613	0,550	0,488
rs2239185	A	0,600	0,574	0,552	0,615	0,500
	G	0,400	0,426	0,448	0,385	0,500
rs1544410 (BsmI)	C	0,618	0,590	0,637	0,526	0,654
	T	0,382	0,410	0,363	0,474	0,346
rs7975232 (ApaI)	G	0,425	0,500	0,475	0,385	0,512
	T	0,575	0,500	0,525	0,615	0,488
rs731236 (TaqI)	A	0,700	0,603	0,625	0,551	0,654
	G	0,300	0,397	0,375	0,449	0,346
rs9729	G	0,412	0,395	0,375	0,295	0,474
	T	0,587	0,605	0,625	0,705	0,526
rs7968585	A	0,433	0,483	0,500	0,645	0,528
	G	0,567	0,517	0,500	0,355	0,472
rs11608702	A	0,395	0,350	0,275	0,295	0,423
	T	0,605	0,650	0,725	0,705	0,577
rs2544040	C	1,000	1,000	1,000	1,000	0,925
	T	0,000	0,000	0,000	0,000	0,075

Fonte: O Autor

Tabela 3.5 : Distribuição das Freqüências alélicas dos seis locos nos polimorfismos do gene VDR. Distribuição das freqüências alélicas dos polimorfismos do gene VDR na população brasileira e nas populações do HapMap.

		Populações			
Locos	Alelos	BRA	CEU	YRI	ASN
rs4077869	A	0,667		0,400	0,978
	G	0,333		0,600	0,022
rs11568820 (Cdx-2)	C	0,684	0,792	0,017	0,506
	T	0,316	0,208	0,983	0,494
rs4516035	A	0,788	0,575	0,992	0,989
	G	0,212	0,425	0,008	0,011
rs10783219	A	0,306	0,331	0,000	0,416
	T	0,694	0,669	1,000	0,584
rs7302235	A	0,703	0,741	0,425	
	G	0,297	0,259	0,575	
rs3890734	C	0,684	0,675	0,877	0,994
	T	0,316	0,325	0,123	0,006
rs2853559	A	0,449	0,424	0,169	0,421
	G	0,551	0,576	0,831	0,579
rs2853564	A	0,372	0,583	0,908	0,579
	G	0,628	0,417	0,092	0,421
rs2254210	C	0,688	0,633	0,658	0,624
	T	0,313	0,367	0,342	0,376
rs10735810 (FokI)	C	0,674	0,525	0,833	0,646
	T	0,326	0,475	0,167	0,354
rs886441	A	0,767	0,808	0,583	0,978
	G	0,233	0,192	0,417	0,022
rs2239179	C	0,443	0,417	0,292	0,247
	T	0,557	0,583	0,708	0,753
rs2248098	A	0,465	0,425	0,381	0,663
	G	0,535	0,575	0,619	0,337
rs2239185	A	0,568		0,542	0,347
	G	0,432		0,458	0,653
rs1544410 (BsmI)	C	0,605	0,525	0,712	0,921
	T	0,395	0,475	0,288	0,079
rs7975232 (ApaI)	G	0,460	0,424	0,375	0,645
	T	0,540	0,576	0,625	0,355
rs731236 (TaqI)	A	0,627	0,526	0,750	0,933
	G	0,373	0,474	0,250	0,067
rs9729	G	0,390	0,414	0,337	0,657
	T	0,610	0,586	0,663	0,343
rs7968585	A	0,525	0,592	0,633	0,326
	G	0,475	0,408	0,367	0,674
rs11608702	A	0,347	0,358	0,183	0,618
	T	0,653	0,642	0,817	0,382
rs2544040	C	0,985	1,000	0,885	1,000
	T	0,015	0,000	0,115	0,000

Fonte: O Autor

Tabela 3.6 : Distribuição das amostras regionais brasileiras segundo o loco, o número de amostras genotipadas (N), heterozigose observada (Ho), heterozigose esperada (He) e valor de p para o teste exato de Fisher para o equilíbrio de Hardy-Weinberg (p-val). Números em negrito indicam p< 0,05.

	BRA_CO				BRA_NE			
Locos	N	Ho	He	EHW p-val	N	Ho	He	EHW p-val
rs4077869	40	0,700	0,455	**0,002**	40	0,700	0,455	**0,001**
rs11568820 (Cdx-2)	40	0,525	0,462	0,686	39	0,538	0,416	0,164
rs4516035	19	0,211	0,332	**0,021**	7	0,143	0,337	**<0,001**
rs10783219	39	0,359	0,444	0,104	40	0,450	0,439	0,753
rs7302235	38	0,342	0,400	0,178	38	0,289	0,347	0,118
rs3890734	40	0,350	0,420	0,138	40	0,425	0,430	0,613
rs2853559	33	0,545	0,496	0,893	34	0,559	0,496	0,730
rs2853564	39	0,564	0,484	0,546	40	0,475	0,447	0,979
rs2254210	40	0,400	0,439	0,299	40	0,475	0,410	0,576
rs10735810 (FokI)	39	0,359	0,444	0,103	39	0,462	0,381	0,418
rs886441	38	0,263	0,229	0,981	39	0,308	0,405	**0,040**
rs2239179	39	0,538	0,500	0,955	38	0,447	0,483	0,355
rs2248098	40	0,600	0,489	0,302	39	0,385	0,495	0,069
rs2239185	35	0,571	0,480	0,508	34	0,441	0,489	0,323
rs1544410 (BsmI)	38	0,395	0,472	0,149	39	0,359	0,484	0,041
rs7975232 (ApaI)	40	0,600	0,489	0,281	40	0,500	0,500	0,693
rs731236 (TaqI)	40	0,450	0,420	0,996	39	0,333	0,479	0,018
rs9729	40	0,475	0,485	0,575	38	0,421	0,478	0,260
rs7968585	30	0,467	0,491	0,428	29	0,414	0,499	0,158
rs11608702	38	0,474	0,478	0,656	40	0,300	0,455	**0,011**
rs2544040	40	0,000	0,000	**<0,001**	39	0,000	0,000	**<0,001**

	BRA_N				BRA_SE			
Locos	N	Ho	He	EHW p-val	N	Ho	He	EHW p-val
rs4077869	40	0,925	0,497	**<0,001**	39	0,487	0,393	0,329
rs11568820 (Cdx-2)	40	0,575	0,447	0,157	40	0,425	0,430	0,601
rs4516035	9	0,222	0,346	**0,023**	19	0,316	0,432	0,065
rs10783219	39	0,333	0,369	0,245	39	0,410	0,405	0,701
rs7302235	40	0,525	0,462	0,659	35	0,314	0,459	**0,016**
rs3890734	39	0,333	0,436	0,061	40	0,400	0,439	0,308
rs2853559	35	0,657	0,485	0,099	38	0,526	0,478	0,828
rs2853564	40	0,325	0,387	0,148	40	0,625	0,500	0,247
rs2254210	40	0,450	0,420	0,979	40	0,475	0,475	0,666
rs10735810 (FokI)	40	0,400	0,439	0,330	39	0,487	0,467	0,889
rs886441	40	0,250	0,320	0,068	39	0,385	0,416	0,354
rs2239179	39	0,615	0,497	0,274	27	0,556	0,497	0,916
rs2248098	40	0,425	0,475	0,254	40	0,400	0,495	0,109
rs2239185	29	0,414	0,495	0,158	39	0,462	0,473	0,539
rs1544410 (BsmI)	40	0,375	0,462	0,095	39	0,436	0,499	0,228
rs7975232 (ApaI)	40	0,550	0,499	0,821	39	0,359	0,473	0,056
rs731236 (TaqI)	40	0,400	0,469	0,187	39	0,436	0,495	0,258
rs9729	40	0,350	0,469	**0,043**	39	0,282	0,416	**0,008**
rs7968585	27	0,333	0,500	**0,020**	38	0,342	0,458	**0,040**
rs11608702	40	0,300	0,399	**0,037**	39	0,487	0,416	0,526
rs2544040	40	0,000	0,000	**<0,001**	40	0,000	0,000	**<0,001**

	BRA_S			
Locos	N	Ho	He	EHW p-val
rs4077869	39	0,462	0,355	0,161
rs11568820 (Cdx-2)	39	0,487	0,393	0,304
rs4516035	12	0,083	0,080	**<0,001**
rs10783219	39	0,436	0,453	0,516
rs7302235	39	0,385	0,393	0,509
rs3890734	34	0,471	0,438	0,939
rs2853559	36	0,556	0,461	0,442
rs2853564	40	0,275	0,462	**0,004**
rs2254210	40	0,375	0,387	0,484
rs10735810 (FokI)	38	0,474	0,450	0,919
rs886441	39	0,462	0,381	0,416
rs2239179	31	0,484	0,467	0,786
rs2248098	40	0,425	0,500	0,171
rs2239185	39	0,436	0,500	0,229
rs1544410 (BsmI)	39	0,385	0,453	0,172
rs7975232 (ApaI)	40	0,475	0,500	0,471
rs731236 (TaqI)	39	0,385	0,453	0,171
rs9729	39	0,436	0,499	0,224
rs7968585	36	0,444	0,498	0,264
rs11608702	39	0,487	0,488	0,678
rs2544040	40	0,150	0,139	0,345

Fonte: O Autor

Tabela 3.7 : Distribuição da população brasileira e das populações do HapMap segundo o loco, o número de amostras genotipadas (N), heterozigose observada (Ho), heterozigose esperada (He) e valor de p para o teste exato de Fisher para o equilíbrio de Hardy-Weinberg (p-val). Números em negrito indicam p< 0,05.

	BRA				CEU			
Locos	N	Ho	He	EHW p-val	N	Ho	He	EHW p-val
rs4077869	198	0,657	0,444	**<0,001**	0	-	-	-
rs11568820 (Cdx-2)	198	0,510	0,432	**0,019**	60	0,350	0,330	0,167
rs4516035	66	0,212	0,334	**0,001**	60	0,517	0,489	0,450
rs10783219	196	0,398	0,425	0,256	59	0,492	0,443	0,322
rs7302235	190	0,374	0,418	0,101	54	0,407	0,384	0,218
rs3890734	193	0,394	0,432	0,151	60	0,550	0,439	0,350
rs2853559	176	0,568	0,495	0,073	59	0,475	0,488	0,432
rs2853564	199	0,452	0,467	0,528	60	0,467	0,486	0,425
rs2254210	200	0,435	0,430	0,949	60	0,400	0,464	0,375
rs10735810 (FokI)	195	0,436	0,439	0,756	60	0,483	0,499	0,442
rs886441	195	0,333	0,358	0,250	60	0,350	0,310	0,192
rs2239179	174	0,529	0,493	0,436	60	0,433	0,486	0,408
rs2248098	199	0,447	0,498	0,107	60	0,383	0,489	0,433
rs2239185	176	0,466	0,491	0,389	0	-	-	-
rs1544410 (BsmI)	195	0,390	0,478	**0,005**	60	0,483	0,499	0,442
rs7975232 (ApaI)	199	0,497	0,497	0,841	59	0,407	0,488	0,422
rs731236 (TaqI)	197	0,401	0,468	**0,025**	58	0,500	0,499	0,440
rs9729	196	0,393	0,476	**0,011**	58	0,414	0,485	0,422
rs7968585	160	0,400	0,499	**0,004**	60	0,417	0,483	0,442
rs11608702	196	0,408	0,453	0,121	60	0,483	0,460	0,375
rs2544040	199	0,030	0,030	0,069	60	0,000	0,000	0,000

	ASN				YRI			
Locos	N	Ho	He	EHW p-val	N	Ho	He	EHW p-val
rs4077869	89	0,045	0,044	**0,022**	60	0,533	0,480	0,417
rs11568820 (Cdx-2)	89	0,472	0,500	0,494	60	0,033	0,033	**0,017**
rs4516035	89	0,022	0,022	**0,011**	60	0,017	0,017	**0,008**
rs10783219	89	0,517	0,486	0,416	60	0,000	0,000	0,000
rs7302235	0	-	-	-	60	0,550	0,489	0,433
rs3890734	85	0,012	0,012	**0,006**	57	0,246	0,215	0,114
rs2853559	89	0,483	0,488	0,421	59	0,305	0,282	0,169
rs2853564	89	0,483	0,488	0,421	60	0,183	0,167	0,092
rs2254210	89	0,416	0,469	0,383	60	0,550	0,450	0,328
rs10735810 (Fok)	89	0,416	0,457	0,354	60	0,300	0,278	0,167
rs886441	89	0,022	0,044	**0,022**	60	0,367	0,486	0,417
rs2239179	89	0,337	0,372	0,247	60	0,350	0,413	0,283
rs2248098	89	0,382	0,447	0,339	59	0,458	0,472	0,381
rs2239185	88	0,443	0,453	0,347	60	0,483	0,497	0,474
rs1544410 (BsmI)	89	0,157	0,145	0,079	59	0,339	0,410	0,297
rs7975232 (ApaI)	86	0,453	0,458	0,356	60	0,517	0,469	0,375
rs731236 (TaqI)	89	0,135	0,126	0,067	58	0,362	0,375	0,259
rs9729	89	0,438	0,451	0,343	52	0,558	0,447	0,330
rs7968585	89	0,449	0,439	0,328	60	0,567	0,464	0,358
rs11608702	89	0,472	0,472	0,382	60	0,300	0,299	0,175
rs2544040	88	0,000	0,000	0,000	52	0,231	0,204	0,125

Fonte: O Autor

Tabela 3.8: Matriz do índice de fixação (Fst) par a par entre as populações. Na diagonal inferior estão os valores de Fst e diagonal superior seus respectivos testes de significância com intervalos de confiança superior e inferior a 95%.

	BRA	YRI	ASN	CEU
BRA		0,171 ic Sup 0,033 ic Inf	0,088 ic Sup 0,028 ic Inf	0,056 ic Sup - 0,001 ic Inf
YRI	0,097		0,286 ic Sup 0,111 ic Inf	0,173 ic Sup 0,035 ic Inf
ASN	0,056	0,196		0,094 ic Sup 0,024 ic Inf
CEU	0,024	0,099	0,053	

Fonte: O Autor

3.2.5 Padrões de Desequilíbrio de Ligação

Os valores de desequilíbrio de ligação foram utilizados para montar as estruturas de blocos baseados nos valores de D'. Para a população brasileira foi observada a formação de dois blocos, um na região promotora do gene e outro que engloba o final da região de transcrição e a região 3'UTR (Figura 3.6). Quando separada por regiões, os padrões de DL permanecem com esses dois blocos, diferenciando apenas nos tamanhos e em pequenos blocos formados na região promotora, codificadora ou 3'UTR. Dentro da população brasileira, a região Centro-Oeste é a que apresenta diferente padrão de blocos haplotípicos, onde o DL é menor e são formados quatro blocos de relativa baixa extensão, entre 1 e 4 kb (Figura 3.6).

Na população brasileira, o maior bloco engloba a região que antecede o exon 4 até a região 3' intergênica, com 21 kb de extensão. O mesmo bloco também está presente nas regiões Norte e Sudeste. Nas regiões Sul e Sudeste este bloco possui 16kb de extensão e chega até a região 3'UTR. Nas regiões Norte e Sul ainda é observado um pequeno bloco de 5kb antes desse bloco maior. Esse bloco contém os polimorfismos BsmI, ApaI e TaqI, os quais se mostraram em alto DL em todas as populações, inclusive no Centro-Oeste com um bloco de 1kb (Figura 3.6).

O segundo bloco é observado na região promotora do gene e se estende por 13 kb na população brasileira e também é observado na amostra do Nordeste. Nas amostras do Sul e Sudeste o bloco formado nessa região é menor, com 9kb e na amostra do Norte são observados dois blocos nessa região, um com 7kb e outro com 2kb. Nas amostras do Nordeste e Centro-Oeste ainda foi observada a formação de um bloco de 5kb que antecede a região codificadora do gene. Regiões com ausência de blocos também foram observadas. Mais precisamente, entre os SNPs rs3890734 e rs2853559 não foi encontrado nenhum bloco em nenhuma das populações, incluindo as do HapMap. Outro ponto onde não foi encontrado blocos foi no FokI, que já foi mencionado em outros estudos como ponto de recombinação no gene (FANG et al., 2005; NEJENTSEV et al., 2004). A figura

3.7 mostra a distribuição dos blocos em cada população e subpopulação de acordo com a posição dos marcadores no gene.

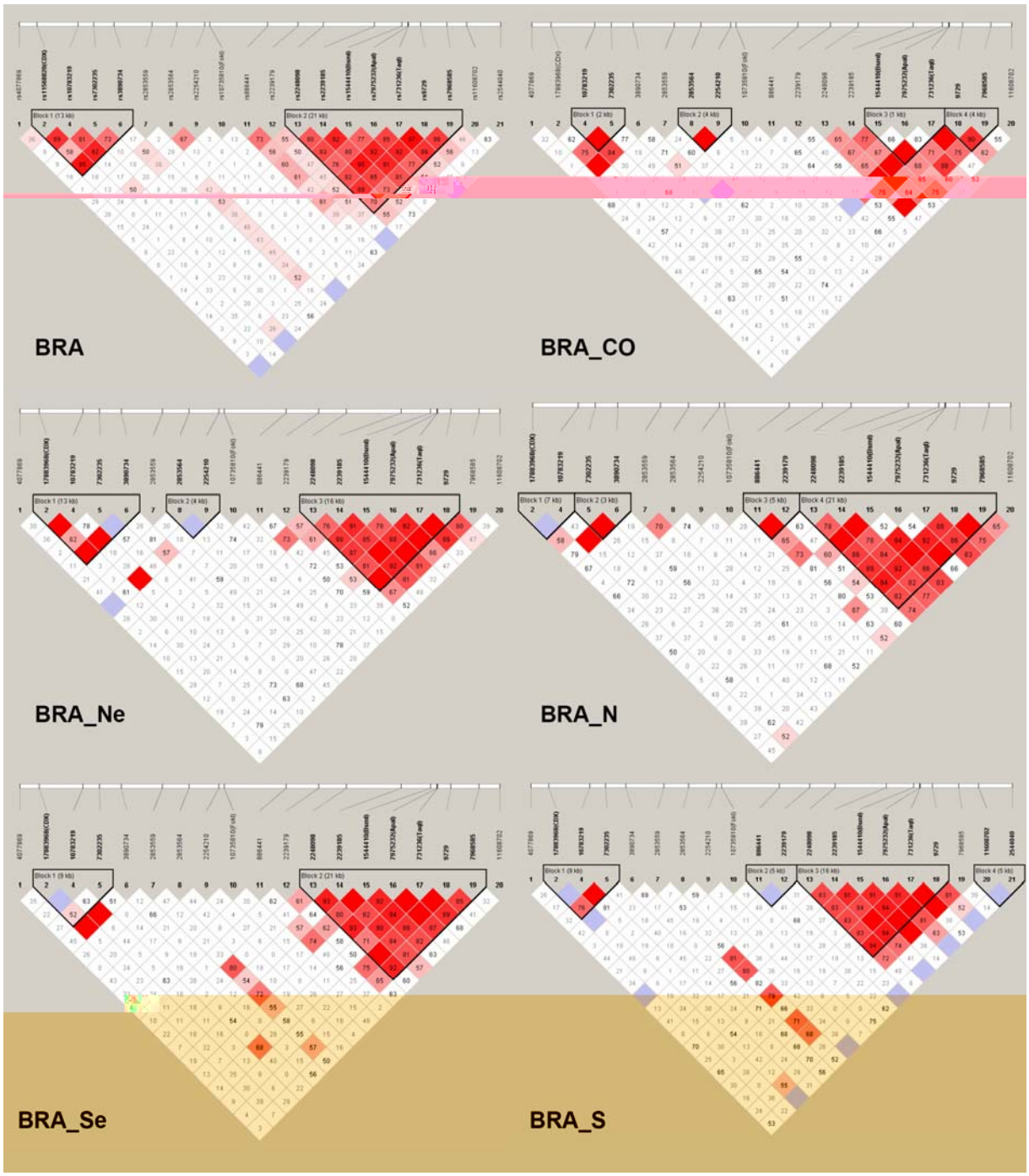


Figura 3.6 : Desequilíbrio de ligação entre os marcadores selecionados na população brasileira e nas amostras regionais.
Fonte: O Autor

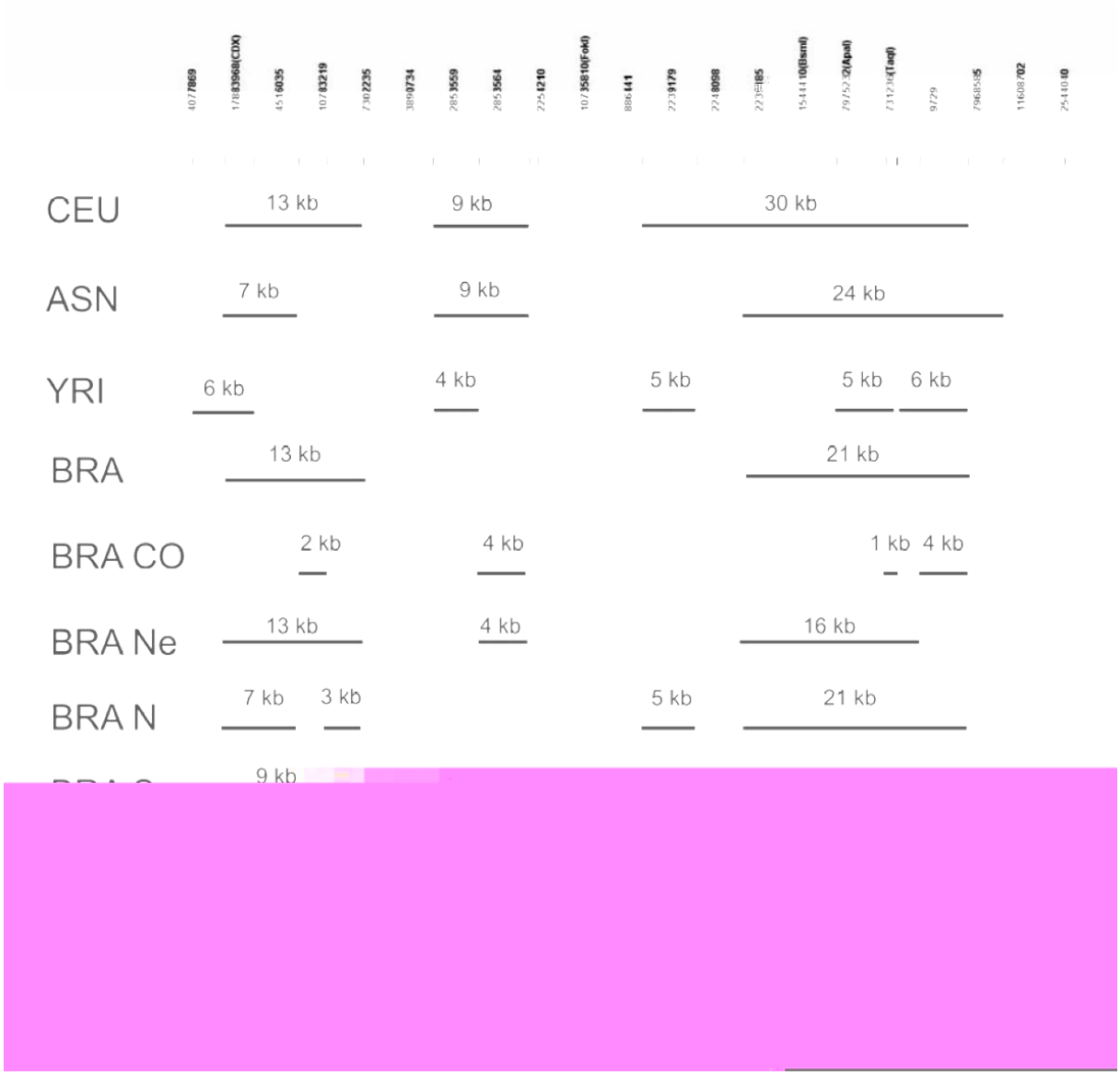


Figura 3.7 : Distribuição esquemática dos blocos haplotípicos em cada população e amostra regional, de acordo com a posição dos marcadores no gene. Linha pontilhada indica regiões de baixo DL onde não foram encontrados blocos haplotípicos.
Fonte: O Autor

A análise de haplótipos realizada quando todos os marcadores foram usados estimou 299 haplótipos sem premissa de recombinação. Neste caso, o programa Phase foi utilizado para estimar dos pontos quentes de recombinação dentro do gene (Figura 3.8) de acordo com a taxa de recombinação basal (0,0022) e os fatores nos quais os pares de marcadores excedem a taxa de recombinação. Em seguida, a estimativa de freqüência haplotípica foi feita separadamente com os dois blocos identificados, denominados 5' e 3' de acordo com suas posições. O bloco 5' incluiu quatro locos, do Cdx-2 ao rs3890734, exceto o rs4516035 por estar com muitos dados faltantes. O bloco 3' inclui os SNPs rs2248098 ao rs7968585. A partir

desses blocos, foram estimados oito haplótipos para o bloco 5' e 14 haplótipos para o bloco 3', utilizando freqüência mínima de 1% (Tabela 3.9). A matriz percentual de *crossing-over* mostra a probabilidade em que um indivíduo carrega cada haplótipo ao mesmo tempo (Tabela 3.10). A tabela 3.11 mostra os valores médios para as estatísticas de DL nas populações do HapMap, na população brasileira e nas suas respectivas regiões definidas como bloco haplotípico.

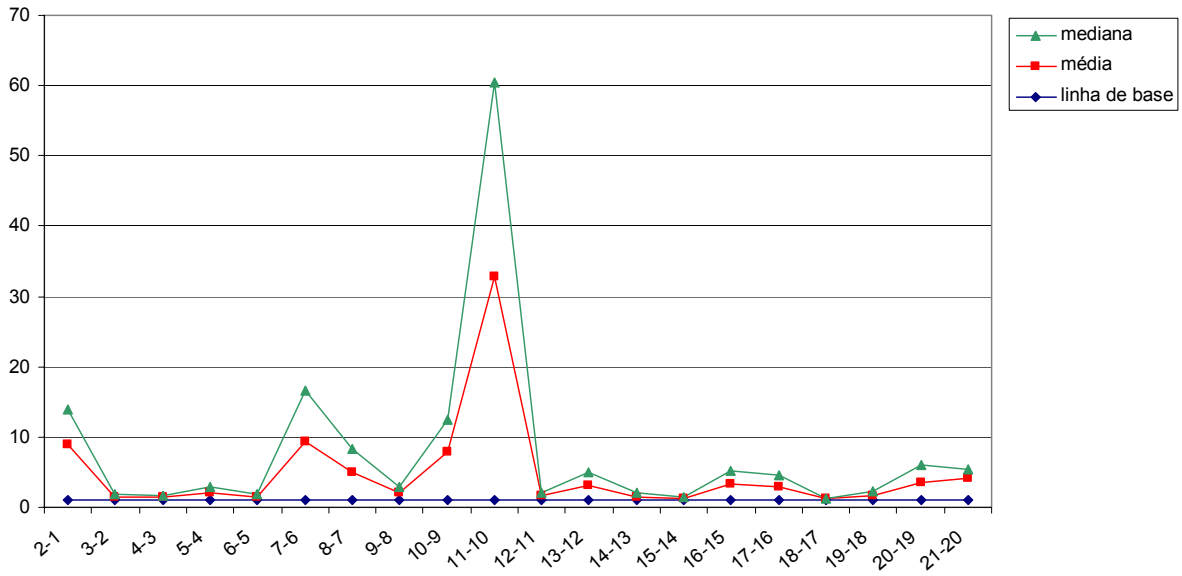


Figura 3.8 : Estimativa de ponto de recombinação entre pares de marcadores. O eixo das ordenadas indica o fator no qual os pares de marcadores excedem a taxa de recombinação basal. Marcadores seguem o sentido 5'-3' do gene, sendo (1) rs4077869 e (21) rs2544040.
Fonte: O Autor

Tabela 3.9 : Identificação dos haplótipos (Hap ID) e suas freqüências na população brasileira nos blocos 5' (A) e 3'(B).

A

Hap ID	Haplótipo	Freq
5'H01	CTAT	0,292
5'H02	CAAC	0,278
5'H03	TTGC	0,225
5'H04	TTAC	0,084
5'H05	CTGC	0,051
5'H06	CTAC	0,038
5'H07	CTGT	0,020
5'H08	CAGC	0,013

B

Hap ID	Haplótipo	Freq
3'H01	AGCGAGG	0,349
3'H02	GATTGTA	0,331
3'H03	GACTATA	0,086
3'H04	GACTATG	0,033
3'H05	GATGGTA	0,031
3'H06	AACTATA	0,027
3'H07	AGCGAGA	0,024
3'H08	AGCGATG	0,022
3'H09	GACGATA	0,020
3'H10	GGCTATA	0,018
3'H11	GATTGTG	0,018
3'H12	GATTATA	0,014
3'H13	AATTGTA	0,014
3'H14	AACGAGG	0,012

Fonte: O Autor

Tabela 3.10 : Matriz percentual de crossing-over entre os blocos haplotípicos.

Haplótipos	3'H01	3'H02	3'H03	3'H04	3'H05	3'H06	3'H07	3'H08	3'H09	3'H10	3'H11	3'H12	3'H13	3'H14
5'H01	0,102	0,107	0,025	0,011	0,000	0,000	0,000	0,000	0,000	0,000	0,010	0,000	0,010	0,000
5'H02	0,128	0,100	0,006	0,000	0,000	0,016	0,002	0,000	0,006	0,000	0,000	0,005	0,000	0,011
5'H03	0,048	0,068	0,026	0,014	0,014	0,000	0,000	0,009	0,000	0,003	0,000	0,005	0,000	0,000
5'H04	0,000	0,026	0,000	0,000	0,010	0,000	0,004	0,000	0,000	0,013	0,004	0,000	0,000	0,000
5'H05	0,022	0,004	0,003	0,000	0,000	0,000	0,005	0,000	0,003	0,000	0,000	0,003	0,000	0,000
5'H06	0,018	0,006	0,000	0,000	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
5'H07	0,000	0,000	0,009	0,004	0,000	0,000	0,005	0,000	0,000	0,000	0,000	0,000	0,000	0,000
5'H08	0,000	0,000	0,003	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000

Fonte: O Autor

Tabela 3.11 : Valores médios para as estatísticas de DL nas populações estudadas e nos blocos haplotípicos da população brasileira.

	D'	LOD	r^2	IC Inf	IC Sup
BRA	0,323	4,898	0,090	0,176	0,512
CEU	0,358	3,835	0,139	0,214	0,547
ASN	0,507	3,510	0,112	0,161	0,681
YRI	0,386	1,427	0,062	0,138	0,631
BRA 5'	0,818	9,592	0,182	0,637	0,900
BRA 3'	0,848	29,767	0,491	0,755	0,909

Fonte: O Autor

Para avaliar o efeito da ancestralidade genômica, foram feitos testes de correlação condicional em todos os haplótipos de cada bloco. O teste de regressão condicional *omnibus* não indicou correlação significativa para o bloco 3' (p = 0,988) e para o bloco 5' (p=0,998). Portanto, os testes de haplótipo específicos não puderam ser realizados, visto que quaisquer resultados seriam não significativos.

3.2.6 Transferibilidade de tagSNPs

As análises de transferibilidade de tagSNPs, feitas par a par entre a população brasileira e as populações do HapMap mostrou que os níveis de ancestralidade influenciam a predição de cobertura da variabilidade genética na população brasileira. As tabelas 3.12 a 3.14 descrevem os tagSNPs de cada população e a predição de variabilidade para cada número de marcadores usados.

O cálculo de perda relativa de variabilidade genética pela transferibilidade de tagSNPs das populações do HapMap para a população brasileira mostrou-se maior quando são usados os marcadores escolhidos para as populações ASN e YRI conforme aumenta-se o número de marcadores utilizados (Figura 3.9). O teste do percentual de variabilidade captada pelos SNPs utilizados no estudo em relação a todos os marcadores disponíveis na fase II do banco de dados do HapMap foi realizado para verificar o potencial do conjunto de SNPs selecionados (Figura 3.10).

Tabela 3.12 : Percentual de variabilidade captado por cada conjunto de marcadores nas populações BRA e YRI.

Num marc	Pop	% Var	rs4077869	Cdx-2	rs4516035	rs10783219	rs7302235	rs3890734	rs2853559	rs2853564	rs2254210	FokI	rs886441	rs2239179	rs2248098	rs2239185	BsmI	ApaI	TaqI	rs9729	rs7968585	rs11608702	rs2544040
2	YRI	64,65														V							V
	BRA	61,57						V										V					
3	YRI	71,85	V														V			V			
	BRA	66,02				V		V							V								
4	YRI	75,98	V								V						V			V			
	BRA	68,46				V		V							V				V				
5	YRI	78,02	V								V		V				V			V			
	BRA	70,38	V	V				V							V				V				
6	YRI	78,67	V								V		V		V		V			V			
	BRA	71,06	V	V				V				V			V				V				
7	YRI	80	V								V		V		V	V	V			V			
	BRA	72,36	V	V				V				V	V			V			V				
8	YRI	81,54	V								V		V	V	V	V	V			V			
	BRA	74,22	V	V				V	V		V	V			V				V				
9	YRI	82,64	V								V	V	V	V	V	V	V			V			
	BRA	76,47	V	V				V	V	V	V	V			V				V				
10	YRI	83,64	V				V				V	V	V	V	V	V	V			V			
	BRA	78,62	V	V				V	V	V	V	V	V			V			V				
11	YRI	84,83	V								V	V	V	V	V	V	V	V	V	V			
	BRA	80,45	V	V				V	V	V	V	V	V	V		V			V				
12	YRI	86,48	V								V	V	V	V	V	V	V	V	V		V	V	
	BRA	81,8	V	V				V	V	V	V	V	V		V		V				V	V	
13	YRI	88,33	V				V				V	V	V	V	V	V	V	V	V		V	V	
	BRA	83,92	V	V				V	V	V	V	V	V	V	V		V				V	V	
14	YRI	90,71	V				V		V		V	V	V	V	V	V	V	V	V		V	V	
	BRA	85,07	V	V				V	V	V	V	V	V	V	V		V			V	V	V	
15	YRI	93,06	V				V	V	V		V	V	V	V	V	V	V	V	V		V	V	
	BRA	86,01	V	V				V	V	V	V	V	V	V	V	V	V			V	V	V	
16	YRI	96	V				V	V	V		V	V	V	V	V	V	V	V	V		V	V	V
	BRA	85,63	V	V		V	V	V	V	V	V	V	V	V	V		V			V	V	V	
17	YRI	97,5	V				V	V	V	V	V	V	V	V	V	V	V	V	V		V	V	V
	BRA	87,31	V	V	V	V	V	V	V	V	V	V	V	V	V		V			V	V	V	
18	YRI	98,33	V				V	V	V	V	V	V	V	V	V	V	V	V	V	V	V	V	V
	BRA	89,95	V	V	V	V	V	V	V	V	V	V	V	V	V	V	V			V	V	V	
19	YRI	98,33	V		V		V	V	V	V	V	V	V	V	V	V	V	V	V	V	V	V	V
	BRA	88,44	V	V	V	V	V	V	V	V	V	V	V	V	V		V	V	V	V	V	V	V
20	YRI	100	V	V	V		V	V	V	V	V	V	V	V	V	V	V	V	V	V	V	V	V
	BRA	97,49	V	V	V	V	V	V	V	V	V	V	V	V	V	V	V	V	V	V	V	V	

Fonte: O Autor

Tabela 3.13 : Percentual de variabilidade captado por cada conjunto de marcadores nas populações BRA e CEU.

Num marc	Pop	% Var	Cdx-2	rs4516035	rs10783219	rs7302235	rs3890734	rs2853559	rs2853564	rs2254210	FokI	rs886441	rs2239179	rs2248098	BsmI	ApaI	TaqI	rs9729	rs7968585	rs11608702	rs2544040
2	CEU	71,37							V										V		
	BRA	59,71							V							V					
3	CEU	76,56	V						V					V							
	BRA	67,4			V		V							V							
4	CEU	80,44	V						V		V			V							
	BRA	70,05			V		V							V			V				
5	CEU	83,69		V		V			V		V			V							
	BRA	70,75			V		V				V			V			V				
6	CEU	85,38		V		V			V		V			V						V	
	BRA	72,13			V		V				V	V		V			V				
7	CEU	87,5		V		V			V		V		V					V		V	
	BRA	74,12			V		V	V	V		V			V			V				
8	CEU	89,7		V		V			V		V	V	V					V		V	
	BRA	76,56			V		V	V	V	V	V			V			V				
9	CEU	91,83		V		V		V		V	V	V	V					V		V	
	BRA	78,94			V		V	V	V	V	V	V		V			V				
10	CEU	94,07		V	V		V		V	V	V	V	V					V		V	
	BRA	80,51			V		V	V	V	V	V	V	V	V			V				
11	CEU	95,21		V	V		V		V	V	V	V	V		V			V		V	
	BRA	82,79			V		V	V	V	V	V	V	V		V				V	V	
12	CEU	96,9		V	V		V		V	V	V	V	V	V	V			V		V	
	BRA	85,07			V		V	V	V	V	V	V	V	V	V				V	V	
13	CEU	98,33		V	V	V	V		V	V	V	V	V	V	V			V		V	
	BRA	86,6			V		V	V	V	V	V	V	V	V	V			V	V	V	
14	CEU	99,33	V	V	V	V	V		V	V	V	V	V	V	V			V		V	
	BRA	85,03			V		V	V	V	V	V	V	V	V	V	V	V		V	V	
15	CEU	99,58	V	V	V	V	V		V	V	V	V	V	V	V			V	V	V	
	BRA	87,19		V	V	V	V	V	V	V	V	V	V	V	V			V	V	V	
16	CEU	100	V	V	V	V	V	V	V	V	V	V	V	V	V			V	V	V	
	BRA	89,95	V	V	V	V	V	V	V	V	V	V	V	V	V			V	V	V	
17	CEU	100	V	V	V	V	V	V	V	V	V	V	V	V	V			V	V	V	V
	BRA	88,19		V	V	V	V	V	V	V	V	V	V	V	V	V	V	V	V		
18	CEU	98,33	V	V	V	V	V		V	V	V	V	V	V	V	V	V	V	V	V	V
	BRA	97,49		V	V	V	V	V	V	V	V	V	V	V	V	V	V	V	V	V	V

Fonte: O Autor

Tabela 3.14 : Percentual de variabilidade captado por cada conjunto de marcadores nas populações BRA e ASN.

Num marc	Pop	% Var	rs4077869	Cdx-2	rs4516035	rs10783219	rs3890734	rs2853559	rs2853564	rs2254210	FokI	rs886441	rs2239179	rs2248098	rs2239185	BsmI	ApaI	TaqI	rs9729	rs7968585	rs11608702	rs2544040
2	ASN	80,84							V												V	
	BRA	63,3					V										V					
3	ASN	88,37				V			V						V							
	BRA	66,95		V			V										V					
4	ASN	91,15				V			V				V		V							
	BRA	69,47		V			V							V				V				
5	ASN	93,78				V			V		V		V		V							
	BRA	70,42	V	V			V							V				V				
6	ASN	94,62				V			V	V	V		V		V							
	BRA	71,32		V			V			V	V			V				V				
7	ASN	95,59				V			V	V	V		V		V	V						
	BRA	73,02		V			V	V	V		V			V				V				
8	ASN	96,54		V		V			V	V	V		V		V	V						
	BRA	75,25		V			V	V	V	V	V			V				V				
9	ASN	96,94		V		V			V	V	V		V	V	V	V						
	BRA	77,3		V			V	V	V	V	V	V			V			V				
10	ASN	97,3		V		V			V	V	V		V	V		V			V		V	
	BRA	79,5	V	V			V	V	V	V	V	V			V			V				
11	ASN	97,63		V		V			V	V	V		V	V		V			V	V	V	
	BRA	81,63	V	V			V	V	V	V	V	V			V			V				
12	ASN	98,03		V		V			V	V	V		V	V	V			V	V	V	V	
	BRA	83,29	V	V			V	V	V	V	V	V		V		V				V	V	
13	ASN	98,39	V	V		V			V	V	V		V	V	V			V	V	V	V	
	BRA	85,93	V	V			V	V	V	V	V	V	V	V		V				V	V	
14	ASN	98,69	V	V		V			V	V	V	V	V	V	V			V	V	V	V	
	BRA	87,6	V	V			V	V	V	V	V	V	V	V		V			V	V	V	
15	ASN	98,88	V	V	V	V			V	V	V	V	V	V	V			V	V	V	V	
	BRA	89,25	V	V			V	V	V	V	V	V	V	V	V	V			V	V	V	
16	ASN	99,16	V	V		V			V	V	V	V	V	V	V	V	V	V	V	V	V	
	BRA	88,32	V	V			V	V	V	V	V	V	V	V		V	V	V	V	V	V	
17	ASN	99,63	V	V	V	V			V	V	V	V	V	V	V	V	V	V	V	V	V	
	BRA	91,29	V	V			V	V	V	V	V	V	V	V	V	V	V	V	V	V	V	
18	ASN	100	V	V	V	V	V		V	V	V	V	V	V	V	V	V	V	V	V	V	
	BRA	88,4	V	V	V	V	V	V	V	V	V	V	V	V		V	V	V	V	V	V	
19	ASN	100	V	V	V	V	V		V	V	V	V	V	V	V	V	V	V	V	V	V	V
	BRA	97,49	V	V	V	V	V	V	V	V	V	V	V	V	V	V	V	V	V	V	V	

Fonte: O Autor

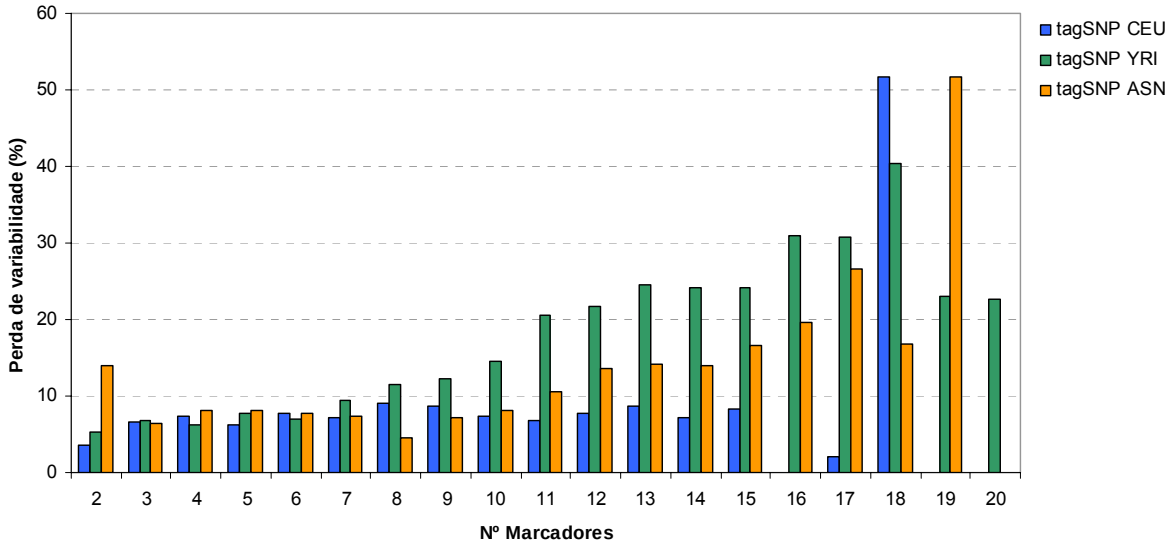


Figura 3.9 : Perda percentual relativa da variabilidade genética da população brasileira captada por cada conjunto de tagSNPs definido em cada população do HapMap.
Fonte: O Autor

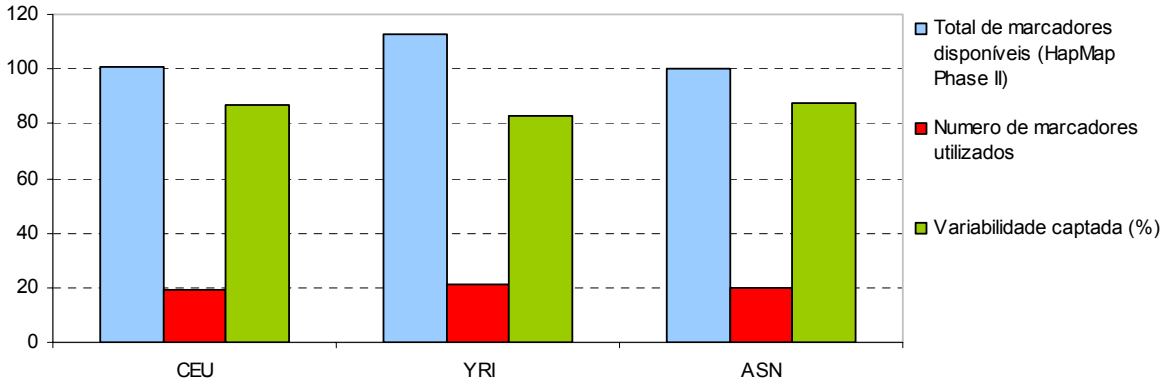


Figura 3.10 : Percentual de variabilidade captada pelos SNPs utilizados no estudo em relação a todos os marcadores disponíveis na fase II do banco de dados do HapMap.
Fonte: O Autor

3.3 DISCUSSÃO

O sistema endócrino da vitamina D no metabolismo humano está envolvido em uma variedade de vias metabólicas, desde controle do metabolismo ósseo, modulação de respostas imunológicas até a regulação da proliferação e diferenciação celular (UITTERLINDEN et al., 2004; VALDIVIELSO; FERNANDEZ, 2006). Variações ao longo dessas vias que alteram a homeostase da vitamina D conseqüentemente levam à predisposição a várias doenças e fenótipos complexos, tais como, osteoporose (BARRETT-CONNOR et al., 2005; FANG et al., 2005; ZMUDA et al., 1999; ZMUDA; CAULEY; FERRELL, 2000), risco de fratura óssea (BARRETT-CONNOR et al., 2005; FANG et al., 2003; FANG et al., 2005; MORRISON et al., 1994; RAMALHO et al., 1998; UITTERLINDEN et al., 2001), diabetes (HAUACHE et al., 1998; MATHIEU; BADENHOOP, 2005; MOTOHASHI et al., 2003; NEJENTSEV et al., 2004; OGUNKOLADE et al., 2002) e câncer (INGLES; ROSS et al., 1997). Nesse sentido, o receptor de vitamina D (*VDR*) é um dos fortes genes candidatos a atuar nessas variações e na provável manifestação desses fenótipos. Um dos maiores problemas nos estudos de associação genética entre o gene *VDR* e fenótipos complexos é a falta de consistência entre resultados de diferentes grupos, gerados por um pequeno número de polimorfismos analisados e, também, pela falta de conhecimento sobre a influência conjunta e a relação entre eles (FANG et al., 2005; MACDONALD et al., 2006; NEJENTSEV et al., 2004; UITTERLINDEN et al., 2004; UITTERLINDEN et al., 1996; ZMUDA; CAULEY; FERRELL, 2000).

Dentre todos os SNPs disponíveis nos bancos de dados para o gene *VDR* e o conjunto formado pelas suas combinações procurou-se selecionar aqueles que poderiam ser utilizados numa análise de genotipagem, análise do DL e estrutura de blocos haplotípicos. Phillips e colaboradores sugeriram que longos blocos com alto DL emergem preferencialmente quando marcadores extensamente espaçados são usados e, assim, utilizando uma densidade crescente de marcadores, blocos mais curtos se tornam visíveis. Portanto, para o gene *VDR*, a escolha do conjunto de SNPs espaçados a cada 5 kb é comparável com outros estudos de DL em outras

regiões (PHILLIPS et al., 2003; STENZEL et al., 2004) e pode prover segurança tanto na determinação dos padrões de DL como na transferibilidade de SNPs das populações do HapMap para a população brasileira.

As genotipagens na população brasileira foram satisfatoriamente realizadas utilizando o protocolo desenvolvido para extensão de base única via SNaPshot™ Multiplex System (Applied Biosystems). Apesar de ter sido verificado problemas de amplificação, que geraram insegurança na genotipagem, como ruídos inespecíficos entre 20 e 26 pares de bases e baixo sinal em rfu em alguns locos, o sucesso de genotipagem para a população brasileira chegou a aproximadamente 93% de todos os genótipos esperados. As fontes desses problemas estão no excesso de primer de PCR na reação de extensão de base única não degradados pela purificação enzimática e ao desbalanço de amplificação de alguns locos, ou na reação de PCR ou na reação de extensão de base única. A taxa de genótipos foi prejudicada principalmente pelo loco rs4516035 que obteve somente 33% dos genótipos e, consequentemente, foi retirado das posteriores análises. Os demais locos variaram entre 80 e 100%, com média de 96% dos genótipos esperados.

O padrão de distribuição das freqüências alélicas na população brasileira mostrou-se com tendência de similaridade ao da população CEU. A análise de variância demonstrou que a diferenciação entre a população brasileira e as populações YRI e ASN é moderada, e com a população CEU é pequena, porém não significativa. Com isso não foi possível identificar subestrutura no gene *VDR* para população brasileira em relação a população de origem européia, indicando que a contribuição de miscigenação pode ter pouco efeito na população brasileira para o gene *VDR*.

Os SNPs selecionados para este estudo mostraram que existe diversidade de blocos entre as populações do HapMap, e que o DL é mais forte na população CEU, de origem européia, depois em asiáticos e por último na população YRI, de origem africana. Este resultado corrobora fortemente anteriores estudos no gene *VDR* (FANG et al., 2005; NEJENTSEV et al., 2004) principalmente entre populações Européias e Africanas.

Para todos marcadores avaliados, o valor médio de D' foi maior nas populações do HapMap do que na população Brasileira. Os resultados médios de r^2, no entanto, foram baixos para todas as populações. Na população brasileira, e suas amostras regionais, foram gerados baixos valores de r^2 e, quando usados os parâmetros de Gabriel (GABRIEL et al., 2002), não foi observada a formação de blocos ou os blocos eram pouco extensos. A média dos valores de D', r^2, LOD e intervalos de confiança para estes dois blocos foram superiores à média de todos os locos. Em relação às populações do HapMap, o padrão do DL na população brasileira se mostrou similar ao da população CEU, como conseqüência da diferenciação medida por Fst, e apresentou dois blocos, definidos como 5' e 3'. As amostras regionais brasileiras não mostraram os mesmos padrões da distribuição dos blocos haplotípicos, mas, contudo foram observados padrões nos limites dos blocos definidos como 5' e 3'.

O bloco da região 5' contém o polimorfismo do Cdx-2 no qual está associado a regulação e expressão do *VDR* no intestino (ARAI et al., 2001; FANG et al., 2003; YAMAMOTO et al., 1999). Este bloco apresenta o mesmo padrão na população brasileira e na população CEU. O mesmo bloco é observado quando separada a amostra da região Nordeste. Nas regiões Sul e Sudeste, o bloco 5' é menor, com extensão de 9kb, enquanto na amostra do Norte são observados dois blocos nessa região gênica, um com 7kb e outro com 2kb. Nas regiões Nordeste e Centro-Oeste ainda foi observada a formação de um bloco de 5kb que antecede a região codificadora do gene. Exceto na subpopulação do Centro-Oeste, alguns pares de marcadores nestes blocos apresentaram altos valores de D', porém baixos valores de LOD (quadrados azuis). Embora o conjunto de marcadores que forma o bloco deva ser avaliado, a atribuição de um bloco deve levar em consideração os fatores estatísticos que determinam tanto o DL quanto a extensão do bloco haplotípico. No caso da região Norte o primeiro bloco (a contar da diração 5'-3') pode ter sido atribuído a fatores adversos da estatística utilizada e não por alto DL. A utilização desse bloco pode ser útil para gerar informações em estudos funcionais que investiguem o papel desses polimorfismos (principalmente Cdx-2) na densidade mineral óssea (FANG et al., 2003; FANG et al., 2005) como também na ativação da transcrição do gene *VDR* no intestino humano (YAMAMOTO et al., 1999).

Um ponto quente de recombinação pode ser definido por uma pequena região onde ocorram *crossing-over* a uma taxa acima de dez vezes maior que suas áreas adjacentes (CRAWFORD; BHANGALE et al., 2004). A análise realizada no programa Phase indicou dois potenciais pontos de recombinação no gene. O primeiro ponto possui uma proporção entre 9 e 18 vezes (média e mediana) da taxa de recombinação basal e foi definido em uma região de 6,5 kb entre os marcadores rs3890734 e rs2853559. O segundo, possui uma proporção muito maior da taxa de recombinação, cerca de 60 vezes mais alta que a taxa de recombinação basal, e foi definido entre os marcadores rs10735810 (FokI) e rs886441 abrangendo aproximadamente 11 kb. Este ponto possui uma extensão que inclui o rs2254210 (anterior ao FokI) com cerca de 8 a 12 vezes a taxa de recombinação basal. Ainda foi identificado um terceiro potencial ponto de recombinação entre os marcadores rs4077869 e o Cdx-2, no entanto, essa estimativa pode estar enviesada devido aos dois marcadores se encontrarem fora do equilíbrio esperado segundo a lei de Hardy-Weinberg (Tabela 3.7).

Os padrões dos blocos haplotípicos na população brasileira e suas amostras regionais revelou baixo desequilíbrio de ligação nessas regiões, indicando a possibilidade de estes pontos serem potenciais pontos quentes de recombinação. As amostras regionais CO e NE apresentaram um bloco de 4 kb de extensão exatamente na depressão entre os dois picos definidos pela análise do Phase, que compreende o bloco formado pelos marcadores rs2853564 e rs2254210. Ainda, não foi observada a formação de blocos haplotípicos em nenhuma das amostras do HapMap. A definição do ponto de recombinação na região do FokI já foi observada em diversas outras populações (FANG et al., 2005; NEJENTSEV et al., 2004), e, portanto, sugere um mecanismo geral de recombinação do gene entre a região promotora e codificadora, que pode ser um dos vários fatores pleiotrópicos de atuação do *VDR* nas diversas características fenotípicas. Nesses outros estudos (FANG et al., 2005; NEJENTSEV et al., 2004), não foi relatado nenhum outro ponto de recombinação, no entanto, a região que antecede o FokI apresentou padrão de DL similar aos encontrados no presente estudo, ou seja, pequena região de baixo DL, seguida de uma região de forte DL (bloco) e uma região de total ausência de DL (FokI). Portanto, esses dados indicam que existe uma região unânime de

recombinação, entre os SNPs FokI e rs886441, e outra menos intensa, entre o rs3890734 e rs2853559.

O bloco da região 3'UTR foi o mais consistente entre as amostras regionais brasileiras e as populações CEU e ASN. Neste bloco estão presentes os polimorfismos BsmI, ApaI, e TaqI, que são amplamente empregados em estudos de associação genética (CARLING et al., 1997; FANG et al., 2003; FANG et al., 2005; FANG et al., 2007; MACDONALD et al., 2006; UITTERLINDEN et al., 2004; UITTERLINDEN et al., 1996; ZMUDA; CAULEY; FERRELL, 2000). No entanto, é grande a divergência de resultados em estudos de associação entre esses polimorfismos e características fenotípicas complexas como, por exemplo, a densidade mineral óssea, atividade física e riscos de fratura óssea (ECCLESHALL et al., 1998; MACDONALD et al., 2006; SHEN et al., 2005; SPOTILA et al., 1996).

A região 3′UTR dos genes é conhecida por estar envolvida na regulação da expressão gênica, especialmente pela regulação da estabilidade do mRNA (RACHEZ; FREEDMAN, 2000). Os polimorfismos BsmI e ApaI são SNPs que se encontram no intron anterior ao exon 9, enquanto TaqI representa uma mutação silenciosa, ou seja, uma mudança no DNA não interfere na proteína codificada. Ainda, o bloco 3' na população brasileira, assim como na população CEU, se estende até a região intergênica, corroborando resultados similares identificados em outros estudos (FANG et al., 2005; NEJENTSEV et al., 2004).

Portanto, existe a possibilidade da inconsistência em vários experimentos estar associada não só ao fato das inúmeras variáveis experimentais, mas também por outros polimorfismos na região 3′ UTR além do BsmI, ApaI, e TaqI, serem os principais candidatos funcionais a essas associações. Nejentsev e colaboradores descreveram que estes três SNPs em particular, não são informativos como tagSNPs, assim, a utilização de outros SNPs no bloco 3' teriam maior chance de captar associação com fenótipos complexos (NEJENTSEV et al., 2004). Desta forma, Fang e colaboradores encontraram maior e mais significativa associação deste bloco com risco de fratura óssea do que simplesmente com o haplótipo formado pelos polimorfismos BsmI-ApaI-TaqI (FANG et al., 2005). Portanto, é provável que, em determinados estudos, o sinal de associação seja captado pelos polimorfismos BsmI, ApaI, e TaqI em alto DL com outros SNPs na região 3'UTR,

mas o mesmo não ocorra em outros trabalhos (CARLING et al., 1997; FANG et al., 2005; MORRISON et al., 1992; OGUNKOLADE et al., 2002).

Visto que o bloco 3' é estatisticamente mais eficaz em detectar associação do que o bloco formado somente pelos polimorfismos BsmI-ApaI-TaqI (FANG et al., 2005; FANG et al., 2007), ainda seria possível utilizar esses marcadores na população brasileira em estudos de associação. Contudo, para futuras investigações epidemiológicas e funcionais do gene *VDR* seria necessário um número maior de SNPs que apresentem substancial DL baseados nos valores de D' e r^2, para que sejam aumentadas as chances de se captar associação verdadeira (FANG et al., 2005; GABRIEL et al., 2002). Nesse sentido, seria válido também a utilização de mais marcadores nas regiões de exons do gene, a fim de identificar polimorfismos funcionais.

Na população brasileira, o bloco formado na região 3' do gene mostrou-se consistente entre as amostras regionais estudadas, exceto na Centro-Oeste, o que indica que mesmo com baixo sinal de DL por r^2 esse bloco ainda é detectado pelos valores de D', LOD e intervalos de confiança relativamente altos. Ainda assim, os polimorfismos BsmI, ApaI, e TaqI estão em DL em todas as amostras regionais brasileiras (incluindo Centro-Oeste). O desequilíbrio de ligação na região definida como bloco haplotípico 3' na amostra regional do Centro-Oeste mostrou-se bastante diferente das demais amostras regionais. No entanto, é observado que os limites dessa região do bloco 3', definido do loco rs2248098 ao rs7968585, possui maiores valores de D' do que os locos que se situam fora desta região, mas, no entanto esses valores não são suficientes para designar um bloco de alto DL pelo método utilizado. No entanto, esse efeito pode ter sido observado devido a forças evolutivas como recombinação e deriva genética nesta amostra regional. Por outro lado, a diversidade das metodologias utilizadas para descrever e identificar blocos haplotípicos não chegaram a um consenso formal de quais critérios devem ser adotados para uma boa descrição da estrutura de blocos haplotípicos (DALY et al., 2001; DAWSON et al., 2002; GABRIEL et al., 2002; KIMMEL; SHAMIR, 2005; LIU, N. et al., 2004; PATIL et al., 2001; PHILLIPS et al., 2003; WALL; PRITCHARD, 2003; WANG, N. et al., 2002; ZHANG et al., 2002). Portanto, outras definições estatísticas poderiam evidenciar a formação do bloco 3' na amostra da

região Centro-Oeste, assim como reestruturar os outros blocos nas demais amostras regionais e na população brasileira.

Os testes de correlação condicional entre ancestralidade e os haplótipos puderam ser realizados com os dados de contribuição ancestral e os dados genotípicos do gene *VDR*. Nesses testes, não foi observado nenhum efeito de estratificação populacional na correlação de ancestralidade com haplótipos. A análise de variância baseada nas freqüências alélicas revelou que não existe diferença significativa entre e dentro das amostras regionais em relação à população brasileira total e nem entre a população brasileira como um todo e a população do HapMap de origem Européia, enquanto a diferença entre as outras foi moderada. Como o padrão de blocos haplotípicos e DL nas populações do HapMap se mostrou diferente entre essas populações. Na amostra da população brasileira utilizada neste estudo, as freqüências dos cinco polimorfismos mais estudados possuem o mesmo padrão de outros estudos realizados no Brasil (GENTIL, 2006; HAUACHE et al., 1998; LAZARETTI-CASTRO et al., 1997; LIMA, 2006; MAISTRO et al., 2004; RAMALHO et al., 1998). Assim, o presente estudo, além de apoiar o padrão achado para os polimorfismos descritos na população brasileira, fornece informação necessária para o aprimoramento dessas investigações nos grupos de pesquisas brasileiros.

A transferibilidade de tagSNPs no gene *VDR* mostrou que a população brasileira perde pouca variabilidade com os marcadores selecionados para a população CEU, por outro lado, a perda de variabilidade é maior quando são escolhidos tagSNPs das populações ASN e YRI. De certa forma, isso é um reflexo do que pode ser observado pelo padrão de distribuição de freqüências alélicas, pelos valores de Fst par a par e pelos padrões de DL. No entanto, a perda de variabilidade genética pode ser substancialmente afetada pela diferença de apenas um marcador, o que torna a aplicabilidade do uso de tagSNPs a partir das populações do HapMap questionável. Por outro lado, a variabilidade das populações do HapMap utilizando os 21 marcadores contra aproximadamente 100 disponíveis na fase II mostrou-se alta, em cerca de 80%. Como não existem ferramentas para predizer essa perda em outras populações, e pelos resultados de transferibilidade

apresentados, sugere-se que, para a população brasileira, devam ser usados o maior número de marcadores a fim de captar a maior variabilidade possível.

Para se ter uma idéia, para captar 80% da variabilidade usando os 21 marcadores são necessários dois marcadores na população ASN, quatro em CEU, sete em YRI e 11 marcadores na população brasileira. Se essa proporção fosse linear, para captar 80% da variabilidade de todos os 100 marcadores disponíveis na fase II do HapMap na população brasileira seriam necessários aproximadamente 50 marcadores, ou seja, mais que o dobro utilizado nas populações do HapMap. No entanto, essa condição é totalmente dependente do conjunto de marcadores escolhidos aliado ao DL entre eles, o que determina a variabilidade. Apesar disso, a escolha de marcadores espaçados a distancias médias mostrou-se uma estratégia eficaz para determinar o padrão de DL, contudo, no caso específico do gene *VDR*, a seleção feita a partir de tagSNPs da população de origem européia providenciou baixa perda relativa de variabilidade na população brasileira.

Nejentsev e colaboradores demonstraram que existe pouca perda de informação na utilização de tagSNPS escolhidos para a população caucasiana de origem Britânica em populações Européias de outros países, como Norueguesas, Finlandesas e Romenas (NEJENTSEV et al., 2004). Embora o padrão de DL nessas populações foi bastante similar, poucas diferenças foram notadas devido à forças evolutivas e histórico demográfico de cada população. Contudo, houve pouca perda de informação entre essas populações (NEJENTSEV et al., 2004), assim como ocorreu entre a população brasileira e a população CEU. Deste modo, em futuros estudos de associação do gene *VDR* na população brasileira, nos quais desejam-se aumentar a densidade de marcadores nas regiões de blocos haplotípicos, a escolha desses marcadores poderá ser feita mediante tagSNPs da população CEU. Por outro lado, devido à heterogeneidade da população brasileira, o controle de estratificação populacional deve sempre ser utilizado a fim de se evitar desvios nos estudos de associação genética, como falhas em detectar associação ou associações espúrias. Além disso, estudos em populações mais estratificadas, como Afro-Brasileiros e populações Indígenas, são necessários para corroborar ou refutar os resultados aqui apresentados.

Portanto, no presente trabalho foi apresentada uma primeira abordagem dos padrões de DL no gene *VDR* e da utilização de tagSNPs de populações de diferentes grupos étnicos para uma amostra da população brasileira, e foi demonstrado que polimorfismos na região promotora 5' e na região 3'UTR constituem blocos haplotípicos na população brasileira. Experimentos funcionais com os haplótipos da região promotora 5' e da região 3'UTR fornecem possíveis explicações moleculares para associação do gene *VDR* com fenótipos complexos como risco de fratura óssea, osteoporose, diabetes, e câncer em diversos estudos (FANG et al., 2005; INGLES; HAILE et al., 1997; NEJENTSEV et al., 2004; ZMUDA; CAULEY; FERRELL, 2000). Portanto, os resultados deste trabalho providenciam relevância para futuras investigações de fenótipos complexos com o gene *VDR* na população brasileira.

CONCLUSÃO

O presente estudo utilizou marcadores genéticos do tipo SNPs com duas finalidades: identificar estruturação genética e investigar os padrões de desequilíbrio de ligação nos genes *PTPN22* e *VDR*, com intuito de providenciar informações necessárias para aplicação em estudos de associação genética envolvendo estes polimorfismos e fenótipos complexos relacionados na população brasileira.

Os resultados obtidos na genotipagem dos marcadores de ancestralidade mostraram que a informação sobre ancestralidade gerada é totalmente dependente do conteúdo de informação e taxa de atribuição correta do marcador e da aplicação conjunta de vários marcadores em uma população miscigenada. Os marcadores avaliados nesse estudo se mostraram eficientes para atribuição de ancestralidade em uma população bi-parental, composta por parentais Europeus e Africanos. Para atender o propósito de ter uma bateria que possa ser aplicada em uma população tri-parental, polimorfismos com predominância em populações indígenas estão sendo selecionados em estudos paralelos a fim de complementar e aprimorar as estimativas de ancestralidade nas amostras da população brasileira.

Para o gene *PTPN22*, os dados de genotipagem e estimativa de freqüências haplotípicas revelaram grande diversidade genética da população brasileira. O loco rs2476601 é relatado na literatura como um importante polimorfismo causal associado a diversas doenças de caráter auto-imune em populações Européias ou com populações miscigenadas com contribuição ancestral Européia. Uma vez que este polimorfismo não é relatado em populações de origem Africana e Asiática, seria interessante de se realizar estudos com a genotipagem desse loco em populações miscigenadas com predominância Africana, do tipo Afro-Brasileira ou Afro-Americana, além de populações indígenas ou ameríndias, para constatar se a ausência do alelo T é verificada nessas populações e poder confrontar os resultados na população miscigenada com predominância Européia.

Ainda no gene *PTPN22*, freqüência dos haplótipos na população brasileira se mostrou diferenciada das populações do HapMap, com tendência para os haplótipos mais freqüentes de todas as três populações étnicas estarem presentes na população brasileira de forma diferenciada da população de origem européia, a qual se mostrou mais próxima geneticamente da população brasileira. A presença de haplótipos raros exclusivos das populações ASN e YRI, e do haplótipo exclusivo da população CEU, carregador do alelo associado a doenças auto-imunes (alelo T do rs2476601) também deu indícios da alta heterogeneidade genética. Os testes de correlação condicional utilizando os dados dos marcadores de ancestralidade indicaram que os haplótipos possuem correlação com os níveis de ancestralidade da população. A investigação para efeitos de seleção mostrou que o gene possui indícios de seleção por efeito carona em varredura seletiva.

No gene VDR a heterogeneidade se mostrou menor que no gene *PTPN22*, no entanto foram observados padrões de distribuição dos blocos haplotípicos similares aos descritos para populações de origem européia. Observou-se ainda, que as amostras regionais apresentam padrões de desequilíbrio de ligação semelhantes, mas que os mesmos podem ser falsamente atribuídos se não forem observadas todas as informações estatísticas com cautela. Os blocos identificados nesse estudo podem servir de base para futuros estudos funcionais ou de associação genética na população brasileira envolvendo o gene *VDR*.

A transferibilidade de SNPs das populações do HapMap para a população brasileira se mostrou diferenciada entre os genes estudados. No gene *PTPN22* houve perda de variabilidade quando usados marcadores selecionados para a população CEU, de origem Européia. Por outro lado, no gene *VDR*, a perda foi maior quando foram usados os marcadores das populações YRI e ASN. Esses resultados indicam que a seleção de SNPs para estudos de associação genética na população brasileira não deve ser generalizada para os tagSNPs das populações do HapMap e, portanto, devem ser realizados estudos com o uso conjunto de várias ferramentas estatísticas a fim de se escolher um conjunto representativo que minimize a perda de variabilidade. Ferramentas estatísticas capazes de inferir um número ótimo de SNPs para múltiplas populações a partir de seus tagSNPs estão em pleno desenvolvimento (DE BAKKER et al., 2006; HALPERIN; KIMMEL;

SHAMIR, 2005; HOWIE et al., 2006), e serão necessárias para futuros estudos de associação por DL na população brasileira.

Portanto, o presente trabalho possui informação de cunho populacional para estudos de contribuição de ancestralidade genômica em populações miscigenadas e para futuros estudos de associação genética da população brasileira nos genes *PTPN22* e *VDR*, notavelmente associados com doenças auto-imunes e com fenótipos complexos das vias de metabolismo ósseo, respectivamente.

REFERÊNCIAS

ABE-SANDES, K. et al. Heterogeneity of the Y chromosome in Afro-Brazilian populations. **Hum Biol**, v.76, n.1, p.77-86, Feb, 2004.

ABREU, B. S. **Miscigenação e estudos de associação na população brasileira:** Portabilidade do HapMap nos genes CETP e HMGCR e correlação entre ancestralidade biogeográfica e perfis lipídicos. Dissertação (Mestrado em Ciências Genômicas e Biotecnologia). Programa de Pós-Graduação *Strictu Senso* em Ciências Genômicas e Biotecnologia, Universidade Católica de Brasília, Brasília, 2007.

ADAMSKY, K. et al. Junctional protein MAGI-3 interacts with receptor tyrosine phosphatase {beta} (RPTP{beta}) and tyrosine-phosphorylated proteins. **J Cell Sci**, v.116, n.7, p.1279-1289, April 1, 2003, 2003.

ALLEBRANDT, K. V. et al. Variability of the paraoxonase gene (PON1) in Euro- and Afro-Brazilians. **Toxicol Appl Pharmacol**, v.180, n.3, p.151-6, May 1, 2002.

ALTSHULER, D. et al. A haplotype map of the human genome. **Nature**, v.437, n.7063, p.1299-320, Oct 27, 2005.

ALVES-SILVA, J. et al. The ancestry of Brazilian mtDNA lineages. **Am J Hum Genet**, v.67, n.2, p.444-61, Aug, 2000.

ARAI, H. et al. A vitamin D receptor gene polymorphism in the translation initiation codon: effect on protein activity and relation to bone mineral density in Japanese women. **J Bone Miner Res**, v.12, n.6, p.915-21, Jun, 1997.

ARAI, H. et al. The polymorphism in the caudal-related homeodomain protein Cdx-2 binding element in the human vitamin D receptor gene. **J Bone Miner Res**, v.16, n.7, p.1256-64, Jul, 2001.

BACA, V. et al. Association analysis of the PTPN22 gene in childhood-onset systemic lupus erythematosus in Mexican population. **Genes Immun**, v.7, n.8, p.693-5, Dec, 2006.

BADANO, J. L.; KATSANIS, N. Beyond Mendel: an evolving view of human genetic disease transmission. **Nat Rev Genet**, v.3, n.10, p.779-89, Oct, 2002.

BADZIOCH, M. D. et al. Summary report: Missing data and pedigree and genotyping errors. **Genet Epidemiol**, v.25 Suppl 1, p.S36-42, 2003.

BAKER, A. R. et al. Cloning and expression of full-length cDNA encoding human vitamin D receptor. **Proc Natl Acad Sci U S A**, v.85, n.10, p.3294-8, May, 1988.

BARNHOLTZ-SLOAN, J. S. et al. Examining population stratification via individual ancestry estimates versus self-reported race. **Cancer Epidemiol Biomarkers Prev**, v.14, n.6, p.1545-51, Jun, 2005.

BARRETT-CONNOR, E. et al. Osteoporosis and fracture risk in women of different ethnic groups. **J Bone Miner Res**, v.20, n.2, p.185-94, Feb, 2005.

BARRETT, J. C. et al. Haploview: analysis and visualization of LD and haplotype maps. **Bioinformatics**, v.21, n.2, p.263-5, Jan 15, 2005.

BASTOS-RODRIGUES, L. et al. The genetic structure of human populations studied through short insertion-deletion polymorphisms. **Ann Hum Genet**, v.70, n.Pt 5, p.658-65, Sep, 2006.

BEGOVICH, A. B. et al. A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (PTPN22) is associated with rheumatoid arthritis. **Am J Hum Genet**, v.75, n.2, p.330-7, Aug, 2004.

BELL, N. H. et al. ApaI polymorphisms of the vitamin D receptor predict bone density of the lumbar spine and not racial difference in bone density in young men. **J Lab Clin Med**, v.137, n.2, p.133-40, Feb, 2001.

BOLDT, A. B. et al. Diversity of the MBL2 gene in various Brazilian populations and the case of selection at the mannose-binding lectin locus. **Hum Immunol**, v.67, n.9, p.722-34, Sep, 2006.

BONILLA, C. et al. Ancestral proportions and their association with skin pigmentation and bone mineral density in Puerto Rican women from New York city. **Hum Genet**, v.115, n.1, p.57-68, Jun, 2004.

BOTSTEIN, D.; RISCH, N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. **Nat Genet**, v.33 Suppl, p.228-37, Mar, 2003.

BOTTINI, N. et al. A functional variant of lymphoid tyrosine phosphatase is associated with type I diabetes. **Nat Genet**, v.36, n.4, p.337-8, Apr, 2004.

BROWN, T. **Genomes**. Manchester: BIOS Scientific. 2002.

BUCHANAN, A. V. et al. Dissecting complex disease: the quest for the Philosopher's Stone? **Int J Epidemiol**, v.35, n.3, p.562-71, Jun, 2006.

BURTON, P. R. et al. Key concepts in genetic epidemiology. **The Lancet**, v.366, n.9489, p.941-951, 2005.

CALLEGARI-JACQUES, S. M. et al. Historical genetics: spatiotemporal analysis of the formation of the Brazilian population. **Am J Hum Biol**, v.15, n.6, p.824-34, Nov-Dec, 2003.

CARDON, L. R.; ABECASIS, G. R. Using haplotype blocks to map human complex trait loci. **Trends Genet**, v.19, n.3, p.135-40, Mar, 2003.

CARDON, L. R.; BELL, J. I. Association study designs for complex diseases. **Nat Rev Genet**, v.2, n.2, p.91-9, Feb, 2001.

CARLING, T. et al. Vitamin D receptor alleles b, a, and T: risk factors for sporadic primary hyperparathyroidism (HPT) but not HPT of uremia or MEN 1. **Biochem Biophys Res Commun**, v.231, n.2, p.329-32, Feb 13, 1997.

CARLTON, V. E. et al. PTPN22 genetic variation: evidence for multiple variants associated with rheumatoid arthritis. **Am J Hum Genet**, v.77, n.4, p.567-81, Oct, 2005.

CARVALHO-SILVA, D. R. et al. The phylogeography of Brazilian Y-chromosome lineages. **Am J Hum Genet**, v.68, n.1, p.281-6, Jan, 2001.

CHAKRABORTY, R. et al. Caucasian genes in American blacks: new data. **Am J Hum Genet**, v.50, n.1, p.145-55, Jan, 1992.

CHOUDHRY, S. et al. Population stratification confounds genetic association studies among Latinos. **Hum Genet**, v.118, n.5, p.652-64, Jan, 2006.

CLOUTIER, J. F.; VEILLETTE, A. Association of inhibitory tyrosine protein kinase p50csk with protein tyrosine phosphatase PEP in T cells and other hemopoietic cells. **Embo J**, v.15, n.18, p.4909-18, Sep 16, 1996.

COLLINS, F. S. et al. A vision for the future of genomics research. **Nature**, v.422, n.6934, p.835-847, 2003.

CRAWFORD, D. C. et al. Evidence for substantial fine-scale variation in recombination rates across the human genome. **Nat Genet**, v.36, n.7, p.700-6, Jul, 2004.

CRAWFORD, D. C. et al. Haplotype diversity across 100 candidate genes for inflammation, lipid metabolism, and blood pressure regulation in two populations. **Am J Hum Genet**, v.74, n.4, p.610-22, Apr, 2004.

CROFTS, L. A. et al. Multiple promoters direct the tissue-specific expression of novel N-terminal variant human vitamin D receptor gene transcripts. **Proc Natl Acad Sci U S A**, v.95, n.18, p.10529-34, Sep 1, 1998.

DALY, M. J. et al. High-resolution haplotype structure in the human genome. **Nat Genet**, v.29, n.2, p.229-32, Oct, 2001.

DAVIDOVICH, O. et al. GEVALT: an integrated software tool for genotype analysis. **BMC Bioinformatics**, v.8, p.36, 2007.

DAWN TEARE, M.; BARRETT, J. H. Genetic linkage studies. **Lancet**, v.366, n.9490, p.1036-44, Sep 17-23, 2005.

DAWSON, E. et al. A first-generation linkage disequilibrium map of human chromosome 22. **Nature**, v.418, n.6897, p.544-8, Aug 1, 2002.

DE BAKKER, P. I. et al. Transferability of tag SNPs in genetic association studies in multiple populations. **Nat Genet**, v.38, n.11, p.1298-303, Nov, 2006.

DE BOER, R. A. et al. Genetics in heart failure: where are we headed? **Congest Heart Fail**, v.12, n.6, p.329-32, Nov-Dec, 2006.

DEVLIN, B.; RISCH, N. A comparison of linkage disequilibrium measures for fine-scale mapping. **Genomics**, v.29, n.2, p.311-22, Sep 20, 1995.

DON, R. H. et al. 'Touchdown' PCR to circumvent spurious priming during gene amplification. **Nucleic Acids Res**, v.19, n.14, p.4008, Jul 25, 1991.

ECCLESHALL, T. R. et al. Lack of correlation between start codon polymorphism of the vitamin D receptor gene and bone mineral density in premenopausal French women: the OFELY study. **J Bone Miner Res**, v.13, n.1, p.31-5, Jan, 1998.

ERHART, M. A. et al. Haplotypes that are mosaic for wild-type and t complex-specific alleles in wild mice. **Genetics**, v.123, n.2, p.405-15, Oct, 1989.

EVANNO, G. et al. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. **Mol Ecol**, v.14, n.8, p.2611-20, Jul, 2005.

FANG, Y. et al. Cdx-2 polymorphism in the promoter region of the human vitamin D receptor gene determines susceptibility to fracture in the elderly. **J Bone Miner Res**, v.18, n.9, p.1632-41, Sep, 2003.

FANG, Y. et al. Promoter and 3'-untranslated-region haplotypes in the vitamin d receptor gene predispose to osteoporotic fracture: the rotterdam study. **Am J Hum Genet**, v.77, n.5, p.807-23, Nov, 2005.

FANG, Y. et al. Vitamin D receptor gene haplotype is associated with body height and bone size. **J Clin Endocrinol Metab**, Jan 9, 2007.

FARACO, J. H. et al. ApaI dimorphism at the human vitamin D receptor gene locus. **Nucleic Acids Res**, v.17, n.5, p.2150, Mar 11, 1989.

FAUCZ, F. R. et al. Polymorphism of LMP2, TAP1, LMP7 and TAP2 in Brazilian Amerindians and Caucasoids: implications for the evolution of allelic and haplotypic diversity. **Eur J Immunogenet**, v.27, n.1, p.5-16, Feb, 2000.

FERREIRA, F. L. et al. Genetic characterization of the population of São Luís, MA, Brazil. **Genet Mol Biol**, v.28, n.01, p.pp. 22-31, 2005.

GABRIEL, S. B. et al. The structure of haplotype blocks in the human genome. **Science**, v.296, n.5576, p.2225-9, Jun 21, 2002.

GENTIL, P. **Interação entre polimorfismos do gene VDR e padrões de atividade física na determinação da DMO de mulheres brasileiras no período pós-**

menopausa. Dissertação (Mestrado em Educação Física). Programa de Pós-Graduação *Strictu Senso* em Educação Física, Universidade Católica de Brasília, Brasília, 2006.

GOMEZ, L. M. et al. PTPN22 C1858T polymorphism in Colombian patients with autoimmune diseases. **Genes Immun**, v.6, n.7, p.628-31, Oct, 2005.

GONG, G.; HAYNATZKI, G. Association between bone mineral density and candidate genes in different ethnic populations and its implications. **Calcif Tissue Int**, v.72, n.2, p.113-23, Feb, 2003.

GONG, G. et al. Bone mineral density-affecting genes in Africans. **J Natl Med Assoc**, v.98, n.7, p.1102-8, Jul, 2006.

GONZALEZ-NEIRA, A. et al. The portability of tagSNPs across populations: A worldwide survey. **Genome Res.**, v.16, n.3, p.323-330, March 1, 2006, 2006.

GREGERSEN, P. K. et al. PTPN22: setting thresholds for autoimmunity. **Semin Immunol**, v.18, n.4, p.214-23, Aug, 2006.

GRIMES, D. S. Are statins analogues of vitamin D? **Lancet**, v.368, n.9529, p.83-6, Jul 1, 2006.

GROSS, C. et al. The presence of a polymorphism at the translation initiation site of the vitamin D receptor gene is associated with low bone mineral density in postmenopausal Mexican-American women. **J Bone Miner Res**, v.11, n.12, p.1850-5, Dec, 1996.

GU, S. et al. Significant variation in haplotype block structure but conservation in tagSNP patterns among global populations. **Eur J Hum Genet**, v.15, n.3, p.302-312, 2007.

GUSELLA, J. F. et al. A polymorphic DNA marker genetically linked to Huntington's disease. **Nature**, v.306, n.5940, p.234-238, 1983.

HALPERIN, E. et al. Tag SNP selection in genotype data for maximizing SNP prediction accuracy. **Bioinformatics**, v.21 Suppl 1, p.i195-203, Jun, 2005.

HAMOSH, A. et al. Online Mendelian Inheritance in Man (OMIM). **Hum Mutat**, v.15, n.1, p.57-61, 2000.

HARDY, G. H. Mendelian proportions in a mixed population. **Science**, v.28, p.49-50, 1908.

HARTL, D. L.; CLARK, A. G. **Principles of Population Genetics**. Sunderland, Massachusetts: Sinauer Associates, Inc. 1997

HAUACHE, O. M. et al. Vitamin D Receptor Gene Polymorphism: Correlation with Bone Mineral Density in a Brazilian Population with Insulin-Dependent Diabetes Mellitus. **Osteoporos Int**, v.8, n.3, p.204-210, 1998.

HILL, W. Estimation of linkage disequilibrium in randomly mating populations. **Heredity**, v.33, n.2, p.229-39, 1974.

HINRICHS, A. L.; SUAREZ, B. K. Genotyping errors, pedigree errors, and missing data. **Genet Epidemiol**, v.29 Suppl 1, p.S120-4, 2005.

HOGGART, C. J. et al. Control of confounding of genetic associations in stratified populations. **Am J Hum Genet**, v.72, n.6, p.1492-1504, Jun, 2003.

HOWIE, B. N. et al. Efficient selection of tagging single-nucleotide polymorphisms in multiple populations. **Hum Genet**, v.120, n.1, p.58-68, Aug, 2006.

IBGE. Estudos e Pesquisas: Informação Demográfica e Socioeconômica. **Síntese de Indicadores Sociais 2006**. Rio de Janeiro. 2006

IKARI, K. et al. Haplotype analysis revealed no association between the PTPN22 gene and RA in a Japanese population. **Rheumatology (Oxford)**, v.45, n.11, p.1345-8, May 11, 2006.

INGLES, S. A. et al. Strength of linkage disequilibrium between two vitamin D receptor markers in five ethnic groups: implications for association studies. **Cancer Epidemiol Biomarkers Prev**, v.6, n.2, p.93-8, Feb, 1997.

INGLES, S. A. et al. Association of prostate cancer risk with genetic polymorphisms in vitamin D receptor and androgen receptor. **J Natl Cancer Inst**, v.89, n.2, p.166-70, Jan 15, 1997.

IOANNIDIS, J. P. Genetic associations: false or true? **Trends Mol Med**, v.9, n.4, p.135-8, Apr, 2003.

JARVINEN, T. L. et al. Vitamin D receptor alleles and bone's response to physical activity. **Calcif Tissue Int**, v.62, n.5, p.413-7, May, 1998.

JONES, G. et al. Current understanding of the molecular actions of vitamin D. **Physiol Rev**, v.78, n.4, p.1193-231, Oct, 1998.

JORDE, L. B. Linkage disequilibrium and the search for complex disease genes. **Genome Res**, v.10, n.10, p.1435-44, Oct, 2000.

KAUFMAN, K. M. et al. Evaluation of the genetic association of the PTPN22 R620W polymorphism in familial and sporadic systemic lupus erythematosus. **Arthritis Rheum**, v.54, n.8, p.2533-40, Aug, 2006.

KAWASAKI, E. et al. Systematic search for single nucleotide polymorphisms in a lymphoid tyrosine phosphatase gene (PTPN22): association between a promoter polymorphism and type 1 diabetes in Asian populations. **Am J Med Genet A**, v.140, n.6, p.586-93, Mar 15, 2006.

KENT, W. J. et al. The human genome browser at UCSC. **Genome Res**, v.12, n.6, p.996-1006, Jun, 2002.

KIM, Y.; NIELSEN, R. Linkage Disequilibrium as a Signature of Selective Sweeps. **Genetics**, v.167, n.3, p.1513-1524, July 1, 2004, 2004.

KIMMEL, G.; SHAMIR, R. GERBIL: Genotype resolution and block identification using likelihood. **Proc Natl Acad Sci U S A**, v.102, n.1, p.158-62, Jan 4, 2005.

KIRCHHEINER, J.; BROCKMOLLER, J. Clinical consequences of cytochrome P450 2C9 polymorphisms. **Clin Pharmacol Ther**, v.77, n.1, p.1-16, Jan, 2005.

KITAGAWA, I. et al. Interplay of physical activity and vitamin D receptor gene polymorphism on bone mineral density. **J Epidemiol**, v.11, n.5, p.229-32, Sep, 2001.

KNOWLER, W. C. et al. Gm3;5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. **Am J Hum Genet**, v.43, n.4, p.520-6, Oct, 1988.

KRUGLYAK, L. The use of a genetic map of biallelic markers in linkage studies. **Nat Genet**, v.17, n.1, p.21-4, Sep, 1997.

KRUGLYAK, L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. **Nat Genet**, v.22, n.2, p.139-44, Jun, 1999.

LADNER, M. B. et al. Association of the single nucleotide polymorphism C1858T of the PTPN22 gene with type 1 diabetes. **Hum Immunol**, v.66, n.1, p.60-4, Jan, 2005.

LANDER, E. S. et al. Initial sequencing and analysis of the human genome. **Nature**, v.409, n.6822, p.860-921, Feb 15, 2001.

LANDER, E. S.; SCHORK, N. J. Genetic dissection of complex traits. **Science**, v.265, n.5181, p.2037-48, Sep 30, 1994.

LAZARETTI-CASTRO, M. et al. Vitamin D receptor alleles and bone mineral density in a normal premenopausal Brazilian female population. **Braz J Med Biol Res**, v.30, p.929-932, 1997.

LEI, S. F. et al. Ethnic difference in osteoporosis-related phenotypes and its potential underlying genetic determination. **J Musculoskelet Neuronal Interact**, v.6, n.1, p.36-46, Jan-Mar, 2006.

LETOVSKY, S. I. et al. GDB: the Human Genome Database. **Nucleic Acids Res**, v.26, n.1, p.94-9, Jan 1, 1998.

LEWIS, P. O.; ZAYKIN, D. **Genetic Data Analysis: Programa de computador para análises de dados genéticos**: Programa de livre distribuição pelos autores na internet. Disponível em: <http://lewis.eeb.uconn.edu/lewishome/software.html>, 2001.

LEWONTIN, R. C. The Interaction of Selection and Linkage. Ii. Optimum Models. **Genetics**, v.50, p.757-82, Oct, 1964.

LEWONTIN, R. C. The detection of linkage disequilibrium in molecular sequence data. **Genetics**, v.140, n.1, p.377-88, May, 1995.

LIMA, R. M. **Estudo de associação entre polimorfismos no gene receptor de vitamina D e massa livre de gordura em brasileiras pós-menopausadas**. Dissertação (Mestrado em Educação Física). Programa de Pós-Graduação *Strictu Senso* em Educação Física, Universidade Católica de Brasília, Brasília, 2006.

LIU, K.; MUSE, S. V. PowerMarker: an integrated analysis environment for genetic marker analysis. **Bioinformatics**, v.21, n.9, p.2128-9, May 1, 2005.

LIU, N. et al. Comparison of single-nucleotide polymorphisms and microsatellites in inference of population structure. **BMC Genet**, v.6 Suppl 1, p.S26, Dec 30, 2005.

LIU, N. et al. Haplotype block structures show significant variation among populations. **Genet Epidemiol**, v.27, n.4, p.385-400, Dec, 2004.

LOHMUELLER, K. E. et al. Variants associated with common disease are not unusually differentiated in frequency across populations. **Am J Hum Genet**, v.78, n.1, p.130-6, Jan, 2006.

LOPEZ, J.; PREZIOSO, V. A better way to optimize: Two-step Gradient PCR. **Eppendorf BioNews**, n.16, 2001.

MACDONALD, H. M. et al. Large-Scale Population-Based Study Shows No Evidence of Association Between Common Polymorphism of the VDR Gene and BMD in British Women. **J Bone Miner Res**, v.21, n.1, p.151-162, 2006.

MAISTRO, S. et al. Vitamin D receptor polymorphisms and prostate cancer risk in Brazilian men. **Int J Biol Markers**, v.19, n.3, p.245-9, Jul-Sep, 2004.

MANIATIS, N. et al. The optimal measure of linkage disequilibrium reduces error in association mapping of affection status. **Hum Mol Genet**, v.14, n.1, p.145-53, Jan 1, 2005.

MARICIC, M. Ethnic variation in bone mineral density: on the road to fracture prevention. **Menopause**, v.12, n.5, p.492-4, Sep-Oct, 2005.

MARRERO, A. R. et al. Heterogeneity of the genome ancestry of individuals classified as White in the state of Rio Grande do Sul, Brazil. **Am J Hum Biol**, v.17, n.4, p.496-506, Jul-Aug, 2005.

MATHIEU, C.; BADENHOOP, K. Vitamin D and type 1 diabetes mellitus: state of the art. **Trends Endocrinol Metab**, v.16, n.6, p.261-6, Aug, 2005.

MCVEAN, G. The structure of linkage disequilibrium around a selective sweep. **Genetics**, p.genetics.106.062828, December 28, 2006, 2006.

MILLER, S. A. et al. A simple salting out procedure for extracting DNA from human nucleated cells. **Nucleic Acids Res**, v.16, n.3, p.1215, Feb 11, 1988.

MIYAMOTO, K. et al. Structural organization of the human vitamin D receptor chromosomal gene and its promoter. **Mol Endocrinol**, v.11, n.8, p.1165-79, Jul, 1997.

MORAES, M. O. et al. Interleukin-10 promoter haplotypes are differently distributed in the Brazilian versus the Dutch population. **Immunogenetics**, v.54, n.12, p.896-9, Mar, 2003.

MORI, M. et al. Ethnic differences in allele frequency of autoimmune-disease-associated SNPs. **J Hum Genet**, v.50, n.5, p.264-6, 2005.

MORRISON, N. A. et al. Prediction of bone density from vitamin D receptor alleles. **Nature**, v.367, n.6460, p.284-7, Jan 20, 1994.

MORRISON, N. A. et al. Contribution of trans-acting factor alleles to normal physiological variability: vitamin D receptor gene polymorphism and circulating osteocalcin. **Proc Natl Acad Sci U S A**, v.89, n.15, p.6665-9, Aug 1, 1992.

MOTOHASHI, Y. et al. Vitamin D receptor gene polymorphism affects onset pattern of type 1 diabetes. **J Clin Endocrinol Metab**, v.88, n.7, p.3137-40, Jul, 2003.

MUELLER, J. C. et al. Linkage disequilibrium patterns and tagSNP transferability among European populations. **Am J Hum Genet**, v.76, n.3, p.387-98, Mar, 2005.

NEJENTSEV, S. et al. Comparative high-resolution analysis of linkage disequilibrium and tag single nucleotide polymorphisms between populations in the vitamin D receptor gene. **Hum Mol Genet**, v.13, n.15, p.1633-1639, August 1, 2004, 2004.

OGUNKOLADE, B. W. et al. Vitamin D receptor (VDR) mRNA and VDR protein levels in relation to vitamin D status, insulin secretory capacity, and VDR genotype in Bangladeshi Asians. **Diabetes**, v.51, n.7, p.2294-300, Jul, 2002.

ONENGUT-GUMUSCU, S. et al. A Haplotype-Based Analysis of the PTPN22 Locus in Type 1 Diabetes. **Diabetes**, v.55, n.10, p.2883-9, Oct, 2006.

OROZCO, G. et al. Association of a functional single-nucleotide polymorphism of PTPN22, encoding lymphoid protein phosphatase, with rheumatoid arthritis and systemic lupus erythematosus. **Arthritis Rheum**, v.52, n.1, p.219-24, Jan, 2005.

PARRA, E. J. et al. Ancestral proportions and admixture dynamics in geographically defined African Americans living in South Carolina. **Am J Phys Anthropol**, v.114, n.1, p.18-29, Jan, 2001.

PARRA, E. J. et al. Estimating African American admixture proportions by use of population-specific alleles. **Am J Hum Genet**, v.63, n.6, p.1839-51, Dec, 1998.

PARRA, F. C. et al. Color and genomic ancestry in Brazilians. **Proc Natl Acad Sci U S A**, v.100, n.1, p.177-82, Jan 7, 2003.

PATIL, N. et al. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. **Science**, v.294, n.5547, p.1719-23, Nov 23, 2001.

PEAKALL, R. O. D.; SMOUSE, P. E. genalex 6: genetic analysis in Excel. Population genetic software for teaching and research. **Mol Ecol Notes**, v.6, n.1, p.288-295, 2006.

PEDROSA, M. A. F. **Composição genética de quatros populações remanescentes de quilombos do Brasil com base em microssatélites e marcadores de ancestralidade**. Dissertação (Mestrado em Biologia Molecular). Departamento de Biologia Celular, Universidade de Brasília, Brasília, 2006.

PELTONEN, L.; MCKUSICK, V. A. Genomics and medicine. Dissecting human disease in the postgenomic era. **Science**, v.291, n.5507, p.1224-9, Feb 16, 2001.

PFAFF, C. L. et al. Information on ancestry from genetic markers. **Genet Epidemiol**, v.26, n.4, p.305-15, May, 2004.

PFAFF, C. L. et al. Adjusting for population structure in admixed populations. **Genet Epidemiol**, v.22, n.2, p.196-201, Feb, 2002.

PFAFF, C. L. et al. Population structure in admixed populations: effect of admixture dynamics on the pattern of linkage disequilibrium. **Am J Hum Genet**, v.68, n.1, p.198-207, Jan, 2001.

PHILLIPS, M. S. et al. Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. **Nat Genet**, v.33, n.3, p.382-7, Mar, 2003.

PIMENTA, J. R. et al. Color and genomic ancestry in Brazilians: a study with forensic microsatellites. **Hum Hered**, v.62, n.4, p.190-5, 2006.

POCIOT, F.; MCDERMOTT, M. F. Genetics of type 1 diabetes mellitus. **Genes Immun**, v.3, n.5, p.235-49, Aug, 2002.

PRITCHARD, J. K.; PRZEWORSKI, M. Linkage disequilibrium in humans: models and data. **Am J Hum Genet**, v.69, n.1, p.1-14, Jul, 2001.

PRITCHARD, J. K.; ROSENBERG, N. A. Use of unlinked genetic markers to detect population stratification in association studies. **Am J Hum Genet**, v.65, n.1, p.220-8, Jul, 1999.

PRITCHARD, J. K. et al. Inference of population structure using multilocus genotype data. **Genetics**, v.155, n.2, p.945-59, Jun, 2000.

PRITCHARD, J. K.; WEN, W. **Documentation for structure software: Version 2**. 2003. Disponível em: <http://pritch.bsd.uchicago.edu/software/readme_2_1/readme.html>. Acesso em Novembro de 2006.

PROBST, C. M. et al. HLA polymorphism and evaluation of European, African, and Amerindian contribution to the white and mulatto populations from Parana, Brazil. **Hum Biol**, v.72, n.4, p.597-617, Aug, 2000.

PURCELL, S. et al. WHAP: haplotype-based association analysis. **Bioinformatics**, v.23, n.2, p.255-6, Jan 15, 2007.

RABON-STITH, K. M. et al. Vitamin D receptor FokI genotype influences bone mineral density response to strength training, but not aerobic training. **Exp Physiol**, v.90, n.4, p.653-61, Jul, 2005.

RACHEZ, C.; FREEDMAN, L. P. Mechanisms of gene regulation by vitamin D (3) receptor: a network of coactivator interactions. **Gene**, v.246, n.1-2, p.9-21, 2000.

RAMALHO, A. C. et al. Fractures of the proximal femur: correlation with vitamin D receptor gene polymorphism. **Braz J Med Biol Res**, v.31, p.921-927, 1998.

RAY, D. et al. Protein tyrosine phosphatase non-receptor type 22 (PTPN22) gene R620W variant and sporadic idiopathic hypoparathyroidism in Asian Indians. **Int J Immunogenet**, v.33, n.4, p.237-40, Aug, 2006.

REICH, D. E. et al. Linkage disequilibrium in the human genome. **Nature**, v.411, n.6834, p.199-204, May 10, 2001.

RISCH, N.; MERIKANGAS, K. The future of genetic studies of complex human diseases. **Science**, v.273, n.5281, p.1516-7, Sep 13, 1996.

ROSENBERG, N. A. et al. Informativeness of genetic markers for inference of ancestry. **Am J Hum Genet**, v.73, n.6, p.1402-22, Dec, 2003.

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

ROZEN, S.; SKALETSKY, H. Primer3 on the WWW for general users and for biologist programmers. **Methods Mol Biol**, v.132, p.365-86, 2000.

SACHIDANANDAM, R. et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. **Nature**, v.409, n.6822, p.928-33, Feb 15, 2001.

SAWYER, S. L. et al. Linkage disequilibrium patterns vary substantially among populations. **Eur J Hum Genet**, v.13, n.5, p.677-86, May, 2005.

SELDIN, M. F. et al. European population substructure: clustering of northern and southern populations. **PLoS Genet**, v.2, n.9, p.e143, Sep 15, 2006.

SEYFERTH, G. **Imigração no Brasil: os preceitos de exclusão** 2000. Disponível em: <http://www.comciencia.br/reportagens/migracoes/migr03.htm>. Acesso em Janeiro de 2006.

SEYFERTH, G. Colonização, imigração e a questão racial no Brasil. **Revista USP**, v.53, p.117-49, 2002.

SHEN, H. et al. Nonreplication in genetic studies of complex diseases--lessons learned from studies of osteoporosis and tentative remedies. **J Bone Miner Res**, v.20, n.3, p.365-76, Mar, 2005.

SHERRY, S. T. et al. dbSNP: the NCBI database of genetic variation. **Nucleic Acids Res**, v.29, n.1, p.308-11, Jan 1, 2001.

SHI, M. et al. Genotype frequencies and linkage disequilibrium in the CEPH human diversity panel for variants in folate pathway genes MTHFR, MTHFD, MTRR, RFC1, and GCP2. **Birth Defects Res A Clin Mol Teratol**, v.67, n.8, p.545-9, Aug, 2003.

SHRIVER, M. D. et al. Skin pigmentation, biogeographical ancestry and admixture mapping. **Hum Genet**, v.112, n.4, p.387-99, Apr, 2003.

SHRIVER, M. D. et al. Ethnic-affiliation estimation by use of population-specific DNA markers. **Am J Hum Genet**, v.60, n.4, p.957-64, Apr, 1997.

SIMINOVITCH, K. A. PTPN22 and autoimmune disease. **Nat Genet**, v.36, n.12, p.1248-9, Dec, 2004.

SLADEK, R. et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. **Nature**, v.445, n.7130, p.881-5, Feb 22, 2007.

SMINK, L. J. et al. T1DBase, a community web-based resource for type 1 diabetes research. **Nucleic Acids Res**, v.33, n.Database issue, p.D544-9, Jan 1, 2005.

SMIT, A. F. A. et al. **RepeatMasker Open-3.0**. 1996. Disponível em: <http://www.repeatmasker.org> Acesso em: Julho de 2005

SMITH, M. W. et al. Markers for mapping by admixture linkage disequilibrium in African American and Hispanic populations. **Am J Hum Genet**, v.69, n.5, p.1080-94, Nov, 2001.

SMITH, M. W. et al. A high-density admixture map for disease gene discovery in african americans. **Am J Hum Genet**, v.74, n.5, p.1001-13, May, 2004.

SMITHIES, O.; WALKER, N. F. Genetic control of some serum proteins in normal humans. **Nature**, v.176, n.4496, p.1265-6, Dec 31, 1955.

SMYTH, D. et al. Replication of an association between the lymphoid tyrosine phosphatase locus (LYP/PTPN22) with type 1 diabetes, and evidence for its role as a general autoimmunity locus. **Diabetes**, v.53, n.11, p.3020-3, Nov, 2004.

SPOTILA, L. D. et al. Vitamin D receptor genotype is not associated with bone mineral density in three ethnic/regional groups. **Calcif Tissue Int**, v.59, n.4, p.235-7, Oct, 1996.

STENZEL, A. et al. Patterns of linkage disequilibrium in the MHC region on human chromosome 6p. **Hum Genet**, v.114, n.4, p.377-85, Mar, 2004.

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

STEPHENS, M.; DONNELLY, P. A comparison of bayesian methods for haplotype reconstruction from population genotype data. **Am J Hum Genet**, v.73, n.5, p.1162-9, Nov, 2003.

STRANGER, B. E. et al. Genome-wide associations of gene expression variation in humans. **PLoS Genet**, v.1, n.6, p.e78, Dec, 2005.

STRANGER, B. E. et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. **Science**, v.315, n.5813, p.848-53, Feb 9, 2007.

SUAREZ-KURTZ, G. et al. Detection of CYP2C9*5 in a white Brazilian subject. **Clin Pharmacol Ther**, v.77, n.6, p.587-8, Jun, 2005.

SVED, J. A. Linkage disequilibrium and homozygosity of chromosome segments in finite populations. **Theor Popul Biol**, v.2, n.2, p.125-41, Jun, 1971.

TAKAHASHI, T. et al. Rosbin: A Novel Homeobox-Like Protein Gene Expressed Exclusively in Round Spermatids. **Biol Reprod**, v.70, n.5, p.1485-1492, May 1,

, 1470, 7.5, p163-692,Jul, 1999.1,

VIEIRA, R. G. et al. Autodenominação de cor de pele e ancestralidade genômica em uma amostra de mulheres Brasileiras em pós-menopausa utilizada em estudos de associação genética. In: CONGRESSO BRASILEIRO DE GENÉTICA, 52., CONGRESSO DE LA ASSOCIACIÓN LATINOAMERICANA DE GENÉTICA, 12. 2006, Foz do Iguaçu. **Resumos**... Foz do Iguaçu: Sociedade Brasileira de Genética. p. 695, Set, 2006. 1 CD-ROM.

VOIGHT, B. F. et al. A map of recent positive selection in the human genome. **PLoS Biol**, v.4, n.3, p.e72, Mar, 2006.

WAKELEY, J.; LESSARD, S. Theory of the effects of population structure and sampling on patterns of linkage disequilibrium applied to genomic data from humans. **Genetics**, v.164, n.3, p.1043-53, Jul, 2003.

WALL, J. D.; PRITCHARD, J. K. Assessing the performance of the haplotype block model of linkage disequilibrium. **Am J Hum Genet**, v.73, n.3, p.502-15, Sep, 2003.

WALTERS, M. R. Newly identified actions of the vitamin D endocrine system. **Endocr Rev**, v.13, n.4, p.719-64, Nov, 1992.

WANG, D. G. et al. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. **Science**, v.280, n.5366, p.1077-82, May 15, 1998.

WANG, N. et al. Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. **Am J Hum Genet**, v.71, n.5, p.1227-34, Nov, 2002.

WEINBERG, W. Über den Nachweis der Verebung beim Menschen. **Jahreshefte des Vereins für vaterländische Naturkunde in Württemberg**, v.64, p.368-382, 1908.

WEISS, K. M.; CLARK, A. G. Linkage disequilibrium and the mapping of complex human traits. **Trends Genet**, v.18, n.1, p.19-24, Jan, 2002.

WILLER, C. J. et al. Tag SNP selection for Finnish individuals based on the CEPH Utah HapMap database. **Genet Epidemiol**, v.30, n.2, p.180-190, 2005.

WRIGHT, S. The interpretation of population structure by F-statistics with special regards to systems of mating. **Evolution**, v.19, p.395-420, 1965.

YAMAMOTO, H. et al. The caudal-related homeodomain protein Cdx-2 regulates vitamin D receptor gene expression in the small intestine. **J Bone Miner Res**, v.14, n.2, p.240-7, Feb, 1999.

YU, N. et al. Larger genetic differences within africans than between Africans and Eurasians. **Genetics**, v.161, n.1, p.269-74, May, 2002.

ZEGGINI, E. et al. Characterisation of the genomic architecture of human chromosome 17q and evaluation of different methods for haplotype block definition. **BMC Genet**, v.6, n.1, p.21, 2005.

ZHANG, K. et al. Randomly distributed crossovers may generate block-like patterns of linkage disequilibrium: an act of genetic drift. **Hum Genet**, v.113, n.1, p.51-9, Jul, 2003.

ZHANG, K. et al. Haplotype block structure and its applications to association studies: power and study designs. **Am J Hum Genet**, v.71, n.6, p.1386-94, Dec, 2002.

ZMUDA, J. M. et al. Vitamin D receptor translation initiation codon polymorphism and markers of osteoporotic risk in older African-American women. **Osteoporos Int**, v.9, n.3, p.214-9, 1999.

ZMUDA, J. M. et al. Molecular epidemiology of vitamin D receptor gene variants. **Epidemiol Rev**, v.22, n.2, p.203-17, 2000.

ZMUDA, J. M. et al. Genetic epidemiology of osteoporosis: past, present, and future. **Curr Osteoporos Rep**, v.3, n.3, p.111-5, Sep, 2005.

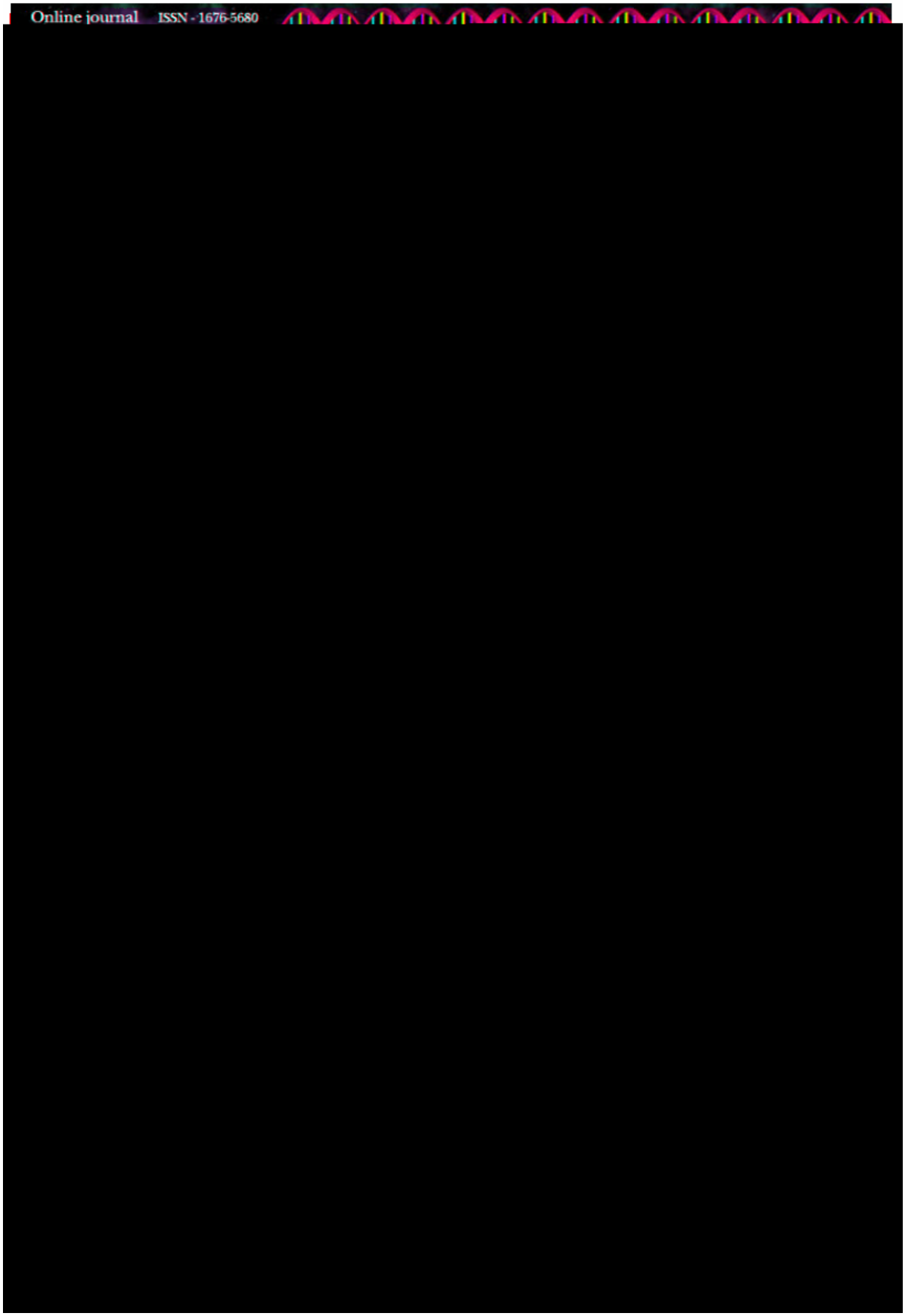
APÊNDICE A

Artigo diretamente relacionado com o desenvolvimento do trabalho, submetido e aceito para publicação na Revista **Genetics and Molecular Research (GMR) (versão não corrigida da revista)**

Título: **A multiplex single base extension protocol for genotyping Cdx2, FokI, BsmI, ApaI and TaqI polymorphisms of the vitamin D receptor gene**

Autores: Tulio Cesar de Lima Lins, Luciana Rollemberg Nogueira, Ricardo Moreno Lima, Paulo Gentil, Ricardo Jacó de Oliveira e Rinaldo Wellerson Pereira

Participação: TCLL conduziu os ensaios de genotipagem em multiplex para extensão de base única, análises genéticas e colaborou em grande parte com o manuscrito.


Online journal ISSN - 1676-5680

zilian postmenopausal women who took part in an ongoing association study carried out by members of our group.

Key words: VDR, SNPs, Multiplex genotyping, Single-base extension

INTRODUCTION

The active metabolite of vitamin D, 1,25-dihydroxyvitamin D3, and the steroid/thyroid hormone nuclear receptor coded by the vitamin D receptor gene (VDR) are the two major players in the vitamin D endocrine system. Besides the essential role in calcium homeostasis and bone metabolism, research in the last two decades has shown that cell differentiation, inhibition of cell growth, immunomodulation, and control of other hormonal systems are also biological processes where the vitamin D endocrine system plays an important role (Dusso et al., 2005). The 1,25-dihydroxyvitamin D3 broad transcriptional action mediated through the vitamin D receptor makes it a strong candidate in association studies from bone phenotypes to cancer (Valdivielso and Fernandez, 2006).

The Cdx2, FokI, ApaI, TaqI, and BsmI polymorphisms are broadly genotyped in VDR/disease association studies (Uitterlinden et al., 2004). The Cdx2 polymorphism is an A to G transition named according to their location at the intestinal specific transcription factor Cdx2-binding site in the VDR 1e promoter region. The A allele is more active than the G allele regarding the Cdx2 transcription factor binding. Thus, the A allele is associated with more intestinal VDR gene transcription (Arai et al., 1997; Yamamoto et al., 1999). The FokI polymorphism is a T to C transition disrupting the start codon and resulting in a vitamin D receptor protein that is three amino acids shorter and more active as a transcription factor (Arai et al., 1997). The other three single nucleotide polymorphisms (SNPs) are in intron 8 and show strong linkage disequilibrium with each other and with other 3' end polymorphisms (Fang et al., 2005). There is no clear evidence of a functional role played by any of the three SNPs at the 3' end and the associations found in some studies probably are related to linkage disequilibrium among them and some functional polymorphic site (Ingles et al., 1997). Cross-sectional and case control studies regarding bone phenotypes where the five polymorphisms are investigated show discordant results. If the difficult to reproduce the associations is related to population genetic background or to poor study design will be cleared by well-designed large scale association study in different populations (Shen et al., 2005; Zmuda et al., 2005, 2006). The implementation of such studies will demand medium to highthrougput automated genotyping methods.

Besides the technical development experimented in the SNP genotyping field (Kwok and Chen, 2003; Syvanen, 2005), most of published association studies genotyping the FokI, BsmI, ApaI, and TaqI use polymerase chain reaction-restriction fragment length polymorphism (PCR-RFLP). The Cdx2 investigation in association studies was first carried out by direct PCR sequencing (Arai et al., 2001), but since Fang et al. (2003) described an allele specific amplification protocol, it has been the method widely used. More recent association studies has genotyped the BsmI, ApaI, TaqI, and FokI using highthrougput TaqMan allelic discrimination assays (Fang et al., 2005). Both PCR-RFLP and allele-specific amplification methodologies are easily implemented when the number of samples to be genotyped is not too high. However, powerful association studies demand dozens of hundreds samples to be genotyped and non-automated meth-

ods turn to be difficult to use when planning such large scale association studies. As PCR-RFLP and allele-specific amplification methods genotype samples individually and normally there is no way to automate the allele-calling process the possibility to mess genotypes from different samples during the data handling must be considered. Another issue regarding methods that genotype one marker per time is DNA sample limitation, which sometimes may occur.

This study presents a straightforward protocol for genotyping the five most investigated VDR polymorphisms. The method is based on multiplex PCR, single-base extension methods and automated allele-calling process over- and automated allele calling. The multiplex fashion a e genotyping errors due to mixing of differ- come issues with DNA sample limitation and possible The method was developed to work with ent sample genotypes during manual data handling. he cost of genotyping. minimum amounts of reagents, which also reduces t

MATERIAL AND METHODS

DNA samples

ıcted from peripheral blood using a modified The DNA samples used in this study were extra ıples represent eight inclusion paternity trios salting out protocol (Miller et al., 1988). The DNA sam d in one ongoing project study of our group, and 7 samples of Brazilian postmenopausal women use

Vitamin D receptor single nucleotide polymorphisms

d here The five SNPs chosen to be part of the multiplex genotyping protocol describe nge of are those broadly investigated in association studies between VDR gene and a diverse ra I. The phenotypes. The SNPs appear in the literature as Cdx2, FokI, BsmI, ApaI, and Taq http:// sequence surrounding each of the five SNPs was easily retrieved from the dbSNP (entifi- www.ncbi.nlm.nih.gov/projects/SNP/) using respectively the following reference SNP id cation numbers: rs11568820, rs10735810, rs1544410, rs7975232, and rs731236.

Polymerase chain reaction primer design

rder to The sequences downloaded from dbSNP were used to design PCR primers in o rithm, amplify fragments harboring each SNP. The primers were designed using Primer3 algo mer3/ which is freely available on the internet (http://frodo.wi.mit.edu/cgi-bin/pri , PCR primer3_www.cgi). As the ApaI and TaqI loci are only 80 bp apart from each other d. The primers for amplification of only one fragment harboring the two SNPs were designe exing. hairpin and primer dimer formation can potentially impair the PCR efficiency and multipl

[illegible] taxi. Tested using the AutoDimer algorithm which was developed specifically to design primers m ting, optimizing the multiplexing success (Vallone and Butler, 2004). PCR primer sequences and nes temperatures are shown in Table 1.

Single-base extension primer design

ense The primers for single-base extension genotyping were designed manually in the s

Table 1. Polymerase chain reaction (PCR) Primers designed to amplify fragments harboring the five vitamin D receptor single nucleotide polymorphosm.

Polymorphism	Forward PCR primer (5'-3')	Reverse PCR primer (5'-3')	Fragment size (bp)	Tm F/R (°C)
Cdx2	CATTGTAGAACATCTTTTGTATCAGGA	GACAAAAAGGATCAGGGATGA	224	59.9/59.0
FokI	GGCCTGCTTGCTGTTCTTAC	TCACCTGAAGAAGCCTTTGC	174	60.0/60.5
BsmI	CCTCACTGCCCTTAGCTCTG	CCATCTCTCAGGCTCCAAAG	209	60.1/59.9
ApaI TaqI	CTGCCGTTGAGTGTCTGTGT	TCGGCTAGCTTCTGGATCAT	242	59.9/59.9

Tm F/R = melting temperature for forward and reverse primers.

or antisense sequence, ending one base adjacent to the SNP to be genotyped. They were checked for primer hairpin and dimer formation with Autodimer. The SBE primer length was arranged in such a way as to distribute the fragments in the range of 20 to 60 bp adding a poly(T) tail on the 5' end of each primer (Table 2).

Table 2. Single-base extension primers for SNaPShot® multiplex reaction.

Polymorphism	Single-base extension primer	Primer size (bp)	Alleles genotyped	DNA strand direction	Reference alleles	Genotyped fragment size (bp)
Cdx2	CCTGAGTAAACTAGCTCAGA	20	A/C	Forward	A/C	22/23
/52 FokI	$(T)_{31}$GCTGGCCGCCATTGCCTCC	50	A/G	Reverse	C/T	51
/46 BsmI	$(T)_{21}$CAGAGCCTGAGTATTGGGAATG	43	C/T	Reverse	A/G	45
/39 ApaI	$(T)_{12}$GTGGTGGGATTGAGCAGTGAGG	34	G/T	Reverse	A/C	38
/31 TaqI	$(T)_9$GCGGTCCTGGATGGCCTC	27	A/G	Reverse	C/T	29

Multiplex polymerase chain reaction amplification

The PCR was optimized to co-amplify the four fragments in one single reaction. The 250 protocol was carried out in 12.5 µL as follows: 1X Taq polymerase buffer, 2.5 mM $MgCl_2$, rase. mM dNTPs, 1.6 mg/mL BSA, 0.50 µM of each primer, 10-40 ng DNA, 1 U of Taq polymer cles: PCR amplification was performed in an ABI9700 thermocycler using the following cy sing denaturation at 95°C for 5 min, followed by 15 cycles of 40 s at 95°C, 40 s at 62°C decrea final 0.5°C per cycle, 40 s at 72°C, and 15 cycles of 40 s at 95°C, 40 s at 54°C, 40 s at 72°C, and extension step at 72°C for 5 min.

Multiplex genotyping by minisequencing

ilka- The PCR-amplified products were purified with exonuclease I (ExoI) and shrimp- ners. line phosphatase (SAP) enzymes in order to eliminate non-incorporated dNTPs and prin SAP The enzymatic purification was carried out in 3 µL PCR mix by adding 1 U ExoI, 0.95 U in at and 0.5X SAP reaction buffer, which were incubated for 90 min at 37°C following 20 m

80°C for enzyme denaturation. The minisequencing was performed using 1.25 µL SNaPshot® Multiplex minisequencing kit reaction mix (Applied Biosystems), 1.25X Big Dye Sequencing Buffer, 1 µL purified PCR product, 1 µL Multiplex primer mix containing 0.2 µM of each single-base extension primer, and sterile autoclaved Milli-Q water up to 5 µL. Single-base extension was performed as follow: 25 cycles of 96°C for 10 s, 50°C for 5 s, and 60°C for 30 s. Single-base extended products were enzymatically purified in order to degrade fluorescent ddNTP not incorporated in the reaction by adding 0.5 U SAP diluted 1:1 with 10X SAP reaction buffer for each reaction, followed by incubation at 37°C for 60 min and a step at 75°C for 20 min. The sample to be electrophoresed on ABI3100 was prepared adding 1 µL purified SBE products in 8.85 µL Hi-Di formamide and 0.15 µL GS120 Liz internal size standard. Samples were electrophoresed on an ABI prism 3100 Genetic Analyzer (Applied Biosystems), setting the equipment to use the SNP36_POP4 default module. Our group uses the ABI3700 POP-6 polymer instead of ABI3100 POP-4 polymer. The electropherograms were analyzed with GeneMapper 3.5 or Genescan Analysis 3.7/Genotyper 3.7 software (Applied Biosystems).

Polymerase chain reaction product sequencing reaction

Direct PCR product sequencing reactions were carried out for samples representing both homozygous and heterozygous genotypes for each VDR SNP. This procedure was performed in singleplex PCR using the same protocol described above. Each singleplex reaction was enzymatically purified using 0.5 U SAP and 0.5 U ExoI over 5 µL PCR product. The sequencing reactions were carried out using 1 µL ExoI/SAP purified PCR amplicons, 0.5 µL Big Dye® Terminator Kit, 3.2 pmol PCR primer, 1.75 µL Big Dye Sequencing Buffer (Applied Biosystems) in a total volume of 10 µL. The sequencing reaction was performed as follows: 96°C for 1 min and 25 cycles of 96°C for 10 s, 50°C for 5 s, and 60°C for 4 min. The sequencing reaction fragments were purified using EDTA/ethanol protocol, where 2.5 µL 125 mM EDTA and 30 µL 100% ethanol were added to each sample. The samples were mixed by inversion and incubated at room temperature for 15 min. After incubation, samples were centrifuged at 4°C at 1650 *g* for 45 min. The supernatant was discarded by inversion and 30 µL 70% ethanol was added to each sample. The samples were centrifuged at 4°C at 1650 *g* for 15 min. The supernatant was discarded by inversion and each sample was allowed to dry at room temperature for 30 min. Afterward, 10 µL Hi-Di formamide was added and samples denatured at 96°C for 3 min and immediately chilled at 4°C for 2 min. Samples were electrophoresed on an ABI prism 3100 Genetic Analyzer (Applied Biosystems), setting the equipment to use the RapidSeq36_POP6 default module. The electropherograms were analyzed using the Seqscape® 2.1 software.

RESULTS

The primers designed to co-amplify four PCR fragments harboring the five SNPs allowed for easy amplification optimization using standard PCR protocol conditions. Enzymatic clean up of PCR products was optimized to work with 3 µL, reduced from the 10 µL recommended by the SNaPshot® Multiplex minisequencing kit. Genotyping the five SNPs using the SNaPshot® Multiplex minisequencing kit was also optimized to minimize the amount of kit spent in each reaction. The reaction described here uses one-fourth of the kit's recommended single-reaction volume in one-half of the final reaction volume recommended by the manufacturer. The

base extension reaction buffering unbalancing due to the kit reaction mix cut down was fixed using the Big Dye Sequencing Buffer (Applied Biosystems). The amount of SAP used to remove the fluorescent ddNTP excess after the single-base extension reaction was also reduced to one-half as a consequence of the reduction in the final reaction volume. Electrophoresis in the ABI3100 was optimized to run with ABI3700 POP™-6 polymer, which costs roughly 50 times less than the ABI3100 POP®-4 polymer.

None of the eight paternity trios showed genotypes that were not in agreement with Mendelian transmission (Table 3). The peak height and fragment size for each allele of the five SNPs genotyped are shown in Figure 1. Direct comparisons of SBE products and sequencing reactions show no difference in homozygous and heterozygous samples for all five loci (Figure 2).

Table 3. Genotype from eight inclusion trios and seven samples of Brazilian postmenopausal women.

Sample	VDR polymorphism genotypes				
	Cdx2	FokI	BsmI	ApaI	TaqI
332M	A/A	C/T	G/A	A/A	C/T
332C	G/A	C/T	G/A	A/A	C/T
332P	G/A	C/C	G/A	C/A	C/T
336M	G/A	C/C	G/A	C/A	C/T
336C	G/A	C/T	G/A	C/A	C/T
336P	G/A	C/T	G/A	C/A	C/T
371M	G/A	C/T	A/A	A/A	C/C
371C	G/G	C/T	A/A	A/A	C/C
371P	G/A	C/T	G/A	C/A	C/T
383M	G/A	C/T	G/A	A/A	C/T
383C	G/A	C/T	G/A	C/A	C/T
383P	G/A	C/C	G/G	C/C	T/T
414M	G/A	C/C	G/G	A/A	C/T
414C	G/G	C/C	G/G	A/A	T/T
414P	G/A	C/C	G/A	A/A	T/T
431M	G/A	C/C	G/A	C/A	C/T
431C	G/A	C/C	G/A	C/A	C/T
431P	G/A	C/T	G/A	C/C	T/T
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
C/T	447P	G/A	C/C	G/A	C/A
C/C	451M	G/G	C/C	A/A	A/A
C/T	451C	G/G	C/T	G/A	C/A
T/T	451P	G/G	C/T	G/G	C/C
C/T	009	G/A	C/C	G/A	C/A
T/T	013	G/G	C/T	G/G	C/A
C/T	014	G/A	T/T	G/G	A/A
C/C	029	G/A	C/T	A/A	A/A
C/C	041	G/A	C/C	G/A	A/A
T/T	042	G/A	T/T	G/G	C/C
T/T	051	A/A	C/C	G/A	C/A

ther, "C" child, and "P" father. Alleles refer to the dbSNP reference allele.

VDR = vitamin D receptor. Sample number indicates the case, and the letter "M" indicates mot

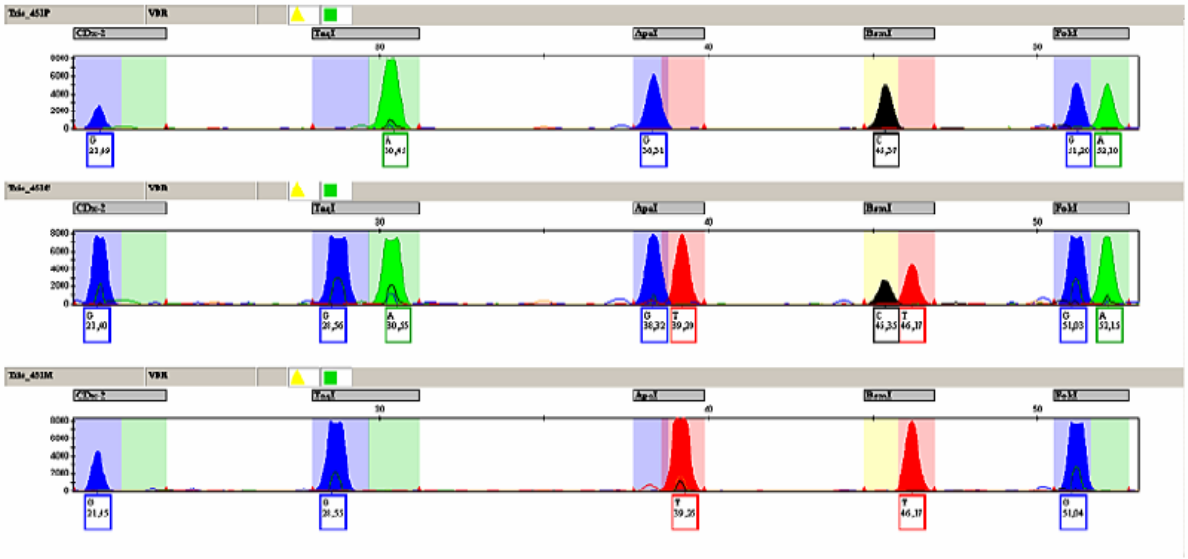


Figure 1. Electropherogram demonstrating allele calling in one genotyped trio (451). Labels below peaks indicate allele calling with respective size in base pairs. The letter "P" indicates father, "C" indicates child, and "M" indicates mother.

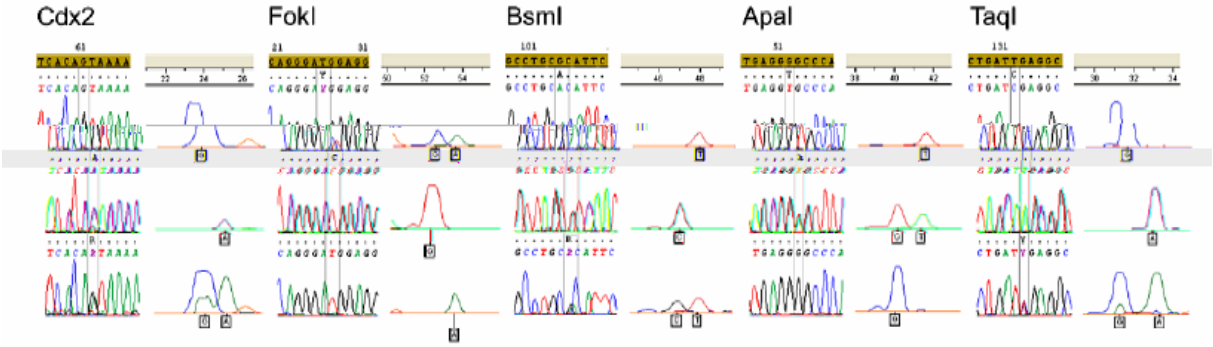


Figure 2. Direct comparison of PCR amplicon sequencing and SNaPshot reaction fragments. Genotyped samples are displayed as follows: Cdx2 (013 G, 051 A, 009 G/A), FokI (013 C/T, 009 C, 042 T), BsmI (029 A, 042 G, 009 G/A), ApaI (014 T, 009 G/T, 042 G), TaqI (041 C, 042 T, 009 C/T).

DISCUSSION

Single-base extension has been shown to be a very straightforward method for SNP genotyping, allowing medium and high throughput method development (Syvanen, 19 99, 2005). Here, we present an optimized protocol based on fluorescent single-base extension combined with capillary electrophoresis that allowed for the multiplex genotyping of the five r nost often used VDR gene polymorphisms in association studies, from bone phenotypes to diffe rent kinds of cancer. Besides optimizing a set of PCR and single-base extension primers that w orked in a multiplex fashion the method presented here introduces protocol changes in the co mmercial single-base extension kit in such a way as to reduce genotyping costs. Amounts of SAP and ExoI were reduced as the volume of the PCR product enzymatically cleaned up was r educed to one-third. Genotyping of the five VDR SNPs worked well even reducing the amo unt of kit

reaction mix to one-fourth and final reaction volume to one-half. The use of ABI3700 POP™-6 polymer instead of the recommended ABI3100 POP™-4 also helped to decrease the genotyping costs, which is essential when planning medium- to large-scale association studies.

The genotypes obtained by single-base extension were in agreement with those obtained by direct PCR amplicon sequencing, showing that the new primers and the protocols optimized during the development of this study are a good alternative for those who plan to use the VDR polymorphisms in large genetic association studies. Besides the advantage of saving time, genotyping the five SNPs in parallel and in an automated way may also be important in decreasing the rate of genotyping errors due to allele calling and data handling.

ACKNOWLEDGMENTS

We are very grateful to Dr. Dario Grattapaglia for the ABI3100 sharing. T.C.L. Lins was supported by a CAPES MSc scholarship. We are grateful to CNPq funding for our work on VDR variation in Brazilians and for L.R. Nogueira support through PIBIC scholarships. Research also supported by Pró-Reitoria de Pós-Graduação da Universidade Católica de Brasília.

REFERENCES

Arai H, Miyamoto K, Taketani Y, Yamamoto H, et al. (1997). A vitamin D receptor gene polymorphism in the translation initiation codon: effect on protein activity and relation to bone mineral density in Japanese women. *J. Bone Miner. Res.* 12: 915-921.

Arai H, Miyamoto KI, Yoshida M, Yamamoto H, et al. (2001). The polymorphism in the caudal-related homeodomain protein Cdx-2 binding element in the human vitamin D receptor gene. *J. Bone Miner. Res.* 16: 1256-1264.

Dusso AS, Brown AJ and Slatopolsky E (2005). Vitamin D. *Am. J. Physiol. Renal Physiol.* 289: F8-28.

Fang Y, van Meurs JB, Berging AP, Hofman A, et al. (2003). Cdx-2 polymorphism in the promoter region of the human vitamin D receptor gene determines susceptibility to fracture in the elderly. *J. Bone Miner. Res.* 18: 1632-1641.

Fang Y, van Meurs JB, d'Alesio A, Jhamai M, et al. (2005). Promoter and 3'-untranslated-region haplotypes in the vitamin D receptor gene predispose to osteoporotic fracture: the Rotterdam study. *Am. J. Hum. Genet.* 77: 807-823.

Ingles SA, Haile RW, Henderson BE, Klonel LN, et al. (1997). Strength of linkage disequilibrium between two vitamin D receptor markers in five ethnic groups: implications for association studies. *Cancer Epidemiol. Biomarkers Prev.* 6: 93-98.

Kwok PY and Chen X (2003). Detection of single nucleotide polymorphisms. *Curr. Issues Mol. Biol.* 5: 43-60.

Miller SA, Dykes DD and Polesky HF (1988). A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res.* 16: 1215.

Shen H, Liu Y, Liu P, Recker RR, et al. (2005). Nonreplication in genetic studies of complex diseases-lessons learned from studies of osteoporosis and tentative remedies. *J. Bone Miner. Res.* 20: 365-376.

Syvanen AC (1999). From gels to chips: "minisequencing" primer extension for analysis of point mutations and single nucleotide polymorphisms. *Hum. Mutat.* 13: 1-10.

Syvanen AC (2005). Toward genome-wide SNP genotyping. *Nat. Genet.* 37: S5-10.

Uitterlinden AG, Fang Y, van Meurs JB, Pols HA, et al. (2004). Genetics and biology of vitamin D receptor polymorphisms. *Gene* 338: 143-156.

Valdivielso JM and Fernandez E (2006). Vitamin D receptor polymorphisms and diseases. *Clin. Chim. Acta* 371: 1-12.

Vallone PM and Butler JM (2004). AutoDimer: a screening tool for primer-dimer and hairpin structures. *Biotechniques* 37: 226-231.

Yamamoto H, Miyamoto K, Li B, Taketani Y, et al. (1999). The caudal-related homeodomain protein Cdx-2 regulates vitamin D receptor gene expression in the small intestine. *J. Bone Miner. Res.* 14: 240-247.

Yamada DM, Shan YT and Maffatte SB (2005). Genetic epidemiology of [illegible]

future. *Curr. Osteoporos. Rep.* 3: 111-115.
Zmuda JM, Sheu YT and Moffett SP (2006). The search for human osteoporosis genes. *J. Musculoskelet. Neuronal Interact.* 6: 3-15.

APÊNDICE B

Produção científica durante o período acadêmico relativo ao mestrado.

Artigos aceitos em periódicos

Tulio Cesar de Lima Lins, Luciana Rollemberg Nogueira, Ricardo Moreno Lima, Paulo Gentil, Ricardo Jacó de Oliveira e Rinaldo Wellerson Pereira. A multiplex single base extension protocol for genotyping Cdx2, FokI, BsmI, ApaI and TaqI polymorphisms of the vitamin D receptor gene. Genetics and Molecular Research, v. 6, n. 2, p. 216-224, 2007.

Paulo Gentil, Ricardo Moreno Lima, Tulio Cesar de Lima Lins, Breno Silva de Abreu, Rinaldo Wellerson Pereira e Ricardo Jacó de Oliveira. Physical activity, Cdx-2 genotype and BMD. International Journal of Sports Medicine, volume 28.

Ricardo Moreno Lima, Breno Silva de Abreu, Paulo Gentil, Tulio Cesar de Lima Lins, Dario Grattapaglia, Rinaldo Wellerson Pereira, Ricardo Jacó de Oliveira. Lack of Association Between Vitamin D Receptor Genotypes and Haplotypes with Fat Free Mass in Postmenopausal Brazilian Women. Journal of Gerontology, volume 62A.

Trabalhos apresentados em congressos

LINS, T. C. L.; ABREU, B. S.; VIEIRA, R. G.; SILVA, M. L.; GENTIL, P.; LIMA, R. M.; OLIVEIRA, R. J.; GRATTAPAGLIA, D.; PEREIRA, R. W. Polimorfismos no gene receptor de vitamina D (VDR) e análise de haplótipos de acordo com ancestralidade genômica individual em mulheres Brasileiras em pós-menopausa. In: CONGRESSO BRASILEIRO DE GENÉTICA, 52., CONGRESSO DE LA ASSOCIACIÓN LATINOAMERICANA DE GENÉTICA, 12. 2006, Foz do Iguaçu. **Resumos**... Foz do Iguaçu: Sociedade Brasileira de Genética . p. 685, Set, 2006. 1 CD-ROM.

ABREU, B. S.; LINS, T. C. L.; VIEIRA, R. G.; SILVA, M. L.; GENTIL, P.; LIMA, R. M.; GRATTAPAGLIA, D.; OLIVEIRA, R. J.; PEREIRA, R. W. . Ancestralidade genômica, autodenominação de cor de pele e suas correlações com densidade mineral óssea (DMO) e índice de massa corporal (IMC). In: CONGRESSO BRASILEIRO DE GENÉTICA, 52., CONGRESSO DE LA ASSOCIACIÓN LATINOAMERICANA DE GENÉTICA, 12. 2006, Foz do Iguaçu. **Resumos**... Foz do Iguaçu: Sociedade Brasileira de Genética . p. 693, Set, 2006. 1 CD-ROM.

VIEIRA, R. G.; SILVA, M. L.; LINS, TCL ; ABREU, B. S.; LIMA, R. M.; GENTIL, P.; OLIVEIRA, R. J.; GRATTAPAGLIA, D.; PEREIRA, R. W. . Autodenominação de cor de pele e ancestralidade genômica em uma amostra de mulheres Brasileiras pós-menopausa utilizada em estudos de associação genética. In: CONGRESSO BRASILEIRO DE GENÉTICA, 52., CONGRESSO DE LA ASSOCIACIÓN LATINOAMERICANA DE GENÉTICA, 12. 2006, Foz do Iguaçu. **Resumos**... Foz do Iguaçu: Sociedade Brasileira de Genética . p. 695, Set, 2006. 1 CD-ROM.

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo