Métodos estatísticos em farmacogenômica

Marcelo Meireles Petenate

DISSERTAÇÃO APRESENTADA
AO
INSTITUTO DE MATEMÁTICA E STATÍSTICA
DA
UNIVERSIDADE DE SÃO PAULO
PARA
OBTENÇÃO DO TÍTULO
DE
MESTRE EM CIÊNCIAS

Programa: Estatística

Orientadora: Profa. Dra. Júlia Maria Pavan Soler

São Paulo, fevereiro de 2011

Métodos estatísticos em farmacogenômica

Esta versão definitiva da dissertação
contém as correções e alterações sugeridas pela
Comissão Julgadora durante a defesa realizada
por Marcelo Meireles Petenate e 17/12/2010

Comissão Julgadora

- Profa. Julia Maria Pavan Soler (Orientadora) – IME-USP
- Suely Ruiz Giolo – UFPR
- Viviana Giampaoli – IME-USP

"Todos os modelos estão errados,

alguns são úteis"

Box

"A resposta certa não importa nada:

o essencial é que as perguntas estejam certas"

Mário Quitanda

Dedico esse trabalho

À minha esposa, Keiko, por existir em minha vida

Aos meus pais, Ângela e Ademir, pelo apoio nos momentos difíceis

Meu irmão Guilherme

Agradecimentos

À minha esposa Keiko pelo companheirismo, paciência e, principalmente, pelo amor durante a elaboração dessa dissertação.

Ao meu pai por ser também meu professor, pelas inúmeras vezes que nós estudamos juntos quando as aulas da graduação pareciam impossíveis e por não ter me deixado desistir com seus conselhos.

À minha mãe, por ser a grande responsável por eu ter me tornado estatístico. Obrigado pela educação e valores transmitidos, principalmente nos momentos mais difíceis. Obrigado também pelos inúmeros incentivos para que eu fizesse o mestrado.

À minha orientadora, Profa. Dra. Júlia M. Paván Soler, pela confiança em mim depositada, orientação e paciência ao longo desse trabalho.

Ao professor Maurício Zevallos, pelas primeiras orientações em minha iniciação científica.

À Edina Miazaki, mais do que uma professora. Obrigado pelos conselhos e grande incentivo dado para eu fazer o mestrado e por ajudar a formar o que eu sou hoje.

Ao laboratório de Cardiologia e Genética Molecular do InCor/USP pela autorização para utilização dos dados reais.

A todos os professores e funcionários do Departamento de Estatística do IME/USP pela colaboração em todos os momentos e valiosos ensinamentos recebidos.

Resumo

Nas pesquisas para o desenvolvimento de novos medicamentos tem sido crescente o interesse por metodologias estatísticas que analisem as respostas aos fármacos levando-se em conta a informação genética dos indivíduos presentes no estudo. Espera-se que com essas metodologias possam ser desenvolvidos novos medicamentos que possibilitem um melhor tratamento aos pacientes, seja por seus compostos serem mais adequados à genética individual, seja pela prescrição de um regime de dose e dosagem mais adequado às características genéticas de um indivíduo.

O presente trabalho explora várias metodologias que incorporam a informação genética no estudo dos fármacos e desafios decorrentes: modelagem da farmacocinética e farmacodinâmica de um composto (Wu & Lin, 2009); análise de dados de famílias (Blangero et al., 1999); correção para estrutura de populações (Price et al., 2006 e 2009) e estudos com modelos animais de populações F_2 (Jiang & Zeng, 1995). Na análise de dados de famílias influentes na resposta a fármacos, resultados são extraídos da teoria do gráfico da variável adicionada aplicada a modelos de componentes de variância. As análises que envolvem correções para estrutura de populações consideram resultados da decomposição espectral de matrizes de dados genéticos e o gráfico Biplot.

Algumas das metodologias supracitadas são aplicadas a dados reais fornecidos pelo Laboratório de Cardiologia e Genética Molecular do InCor/USP. Dentre eles dados sobre a pressão arterial de diversas famílias da cidade de Baependi, no interior de Minas Gerais; dados de marcadores do tipo SNP de uma amostra de indivíduos não relacionados da região sudeste do Brasil e de 11 populações disponíveis no projeto HapMap; e medidas da pressão arterial de ratos provenientes de uma população F_2 nas condições basal, pós dieta de sal e pós tratamento com medicamento anti-hipertensivo. As análises foram realizadas usando, principalmente, os recursos computacionais do aplicativo SAS.

Palavras-chave: Farmacogenômica, delineamentos com famílias, estrutura de população, modelos mistos, biplot.

Abstract

In research for the development of new drugs, the interest in statistical methods to analyze the responses to drugs taking into account the genetic information of individuals in the study has been increasing. It is hoped that with these methodologies it can be developed new drugs that enable better treatments to patients, either because their compounds are more suitable for individual genetic, either by prescribing a dose and dosage regimen most appropriate to the genetic characteristics of an individual.

This work explores various methodologies that incorporate genetic information in the study of drugs and the challenges arising from it: modeling of pharmacokinetics and pharmacodynamics of a compound (Wu & Lin, 2009), analysis of family data (Blangero et al., 1999); correction to population structure (Price et al., 2006 and 2009) and studies in animal models of F_2 populations (Jiang & Zeng, 1995). In analyzing data from influential families in the response to drugs, results are extracted from the added variable plot theory applied to the variance component models. Analyses involving corrections for population structure consider results of spectral decomposition of matrices of genetic data and the Biplot graph.

Some of the above methodologies are applied to real data provided by the Laboratory of Molecular Cardiology and Genetics of Incor/USP: data on blood pressure of several families in Baependi city, Minas Gerais; data of SNP markers in a sample of unrelated individuals from southeastern Brazil and 11 populations available in the HapMap project; and measures of the blood pressure of rats from an F_2 population under basal conditions, post-salt diet and after treatment with antihypertensive medication. Analyses were performed using mainly the computational resources of SAS application.

Keywords: Pharmacogenomics, family designs, population structure, mixed models, biplot.

Sumário

Lista de Tabelas

Lista de Figuras

1. Introdução

Genética (do grego *genno*; fazer nascer) é a ciência que estuda a hereditariedade, ou seja, as características que são passadas de pais para filhos. A unidade fundamental física e funcional de hereditariedade é o gene, que foi primeiramente definido pelo experimento de Mendel com ervilhas em 1865. No início do século XX Thomas Morgan descobre que os "fatores hereditários" se encontram nos cromossomos e, por essa contribuição, é contemplado com o Prêmio Nobel de 1933. Esses estudos criam as bases para que os estudos de associação genética sejam realizados. Nesses estudos, a Estatística é uma ferramenta imprescindível, e em Kempthorne (1957) temos um reconhecimento dessa interdisciplinaridade "O ponto de partida da Genética, assim como a conhecemos no presente momento, foi a descoberta, por Gregor Mendel, de que as freqüências de tipos de descendentes de cruzamentos híbridos eram freqüências estatísticas... Parte da genialidade de Mendel consistiu no reconhecimento de que esse fenômeno tinha uma surpreendente semelhança com o ato de jogar moedas ou dados... O tremendo desenvolvimento da Genética no último século é uma boa evidência de que essa foi uma decisão sábia".

É conhecido que quase todos os fenômenos ou processos biológicos, incluindo a resposta à droga de um paciente, envolvem um componente genético. Como resultado desse fato é de extrema importância relacionar a aplicação dos medicamentos com a genética do indivíduo. Para alguns indivíduos, mesmo nas doses atualmente recomendadas, nota-se toxicidade. Para outros, com a mesma dose, não se nota efeito de toxicidade nem efeito benéfico da droga e, possivelmente, uma dose maior pudesse ser prescrita, pois esse indivíduo talvez tenha um metabolismo acelerado para o composto ativo desse medicamento. Assim, é esperado que com a inclusão da informação genética nos estudos de farmacocinética (o que o corpo faz com a droga) e farmacodinâmica (o que a droga faz com o corpo) do composto seja possível prescrever doses adequadas à composição genética individual (Wu & Lin, 2009).

O termo Farmacogenética foi cunhado por Friedrich Vogel em 1959 para descrever o estudo da variabilidade na resposta à droga devido a fatores hereditários. O termo farmacogenômica tem sido utilizado para pontuar que o genoma é mais do que um agregado de genes (Kirk et al., 2008). O primeiro estudo de farmacogenética de que se tem notícia é datado de 1932 e demonstrava que certos químicos reagem de forma diferente dependendo da genética individual. O primeiro exemplo de diferenças na metabolização de uma droga devido a fatores genéticos foi produzido na década de 1950 por Sweeney (2005) e podemos

pontuar esse momento como o início da farmacogenética moderna. Atualmente, as pesquisas farmacogenômicas lidam com o problema da personalização da terapia e é esperado para o futuro o desenvolvimento de novos medicamentos baseados na genética individual.

Como a informação genética não é levada em conta na grande maioria dos estudos farmacológicos atuais, os compostos aprovados são somente aqueles que produzem um efeito satisfatório para a resposta média populacional, ou seja, temos o desenvolvimento de drogas pela "média", independente da genética individual. É esperado que quando os estudos de desenvolvimento de drogas levarem em conta as informações genéticas muitos medicamentos poderão ser desenvolvidos para diferentes estratos genéticos da população. Isso é de extrema importância tanto para a população doente, que poderá se beneficiar de medicamentos novos e personalizados, quanto para a indústria farmacêutica, que poderá desenvolver novos compostos que antes eram estudados em fases pré-clínicas, se mostravam promissores, mas eram descartados devido ao não funcionamento para a população como um todo.

Outro fator importante a ser destacado é que a população brasileira é sabidamente miscigenada. Pode-se destacar a influência africana, européia e nativa como as principais fontes genéticas de constituição da população brasileira. Por esse motivo o *background* genético dessa população é diferente da população de outros países, o que sugere que muitas drogas desenvolvidas no exterior podem não ser eficientes aqui e muitos dos compostos rejeitados lá podem se mostrar eficazes aqui (Parra et al., 2003; Gonçalves et al., 2008). Além disso, miscigenação na população pode causar correlações espúrias entre alguma resposta ao fármaco e determinados estratos genéticos. Como a população brasileira é bastante miscigenada torna-se indispensável o estudo de métodos que corrijam esse problema (Price et al., 2006; Tiwari et al., 2008). Isso mostra a importância do desenvolvimento dessa área de pesquisa no Brasil.

Como motivação, as aplicações apresentadas nesse trabalho utilizam dados reais coletados em três estudos diferentes. O primeiro deles trata de um estudo com 119 famílias do município de Baependi, no interior de Minas Gerais. Esse estudo foi delineado com o intuito de pesquisar genes ligados à regulação da pressão arterial na população brasileira e o banco de dados possui informação da pressão arterial do indivíduo, idade, sexo, índice de massa corporal, se toma ou não toma algum medicamento para pressão, dentre outras variáveis. No presente trabalho o interesse foi ilustrar como o efeito geral de medicamento pode ser, inicialmente, decomposto em efeito relativo a fatores ambientais residuais e fatores genéticos. Esse fator genético pode também ser decomposto em efeitos genéticos familiares. A

ferramenta utilizada para tal decomposição é a teoria do gráfico da variável adicionada para modelos mistos.

Outro conjunto de dados aqui analisados é referente a uma amostra da população Sudeste do Brasil e do banco de dados público do *Haplotype Map* (www.hapmap.org), uma iniciativa com o objetivo de construir um mapa dos haplótipos humanos. Na análise desses dados é possível visualizar a grande miscigenação da população brasileira, comparada com outras populações mundiais, com forte ancestralidade européia e africana. Como a miscigenação é uma importante fonte de correlações espúrias, esse fato evidência a necessidade de um cuidado especial quando se estuda a população brasileira com o objetivo de realizar estudos de associação entre genes e doenças.

Para entender como funciona a miscigenação em populações experimentais, utilizamos os dados de um estudo do Laboratório de Cardiologia e Genética Molecular do Instituto do Coração de São Paulo (InCor-USP), com 221 ratos da geração F_2 resultantes de um cruzamento controlado entre linhagens de animais normotensos e hipertensos, de forma que o *background* genético fosse teoricamente idêntico, diferindo apenas nos genes controladores da pressão arterial. Algumas das variáveis coletadas foram: pressão arterial basal, pressão arterial pós-sal e pressão arterial após administração do medicamento Captopril, que é uma das drogas mais utilizadas no tratamento de doenças coronárias (Goodman e Gilman, 2005). Todos os animais foram genotipados utilizando um mapa com 182 marcadores moleculares espalhados pelos 21 cromossomos dos ratos (ver Figura 5. 6).

No capítulo 2 encontram-se alguns conceitos de Genética fundamentais para o entendimento das metodologias desenvolvidas subseqüentemente. Os conceitos farmacológicos e suas junções com a Genética, que criam a área da farmacogenética são apresentados no Capítulo 3. O Capítulo 4 apresenta um estudo realizado na cidade mineira de Baependi, onde informações, incluindo pressão arterial e utilização de medicamento foram coletadas em 1712 indivíduos. Nesse capítulo é descrito detalhadamente o gráfico da variável adicionada para modelos mistos, uma ferramenta estatística que foi utilizada nesse trabalho para decompor os efeitos genéticos familiares da medicação. Como será apresentado no Capítulo 2, a estrutura de uma população pode ser um causador de correlação espúria em estudos genéticos e para estudar suas possíveis implicações o Capítulo 5 detalha metodologias para lidar com esse problema e apresenta estudos, com humanos e animais, sobre esse problema. Finalmente, no Capítulo 6, são apresentadas as considerações finais, com conclusões obtidas com o presente trabalho bem como sugestões para futuros estudos.

2. Conceitos de Genética e Genômica

Para ajudar na compreensão da teoria desenvolvida neste trabalho faz-se necessária a introdução de alguns conceitos importantes de Genética.

2.1. Genoma

Todo organismo possui um genoma que contém a informação biológica necessária para sua sobrevivência e reprodução. A maioria dos genomas, incluindo o humano, é composto de DNA (ácido desoxirribonucléico), mas alguns poucos vírus têm RNA (ácido ribonucléico). O genoma humano, o qual é típico dos genomas de todos os animais multicelulares diplóides, é representado na Figura 2. 1 e consiste em duas partes distintas: O genoma nuclear, com aproximadamente 3 bilhões de nucleotídeos (é a unidade básica do DNA) doados de cada genoma parental (pai e mãe), está dividido em 23 pares de moléculas chamadas cromossomos, que variam entre 50 milhões e 260 milhões de nucleotídeos. Dos 23 pares de cromossomos existentes, 22 são ditos autossomos e um par é composto pelos cromossomos sexuais, o X e o Y, que são responsáveis pela determinação do sexo do indivíduo, totalizando 46 cromossomos. Genericamente falando, o gene é um conjunto de nucleotídeos que são responsáveis pela transmissão de alguma característica ao indivíduo e aproximadamente 30 mil genes estão presentes no genoma nuclear humano. A segunda parte, o genoma mitocondrial, é uma molécula de DNA circular que contém cerca de 40 genes (Farah, 1997; Pierce, 2005).

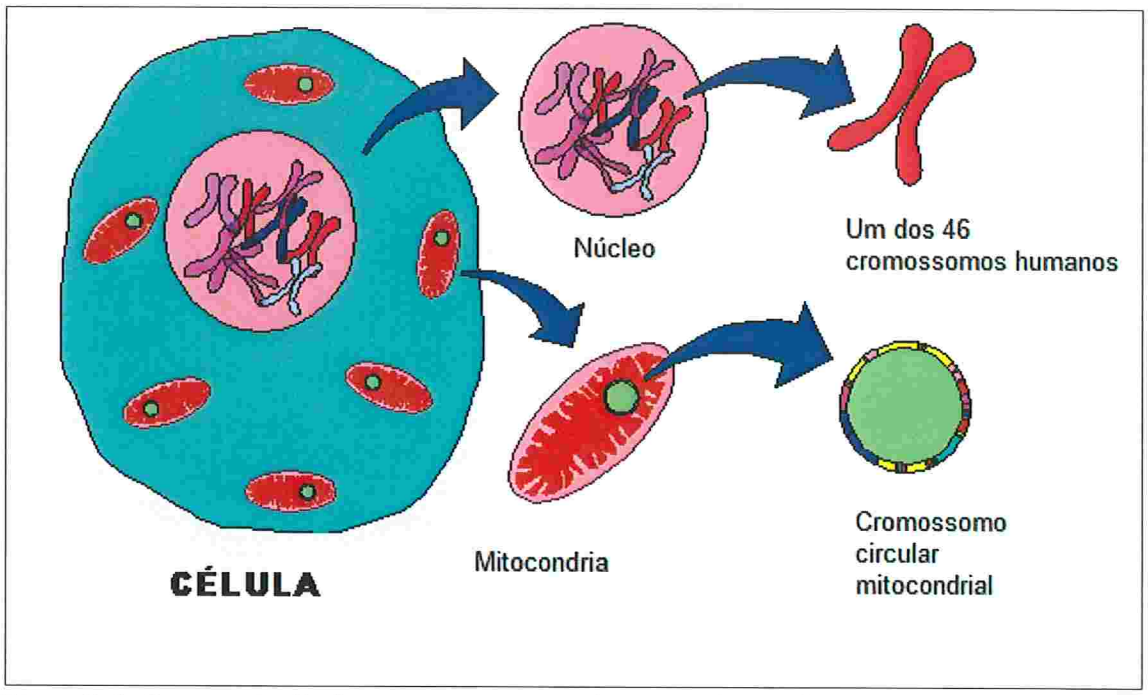

Figura 2. 1: Representação dos cromossomos nucleares e mitocondriais.

A maioria das nossas células possui sua própria cópia do genoma. As células somáticas, que compõem a grande maioria das nossas células, são diplóides e, portanto, têm duas cópias de cada autossomo, que com os dois cromossomos sexuais, XX para as fêmeas e XY para os machos, resultam em um total de 46 cromossomos (ver Figura 2. 2).

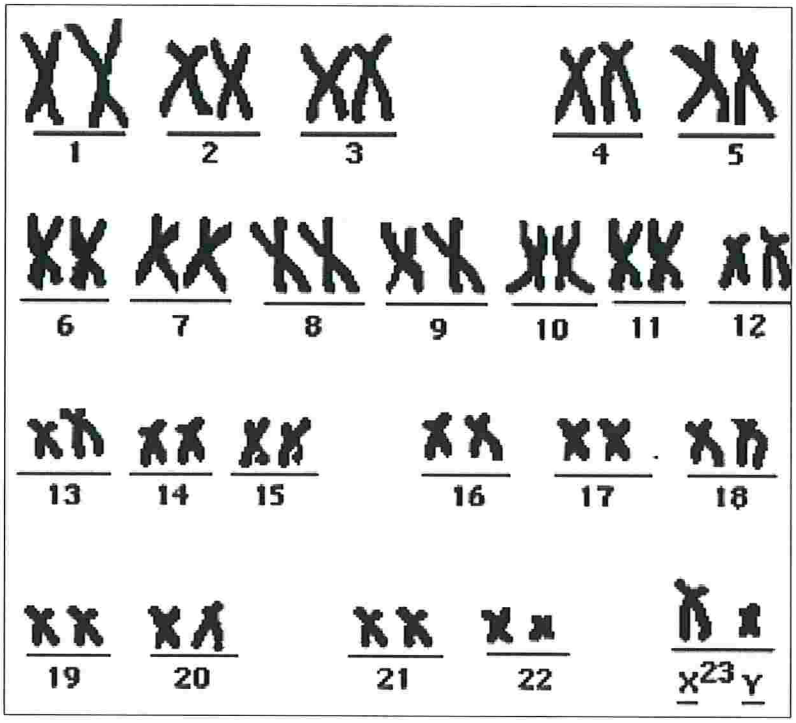

Figura 2. 2: Cariótipo humano (na meiose).

O genoma é uma "fábrica" de informação biológica, mas sozinho não é capaz de

mandar essa informação para a célula. A utilização da informação contida no genoma requer a atividade coordenada de enzimas e de outras proteínas, as quais participam de uma complexa rede de reações bioquímicas chamada de expressão genômica (Figura 2. 3). O produto inicial da expressão genômica é o transcriptoma, uma coleção de moléculas de RNA derivadas de regiões do DNA (genes) codificadoras de proteínas das quais a informação biológica é solicitada pela célula em um particular instante. O transcriptoma é produzido pelo processo chamado de transcrição, no qual genes individuais são copiados em moléculas de RNA. O segundo produto da expressão genômica é o proteoma, o repertório de proteínas da célula, o qual especifica a natureza das reações bioquímicas que a célula é capaz de realizar. As proteínas que compõem o proteoma são sintetizadas por translação das moléculas individuais de RNA presentes no transcriptoma (Brown, 2007; Falconer & Mackay, 1996).

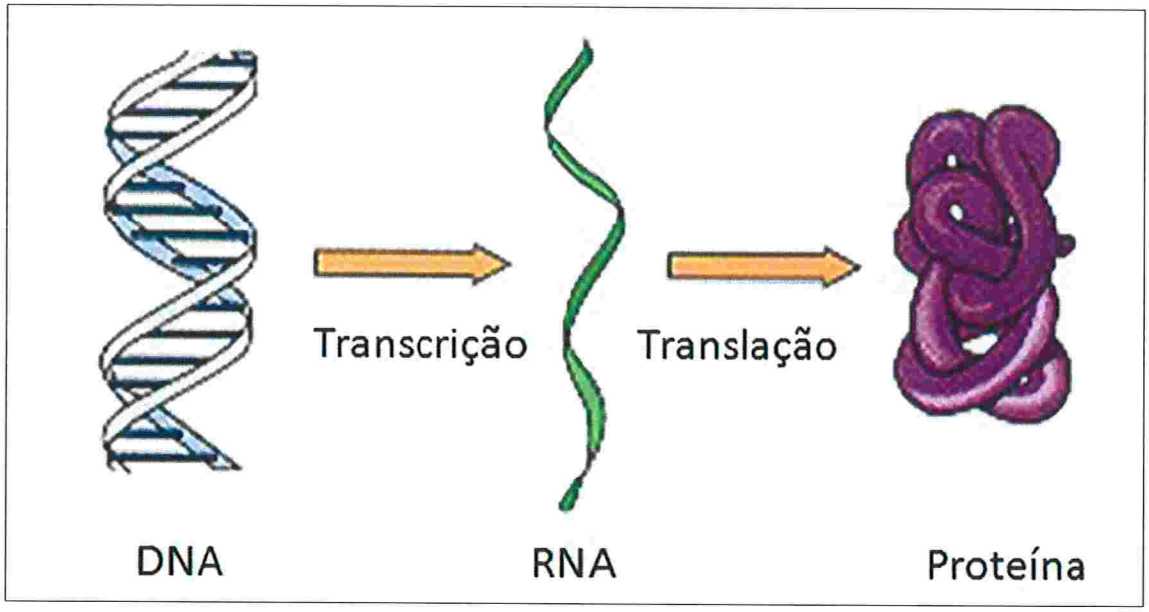

Figura 2. 3: Expressão genômica.

A molécula de DNA é composta pelos nucleotídeos Adenina (A), Citosina (C), Guanina (G) e Timina (T) que são ligados e estabilizados por ligações de hidrogênio. Em um DNA a Guanina se liga com a Citosina, enquanto que a Adenina se liga com a Timina. Esse arranjo de dois nucleotídeos complementares é chamado *par de bases*. Em organismos vivos, o DNA não existe como uma molécula única (cadeia simples), mas sim como um par de moléculas firmemente associadas (Watson & Crick, 1953; Berg et al., 2002). As duas longas cadeias de DNA enrolam-se como uma "trepadeira" formando uma dupla hélice (ver Figura 2. 4).

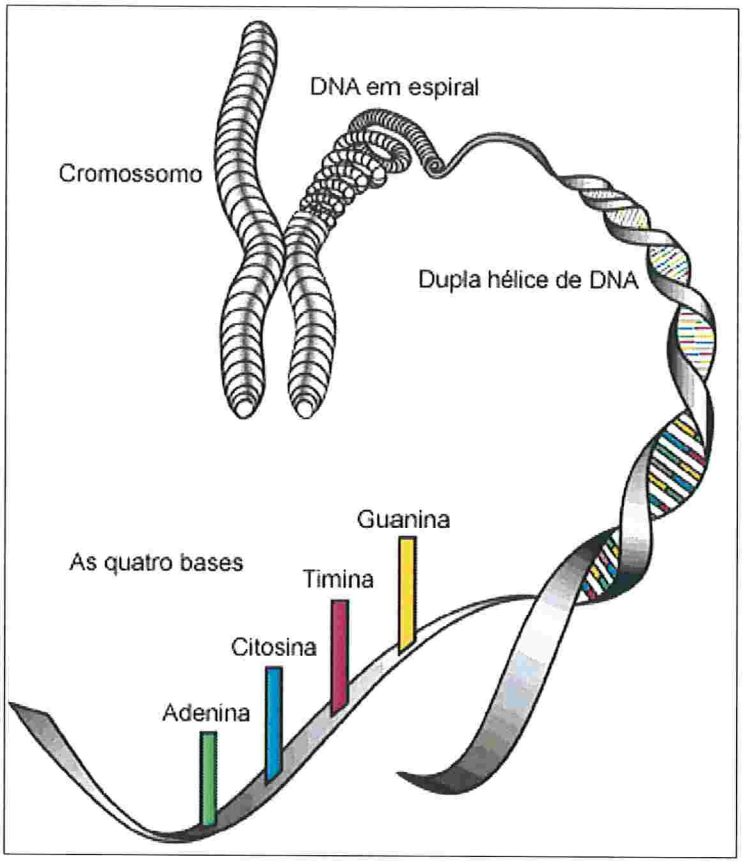

Figura 2. 4: Esquematização do cromossomo, DNA e suas bases de nucleotídeos.

Gene é um termo geral que significa a entidade física transmitida de pai para filho durante o processo de reprodução e que influencia características hereditárias (Falconer & Mackay, 1996) e é um conjunto de pares de bases que transmite alguma característica (fenótipo), observável ou não. Os fenótipos podem ser qualitativos (ou mendelianos) – definidos deterministicamente pela presença de um determinado gene, como no caso da anemia falciforme – ou quantitativos – que são controlados por dois ou mais genes e suas interações, entre si e com o meio ambiente, apresentando assim característica de variação no grau, tal como a hipertensão. Os locos (grandes regiões do genoma) que controlam traços quantitativos são chamados de Loco de Traço Quantitativo – QTL (do inglês *Quantitative Trait Locus*). O tamanho médio dos genes é de cerca de 3000 bases, mas os tamanhos variam consideravelmente e o maior gene, Distrofina, tem cerca de 2,4 milhões de bases. Como dito anteriormente, o número estimado de genes no genoma humano é 30.000. As funções ainda são desconhecidas para muitos dos genes já descobertos. O cromossomo 1, com cerca de 2970 genes, tem a maioria dos genes e o cromossomo Y, com cerca de 230 genes, a minoria. Quase todas as bases de nucleotídeos (99,9%) são iguais para as pessoas (Brown, 1997; Berg et al., 2001).

Cada mudança na seqüência de nucleotídeos em regiões codificadoras pode alterar a transcrição de proteína e, portanto, alterar a função genética. São essas alterações que se procura quando o genoma de algum indivíduo é seqüenciado e se dá o nome de alelo para cada forma alternativa de um gene. Do ponto de vista Estatístico, para certa região do DNA (loco) pode-se definir uma variável preditora classificatória, com n categorias, em que n é o número de diferentes formas alélicas existentes. Um indivíduo diplóide, como os humanos, tem dois alelos em um determinado loco uma vez que os cromossomos ocorrem em pares homólogos, um originário da mãe, o outro do pai. Como notação, para representar os alelos em um loco serão utilizadas letras maiúsculas com os respectivos índices indicando a forma alternativa, por exemplo, A_i e A_j. Uma linhagem é homozigota em algum dado gene se os alelos provenientes do pai e da mãe são idênticos, ou seja, no caso dialélico ele é A_1A_1 ou A_2A_2. Os indivíduos heterozigotos são aqueles que carregam um alelo de cada tipo (por exemplo, A_1A_2). Caso o alelo A_1 seja dominante para o alelo A_2 (equivalentemente A_2 é recessivo para A_1), então tanto o indivíduo A_1A_1 quanto o indivíduo A_1A_2 são fenotipicamente idênticos e a expressão da característica "carregada" pelo alelo A_2 só será manifestada caso o indivíduo seja A_2A_2. Por outro lado, em alelos codominantes, os fenótipos manifestados por indivíduos com genótipos diferentes (A_1A_1, A_1A_2 e A_2A_2, no caso dialélico) são também diferentes. Assim, locos multialélicos podem ser responsáveis por definirem um número maior de classes fenotípicas na população, juntamente com a interação entre dois ou mais locos (efeito de epistasia) e com o ambiente, caracterizando assim os fenótipos contínuos.

A meiose é o processo pelo qual são gerados os gametas, que são as células reprodutoras. Nesse processo os cromossomos homólogos se duplicam, entrelaçam-se e sofrem quebras. Nesse ponto de quebra o cromossomo junta-se com seu homólogo, formando assim um cromossomo híbrido e gerando variabilidade genética (Figura 2. 5). A esse evento dá-se o nome de recombinação gênica ou *crossing-over*. É importante ressaltar que quanto mais próximo dois locos estiverem entre si, menor a chance de recombinação e esses locos tendem a ser transmitidos juntos. Quando dois alelos estão distantes a chance de ocorrer recombinação é maior, gerando alelos recombinantes (Falconer & Mackay, 1996). Desse modo, o mecanismo de recombinação é função da distância entre dois locos (Lynch & Walsh, 1998). Podemos encontrar diferentes funções de distância genética em Lange (1997).

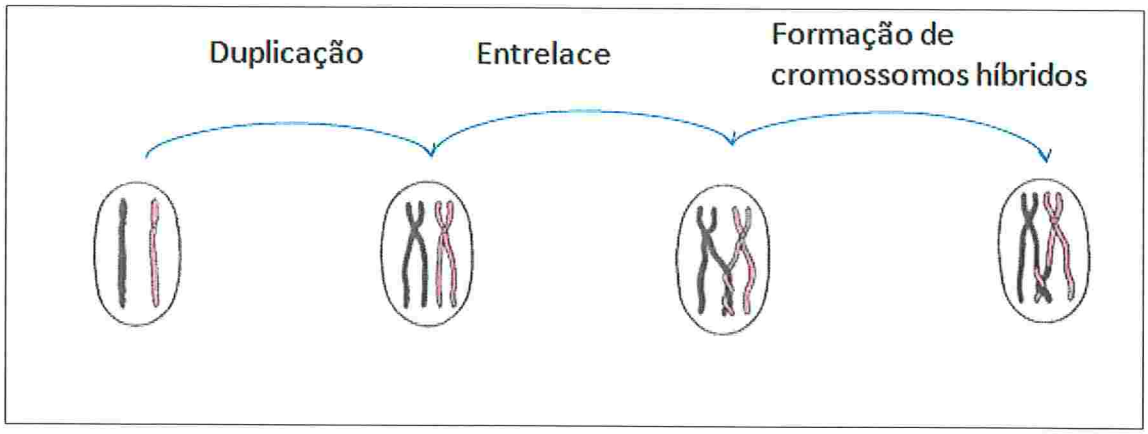

Figura 2. 5: Recombinação gênica.

2.2. Mapeamento Genético

Os métodos estatísticos para mapear QTLs, isto é, estimar suas posições no genoma e seus efeitos sobre um fenótipo, são baseados, principalmente, em duas medidas de associação ou de dependência entre locos: desequilíbrio de ligação e ligação (Falconer & Mackay, 1996). O primeiro é uma medida de associação probabilística entre alelos de locos diferentes e é usada em estudos observacionais de associação entre locos genéticos e doenças. A ligação entre locos depende dos eventos de recombinação entre locos, sendo muito utilizada quando dados de famílias ou de cruzamentos controlados são coletados (Lander & Botstein, 1989; Jansen & Stam, 1994; Zeng, 1994; Blangero et al., 2001).

Para realizar um mapeamento é necessário um mapa de marcadores que cubra adequadamente o genoma e a variação na característica fenotípica quantitativa entre as populações amostradas.

Amostragem do Genoma

Os mapas de cromossomos são uma maneira natural de organizar os dados genéticos sobre os cromossomos de uma maneira muito similar aos mapas ordinários (cartógrafos), que organizam dados geográficos sobre continentes, países e cidades, e são construídos por meio de uma amostragem dos locos do genoma. Os mapas genéticos são um tipo de mapa cromossômico usado em mapeamento de genes e são construídos a partir das distâncias entre pares de genes sendo definidas em termos de probabilidade de recombinações. Como instrumento de localização em um mapa genético são utilizados os biomarcadores moleculares (veja um exemplo na Figura 2. 6). Observa-se que cada cromossomo é representado por um

arranjo linear de marcadores cuja distância entre eles é dada em unidades de centimorgans (cM), ocupa uma posição fixa no cromossomo e tem um nome específico. Em sua definição mais geral, um biomarcador é qualquer "coisa" que possa ser utilizada como um indicador de algum estado biológico de um organismo. No caso de marcadores genéticos, são locos do cromossomo com posições conhecidas (em geral, em unidades de centimorgan) sobre os quais os indivíduos são genotipados (classificados de acordo com os alelos que possuírem). Os dois tipos mais comuns de marcadores são os microsatélites e os SNP (*Single Nucleotide Polymorphism*), (Speed & Zhao, 2003; Liu, 1998).

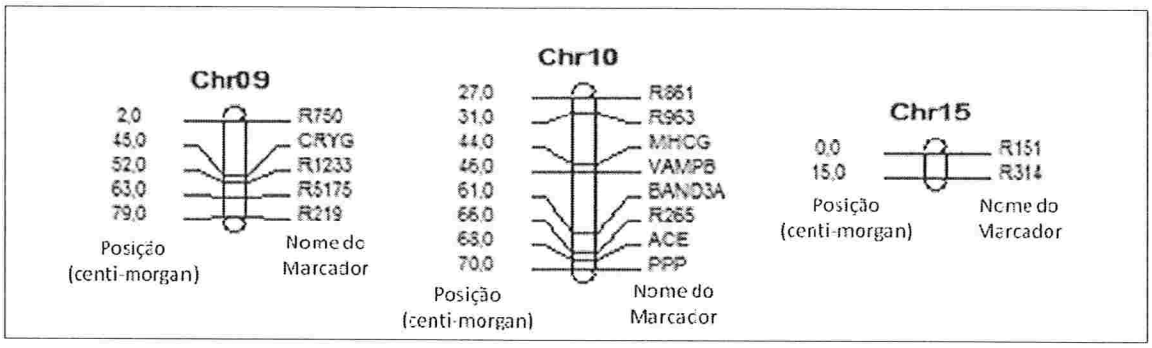

Figura 2. 6: Exemplo de um mapa de marcadores.

Os microsatélites são seqüências de não mais do que cerca de 13 bases repetidas n vezes, com n variando tipicamente entre 10-30 vezes, assim esses marcadores não costumam passar de 300 pares de bases. Um exemplo é apresentado na Figura 2. 7. Como na mitose a informação genética dos genitores é passada através de longos pedaços de DNA, a chance de que dois indivíduos parentes compartilhem o mesmo genótipo para determinado loco é muito maior do que entre indivíduos não relacionados e, por esse motivo, esse tipo de marcador é particularmente útil em delineamentos com famílias, onde o interesse é relacionar fenótipos e genótipos através de correlações familiares. Essa técnica de genotipagem é muito trabalhosa – exige que somente um marcador seja genotipado por vez por indivíduo – sendo assim uma técnica cara – e a genotipagem do marcador é obtida de forma direta por essa técnica. Ainda, os mapas de marcadores microsatélites são esparsos, com geralmente 200 marcadores para representar o genoma (Blangero et. al, 2002; Viana, 2003).

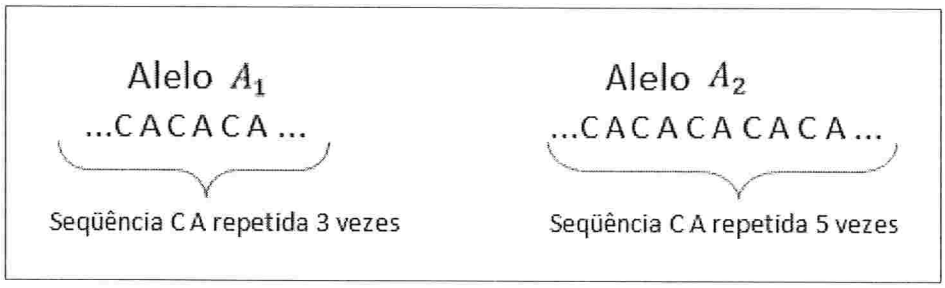

Figura 2. 7: Exemplo de um marcador microsatélite.

Os SNPs são posições em um genoma onde alguns indivíduos têm um nucleotídeo (e.g., um G) e outros têm um nucleotídeo diferente (e.g., um A), ver Figura 2. 8. Mapas desse tipo de marcadores são muito densos, sendo que existem cerca de 10 milhões de SNPs no genoma humano e atualmente estão disponíveis plataformas de SNPs com cerca de 1 milhão destes marcadores. O custo dessa técnica é relativamente baixo, uma vez que todos os marcadores são genotipados de uma só vez para cada indivíduo, o que tem levado à sua grande popularidade. Para a genotipagem são coletadas centenas de milhares de amostras do genoma (sondas) que são colocadas em um "chip" que é então hibridizado e o resultado é uma intensidade luminosa. Essa intensidade luminosa é lida por algum algoritmo de agrupamento que então assinala cada indivíduo a um genótipo e, portanto, a genotipagem se dá de forma indireta (Rapley & Harbon, 2004).

A chance de que dois indivíduos compartilhem um SNP é independente do grau de parentesco. Portanto, esses marcadores não têm sido indicados para estudos com famílias, mas têm se mostrado bastante úteis em estudos observacionais caso-controle onde indivíduos, geralmente sem grau de parentesco, são amostrados.

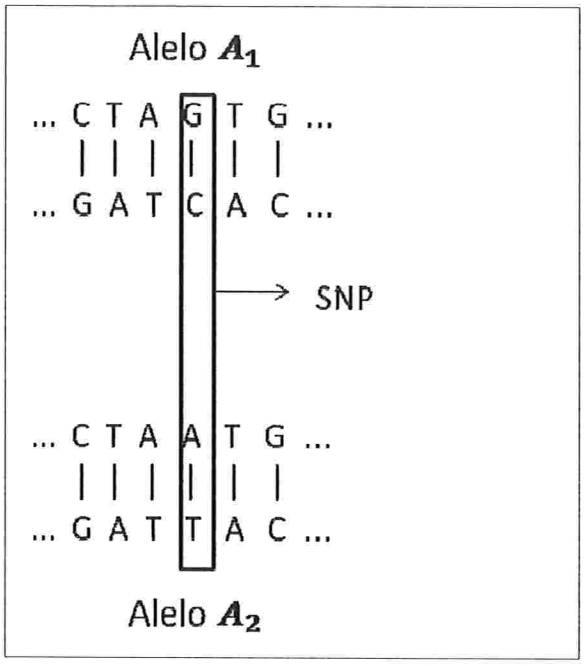

Figura 2. 8: Exemplo de um marcador SNP.

Amostragem de populações comumente utilizadas em estudos genéticos

De forma geral, podemos dividir os estudos genéticos (assim como os estudos de outras áreas) pela maneira como os dados são obtidos: populações experimentais (experimentos planejados), em que por questões éticas são utilizados animais ou plantas; e populações naturais (estudos observacionais), nos quais os estudos com humanos se encaixam.

Nas populações experimentais os cruzamentos são controlados e ocorrem entre indivíduos do mesmo grupo ou até da mesma família, bem como com indivíduos de espécies diferentes (principalmente no caso de plantas), (Liu, 1998). A maioria dos experimentos desse tipo parte de duas linhagens alternativas, P_1 (de genótipo A_1A_1 para um conjunto de locos supostamente associados com alguma característica que diferencia as linhagens) e P_2 (de genótipo A_2A_2), que cruzadas geram a geração F_1 (primeira geração de descendentes). Essa geração F_1 é composta somente por indivíduos idênticos, todos heterozigotos A_1A_2, e apresentam **desequilíbrio de ligação** completo, que é a associação probabilística de alelos, para todos os genes divergentes (dos mesmos cromossomos) entre as linhagens parentais, o que por conseqüência permite detectar **ligação** entre dois locos cromossômicos (Lynch & Walsh, 1998).

Se cruzarmos essa geração (F_1) com um de seus genitores, P_1 ou P_2, obtemos um

delineamento do tipo *Backcross*, ou retro-cruzamento. A Figura 2. 9 apresenta o esquema de geração de tal delineamento. Se o cruzamento for com P_1 então os genótipos gerados na próxima geração serão A_1A_1 e A_1A_2, caso o cruzamento seja com P_2 serão gerados os genótipos A_2A_2 e A_1A_2. Esse tipo de experimento é interessante quando o objetivo principal é a estimação do efeito aditivo entre alelos de um loco (Wu & Lin, 2009).

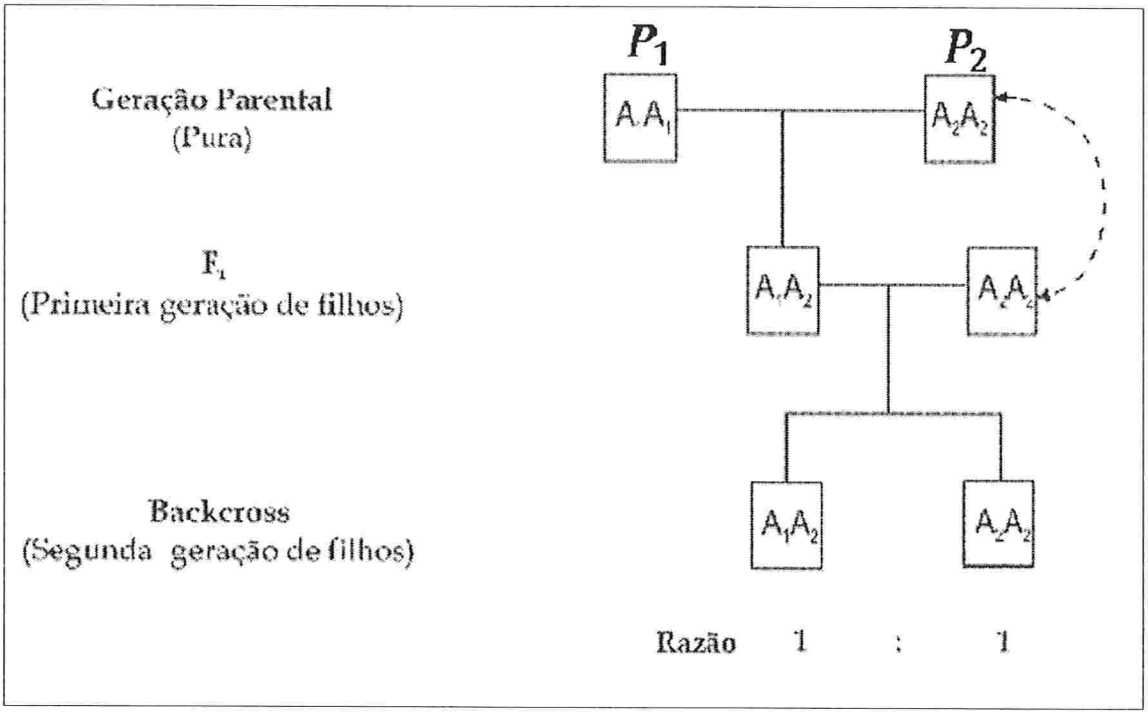

Figura 2. 9: Esquema representativo dos cruzamentos envolvidos na obtenção de uma geração de Retrocruzamento.

Um dos desenhos experimentais mais utilizados com animais é o Delineamento F_2 (Figura 2. 10). Nesse tipo de estudo a geração F_1 é cruzada entre si gerando a geração F_2 que terá, pela segunda lei de Mendel, os genótipos A_1A_1, A_2A_2, e A_1A_2 com freqüências ¼, ¼ e ½, respectivamente. Assim, esse delineamento gera três genótipos para cada loco cromossômico, o que permite a estimação dos efeitos aditivos e de dominância de um loco e o torna mais atraente do que o experimento *Backcross* para muitas situações (Lynch & Walsh, 1998).

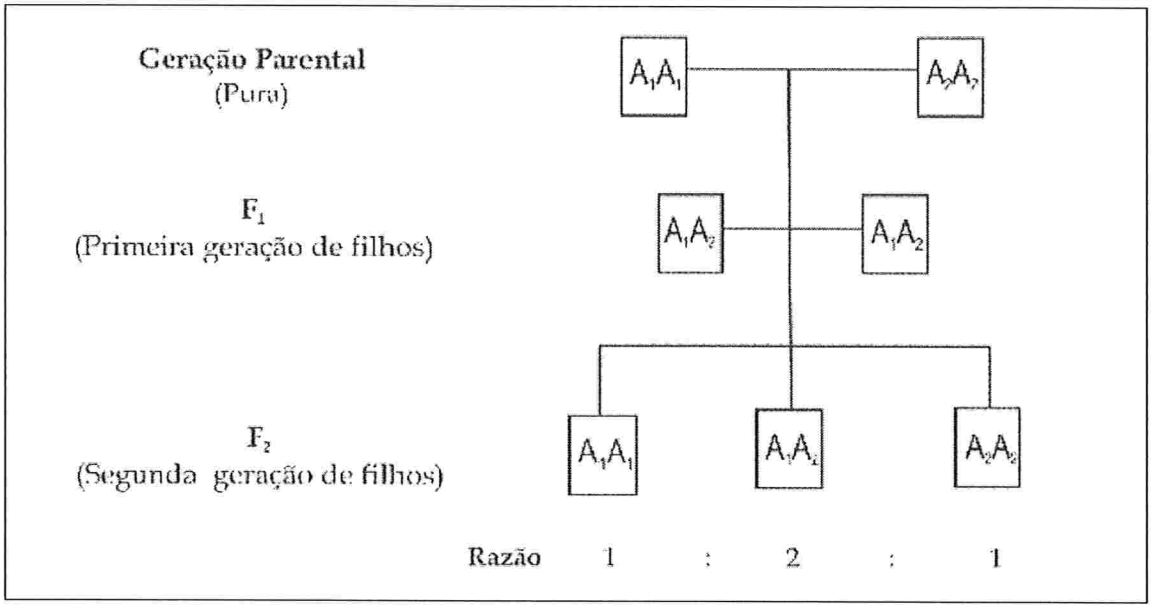

Figura 2. 10: Esquema representativo dos cruzamentos envolvidos na obtenção de uma geração F_2.

A vantagem dos cruzamentos controlados é a possibilidade da escolha dos genitores e esquema do cruzamento. Ao escolhermos pais divergentes para o fenótipo de interesse existe uma grande chance de que exista também um diferença genotípica, mesmo levando em conta as possíveis influencias do meio ambiente na característica de interesse. Além disso, esses cruzamentos criam o desequilíbrio de ligação entre locos que é necessário para a detecção dos efeitos genéticos e mapeamento de genes de todo o genoma com base na informação de um conjunto de marcadores (Liu, 1998; Doerge et al., 1997).

O contraponto para a utilização de delineamentos desses tipos é que, infelizmente, a diferença entre as características fenotípicas dos genitores reflete o efeito geral genético e não os efeitos individuais dos genes. Em termos estatísticos temos efeitos de confundimento ou colinearidade. Os mecanismos genéticos de reprodução nos fornecerão descendentes com uma nova combinação alélica, geradas por segregação independente de diferentes cromossomos e por recombinação entre cromossomos. Portanto, nos descendentes, os genes em cromossomos diferentes são variáveis aleatórias independentes (genes não ligados). Em contraste, os genes nos mesmos cromossomos serão estatisticamente dependentes ou em associação de ligação, apesar dessa dependência ser desprezível se eles estiverem suficientemente longe, e nós falamos de genes ou locos ligados. Genes não ligados serão fatores ortogonais (em populações finitas) e genes fortemente ligados serão fatores com alto grau de colinearidade. Qualquer recombinação entre dois genes próximos aumenta nossa chance de dissecar seus efeitos (Jansen, 2003).

Quando não é possível realizar cruzamentos controlados, as populações naturais são uma alternativa. Elas caracterizam-se pelo cruzamento aleatório. Nesta classe classificam-se os experimentos de acordo com o plano de amostragem. No delineamento com trios, por exemplo, são selecionadas pessoas portadoras de determinada doença e seus pais e a análise é então conduzida de forma condicional, com cada "trio" familiar sendo considerada uma unidade a parte. Nos estudos com famílias estendidas, a partir de um probando (indivíduo que "abre" a amostra de uma família), são genotipados, tanto quanto possível, todos os indivíduos vivos de uma família. Já nos estudos caso-controle são selecionados indivíduos sem relação de parentesco, alguns possuindo determinada característica e outros não e procuram-se genes ligados com essa característica.

2.3. Efeitos Genéticos

O modelo clássico genético, para doenças não mendelianas, pode ser formulado, de maneira geral, como

$$\boldsymbol{Y = G + e} \qquad (2.1)$$

em que Y representa o fenótipo (variável resposta), G é o efeito determinado por genes e e o efeito residual. Assim, o fenótipo é determinado pelos genes que o indivíduo carrega mais suas interações com o meio onde vive e demais variáveis, por exemplo, alimentação, exposição ao sol, idade, sexo e assim por diante. Além disso, é conhecido que os genes não regulam as variáveis de forma homogênea (há efeito de interação entre diferentes locos e com o meio ambiente), além dos efeitos para a maioria dos locos já identificados serem pequenos (Schuster & Cruz, 2004), o que torna os estudos de mapeamento genético extremamente complexos.

Dependendo do tipo de população que se esteja estudando, o efeito genético é modelado de diferentes maneiras. No caso das populações experimentais considera-se o efeito genético como fixo, uma vez que todos os indivíduos possuem o mesmo *background* genético e podem ser pensados como vindos de uma mesma família. No caso de um loco com 2 alelos alternativos (A_1 e A_2) temos um fator com 3 níveis (A_1A_1, A_1A_2 e A_2A_2), o que resulta em 2 graus de liberdade para o estudo de seu efeito e, portanto, podemos estimar dois contrastes ortogonais entre médias. Dois contrastes bastante utilizados em genética são os

efeitos genéticos aditivos e de dominância (Falconer & Mackay, 1996). O primeiro refere-se a um efeito que pode ser predito linearmente pela diferença entre as médias dos dois grupos homozigotos, enquanto o efeito de dominância representa um efeito que não pode ser predito linearmente, sendo o resíduo genético da interação entre os alelos A_1 e A_2. Considerando que uma população esteja classificada para certo loco segundo as classes genotípicas A_1A_1, A_1A_2 e A_2A_2, e que suas respectivas médias fenotípicas sejam $\mu_{A_1A_1}$, $\mu_{A_1A_2}$ e $\mu_{A_2A_2}$, respectivamente, então os efeitos aditivo (a) e de dominância (d) são dados pelos seguintes contrastes ortogonais (Lynch & Walsh, 1998):

$$a = \frac{\mu_{A_1A_1} - \mu_{A_2A_2}}{2}$$

$$d = \frac{(2\mu_{A_1A_2} - \mu_{A_1A_1} - \mu_{A_2A_2})}{2} = \mu_{A_1A_2} - \left(\frac{\mu_{A_1A_1} + \mu_{A_2A_2}}{2}\right).$$

Caso os dados sejam provenientes de um estudo com famílias, essas famílias possivelmente carregam diferentes *background* genéticos, portanto é natural assumir que os níveis dos fatores genéticos que atuam nos indivíduos são fatores aleatórios, isto é, os níveis do fator genético que ocorrem nas famílias amostradas são considerados como uma amostra dos possíveis níveis presentes na população (Blangero et al., 1999). Assim o modelo de componentes de variância é adequado para acomodar tal estrutura.

2.4. Estrutura de Populações

Genericamente falando, a estrutura populacional ou estratificação é a organização de uma população em sub-populações que diferem nas freqüências alélicas devido a cruzamentos não aleatórios, tamanho de população finito e/ ou barreiras geográficas. Existem diversas razões que levam a essa estratificação. Duas populações separadas por fronteiras geográficas ou culturais por muitas gerações, oscilações, mutações espontâneas, diferentes pressões de seleção, entre outros fatores, podem levar a diferenças alélicas entre indivíduos da população (Tiwari et al., 2008). Ainda, populações miscigenadas devem, provavelmente, exibir freqüências alélicas diferentes, caso esse fato esteja presente nas populações parentais (veja Ewens & Spielman, 1995).

Os estudos de associação caso-controle têm sido uma técnica bastante empregada para o mapeamento de genes associados na regulação do fenótipo de interesse. Esses estudos se baseiam na premissa de que um grupo que possui determinado fenótipo deverá ter

associado a esse fenótipo uma carga genética diferente do que aqueles que não possuem esse traço. Essa premissa é adequada em estudos em que a seleção dos indivíduos é aleatória, uma vez que o desequilíbrio de ligação entre marcadores não ligados se decompõe muito rapidamente ao longo do tempo (Pritchard & Rosenberg, 1999).

Entretanto, quando uma subdivisão da população está presente, é possível encontrar associação estatística entre o fenótipo e locos arbitrários que não têm nenhuma ligação física com o loco causal (veja Ewens & Spielman, 1995). Tais associações acontecem por causa da subdivisão da população (ou qualquer outra forma de cruzamento não aleatório) permitindo que a freqüência alélica dos marcadores varie entre os segmentos da população, como resultado da oscilação genética (Slatkin, 1991). Com isso, qualquer doença que seja mais prevalente em uma população estará em associação com quaisquer alelos que tiverem alta freqüência nessa população, causando assim uma associação espúria e efeito de confundimento. A Figura 2. 11 mostra a ocorrência do paradoxo de Sympson em um possível estudo. Nela vemos que existe diferença entre a média fenotípica em um determinado loco para os alelos A_1A_1 e A_2A_2, mas todos os indivíduos genotipados como A_1A_1 pertencem a uma população diferente dos indivíduos genotipados como A_2A_2, mostrando que o efeito desse loco na variável resposta está confundido com a variável latente que estratifica a população. É importante ressaltar que em estudos com famílias, intuitivamente, estimamos uma média para cada família. Como dentro de cada família os indivíduos têm o mesmo histórico familiar, não é necessária a correção da estrutura populacional no estudo.

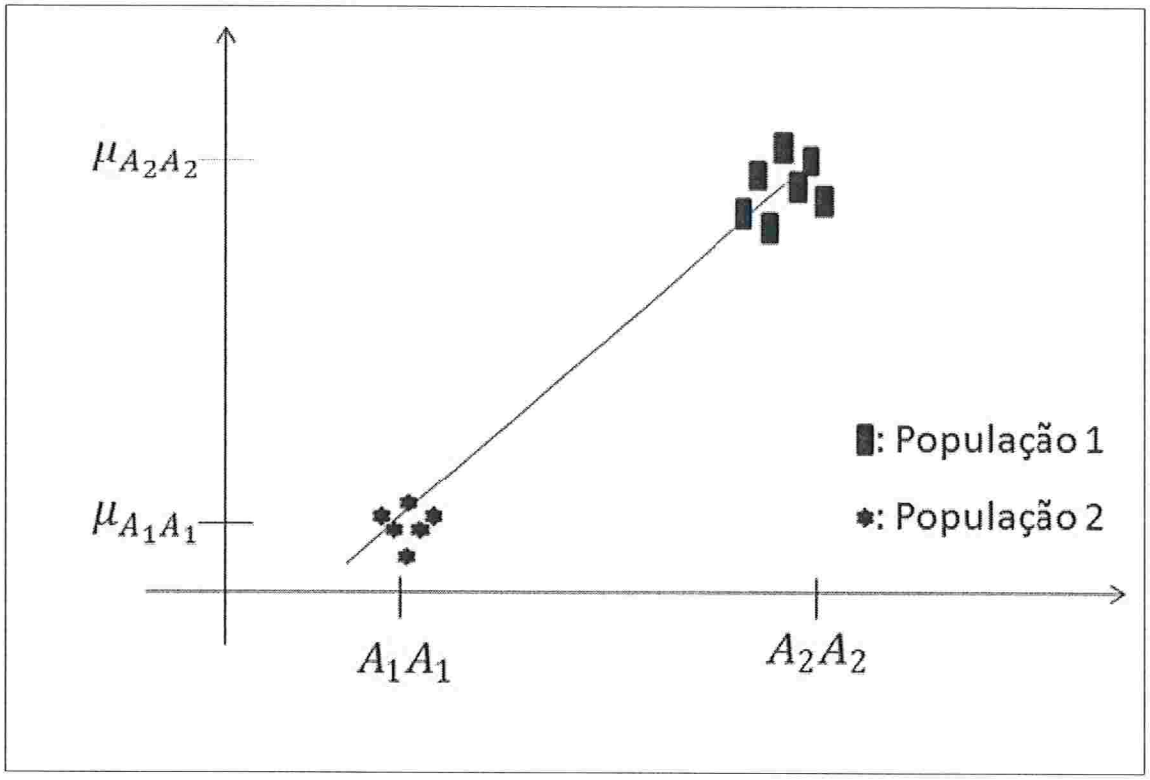

Figura 2. 11: Figura ilustrativa de um efeito de confundimento causado por estratificação da população.

Um ponto importante a respeito desse problema é que a severidade do problema de associação espúria aumenta conforme o tamanho da amostra; assim a existência de subpopulações será um problema em estudos de associação de grande porte, além do fato de que o cruzamento entre populações divergentes tem crescido rapidamente em muitas populações humanas (Pritchard & Rosenberg, 1999).

O controle genômico é uma técnica que corrige esse problema através da seleção de grupos caso e controle pareados para um conjunto de marcadores genômicos (Shmulewitz et. al., 2004; & Devlin & Roeder, 1999). Outra ferramenta, que assinala cada unidade amostral a um grupo (sub-população), também lida com o problema de estratificação (Pritchard et. al., 2000). No Capítulo 5 é descrita uma técnica que trata o problema de sub-populações utilizando componentes principais para ajustar as estatísticas de associação (Price et. al., 2006).

O reconhecimento das diferenças inter-étnicas na resposta à droga, por exemplo, é útil no delineamento e interpretação de ensaios clínicos, pois é uma fonte oculta de estrutura de populações que pode levar a associações espúrias entre genótipo e fenótipo em estudos farmacogenômicos, portanto é importantíssimo controlar esse efeito de confundimento. Os geneticistas têm mostrado que a definição genética de raça é um problema extremamente

complexo em populações miscigenadas. Características fenotípicas observáveis como cor da pele, por exemplo, não definem raça. Existem estudos mostrando que no Brasil algumas pessoas brancas apresentam ancestralidade genética semelhante a indivíduos africanos, enquanto que muitos negros assemelham-se geneticamente aos europeus (Suarez-Kurtz, 2005). Ainda, existe a questão da ancestralidade local, em que um indivíduo tem características genéticas mais semelhantes aos europeus em alguns cromossomos e aos africanos em outros. Devemos, portanto, corrigir o problema de estrutura localmente, pois uma correção global também pode levar a efeitos de confundimento.

No Brasil, e na América em geral, os ameríndios nativos, imigrantes europeus e africanos, contribuíram para a formação da população dos dias atuais. Enquanto em algumas regiões do Brasil a ancestralidade africana em brancos varia entre 13% até 32% (Parra et al., 2003), a contribuição européia em negros é estimada entre 21% e 38% e a dos ameríndios, tanto para brancos quanto para negros pode chegar até 50% nos estados do norte (Salzano & Bortolini, 2002), evidenciando que o problema de estrutura de populações no Brasil é de extrema relevância.

3. Conceitos Farmacológicos Aplicados em Genômica: Farmacogenômica

Enquanto uma terapia medicamentosa pode funcionar para a maioria dos sujeitos de uma dada população, podem existir pacientes para os quais não existe esse efeito. Ainda, alguns pacientes podem apresentar efeitos colaterais (toxicidade), enquanto em outros esse efeito não é manifestado. A Farmacogenômica é o estudo dessas variações na resposta a fármacos e na determinação de quais mutações genéticas levam a isso. Com o conhecimento de quais variações na droga levam a quais respostas dos pacientes a esperança é que os médicos possam prescrever melhores tratamentos aos seus pacientes e é a isso que chamamos de medicina personalizada (Kirk et al., 2008). A Farmacogenômica mudou a forma com que os ensaios clínicos têm sido conduzidos, tal como a genotipagem de todos os pacientes em um dado estudo ou especificando critérios de inclusão para grupos em estudos baseados em biomarcadores e também pode ser utilizada em todas as fases dos ensaios clínicos (Frost & Sulivan, 2004).

Voltemos ao modelo (2. 1), que traz o modelo clássico genético. No contexto Farmacogenômico o fenótipo representará alguma resposta do organismo a um determinado composto medicamentoso, como, por exemplo, a pressão arterial sistólica, para medicamentos relacionados à hipertensão, ou a concentração desse medicamento no plasma sanguíneo, caso o estudo seja relacionado a propriedades desse composto. Já a informação *resíduo* poderá ser decomposta em variáveis que representam a dose, o tempo entre doses, se a pessoa toma ou não medicamento, sexo, idade, entre outros. Por fim, a informação sobre o *genótipo* do indivíduo continuará dependendo do tipo de marcador (ou loco cromossômico) utilizado, e com essa variável serão construídos os estratos genéticos da população sob análise. O seguinte modelo farmacogenômico é resultante dessas adaptações (Wu & Lin, 2009):

$$\boldsymbol{Y = G + f(CA) + G * f(CA) + e} \qquad (3.\ 1)$$

em que Y representa alguma característica do fármaco (efeito, concentração, etc.), G representa o genótipo do indivíduo, $f(CA)$ representa alguma característica de aplicação desse fármaco (tais como dose, tempo após a aplicação do medicamento, etc.), $G * f(CA)$

representa interações entre o genótipo e as características de aplicação desse fármaco e *e* representa o resíduo ambiental e suas possíveis interações.

3.1. Etapas para Aprovação da Comercialização de um Medicamento

Em uma visão macro, existem duas etapas no desenvolvimento de um medicamento: a etapa pré-clínica, onde um novo composto é encontrado – a partir de algum efeito desejável que ele exerce sobre alguma função da célula – e testado em animais; e a etapa clínica, onde esse composto é testado em seres humanos (Chow & Liu, 2004).

Para que um composto seja aprovado pelos órgãos regulamentadores para ser utilizado pela população exigi-se que ele funcione adequadamente bem para o público alvo (alta eficácia), nesse caso todos os seres humanos. Muito poucos compostos atingem essa exigência (aproximadamente dois por cento de todos os compostos descobertos ou testados são aprovados como medicamentos) e uma das causas desse fato é a alta variabilidade na resposta dos indivíduos a esses compostos – funciona bem para alguns e mal para outros. Ao incluirmos a informação genética no estudo farmacológico pode-se descobrir os indivíduos para os quais o composto funciona de forma específica (estratos genéticos). Com isso, a indústria farmacêutica pode solicitar a aprovação de certo medicamento para um determinado grupo de indivíduos, aumentando assim a taxa de compostos que são aprovados para comercialização, beneficiando tanto os doentes quanto a indústria farmacêutica.

Etapa pré-clínica

Através de modelagem computacional ou entendimento dos mecanismos da doença é possível investigar a atividade biológica de uma molécula e então tentar desenvolver um composto (promissor) que afete esses mecanismos. Antes de esses compostos serem testados em seres humanos eles são utilizados em animais e um modelo animal é desenvolvido para avaliação de efeitos desejáveis (eficácia) e indesejáveis (toxicidade), onde os experimentos planejados são os mais adequados. Compostos que se mostrarem seguros e efetivos passam então a serem candidatos em estudos com humanos, após a aprovação pelos órgãos regulamentadores (no Brasil é a Agência Nacional de Vigilância Sanitária, a ANVISA). Os testes em seres humanos (etapa clínica) consistirão de 4 fases (Senn, 2007).

Etapa Clínica

Fase 1: Geralmente entre 20 a 80 pessoas (normalmente voluntários saudáveis e jovens do sexo masculino) são selecionadas para participar dos testes e a segurança e toxicidade em seres humanos é avaliada. Delineamentos planejados são os mais indicados. Nessa fase é determinada a dose na qual os efeitos tóxicos começam a aparecer. Também se inicia o estudo da farmacocinética da droga (quanto tempo demora em ser absorvida, metabolizada e excretada). Verificar se o medicamento funciona ou não e em qual dosagem são perguntas respondidas em outras fases. Entre 60-70% dos medicamentos passam da fase1 para a fase 2.

Fase 2: Determina se o componente é ativo contra a enfermidade alvo e também se tenta avaliar sua toxicidade. Novamente, os experimentos planejados são os mais adequados e procura-se estimar as curvas de dose-resposta da droga. Mais ou menos 100 pessoas são tratadas com o componente durante um período geralmente maior do que 1 ano. Um objetivo adicional é determinar uma aplicação ótima de dose-resposta. Nessa fase os critérios de inclusão continuam sendo muito rígidos. Aproximadamente 40% dos medicamentos passam da fase 2 para a fase 3.

Fase 3: Avalia o efeito da droga em uma população maior (muitas vezes entre 100 e 1000 indivíduos) e mais heterogênea, em uma tentativa de aprovar a utilização da droga proposta. Esta fase também compara a droga com os tratamentos existentes, com um placebo ou ambos. Os estudos podem envolver muitos centros clínicos e experimentos planejados duplo-cego são indicados. O objetivo é verificar a eficácia e detectar os efeitos bons e ruins, que podem não ter sido observados durante as fases 1 e 2. Quando os dados coletados são suficientes para justificar e solicitar a aprovação da droga, uma aplicação de nova droga é submetida ao órgão regulamentador.

Fase 4: Esses estudos ocorrem depois que a droga é aprovada e comercializada. Eles estão em andamento e envolvem grandes populações, por isso os estudos observacionais são normalmente utilizados. Muitas vezes, as sub-populações especiais (por exemplo, mulheres grávidas, crianças, idosos) são estudadas. A Fase 4 também inclui relatórios atualizando os efeitos adversos (Senn, 2007; Chow & Liu, 2004; Wang & Bakhai, 2006).

Uma observação interessante é que o processo desde o desenvolvimento inicial até a aprovação da droga muitas vezes leva 10 anos ou mais. Como dito anteriormente, apenas 2%

das drogas candidatas entra para estudos humanos e apenas 20% das que entram na fase 1 são aprovados para comercialização.

3.2. Estudos Farmacogenômicos

Muitas vezes, indivíduos com um metabolismo mais rápido chegam a nem usufruir dos efeitos da droga, uma vez que ela "entra" e "sai" do corpo muito rapidamente, não permitindo assim que a droga tenha tempo para agir em seu sítio alvo. Já para outros, com um metabolismo mais lento, uma pequena dose é capaz de causar efeitos tóxicos. Já é conhecido que esse e muitos outros processos fisiológicos que interferem na eficácia/ toxicidade de um medicamento são influenciados pela genética dos indivíduos. Portanto, além do aumento da taxa de fármacos aprovados para comercialização, outra utilização da farmacogenômica será no auxílio da definição da dose e dosagem para cada estrato genético. Além disso, o custo para a aprovação de medicamentos pode diminuir até 35% e o tempo de desenvolvimento ser 15% menor através da pré-seleção de candidatos aos ensaios clínicos (Frost & Sulivan, 2004). A Figura 3. 1 mostra uma população composta por indivíduos heterogêneos com relação ao efeito da droga no organismo. Por meio da farmacogenômica tenta-se estratificar (geneticamente) os indivíduos de acordo com a toxicidade e eficácia do medicamento. No caso desse exemplo os indivíduos no quadrante inferior direito (droga não toxica e benéfica) e no quadrante superior esquerdo (droga tóxica, mas benéfica) seriam indicados para receber o medicamento, prescrevendo doses mais baixas para esses últimos em razão da toxicidade.

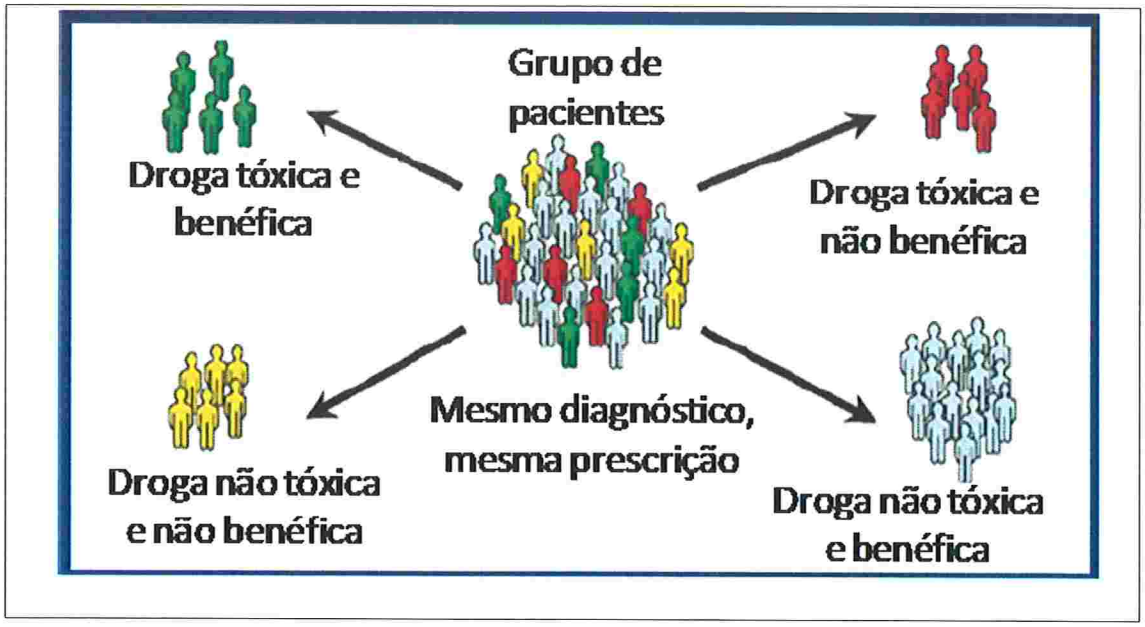

Figura 3. 1: Pacientes estratificados geneticamente de acordo com o efeito de medicamento.

Para mapear respostas a drogas, é necessário o entendimento dos mecanismos de deposição e ação da droga, e as relações entre concentração da droga e seu efeito. Propriedades farmacológicas da droga incluem interações entre os químicos e os tecidos vivos, os quais podem ser quantificados por dois diferentes, mas interativos processos bioquímicos – a **farmacocinética** (PK – do inglês pharmacokinetics) e a **farmacodinâmica** (PD – do inglês pharmacodynamics) (Derendorf & Meibohm, 1999). Simplificadamente, pode-se dizer que a farmacocinética seria o estudo do que o corpo faz com a droga, ao passo que a farmacodinâmica seria o estudo do que a droga faz com o corpo.

A intensidade da resposta terapêutica produzida por uma droga é relacionada à concentração dessa droga no sítio de ação, o qual, por sua vez, é afetado por inúmeros fatores incluindo genes e meio-ambiente. Assim, por meio da identificação desses fatores e do controle dos mecanismos e dos aspectos farmacocinéticos e farmacodinâmicos da resposta à droga, nós somos capazes de determinar o regime de dosagem para um indivíduo que seja provável de atingir a resposta terapêutica desejável com o mínimo risco de efeitos tóxicos (Wu & Li. 2009). Outra pergunta importante a ser respondida é, após a definição da dose/dosagem, qual é a eficácia e a toxicidade do medicamento. Os fenótipos analisados para se responder a essa questão são relacionados a **resultados clínicos**, os quais, novamente, poderão estar relacionados com fatores genéticos. A Figura 3. 2 esquematiza esses fenótipos.

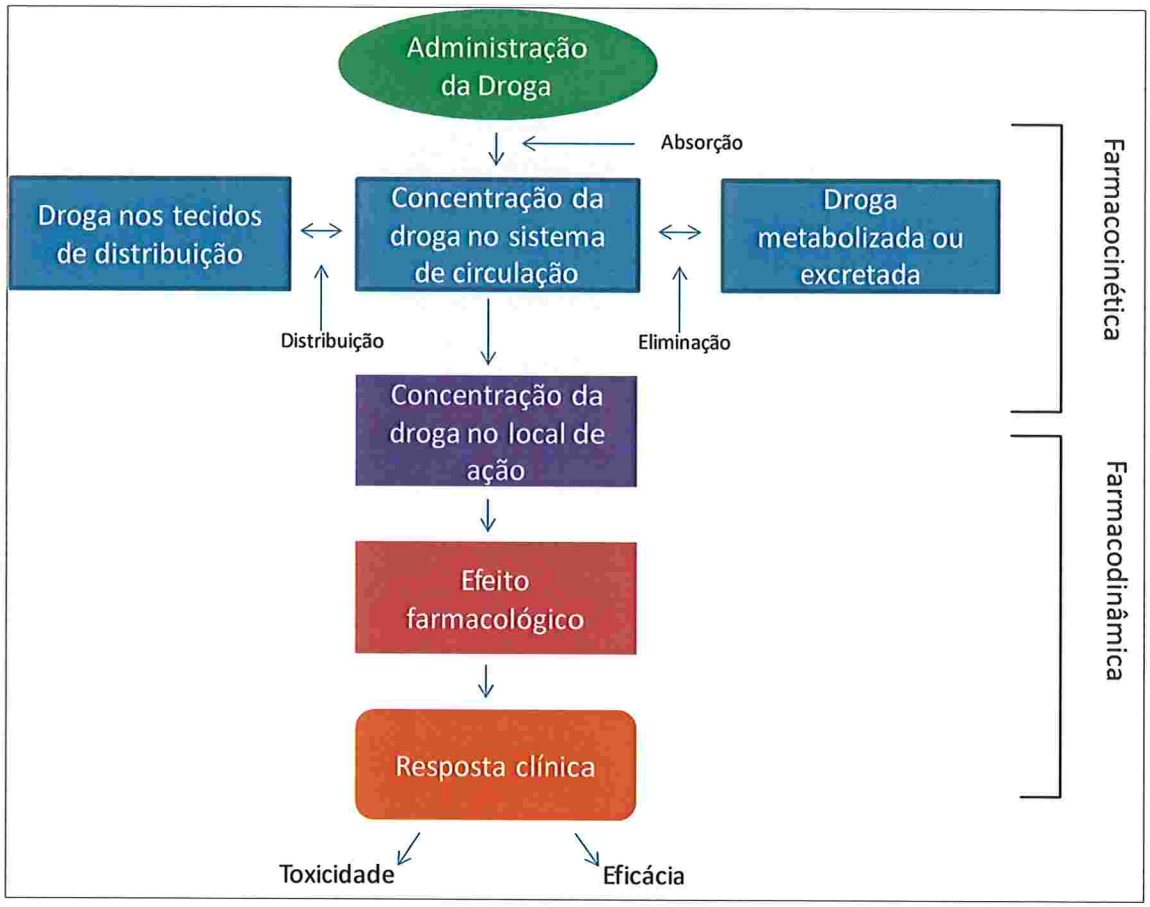

Figura 3. 2: Esquematização dos processos farmacocinéticos e farmacodinâmicos.

Farmacocinética

A farmacocinética da droga determina o início, a duração e a intensidade do seu efeito. A Figura 3. 3 mostra uma curva clássica obtida em experimentos farmacocinéticos. A farmacocinética é composta de quatro processos principais: absorção, distribuição, metabolização e eliminação. Variações nesses processos podem causar mudanças na eficácia do medicamento e uma breve descrição de cada um deles é dada a seguir (Wu & Lin, 2009).

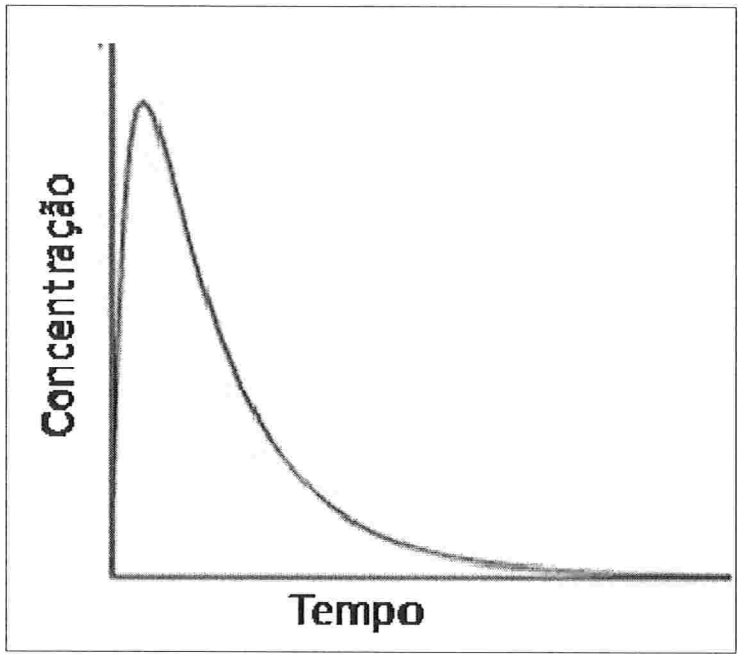

Figura 3. 3: Exemplo de uma curva representando a concentração da droga no organismo ao longo do tempo.

Absorção: é determinada por suas propriedades físico-químicas, formulação e via de administração.

Distribuição: depois que a droga entra em circulação sistêmica, ela é distribuída para os tecidos do corpo. A distribuição é geralmente irregular devido a diferenças na perfusão de sangue, vinculação do tecido (por exemplo, por causa do teor de lipídios), pH local e permeabilidade das membranas celulares.

Metabolização: o objetivo da metabolização é facilitar a excreção da droga e o fígado é o principal responsável por esse processo. Geralmente o metabolismo inativa a droga, mas alguns metabólitos da droga podem ser farmacologicamente ativos. As taxas de metabolismo das drogas variam de paciente para paciente. Os pacientes nos quais a droga é metabolizada muito rapidamente podem não obter o efeito terapêutico, já para pacientes que metabolizam a droga muito lentamente podem ter efeitos tóxicos, mesmo com doses usuais.

Eliminação: os rins e o sistema biliar são os principais responsáveis pela excreção da droga. Em geral a contribuição do intestino, da saliva, do suor, do leite materno e dos pulmões na excreção é pequena, exceção feita à exalação de anestésicos voláteis.

No caso do Captopril, medicamento hipotensor, após a administração oral de doses terapêuticas, a absorção é rápida e níveis sangüíneos máximos são atingidos em cerca de uma hora. A absorção mínima média é de aproximadamente 75% e em um período de 24 horas,

mais de 95% da dose absorvida é excretada pela urina. No entanto, em pacientes com insuficiência renal ocorre retenção de Captopril (ver a bula do Captopril).

Voltando à Equação (3. 1), no caso da farmacocinética o fenótipo avaliado será a concentração da droga, que é comumente medida no plasma sanguíneo em vários instante do tempo, resultando na seguinte equação geral (referência)

$$\boldsymbol{Y = G + f(tempo) + G * f(tempo) + e} \qquad (3.\ 2)$$

em que Y representa a concentração da droga, G o genótipo do indivíduo, que permitirá modelar interceptos diferentes para cada genótipo; $f(tempo)$ alguma função do tempo decorrido desde a aplicação de alguma dose; $G * f(tempo)$ alguma possível interação entre o genótipo e o tempo, permitindo que diferentes efeitos aconteçam para os estratos ao longo do tempo; e e representa o resíduo ambiental e suas interações.

O delineamento de estudos PK envolve a administração da droga e a avaliação da concentração e metabolização da droga ao longo do tempo a partir de fluídos biológicos tais como sangue, plasma, urina ou saliva em tempos pré-especificados (Wu e Lin, 2009). Para situações em que a concentração medida (ou concentração no plasma, c_p) é proporcional à concentração da droga no sítio de ação, temos uma ligação denominada direta e utilizamos o modelo a seguir

$$\boldsymbol{c_p} = \begin{cases} \dfrac{d}{V_d} e^{-K_e t} & \textbf{para um bólus intravenoso} \\ \dfrac{dk_a}{V_d(k_a - k_e)} \left(e^{-k_a t} - e^{-k_e t}\right) & \textbf{para uma absorção de ordem um} \\ \dfrac{k_0}{k_e V_d} \left(e^{k_e \tau} - 1\right) e^{-k_e t} & \textbf{para uma absorção de ordem zero} \end{cases} \qquad (3.\ 3)$$

onde d é a dose biodisponível, V_d é o volume de distribuição (o qual é constante para uma droga conhecida), k_e é a constante da taxa de eliminação, k_a é a constante da taxa de absorção de primeira ordem, k_0 é a constante da taxa de absorção de ordem zero, τ é a duração da absorção de ordem 0 ($\tau = t$ durante a absorção e constante na fase pós-absorção) e t é o tempo após a última dose ter sido administrada.

No caso da abordagem indireta, assume-se a existência de um compartimento hipotético e baseia-se na concentração (c_e) nesse compartimento que hipoteticamente causa o efeito da droga. Essa ligação é descrita por

$$c_e = \begin{cases} \dfrac{dk_{co}}{V_d(k_{co}-k_e)}\left(e^{-k_c t}-e^{-k_{co}t}\right) & \textbf{para um bólus intravenoso} \\ \dfrac{dk_a k_{co}}{V_d}\left[\dfrac{e^{-k_c t}}{(k_a-k_e)(k_{co}-k_e)}-\dfrac{e^{-k_a t}}{(k_e-k_a)(k_{co}-k_a)}\right. & \\ \left.+\dfrac{e^{-k_{co}t}}{(k_e-k_{co})(k_a-k_{co})}\right] & \textbf{para uma absorção de ordem um} \\ \dfrac{k_0}{k_e V_d(k_{co}-k_e)}\left[k_{co}\left(e^{k_e\tau}-1\right)e^{-k_e t}-k_c\left(e^{k_{co}\tau}-1\right)e^{-k_{co}t}\right] & \\ & \textbf{para uma absorção de ordem zero} \end{cases} \quad (3.4)$$

onde k_{co} é a constante da taxa para a distribuição da droga a partir do compartimento no qual a concentração da droga causa o efeito. A Figura 3. 4 apresenta um possível exemplo de uma análise farmacocinética/Farmacogenômica. Nela vemos três curvas diferentes para cada um dos três genótipos possíveis, que diferem na concentração máxima e no tempo de excreção pelo organismo. Métodos para estimação dessas curvas são discutidos no final dessa seção.

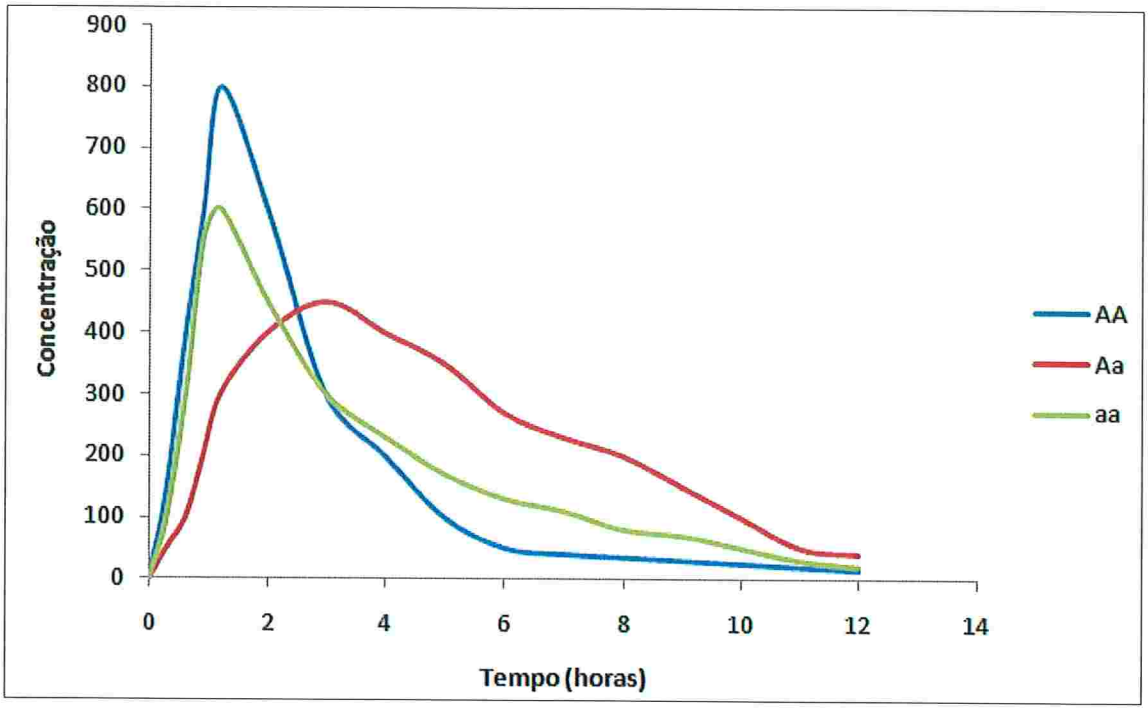

Figura 3. 4: Exemplo de três possíveis curvas de concentração para três genótipos diferentes.

Farmacodinâmica

A farmacodinâmica envolve o estudo de como a concentração da droga afeta o **alvo** da droga – molécula chave para a condição de doença ou sobrevivência de um agente patogênico microbiano. A variação genética nas moléculas chave pode causar diferenças mensuráveis na resposta à droga de um organismo.

É comum analisar o efeito da droga contra o logaritmo da concentração dessa droga no sítio de ação. A Figura 3. 5 apresenta a forma básica da curva que relaciona essas duas quantidades e percebe-se uma assíntota, chamada de E_{max} que representa o efeito máximo da droga. Outro possível parâmetro dessa curva é o EC_{50}, que é a concentração da droga que correspondente a 50% do efeito máximo.

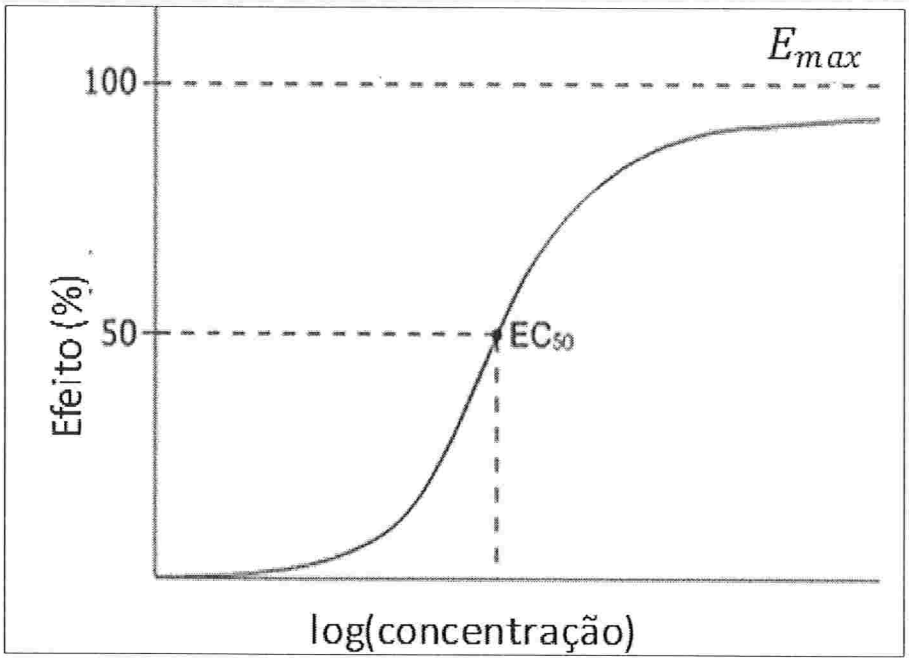

Figura 3. 5: Curva do efeito pela concentração (logaritmo), mostrando os parâmetros $\boldsymbol{EC_{50}}$ e $\boldsymbol{E_{max}}$.

O conceito de potência envolve novamente o parâmetro EC_{50} e quanto antes, ou a uma menor concentração, for obtido 50% do efeito da droga, mais potente ela será (Figura 3. 6). As drogas mais eficazes são aquelas para as quais o efeito máximo, E_{max}, é maior (Figura 3. 7).

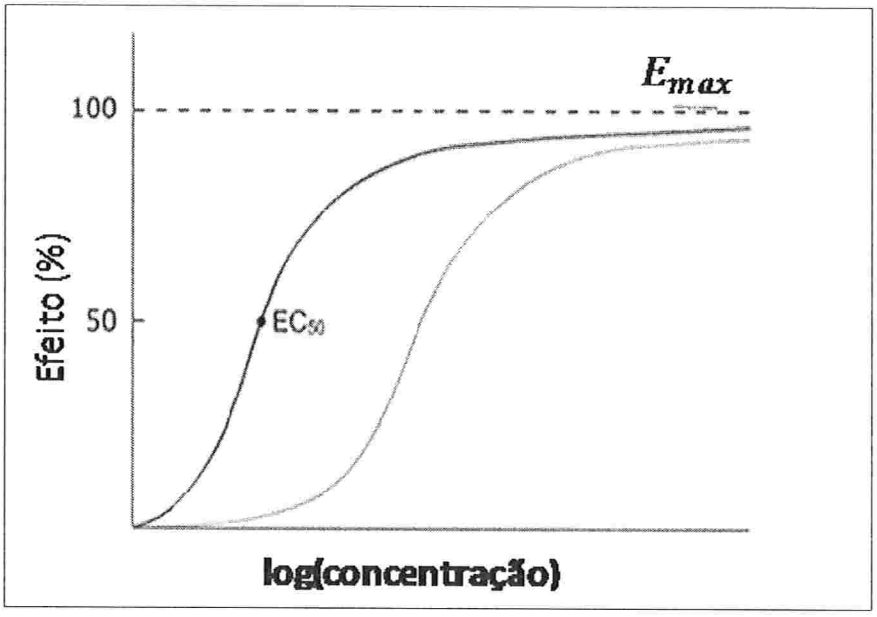

Figura 3. 6: Curva do efeito pela concentração (logaritmo), exemplificando o conceito de potência da droga.

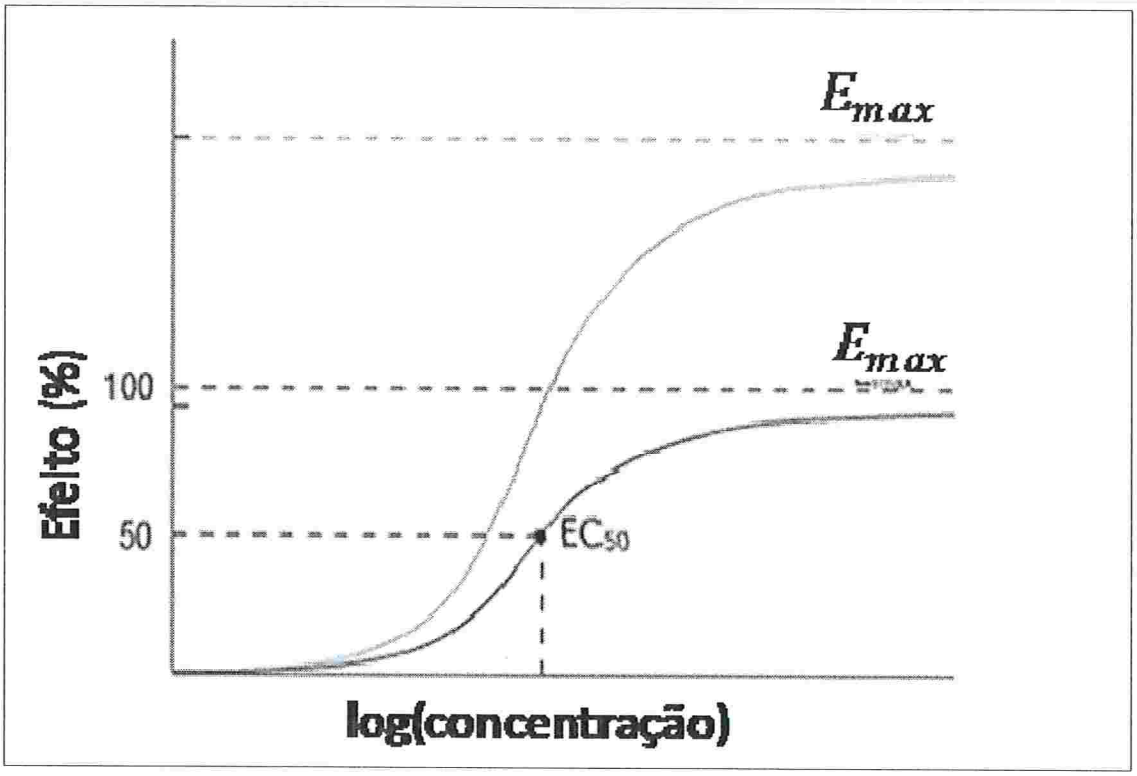

Figura 3. 7: Curva do efeito pela concentração (logaritmo), exemplificando o conceito de eficácia da droga.

Finalmente, o índice terapêutico pode ser medido como a razão do EC_{50} do efeito não desejável da droga dividido pelo EC_{50} do efeito desejável da droga. Quanto maior o índice terapêutico, melhor a droga.

Voltando a Equação (3. 1), no caso da farmacodinâmica o fenótipo avaliado será o efeito farmacológico da droga no sítio de ação da droga e como covariável utiliza-se a concentração dessa droga, resultando no seguinte modelo geral

$$\boldsymbol{Y = G + f(C) + G * f(C) + e} \qquad (3.\ 5)$$

onde Y representa o efeito, que é geralmente medido no organismo como um todo e tradicionalmente utiliza-se uma medida diretamente relacionada ao efeito da droga – para o Captopril utiliza-se, por exemplo, a concentração sérica de angiotensina II, pode-se também substituir o efeito por uma medida clínica, como, por exemplo, a pressão arterial sistólica, novamente no caso do Captopril; G representa o genótipo do indivíduo, que permitirá modelar interceptos diferentes para cada estrato populacional; $f(C)$ é alguma função da concentração;

$G * f(C)$ representa as interações entre genótipo e concentração, que permitem modelar diferentes características das curvas; e e é o residual ambiental e suas interações.

Uma forma para a função da concentração encontrada na Equação (3. 6) com parâmetros com significado biológico importante foi descrito por Giraldo (2003), dada por

$$\boldsymbol{E(c) = E_0 + \frac{E_{max} c^H}{EC_{50}^H + c^H} \approx \frac{E_{max} c^H}{EC_{50}^H + c^H}} \qquad (3.\ 6)$$

com E_{max} sendo o efeito máximo assintótico (limite), H representando o parâmetro de inclinação e EC_{50} é a concentração da droga que resulta em 50% do efeito máximo.

Um delineamento interessante para estudar a farmacodinâmica consiste na administração de doses crescentes da droga, até a dose máxima pré-determinada, através de modelos animais ou fase clínica anterior, ser alcançada.

Modelando a Farmacodinâmica e Farmacocinética Conjuntamente

Uma tendência crescente tem sido a introdução da modelagem PK/PD conjuntamente, uma abordagem na qual os dados PK e PD são gerados no mesmo estudo e, então, modelados conjuntamente para quantificar melhor a relação entre os efeitos farmacológicos da droga e sua dose assim como o tempo após a droga ter sido administrada (Hochhaus e Derendorf, 1995). O modelo geral resultante pode ser escrito como

$$\boldsymbol{Y = G + f\big(g(t)\big) + e} \qquad (3.\ 7)$$

onde $g(t)$ é uma função que relaciona a concentração e o tempo e os demais componentes estão definidos como anteriormente citado.

Em um estudo conjunto PK/PD, os efeitos da droga são medidos em uma série de instantes de tempo depois de a droga ter sido administrada com diversos níveis de dosagem (Wu e Lin, 2009). Para modelarmos os processos PK e PD conjuntamente basta substituir as equações farmacocinéticas (3. 3) ou (3. 4) no modelo farmacodinâmico, encontrado na equação (3. 6). Por exemplo, no caso um bólus intravenoso sem um compartimento de efeito, pode-se expressar o modelo PK/PD como

$$E = \frac{E_{max} d e^{-k_e t}}{EC_{50} V_d + d e^{-k_e t}} \,. \qquad (3.\ 8)$$

Estimação

Na maioria das aplicações práticas de genética encontra-se o problema da alta/baixa dimensionalidade, ou seja, alta dimensionalidade nas variáveis genotípicas (muitos marcadores) e baixa dimensionalidade nos sujeitos submetidos ao teste. Para transpor tal dificuldade pode-se utilizar um procedimento multi-estágios, por exemplo, estimar um modelo para cada marcador, colocar em um gráfico o perfil da verossimilhança encontrada e procurar por "picos" nesse gráfico.

Para a estimação desses modelos (Farmacocinético, Farmacodinâmico ou conjunto) pode-se utilizar metodologia de inferência em modelos não-lineares mistos, para acomodar a estrutura de covariância, uma vez que o estudo é feito, geralmente, com medidas repetidas. Davidian & Giltinan (1995) e Vonesh & Chinchilli (1997) apresentam detalhadamente a teoria envolvida em tais modelos. Outra possibilidade, descrita em Wu & Lin (2009) é a utilização do Mapeamento Funcional.

Respostas Clínicas

Delineamentos desse tipo são realizados quando o interesse é medir eficácia e toxicidade do medicamento na população. Voltando à Equação (3. 1), nesse caso os fenótipos avaliados serão respostas clínicas, como batimento cardíaco, para uma população de estudo, resultando no seguinte modelo geral,

$$\boldsymbol{Y = G + e} \qquad (3.\ 9)$$

em que Y representa a resposta clínica, G representa genótipo do indivíduo e e e resíduo ambiental e suas interações.

Todos os tipos de estudos farmacológicos (farmacodinâmica, farmacocinética e respostas clínicas) são de extrema relevância no desenvolvimento de medicamentos e merecem ampla discussão de métodos estatísticos para análises dos seus resultados. Nesse

trabalho, devido a disponibilidade dos dados, os fenótipos analisados nas aplicações contidas nos capítulos 5 e 6 são relativos a respostas clínicas, com objetivo principal de encontrar diferença entre a eficácia do medicamento em extratos genéticos. No capítulo 5 o estudo se encaixaria em um de fase 3 ou 4, enquanto que a aplicação do capítulo 6 seria com modelos animais.

4. Análise de Dados de Famílias

A doença arterial coronária é uma doença com pré-disposição genética clara – envolvendo interações metabólicas, neuroendócrinas e genéticas – e com participação importante de fatores ambientais (Shmulewitz et al., 2006; Wang, 2005). A herdabilidade é a quantidade de variação presente nos dados que pode ser explicada por fatores genéticos e acredita-se que ela difira para os fatores de risco cardiovasculares entre populações por causa das diferentes distribuições dos fatores de riscos ambientais, bem como das particulares constituições genéticas de diferentes populações humanas. Como a prevalência de doenças cardíacas é muito alta (cerca de 30% na maioria das populações), vários estudos foram conduzidos ao redor do mundo para estimação da herdabilidade e os resultados encontrados têm sido divergentes. Um possível motivo para as diferenças nesses resultados é a diferença na história genealógica e hábitos culturais entre as populações que têm sido utilizadas (McQueen et al, 2003).

Os dados aqui analisados pertencem a um estudo coordenado pelo Laboratório de Cardiologia e Genética Molecular do Instituto do Coração de São Paulo, InCor-USP, e tem como objetivo a caracterização de marcadores genéticos associados com a hipertensão arterial na população brasileira, que sabidamente possuí características genéticas diferentes (devido à sua miscigenação) de todas as outras. Decidiu-se por amostrar famílias na cidade mineira de Baependi, pois devido a históricos de migração e imigração acredita-se que os mineiros representem um perfil "médio" do brasileiro. Outro fator positivo da cidade de Baependi é a não existência histórica de grande migração para outras cidades, o que facilita com que um grande número de indivíduos de cada família selecionada seja estudado sem grandes custos.

Para selecionar as famílias que entrariam no estudo realizou-se uma amostragem em múltiplos estágios na cidade de Baependi, localizada na área rural do estado de Minas Gerais (752 Km^2 e 18072 habitantes), entre dezembro de 2005 e janeiro de 2006. Inicialmente foram escolhidos 11 distritos censitários, de um total de 12. A próxima etapa foi selecionar aleatoriamente residentes de cada um desses distritos – primeiramente selecionando a rua e depois o domicílio, sistematicamente. Finalmente, o critério de elegibilidade foi qualquer indivíduo do domicílio selecionado de 18 anos ou mais. Uma vez selecionado o probando, todos os seus parentes até terceiro grau foram convidados a participar do estudo, além dos respectivos cônjuges. Totalizou-se 119 famílias (1712 indivíduos) no estudo.

Os fenótipos de interesse nesse estudo são variáveis relacionadas a fatores de risco cardiovasculares, em particular, nesse trabalho, será analisada a pressão sistólica do indivíduo (*SBP*), e as covariáveis levantadas no questionário aplicado são: o gênero (*gênero*), a idade (*idade*), o índice de massa corporal (*IMC*) e se toma algum medicamento para pressão (*medicamento*), entre outras. Do ponto de vista farmacológico, essa análise poderia se encaixar em um estudo de fase 3 ou 4, onde o tipo de fenótipo analisado é uma resposta clínca (capítulo 3).

4.1. Metodologia

Para as análises dos dados foram utilizadas, principalmente, duas técnicas estatísticas: modelos mistos e gráficos da variável adicionada (Hilden-Minton, 1995; McCulloch & Searle, 2001; Searle et al., 1992; Nobre, 2004). A seguir apresentamos os principais resultados dessas técnicas.

Um modelo linear misto pode ser escrito como

$$\boldsymbol{Y_i = X_i\beta + Z_i\gamma_i + e_i,} \quad \boldsymbol{i = 1, \dots, c} \tag{4.1}$$

onde

Y_i: vetor $(n_i \times 1)$ de respostas observadas para a i-ésima unidade amostral (por exemplo, famílias);

$\boldsymbol{X_i}$: matriz $(n_i \times p)$ de especificação dos efeitos fixos para a i-ésima família;

β: vetor $(p \times 1)$ de parâmetros dos efeitos fixos;

$\boldsymbol{Z_i}$: matriz $(n_i \times q)$ de especificação dos efeitos aleatórios, conhecida e de posto completo;

γ_i: vetor $(q \times 1)$ de variáveis latentes, comumente denominados efeitos aleatórios, as quais refletem o efeito individual da i-ésima unidade amostral;

e_i: representa o vetor $(n_i \times 1)$ de efeitos residuais.

Fazendo $\boldsymbol{Y} = (\boldsymbol{Y}_1^T, \dots, \boldsymbol{Y}_c^T)^T$, $\boldsymbol{X} = (\boldsymbol{X}_1^T, \dots, \boldsymbol{X}_c^T)^T$, $\boldsymbol{Z} = diag(\boldsymbol{Z}_1^T, \dots, \boldsymbol{Z}_c^T)^T$, $\boldsymbol{\gamma} = (\gamma_1^T, \dots \gamma_c^T)^T$ e $\boldsymbol{e} = (e_1^T, \dots, e_c^T)^T$, o modelo 4.1 pode ser escrito na forma matricial como

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Z\gamma} + \boldsymbol{e}. \tag{4. 2}$$

Em geral assume-se que γ e e são não correlacionados e que $\mathbb{E}(\gamma) = \mathbb{E}(e) = \mathbf{0}$ com matriz de covariância

$$\boldsymbol{Cov}\begin{bmatrix}\boldsymbol{\gamma}\\ \boldsymbol{e}\end{bmatrix} = \begin{bmatrix}\boldsymbol{\Delta} & \mathbf{0}_{\boldsymbol{cq\times n}}\\ \mathbf{0}_{\boldsymbol{n\times cq}} & \boldsymbol{\Sigma}\end{bmatrix} \tag{4. 3}$$

com $\mathbf{0}_{(cq\times n)}$ representando uma matriz nula de ordem $(cq \times n)$ e $\boldsymbol{\Delta}$ e $\boldsymbol{\Sigma}$ são matrizes quadradas positivas definidas de ordem cq e $n = \sum_{i=1}^{c} n_i$, que representam as matrizes de covariâncias dos efeitos aleatórios γ e e, respectivamente.

No modelo (4. 2) os efeitos fixos são utilizados para modelar o valor esperado da variável resposta Y, enquanto que os efeitos aleatórios são utilizados para modelar sua estrutura de covariâncias. Usualmente assume-se que γ e e têm distribuição normal cq e n-variada, respectivamente. Fazendo $\boldsymbol{\xi} = \boldsymbol{Z}\gamma + \boldsymbol{e}$, obtêm-se

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\xi} \tag{4. 4}$$

e essas especificações implicam que $\boldsymbol{\xi}$ tem distribuição normal n-variada com vetor de médias $\mathbf{0}_{\boldsymbol{n}}$ e matriz de covariâncias

$$\boldsymbol{\Omega} = \boldsymbol{Z\Delta Z^T} + \boldsymbol{\Sigma}. \tag{4. 5}$$

Em geral, $\boldsymbol{\Delta}$ e $\boldsymbol{\Sigma}$ são (parcialmente) desconhecidas e é necessária a imposição de uma estrutura para sua estimação. Às vezes é comum colocar um parâmetro de dispersão em evidência, fazendo $\boldsymbol{\Delta} = \sigma^2\boldsymbol{D}$ e $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{R}$, com $\boldsymbol{D}$ e $\boldsymbol{R}$ denotando matrizes positivas definidas, resultando em

$$\boldsymbol{\Omega} = \boldsymbol{\sigma^2}(\boldsymbol{ZDZ^T} + \boldsymbol{R}). \tag{4. 6}$$

Diferentes estruturas podem ser encontradas na literatura para $\boldsymbol{D}$ e $\boldsymbol{R}$; veja, por exemplo, Searle et al. (1992), Singer & Andrade (2000), Wolfinger (1993).

Os modelos mistos têm se apresentado como ferramentas extremamente úteis no mapeamento genético com dados de famílias (Blangero et al., 2001). A seguir apresenta-se o modelo poligênico, uma adaptação do modelo (misto) de componentes de variância para o caso genético de dados de famílias.

Modelo Poligênico

Neste modelo não se utiliza dados de marcadores e a estrutura familiar é utilizada para modelar o efeito genético. Especifica-se o modelo poligênico por meio da formulação do seguinte modelo misto dado por (Amos, 1994; Blangero et al., 2001),

$$\boldsymbol{Y_f = X_f\beta + g_f + e_f,\ \ f = 1, \dots, F} \qquad (4.7)$$

em que Y_f, de dimensão $(n_f \times 1)$, representa o vetor de respostas fenotípicas da f-ésima família e, fazendo correspondência com o modelo (4. 1), $\boldsymbol{Z_f = I_{n_f}}$, o efeito aleatório γ é denotado por g, a soma de todos os efeitos aditivos genéticos (poligene) que induzem à estrutura de correlação entre indivíduos parentes (Fisher, 1918). Assume-se que os efeitos aleatórios g_f e e_f são não correlacionados com vetor de média zero e covariância entre as variáveis resposta para os indivíduos i e i' dada por

$$\boldsymbol{Cov(y_{fi}, y_{f'i'}) = \begin{cases} \sigma_g^2 + \sigma_e^2, & para\ i = i' \\ 2\phi_{ii'}\sigma_g^2, & para\ i \neq i'\ e\ f = f'\ (relacionados) \\ 0, & para\ i \neq i'\ e\ f \neq f'\ (não\ relacionados) \end{cases}}$$

onde $2\phi_{ii'}$ é dado por $\left(\frac{1}{2}\right)^r$, com r sendo o grau de parentesco entre dois indivíduos. Assim, a matriz de covariâncias de $Y = (Y_1^T, \dots, Y_F^T)$ é bloco diagonal dada por:

$$\boldsymbol{\Omega = 2\Phi\sigma_g^2 + I\sigma_e^2.} \qquad (4.8)$$

A dedução da forma do **melhor estimador linear não viesado** (BLUE – *best linear unbiased estimator*) de β e do **melhor preditor linear não viesado** (BLUP – *best linear unbiased estimator*) de g podem ser encontradas em Blangero et al. (2001) e McCulloch & Searle (2001). São eles

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X^T \Omega^{-1} X})^{-1} \boldsymbol{X^T \Omega^{-1} Y} \quad (4.9)$$

$$\widehat{\boldsymbol{g}} = 2\boldsymbol{\Phi}\sigma_g^2 \boldsymbol{Z^T Q Y} = 2\boldsymbol{\Phi}\sigma_g^2 \boldsymbol{Q Y} \quad (4.10)$$

em que $\boldsymbol{Q} = \boldsymbol{\Omega}^{-1} - \boldsymbol{\Omega}^{-1}\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}$, considerando $\boldsymbol{\Omega}$ conhecida ou substituída por uma estimativa.

Propriedades de $\hat{\beta}$ e $\hat{g}$ podem ser encontradas em Henderson (1975) e McCulloch & Searle (2001). Uma pergunta interessante a ser respondida pelo modelo é a proporção de variabilidade existente nos dados que pode ser explicada por fatores genéticos. Representa-se essa medida pela herdabilidade, definida pelo coeficiente de correlação intra-classe,

$$\boldsymbol{h_g^2} = \frac{\boldsymbol{\sigma_g^2}}{\boldsymbol{\sigma_g^2 + \sigma_e^2}} = \frac{\boldsymbol{\sigma_g^2}}{\boldsymbol{\sigma_y^2}} \quad . \quad (4.11)$$

Estimadores dos componentes de variância são obtidos pelo método da máxima verossimilhança restrita (Blangero et al., 2001; Searle et al, 1992). Uma hipótese de interesse é $H_0: \sigma_g^2 = 0$, ou seja, o traço em estudo não é regulado determinantemente por fatores genéticos. Testamos essa hipótese através de $H_0: \sigma_g^2 = 0 \ \ vs \ H_1: \sigma_g^2 > 0$. Sob H_0 a estatística razão de verossimilhanças é assintoticamente distribuída como uma mistura $(1/2):(1/2)$ de χ_1^2 e χ_0^2 (Self & Liang, 1987).

Gráfico da variável adicionada para efeitos fixos

Quando incluímos uma nova covariável em um modelo estatístico nós reduzimos, em menor ou maior grau, o componente aleatório desse modelo. No modelo poligênico isso significa que o quadrado médio do erro referente ao componente $\xi = g + e$ será reduzido. Se essa redução for estatisticamente significativa, teremos um dos seguintes casos:

I. A redução no componente g não é significativa, mas é significativa no componente e;

II. A redução no componente g é significativa e no componente e não;

III. Tanto a redução no componente g quanto a redução no componente e são significativas.

Quando acontece o primeiro caso, concluímos que a covariável incluída está correlacionada com o resíduo ambiental e não está correlacionada com o componente genético, resultando assim no aumento do valor da herdabilidade (Equação (4. 11)). Quando o segundo caso se apresenta, tem-se a indicação de que a covariável adicionada está correlacionada somente com o efeito genético e o resultado é a diminuição da herdabilidade. No terceiro caso existe correlação entre a covariável com ambos os efeitos aleatórios. O gráfico da variável adicionada permitirá uma avaliação **descritiva** dessas situações.

Suponha que $\boldsymbol{X_1}$ seja a matriz de covariáveis fixas já presentes no modelo e o interesse seja estudar a influência que a inclusão da covariável X_2 exerce nos componentes de variância do modelo. O modelo restrito pode ser escrito como

$$\boldsymbol{Y = X_1\beta_1 + Zg + e = X_1\beta_1 + \xi,} \qquad (4.\ 12)$$

enquanto que o modelo irrestrito tem a forma

$$\boldsymbol{Y = X_1\beta_1^* + X_2\beta_2 + Zg^* + e^* = X_1\beta_1^* + X_2\beta_2 + \xi^*} \qquad (4.\ 13)$$

e a pergunta que queremos responder é se a diferença entre $\boldsymbol{\xi}$ e $\boldsymbol{\xi^*}$ é "grande". É possível demonstrar (veja, por exemplo, Nobre (2004); Hilden-Minton (1995)) que o BLUE de β_2 no modelo irrestrito é dado por

$$\boldsymbol{\widehat{\beta}_2 = \left(X_2^T Q_1 X_2\right)^{-1} X_2^T Q_1 Y}\ , \qquad (4.\ 14)$$

em que $\boldsymbol{Q_1 = \Omega^{-1} - \Omega^{-1}X_1\left(X_1^T\Omega^{-1}X_1\right)^{-1}X_1^T\Omega^{-1}}$.

Seja $e(x_2|x_1)$ o resíduo da regressão em que X_2 é a variável resposta e $\boldsymbol{X_1}$ a preditora, que representa a parte da variabilidade de $\boldsymbol{X_2}$ que não pode ser explicada por $\boldsymbol{X_1}$. O resíduo $e(y|x_1)$ é definido como o resíduo da regressão quando $\boldsymbol{Y}$ é a variável resposta e $\boldsymbol{X_1}$ a preditora, que representa a quantidade da variabilidade de $\boldsymbol{Y}$ que não pode ser explicada por $\boldsymbol{X_1}$. É direta a conclusão de que se existir correlação entre $e(x_2|x_1)$ e $e(y|x_1)$, então existe "informação" em $\boldsymbol{X_2}$ que não existe em $\boldsymbol{X_1}$ e que ajuda a explicar Y e, portanto, o modelo irrestrito é o mais adequado.

Em (Nobre 2004) mostra-se que os resíduos $e(x_2|x_1)$ e $e(y|x_1)$ podem ser escritos,

respectivamente, como

$$\boldsymbol{e(x_2|x_1) = \Omega^{1/2} Q_1 X_2} \tag{4. 15}$$

e

$$\boldsymbol{e(y|x_1) = \Omega^{1/2} Q_1 Y.} \tag{4. 16}$$

O gráfico da variável adicionada é definido como o gráfico de dispersão dos resíduos $e(x_2|x_1)$ (eixo horizontal) e $e(y|x_1)$ (eixo vertical). Adiciona-se ainda uma reta da regressão sem intercepto de $e(y|x_1)$ como variável resposta e $e(x_2|x_1)$ como preditora. Caso a inclinação dessa reta seja significativamente diferente de 0, ou seja, diferente da reta horizontal, então existe evidência de que a covariável X_2 deve ser incluída no modelo. É conhecido da teoria de modelos lineares (Graybill, 1976) que a reta de regressão sem o intercepto é estimada por

$$\begin{aligned}
&[e(x_2|x_1)^T e(x_2|x_1)]^{-1} e(x_2|x_1)^T e(y|x_1) \\
&\qquad = \left(X_2^T Q_1^T \left(\Omega^{\frac{1}{2}}\right)^T \Omega^{\frac{1}{2}} Q_1 X_2 \right)^{-1} X_2^T Q_1^T \left(\Omega^{\frac{1}{2}}\right)^T \Omega^{\frac{1}{2}} Q_1 Y \\
&\qquad = \left(X_2^T Q_1 \Omega Q_1 X_2\right)^{-1} X_2^T Q_1 \Omega Q_1 Y \\
&\qquad = \left(X_2^T Q_1 X_2\right)^{-1} X_2^T Q_1 Y = \widehat{\beta}_2
\end{aligned} \tag{4. 17}$$

pois $\boldsymbol{Q_1 \Omega Q_1 = Q_1}$ e, portanto, concluímos que o estimador do coeficiente de inclinação da regressão de $e(x_2|x_1)$ sobre $e(y|x_1)$ é igual ao BLUE de β_2 no modelo irrestrito (ver Equação (4. 14)). A utilidade do gráfico da variável adicionada vem dessa e de outras propriedades notáveis (Friendly, 1991), por exemplo:

I. O resíduo da reta de regressão desse gráfico é idêntico ao resíduo de Y no modelo irrestrito;

II. A correlação simples entre $e(x_2|x_1)$ e $e(y|x_1)$ é igual à correlação parcial entre Y e X_2 corrigida para as variáveis contidas em $\boldsymbol{X_1}$;

III. Os valores $[e_i(x_2|x_1)]^2$, resíduo referente a i-ésima unidade amostral, são proporcionais à alavanca parcial adicionada em h_{ii}, o i-ésimo elemento da diagonal da matriz de projeção, também conhecida como matriz *hat* (Cook & Weisberg, 1981), pela inclusão de X_2 no modelo. Observações de alta alavancagem em X_2 são aquelas com valores extremos em $e_i(x_2|x_1)$.

Decomposição do gráfico da variável adicionada

Quando trabalhamos com um modelo misto poligênico (de componentes de variância) temos dois efeitos aleatórios e é interessante estudar como a inclusão da covariável $\boldsymbol{X_2}$ influência cada um desses efeitos. Hilden-Minton (1995) propõe a decomposição do gráfico da variável adicionada através da substituição, nas Equações (4. 15) e (4. 16), de $\Omega^{1/2}$ por uma matriz $\boldsymbol{A}$ de dimensão $(n+n)\times n$ tal que $\boldsymbol{\Omega = A^T A}$. Em particular, ele sugere utilizar

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{I_n}(\boldsymbol{h_e^2})^{\frac{1}{2}} \\ \left(\boldsymbol{h_g^2 2\Phi}\right)^{\frac{1}{2}} \end{bmatrix}. \qquad (4.\ 18)$$

O gráfico da variável adicionada decomposto corresponde ao gráfico de dispersão dos componentes $\boldsymbol{AQ_1}Y$ contra $\boldsymbol{AQ_1}X_2$, que são

$$\boldsymbol{A}\times[\boldsymbol{Q_1Y}\quad \boldsymbol{Q_1X_2}] = \begin{bmatrix} \boldsymbol{e^e(y|x_1)} & \boldsymbol{e^e(x_2|x_1)} \\ \boldsymbol{e^g(y|x_1)} & \boldsymbol{e^g(x_2|x_1)} \end{bmatrix}$$
$$= \begin{bmatrix} \left(\frac{\sigma_e^2}{\sigma_y^2}\right)^{1/2}\boldsymbol{Q_1Y} & \left(\frac{\sigma_e^2}{\sigma_y^2}\right)^{1/2}\boldsymbol{Q_1X_2} \\ \left(\frac{\sigma_g^2}{\sigma_y^2}2\Phi\right)^{1/2}\boldsymbol{Q_1Y} & \left(\frac{\sigma_g^2}{\sigma_y^2}2\Phi\right)^{1/2}\boldsymbol{Q_1X_2} \end{bmatrix}. \qquad (4.\ 19)$$

O gráfico de $e^e(x_2|x_1)$ versus $e^e(y|x_1)$ é chamado por Hilden-Minton (1995) de **gráfico da variável adicionada intra-unidades** ("within-unit"), ou intra-famílias no caso do modelo poligênico, e leva em conta a alteração no componente residual. O gráfico de $e^g(x_2|x_1)$ versus $e^g(y|x_1)$ é denominado de **gráfico da variável adicionada entre - unidades** ("between-unit"), ou entre - famílias, e leva em conta a mudança no componente genético do modelo. Os "novos" coeficientes de regressão, estimados pelo modelo de regressão sem

intercepto, de $e^e(x_2|x_1)$ sobre $e^e(y|x_1)$ e de $e^g(x_2|x_1)$ sobre $e^g(y|x_1)$ são, respectivamente,

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_2^e &= [\boldsymbol{e}^e(\boldsymbol{x}_2|\boldsymbol{x}_1)^T\boldsymbol{e}^e(\boldsymbol{x}_2|\boldsymbol{x}_1)]^{-1}\boldsymbol{e}^e(\boldsymbol{x}_2|\boldsymbol{x}_1)^T\boldsymbol{e}^e(\boldsymbol{y}|\boldsymbol{x}_1)\\ &= \left(\boldsymbol{X}_2^T\boldsymbol{Q}_1\boldsymbol{h}_e^2\boldsymbol{Q}_1\boldsymbol{X}_2\right)^{-1}\boldsymbol{X}_2^T\boldsymbol{Q}_1\boldsymbol{h}_e^2\boldsymbol{Q}_1\boldsymbol{Y} \qquad (4.20)\\ &= \left(\boldsymbol{X}_2^T\boldsymbol{Q}_1\boldsymbol{\sigma}_e^2\boldsymbol{Q}_1\boldsymbol{X}_2\right)^{-1}\boldsymbol{X}_2^T\boldsymbol{Q}_1\boldsymbol{\sigma}_e^2\boldsymbol{Q}_1\boldsymbol{Y}\end{aligned}$$

e

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_2^g &= [\boldsymbol{e}^g(\boldsymbol{x}_2|\boldsymbol{x}_1)^T\boldsymbol{e}^g(\boldsymbol{x}_2|\boldsymbol{x}_1)]^{-1}\boldsymbol{e}^g(\boldsymbol{x}_2|\boldsymbol{x}_1)^T\boldsymbol{e}^g(\boldsymbol{y}|\boldsymbol{x}_1)\\ &= \left(\boldsymbol{X}_2^T\boldsymbol{Q}_1\boldsymbol{2\Phi h}_g^2\boldsymbol{Q}_1\boldsymbol{X}_2\right)^{-1}\boldsymbol{X}_2^T\boldsymbol{Q}_1\boldsymbol{2\Phi h}_g^2\boldsymbol{Q}_1\boldsymbol{Y} = \qquad (4.21)\\ &\left(\boldsymbol{X}_2^T\boldsymbol{Q}_1\boldsymbol{2\Phi\sigma}_g^2\boldsymbol{Q}_1\boldsymbol{X}_2\right)^{-1}\boldsymbol{X}_2^T\boldsymbol{Q}_1\boldsymbol{2\Phi\sigma}_g^2\boldsymbol{Q}_1\boldsymbol{Y}.\end{aligned}$$

Note que,

$$\begin{aligned}&\boldsymbol{X}_2^T\boldsymbol{Q}_1\boldsymbol{\sigma}_e^2\boldsymbol{Q}_1\boldsymbol{Y} + \boldsymbol{X}_2^T\boldsymbol{Q}_1\boldsymbol{2\Phi\sigma}_g^2\boldsymbol{Q}_1\boldsymbol{Y}\\ &= \boldsymbol{X}_2^T\boldsymbol{Q}_1\left(\boldsymbol{I}_n\boldsymbol{\sigma}_e^2 + \boldsymbol{2\Phi\sigma}_g^2\right)\boldsymbol{Q}_1\boldsymbol{Y} \qquad (4.22)\\ &= \boldsymbol{X}_2^T\boldsymbol{Q}_1\boldsymbol{\Omega}\boldsymbol{Q}_1\boldsymbol{Y} = \boldsymbol{X}_2^T\boldsymbol{Q}_1\boldsymbol{Y},\end{aligned}$$

e que,

$$\boldsymbol{X}_2^T\boldsymbol{Q}_1\boldsymbol{2\Phi\sigma}_g^2\boldsymbol{Q}_1\boldsymbol{X}_2 + \boldsymbol{X}_2^T\boldsymbol{Q}_1\boldsymbol{\sigma}_e^2\boldsymbol{Q}_1\boldsymbol{X}_2 = \boldsymbol{X}_2^T\boldsymbol{Q}_1\boldsymbol{X}_2. \qquad (4.23)$$

Com isso podemos escrever o BLUE de β_2 como

$$\widehat{\boldsymbol{\beta}}_2 = \frac{\widehat{\boldsymbol{\beta}}_2^e[\boldsymbol{e}^e(\boldsymbol{x}_2|\boldsymbol{x}_1)^T\boldsymbol{e}^e(\boldsymbol{x}_2|\boldsymbol{x}_1)] + \widehat{\boldsymbol{\beta}}_2^g[\boldsymbol{e}^g(\boldsymbol{x}_2|\boldsymbol{x}_1)^T\boldsymbol{e}^g(\boldsymbol{x}_2|\boldsymbol{x}_1)^T]}{[\boldsymbol{e}^e(\boldsymbol{x}_2|\boldsymbol{x}_1)^T\boldsymbol{e}^e(\boldsymbol{x}_2|\boldsymbol{x}_1)] + [\boldsymbol{e}^g(\boldsymbol{x}_2|\boldsymbol{x}_1)^T\boldsymbol{e}^g(\boldsymbol{x}_2|\boldsymbol{x}_1)^T]}. \qquad (4.24)$$

Notamos, ainda, que

$$[e^e(x_2|x_1)^T e^e(x_2|x_1)] + [e^g(x_2|x_1)^T e^g(x_2|x_1)^T]$$

$$= \left[X_2^T Q_1 h_e^2 Q_1 X_2\right] + \left[X_2^T Q_1 2\Phi h_g^2 Q_1 X_2\right]$$

$$= X_2^T Q_1 \left[I \times h_e^2 + 2\Phi h_g^2\right] Q_1 X_2^T \qquad (4.\ 25)$$

$$= \frac{1}{\sigma^2} X_2^T Q_1 \Omega Q_1 X_2^T = \frac{1}{\sigma^2} e(x_2|x_1)^T e(x_2|x_1).$$

Podemos definir o seguinte modelo linear geral:

$$e_i(y|x_1) = e_i(x_2|x_1)\,\beta_2 + \epsilon_i \qquad (4.\ 26)$$

com ϵ_i variáveis aleatórias independentes e normalmente distribuídas de média 0 e variância σ^2. Graybill (1976) mostra que a variância desse modelo condicional pode ser calculada por

$$V\left[\hat{\beta}_2\right] = V\left[\hat{\beta}_2\right]_{e(y|x_1)|e(x_2|x_1)} = \sigma^2[e(x_2|x_1)^T e(x_2|x_1)]^{-1}. \qquad (4.\ 27)$$

Portanto, o denominador da Equação (4. 24) é igual ao inverso da variância de $\hat{\beta}_2$, quando esse é estimado pelo gráfico da variável adicionada. Note ainda que $V\left[\hat{\beta}_2^e\right] = 1/\sigma^2[e^e(x_2|x_1)^T e^e(x_2|x_1)]$ e que $V\left[\hat{\beta}_2^g\right] = 1/\sigma^2[e^g(x_2|x_1)^T e^g(x_2|x_1)^T]$, também quando esses são calculados pelas correspondentes decomposições dos gráficos da variável adicionada. Substituindo esse resultado na Equação (4. 24) tem-se,

$$\hat{\beta}_2 = \frac{\hat{\beta}_2^e V\left[\hat{\beta}_2^e\right]^{-1} + \hat{\beta}_2^g V\left[\hat{\beta}_2^g\right]^{-1}}{V\left[\hat{\beta}_2\right]^{-1}} = p_e \hat{\beta}_2^e + p_g \hat{\beta}_{2,g}\ , \qquad (4.\ 26)$$

em que os pesos p_e e p_g são proporções entre a variância total e a "decomposta".

Além de descobrir qual é o efeito da inclusão de X_2 em cada componente aleatório do modelo, pode ser interessante estudar se esse efeito é igual para todas as famílias. Suponhamos, por exemplo, que a inclusão de uma nova covariável, a qual indica se a pessoa toma ou não um determinado medicamento para hipertensão, causa uma diminuição em σ_g^2.

Vimos anteriormente que isso indica que essa covariável, que denominaremos

medicamento, está correlacionada com a variável resposta pressão sistólica. Essa correlação se manifestará através da significância do parâmetro $\hat{\beta}_2$, na Equação (4. 13), que aumentará ou diminuirá a estimativa da pressão média dos indivíduos. Uma importante questão médica seria responder se esse efeito é ou não o mesmo para todas as famílias. Para responder tal pergunta podemos decompor ainda mais o gráfico da variável adicionada decompondo-o em efeitos familiares.

Trabalhando primeiramente com o efeito aleatório genético, gráfico entre - unidades, ordenamos os vetores $e^g(x_2|x_1)$ e $e^g(y|x_1)$ de tal maneira que os seus n_1 primeiros componentes correspondam à família codificada como família 1, os n_2 componentes seguintes correspondam à família 2 e assim por diante. O particionamento obtido vem a seguir:

$$\boldsymbol{e^g(x_2|x_1)^T} = \left[\boldsymbol{e^g_{11}(x_2|x_1)}, \dots, \boldsymbol{e^g_{1n_1}(x_2|x_1)}, \dots, \boldsymbol{e^g_{F1}(x_2|x_1)}, \dots, \boldsymbol{e^g_{Fn_F}(x_2|x_1)}\right) \boldsymbol{e} \tag{4. 27}$$

$$\boldsymbol{e^g(y|x_1)^T} = \left[\boldsymbol{e^g_{11}(y|x_1)}, \dots, \boldsymbol{e^g_{1n_1}(y|x_1)}, \dots, \boldsymbol{e^g_{F1}(y|x_1)}, \dots, \boldsymbol{e^g_{Fn_F}(y|x_1)}\right], \tag{4. 28}$$

em que, por exemplo, $e^g_{ij}(x_2|x_1)$ é o j-ésimo elemento da i-ésima família. Utilizaremos, ainda, a notação $e^g_i(x_2|x_1)^T = \left[e^g_{i1}(x_2|x_1), \dots, e^g_{in_i}(x_2|x_1)\right]$. É fácil perceber que o coeficiente de regressão do gráfico da variável adicionada decomposto referente somente à i-ésima família é dado por

$$\widehat{\boldsymbol{\beta}}^g_{2i} = \left[\boldsymbol{e^g_i(x_2|x_1)^T e^g_i(x_2|x_1)}\right]^{-1} \boldsymbol{e^g_i(x_2|x_1)^T e^g_i(y|x_1)}. \tag{4. 29}$$

É imediata a conclusão de que

$$\widehat{\boldsymbol{\beta}}_2^g = \frac{\sum_{i=1}^{F} \widehat{\boldsymbol{\beta}}_{2i}^g \left[\boldsymbol{e}_i^g(x_2|x_1)^T \boldsymbol{e}_i^g(x_2|x_1) \right]}{\sum_{i=1}^{F} \left[\boldsymbol{e}_i^g(x_2|x_1)^T \boldsymbol{e}_i^g(x_2|x_1) \right]}$$

$$= \frac{\sum_{i=1}^{F} \widehat{\boldsymbol{\beta}}_{2i}^g \left[\boldsymbol{e}_i^g(x_2|x_1)^T \boldsymbol{e}_i^g(x_2|x_1) \right]}{\boldsymbol{e}^g(x_2|x_1)^T \boldsymbol{e}^g(x_2|x_1)} = \sum_{i=1}^{F} \widehat{\boldsymbol{\beta}}_{2i}^g \boldsymbol{w}_{g,i}, \tag{4. 30}$$

em que, novamente,

$$\boldsymbol{w}_{g,i} = \frac{\boldsymbol{V}\left[\widehat{\boldsymbol{\beta}}_{2,i}^g\right]^{-1}}{\boldsymbol{V}\left[\widehat{\boldsymbol{\beta}}_2^g\right]^{-1}}, \tag{4. 31}$$

ou seja, $\hat{\beta}_2^g$ pode ser decomposto em uma combinação linear de efeitos correspondentes a cada família. Por simetria, chega-se em

$$\widehat{\boldsymbol{\beta}}_2^e = \frac{\sum_{i=1}^{F} \widehat{\boldsymbol{\beta}}_{2i}^e \left[\boldsymbol{e}_i^e(x_2|x_1)^T \boldsymbol{e}_i^e(x_2|x_1) \right]}{\sum_{i=1}^{F} \left[\boldsymbol{e}_i^e(x_2|x_1)^T \boldsymbol{e}_i^e(x_2|x_1) \right]}$$

$$= \frac{\sum_{i=1}^{F} \widehat{\boldsymbol{\beta}}_{2i}^e \left[\boldsymbol{e}_i^e(x_2|x_1)^T \boldsymbol{e}_i^e(x_2|x_1) \right]}{\boldsymbol{e}^e(x_2 x_1)^T \boldsymbol{e}^e(x_2|x_1)} = \sum_{i=1}^{F} \widehat{\boldsymbol{\beta}}_{2i}^e \boldsymbol{w}_{e,i}, \tag{4. 32}$$

com

$$\boldsymbol{w}_{e,i} = \frac{\boldsymbol{V}\left[\widehat{\boldsymbol{\beta}}_{2,i}^e\right]^{-1}}{\boldsymbol{V}\left[\widehat{\boldsymbol{\beta}}_2^e\right]^{-1}}. \tag{4. 33}$$

Finalmente, reconstruindo $\hat{\beta}_2$, tem-se

$$\widehat{\boldsymbol{\beta}}_2 = \boldsymbol{p}_g \times \widehat{\boldsymbol{\beta}}_2^g + \widehat{\boldsymbol{\beta}}_2^e$$

$$= \boldsymbol{p}_g \times \sum_{i=1}^{F} \widehat{\boldsymbol{\beta}}_{2,i}^g \boldsymbol{w}_{g,i} + \boldsymbol{p}_e \times \sum_{i=1}^{F} \widehat{\boldsymbol{\beta}}_{2i}^e \boldsymbol{w}_{e,i} \tag{4. 34}$$

$$= \frac{V[\hat{\beta}_2^g]^{-1}}{V[\hat{\beta}_2]^{-1}} \times \sum_{i=1}^{F} \hat{\beta}_{2,i}^g \frac{V[\hat{\beta}_{2,i}^g]^{-1}}{V[\hat{\beta}_2^g]^{-1}} + \frac{V[\hat{\beta}_2^e]^{-1}}{V[\hat{\beta}_2]^{-1}} \times \sum_{i=1}^{F} \hat{\beta}_{2i}^e \frac{V[\hat{\beta}_{2,i}^e]^{-1}}{V[\hat{\beta}_2^e]^{-1}}.$$

Ressalta-se que o efeito de família é de maior interesse no componente β_2^g e não no componente β_2^e.

4.2. Aplicação: Projeto Corações de Baependi, Minas Gerais

Nesta seção são apresentados os resultados da aplicação das técnicas descritas anteriormente aos dados das famílias de Baependi. Essa aplicação se encaixa em estudo de fase 3 ou 4, onde o principal objetivo seria a avaliação de eficácia do medicamento. Para obtenção dos resultados utilizou-se o aplicativo SAS (programas no apêndice).

Com o objetivo de correção da curtose da distribuição do *SBP*, aplicou-se a transformação logarítmica, em particular, $7lnSBP = 7 * \ln(SBP)$. A densidade estimada por kernel (ver, Silverman (1986), para maiores detalhes) dessa nova variável é apresentada na Figura 4. 1 e pode-se perceber que a hipótese de que a variável transformada apresenta distribuição simétrica é razoável.

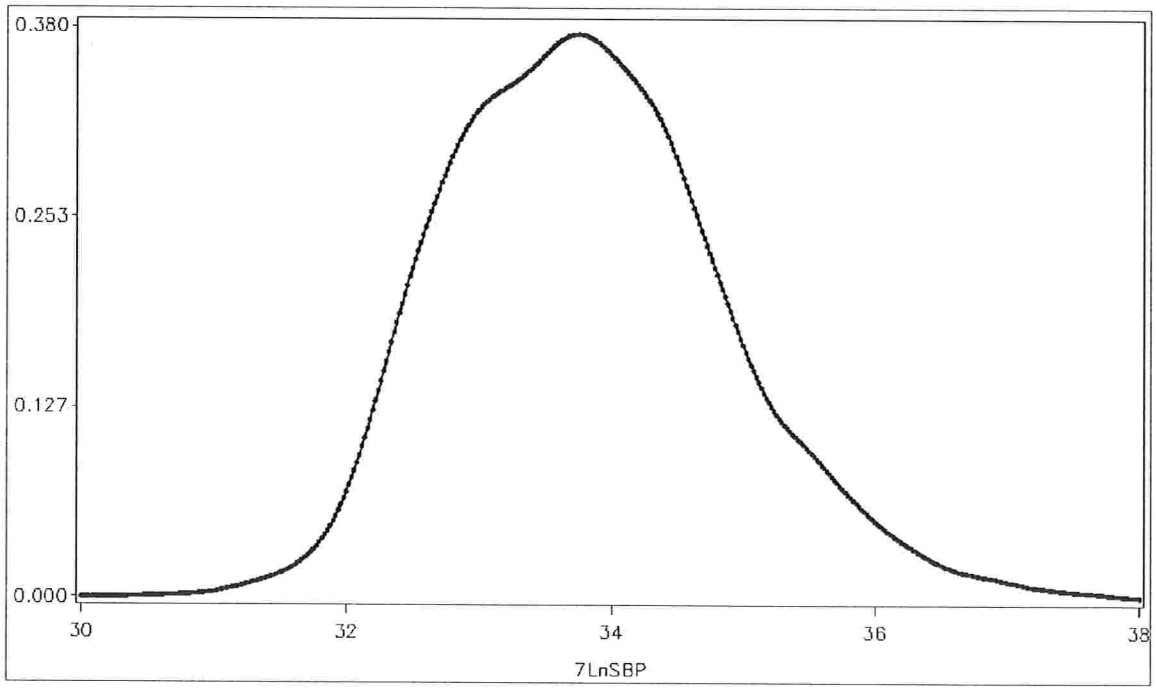

Figura 4. 1: Estimador de kernel da densidade de $\boldsymbol{7lnSBP}$.

Com o interesse em estudar o efeito que a inclusão da covariável *medicamento* exerce no componente poligênico, quatro modelos serão comparados:

Modelo 1: nenhuma covariável

Modelo 2: com a covariável *gênero*

Modelo 3: com as covariáveis *gênero*, *idade*, *gênero*idade* e *IMC*;

Modelo 4: com todas as covariáveis do Modelo 3 além da inclusão da covariável *medicamento*.

A Tabela 4. 1 apresenta as estimativas dos componentes de variância e da herdabilidade poligênica para os quatro modelos mencionados. Essas estimativas foram inicialmente obtidas por Oliveira et al. (2008) e, nesse trabalho, foram novamente obtidas. A herdabilidade (h_g^2) quando nenhuma covariável é incluída no modelo é de aproximadamente 0,15, mesmo valor encontrado para o modelo com somente a covariável *gênero*. Para os modelos 3 e 4 as correspondentes herdabilidades apresentam valores superiores (aproximadamente 0,26 e 0,21, respectivamente), o que mostra que as covariáveis incluídas ajudam a explicar a variabilidade ambiental presente na pressão sistólica arterial dos indivíduos. Destaca-se o fato do componente poligênico ter diminuído com a inclusão da covariável *medicamento* (do modelo 3 para o modelo 4) indicando que existe correlação entre a covariável *medicamento* e o componente herdável .

Tabela 4. 1: Estimativas dos parâmetros aleatórios.

Modelos	σ_y^2	h_g^2	σ_g^2	σ_e^2
1	1,056	0,153	0,162	0,894
2	1,025	0,156	0,160	0,865
3	0,725	0,259	0,187	0,538
4	0,636	0,206	0,130	0,506

Para estudar como acontece essa correlação consideraremos, inicialmente, o gráfico da variável adicionada do modelo 3 (restrito) contra o modelo 4 (irrestrito). A Figura 4. 2 apresenta o gráfico da variável adicionada intra-unidades (componente residual) e uma reta da regressão sem intercepto dos $e_i^e(y|x_1)$ contra os $e_i^e(x_2|x_1)$. A Figura 4. 3 apresenta o gráfico da variável adicionada entre - unidades (componente genético) com a reta da regressão sem intercepto dos $e_i^g(y|x_1)$ contra os $e_i^g(x_2|x_1)$. A Figura 4. 4 apresenta o gráfico da variável adicionada para os componentes residual e genético sobrepostos, com as respectivas retas de regressão ajustadas. A primeira observação a ser destacada é o fato de que ambos os

coeficientes de regressão são positivos ($\hat{\beta}_2^e$=0,74 e $\hat{\beta}_2^g$=0,88), indicando que a média da pressão para os indivíduos que tomaram medicamento é maior do que a média da pressão dos indivíduos que não tomaram medicamento. Destaca-se também o fato de que a inclinação da reta do gráfico intra - unidades é menor do que a inclinação da reta do gráfico entre – unidades. Uma possível interpretação médica para essa correlação é que a covariável medicamento está associada a fatores genéticos.

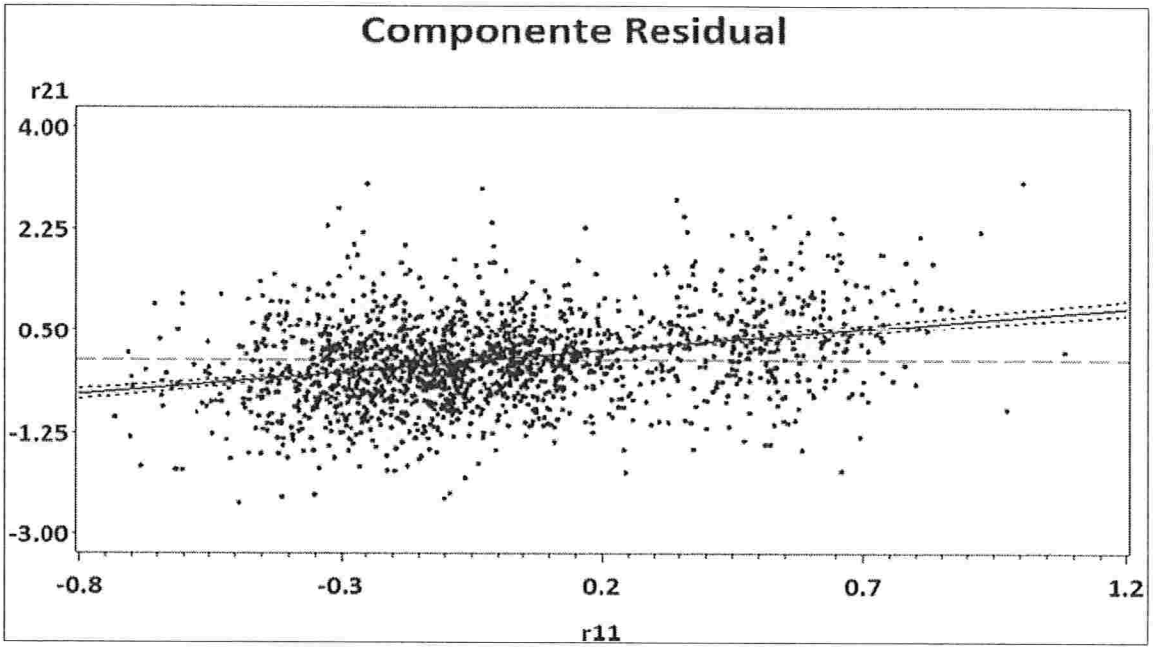

Figura 4. 2: Gráfico da variável adicionada relativo ao componente residual.

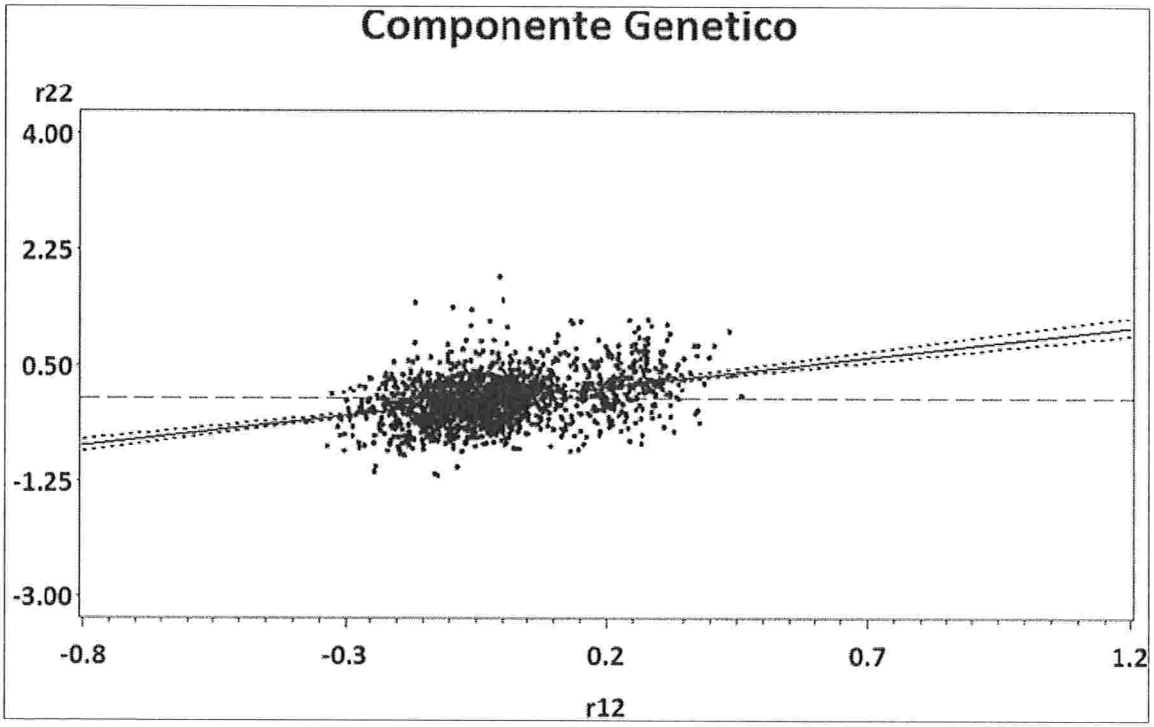

Figura 4. 3: Gráfico da variável adicionada relativo ao componente poligênico.

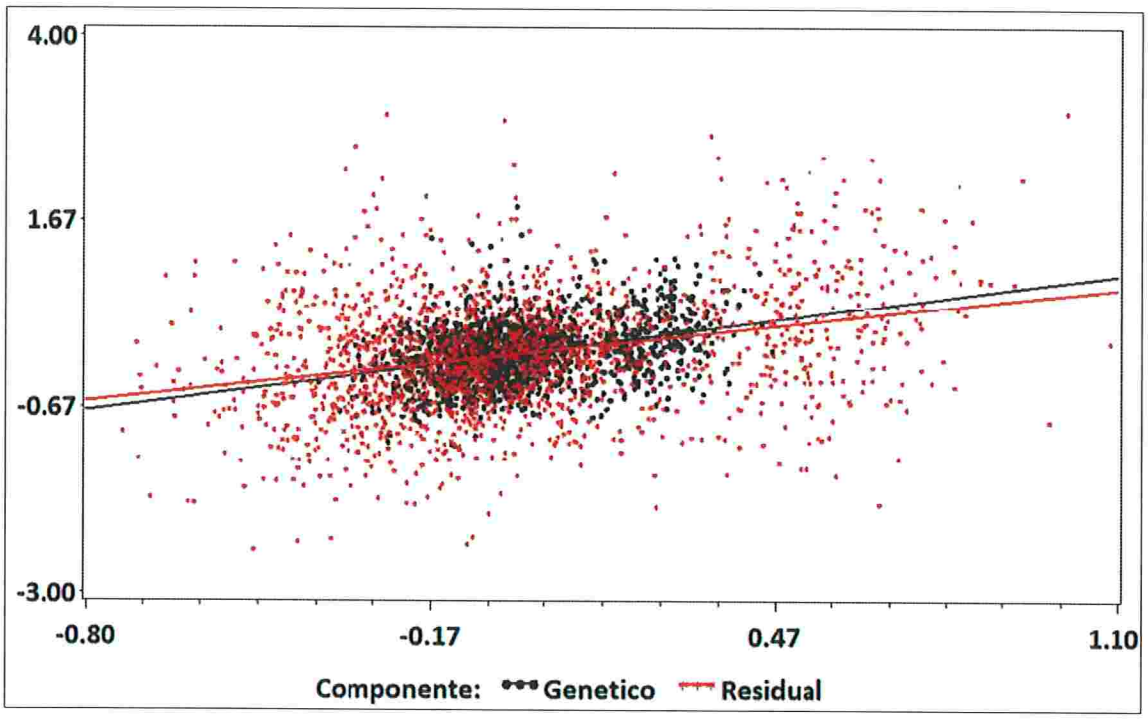

Figura 4. 4: Gráfico da variável adicionada para os componentes residual e poligênico sobrepostos.

Para estudar como acontece a influência do medicamento em cada família, a Figura 4. 5 apresenta o gráfico da variável adicionada entre - unidades (componente genético) decomposto por famílias. Observa-se que existem famílias para as quais o coeficiente angular da regressão sem intercepto é positivo, em outras esse coeficiente é próximo de zero e ainda observa-se algumas com coeficiente negativo. Se o efeito do medicamento fosse parecido para todas as famílias, esperar-se-ia que todas as retas estivessem na mesma direção, ou seja, que as regressões sem intercepto para todas as famílias tivessem coeficientes angulares parecidos.

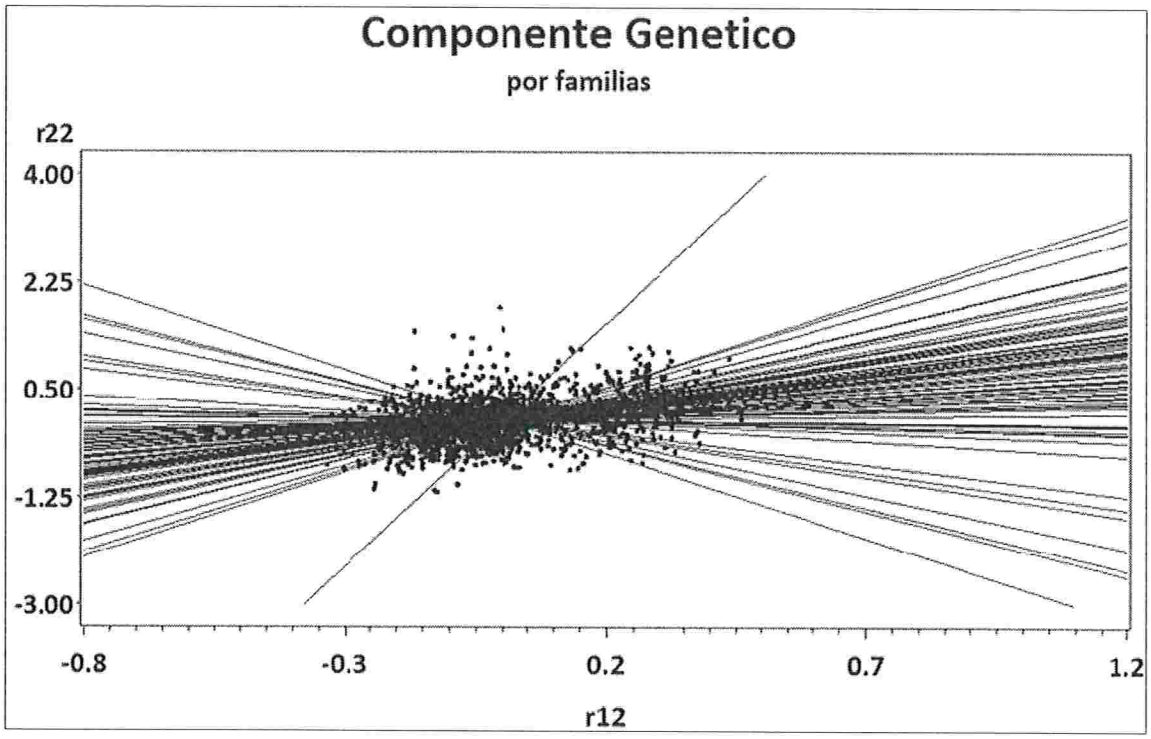

Figura 4. 5: Gráfico da variável adicionada para o componente poligênico decomposto por famílias.

Uma análise direta da Figura 4. 5 nos diria que a família 24 deve ser analisada mais detalhadamente, uma vez que apresenta o maior coeficiente angular (7,8). Voltando às Equações (4. 30) e (4. 31), percebe-se que o "peso" que $\hat{\beta}_{2,i}^{g}$ exercerá em $\hat{\beta}_{2}^{g}$ depende somente de sua variância populacional. Quanto maior for esse parâmetro, menor o seu peso. Para estimar essa quantidade utilizaremos a variância amostral segundo o modelo linear geral, ou seja,

$$\widehat{\boldsymbol{V}}\left[\widehat{\boldsymbol{\beta}}_{2,i}^{\boldsymbol{g}}\right] = [\boldsymbol{e}_{\boldsymbol{i}}^{\boldsymbol{e}}(\boldsymbol{x}_2|\boldsymbol{x}_1)^{\boldsymbol{T}}\boldsymbol{e}_{\boldsymbol{i}}^{\boldsymbol{e}}(\boldsymbol{x}_2|\boldsymbol{x}_1)]\boldsymbol{QME}_{\boldsymbol{i}} \qquad (4.\ 35)$$

em que $QME_i^g = \sum_{j=1}^{n_i} \frac{\left(e_i^g(y|x_1)^T - e_i^g\widehat{(y|x_1)^T}\right)^2}{n_i}$ é o Quadrado Médio do Erro (Graybill, 1976).

Procuraremos assim, famílias para as quais $\hat{\beta}_{2,i}^{g}$ e $\hat{V}\left[\hat{\beta}_{2,i}^{g}\right]$ sejam, respectivamente, relativamente alto e baixo, simultaneamente. De acordo com esse critério, a Tabela 4. 2 apresenta $\hat{V}\left[\hat{\beta}_{2,i}^{g}\right]$ e $\hat{\beta}_{2,i}^{g}$ das 25 famílias com menor variância. Dentre esses valores, procuraremos os que mais se destacam, ou seja, são mais diferentes de $\hat{\beta}_{2}^{g}$, cujo valor estimado foi de 0,88 ($\hat{V}\left[\hat{\beta}_{2}^{g}\right] = 0{,}0036$). A Tabela 4. 3 apresenta alguns percentis para essas quantidades.

Tabela 4. 2: $\widehat{V}[\widehat{\beta}_{2,i}^{g}]$ e $\widehat{\beta}_{2,i}^{g}$ para as 25 famílias da amostra com menor variância no modelo 4.26.

Família	Número de observações não faltantes	$\widehat{V}[\hat{\beta}_{2,i}^{g}]$	$\hat{\beta}_{2,i}^{g}$
30	116	0,04	1,49
28	155	0,04	0,95
122	3	0,05	0,56
27	41	0,05	0,53
68	105	0,06	0,83
36	82	0,08	1,51
15	77	0,08	0,74
5	33	0,09	0,53
20	74	0,09	1,04
11	44	0,11	0,60
7	19	0,11	1,16
105	22	0,11	0,43
101	19	0,11	0,39
16	21	0,13	-0,14
61	18	0,13	1,52
43	18	0,13	0,32
74	22	0,15	1,88
103	3	0,15	1,04
70	15	0,15	1,28
19	21	0,16	1,03
83	21	0,17	1,42
59	16	0,17	0,19
95	10	0,17	-0,07
55	22	0,18	0,26
98	14	0,18	1,29

Tabela 4. 3: Percentis para $\widehat{\boldsymbol{V}}[\widehat{\boldsymbol{\beta}}_{2,i}^{g}]$ e $\widehat{\boldsymbol{\beta}}_{2,i}^{g}$ para todas as famílias da amostra.

	$\hat{V}[\hat{\beta}_{2,i}^{g}]$	$\hat{\beta}_{2,i}^{g}$
Mínimo	0,038603	-2,7808
P1	0,062218	-1,08925
P5	0,093564	-0,27782
P10	0,170605	0,2787
P90	0,911149	1,276475
P95	2,801081	1,685158
P99	5,137832	1,985681
Máximo	28,17087	7,813424

Como um estudo descritivo, a Figura 4. 6 apresenta o gráfico da variável adicionada do componente genético decomposto por famílias. Nele podemos encontrar 3 retas de regressão sem intercepto estimadas, juntamente com o respectivo intervalo com 80% de confiança: a geral para todas as famílias, a da família 30, que dentre as 25 de menor variância é a que apresenta a maior inclinação positiva e a da família 16, que apresenta o menor coeficiente para as de maior peso.

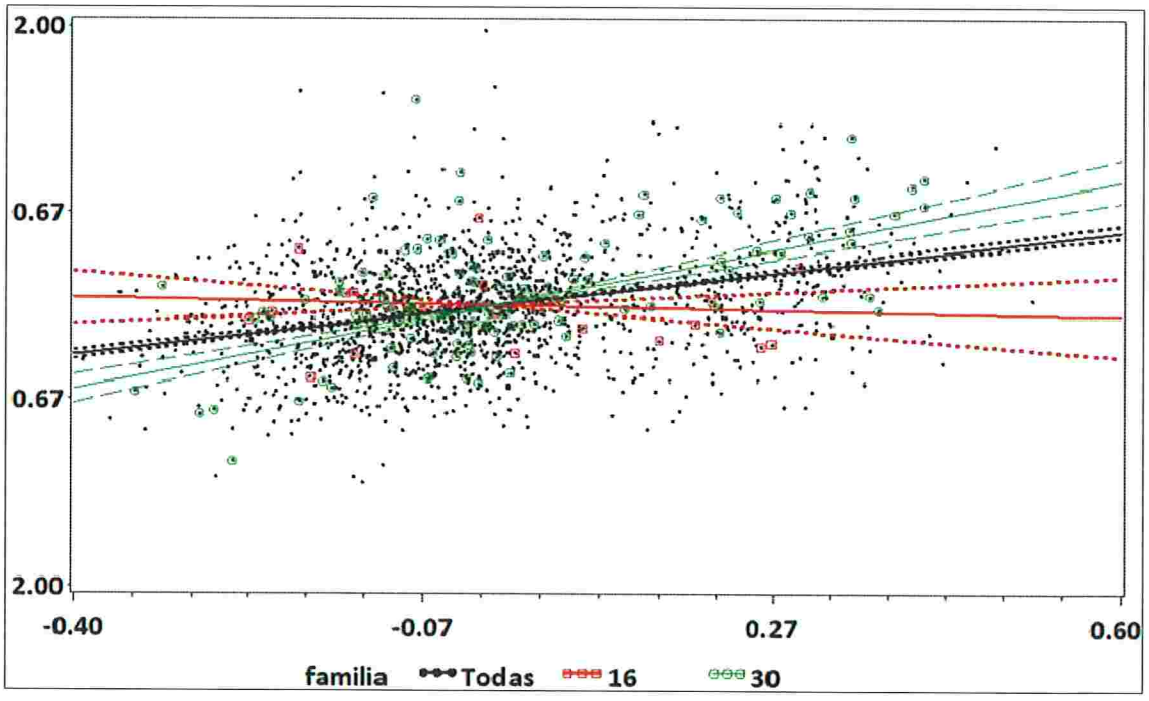

Figura 4. 6: Gráfico da variável adicionada com as retas de regressão sem intercepto para todas as famílias, família 16 e família 30.

Vemos que a variável medicamento tem diferentes efeitos para algumas famílias,

inclusive efeito contrário, como é o caso da família 16. Poderíamos ainda olhar os heredogramas dessas famílias (30 e 16). A família 30 é composta por 116 indivíduos, o que torna a visualização por heredograma bastante difícil. Assim, para título de ilustração, ao invés da família 30 utilizaremos a família 61 ($\hat{V}\left[\hat{\beta}_{2,i}^{g}\right] = 0{,}13$ e $\hat{\beta}_{2,i}^{g} = 1{,}52$). Percebemos, na Figura 4. 7, que apresenta o heredograma da família 16 ($\hat{V}\left[\hat{\beta}_{2,i}^{g}\right] = 0{,}13$ e $\hat{\beta}_{2,i}^{g} = -0{,}14$), que os indivíduos que tomam medicamento (cor branca e sem "NA") têm a pressão ligeiramente inferior àqueles que não tomam medicamento (cor preta), um efeito contrário ao que pode ser percebido na Figura 4. 8, que apresenta o heredograma da família 61, em que as pessoas que tomam medicamento têm pressão média maior do que as que não tomam. Os heredogramas apresentam ainda o gênero (masculino quadrado e feminino oval), a idade (primeira informação abaixo do indivíduo), a pressão arterial sistólica (segunda informação abaixo do indivíduo) e, ainda, se a informação de tomar ou não medicamento não foi obtida (NA).

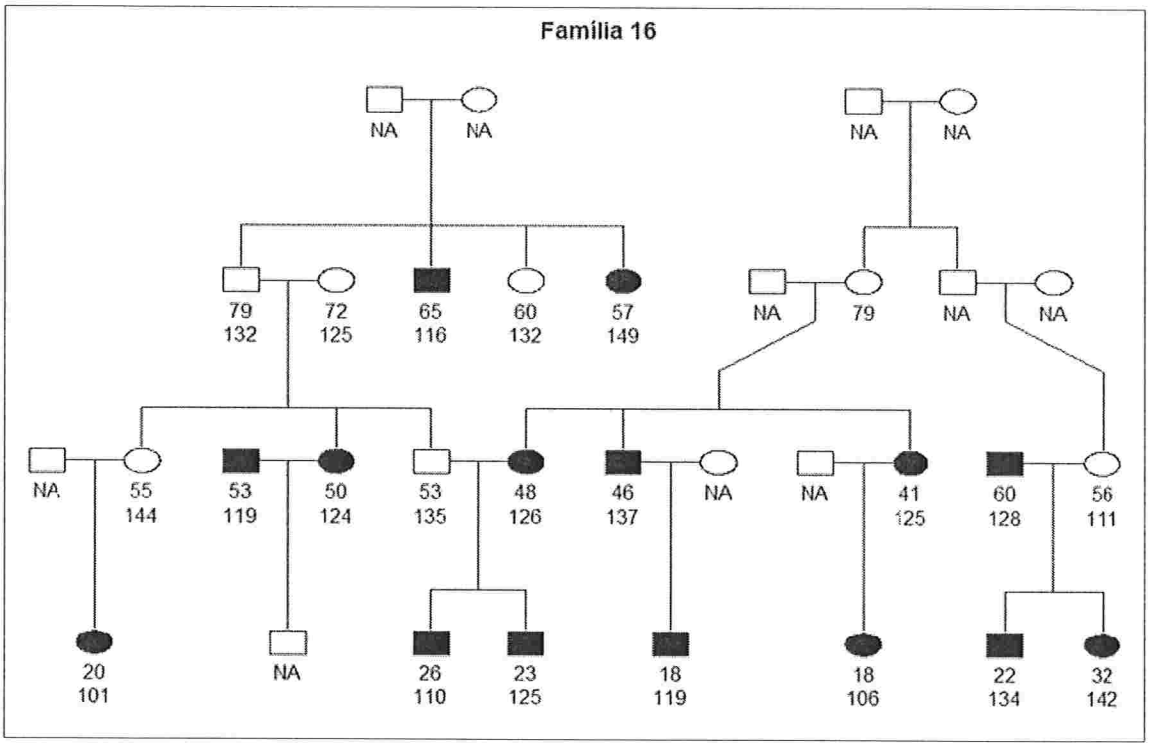

Figura 4. 7: Heredograma da família 16, onde as pessoas do sexo masculino são representadas pelo símbolo quadrado e as do sexo feminino pelo símbolo oval. Abaixo de cada símbolo são apresentadas a idade e pressão arterial.

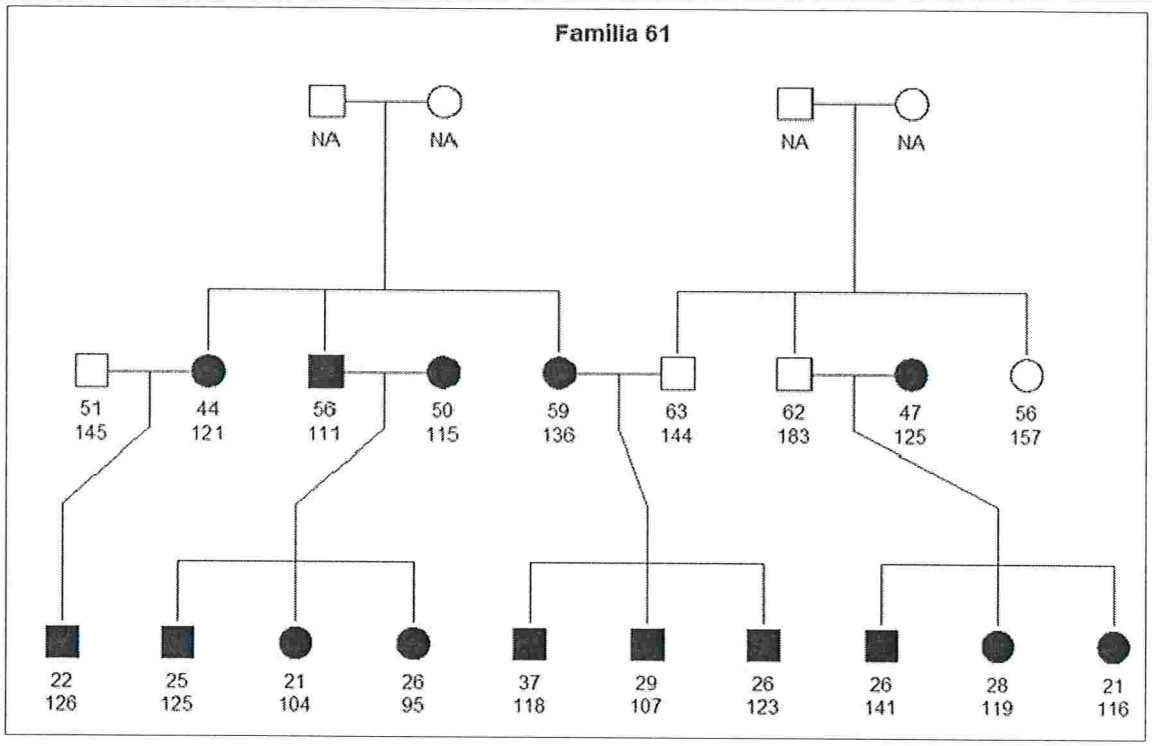

Figura 4. 8: Heredograma da família 61, onde as pessoas do sexo masculino são representadas pelo símbolo quadrado e as do sexo feminino pelo símbolo oval. Abaixo de cada símbolo são apresentadas a idade e pressão arterial.

Poderíamos também estimar o gráfico da variável adicionada comparando o modelo 1 (sem nenhuma covariável) com o modelo 2 (com a covariável *sexo*). A Figura 4. 9 apresenta esse gráfico. Os parâmetros estimados foram ($\hat{\beta}_2^e$=0,36 e $\hat{\beta}_2^g$=0,33).

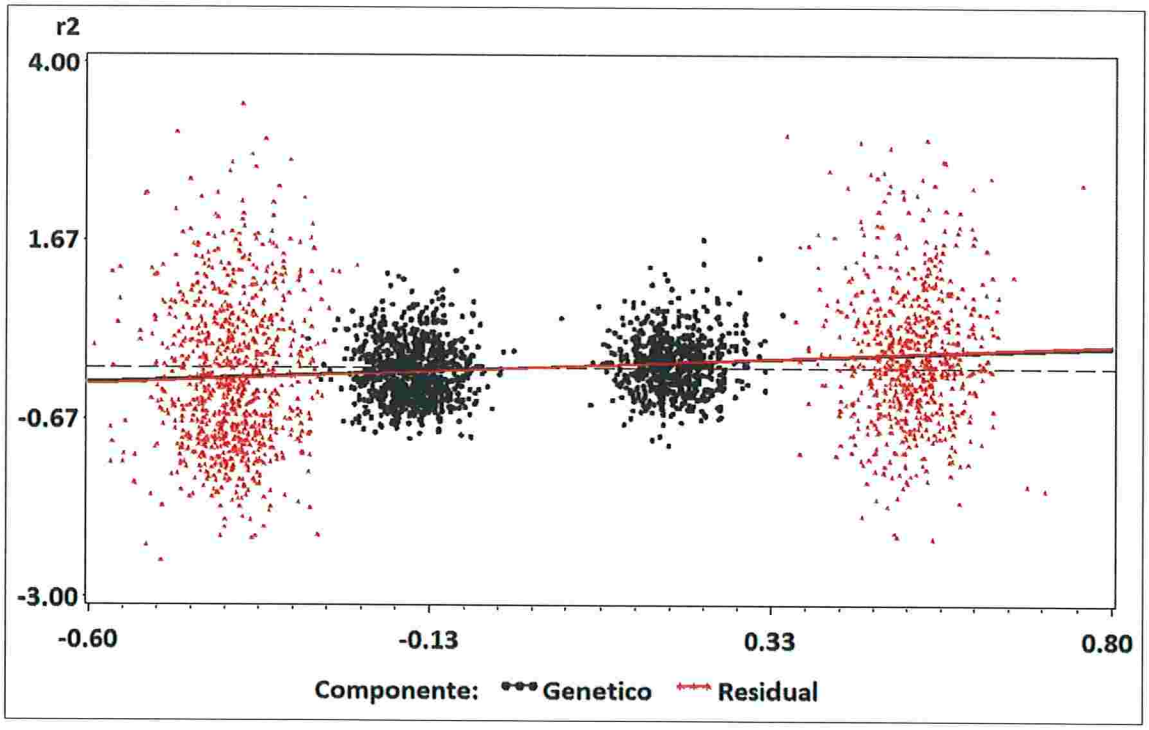

Figura 4. 9: Gráfico da variável adicionada para os componentes residual e poligênico sobrepostos (modelo 1 versus modelo 2).

Comentários

Esse estudo mostra diferenças entre o efeito de medicamento, em que efeito está definido nesse trabalho como a diferença no valor esperado entre a pressão arterial sistólica das pessoas que tomam medicamento versus as pessoas que não tomam medicamento, nas diversas famílias. Uma etapa posterior de estudo sugerida é a detecção de quais são os QTLs responsáveis por essa diferença de efeito.

O delineamento aqui utilizado encaixa-se em um estudo farmacológico de fase 4 com o objetivo de estudar famílias influentes no efeito de medicamento. Essa informação poderia ser para médicos avaliarem famílias que podem necessitar de outro regime de dosagem do medicamento ou de troca da medicação.

Algumas adaptações ao estudo podem ser sugeridas para obtenção de resultados mais informativos. Para um estudo de fase 2 ou 3, poder-se-ia sortear aleatoriamente quais famílias receberiam um determinado medicamento ativo e as outras famílias receberiam um placebo. Isso permitiria uma definição de efeito do medicamento como a eficácia do medicamento contra um placebo.

5. Estrutura Populacional

No Capítulo 2 discutiu-se o conceito de estrutura de populações (estratos genéticos) e as possíveis implicações de associações espúrias em estudos de associação Farmacogenômicos. Ainda, foi salientada a grande chance de ocorrência de tal evento na população brasileira, sabidamente miscigenada. Nesse capítulo o interesse é discutir métodos utilizados para correção de estrutura de populações e alguns resultados encontrados em populações humanas e de ratos F2.

5.1. Métodos para Corrigir o Efeito da Estratificação Genética

A estratificação populacional é um problema de extrema relevância nos estudos Farmacogenômicos, como foi apresentado no capítulo 2. Para lidar com a estratificação existem vários métodos. Dois métodos consagrados pela literatura são o controle genômico e a associação estruturada. O controle genômico, (veja Shmulewitz et al., 2004; Devlin & Roeder, 1999), corrige a estratificação por meio do ajuste de estatísticas de associação em cada marcador por um fator uniforme geral de inflação. Entretanto, alguns marcadores diferem em sua freqüência alélica mais do que outros, de modo que o ajuste uniforme aplicado pelo controle genômico pode ser insuficiente. O outro método, associação estruturada (veja, Pritchard et al., 2000), assinala um grupo para cada uma das unidades amostrais e, então, agrega evidência de associação dentro de cada cluster. Entretanto, esse método não pode ser aplicado atualmente a estudos genômicos de associação devido ao seu alto custo computacional em grandes conjuntos de dados (Price et al., 2006).

Um novo método foi desenvolvido por (Price et. al., 2006) e utiliza-se da técnica de componentes principais. Seja $G = (g_{ij})$ uma matriz de dimensão $N \times M$ em que N representa o número de indivíduos na amostra, M o número de marcadores amostrados de cada indivíduo e g_{ij} representa a informação genotípica do indivíduo i no marcador j. Considere a matriz $X_{N\times M} = (x_{ij})$ resultante da seguinte normalização nas colunas da matriz $G_{N\times M}$

$$x_{ij} = \left(\frac{g_{ij} - \overline{g_j}}{s_j}\right)$$

em que $\overline{g_j}$ é a média aritmética e s_j o desvio padrão da variável j (j-ésimo marcador).

É possível obter duas matrizes a partir da matriz X. A primeira, denotada por $\Sigma_{M\times M} = X^T X$, consiste na matriz de covariâncias entre os M marcadores e, a outra, denotada por $\Psi_{N\times N} = XX^T$, a matriz de covariâncias entre os N indivíduos. Em estudos caso-controle de associação genética, para o ajuste das medidas de associação devido ao efeito de estrutura de população, Zhang et al. (2003) sugerem a utilização da matriz $\Sigma_{M\times M}$ (Análise de Componentes Principais) para o cálculo da ancestralidade, enquanto Price et al. (2006) utiliza a matriz $\Psi_{N\times N}$ (Análise de Coordenadas Principais). Gower (1996) descreve a dualidade e equivalência analítica das duas técnicas anteriores. Na maioria das aplicações atuais o espaço de marcadores é muito maior do que o espaço dos indivíduos, o que torna o método proposto por Price et al. (2006) mais interessante do ponto de vista computacional.

Após aplicar a técnica de Coordenadas Principais à matriz X obtêm-se os coeficientes de ancestralidade. Os principais componentes de ancestralidade são então incorporados como covariáveis na análise de regressão (logística, por exemplo) corrigindo o efeito da estrutura da população. Intuitivamente, isso cria um conjunto de casos e controles balanceados (Price et. al., 2006).

Esses coeficientes de ancestralidade obtidos por esses dois métodos são globais, isto é, são uma medida resumo do genoma inteiro. Considerando que o genoma dos indivíduos pertencentes a populações miscigenadas é composto por segmentos cromossômicos de ancestralidades distintas é interessante calcular a ancestralidade local em tais indivíduos. O método implementado no algoritmo HAPMIX, proposto em Price et al. (2009), baseia-se em duas populações parentais e emprega um modelo genético populacional baseado em cadeias de Markov ocultas para calcular a ancestralidade local, utilizando dados de alta resolução (SNPs, por exemplo).

Aplicação 1: Ancestralidade na População Brasileira

Historicamente, tem acontecido na população brasileira um grande grau de miscigenação, sendo ela uma das mais miscigenadas do mundo, possuindo ascendência principalmente européia, africana e nativa, e pouco é conhecido sobre sua estrutura genética. Pode-se dividir o Brasil em cinco regiões geográficas e as características de miscigenação diferem entre elas. No norte predomina a ascendência ameríndia, com um pouco de africana também. A população do nordeste e centro-oeste tem origem na população africana e européia. Na região Sul e Sudeste nota-se uma predominância de ancestralidade européia,

mas também se pode notar algumas regiões com bastante ascendência africana e asiática. Entre essas cinco regiões, a Sudeste foi por décadas o maior centro de atração para imigrantes e migrantes, sendo possivelmente a região que apresenta o maior grau de miscigenação racial (Giolo et al., 2009; Gonçalves et al., 2008).

Os dados aqui estudados foram primeiramente analisados em Giolo et al. (2009) e são compostos de uma amostra de 138 indivíduos da população do sudeste do Brasil avaliados genotipicamente por meio da plataforma de SNPs da Affymetrix 6.0 (com cerca de 1 milhão de SNPs) e 991 indivíduos das seguintes populações do banco de dados do Haplotype Map (HapMap) fase III (veja International HapMap Project, Phase III; Duan et al., 2008): ancestrais do sudeste da África (ASW); residentes de Utah, EUA, com ancestralidade do norte e do oeste da Europa a partir da coleta CEPH (CEU); Chineses da etnia Han de Pequim, na China (CHB); Chineses da Metrópole de Denver, no Colorado, EUA (CHD); descendentes do povo Gujarati que vivem em Houston, Texas, EUA (GIH); Japoneses de Tóquio, no Japão (JPT); Luhya no Webuye, no Quênia (LWK); ascendência mexicana em Los Angeles, na California, EUA (MEX); Maasai no Kinyawa, no Quênia (MKK); Toscanos na Itália (TSI) e os Yoruba em Ibadan, na Nigéria (YRI). Para combinar esses 12 bancos de dados foi necessário excluir os SNPs que não eram comuns entre eles, resultando em 365.116 SNPs comuns e 1129 indivíduos não relacionados para análise.

Utilizando a metodologia proposta em Price et al. (2006), Giolo et al. (2009) aplicaram a técnica de Coordenadas Principais aos dados descritos anteriormente. A Figura 5. 1 apresenta o valor de cada uma das vinte primeiras coordenadas principais. Nota-se que após a terceira, os valores são aproximadamente constantes.

Na Figura 5. 2 tem-se o gráfico de dispersão entre as duas primeiras coordenadas principais, com a respectiva legenda das populações. As populações HapMap referentes aos diferentes continentes estão claramente separadas pelos eixos de ancestralidade. A população brasileira (sudeste do Brasil) apresenta uma variação contínua entre esses dois eixos, com indivíduos mais próximos das populações do continente africano e europeu, indicando uma importante miscigenação entre esses povos, que durante séculos originou a população brasileira.

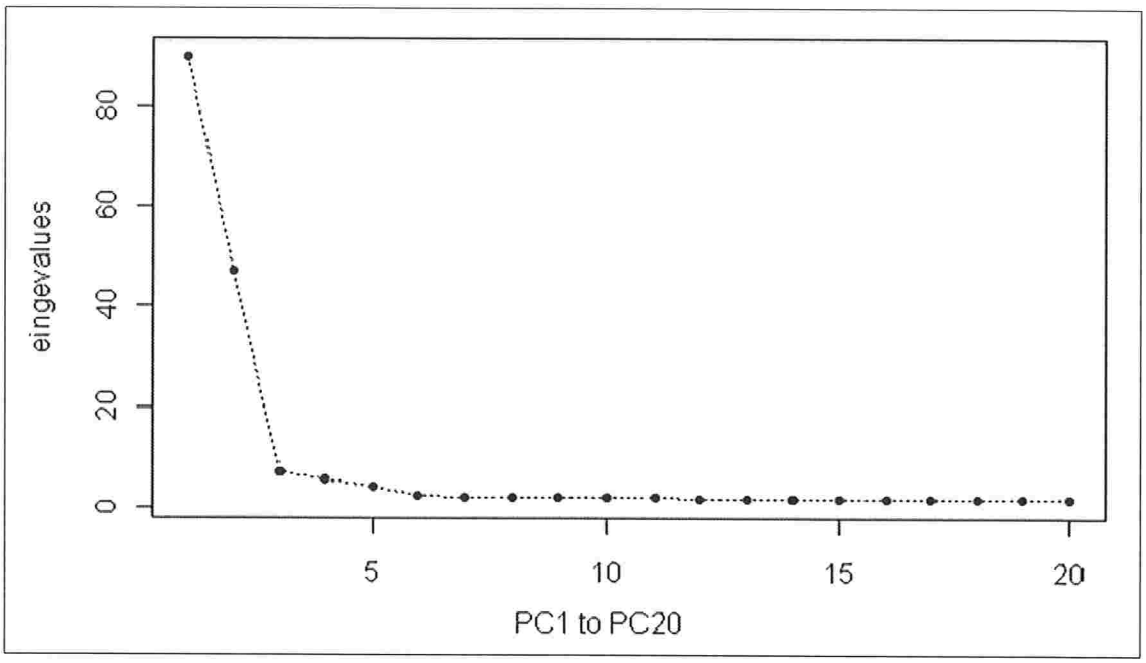

Figura 5. 1: Autovalores associados com as 20 primeiras coordenadas principais (eixos de variação).

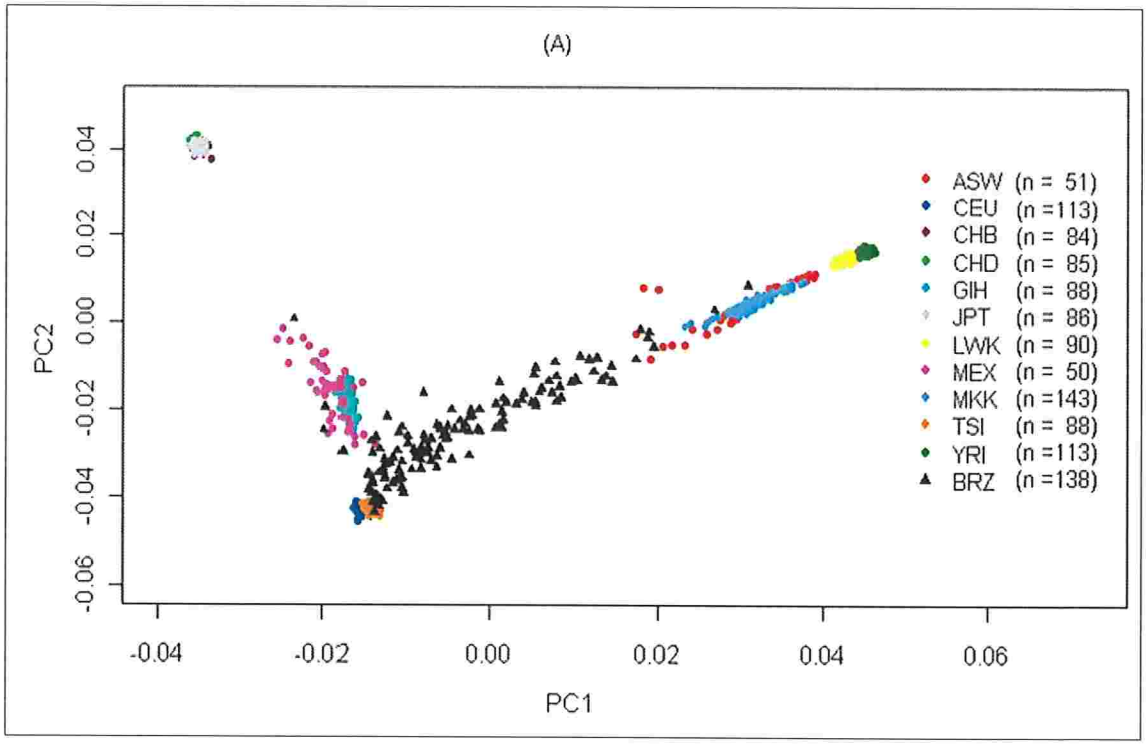

Figura 5. 2: Projeção dos 1129 indivíduos nos seus primeiro e segundo eixos de variação.

Pela Figura 5. 3, onde se encontra um gráfico de dispersão tri-dimensional, com as três primeiras coordenadas principais, vê-se que o terceiro componente principal consegue separar claramente a população mexicana dos descendentes de Gujarati. Caso os indivíduos aqui analisados pertencessem a um estudo Farmacogenômico, utilizaríamos os valores estimados

dessas coordenadas principais como covariáveis em estudos de associação para corrigir o problema de estrutura de populações na amostra.

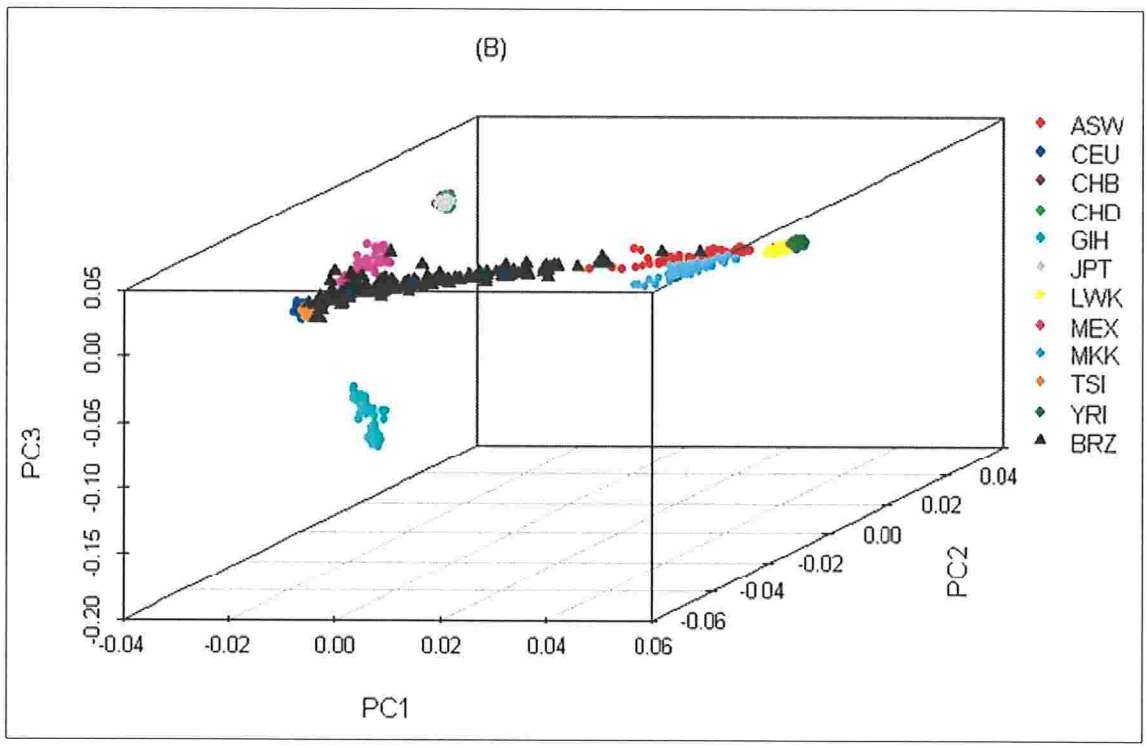

Figura 5. 3: Projeção dos 1129 indivíduos nos seus primeiro, segundo e terceiro eixos de variação.

Para aumentar o entendimento sobre a contribuição de populações ancestrais para os eventos de ancestralidade, foram estimadas as probabilidades ancestrais dos 138 indivíduos brasileiros amostrados utilizando o método HAPMIX considerando as populações CEU (européia) e YRI (africana) como supostas populações parentais de referência, gerando uma estimativa probabilística da ancestralidade de cada indivíduo em cada loco (SNP) (Giolo et al., 2009). A Figura 5. 4 apresenta o número esperado de cópias de alelos da população CEU para um indivíduo selecionado na amostra, utilizando o método HAPMIX (Price et al., 2009). A partir dessa figura podemos observar que o genoma desse indivíduo é composto por segmentos de cromossomos com ancestralidades distintas.

Em caso de um estudo Farmacogenômico, se suspeitássemos que os indivíduos do estudo possuíssem ancestralidade distinta em determinada região cromossômica poder-se-ia utilizar os coeficientes de ancestralidade locais aqui calculados para corrigir esse problema nos marcadores dessa região.

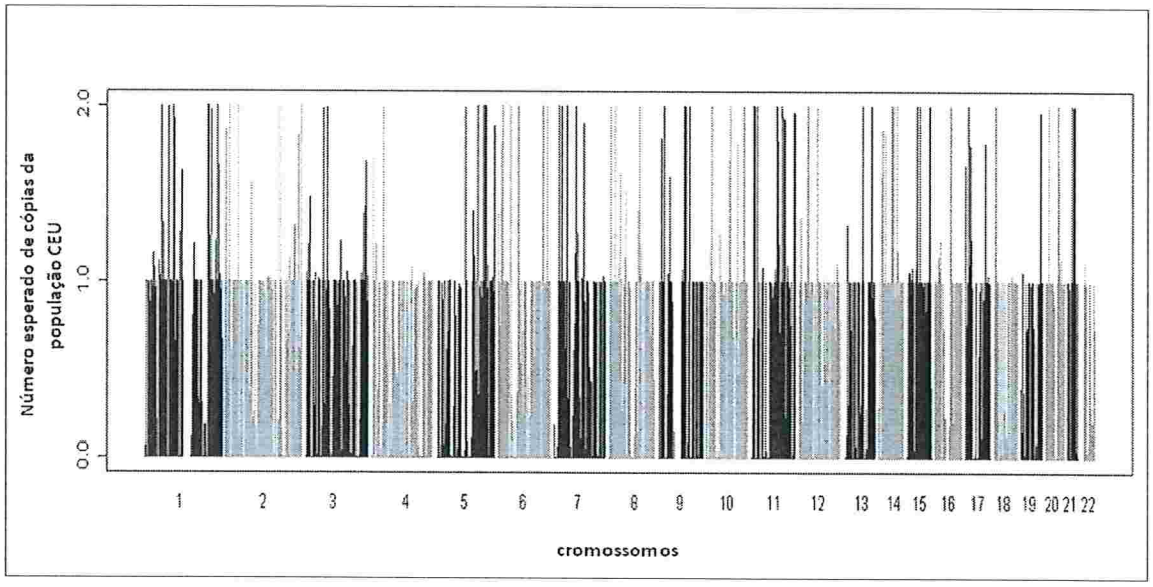

Figura 5. 4: Número esperado de alelos CEU para um indivíduo brasileiro para os 22 cromossomos autossomos.

Similarmente, a Figura 5. 5 mostra o número esperado de alelos CEU para todos os 138 indivíduos da amostra, utilizando o método HAPMIX (Price et al., 2009). Novamente podemos notar que o genoma dessa população é composto por cromossomos de ancestralidade distinta, com relevante miscigenação entre europeus e africanos, permitindo a conclusão de que hoje existe forte ancestralidade européia e africana na população brasileira proveniente de miscigenações que vêm acontecendo por décadas.

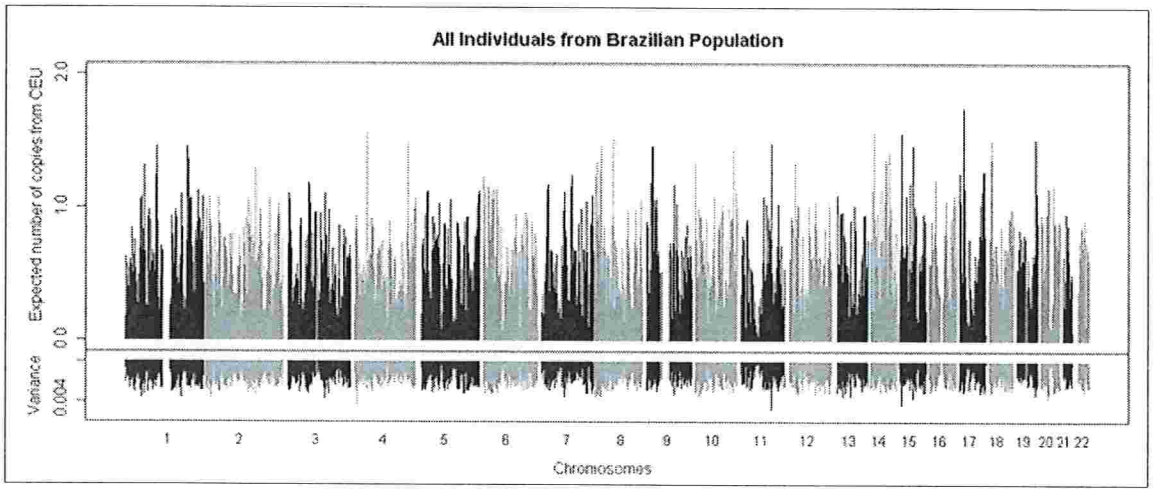

Figura 5. 5: Número esperado de alelos CEU para todos os 138 indivíduos da amostra, com sua respectiva variância, para os 22 cromossomos autossomos.

Os resultados presentes nessa seção mostram que a grande miscigenação que aconteceu na população brasileira desempenha um papel fundamental na sua variação e caracterização genética, tornando-a diferente das outras populações. Esse fato indica que

estudos farmacológicos devem ser desenvolvidos com a população brasileira tanto para melhorar os efeitos farmacológicos das drogas já existentes, através de um regime de dose e dosagem mais indicado à sua constituição genética, quanto para aprovar novos medicamentos.

Aplicação 2: Variabilidade Genética em Populações F2

Como visto no capítulo 2 os modelos animais desempenham um importante papel no desenvolvimento de medicamentos e é de extrema importância entender como os dados produzidos em estudos com animais devem ser analisados. Podemos pensar em ratos F2 como miscigenados (recentemente) a partir de duas populações. Para entender como funciona a estrutura de populações em cruzamentos controlados onde se conhece a informação genética dos ancestrais (pais) aplicamos a técnica para visualização de ancestralidade descrita na seção anterior.

Os dados aqui analisados já foram previamente estudados em Schork et al. (1995) e Duarte (2007) e referem-se a um experimento realizado no Laboratório de Cardiologia e Genética Molecular do Instituto do Coração de São Paulo, InCor-USP, onde uma linhagem de ratos normotensos, Brown-Norway (BN), foi cruzada com outra de ratos hipertensos, Spontaneously hipertensive rat (SHR), gerando 221 ratos da geração F2, que possuem características de *background* genético praticamente idênticas, diferindo somente nos determinantes genéticos da pressão arterial. Foram coletadas 23 variáveis, entre elas: pressão basal, pressão sistólica antes e pós-sal, pressão diastólica antes e pós-sal, pressão após o uso do medicamento Captopril e o peso do animal. Segundo a definição médica, ratos com pressão arterial sistólica igual ou superior a 140 mmHg são considerados hipertensos, enquanto ratos com pressão arterial sistólica inferior a esse limite são ditos normotensos. Além disso, para o mapeamento genético, foi utilizado um mapa com 182 marcadores espalhados nos 21 cromossomos dos ratos (ver Figura 5. 6). Essa aplicação utiliza como fenótipo respostas clínicas (capítulo 3).

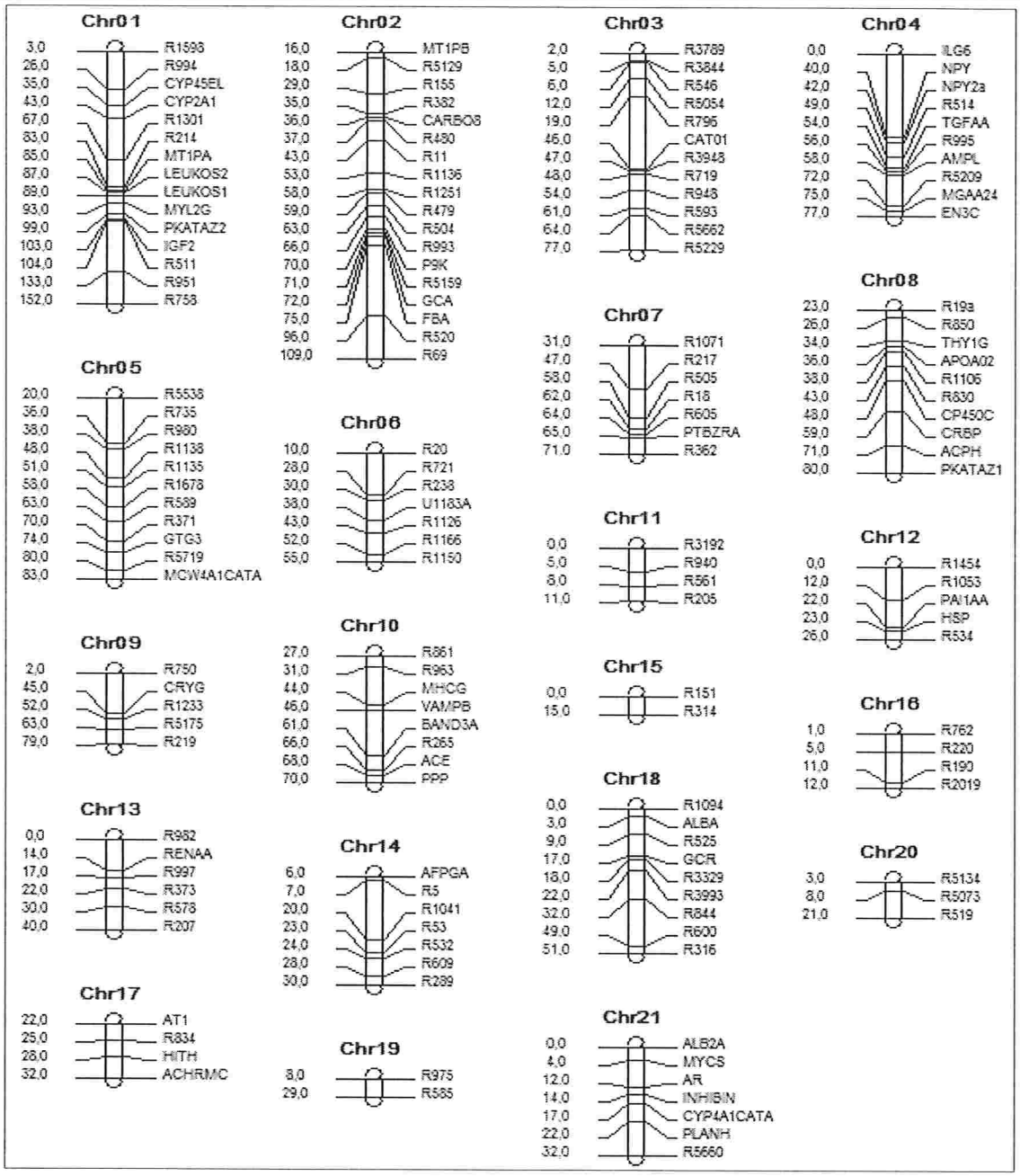

Figura 5. 6: Mapa de marcadores utilizado para genotipagem dos ratos F_2.

Foram excluídos da análise 43 (19% do total) dos marcadores que tinham 30% ou mais de dados faltantes e 42 (19% do total) ratos foram excluídos da análise por apresentarem 30% ou mais de dados faltantes. Dos dados resultantes, ainda restaram 10% de dados faltantes e esses foram imputados utilizando as probabilidades genotípicas de marcadores adjacentes, descritas em Haley & Knott (1992).

A título de descrição dos dados, a Figura 5. 7 exibe o número de alelos para cada marcador de um determinado rato selecionado aleatoriamente na amostra. Nos cromossomos 13, 14 e 15 deste animal todos os marcadores carregam um alelo de risco. Isso se deve ao fato

do DNA ser passado dos pais para os filhos em grandes pedaços.

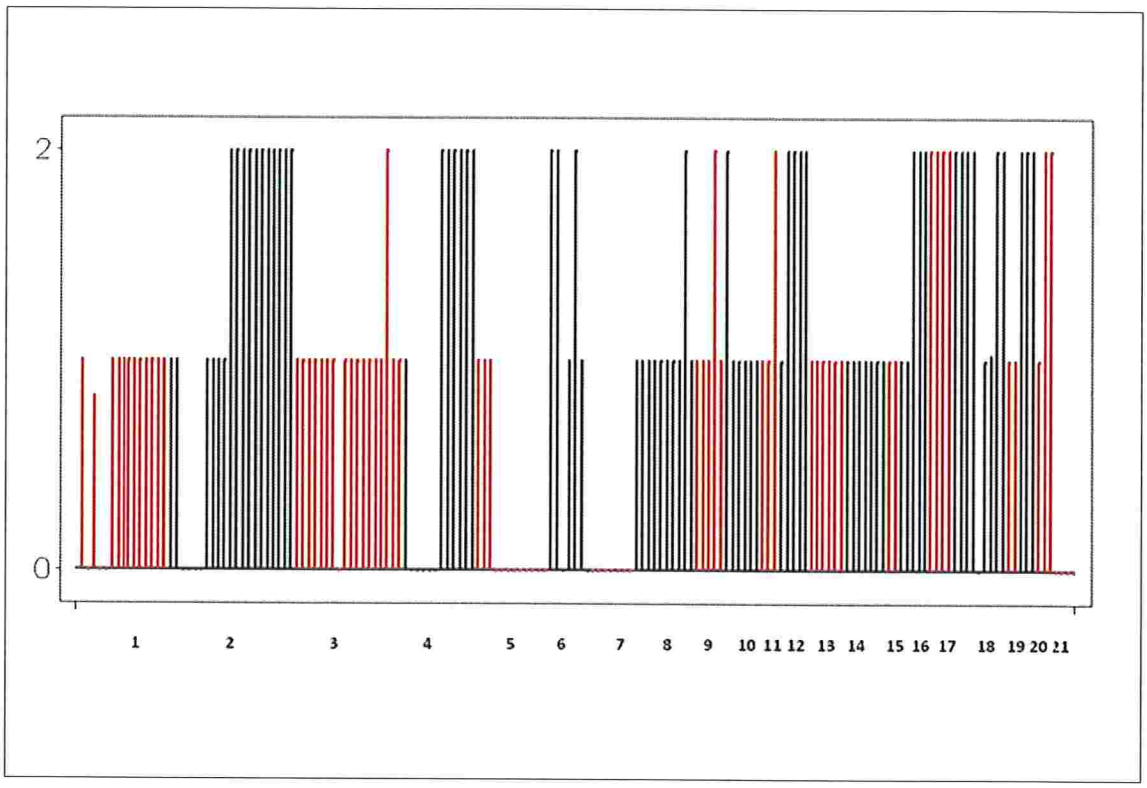

Figura 5. 7: Número de alelos de risco para um rato F_2 da amostra.

A Figura 5. 8 apresenta a média amostral de alelos de risco. Nota-se que essa média, em todos os marcadores, está muito próxima de 1. Isso se deve ao fato de que, de acordo com a segunda lei de Mendel (seção 2.3), os genótipos AA, Aa e aa serão gerados, em um experimento F_2, na razão $1:2:1$, assim a esperança do número de alelos de risco ($E(A)$) pode ser calculada por

$$E(A) = 2 \times \frac{1}{4} + 1 \times \frac{1}{2} + 0 \times \frac{1}{4} = 1 .$$

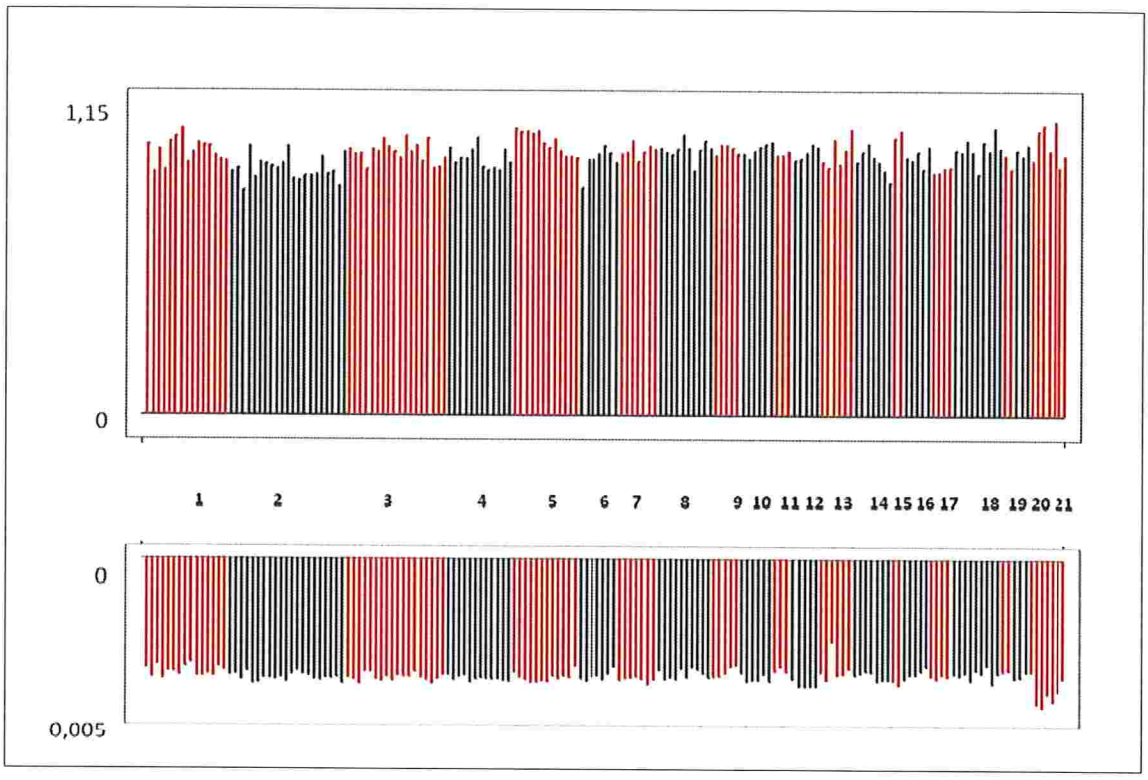

Figura 5. 8: Média amostral de alelos de risco e correspondentes variâncias para os ratos F_2 para os 21 cromossomos.

A Figura 5. 9 apresenta as duas primeiras coordenadas principais, onde não é possível notar nenhuma estratificação clara entre os indivíduos e, a partir, disso podem ser levantadas duas hipóteses:

I. Todos os indivíduos compõem uma mesma população;

II. Os indivíduos compõem três populações ou estratos, dois deles próximos a cada um dos genitores e o terceiro como uma "mistura" dos genitores. Apesar disso, a variabilidade genética é muito grande e contínua e por essa razão não se consegue distinguir claramente essas três populações.

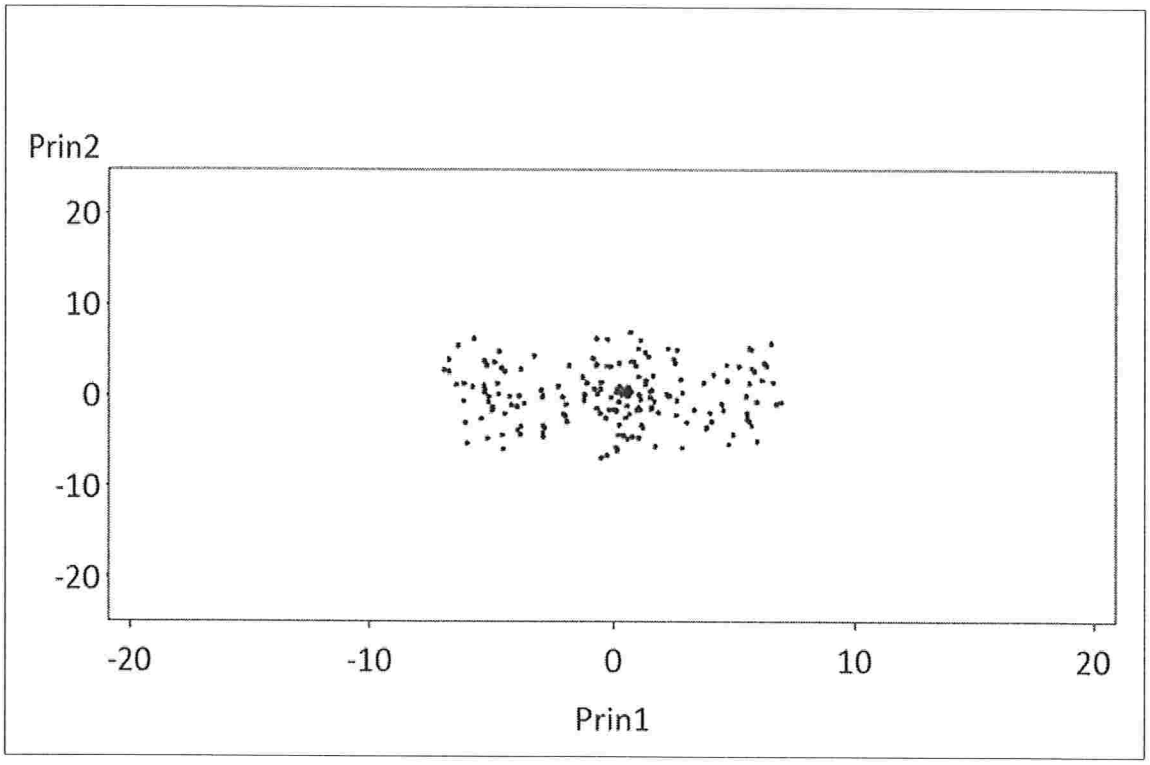

Figura 5. 9: Projeção dos ratos F_2 da amostra nos seus primeiro e segundo eixos de variação obtidos a partir da técnica de coordenadas principais.

Um dos "avôs" dessa geração F_2 é um indivíduo com valor 0 para todos os marcadores, o homozigoto recessivo (aa), e o outro com valor 2 para todos os marcadores, o homozigoto dominante (AA). A Figura 5. 10 apresenta novamente as duas primeiras coordenadas principais, com a inclusão dessas duas "novas" observações. Os avôs se localizaram em direções opostas, e os indivíduos da geração F_2 dispersos entre eles. Esse padrão é diferente do observado nas Figura 5. 2 e Figura 5. 3, onde a população brasileira se distribui mais uniformemente entre as duas populações parentais, diferente dos ratos F_2, que se concentram mais no "centro". Ainda, uma possível explicação para o fato de não se conseguir distinguir grupos é porque o número de marcadores não é suficiente nem informativo para a técnica empregada (Price et al., 2006) ou o número de gerações depois da miscigenação é pequeno (Ewens & Spielman, 1995).

Cruzamentos controlados entre linhagens divergentes oferecem uma configuração ideal para se detectar e mapear QTLs através de associações entre marcadores e traço, uma vez que a geração F_1 é idêntica entre si e apresenta desequilíbrio de ligação completo (associação não aleatória entre alelos em dois ou mais locos, não necessariamente no mesmo cromossomo) para todos os genes divergentes entre as linhagens (Lynch & Walsh, 1998). As duas linhagens presentes nesse estudo foram construídas com o interesse de que elas

divergissem geneticamente apenas para os genes reguladores da pressão arterial. Mas como isso é feito? Em plantas, por exemplo, podemos utilizar a estratégia de descendentes de uma única semente, em que uma única semente de cada planta da cada geração é selecionada aleatoriamente e plantada para gerar descendentes. A esperança do número de heterozigotos cai pela metade em cada geração. Depois de 8 ou 10 cruzamentos obtém-se uma linhagem (quase) homozigota. Obviamente não podemos fazer isso com animais e, portanto, um cruzamento entre irmãos é o mais indicado para a criação de uma linhagem (Jansen, 2003). Como essas linhagens foram construídas com o objetivo de garantir que elas diferissem exclusivamente nos genes determinantes de hipertensão, é esperado que esses ratos F_2 não apresentem "problema" de estrutura de populações.

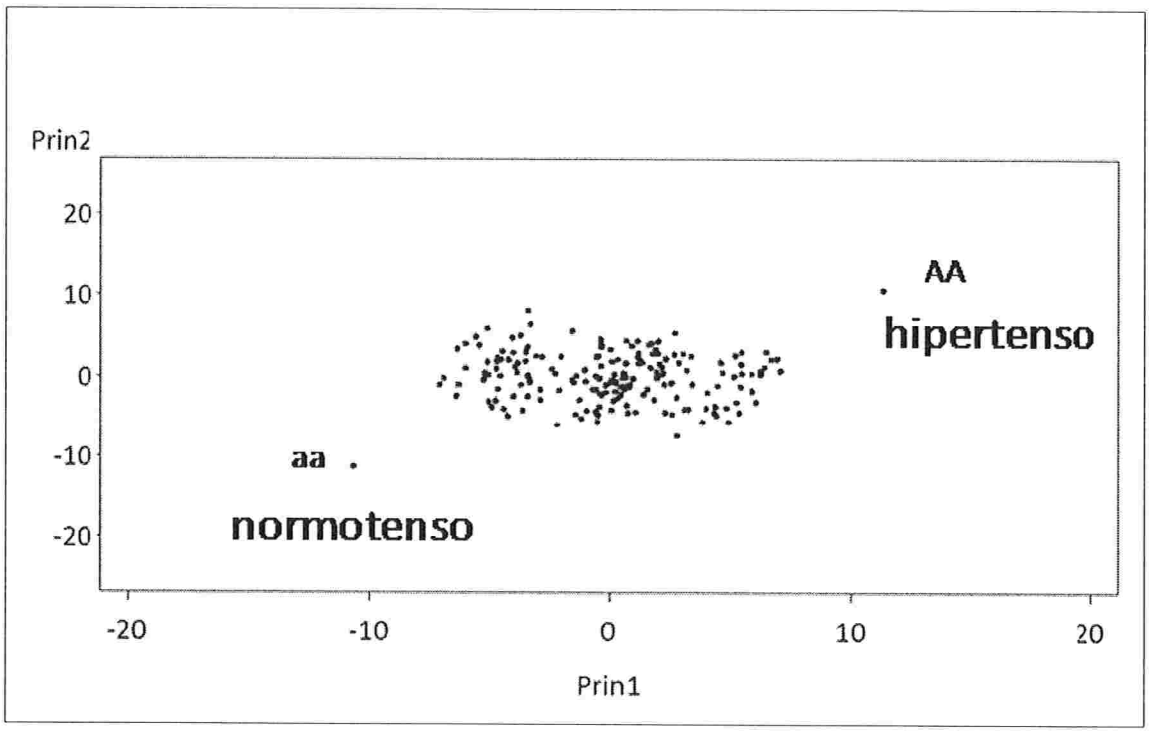

Figura 5. 10: Projeção dos ratos F_2 da amostra de dos ratos parentais nos seus primeiro e segundo eixos de variação obtidos a partir da técnica de coordenadas principais.

Para representar conjuntamente os indivíduos e os marcadores, e entendermos quais marcadores estão mais correlacionados com cada sub-população (caso elas existam), podemos utilizar os Biplots (Gabriel, 1971; Johnson & Wichern, 2002). A Figura 5. 11 exibe essa técnica aplicada aos dados. Os pontos indicam os ratos F_2 e as setas os marcadores. Notam-se três grupos bem definidos de marcadores distribuídos quase simetricamente entre os pontos parentais. Os animais F_2 distribuem-se uniformemente no "centro" desses três grupos.

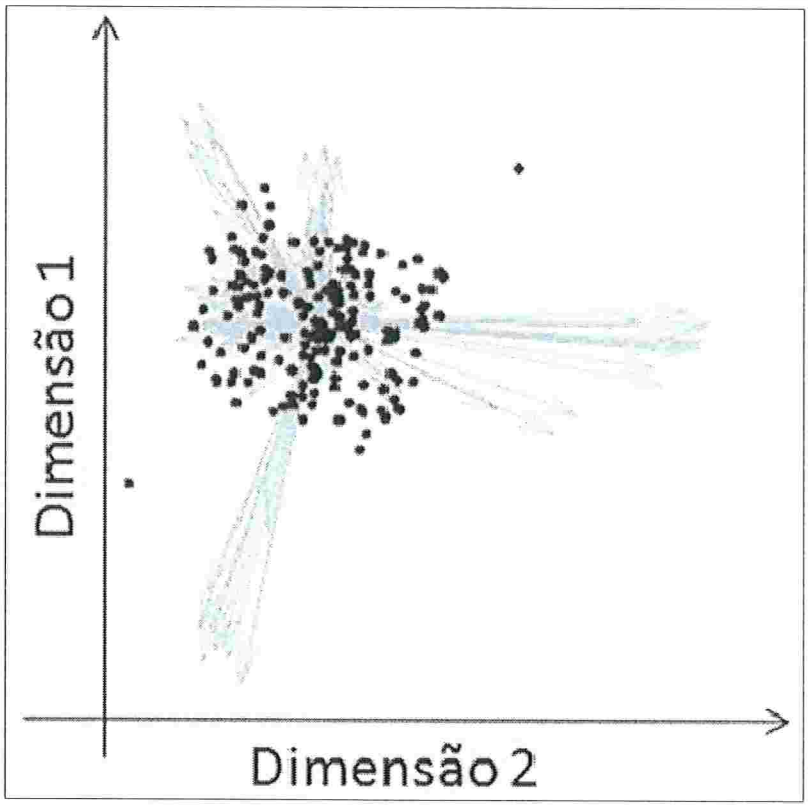

Figura 5. 11: Projeção dos ratos F_2 da amostra e dos ratos parentais bem como dos marcadores moleculares nos seus primeiro e segundo eixos de variação obtidos a partir da técnica de Biplot.

Dando continuidade à exploração da estrutura populacional de populações F_2 o interesse agora é detectar genes reguladores do efeito farmacológico do Captopril a partir desses dados, trabalho iniciado em Duarte (2007). Para tanto os seguintes fenótipos serão estudados:

Y_{SBP}: medidas de pressão arterial sistólica dos ratos em condição basal;

Y_{SBPS}: medidas de pressão arterial sistólica dos ratos após a intervenção com sal;

$Y_{Captopril}$: medidas de pressão arterial sistólica dos ratos após administrado o medicamento anti-hipertensivo chamado Captopril.

Tabela 5. 1: Correlações entre as medidas de pressão.

	SBPS	*Captopril*
SBP	0.423	0,382
SBPS		0,573

A análise da Tabela 5. 1, que apresenta os valores das correlações (de Pearson) entre as medidas, mostra que as correlações são sempre positivas e fracas ou moderadas (sempre menores do que 0,6). A Figura 5. 12 apresenta os perfis individuais da resposta para os três

traços em estudo. Nota-se que a tendência geral é de aumento na pressão após a administração de sal. Apesar disso, em alguns ratos, cuja pressão basal era alta, acontece uma queda na pressão após a administração de sal e isso pode ser explicado pela presença de algum gene protetor, isso é, que regula aumentos muito elevados da pressão. Ainda, da Figura 5. 12, percebe-se que a tendência é de queda da pressão após administração do medicamento Captopril. Após essa queda, a pressão arterial atinge valores ainda mais baixos do que a pressão basal. Novamente, alguns ratos não seguem o padrão e a pressão para eles aumenta após a utilização do Captopril. Esse fato pode ser explicado pela existência de genes que influenciam a eficácia do medicamento produzindo respostas atípicas.

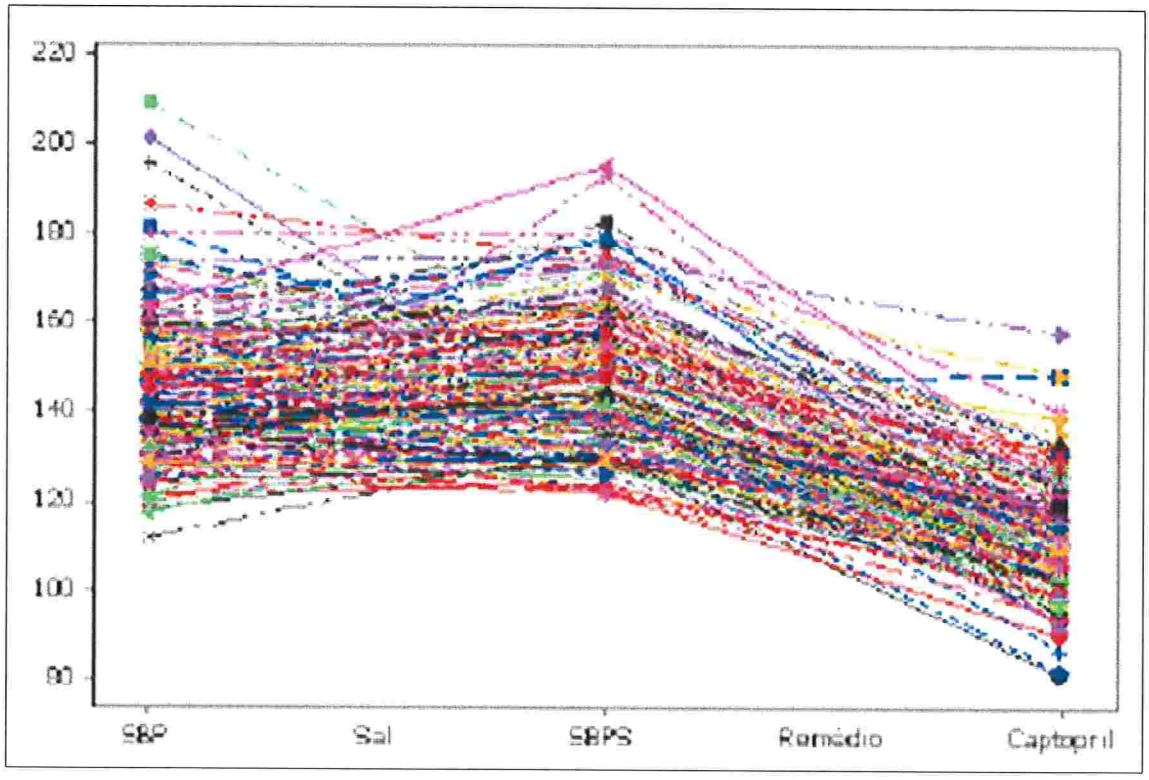

Figura 5. 12: Perfis individuais de resposta para a pressão arterial medida em três etapas.

Para encontrar QTLs em associação com a regulação do efeito de Captopril foi utilizado o modelo multivariado intervalar composto (Jansen, 1993; Lander & Botstein, 1989 ; Duarte, 2007). Nesse modelo testa-se a hipótese da não existência de efeito genético em cada posição fixada do genoma e o perfil da estatística da razão de verossimilhanças é colocado em um gráfico para avaliação de possíveis sinais de efeito genético (Zeng, 1994).

A Figura 5. 13 apresenta o perfil da estatística razão de verossimilhanças, na escala log_{10} (chamada de lod-escore). Devido ao problema de falsos positivos em testes múltiplos, Kao et al. (1999) sugerem a utilização do valor crítico 2,5 para a estatística lod-escore. Na

Figura 5. 13 destacam-se sinais de ligação significativa nos cromossomos 1, 2, 4, 5, 8, 16 e 21, onde o cromossomo 21 é o sexual. Quanto aos sinais referentes aos cromossomos 1, 2, 4, 5, 8 (primeiro QTL) e 16, outros estudos já têm feito referência a essas regiões (Schork et. al. 1995; Duarte, 2007), as quais estão associadas ao efeito do sal na pressão arterial. Ressalta-se que o sinal do segundo QTL do cromossomo 8, cujo código é ACPH, somente é significativo quando o fenótipo pressão arterial após administração de Captopril é incluído na análise (Duarte, 2007).

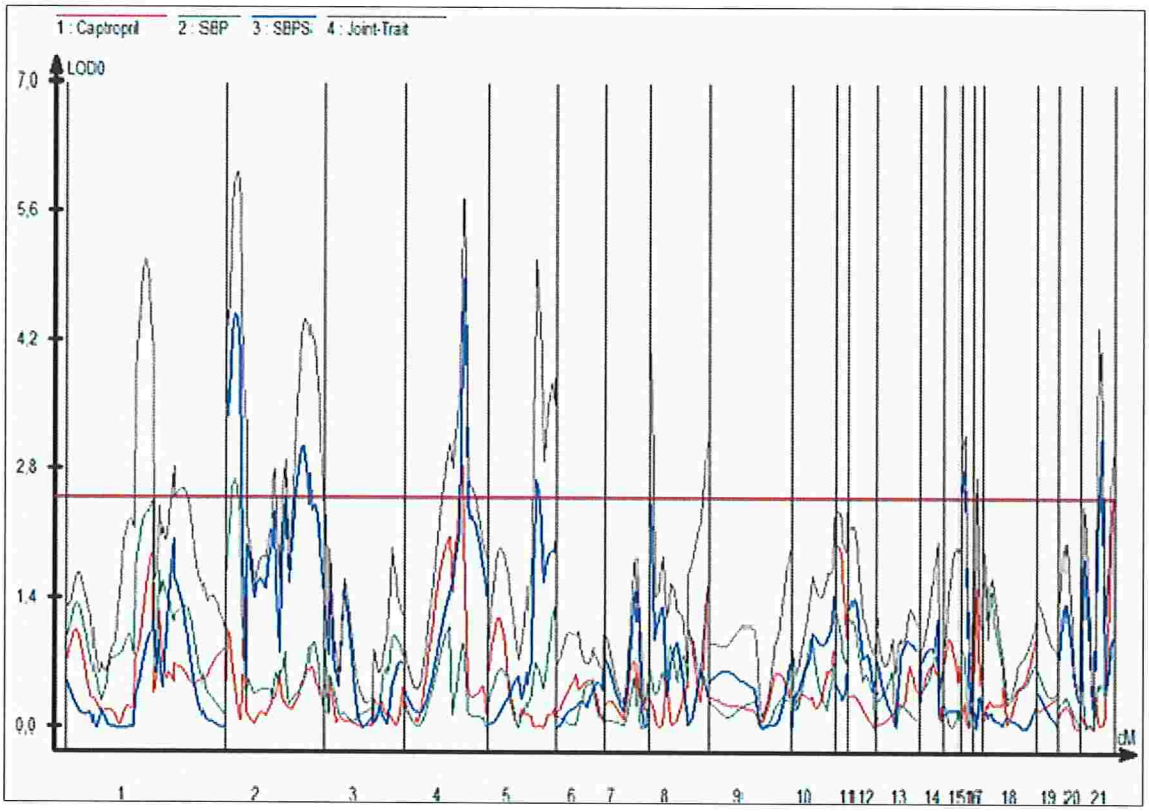

Figura 5. 13: Perfil da estatística lod-escore para o modelo multivariado.

A Tabela 5. 2 apresenta as estimativas dos efeitos aditivos e de dominância para cada um dos 8 QTLs significativos encontrados (excluindo-se o QTL do cromossomo sexual) e cada um dos fenótipos analisados. A primeira coluna mostra o cromossomo (Cr), a segunda coluna o marcador mais próximo (M) e a terceira coluna a posição (em Morgan) do QTL encontrado (Poisção). Para cada fenótipo temos na primeira coluna a estatística lod-escore marginal (LR), na segunda coluna a estimativa do efeito aditivo (a) e na terceira coluna a estimativa do efeito de dominância (d). A última coluna da Tabela 5. 2 apresenta a estatística lod-escore conjunta (LR conjunta) para os três fenótipos.

Tabela 5. 2: Estimativas dos efeitos genéticos do modelo multivariado.

Cr	M	Posição	SBP			SBPS			Captopril			LR Conjunta
			LR	a	d	LR	a	d	LR	a	d	
1	M1301	0,72	10,6	7,6	-0,8	3,7	3,1	3,1	6,3	1,4	5,6	23
2	R155	0,13	10,5	6,3	-2,4	18,8	6,8	-0,4	0,5	0,5	0,7	27
2	FBA	0,71	2,0	2,9	1,0	14,0	7,4	-2,8	2,1	-0,9	3,6	20
4	R514	0,53	4,2	3,7	-1,9	22,4	8,6	-0,3	9,7	4,9	-0,8	26
5	R589	0,43	3,3	0,7	5,0	12,4	4,1	-4,3	0,1	-0,3	-0,4	23
8	R19a	0,0001	3,0	1,0	4,6	11,1	45,0	-2,5	1,5	0,0	2,5	19
8	**ACPH**	**0,48**	**2,5**	**-0,7**	**-4,6**	**3,1**	**-2,8**	**0,4**	**1,4**	**1,7**	**-1,0**	**10**
16	R762	0,0001	4,0	3,4	0,2	11,9	4,6	-3,3	0,6	0,8	0,0	14

Selecionado o QTL localizado no cromossomo 8 e de posição 0,48 M, na Figura 5. 14, temos o Biplot dos dados de ratos e marcadores, onde agora cada ponto foi colorido de acordo com o número de alelos de risco que o rato carregava no cromossomo 8. Os ratos com 2 alelos de risco receberam a cor preta, os heterozigotos a cor azul e os homozigotos recessivos a cor vermelha. Além disso, o marcador ACPH do cromossomo 8 foi ressaltado. Não é possível visualizar nenhum padrão entre as cores dos pontos e a distância entre os três grupos de marcadores representados pelas setas.

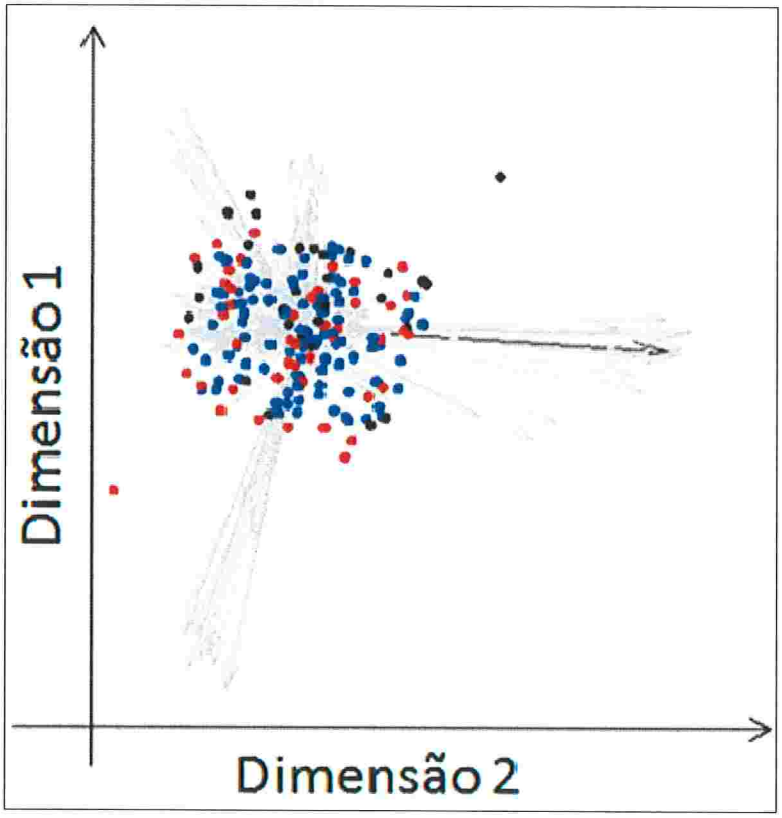

Figura 5. 14: Projeção dos ratos F_2 da amostra de dos ratos parentais, genotipados para o marcador ACPH, nos seus primeiro e segundo eixos de variação obtidos a partir da técnica de Biplot, com o marcador ACPH destacado.

A Figura 5. 15 ilustra os efeitos genotípicos estimados pelo modelo multivariado intervalar composto. Nota-se que para todos os genótipos, as médias estimadas para a pressão após administração de Captopril são bastante inferiores ao limite de definição de hipertensão. Ressalta-se que isso ocorreu para todos os outros QTLs encontrados e mostrados na Tabela 5. 2. Uma combinação dos seguintes fatos pode explicar esse resultado. O primeiro deles é que, nesse estudo, a administração de Captopril foi feita acima das doses recomendadas, não se preocupando com efeitos de toxicidade e talvez seja por essa alta dosagem que a pressão média para todas as categorias genotípicas está abaixo do limite de definição de hipertensão. O segundo fato baseia-se no que foi discutido no capítulo 3. Como os ensaios clínicos para aprovação dos medicamentos em uso ainda não utilizam informações genotípicas, para serem aprovados eles devem funcionar para a grande maioria da população, o que também explica os resultados obtidos.

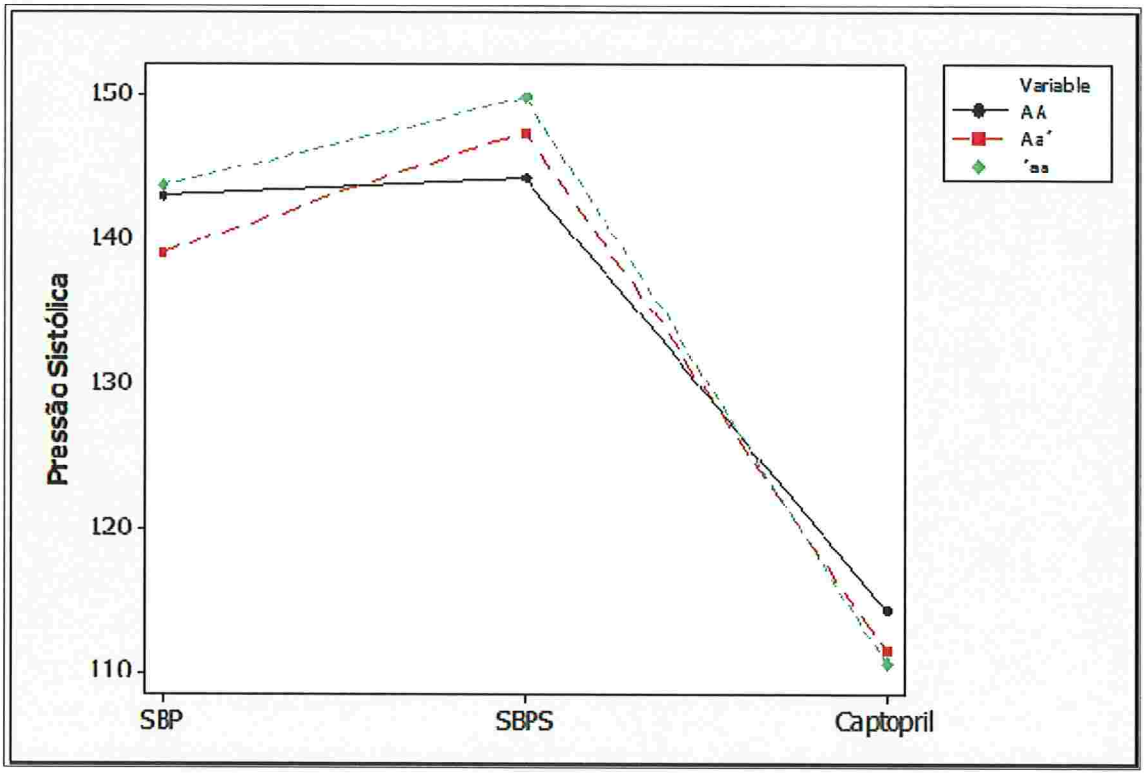

Figura 5. 15: Perfis de médias estimadas para o marcador ACPH do cromossomo 8.

6. Considerações finais

O interesse pelo desenvolvimento de metodologias estatísticas na área da farmacogenômica está cada vez maior, já que são fundamentais para o desenvolvimento de novos medicamentos e definição do regime de aplicação de medicamentos que já estão no mercado mais adequados às características individuais (genéticas) de cada indivíduo. Nesse trabalho foram consideradas diversas metodologias e conceitos importantes para os estudos farmacogenômicos. As aplicações se restringiram aos bancos de dados disponíveis para análise. Nesse sentido, o ajuste de modelos não lineares, do tipo considerado no capítulo 3, não foi abordado. As seguintes conclusões foram obtidas a partir desse trabalho:

1. Grande importância de inclusão de dados genéticos nos estudos farmacológicos para selecionar drogas e doses que aumentem os efeitos benéficos e diminuam os efeitos tóxicos das drogas de acordo com a genética individual;

2. A técnica do gráfico da variável adicionada tem grande potencial de aplicação na área de Farmacogenômica devido à sua capacidade de discriminar famílias (ou grupos genéticos) influentes no efeito do medicamento;

3. O problema de estrutura de populações deve sempre ser levado em conta nos estudos de associação (caso-controle) Farmacogenômicos para evitar associações espúrias. Devido a isso, torna-se extremamente importante desenvolver ensaios clínicos envolvendo as populações com diferentes histórias genéticas, em particular a brasileira, conhecidamente miscigenada;

4. Não necessidade de correção para estrutura de populações em cruzamentos controlados do tipo F_2;

5. A técnica de componentes principais (ou coordenadas principais), bem como o gráfico Biplot são úteis na visualização e correção do problema de estrutura de populações.

A seguir destacam-se alguns tópicos de pesquisas futuras a serem exploradas nas análises de delineamentos farmacogenômicos:

- Continuidades da análise dos dados das famílias de Baependi para entender quais são os locos cromossômicos responsáveis pelo efeito do medicamento no controle da pressão arterial. Para isso são aguardados dados de genótipos de marcadores

moleculares (SNPs, por exemplo);

- Obtenção de dados reais ou simulados para ajuste de modelos não lineares para os processos farmacocinéticos e farmacodinâmicos;

- A Farmacogenômica traz motivações para que diferentes fenótipos sejam construídos para avaliação do efeito genético, por exemplo, a velocidade de decréscimo do valor da pressão.

Como último resultado sugere-se que delineamentos com famílias sejam incluídos em ensaios clínicos para o desenvolvimento de medicamentos, estratégia que não é comum nos estudos de fármacos atuais. Esses estudos podem ser realizados através da seleção de famílias (brasileiras) "informativas" para fenótipos de interesse farmacológico, por exemplo, nas quais a doença tem segregado por gerações. Esse tipo de estudo seria útil em todas as fases clínicas de desenvolvimento da droga. Em um estudo de fase 4, por exemplo, poder-se-ia genotipar as famílias selecionadas, considerando diferentes plataformas de marcadores moleculares (como microsatélites ou SNPs), e seus valores fenotípicos seriam acompanhados por anos, atualizando resultados adversos ou gerando fenótipos mais complexos.

Apêndice A

Nesse apêndice segue o código fonte do programa implementado no SAS utilizado nesse trabalho para o cálculo dos resíduos utilizados para a construção do gráfico da variável adicionada. O procedimento utilizado foi o PROC IML.

```
/*                QUANTIDADES QUE SERÃO UTILIZADAS:                */

/*    -sigma2_g:componente de variância associado ao componente  */
/*    genético no modelo restrito                                */

/*    -sigma2_e:componente de variância associado ao componente  */
/*    residual no modelo restrito                                */

/*    -phi2_dataset: conjunto de dados com as entradas da matriz */
/*    de parentesco                                              */

/*    -Z_dataset: conjunto de dados com as entradas da matriz    */
/*    identidade                                                 */

/*    -Y_dataset: conjunto de dados com as entradas do vetor Y   */

/*    -X2_dataset: conjunto de dados com as entradas do vetor X2 */

/*    -X1_dataset: conjunto de dados com as entradas da matriz X1*/


proc iml;
start added_plot;

      Var=sigma2_g*phi2+sigma2_e*Z;
/*    Calcula a matriz de covariância do modelo poligênico       */

      call svd(uVar,qVar,vVar,Var);
      sqr_Var=uVar*diag(sqrt(qVar))*vVar`;
/*    Calcula a raiz quadrada da matriz de covariância do modelo */
/*    poligênico utilizando a decomposição de Cholesky           */

      InvSQR_M=sqr_Var/(sigma2_y**0.5);
      M=inv(h2_g*phi2+h2_e*Z);
/*    Calcula a matriz M                                         */

      inv_X1MX1=inv(X1`*M*X1);
/*    Calcula a inversa da matriz X1´*M*X1                       */

      Q1=M-M*X1*inv_X1MX1*X1`*M;
/*    Calcula a matriz de Q1                                     */

      call svd(uphi2,qphi2,vphi2,phi2);
      sqr_phi2=uphi2*diag(sqrt(qphi2))*vphi2`;
/*    Calcula a raiz quadrada da matriz phi2 utilizando a        */
/*    decomposição de Cholesky                                   */
```

```
      r21=h_e*Q1*Y;
      r11=h_e*Q1*X2;
/*    Calcula os resíduos intra-unidades                          */

      r22=h_g*sqr_phi2*Q1*Y;
      r12=h_g*sqr_phi2*Q1*X2;
/*    Calcula os resíduos entre-unidades                          */

      create res1 from r1;append from r1;
      create res2 from r2;append from r2;
      create res11 from r11;append from r11;
      create res21 from r21;append from r21;
      create res22 from r22;append from r22;
      create res12 from r12;append from r12;
/*    Cria os datasets utilizando os vetores criados durante o    */
/*    PROC IML                                                    */

finish added_plot;

use phi2_dataset; read all into phi2;
/*    Cria o vetor phi2 utilizando o dataset phi2_dataset         */

use Z_dataset; read all into Z;
/*    Cria o vetor phi2 utilizando o dataset Z_dataset            */

use Y_dataset; read all into Y;
/*    Cria o vetor phi2 utilizando o dataset SBP_dataset          */

use X2_dataset; read all into X2;
/*    Cria o vetor phi2 utilizando o dataset SBP_dataset          */

use X1_dataset;read all into X1;
/*    Cria o vetor X1 utilizando o dataset SBP_dataset            */

run added_plot;
quit;
```

Apêndice B

Nesse apêndice segue o código fonte do programa implementado no SAS utilizado nesse trabalho para a imputação dos dados faltantes dos ratos F_2.

```
/*                QUANTIDADES QUE SERÃO UTILIZADAS:                */
/*    -ratos: conjunto de dadas com as informações genotípicas     */
/*    dos ratos da amostra, onde R5--MCW4A1CATAa são o primeiro     */
/*    e último marcador, respectivamente                            */


/*excluir ratos e marcadores com mais de 30% de dados faltantes*/
data ratos_a;
      set ratos;
      array a(208) R5--MCW4A1CATAa;
            do i=1 to 208;
                  if a(i)=0 then a(i)=.;
                  else if a(i)=1 then a(i)=0;
                  else if a(i)=2 then a(i)=2;
                  else if a(i)=3 then a(i)=1;
                  else if a(i)=4 then a(i)=.;
                  else if a(i)=5 then a(i)=.;
            end;
      missing=nmiss(of R5--MCW4A1CATAa)/208;
      n=_n_;
      if missing>0.3 then n=.;
      drop missing i;
run;
data n1;
      set ratos_a (keep=n);
run;
proc transpose data=ratos_a out=ratos_a_transp;
run;
data ratos_a_transp;
      set ratos_a_transp;
      if _name_="n" then delete;
      n=_n_;
      missing=nmiss(of col1--col188)/188;
      if missing>0.3 then n=.;
run;
data n2;
      set ratos_a_transp;
      keep n;
run;

data ratos;
      set ratos0 (keep=R5--MCW4A1CATAa);
run;
data ratos;
      merge n1 ratos;
      if n=. then delete;
      drop n;
run;
```

```
proc transpose data=ratos out=ratos_aa;
run;
data ratos_aa;
      merge n2 ratos_aa;
      if n=. then delete;
      drop _label_ n;
run;
proc transpose data=ratos_aa out=ratos;
run;
data ratos;
      set ratos;
      drop _name_;
run;


/*    criando o conjunto de dados com as primeiras colunas sendo */
/*    os marcadores e as colunas seguintes sendo a distancia     */
/*    entre os marcadores                                        */

/*    O valor 179 se refere ao numero de indivíduos              */

data distancia1;
      set ratos;
      if distancia=0 then distancia=0.00001;
      keep distancia;
proc transpose data=distancia1 out=distancia1;
data distancia1;
      set distancia1;
      do i=1 to 179;output;
      end;
      drop _name_ i;
run;
data ratos_ordem_cr_resultantes;
      set ratos;
      keep cromossomo--distancia;
run;
proc transpose data=ratos out=ratos prefix=var;
      var col1--col179;run;
data ratos;
      merge ratos(drop=_name_) distancia1;
run;


/*    imputar os dados utilizando a tabela com as probabilidades */
/*    genotípicas de marcadores adjacentes de Haley & Knott(1992)*/

data ratos;
      set ratos;
      array a(166) var1--var166;
      array b(166) col1--col166;
      do i=2 to 165;

      if a(i)=20 and a(i-1)=0 and a(i+1)=0 then a(i)=sum(  b(i-1)*(1-
b(i-1))*b(i)*(1-b(i)),  2*b(i-1)*b(i-1)*b(i)*b(i))/sum(    b(i-1)*(1-
b(i-1))*b(i)*(1-b(i)),b(i-1)*b(i-1)*b(i)*b(i));

      if a(i)=20 and a(i-1)=0 and a(i+1)=1 then a(i)=sum(  (1-b(i-
1))*b(i-1)*((1-b(i))*(1-b(i))+b(i)*b(i)), 2*b(i-1)*b(i-1)*b(i)*(1-
b(i)))/sum( (1-b(i-1))*b(i-1)*((1-b(i))*(1-b(i))+b(i)*b(i)),b(i-1)*b(i-
1)*b(i)*(1-b(i)));
```

```
      if a(i)=20 and a(i-1)=0 and a(i+1)=2 then a(i)=sum(b(i-1)*(1-
b(i-1))*b(i)*(1-b(i)),2*(1-b(i-1))*(1-b(i-1))*(1-b(i))*(1-b(i)))/sum(
      b(i-1)*(1-b(i-1))*b(i)*(1-b(i)),(1-b(i-1))*(1-b(i-1))*(1-
b(i))*(1-b(i)));

      if a(i)=20 and a(i-1)=1 and a(i+1)=0 then a(i)=sum(((1-b(i-
1))*(1-b(i-1))+b(i-1)*b(i-1))*b(i)*(1-b(i)),2*(1-b(i-1))*b(i-
1)*b(i)*b(i))/sum(((1-b(i-1))*(1-b(i-1))+b(i-1)*b(i-1))*b(i)*(1-
b(i)),(1-b(i-1))*b(i-1)*b(i)*b(i));

      if a(i)=20 and a(i-1)=1 and a(i+1)=1 then a(i)=sum((1-b(i-
1))*(1-b(i-1))*(1-b(i))*(1-b(i))+b(i-1)*b(i-1)*b(i)*b(i)+(1-b(i-
1))*(1-b(i-1))*b(i)*b(i)+b(i-1)*b(i-1)*(1-b(i))*(1-b(i)),2*(1-b(i-
1))*b(i-1)*(1-b(i))*b(i))/sum((1-b(i-1))*(1-b(i-1))*(1-b(i))*(1-
b(i))+b(i-1)*b(i-1)*b(i)*b(i)+(1-b(i-1))*(1-b(i-1))*b(i)*b(i)+b(i-
1)*b(i-1)*(1-b(i))*(1-b(i)),(1-b(i-1))*b(i-1)*(1-b(i))*b(i));

      if a(i)=20 and a(i-1)=1 and a(i+1)=2 then a(i)=sum(((1-b(i-
1))*(1-b(i-1))+b(i-1)*b(i-1))*b(i)*(1-b(i)),2*(1-b(i-1))*b(i-
1)*b(i)*b(i))/sum(((1-b(i-1))*(1-b(i-1))+b(i-1)*b(i-1))*b(i)*(1-
b(i)),(1-b(i-1))*b(i-1)*b(i)*b(i)   );

      if a(i)=20 and a(i-1)=2 and a(i+1)=0 then a(i)=sum(b(i-1)*(1-
b(i-1))*b(i)*(1-b(i)),2*(1-b(i-1))*(1-b(i-1))*b(i)*b(i))/sum(b(i-
1)*(1-b(i-1))*b(i)*(1-b(i)),(1-b(i-1))*(1-b(i-1))*b(i)*b(i));

      if a(i)=20 and a(i-1)=2 and a(i+1)=1 then a(i)=sum((1-b(i-
1))*b(i-1)*((1-b(i))+(1-b(i))*b(i)*b(i)),2*(1-b(i-1))*(1-b(i-
1))*b(i)*(1-b(i)) )/sum((1-b(i-1))*b(i-1)*((1-b(i))+(1-
b(i))*b(i)*b(i)), (1-b(i-1))*(1-b(i-1))*b(i)*(1-b(i)) );

      if a(i)=20 and a(i-1)=2 and a(i+1)=2 then a(i)=sum(b(i-1)*(1-
b(i-1))*b(i)*(1-b(i)),2*(1-b(i-1))*(1-b(i-1))*(1-b(i))*(1-
b(i)))/sum(b(i-1)*(1-b(i-1))*b(i)*(1-b(i)),(1-b(i-1))*(1-b(i-1))*(1-
b(i))*(1-b(i))    );

      if a(i)=10 and a(i-1)=2 and a(i+1)=2 then a(i)=b(i-1)*(1-b(i-
1))*b(i)*(1-b(i))/sum(b(i-1)*(1-b(i-1))*b(i)*(1-b(i)),b(i-1)*b(i-
1)*b(i)*b(i));

      if a(i)=10 and a(i-1)=2 and a(i+1)=1 then a(i)=(1-b(i-1))*b(i-
1)*((1-b(i))*(1-b(i))+b(i)*b(i))/sum((1-b(i-1))*b(i-1)*((1-b(i))*(1-
b(i))+b(i)*b(i)), b(i-1)*b(i-1)*b(i)*(1-b(i)));

      if a(i)=10 and a(i-1)=2 and a(i+1)=0 then a(i)=b(i-1)*(1-b(i-
1))*b(i)*(1-b(i))/sum(b(i-1)*(1-b(i-1))*b(i)*(1-b(i)),(1-b(i-1))*(1-
b(i-1))*(1-b(i))*(1-b(i)));if a(i)=10 and a(i-1)=1 and a(i+1)=2 then
a(i)=((1-b(i-1))*(1-b(i-1))+b(i-1)*b(i-1) )*b(i)*(1-b(i))/sum(((1-b(i-
1))*(1-b(i-1))+b(i-1)*b(i-1)  )*b(i)*(1-b(i)),  (1-b(i-1))*b(i-
1)*b(i)*b(i));

      if a(i)=10 and a(i-1)=1 and a(i+1)=1 then a(i)=sum((1-b(i-
1))*(1-b(i-1))*(1-b(i))*(1-b(i))+b(i-1)*b(i-1)*b(i)*b(i)+(1-b(i-
1))*(1-b(i-1))*b(i)*b(i)+b(i-1)*b(i-1)*(1-b(i))*(1-b(i)),0*(1-b(i-
1))*b(i-1)*(1-b(i))*b(i))/sum((1-b(i-1))*(1-b(i-1))*(1-b(i))*(1-
b(i))+b(i-1)*b(i-1)*b(i)*b(i)+(1-b(i-1))*(1-b(i-1))*b(i)*b(i)+b(i-
1)*b(i-1)*(1-b(i))*(1-b(i)),(1-b(i-1))*b(i-1)*(1-b(i))*b(i));

      if a(i)=10 and a(i-1)=1 and a(i+1)=0 then a(i)=((1-b(i-1))*(1-
b(i-1))+b(i-1)*b(i-1))*b(i)*(1-b(i))/sum( ((1-b(i-1))*(1-b(i-1))+b(i-
1)*b(i-1))*b(i)*(1-b(i)),     (1-b(i-1))*b(i-1)*b(i)*b(i));
```

```
      if a(i)=10 and a(i-1)=0 and a(i+1)=2 then a(i)=b(i-1)*(1-b(i-
1))*b(i)*(1-b(i))/sum(b(i-1)*(1-b(i-1))*b(i)*(1-b(i)),     (1-b(i-
1))*(1-b(ib(i)*b(i));

      if a(i)=10 and a(i-1)=0 and a(i+1)=1 then a(i)=(1-b(i-1))*b(i-
1)*((1-b(i))*(1-b(i))+b(i)*b(i))/sum((1-b(i-1))*b(i-1)*((1-b(i))*(1-
b(i))+b(i)*b(i)), (1-b(i-1))*(1-b(i-1))*b(i)*(1-b(i)) );

      if a(i)=10 and a(i-1)=0 and a(i+1)=0 then a(i)=b(i-1)*(1-b(i-
1))*b(i)*(1-b(i))/ sum( b(i-1)*(1-b(i-1))*b(i)*(1-b(i)),(1-b(i-1))*(1-
b(i-1))*(1-b(i))*(1-b(i)));

      if a(i)=. and a(i-1)=2 and a(i+1)=2 then  a(i)=sum(0*(1-b(i-
1))*(1-b(i-1))*(1-b(i))*(1-b(i)), 1*b(i-1)*(1-b(i-1))*b(i)*(1-
b(i)),2*b(i-1)*b(i-1)*b(i)*b(i))/sum((1-b(i-1))*(1-b(i-1))*(1-
b(i))*(1-b(i)),b(i-1)*(1-b(i-1))*b(i)*(1-b(i)),b(i-1)*b(i-
1)*b(i)*b(i));

      if a(i)=. and a(i-1)=2 and a(i+1)=1 then  a(i)=sum(0*(1-b(i-
1))*(1-b(i-1))*b(i)*(1-b(i)), 1*b(i-1)*(1-b(i-1))*(1-b(i))*(1-
b(i))+b(i-1)*(1-b(i-1))*b(i)*b(i),2*b(i-1)*b(i-1)*b(i)*(1-b(i)))/sum(
      (1-b(i-1))*(1-b(i-1))*b(i)*(1-b(i)),b(i-1)*(1-b(i-1))*(1-
b(i))*(1-b(i))+b(i-1)*(1-b(i-1))*b(i)*b(i),b(i-1)*b(i-1)*b(i)*(1-
b(i)));

      if a(i)=. and a(i-1)=2 and a(i+1)=0 then a(i)=sum(   0*(1-b(i-
1))*(1-b(i-1))*b(i)*b(i), 1*b(i-1)*(1-b(i-1))*b(i)*(1-b(i)),2*b(i-
1)*b(i-1)*(1-b(i))*(1-b(i)))/ sum(  (1-b(i-1))*(1-b(i-
1))*b(i)*b(i),b(i-1)*(1-b(i-1))*b(i)*(1-b(i)),b(i-1)*b(i-1)*(1-
b(i))*(1-b(i)));

      if a(i)=. and a(i-1)=1 and a(i+1)=2 then a(i)=sum(   0*b(i-
1)*(1-b(i-1))*(1-b(i))*(1-b(i)), 1*(1-b(i-1))*(1-b(i-1))*b(i)*(1-
b(i))+b(i-1)*b(i-1)*b(i)*(1-b(i)),2*b(i-1)*(1-b(i-
1))*b(i)*b(i))/sum(b(i-1)*(1-b(i-1))*(1-b(i))*(1-b(i)),(1-b(i-1))*(1-
b(i-1))*b(i)*(1-b(i))+b(i-1)*b(i-1)*b(i)*(1-b(i)), b(i-1)*(1-b(i-
1))*b(i)*b(i));

      if a(i)=. and a(i-1)=1 and a(i+1)=1 then a(i)=sum(   0*b(i-
1)*(1-b(i-1))*b(i)*(1-b(i)),1*b(i-1)*b(i-1)*b(i)*b(i)+b(i-1)*b(i-
1)*(1-b(i))*(1-b(i))+(1-b(i-1))*(1-b(i-1))*b(i)*b(i)+(1-b(i-1))*(1-
b(i-1))*(1-b(i))*(1-b(i)), 2*b(i-1)*(1-b(i-1))*b(i)*(1-b(i)))/sum(
      b(i-1)*(1-b(i-1))*b(i)*(1-b(i)), b(i-1)*b(i-1)*b(i)*b(i)+b(i-
1)*b(i-1)*(1-b(i))*(1-b(i))+(1-b(i-1))*(1-b(i-1))*b(i)*b(i)+(1-b(i-
1))*(1-b(i-1))*(1-b(i))*(1-b(i)), b(i-1)*(1-b(i-1))*b(i)*(1-b(i)));

      if a(i)=. and a(i-1)=1 and a(i+1)=0 then  a(i)=sum(   0*b(i-
1)*(1-b(i-1))*b(i)*b(i), 1*(1-b(i-1))*(1-b(i-1))*b(i)*(1-b(i))+b(i-
1)*b(i-1)*b(i)*(1-b(i)), 2*b(i-1)*(1-b(i-1))*(1-b(i))*(1-b(i)))/ sum(
      b(i-1)*(1-b(i-1))*b(i)*b(i),  (1-b(i-1))*(1-b(i-1))*b(i)*(1-
b(i))+b(i-1)*b(i-1)*b(i)*(1-b(i)), b(i-1)*(1-b(i-1))*(1-b(i))*(1-
b(i)));

      if a(i)=. and a(i-1)=0 and a(i+1)=2 then  a(i)=sum(0*b(i-1)*b(i-
1)*(1-b(i))*(1-b(i)), 1*b(i-1)*(1-b(i-1))*b(i)*(1-b(i)),2*(1-b(i-
1))*(1-b(i-1))*b(i)*b(i))/sum(     b(i-1)*b(i-1)*(1-b(i))*(1-
b(i)),b(i-1)*(1-b(i-1))*b(i)*(1-b(i)),(1-b(i-1))*(1-b(i-
1))*b(i)*b(i));
```

```
      if a(i)=. and a(i-1)=0 and a(i+1)=1 then a(i)=sum(   0*b(i-
1)*b(i-1)*b(i)*(1-b(i)),1*b(i-1)*(1-b(i-1))*(1-b(i))*(1-b(i))+b(i-
1)*(1-b(i-1))*b(i)*b(i), 2*(1-b(i-1))*(1-b(i-1))*b(i)*(1-b(i)))/ sum(
      b(i-1)*b(i-1)*b(i)*(1-b(i)), b(i-1)*(1-b(i-1))*(1-b(i))*(1-
b(i))+b(i-1)*(1-b(i-1))*b(i)*b(i), (1-b(i-1))*(1-b(i-1))*b(i)*(1-
b(i)));

      if a(i)=. and a(i-1)=0 and a(i+1)=0 then a(i)=sum(   0*b(i-
1)*b(i-1)*b(i)*b(i), 1*b(i-1)*(1-b(i-1))*b(i)*(1-b(i)),2*(1-b(i-
1))*(1-b(i-1))*(1-b(i))*(1-b(i)))/ sum( b(i-1)*b(i-1)*b(i)*b(i),b(i-
1)*(1-b(i-1))*b(i)*(1-b(i)),(1-b(i-1))*(1-b(i-1))*(1-b(i))*(1-b(i)));

      end;
      keep var1--var166;
run;

data ratos;
      set ratos;
      array a(166) var1--var166;
            do i=1 to 166;
                  if a(i)=. then a(i)=1;
                  else if a(i)=10 then a(i)=2/3;
                  else if a(i)=20 then a(i)=4/3;
      end;
      drop i;
run;
```

Referências Bibliográficas

[1] Amos, C. I. (1994). Robust Variance-Components Approach for Assessing Genetic Linkage in Pedigrees. *Am. J. Hum. Genet.* 54(**3**): 535-543.

[2] Berg J., Tymoczko J. and Stryer L. (2002) *Biochemistry.* W. H. Freeman and Company

[3] Blangero, J.; Williams, J.T.; Itorria, S. J.; Almasy, L. (1999). Oligogenic Model Selection Using the Bayesian Information Criterion: Linkage Analysis of the P300 Cz Event-Related Brain Potential. *Genetic Epidemiology*. **17(Suppl I)**: S67-S72.

[4] Blangero, J.; Williams, J.T.; Almasy, L. (2001). Variance Component Methods for Detecting Complex Trait Loci. *Advances in Genetics.* **42**: 151-181.

[5] Chow, S.C.; Liu, J-P (2000). Design and Analysis of Bioavailability and Bioequivalence Studies. New York: Marcel Dekker.

[6] Chow, S.; Liu, J. (2004). *Design and Analysis of Clinical Trials*, John Wiley & Sons.

[7] Cotterman, C. 1941. The correlation between relatives in a random-mating population. *Sci Month*. **53**: 227234.

[8] Cook, R. D. e S. Weisberg (1982). *Residuals and Influence in Regression*. London: Chapman and Hall.

[9] Cook, R. D. e S. Weisberg (1994). *An Introduction to Regression Graphics*. Nova Yorque: John Wiley & Sons, INC.

[10] Darendorf, H. e Meibohm, (1999). Modeling of pharmacokinetics/pharmacdynamics (PK/PD) relationships: concepts and perspectives. *Pharmaceutical Research* **16**, 176-185.

[11] Davidian, M. and Giltinan, D.M. (1995), *Nonlinear Models for Repeated Measurement Data*, New York: Chapman & Hall.

[12] Doerge, R.W., Zeng, Z.B. & Weir, B.S. (1997). Statistical issues in the search for genes affecting quantitative traits in experimental populations. *Statistical Science*.**12**, 195-219.

[13] Devlin, B. & Roeder, K. (1999). Genomic control for association studies. Biometrics **55**:997–1004

[14] Duan S, Zhang W, Cox NJ, Dolan ME (2008). FstSNP-HapMap3: a database of SNPs with high population differentiation for HapMap3. *Bioinformatics* 3(**3**): 139-141.

[15] Duarte, N. E. (2007). *Análise multivariada no mapeamento genético de traços quantitativos*. Dissertação de mestrado. IME/USP, São Paulo

[16] Ewens, W. & Spielman, R. (1995). The Transmission/Disequilibrium Test: History, Subdivision, and Admixture. *Am. J. Hum. Genet*. **57**: 445-464.

[17] Falconer, D.S., Mackay, T. F C. (1996). *Introduction to quantitative genetics*. London: Oliver and Boyd Ltd.

[18] Farah, S.B. (1997). *DNA segredos e mistérios*. São Paulo: Sarvier.

[19] Fisher, R. T. (1918). The correlation between relatives on the supposition of Mendelian Inheritance. *Trans. R. Soc. Edinburg.* **52**: 399-433.

[20] Friendly, M. (1991). *SAS® System for Statistical Graphics*. Cary, NC: SAS Institute Inc.

[21] Frost & Sullivan. (2004). Strategic analysis of biomarkers in clinical trials.

[22] Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika* **58**, 453-467.

[23] Giolo, S; Soler, J.; Greenway, S.; de Andrade, M.; Seidman, J.; Seidman, C; Krieger, J.; Pereira, A. (2009) Brazilian Urban Population Genetic Structure Reveals a High Degree of Admixture , artigo sob submissão

[24] Giraldo, J. (2003). Empirical models and Hill coeficients. *Trends in Pharmacological Sciences.* **24**: 63-65.

[25] Gonçalves VF, Carvalho CMB, Bartolini MC, Bydlowski SP, Pena, SDJ (2008). The Phylogeography of African Brazilians. *Human Heredity* **65**:23–32.

[26] Goodman & Gilman. *As bases farmacológicas da terapêutica.* [tradução da 10. ed. original, Carla de Melo Vorsatz. *et al*] Rio de Janeiro: McGraw-Hill, 2005.

[27] Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* **53**: 325-338.

[28] Graybill, F. A. (1976). *Theory and Application of the Linear Model*. California: Wadsworth Publishing Company.

[29] Gregory S, et al. (2006). The DNA sequence and biological annotation of human chromosome 1. *Nature* 441 (**7091**): 315–21.

[30] Haley, C. S. & Knott, S. A. (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* **69**: 315-324

[31] Hartl, D. & Clark, A. (1997). *Principles of Population Genetics*. Sinauer.

[32] Henderson, C.R. (1990), Statistical Method in Animal Improvement: Historical Overview. *Advances in Statistical Methods for Genetic Improvement of Livestock*, New York: Springer-Verlag, 1 - 14.

[33] Henderson, C.R. (1975). Best Linear Unbiased Estimation and Prediction under a Selection Model. *Biometrics* **31**: 423-447.

[34] Hilden-Minton, J.A. (1995). *Multilevel Diagnostics for Mixed and Hierarchical Linear Models*. PhD Thesis. University of California, Los Angeles.

[35] International HapMap Project, Phase III. Available at http://www.sanger.ac.uk/humgen/hapmapap3

[36] Jansen, R.C. (1993). Maximum likelihood in a generalized linear finite mixture model by using the EM algorithm. *Biometrics*. **49**, 227-231.

[37] Jansen, R.C. & Stam, P. (1994). High resolution of quantitative traits into multiple loci via interval mapping. *Genetics*. **136**: 1447-1455.

[38] Jansen, R. C. (2003). *Quantitative Trait Loci in Inbreed Lines*, no Handbook of Statistical Genetics.

[39] Jiang, C. & Zeng, Z.B. (1995). Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics* **140**: 1111-1127.

[40] Johnson, R. A. & Wichern, D. W. (2002). *Applied Multivariate Statistical Analysis*. New Jersey: Prentice Hall.

[41] Kao, C. H. & Zeng, B. Z. (1999). Modelling Epistasis of Quantitative Trait Loci Using Cockerham's Model. *Genetics*. **160**, 1243-1261.

[42] Kempthorne, O. (1957). *An Introduction to Genetic Statistics*. Nova Yorque: John Wiley & Sons, Inc.

[43] Kirk, R.; Hung, J.; Horner, S.; Perez, J. (2008). Implications of Pharmacogenomics for Drug Development. *Exp. Biol Med.* **233**:1484-1497.

[44] Lander, E. S. & Botstein, D. (1989). Mapping mendelian factors underlying quantitativestraits using RFLP linkage maps. *Genetics*. **121**, 185-199.

[45] Lange, K. (1997). Mathematical and statistical methods for genetic analysis. New York: Springer.

[46] Littell, R.C., Milliken, G.A., Stroup, W.W., and Wolfinger, R.D. (1996), *SAS System for Mixed Models*. Cary, NC: SAS Institute Inc.

[47] Liu, B.H. (1998). Statistical Genomics: Linkage, Mapping, and QTL Analysis. Boca Raton FL: CRC Press.

[48] Lynch, M. & Walsh, B. (1998). *Genetics and analysis of quantitative traits*. Sunderland. Sinauer Associates.

[49] McCulloch, C.E. & Searle, S.R. (2001). *Generalized, Linear, and Mixed Models*. Nova Iorque: John Wiley & Sons.

[50] McQueen MB, Bertram L, Rimm EB, Blacker D, Santangelo SL (2003): A QTL genome scan of the metabolic syndrome and its component traits. *BMC Genet* 4 (**1**):S96.

[51] Mendel, G. (1986). Experiments in a monastery garden. *American Zoologist*. 26(**3**):749-52.

[52] Mood, A.; Graybill; F. Boes, D. (1974). *Introduction to the Theory of Statistics*, Singapura: McGraw-Hill Book Company

[53] Nobre, J. S. (2004). *Métodos de Diagnóstico para Modelos Lineares Mistos*. Dissertação de mestrado. IME/USP, São Paulo.

[54] Oliveira, C.; Pereira, A.; Andrade, M.; Soler, J.; Krieger, J.(2008): Heritability of cardiovascular risk factors in a Brazilian population: Baependi Heart Study. *BMC Medical Genetics* 9:32

[55] Oliveira, P. (2008). *Aplicação do algoritmo genético no mapeamento de genes epistáticos em cruzamentos controlados*. Tese para obtenção do grau de Doutor em Ciências. Área de concentração: Estatística. Instituto de Matemática e Estatística da Universidade de São Paulo. São Paulo.

[56] Parra, F.C. et al. (2003) Color and genomic ancestry in Brazilians. *Proc. Natl. Acad. Sci. U. S. A.* **100**: 177–182

[57] Pierce, B. (2005). *Genetics: A Conceptual Approach*.EUA: Freeman

[58] Price, A.; Patterson, N.; Plenge, R.; Weinblatt, M.; Shadick, N.; & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetcs*. **38**: 904-909.

[59] Price,A.; Tandon, A.; Petterson, N.; Barnes, K.; Rafael, N.; Ruczinski, I.; Beaty, T; Mathias, R; Reich, D.; Myers, S. (2009). Sensitive Detection of chromosomal Segments of Distinct Ancestry in Admixed Populations. *Plos Genetics*, 4(**1**):e236.

[60] Pritchard, J. & Rosenberg, N. (1999): Use of Unlinked Genetic Markers to Detect Population Stratification in Association Studies. Am. J. Hum. Genet. **65**:220-228.

[61] Pritchard, J.K. et al. (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959.

[62] Rapley, R. & Harbon, S. (2004). *Molecular Analysis and Genome Discovery*. Chichester. John Wiley & Sons LTDA.

[63] Sall, J., Lehman, A. e Creighton, L. (2001). *JMP® Start Statistics*. NC: SAS Institute Inc.

[64] Salzano, F.M. & Bortolini, M.C. (2002). *The Evolution and Genetics of Latin American Populations*, Cambridge University Press

[65] Schork, N.J., Krieger, J.E., Trolliet, M.R., Franchini, K.G., Koike, G., Kriger, E.M., Lander, E.S., Dzau, V.J. & Jacob, H.J. (1995). A biometrical genome search in rats reveals the multigenetic basic of blood pressure variation. *Genome Research*. **5**, 164-172.

[66] Schuster, I. & Cruz, C. (2004). Estatística Genômica Aplicada a Populações Derivadas de Cruzamentos Controlados. Editora UFV, Viçosa MG.

[67] Searle, S. R., Casella, G., e McCulloch, C.E. (1992), *Variance Components*, New York: John Wiley & Sons, Inc

[68] Self, S. G. Liang, K. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio test under nonstandard conditions *Am. Stat.Assoc*. **82**, 605-610.

[69] Senn, S. (2007). *Statistical Issues in Drug Development*. New York: John Wiley & Sons, Inc

[70] Shmulewitz, D.; Zhang, J., Greenberg, D. (2004). Case-Control Associations Studies in Mixed Populations: Correcting Using Genomic Control. *Human Heredity*. **58**:145-153

[71] Shmulewitz D, Heath SC, Blundell ML, Han Z, Sharma R, Salit J, Auerbach, SB, Signorini S, Reslow JL, Stoffel M, Friedman JM (2006). Linkage analysis of quantitative traits for obesity, diabetes, hypertension and dyslipidemia on the island of Kosrae, Federated States of Micronesia. *Proc Natl Acad Sci EUA* **103**:3502-3509.

[72] Silverman, B. W. (1986), *Density Estimation*, Nova Iorque: Chapman and Hall

[73] Singer, J.M. & Andrade, D.F. (2000). Analysis of longitudinal data. Handbook of Statistics, Volume 18: Bio-environmental and Public Health Statistics. Eds. P.K.

[74] Slatkin, M. (1991). Inbreeding coefficients and coalescence times. *Genet. Res*. **58**:167-175.

[75] Suarez-Kurtz, G. (2005). Pharmacogenomics in admixed populations. *TRENDS in Pharmacological Sciences*. **26**:196-201.

[76] Sweeney BP (2005). Pharmacogenomics: the genetic basis for variability in drug response. In: Cashman JN, Grounds RM, Eds. Recent Advances in Anaesthesia and Intensive Care. United Kingdom: *Cambridge University Press* **23**:1–34.

[77] Tiwari, H.; Barnholtz-Sloan, J.; Weneiger, N.; Padilla, M; Vaughan, L & Allison, D. (2008). Review and Evaluation of Methods Correcting for Population Stratification with a Focus on Underlying Statistical Principles. *Hum. Hered*. **66**:67-86.

[78] Viana, J M.; Cruz, C. D.; Barros, E. G. (2003). *Genética. Volume 1 – Fundamentos*, Editora UFV, Viçosa – MG.

[79] Vonesh, E.F. and Chinchilli, V.M. (1997), Linear and Nonlinear Models for the Analysis of Repeated Measurements. New York: Marcel Dekker.

[80] Wang Q. Molecular genetics of coronary artery disease (2005). *Curr Opin Cardiol* 2005;**20**:182-8.

[81] Wang D, Bakhai A (2006). Clinical Trials: A Practical Guide to Design, Analysis, and Reporting. London: Remedica.

[82] Watson J & Crick F (1953). Molecular structure of nucleid acids: a structure for deoxyribose nucleid acid. *Nature* 171 (**4356**): 737–8

[83] Weisber, S. (1985). *Applied Linear Regression*, Nova Iorque: John Wiley & Sons, INC

[84] Wolfinger, R. (1993). Covariate structure selection in general mixed models. *Communications in Statistics – Simulation*. **22**: 1079-1106.

[85] Wu, R. & Lin, M. (2009). *Statistical and Computational Pharmacogenomics*. Florida: Chapman & Hall.

[86] Zeng, Z.B. (1993). Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. *Procedures of the Natinal Academics Science USA*. **90**, 10972-10976.

[87] Zeng, Z.B. (1994). Precision mapping quantitative trait loci. *Genetics*. **136** , 1457-1468.

[88] Zhang, S., Zhu, X. & Zhao, H. (2003). On a Semiparametric Test to Detect Associations Between Quantitative Traits and Candidate Genes Using Unrelated Individuals. *Genetic Epidemiology*. **24**, 44-56.