UMA METODOLOGIA PARA A DESCOBERTA DE MARCADORES GENÉTICOS EM ESTUDOS DE ASSOCIAÇÃO

Margarita Ramona Ruiz Olazar

Tese de Doutorado apresentada ao Programa de Pós-graduação em Engenharia Elétrica, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Doutor em Engenharia Elétrica.

Orientador: Prof. Eugenius Kaszkurewicz

Rio de Janeiro
Maio de 2013

UMA METODOLOGIA PARA A DESCOBERTA DE MARCADORES GENÉTICOS EM ESTUDOS DE ASSOCIAÇÃO

Margarita Ramona Ruiz Olazar

TESE SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTOR EM CIÊNCIAS EM ENGENHARIA ELÉTRICA.

Examinada por:

_____
Prof. Eugenius Kaszkurewicz, D.Sc.

_____
Prof. Amit Bhaya, Ph.D.

_____
Prof. Alberto Martin Rivera Davila, D.Sc.

_____
Prof. Beatriz de Souza Leite Pires de Lima, D.Sc.

_____
Prof. Andre Ponce de Leon F. de Carvalho, D.Sc.

RIO DE JANEIRO, RJ - BRASIL

MAIO DE 2013

Olazar, Margarita Ruiz

Uma metodologia para a descoberta de marcadores genéticos/ Margarita Ramona Ruiz-Olazar. – Rio de Janeiro: UFRJ/COPPE, 2013.

XVI, 133 p.: il.; 29,7 cm.

Orientador: Eugenius Kaszkurewicz

Tese (doutorado) – UFRJ/ COPPE/ Programa de Engenharia Elétrica, 2013.

Referencias Bibliográficas: p. 115-124.

1. Estudos de associação do genoma inteiro. 2. Polimorfismo de nucleotídeo único. 3. Informação mútua 4. Algoritmos genéticos. I. Kaszkurewicz, Eugenius. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Elétrica. III. Titulo.

Em memória de minha irmã, Carmen Mercedes Ruiz Olazar (1962–2013).

AGRADECIMENTOS

A Deus, por todas as oportunidades e bênçãos na minha vida.

A minha família que sempre acreditou no meu potencial e me deu seu apoio incondicional em todos os momentos de minha existência. A minha mãe, meus irmãos e irmãs, meus sobrinhos e sobrinhas, por me dar sempre a força e o alento necessários nestes anos longe de casa.

Ao professor Eugenius Kaszkurewicz, quem me acolheu como sua orientada e possibilitou que a minha experiência como aluna de doutorado fosse a mais proveitosa possível. Agradeço-lhe pelo inestimável apoio e pela dedicação do seu tempo determinantes para o sucesso deste trabalho.

Ao professor Amit Bhaya, pela ajuda e esforço para que esta pesquisa pudesse ter um curso proveitoso e produtivo.

Aos colegas Marcelo Ribeiro Alves da Fiocruz e Raquel Barbosa do INCA por sua amizade, carinho, colaboração sincera e preciosa ajuda que tornaram esta pesquisa apaixonante.

A meus amigos, do NACAD - Núcleo de Atendimento de Computação de Alto Desempenho, por sua cordial acolhida, amizade e ajuda.

À Universidade Federal do Rio de Janeiro, e em especial ao Programa de Engenharia Elétrica da COPPE por ter contribuído para minha formação e ter recebido e dado oportunidade a uma estudante paraguaia.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPQ) pelo fundamental apoio financeiro para o cumprimento desta pesquisa.

Meus sinceros agradecimentos!

Resumo da Tese apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Doutor em Ciências (D.Sc.)

UMA METODOLOGIA PARA A DESCOBERTA DE MARCADORES GENÉTICOS EM ESTUDOS DE ASSOCIAÇÃO

Margarita Ramona Ruiz Olazar

Maio/2013

Orientador: Eugenius Kaszkurewicz

Programa: Engenharia Elétrica

Este trabalho desenvolve uma metodologia para ajudar a descobrir marcadores genéticos (ex. SNPs, do inglês, *Single Nucleotide Polymorphims*) em Estudos de associação do genoma inteiro (GWAS, do inglês, *Genome Wide Association Studies*), abrangendo desde aspectos fundamentais do controle de qualidade dos dados até a identificação dos haplótipos potenciais de risco de desenvolvimento da doença estudada. Foram feitos testes com 82 conjuntos de dados de diferentes modelos epistáticos gerados através de simulações e também com 5 conjuntos de dados reais de doenças complexas (Diabetes Mellitus tipo 1, Diabetes Mellitus tipo 2, Desordem bipolar, Hipertensão e Doença arterial coronária), estes dados são provenientes da Wellcome Trust Case Control Consortium (WTCCC) do Reino Unido. Para identificar os SNPs que interagem com a doença estudada foi desenvolvido um algoritmo, chamado ***MIGA-2L***, que esta baseado na teoria da informação mútua em combinação com um algoritmo genético executado sobre mascaras de grupos de SNPs com o objetivo de otimizar a busca. Também foi feita uma análise comparativa do ***MIGA-2L*** com o programa ***Plink***, executado sobre um cluster SGI Altix ICE 8400 utilizando os conjuntos de dados mencionados anteriormente. Os resultados obtidos, mostrados tanto com medidas de desempenho computacionais como epidemiológicas, confirmam que a metodologia proposta pode ser uma ferramenta computacional útil e rápida para realizar GWAS em dados reais.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Doctor of Science (D.Sc.)

A METHODOLOGY TO DISCOVER GENETIC MARKERS IN ASSOCIATION STUDIES

Margarita Ramona Ruiz Olazar

May/2013

Advisor: Eugenius Kaszkurewicz

Department: Electrical Engineering

This work presents a methodology to discover genetic markers (SNPs) in GWAS covering from fundamental aspect of data quality control until the identification of the haplotypes that suggest risk of developing of the disease under study. The presented methodology is based on workflow technologies to take advantage of the flexible characteristics offered by the workflow engine to model and manage resources and reduce the time needed to perform the complex analysis involved in the fundamental steps in a GWAS, starting from raw data. A algorithm for test interaction SNP-SNP was developed, called ***MIGA-2L***, that is based on mútual information in combination with a genetic algorithm that runs on masks of groups of SNPs to optimize the search. The methodology was tested on 82 different epistatic models of simulated datasets and also on five WTCCC dataset (Diabetes Type 1, Diabetes Type 2, Coronary artery disease, Hypertension and Bipolar disorder) from United Kingdom. A comparative analysis of the algorithm ***MIGA-2L*** was performed with the program ***Plink*** which is typically used in this type of association studies, these experiment were ran on a cluster SGI Altix ICE 8400 using the dataset mentioned above. Showing these results with computational as epidemiologic performance measures. The results obtained confirm that this methodology can be a useful computational tool to perform genome-wide case-control studies on real datasets.

ÍNDICE

LISTA DE FIGURA S

LISTA DE TABELAS

SÍMBOLOS E ABREVIATURAS

DNA *Acido Desoxirribonucléico*

EA *Algoritmos Evolucionários*

GWAS *Estudos de associação do genoma inteiro*

GWA *Associação genômica ampla*

LD *Desequilíbrio de ligação*

EHW *Equilíbrio de Hardy-Weinberg*

SNP *Polimorfismo de Nucleotídeo único/simples*

AG *Algoritmo genético*

NN *Redes Neurais*

WTCCC *Welcome Trust Case-Control Consortium*

MAF *Frequência do menor alelo*

MDR *Multifactor Dimensionality Reduction*

DENOMINAÇÕES E CONCEITOS BÁSICOS

Alelo: um possível estado de um loco polimórfico. Por exemplo, um SNP poderia ter alelos G e T.

Doença complexa: Um fenótipo com uma etiologia multifatorial, freqüentemente consistindo de vários componentes genéticos e ambientais.

Equilíbrio de ligação (LD): Associação de alelos em dois loci devido a um fenômeno diferente da chance aleatória.

Estudos de associação: Buscam relacionar um marcador genético particular com uma doença complexa em uma população.

Fenótipo: são as características observáveis de um indivíduo como, por exemplo: cor de olhos, pressão sanguínea, ou presença de uma doença particular.

Genome-wide association study (GWAS): são Estudos de associação do genoma inteiro no qual 100,000 ou mais marcadores SNPs são testados em amostras individuais de DNA.

Genótipo: um conjunto de alelos presentes em um loco particular. Genótipos humanos têm dois alelos, um herdado de cada parente.

Genotipagem: é um processo de determinação do genótipo ou conteúdo genômico, na forma do DNA, específico de um organismo biológico, mediante um procedimento de laboratório.

Haplótipo: Um haplótipo é uma combinação de alelos em loci adjacentes, que fazem parte do mesmo cromossomo e são transmitidos juntos. Um haplótipo pode ser formado por um ou vários alelos, ou até pelo cromossomo inteiro.

Heterogeneidade genética: Na clínica, a heterogeneidade genética se refere à presença de vários defeitos genéticos que causam a mesma doença, frequentemente devido a mutações em locais diferentes no mesmo gene, um achado comum a muitas doenças humanas, inclusive a Doença De Alzheimer, Fibrose Cística, Deficiência De Lipoproteína Lípase familiar e Neuropatias Policísticas.

Lócus: (do latim "lugar", no plural loci) é o local fixo num cromossomo onde está localizado determinado gene ou marcador genético.

Marcador genético: é um gene ou posição no genoma que existe em dois ou mais alelos distinguíveis e cuja herança pode ser seguida através de um cruzamento genético, permitindo mapear a posição de um gene a determinar. Ex. SNPs.

Penetrância: é um termo utilizado em genética para descrever a proporção de indivíduos portadores de uma variação particular de um gene, que igualmente se expressa no seu fenótipo.

Polimorfismos genéticos: são variantes do genoma que aparecem por mutações em alguns indivíduos, se transmitem à descendência e adquirem certa frequência na população após múltiplas gerações. Os polimorfismos mais frequentes são os de base única chamados SNPs.

Polimorfismo de nucleotídeo simples (SNP): um loco com uma única substituição de base. Devido a sua abundância e fácil detecção, SNPs são usados frequentemente como marcadores em Estudos de associação do genoma inteiro.

Principio de Hardy-Weinberg (EHW): A situação na qual as frequências dos alelos e genótipos permanecem constantes em uma população durante gerações. Quando no EHW, as frequências dos alelos A e B para um loco bi-alélico em uma população diplóide são esperados a estar relacionados a seus genótipos por $Pr(AA) = Pr^2(A)$, $Pr(AB) = Pr(A)Pr(B)$, $Pr(BB) = Pr^2(B)$.

CAPÍTULO 1: INTRODUÇÃO

Este capítulo aborda a importância dos Estudos de associação do genoma inteiro para a descoberta de marcadores genéticos de doenças no contexto das ciências biológicas e ciência computacional e suas principais contribuições para estudos do mesmo gênero. Na última seção se detalha a estrutura deste documento de uma maneira geral.

1.1 MOTIVAÇÃO

Os Estudos de associação do genoma inteiro, do inglês *Genome Wide Association Studies* (GWAS), são uma forma relativamente nova de identificar genes envolvidos em doenças humanas. Os cientistas procuram pequenas variações ou polimorfismos no genoma que ocorrem mais frequentemente em pessoas com uma determinada doença do que em pessoas sem a doença. Cada estudo pode analisar centenas ou milhares destes polimorfismos ao mesmo tempo.

Habitualmente, neste tipo de análise são utilizados polimorfismos de nucleotídeo simples ou SNPs (pronunciado "snips") que são a forma de variação mais frequente no Genoma, que acontece quando um dos nucleotídeos (A,C,G,T) difere entre indivíduos em um determinado local cromossômico (*lócus*). Os SNPs não causam doenças, eles ajudam a estabelecer localizações, no genoma, de algum fator genético que contribui à variabilidade. Estes pontos de referencia são conhecidos como marcadores genéticos e facilitam a navegação no genoma humano (Figura 1.1).

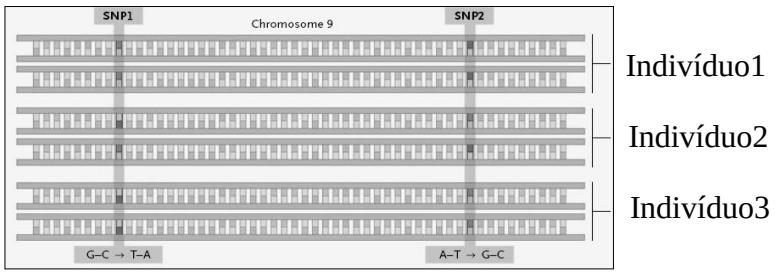


Figura 1.1: Este painel mostra a sequencia de genótipos de 3 indivíduos. Pode-se observar dois SNPs localizados em uma pequena porção do cromossomo 9. Fonte: Manolio 2010, N Engl J Med.

Pesquisadores procuram por associação entre um certo traço de interes ou fenótipo e alelos ou genótipos em um lócus genético com o objetivo de determinar se um alelo particular associa-se com certa doença na população como um todo. Esta associação pode indicar uma relação causal direta que permitiria investigar os mecanismos e vias metabólicas (chamados *pathways*) na progressão da doença ou pode indicar uma relação indireta que pode ajudar a localizar a variante causal.

GWAS já identificaram SNPs de várias doenças complexas, incluindo Parkinson [LI et al, 2012], Câncer de mamas [Ghoussaini et al, 2012; Easton et al. 2007], Artrite reumatoide [Kurreeman et al, 2012; Plenge et al. 2007] e Diabetes tipo 1 e tipo 2 [LU et al, 2012; Fagerholm et al, 2012; Todd et al, 2007; Zeggini et al. 2008]. Um estudo muito bem sucedido é da Wellcome Trust Case-Control Consortium (WTCCC), onde uma população de indivíduos de 7 diferentes doenças foram analisadas, encontrando 24 sinais de associações independentes em 6 das 7 doenças estudadas [WTCCC 2007]. Apesar de os primeiros resultados em GWAS terem sido relatados em 2005 [Klein et al , 2005] e 2006 [Dewan et al, 2006], a comunidade científica considera este estudo da WTCCC, publicado no journal *Nature* em 2007, como o ponto de partida dos estudos GWAS.

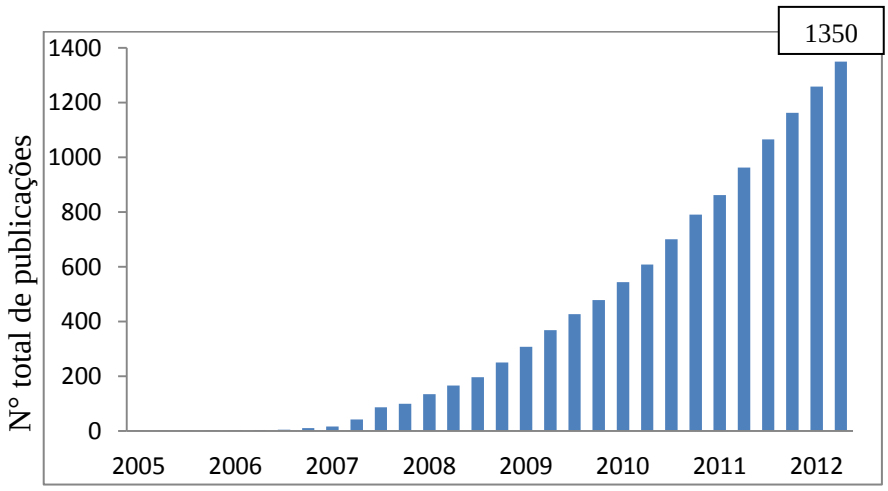


Figura 1.2: Esta Figura apresenta uma gráfico de GWAS realizados nos últimos anos. Os GWAS considerados aqui incluem aqueles estudos com ao menos 100,000 SNPs na fase inicial. (Fonte http://www.genome.gov/26525384)

Até julho de 2012, foram relatados em publicações de jornais científicos mais de 1,300 estudos GWAS, como mostrado na Figura 1.2, neles são reportados aproximadamente 2,000 loci significativamente e fortemente associados com uma ou mais características complexas [Visscher et al, 2012].

Os pesquisadores esperam descobrir mais SNPs associados com doenças crônicas, assim como entender a forma como estas variações afetam a resposta de uma pessoa a certas drogas e os efeitos produzidos pelas interações entre os genes e o meio ambiente. Para este propósito, se torna essencial em GWAS o uso de softwares especializados para lidar com a grande dimensionalidade dos dados e realizar um grande número de testes para investigar associações diminuindo a complexidade computacional.

1.2 ESTUDOS DE ASSOCIAÇÃO DO GENOMA INTEIRO

Com a conclusão do Projeto Genoma Humano em 2003 [*http://www.genome.gov/*] e o inicio do Projeto Internacional HapMap em 2002 [*http://hapmap.ncbi.nlm.nih.gov/*], os pesquisadores têm agora um conjunto de ferramentas de investigação que tornam possível pesquisas das contribuições genéticas de doenças comuns. As ferramentas incluem bancos de dados que contêm a referência da sequencia do genoma humano [*http://www.ensembl.org/*;*http://www.ncbi.nlm.nih.gov/genome*;*http://www.ebi.ac.uk/*; *http://genome.ucsc.edu/*], também catálogos da variação genética humana (mapas dos marcadores genéticos)[*http://www.genome.gov/*] e um conjunto de novas tecnologias de alto rendimento de genotipagem (ex. Affymetrix e Illumina) que podem, de forma rápida e precisa, coletar amostras do genoma de um indivíduo[*http://www.affymetrix.com/*; *http://www.illumina.com*].

Estes avanços contribuíram para um considerável incremento de publicações científicas. Nos últimos 4 ou 5 anos uma série de GWAS de alto perfil, principalmente em desenhos casos/controles, como foi mostrado na Figura 1.2. Desafortunadamente os resultados não foram tão úteis como se esperava [Juyal et al, 2011]. Um número substancial de recentes estudos GWAS indicam que para a maioria das doenças, apenas algumas variantes comuns estão envolvidas, e os SNPs associados explicam apenas uma pequena fração do risco genético [Cantor et al., 2010] [Hindorff et al., 2010]. A proporção da variação genética explicada por SNPs significativamente associados é geralmente baixa (tipicamente menos de 10%) para muitas características complexas. Em Doenças tais como a doença de Crohn e a esclerose múltipla, e para as características quantitativas tais como a altura e traços de lípidos, entre 10% e 20% da variância genética foi contabilizada [Visscher et al,

2012]. Porém, em comparação com a era pré-GWAS, a proporção da variação genética explicada por variantes recém-descobertas que são segregadas na população é grande.

A maioria dos estudos são focados na análise de um único lócus, que avalia diretamente associação entre um SNP e a variante fenotípica. Sem embargo, considerar o efeito de interações gene-gene ou SNP-SNP, assim como as interações do gene-ambiente podem também desempenhar um papel significativo na determinação do fenótipo. Este fenômeno de interes biológico, chamado epistasis, ajudaria a identificar genes que interagem para causar doenças e para entender os mecanismos e vias metabólicas na progressão dessas doenças [Cordell 2002; Cordell 2009]. Neste trabalho, estamos focados na utilização de métodos de detecção de interacção SNP-SNP ou de deteção de "epistasia estatística" para a descoberta de marcadores genéticos associados à doenças, de acordo com a definição de [Phillips 2008].

Complexidade computacional

É importante notar que o número das variantes genéticas (ex. SNPs) descobertas está fortemente correlacionado com o tamanho da amostra experimental. Futuras pesquisas em GWAS terão que atingir um limiar mínimo de tamanho da amostra para detectar variantes com alto índice de confiança. Isto levanta um desafio computacional grande na identificação de interações genéticas que estão potencialmente associadas a doenças. Na Figura 1.3, pode-se perceber a tendência de que um tamanho cada vez maior da amostra irá aumentar o número de variantes descobertas.

Neste contexto, muitas abordagens computacionais baseadas em métodos estatísticos foram propostas. Com poucas exceções, elas caem dentro de uma de duas categorias. As que explicitamente testam cada interação possível entre marcadores, ou seja, procuram em todo o espaço de busca, e aquela que evitam uma enumeração exaustiva do espaço de busca. Dentro desta ultima categoria se encontram os algoritmos estocásticos que realizam uma investigação probabilística do espaço de busca e os algoritmos gulosos que simplesmente fazem a melhor escolha baseado na informação disponível. Em particular, os que realizam um teste completo de todas as

possíveis interações entre marcadores genéticos são computacionalmente complexos e inaceitáveis.

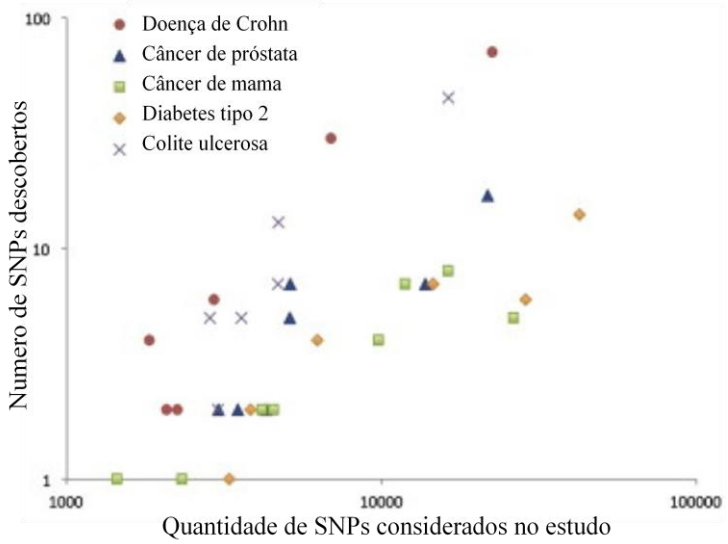


Figura 1.3 O gráfico apresenta uma relação entre o número de achados e a quantidade de SNPs considerados em GWAS. Fonte: Visscher et al, 2012.

A quantidade de testes necessários para investigar interação de marcadores (por exemplo SNPs) em GWAS depende da quantidade de marcadores considerado no estudo, como também do número de loci investigado. O número de testes pode-se calcular com a equação 1.1, onde *L* é o número total de marcadores e *n* é número de loci considerado .

$$N° \, Testes = \frac{L*(L-1)}{n!} \qquad 1.1$$

Por exemplo, um conjunto de dados com 500,000 mil marcadores precisará aproximadamente 125 bilhões de testes investigando 2 loci. Com um computador executando 1000 testes por segundo, deveríamos esperar aproximadamente 238 anos para verificar os resultados. A complexidade aumenta exponencialmente com o número de marcadores e a quantidade de loci considerados. Na prática, esta estratégia é limitada a um pequeno conjunto de marcadores. No capítulo 4 serão descritas algumas abordagens computacionais que ajudam a investigar interações de marcadores com doenças comuns.

1.3 IDENTIFICAÇÃO DE INTERAÇÃO DE SNP-SNP

Uma variedade de métodos foram desenvolvidos na busca de efeitos de interação em doenças complexas, incluindo análise de regressão [Marchini et al. 2005; Kooperberg et al. 2005; Park et al. 2008; Yang C et al. 2010; Purcell et al. 2007], inferência Bayesiana [Zhang et al. 2007], técnicas de aprendizado de máquina [Ritche MD 2001; Breiman 2001], cálculo de entropia [Miller DJ et al. 2009], entre outros. Abordagens interessantes são os métodos usados na teoria da informação que podem ser utilizados em análise genômica para proporcionar uma visão mais significativa do processo genético. Basicamente, a ideia é que usando um modelo de transferência de informação entre certos polimorfismos no genoma humano (SNPs) e certas doenças (ex. Alzheimer), a informação mútua de Shannon [Cover & Thomas, 2006] pode identificar os SNPs potenciais de associação com a doença. Alguns trabalhos relacionados podem-se encontrar em [Hagenauer et al. 2004; Jakulin & Bratko, 2004; Moore et al. 2006].

Mesmo escolhendo um método robusto para testar interações entre marcadores, uma busca exaustiva pode resultar em execuções muito lentas quando a dimensionalidade dos dados aumenta. Muitos cientistas carecem dos recursos computacionais requeridos para implantar estes métodos na escala genômica. Apesar do aumento do numero de publicações propondo novas metodologias, algumas simplesmente ajustam o tamanho do conjunto de dados filtrando um grupo pequeno de marcadores para investigar interação, esta estratégia muitas vezes falha ao excluir SNPs que apresentam interações importantes e, portanto, não conseguem identificar todos os marcadores que conferem risco.

Para abordar estas deficiências, este trabalho apresenta uma metodologia que viabiliza o uso de marcadores genéticos em escala genômica abrangendo desde o controle de qualidade dos dados até a seleção e classificação dos haplótipos potenciais que indicam associação com a doença estudada. A metodologia proposta utiliza um motor de workflow de forma a modelar e otimizar a execução das várias etapas do processo envolvido em GWAS. Desta forma, visamos proporcionar uma metodologia flexível, rápida e robusta que pode manipular dados de alta dimensionalidade e integrar vários tipos de programas e métodos bioestatísticos para controle de qualidade e análise de interação, podendo ser estendido para estudos posteriores tais

como replicação e meta-análise, assim como estudos de identificação de vias metabólicas e de alvos para drogas.

Na fase de análise da interação propriamente dita foi utilizado um algoritmo genético (AG) para a otimização da busca e identificação de associações relevantes. O AG é um paradigma evolucionário [Goldberg, 1989], onde um algoritmo realiza uma busca estocástica baseado no processo da evolução Darwiniana a fim de encontrar soluções para problemas computacionalmente complexos. Os AGs são apropriados para estudos GWAS já que através de um processo aleatório, tal como seleção natural, mutação e cruzamento, investigam só um subconjunto de todas as possíveis interações. Contudo, havendo descoberto uma interação importante, ele é capaz de preservar este padrão em futuras gerações.

Na literatura foi possível encontrar algumas abordagens que utilizam paradigmas evolucionários, alguns deles para melhorar o desempenho dos métodos de busca estocástica. Neste contexto, pode-se citar a programação genética [Moore et al, 2004] e AG combinado com o algoritmo de colônia de formigas [Greene et al., 2008]. Também foi implementado um método que usa evolução gramatical de redes neurais [Motsinger-Reif et al., 2008] onde um algoritmo evolucionário '*grammatical evolution*' é utilizado para construir redes neurais (NN) e selecionar os SNPs associados à doenças.

Do mesmo modo, Shah et al [Shah et al., 2004] utilizaram um algoritmo genético para realizar seleção característica construindo árvores de decisão e Clark et al. [Clark et al., 2005; Clark et al., 2008] utilizaram um AG para construir árvores de decisão de expressões booleanas construídas com blocos de SNPs. Em uma publicação recente, um AG foi implementado como uma estratégia adaptativa evolutiva em combinação com uma abordagem baseada em desequilíbrio de ligação para identificar interação de *loci* [Fontanarosa, Yang; 2010].

A metodologia aqui apresentada compara dois grupos de indivíduos, aqueles relacionados com a doença (chamado casos) e aqueles sadios (chamados controles). Foram realizados experimentos utilizando dados simulados de 82 modelos epistáticos diferentes, assim como experimentos com dados reais de genótipos de 5 doenças, com amostras originadas no Reino Unido e fornecidos pela WTCCC [http://www.wtccc.org.uk/].

1.4 ESTRUTURA DO DOCUMENTO

No próximo capítulo será feita uma revisão da base biológica necessária para o entendimento do trabalho, abordando os fundamentos da biologia e genética. No capítulo 3 será apresentado o estudo de associação genômica ampla e conceitos da genética de populações. Posteriormente, no capítulo 4 serão descritos alguns métodos computacionais categorizados de acordo com as estratégias de otimização que utilizam. Seguidamente, no capítulo 5 será explicado o modelo proposto para o problema de descoberta de SNPs potencialmente associados a doenças, descrevendo cada passo da metodologia empregada. O capítulo 6, apresenta os resultados dos experimentos realizados tanto com dados simulados como com dados reais. Finalmente, no capítulo 7 são apresentadas as conclusões e os trabalhos futuros que podem seguir a este trabalho de tese.

1.5 SÍNTESE DO CAPÍTULO

GWAS é uma área de pesquisa crescente que ainda apresenta muitos desafios a serem resolvidos. Como foi destacado neste primeiro capítulo, ao longo desta ultima década, muito se avançou no desenvolvimento de técnicas que ajudam ao estudo dos dados genômicos, os quais levam em consideração o impacto em problemas de saúde. Com o intuito de explicar a motivação deste trabalho de tese foi apresentada aqui uma revisão geral de pesquisas sobre GWAS, assim como trabalhos desenvolvidos dentro do enfoque dos estudos caso-controle.

Cabe resaltar a importância dos estudos genômicos de grande escala como fonte primordial de dados para a construção de indicadores de saúde. Embora se tenha conhecimento acerca das inúmeras imperfeições neste tipo de análise, o uso cada vez mais amplo contribuirá, certamente, para o seu aprimoramento, o que necessariamente depende da utilização de ferramentas válidas que estejam livre de erros metodológicos na sua concepção, desenho, implementação, e no processo de análise dos dados.

CAPÍTULO 2: GENÉTICA HUMANA BÁSICA

Neste capítulo se apresenta um resumo de genética humana básica. A informação descrita aqui foi baseada nas notas providas pelo curso de "*Statistic in Genetics*" [Almgen et al, 2003] e no livro de Sham, "*Statistic in human genetics*" [Sham, 1998]

2.1 A ESTRUTURA DO MATERIAL GENÉTICO HUMANO

O genoma humano é a totalidade da informação genética que possui um organismo em particular e que codifica para ele. O material genético em humanos está presente em cada célula do corpo. A porção principal é contida em cromossomos localizados no núcleo da célula e uma pequena parte restante é localizada dentro da mitocôndria.

2.1.1 CROMOSSOMOS SEXUAIS E AUTOSSÔMICOS

O núcleo de toda célula somática (ex. todas as células, à exceção dos óvulos e espermatozóides) normalmente contém 23 pares de cromossomos, subdivididos em 22 pares de cromossomos autossômicos e um par de cromossomos sexuais (representado por dois cromossomos X em mulheres, e um cromossomo X e um Y em homens), um total de 46 cromossomos. Em células sexuais (óvulo e espermatozóides), não existem cromossomos pares. Uma fotografia de microscópio dos cromossomos masculinos é mostrada na Figura 2.1. O genoma haplóide (ou seja, com uma única representação de cada par) tem um comprimento total aproximado de 3,2 Giga de pares de bases de DNA (3,2 Gpb) que contém entre 20,000 a 25,000 genes [http://www.genome.org].

Para simplificar, não serão consideradas análises de características ligadas ao sexo, ex. características onde os genes contribuintes estão localizados nos cromossomos X ou Y.

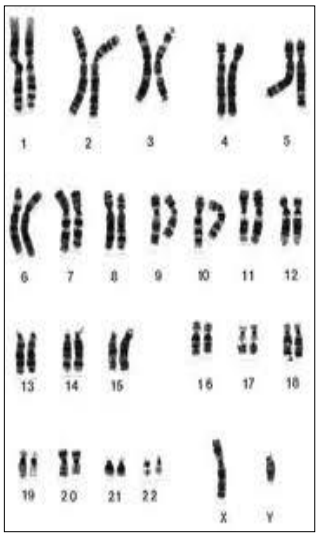
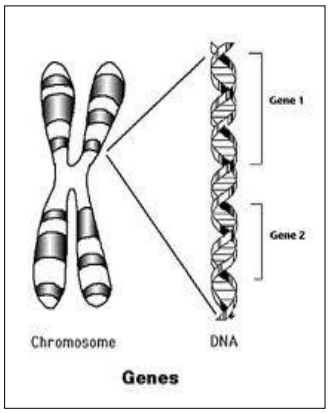


Figura 2.1: Cariograma dos cromossomos masculinos. (Fonte: Almgen et al, 2003)

A estrutura linear dos cromossomos

Cada cromossomo é composto de DNA o qual tem uma estrutura linear e é essencialmente uma sequencia de pares de bases complementares, os quais estão ligados entre si por limites químicos. As quatro bases de DNA são moléculas chamadas adenina, guanina, citosina, e timina, abreviadas por A, G, C e T respectivamente. Cada uma dessas bases pode formar um par complementar com uma e somente uma outra base, então poderia haver quatro diferentes pares de bases complementares: A-T, G-C, T-A e C-G (a ordem das bases não importa). Note que, os pares de base complementares podem se conectar em qualquer ordem no cromossomo. Uma representação simplificada da estrutura de um cromossomo em pares de bases é mostrada na Figura 2.2.

Cada cromossomo poderia ser descrito por sua sequencia de DNA, ex. G-G-A-C-T-A-A e poderia ser visto como um vetor de letras com alfabeto {A, C, G, T}. Estas sequências poderiam diferir em comprimento de cromossomo para cromossomo e, de fato, na ordem de pares de bases. No total, existem aproximadamente 3,000,000,000 (3×10^9) pares de bases no genoma humano (ex. em todos os cromossomos).



G C
| |
G C
| |
A T
| |
C G
| |
T A
| |
A T
| |
A T
⋮ ⋮

Figura 2.2: estrutura simplificada do cromossomo. (Fonte: Almgen et al, 2003). Note que duas fitas de DNA formadas pelas bases estão conectadas linearmente (ex. G-G-A-C-T-A-A e C-C-T-G-A-T-T).

Pares de cromossomos homólogos

Lembrando que os cromossomos autossômicos são dispostos em pares, eles são chamados pares homólogos de cromossomos. Como o termo "homólogo" diz, os cromossomos do par são muito similares na estrutura (ex. em comprimento e na sequencia de pares de bases). No entanto, eles não são cópias completas um do outro. Um cromossomo é herdado da mãe, outro do pai, e eles são diferentes devido à existência de variações ou polimorfismos genéticos.

2.2 POLIMORFISMOS GENÉTICOS

Ao se comparar a sequencia genética de diferentes indivíduos da população no mesmo cromossomo, pode-se notar que grande parte da sequencia genética é completamente similar para todos os indivíduos (de fato, há partes assemelhando-se à estrutura genética encontrada em animais, por exemplo, em rato, ilustrando o desenvolvimento evolucionário da espécie humana). No entanto, é também evidente que a sequencia de pares de bases varia de indivíduo para indivíduo.

Os seres humanos do mesmo sexo compartilham uma porcentagem muito elevada (em torno de 99%) de sua sequencia de DNA, o que permite trabalhar com uma única sequencia de referência, onde pequenas variações genômicas fundamentam boa parte da variabilidade fenotípica interindividual. Uma variação no genoma, por

substituição, deleção ou inserção, se denomina polimorfismo genético. Nem todos os polimorfismos genéticos provocam uma alteração na sequencia de uma proteína ou de seu nível de expressão, ou seja, muitos são silenciosos e carecem de expressão fenotípica.

Tipos de polimorfismo genético

Muitas vezes, na sequencia de DNA de indivíduos, uma das "letras" do alfabeto de pares de bases (ex. uma A) na sequencia de pares de bases no cromossomo será substituída por outra "letra" (ex. uma C). Este tipo de polimorfismo é chamado polimorfismo de nucleotídeo simples ou SNP. Em outros casos a mesma sequencia de "letras" poderia ser repetida uma quantidade de vezes diferente de um indivíduo a outro. Por exemplo, a sequencia em um indivíduo poderia ser GGACTAA[ACTT] (uma repetição de ACTT) e a sequencia em outro indivíduo poderia ser GGACTAA[ACTT][ACTT][ACTT] (três repetições de ACTT). Um polimorfismo genético deste tipo é chamado microsatélite autossômico (também conhecidos pela sua sigla em inglês STRs - Short Tandem Repeats). De modo geral, podem-se encontrar também polimorfismos genéticos onde uma sequencia de uma ou mais "letras" parece estar excluída ou inserida na sequencia original, polimorfismos deste tipo são chamados polimorfismos de inserção-deleção ou INDEL. Note que isto implica que o comprimento do mesmo cromossomo pode variar entre os indivíduos, no entanto, essa variação no comprimento é bastante insignificante em relação ao comprimento do cromossomo inteiro.

Polimorfismo de Nucleotídeo Único (SNP)

Como seu nome sugere, um SNP se refere a uma única diferença de nucleotídeo (A, T, C, ou G) no genoma de uma população, nos quais são focados a maioria dos estudos GWAS. Os SNPs são a forma mais abundante de variação encontrada no genoma humano (o genoma humano tem aproximadamente entre 10 a 20 milhares de SNPs [http://www.genome.gov]). Dada sua importância, na atualidade existe um projeto internacional (http://www.hapmap.org) para catalogar em grande escala os SNPs do genoma humano. Neste contexto, a denominação de SNP

frequentemente se restringe a aqueles polimorfismos de um único nucleotídeo nos que o alelo menos frequente aparece em ao menos 1% da população.

A Figura 2.3 mostra uma ilustração de SNPs no genoma. Existem dois tipos de nucleotídeos possíveis em um SNP específico. Por exemplo, no SNP1 somente aparecem "A" e "C" e no SNP2 aparecem somente "G" e "A". O nucleotídeo com maior frequência na população é chamado "alelo maior" e o outro "alelo menor". Por exemplo, o alelo maior e menor para o SNP1 são "C" e "A", respectivamente. Em genética, uma combinação de alelos em diferentes *loci* sobre o cromossomo que são transmitidos (ex. herdados) juntos é referenciado como um haplótipo. Se os seis SNPs na Figura 2.3 são herdados juntos, então existem dois haplótipos para a primeira amostra, "AGCCCA" herdada do seu pai e "CGCCCA" da sua mãe. De modo similar, dois haplótipos para a segunda amostra: "CATGCA" do seu pai e "CGCCCA" da sua mãe.

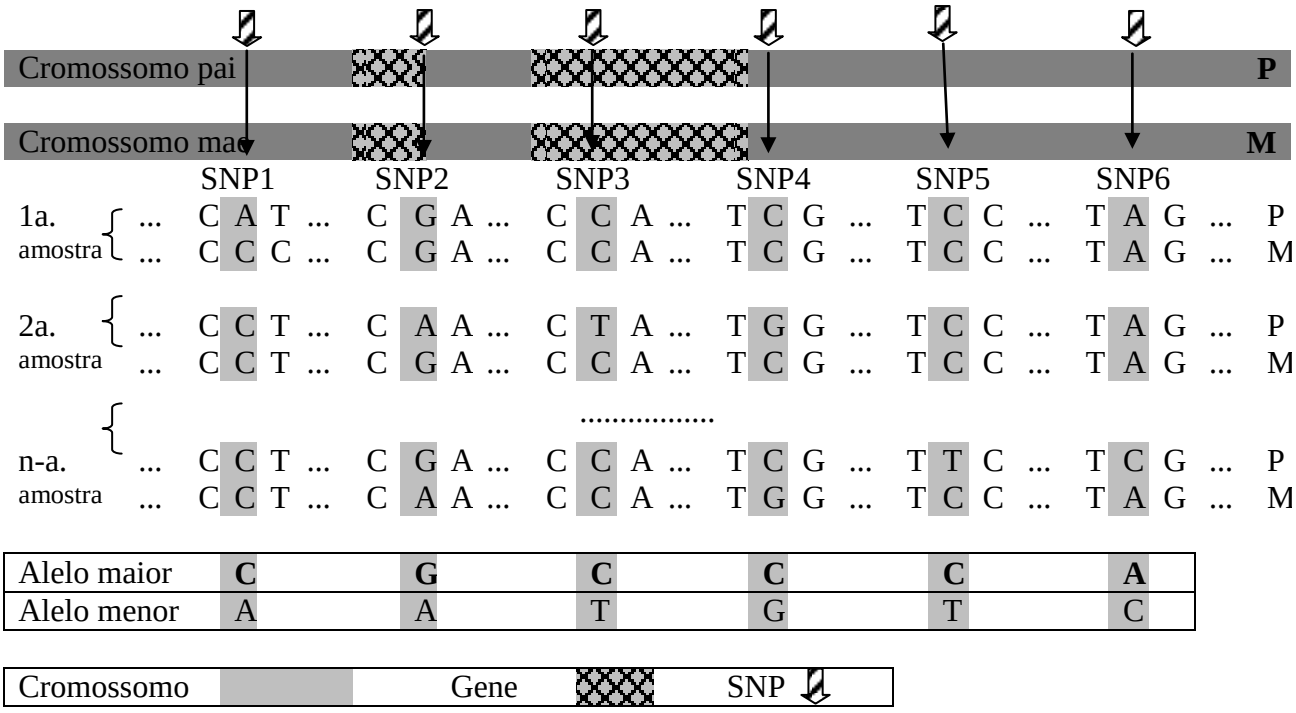


Figura 2.3. Uma ilustração de uma população de amostras de 6 SNPs em um cromossomo no genoma. As setas indicam a posição de cada SNP no cromossomo. A letra **P** referencia ao cromossomo do Pai e a letra **M** da Mãe.

Para um conjunto de SNPs, os haplótipos não podem diretamente ser observados em GWAS. Isto causa algumas incertezas quando queremos conhecer os haplótipos de um arquivo de genótipos. A estimação da fase alélica de haplótipos do conjunto de dados de genótipos é chamado "*phasing*". Na segunda amostra na Figura

2.3, um conjunto de dados de SNPs será observado como na Tabela 2.1, neste conjunto de dados, não sabemos qual alelo “A” no SNP2 vem do pai ou da mãe.

Amostra ID	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6
2	CC	AG	TC	GC	CC	AA

Tabela 2.1: Conjunto de dados de genótipos de 6 SNPs observados na amostra 2 da Figura 2.3

2.3 POSIÇÕES CROMOSSÔMICAS E LOCI GENÉTICOS

Como o cromossomo é uma estrutura linear, faz sentido falar sobre posições ao longo do cromossomo. Uma maneira para definir a posição cromossômica é simplesmente contar o número de pares de bases a partir de uma origem bem definida do cromossomo, tal como seu final (o telômero) ou seu centro (o centrômero). Como o número de pares de base é muito grande, pode ser expresso como uma medida de quilo pares de bases (1Kbp) ou em mega pares de bases (1 Mpb).

Outra maneira de definir um local ou região específica do cromossomo é olhando para a sequencia de pares de bases. No fim das contas, uma sequencia de 20 ou mais pares de bases de comprimento é na maioria dos casos suficiente para a identificação única de uma localização cromossômica, o que significa que tal sequencia só aparece uma vez em um lugar do genoma. Assim, se escolhemos uma sequencia de 20 bp, que é compartilhada por todos os indivíduos na população, esta pode ser usada para apontar uma região específica do cromossomo, que tem estrutura e função semelhante nos indivíduos.

Tendo estabelecido uma forma de apontar um local específico no cromossomo, pode-se agora definir o conceito de um lócus genético, como um trecho de DNA em um lugar particular em um cromossomo específico que está sendo analisado em sua variabilidade. Note que a sequencia não tem que ter um comprimento fixo. Assim, lócus genético é uma região cromossômica bem definida em algum local específico do cromossomo. Pesquisadores estão normalmente interessados em analisar as variações de pares de bases da sequencia nessa região nos indivíduos da população.

Alelos e genótipos em loci genético.

As variantes de sequências diferentes que podem ocorrer em um lócus genético são chamados alelos. Por exemplo, no caso de um lócus com um SNP correspondente a uma substituição de C por T poderíamos definir dois alelos: o primeiro corresponde a uma sequencia de pares de bases contendo um C, e o segundo corresponde a uma sequencia de pares de bases contendo um T. Na genética clássica Mendeliana os dois alelos são normalmente identificados por *A* e *a*.

Claramente, o número de diferentes variantes de pares de bases da sequencia encontrada nos indivíduos depende do comprimento do lócus genético - uma região longa tem maior probabilidade de dar origem a uma grande quantidade de variantes de sequências genéticas em relação a uma região curta. Na prática, é importante definir um lócus genético de forma que o número correspondente de alelos seja manejável.

Os cromossomos autossômicos são dispostos em pares homólogos. Em um único indivíduo um lócus genético está, portanto, presente em ambos cromossomos. Uma descrição completa da estrutura genética neste lócus requer portanto a especificação de dois alelos (um de cada cromossomo). A combinação de dois alelos desde os dois cromossomos é chamada o genótipo do indivíduo nesse lócus genético. Retornando ao exemplo do SNP com C/T, substituindo pelos alelos *A* e *a*, três genótipos são possíveis: *AA, Aa e aa*. Os genótipos com dois alelos idênticos são chamados homozigotos, enquanto o genótipo com diferentes alelos é chamado heterozigoto.

2.4 O QUE PODE SER MEDIDO NO LABORATÓRIO

A tecnologia moderna da genética molecular disponível em laboratório tem importantes limitações no que diz respeito a análise da sequencia genética. Em particular, o processo de obter uma sequencia genética completa do genoma de um indivíduo é ainda muito custoso e lento. Por isto, a maioria das técnicas usadas hoje em dia considera um lócus genético por vez. No entanto, estas técnicas não permitem análises separadas dos cromossomos que formam um par homólogo. Por esse motivo, o resultado de um análise de laboratório de uma sequencia genética em um particular lócus é um genótipo, ex. *AA, Aa ou aa* para um lócus com alelos *A* e *a*.

Se nosso interesse é procurar vários *loci*, isto tem que ser feito separadamente para cada lócus, um lócus por vez. Suponha, que no primeiro cromossomo em um par homólogo temos um alelo *A* no primeiro lócus e um alelo *B* no segundo lócus, e no segundo cromossomo, no mesmo par, temos um alelo *a* e outro alelo *b* nos dois loci respectivamente como é mostrado na Figura 2.4.

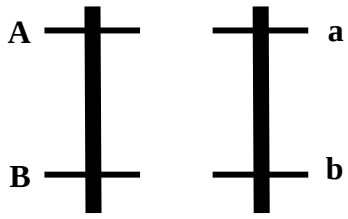


Figura 2.4: loci di-alélico

A informação sobre *A* e *B* que estão no mesmo cromossomo é importante em algumas situações. Se diz que os dois alelos *A* e *B* formam um haplótipo *AB* o qual corresponde a um alelo no lócus combinado consistindo do lócus 1 e lócus 2.

Porém, no laboratório a análise separada dos dois *loci* deveria resultar em um genótipo *Aa* no primeiro lócus e um genótipo *Bb* no segundo lócus (a ordem dos alelos não pode ser determinada), sem qualquer informação sobre si *A* e *B* estão no mesmo cromossomo. De fato, a mesma informação do genótipo deveria ter sido obtida se o primeiro cromossomo carregava alelos *A* e *b*, correspondente ao haplótipo *Ab* e o segundo cromossomo carregava os alelos *a* e *B*, correspondente ao haplótipo *aB*. Fazendo uma análise lócus por lócus se diz que perdemos informação sobre a fase dos alelos nos dois loci, o qual é uma limitação importante dos métodos de laboratório.

2.5 MARCADORES GENÉTICOS

O último objetivo da análise de certa característica genética é estabelecer a localização de algum fator genético contribuindo para a variação da característica ou fenótipo. Uma maneira de fazer isto é descrever a localização do novo lócus que influencia a característica relacionando-a com algum *loci* de localização bem estabelecido.

Um marcador genético, ou marcador molecular, é um segmento de DNA com uma localização física identificável (lócus) em um cromossomo e cuja herança genética pode-se rastrear. Deverá ser viável e eficiente determinar o genótipo de um marcador genético em um indivíduo utilizando métodos laboratoriais, como a *genotipagem*.

Para este efeito, um esforço considerável dos pesquisadores foi gasto na criação de marcos nos cromossomos, que são chamados de marcadores genéticos. Até o ano 2011 foram publicados mais de 1,000 GWAS com aproximadamente 2,000 achados. Na Figura 2.5, é apresentado por cada cromossomo, marcadores genéticos associados a doenças. [NHGRI GWAS Catalog]

2.6 OBTENÇÃO DE AMOSTRAS BIOLÓGICAS

A maioria das células de eucariotos contém o conteúdo completo de todos os cromossomos do organismo. Como a complexidade dos organismos cresce, assim também os diferentes tipos de células de um organismo. Com exceção das células sexuais haplóides, todas as células nucleadas têm o conteúdo total de cromossomos que todas as células têm. Isto faz possível coletar DNA para o propósito de genotipagem utilizando amostras biológicas de diferentes formas, como sangue, cabelo, pele, como também saliva.

A extração do DNA começa tão pronto quanto a célula de origem é recuperada em sua fonte. Esta amostra contém milhares de células completas com não somente DNA, mas também outros materiais intra e extra celulares. Os derivados celulares mais perigosos que são coletados com cada amostra são enzimas que quebram e digerem o DNA. Portanto, não importa o tipo de material biológico coletado, a purificação desse material é importante para a fidelidade do DNA que será extraído.

Figura 2.5: Catalogo de marcadores genéticos (SNPs) associados a doenças. Na Figura superior pode-se observar os cromossomos com loci sugeridos por GWAS. Na Figura inferior são referidas as doenças cujos *loci* são assinalados nos cromossomos da Figura superior. Este catalogo considera GWAS desde o ano 2009 até 2012. Fonte: NHGRI GWA Catalog. http://www.genome.gov/GWAStudies/

Genotipagem

A quantidade total de informação, de cadeia dupla, do DNA puro afeta a fidelidade da genotipagem, independentemente da tecnologia usada. Toda tecnologia de genotipagem baseia-se no fato de que o DNA pode ser teoricamente replicado na direção 5' a 3' infinitamente. Para efeitos de genotipagem, a amplificação de um local específico ou de todo o genoma é essencial, afim de ter sinais suficientemente grandes para que as tecnologias atuais possam ser executadas confiavelmente.

A genotipagem mede a composição alélica específica do indivíduo. Um alelo, como foi mencionado anteriormente, é um membro par de um gene em uma área específica de um cromossomo específico. O objetivo da genotipagem é encontrar um traço ou característica especifica de uma pessoa, planta ou animal. Cada gene tem dois traços e três combinações possíveis para esse traço. Os principais métodos para realizar genotipagem para deteção de SNPs são: a reação em cadeia polimerase (PCR), alelo-específico de oligonucleótidos (ASO) e ensaios de microarranjos de DNA [NEALE M. et al, 2008].

Microarranjos de DNA, ou DNA-chip, consiste num arranjo pré-definido de moléculas de DNA (fragmentos de DNA genômico, cDNAs ou oligonucleotídeos) quimicamente ligadas a uma superfície sólida, usualmente lâminas de microscópio revestidas com compostos que conferem carga positiva. Existem várias plataformas comerciais de microarranjos tais como Affymetrix, Illumina, Agilent, AppliedBiosystems, Incyte/Stanford etc. Por exemplo, a tecnologia desenvolvida pela Affymetrix (Figura 2.6) é atualmente utilizada para analisar mais de um milhão de SNPs sobre um chip. Esta tecnologia também usa uma quantidade mínima de DNA por genótipo, requerendo só 250 ng para completar todo o arranjo (http://www.affymetrix.com/).

Para fins computacionais, os dados brutos de SNPs, vindos da genotipagem são mostrados como letras (ex. aa, aA, AA) que definem os alelos observados em cada indivíduo, ou em forma de números (0, 1, 2). Diferentes abordagens para deteminação do genotipo SNP são adaptadas para diferentes tecnologias, na maioria delas a determinação das variantes do genótipos é tipicamente realizada pela análise de clusterização [NEALE M. et al, 2008]. Como as tecnologias de SNP focam na análise de alta dimensionalide dos dados, uma inspeção visual sobre a determinação do genótipo para todos os marcadores é irrealista. Como em qualquer procedimento estatístico, os erros da técnica de cluster são uma armadilha potencial.

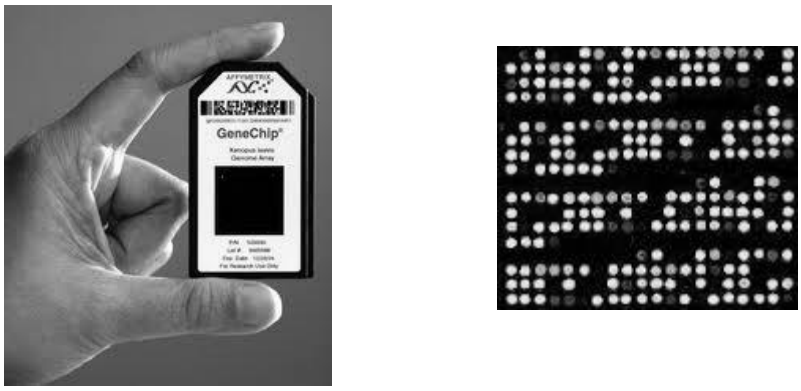


Figura 2.6: Imagem do chip Affymetrix. A direita são mostradas as sinais obtidas no processo experimental.

Portanto, para qualquer SNP mostrando um sinal de associação significante com a doença estudada, as imagenes atuais de intensidade providas pela genotipagem devem ser inspecionadas. Também, fazer a genotipagem dos SNPs com outra plataforma ou sobre a fita oposta pode adicionar mais confibilidade ao resultado.

2.7 FENÓTIPO

O termo "fenótipo" (do grego *pheno*, evidente, brilhante, e *typos*, característico) é empregado para designar as características apresentadas por um indivíduo, sejam elas morfológicas, fisiológicas e comportamentais. Também fazem parte do fenótipo características microscópicas e de natureza bioquímica, que necessitam de testes especiais para a sua identificação.

Entre as características fenotípicas visíveis, pode-se citar a cor de uma flor, a cor dos olhos de uma pessoa, a textura do cabelo, a cor do pelo de um animal, etc. Já o tipo sanguíneo e a sequência de aminoácidos de uma proteína são características fenotípicas revelada apenas mediante testes especiais. O fenótipo de um indivíduo sofre transformações com o passar do tempo. Por exemplo, à medida que envelhecemos o nosso corpo se modifica. Fatores ambientais também podem alterar o fenótipo: se ficarmos expostos à luz do sol, nossa pele escurecerá.

O termo "genótipo" (do grego *genos*, originar, provir, e *typos*, característica) refere-se à constituição genética do indivíduo, ou seja, aos genes que ele possui.

Fenótipo e genótipo e ambiente em interação

O fenótipo resulta da interação do genótipo com o ambiente. Consideremos, por exemplo, duas pessoas que tenham os mesmos tipos de alelos para pigmentação da pele; se uma delas toma sol com mais frequência que a outra, suas tonalidades de pele, fenótipo, são diferentes.

Um exemplo interessante de interação entre genótipo e ambiente na produção do fenótipo é a reação dos coelhos da raça Himalaia à temperatura. Em temperaturas baixas, os pelos crescem pretos e, em temperaturas altas, crescem brancos. A pelagem normal desses coelhos é branca, menos nas extremidades do corpo (focinho, orelha, rabo e patas), que, por perderem mais calor e apresentarem temperatura mais baixa, desenvolvem pelagem preta.

O fenótipo é qualquer característica mensurável e podem ser discretos ou contínuos. Os fenótipos em geral não são passados de uma geração a outra, os Gametas não. Os gametas são o mecanismo de transferência de informação genética. Estes gametas sempre ocorrem em algum contexto ambiental para produzir os fenótipos.

A grande maioria de fenótipos não tem categorias discretas e não tem um gene que seja necessário e suficiente para explicar sua variação. Hardy e Weinberg (no ano 1908) ajudaram a estabelecer que várias características eram Mendelianas. Mesmo assim, a maioria dos caracteres quantitativos não pode ser vista em um enfoque Mendeliano simples. A maioria dos caracteres que eles estudaram eram quantitativas. Por tanto, muitos científicos dessa época acreditavam que um mecanismo alternativo e mais importante de hereditariedade existia, além do Mendelismo. O Mendelismo não foi capaz de explicar os padrões de herança para a grande maioria da variação fenotípica.

Duas formas não mutuamente excludentes de genótipos discretos produzirem fenótipos contínuos são: Variação ambiental e Poligenes.

Na variação ambiental o mesmo genótipo pode responder diferentemente a alterações no ambiente. Por exemplo, indivíduos com o mesmo genótipo podem apresentar variação no fenótipo em função de influências ambientais. A forma como um genótipo responde ao meio ambiente é chamada de norma de reação daquele fenótipo. Um exemplo, são as mutações em Drosophila melanogaster que conferem

tamanhos diferentes aos olhos destas moscas. No entanto, este tamanho também varia em função da temperatura a que as moscas estão expostas.

Na Poligenes, Ronald A. Fisher [FISHER, 1918] observou que quando vários loci estão determinando um fenótipo, várias classes diferentes podem ser produzidas pela conjunção de diferentes alelos neste loci distintos. Dessa forma, quanto mais loci estiverem controlando um caráter, maior a possibilidade de formação de fenótipos com valores distintos.

Portanto, apesar de sua definição aparentemente simples, o conceito de fenótipo apresenta algumas sutilezas: Primeiro, a maior parte das moléculas codificadas no material genético, que consequentemente são parte do fenótipo, não são visíveis na aparência do organismo, ainda que sejam observáveis. Um bom exemplo é o tipo sanguíneo em humanos. Segundo, o fenótipo não é meramente um produto do genótipo, mas é influenciado em graus variáveis pelo ambiente.

Além disso, vale lembrar que a hereditariedade não está restrita ao DNA nuclear, já que a mitocôndria também apresenta o seu próprio DNA. Ao expandir o conceito de genótipo incluindo outros elementos hereditários, ampliamos também o conceito de fenótipo.

2.8 SÍNTESE DO CAPÍTULO

Procurou-se aqui introduzir alguns conceitos essenciais sobre a estrutura do material genético humano. Um conceito importante é o polimorfismo genético, o qual determina a variabilidade no genoma humano e ajuda a estudar a diferença que existe entre seres humanos. A principal fonte de variabilidade nos genomas dos seres humanos procede das variações de um único nucleotídeo, conhecido como SNP, nos quais são focados a maioria dos estudos GWAS. Foi também aqui abordado o processo requerido para obter os genótipos destes SNPs desde amostras biológicas para estudos GWA. Este capítulo não pretende fazer uma descrição de forma exaustiva, mas sim uma tentativa de fornecer uma visão e contexto para a criação de dados genéticos essenciais para a compreensão deste trabalho.

CAPÍTULO 3: ESTUDOS DE ASSOCIAÇÃO DO GENOMA INTEIRO - GWAS

Existem dois métodos analíticos principais para mapear genes envolvidos em traços humanos e de susceptibilidade a doenças, eles são ligação e associação. Os métodos de associação provêm maior poder e resolução que análises de ligação [Risch and Merikangas, 1996]. A ideia básica dos GWAS é rastrear todo o genoma procurando associações com certas doenças. A motivação é que tais associações podem fornecer novos candidatos para as variantes nos genes causais (ou em seus elementos regulatórios) que desempenham um papel para o fenótipo de interesse. No contexto clínico isto pode eventualmente levar a uma melhor compreensão dos componentes genéticos de doenças e seus factores de risco. Neste capítulo descrevemos os conceitos básicos para entender GWAS e a genética de populações.

3.1 ESTUDOS DE LIGAÇÃO

Os estudos de ligação (LD) são também conhecidos como estudos de desequilíbrio de ligação (do inglês, *Linkage disequilíbrium*). Em populações genéticas, LD é a associação não randômica dos alelos em dois ou mais loci, não necessariamente no mesmo cromossoma. O desequilíbrio de ligação descreve uma situação em que algumas combinações de alelos ou marcadores genéticos ocorrem mais ou menos frequentemente numa população do que seria esperado pela formação aleatória de haplótipos a partir de alelos baseados nas suas frequências. Associações não aleatórias entre polimorfismos em loci diferentes são medidas pelo grau de desequilíbrio de ligação. Por exemplo, alelos dos SNPs que residem próximos uns dos outros no cromossomo freqüentemente ocorrem em combinações não randômicas devido à infreqüente recombinação. Isto implica em dizer que alelos que estão em desequilíbrio de ligação migram juntos na divisão celular.

3.2 ESTUDOS DE ASSOCIAÇÃO

Em Estudos de associação do genoma inteiro, ou GWAS, os pesquisadores procuram por certos alelos que predispõem seus carreadores a certas doenças. A

abundancia de SNPs e a facilidade provida pelas tecnologias de genotipagem fazem com que estes marcadores genéticos sejam a escolha da maioria dos estudos de associação. Técnicas de genotipagem de alto rendimento estão evoluindo rapidamente e conseguem hoje que aproximadamente 1 milhão de SNPs sejam genotipados [Spencer et al., 2009]. Ao mesmo tempo, o custo da genotipagem de SNPs caiu dramaticamente fazendo os estudos de associação com milhares de pacientes uma realidade. As estimativas sugerem que com 500 mil SNPs, 85-92% da variação comum no genoma da população caucasiana será capturada [WTCCC, 2007]. Por exemplo, a *Wellcome Trust Case Control Consortium* realizou um estudo de associação de um lócus, em sete doenças comuns com um total de 14,000 pacientes cujos resultados foram depois replicados com sucesso. Assim, GWA é de longe o método mais detalhado e completo de investigar o genoma inteiro atualmente disponível.

Geralmente, estudos de associação são divididos em duas categorias: estudos baseados em família e estudos baseados na população. Nos estudos baseados em famílias, dados sobre os indivíduos afetados e seus pais são coletados. Então, é realizada uma procura dos alelos que são transmitidos de pais para seus filhos afetados, com mais frequência do que seria esperado ao acaso. Os estudos baseados na população são compostos de indivíduos afetados não relacionados (chamados de casos) e indivíduos saudáveis não relacionados (chamados de controles), por isso também são conhecidos como estudos "caso-controle". Neste tipo de estudo se procura alelos cuja frequência entre os casos é diferente de sua frequência entre os controles. Descobrir um SNP associado poderia significar a causa direta do desenvolvimento de certa doença, mas alternativamente, pode significar apenas uma ligação genética ao SNP causal. Portanto, uma investigação mais aprofundada e um mapeamento fino das áreas em torno de SNPs associados geralmente são necessários.

3.3 ESTUDOS CASO-CONTROLE

Os estudos caso-controle são os mais comuns na literatura científica. Muitos autores afirmam que eles são a abordagem mais poderosa e eficiente, garantindo robustez quando se estuda um grande número de SNPs [Ioannidis et al., 2001]. De uma perspectiva epidemiológica, a principal limitação desta abordagem é que às vezes levam ao surgimento de falsos positivos [Cardon L.R. & Bell, 2001]. Por outro

lado, os estudos baseados em família têm a vantagem sobre os estudos de base populacional, de que são robustos contra a miscigenação e estratificação da população, e permitem tanto testes de ligação como de associação. Além disso, o fato de que eles contêm informação tanto dentro como entre famílias, o qual prove benefícios substanciais em termos de múltiplos testes de hipóteses, especialmente no contexto de estudos de associação de todo o genoma. A limitação deste tipo de estudos é que precisa de dados de genótipos de muitos indivíduos com relação parenteral e este tipo de dados são mais difíceis de conseguir. Por isso estes estudos são feitos geralmente complementando os estudos caso-controle.

Um aumento na frequência de um alelo ou genótipo em casos comparados com os controles indica que a presença desse genótipo pode aumentar o risco da doença. A Figura 3.1 apresenta um esquema de um estudo caso-controle. O principal problema neste tipo de estudo é garantir uma boa correspondência entre a base genética dos casos e controles, de modo que qualquer diferença genética entre eles esteja relacionada com a doença em estudo e não a uma amostragem tendenciosa. Claramente, os casos e controles devem ser de grupos étnicos similares. Muitas diferenças genéticas sutis podem ser evitadas através da coleta de controles da mesma área geográfico dos casos, ou através da coleta de informações sobre o local de nascimento dos avôs de modo a verificar uma distribuição semelhante entre casos e controles.

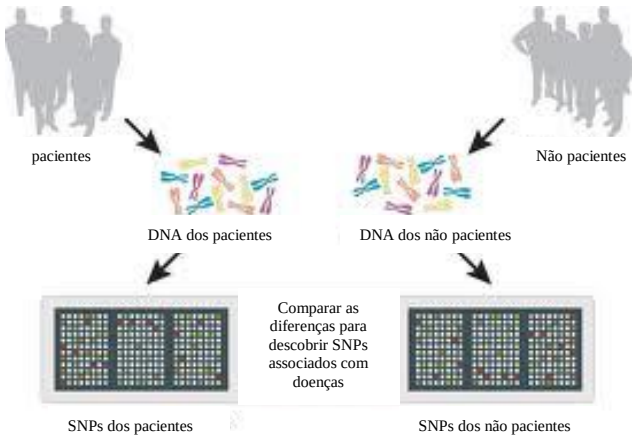


Figura 3.1: A força da associação entre cada SNPs e a doença é calculada com base na prevalência de cada SNPs nos pacientes (casos) e nos não pacientes (controles).

3.4 ETAPAS DE UM GWAS

Os estudos GWAS caso-controle, tipicamente consideram 5 passos fundamentais na análise:

1. Primeiramente, um bom planejamento do estudo tem que ser feito, onde se realiza a seleção de um grande número de indivíduos com a doença a ser estudada e do grupo de comparação, ou seja dos controles;
2. Posteriormente começa a coleta das amostras e o processo de genotipagem do DNA de cada indivíduo selecionado no desenho do estudo;
3. Depois será feito um controle de qualidade sobre os dados brutos vindos da genotipagem, verificando e corrigindo os erros do processo de genotipagem;
4. Logo após será realizada a análise de associação (ex. teste estatístico) entre os SNPs (aqueles que passaram o controle de qualidade) e a doença. Aqui é realizado um teste para cada SNPs ou múltiplos testes;
5. Finalmente, para confirmar os sinais de associação positiva desde o estudo inicial, é essencial replicar os resultados em uma amostra de uma população independente. Também deve ser realizada uma análise funcional dos SNPs identificados.

Além disso uma meta-análise pode incrementar o poder em detectar variantes mais raras de efeito modesto utilizando uma amostra maior que no estudo individual.

GWAS pode ser também utilizado para análises de interação gene-gene, para detecção de haplótipos de alto risco, associação entre SNPs e expressão do gene (ex. quantidade de proteínas para o qual um gene codifica).

3.5 CONCEITOS DA GENÉTICA DE POPULAÇÕES

Nesta seção serão abordados alguns conceitos importantes da genética de populações que respondem a questões fundamentais que os epidemiologistas genéticos consideram em GWAS.

3.5.1 PRINCÍPIO DE HARDY-WEINBERG OU EQUILIBRIO DE HARDY WEINBERG

O Equilíbrio de Hardy-Weinberg (EHW), é o princípio base da genética de populações, que diz que em uma população suficientemente grande e na ausência de seleção, migração e mutação, a frequência com que ocorre um alelo, permanecerá constante ao passar das gerações [Hoffee, 2000].

Para se entender melhor consideremos um caso simples, um único lócus gênico com dois alelos, sendo o alelo dominante "*A*" e o alelo recessivo "*a*", com frequências alélicas p e q respectivamente, sendo a freq(*A*) = p e a freq(*a*) = q onde p+q = 1. Então, considerando que os alelos dos controles no conjunto de dados estão em equilíbrio de Hardy-Weinberg, teremos a freq(*AA*) = p^2 para os homozigotos *AA* (dominantes) na população, a freq(*aa*) = q^2 para os homozigotos *aa* (recessivos), e freq(*Aa*) = 2pq para os heterozigotos.

Genótipos de controles devem estar em equilíbrio de HW. Desvios do EHW podem ser devidos a endogamia, estratificação ou seleção da população. Também podem ser um sintoma de associação com a doença. Desvios aparentes do EHW podem surgir na presença de um polimorfismo de deleção comum, também devido a um sítio *PCR-primer* mutante ou devido a uma tendência de erroneamente ler um heterozigoto como homozigoto. Geralmente, os pesquisadores testam o EHW de modo a avaliar a qualidade dos dados e descartam aqueles loci que, por exemplo, tem um desvio do EHW entre controles com um nível de significância de $\alpha=10^{-3}$ ou 10^{-4}. [Baldwin, 2006]

3.5.2 HERDABILIDADE

Vale lembrar que não é apenas a carga genética que uma pessoa possui que irá determinar a ocorrência de uma doença: fatores ambientais, como dieta, fumo, agentes infecciosos, também estão envolvidos no processo. Na genética o conceito de fenótipo é definido como F = G + A, onde F é o Fenótipo, G os genes e A o ambiente, ou seja, a expressão do fenótipo é dependente da expressão dos genes e dos fatores ambientes. Então, a herdabilidade é calculada através da Equação:

$$H^2 = \frac{Var(G)}{Var(F)}$$

Estimado pela razão entre as variâncias do genótipo (Var(G)) e Fenótipo (Var(F)), que mede a proporção da variação fenotípica que pode ser herdada em uma população, ou seja, o quanto que o fator genético influencia sobre um fenótipo.

A estimativa da herdabilidade analisa a contribuição relativa da diferença entre os fatores genéticos e não genéticos para a variância fenotípica total em uma população. Sendo assim, se uma doença ou fenótipo possui uma herdabilidade de 0,5, isso significa que 50% de toda variação fenotípica da doença é devido à constituição genética, mas isto não significa que 50% deles são causados pela genética.

Avanços científicos na genética molecular têm aumentado o conhecimento das variações genéticas que contribuem na ocorrência de doenças na população. Vários genes já foram identificados como os da fibrose cística, algumas formas de câncer, dentre os quais pelo menos cerca de 5% destes já possuem mecanismos de herdabilidade explicados [TOMLINSON et al, 2011].

3.5.3 DESEQUILÍBRIO DE LIGAÇÃO (LD)

Além dessas propriedades biológicas já citadas nas seções anteriores, também se pode utilizar o princípio de Desequilíbrio de Ligação. O princípio de LD é estudado para definir a associação existente entre SNPs, também definido como uma associação não-aleatória de SNPs [Ardlie et al., 2002]. Quando dois ou mais alelos específicos, em loci distintos, em um mesmo cromossomo são mais freqüentes em conjunto do que separados, então os loci estão em desequilíbrio.

Análises de LD permanecerão cruciais para o planejamento dos estudos de associação até que o re-seqüenciamento de todo o genoma se torne rotineiramente viável. Atualmente, só alguns dos mais de 10 milhões de polimorfismos humanos comuns são considerados em qualquer estudo. Se um polimorfismo causal não é genotipado, podemos ainda esperar detectar seu efeito através de LD com os polimorfismos que foram genotipados. No entanto, LD é um fenômeno não-quantitativo: não há escala natural para medi-la. Entre as medidas que têm sido propostas para haplótipos de dois lócus, as duas mais importantes são D' e r^2.

A medida D' calcula o desequilíbrio pela diferença entre a frequência observada entre um par de loci, P_{AB} e a frequência esperada entre os alelos separados P_A e P_B [Ardlie et al., 2002].

Então D’ é dado pela formula:

$$D' = P_{AB} - P_A \times P_B$$

A medida r^2 reflete o poder estatístico para detectar LD: nr^2 é o teste estatístico de Pearson em uma Tabela 2x2 de contagem de haplótipos. O valor máximo que r^2 pode atingir é fortemente determinado pelas frequências alélicas nos dois lócus [Wray, 2005]. Quanto mais diferentes as frequências alélicas, menor o valor de r^2. Assim, como a maioria dos SNP genotipados são comuns, se os variantes são raros, r^2 será baixo. Portanto, um r^2 baixo é necessário para detectar o LD entre os marcadores.

3.5.4 AUSÊNCIA DE GENÓTIPO

Em GWAS onde se investiga associação de um SNP com certa doença, se poucos genótipos estiverem faltando, não há muito problema. Já na análise de *loci*, os dados ausentes podem ser mais problemáticos porque muitos indivíduos podem ter um ou mais genótipos faltando. Isto pode conduzir a resultados espúrios. Uma solução conveniente é fazer imputação destes dados. Uma estratégia frequentemente utilizada é realizar substituição dos genótipos faltantes com valores previstos baseando-se em um painel de referencia, como a fornecida pelo International Hapmap Consortium. Em esta estratégia tenta se identificar os haplótipos comuns obtidos do cruzamento entre o painel de haplótipos proporcionado pelo Hapmap e os haplótipos dos indivíduos do estudo. Então utiliza se os haplótipos compartilhados para imputar os alelos em falta nos indivíduos do estudo. [Anderson et al., 2010; Marchini&Howie, 2010].

3.5.5 ESTRATIFICAÇÃO DA POPULAÇÃO

Os estudos caso-controle assumem que qualquer diferença nos genótipos de SNPs entre casos e controles é unicamente devido a sua diferença no status da doença, e não devido a qualquer diferença de fundo genético.

Este pressuposto é fundamental para um estudo bem sucedido, mas é difícil de garantir tanto na fase da concepção de um estudo, ou na fase de análise. O problema surge se a população subjacente é na verdade uma mistura de populações

ancestralmente distintas com diferentes valores de prevalência da doença e frequência de SNPs.

Por exemplo, com duas populações onde a população 1 tem uma alta prevalência da doença e (independentemente) uma frequência de alelos maior no SNP que a população 2, teremos uma maior frequência de alelo de SNP em casos que nos controles, que resultará em uma associação espúria entre o SNP e a doença.

A falta de sucesso na replicação em muitos estudos de associação de doenças pode ser devido à estratificação da população, mas há pouca evidência de que as diferenças genéticas entre populações são suficientes para levar a estes resultados. Diversas populações genéticas, como os africanos e caucasianos têm prevalências de doenças diferentes e por tanto também diferem nas frequências alélicas de seus SNPs. Ex. Hipertensão, câncer de próstata não seriam analisados em um estudo genético considerando essas populações em conjunto. [Hirschhorn, 2002]

3.5.6 EPISTÁSIS

Epistasis foi originalmente definido por Bateson como a expressão de um alelo em um lócus mascarado por um alelo em outro lócus [Bateson 1910]. Este conceito foi posteriormente explicado em uma maneira estatística por Fisher [Kempthorne 1968] como qualquer desvio estatístico da combinação aditiva de dois loci em seus efeitos sobre o fenótipo. A definição de Fisher permite que a epistasia seja quantificada em formas diferentes baseado em seu significado biológico determinado por Bateson. A existência de epistasis é amplamente reconhecida como fundamentalmente importante para a compreensão da estrutura e função de pathways e da dinâmica evolucionária de sistemas genéticos complexos [Phillips 2008].

Epistasia é uma medida da força de interações epistáticas. Interações epistáticas são interações não-aditivas entre alelos, locos, ou mutações. Isto é, se o efeito combinado de um par de mutações não é o que se espera de seus efeitos individuais, então pode-se dizer que há epistasia entre estas duas mutações. Epistasia descreve como interações entre genes podem afetar fenótipos. Os genes podem mascarar a presença do outro ou se combinar para produzir uma característica totalmente nova.

Um clássico exemplo de epistasia é a cor da pelagem que resulta do cruzamento de dois camundongos. Na Figura 3.2, dois dos *loci* responsáveis pela cor da pelagem em camundongos são:

	AA	Aa	aa
BB			
Bb			
bb			

Figura 3.2: Epistasia envolvendo a cor da pelagem em camundongos, alelos no lócus A, alteram o efeito do fenótipo dos alelos no lócus B.

Locus A: afeta as etapas iniciais de produção de uma enzima responsável pela produção de pigmentos:.

- alelo dominante (A) - a produção de pigmento normal;
- alelo recessivo (a) - homozigotos bloqueiam toda produção de pigmentos e são albinos.

Lócus B: determina se a pelagem tem bandas.

- alelo dominante (B) - resulta em pelagem com bandas e pelagem cutia (marrom). Mostrado na Figura 3.2 com uma cor cinza;
- alelo recessivo (b) - homozigotos não têm bandas e sua pelagem é negra.

3.5.7 Efeitos Epistáticos e Principais

Além de respeitar o equilíbrio de Hardy-Weinberg, os dados também levam em consideração a influência de um ou mais SNPs sobre uma doença. Este tipo de variação genética que possui alcance individual suficiente para influenciar sobre uma doença ou fenótipo é conhecida como efeito principal [Yang et al., 2010].

Considerando a complexidade envolvida no mecanismo de regulação no genoma humano e nas diferentes formas de manifestações de doenças e susceptibilidade, é amplamente aceito que doenças complexas ou multifatoriais sejam normalmente causadas por influência de múltiplas variações genéticas, ou seja, pelo efeito combinado de vários SNPs. Este tipo de variação genética que influencia de forma combinada é conhecido como efeito epistático ou interação, apresentado por SNPs com pouco ou nenhum efeito individual, mas que apresentam forte influência quando estão atuando em conjunto. Basicamente, uma interação entre dois SNPs acontece quando seu efeito conjunto não pode ser entendido como a soma de seus efeitos individuais.

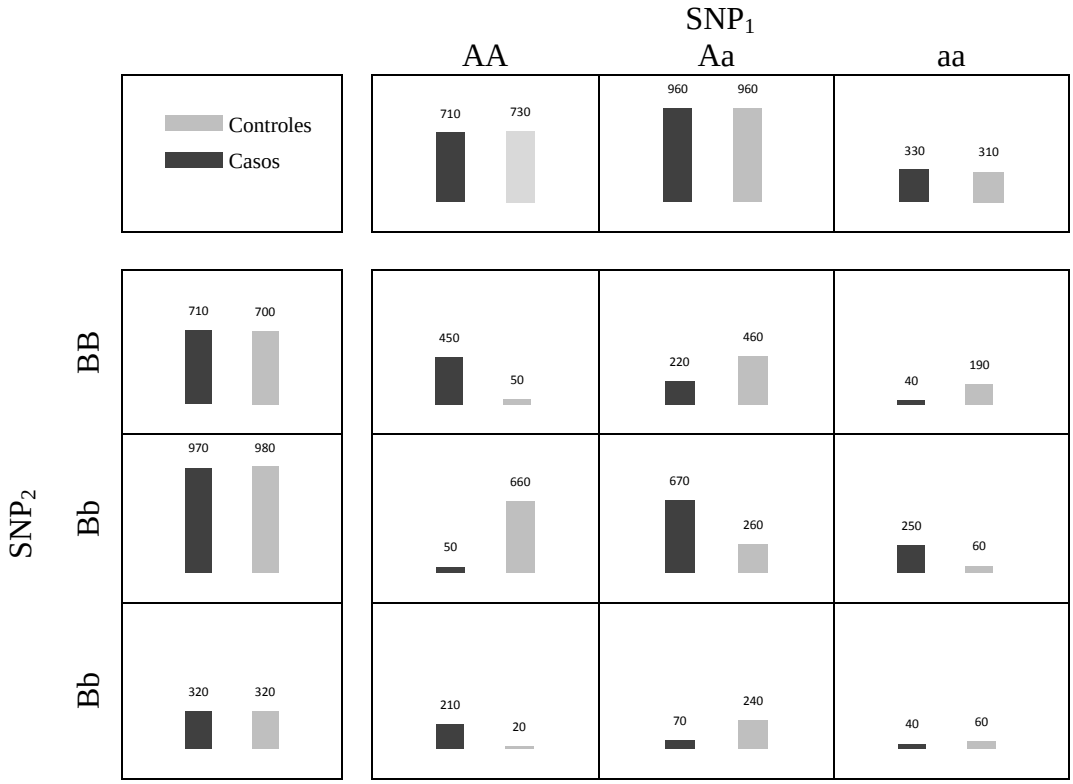


Figura 3.3: o gráfico mostra 2000 casos e 2000 controles onde o SNP_1 e o SNP_2 têm a mesma distribuição em casos e controles, mas sua distribuição conjunta é significativamente diferente entre casos e controles.

Um exemplo extremo é mostrado na Figura 3.3. Pode-se observar que as quantidade de genótipos são quase as mesmas no SNP1 para os casos e controles.

Assim, seu efeito individual é muito fraco ou seja não ajuda a distinguir casos e controles. Similarmente, o efeito individual do SNP2 também é fraco. Todavia, seu efeito conjunto é forte já que as quantidades de genótipos aparecem completamente diferentes para cada combinação de genótipos. Este fenômeno é chamado "efeito marginal fraco com interação forte". Uma estratégia muito comum primeiro seleciona aqueles SNPs com grande efeito marginal para depois realizar o teste de interação somente entre esses SNPs selecionados. Claramente esta estratégia não identificará SNPs com o chamado "efeito marginal fraco porém com forte interação".

3.6 SÍNTESE DO CAPÍTULO

Procurou-se neste capítulo mostrar uma introdução sobre GWAS, descrevendo os tipos de estudos de associação. De maneira especial foram abordados os estudos do tipo caso-controle e as principais etapas consideradas na análise. Na última parte do capítulo foram revistos conceitos fundamentais da genética de populações que são importantes para a abordagem GWAS. Estes conceitos ajudam o pesquisador a entender as questões principais que subjazem a epidemiologia genética. Embora o capítulo possa aparentar uma certa complexidade, vale dizer que se buscou abordar os conceitos necessários para compreender este trabalho de tese.

CAPÍTULO 4: ABORDAGENS COMPUTACIONAIS PARA DETECÇÃO DE INTERAÇÃO DE SNPS EM DOENÇAS

Como foi mencionado no capítulo anterior, a epistasia é reconhecida como fundamentalmente importante para a compreensão do mecanismo da doença que causa a variação genética. Nos últimos 5 anos, o número de estudos GWAS publicados aumentou consideravelmente. Isto se deveu aos avanços nas ferramentas para pesquisa genética, como banco de dados, e tecnologias para genotipagem cada vez mais precisas e acessíveis, contribuindo aos avanços no desenvolvimento de softwares para análise GWAS.

Não existe um paradigma para a análise de interação de SNPs em GWAS. Fazendo uma revisão na literatura sobre os métodos disponíveis, foram apontados dois grupos de abordagens de acordo a sua estratégia de busca: métodos baseados na busca exaustiva e métodos baseados na busca não exaustiva também conhecidos como busca estocástica e/ou heurística. A Figura 4.1 apresenta um diagrama com alguns métodos encontrados na literatura. Nas seções seguintes alguns deste métodos serão apresentados.

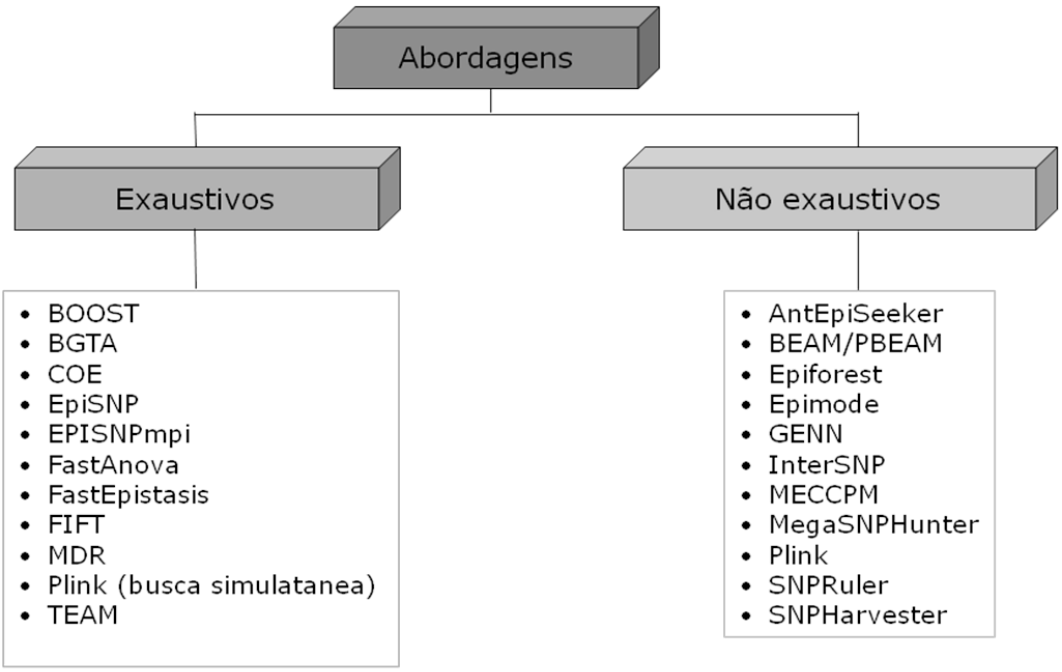


Figura 4.1: Classificação das abordagens computacionais para detectar interação em GWAS.

4.1 MÉTODOS DE BUSCA EXAUSTIVA

A busca exaustiva, enumera todos os k-lócus de interações possíveis entre SNPs para identificar o efeito, ou efeitos, que melhor predizem o desfecho fenotípico. Esta propriedade exaustiva leva à sua característica mais importante do ponto de vista computacional. Embora seja viável, mesmo para os maiores conjuntos de dados disponíveis hoje, utilizando, por exemplo, computação paralela, as generalizações para investigar interações de ordem superior são excessivamente demoradas. Nesta seção será realizado um breve resumo de algumas abordagens exaustivas disponíveis na literatura.

BGTA (Backward genotype-trait association) [Zheng et al, 2006]

Usa um procedimento de varredura tipo *boopstrap* para selecionar marcadores. Aqueles marcadores que retornam uma frequência maior aplicando um procedimento de regressão *log quantile-quantile* são considerados marcadores associados com a doença.

Nesta abordagem, um algoritmo de varredura seleciona um conjunto de marcadores relevantes que exibem sinais de associação com certa doença. Em cada iteração, o algoritmo remove os marcadores que têm uma contribuição mínima na informação de associação.

BOOST (BOolean Operation-based Screening and Testing) [Wang X. et al, 2010]

Pertence aos métodos que investigam epistasia estatística para o descobrimento de interações desconhecidas gene-gene que subjazem às doenças complexas. Permite a análise de todos os pares de interações de marcadores de estudos GWA caso-controle de uma maneira rápida. O método utiliza uma representação booleana dos genótipos para obter uma Tabela de contingência mais eficiente utilizando operações booleanas.

Sobre a base da equivalência entre o modelo log-linear e seu correspondente modelo de regressão logística, BOOST constrói um teste estatístico usando um modelo de associação homogêneo e um modelo saturado, estes modelos são maximizados através de estimação de máxima verossimilhança (do inglês, *maximum-*

likelihood estimation - MLE). A nova medida de interação é dada pela diferença desses dois modelos. A representação booleana dos dados de genótipos ajuda na eficiência de CPU porque só envolve valores booleanos e permite usar operações lógicas rápidas (bitwise) para obter as Tabelas de contingência. O programa esta disponível em http://bioinformatics.ust.hk/BOOST.html

EPISNP e EPISNPmpi [Ma et al., 2008]

Os programas EPISNPmpi e EPISNP foram desenvolvidos para detectar efeito de um único lócus e efeitos epistáticos de SNP em características quantitativas em GWAS. Também incluem investigação de efeitos de três lócus para cada SNP e cinco efeitos de epistasia para cada par de SNPs, com base no modelo estendido de Kempthorne (1954). EPISNPmpi é um programa de computação paralela para investigar epistasia em GWAS em supercomputadores e clusters, utilizando um modelo de regressão linear.

O método estatístico implementado para detectar efeitos de 1-lócus e epistasia em computação paralela e serial utiliza um modelo linear geral para investigar os efeitos de cada SNP e cada par de SNPs, e está baseado no modelo de Kempthorne para investigar efeitos aditivos e de dominância de cada SNP investigado e de cada par de SNPs. Uma análise de mínimos quadrados de dois passos é usada para implementar o teste estatístico. O primeiro passo corrige o valor fenotípico causado por efeitos sistemáticos de gênero e idade. O segundo passo investiga os efeitos epistáticos e de 1-lócus usando os valores fenotípicos corrigidos. Esta análise de dois passos estima e remove efeitos sistemáticos de uma vez, e então consegue uma considerável vantagem computacional quando o número de SNPs é grande.

FastEpistasis [**SCHÜPBACH et al., 2010**]

FastEpistasis é uma ferramenta de software, capaz de computar testes de epistasia para um grande número de pares de SNPs. È uma extensão paralela eficiente para o módulo de epistasia desenvolvido no PLINK. Ele investiga efeitos epistáticos usando um método de regressão linear normal de resposta quantitativa sobre os efeitos marginais de cada SNP e um efeito de interação do par de SNP.

FastEpistasis otimiza os cálculos, dividindo as tarefas da análise em três aplicações distintas: pré computação, núcleo ou computacional e pós-computação. A fase de pré computação carrega arquivos de dados no formato binário PLINK, reformata os dados para realizar computações mais rápidas e reduz o número de condições para verificar na fase computacional. A fase núcleo é desenhada para realizar computação embaraçosamente paralela, através de iteração de testes de epistasis em pares de SNPs. A computação está baseada na aplicação da decomposição QR para derivar estimações de mínimos quadrados do coeficiente de interação e de seu erro padrão. Uma fase opcional de pós computação é fornecida para agregar resultados de cada processador ou núcleo, podendo incluir detalhada informação de cada SNP, computar *p-valoress* de cada teste, e converte arquivos de texto.

FIFT (Focused Interaction Testing Framework) [Millstein et al, 2006]

FIFT foi desenvolvido para identificar a susceptibilidade de genes envolvidos em interações epistáticas de estudos caso-controle de genes candidatos. Nesta abordagem, testes de razão de verossimilhança são realizados em etapas, que vão aumentando segundo a ordem de interação considerada. Realiza uma redução do número de testes fazendo uma varredura das combinações de genes utilizando o teste estatístico qui-quadrado.

MDR [Ritchie et al., 2001]

O método *Multifactor Dimensionality Reduction* (MDR) foi inicialmente proposto por Ritchie et al. (2001), ela é uma abordagem de mineração de dados não-paramétrica que não assume nenhum modelo genético para detectar e caracterizar as combinações entre variáveis genéticas e ambientais que interagem para influenciar a variável classe (caso-controle). MDR procura exaustivamente o espaço de k combinações de marcadores e constrói um classificador para cada combinação. Todo o processo de classificação é realizado utilizando validação cruzada estratificada com fator 10.

MDR identifica k formas de interações através da busca exaustiva e avalia a associação entre cada interação e a doença por meio de validações cruzadas. Desde a

descrição inicial do MDR feita por Ritchie et al. (2001), muitas modificações e extensões tem sido propostas, mas a ideia principal da abordagem é baseada na redução do espaço de representação dos dados, tornando mais fácil para outros métodos detectarem interações. A ferramenta de código aberto escrita na linguagem Java foi implementada e descrita por Ritchie et al. (2003b), ela é capaz de tratar tanto dados de caso e controle como também dados baseados em família. Como resultado do esforço na difusão da metodologia proposta e fácil acesso à ferramenta, MDR é atualmente um dos métodos mais amplamente utilizado para detectar interação entre marcadores ou epistasis, como é evidenciado pelas 378 publicações encontradas no Pubmed, buscando todos os campos (*all fields*) por "*Multifactor Dimensionality Reduction*".

A premissa básica do MDR é reduzir a multi-dimensionalidade do espaço de busca para uma variável preditora de uma única dimensão. Análise de genótipos multi-lócus são agrupadas em categorias de alto e baixo risco para seguidamente resultar em uma dimensão. MDR é um algoritmo de quatro passos como mostrado na Figura 4.2.

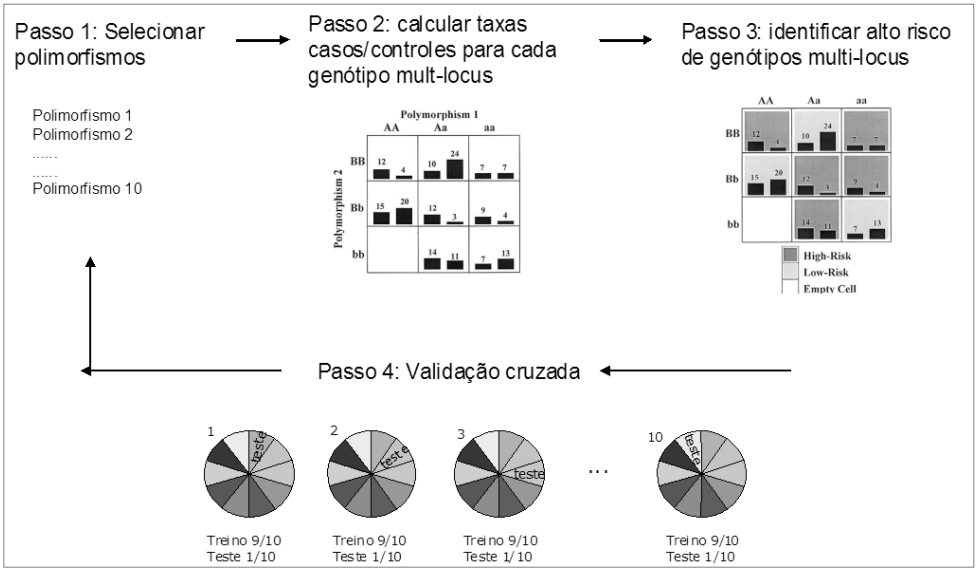


Figura 4.2: Visão geral do algoritmo MDR. Basicamente, uma Tabela de contingência é construída para cada SNP-SNP de forma a fazer uma classificação dos genótipos em alto ou baixo risco. Finalmente, aqueles genótipos que minimizam o erro de predição sobre os k valores analisados são selecionados como o melhor modelo.

O algoritmo possui 4 passos principais:

Passo 1. Selecionar k fatores (k variáveis para modelar, ex. dois lócus: SNP_i x SNP_j). O conjunto selecionado deve ser pequeno para fazer o passo 2 tratável;

Passo 2. Calcular as taxas caso-controle para cada genótipo multi-lócus. Uma Tabela é construída com uma célula para cada genótipo multi-lócus. Por exemplo, para marcadores bi-alélicos tal como SNP para o qual a Tabela terá dimensão 3x3. Como na Figura 4.2;

Passo 3. Identificar genótipos multi-lócus de alto risco. Para certo limiar T, etiquetar todas as células com taxa caso/controle $R \geq T$ como alto risco, aqueles com $R<T$ como baixo risco, e ignorar as células vazias. Geralmente o limiar T =1.0;

Passo 4. Estimar o erro de predição usando validação cruzada estratificada de fator 10. Para cada modelo resultante, o erro de predição é determinado.

Uma vez que estes passos são completados para todos os k valores considerados (todas as combinações do conjunto de dados), o conjunto de fatores que minimiza o erro de predição sobre os k valores analisados é selecionado como o melhor modelo.

Este tipo de método de busca exaustiva trabalha bem com um problema de tamanho pequeno. Em GWAS, aplicação direta deste método é computacionalmente proibitiva. É preciso uma filtragem efetiva para reduzir significativamente o número de SNPs de modo que a busca exaustiva seja computacionalmente factível. Estudos comprovam que MDR tem melhor desempenho que regressão logística para doenças comuns [Zhang et al, 2009]. Esta ferramenta encontra-se disponível em [http://sourceforge.net/projects/mdr/].

PLINK [Purcell et al, 2007]

Plink é um software livre de código aberto, que proporciona um conjunto de ferramentas de análise de associação do genoma completo. Este método é o mais utilizado em GWAS e por isso considerado como "o estado da arte", como é evidenciado pelas 2,334 citações em artigos científicos, quando procurado na

biblioteca de medicina *Pubmed Central* [*http://www.ncbi.nlm.nih.gov/pubmed*] na data 20 de maio de 2013. Foi projetado para realizar uma série de análises básicas em dados de alta dimensionalidade de forma computacionalmente eficiente. Além de suas outras funções, ele pode ser usado para investigar epistasia estatística.

Para identificar as interações entre todos os pares de *loci*, usa uma estratégia simples analisando exaustivamente cada combinação SNP-SNP através de um modelo de interação completa baseada em regressão logística. Esta estratégia, implementada primeiramente por Marchini et al (2005), foi revisada por Ionita and Man(2006), que adicionaram uma busca condicional que posteriormente foi implementado no Plink por Purcell em 2007. Atualmente o Plink proporciona métodos tanto de busca exaustiva como de busca não exaustiva.

A busca por associação de 2 loci com certo fenótipo é realizada pelo modulo epistasis. Plink aplica regressão logística por cada avaliação de lócus x lócus aplicada segundo a equação 4.1. A regressão logística é uma adaptação da regressão linear na qual uma transformação logarítmica "logit" é usada para permitir a analise de um Fenótipo binário (ex. estado de caso ou controle). Na equação 4.1, $\boldsymbol{p}$ é a probabilidade de ter a doença, β_0 representa o efeito nulo, β_1 e β_2 representa o efeito principal de cada lócus sobre o fenótipo, e β_3 representa o termo de interação. As variáveis x_1 e x_2 contêm informação sobre o genótipo nos dois lócus e podem ser codificados de formas diferentes, por exemplo, -1, 0 e 1 para homozigoto recessivo, heterozigoto e homozigoto dominante respectivamente. O termo de interação ($x_1 * x_2$) pode também ser codificado de formas diferentes:

$$logit(p) = ln\frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 * x_2) \quad [4.1]$$

Os coeficientes β são estimados para cada SNP como também para a interação entre estes. Este procedimento pode ser usado sistematicamente para comparar diferentes modelos genéticos e para investigar se múltiplos SNPs têm efeitos independentes sobre o fenótipo ou se estão em desequilíbrio de ligação um como o outro. Plink avalia os seguintes modelos genéticos, descritos por [Marchini et al, 2005] :

Modelo multiplicativo: em analises de 1 lócus o modelo multiplicativo é avaliado entre e dentro loci, onde assume se um aumento do risco à doença adicionando uma variante em cada marcador multiplicativamente.

Interação de dois lócus, efeito limiar: em analise onde um lócus não é suficiente, se assume que a presença de variantes de risco a partir de ambos marcadores, elevam o risco que aumenta para um nível constante.

Interação de dois lócus, efeito multiplicativo: em analise onde um lócus não é suficiente se assume que alelos de risco múltiplos desde diferentes loci incrementam o risco linearmente.

Algoritmo

a) Avaliar todos os pares de loci.
b) Para cada par de loci x_1 e x_2, avaliar o modelo de interação segundo a equação 4.1:

$$logit(p) = ln\frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 * x_2),$$

c) Estimar OR (Odds Ratio). Para estimar a OR no modelo de interação de lócus x lócus, Plink preenche uma Tabela com as frequências alélicas da combinação dos loci e posteriormente colapsa essa Tabela 3x3 em uma Tabela 2 x 2 como é mostrada na Figura 4.3
d) Aplicar o teste Z score. Equação 4.2
e) Realizar o controle de testes múltiplos aplicando correção de Bonferroni ou FDR.

Lócus G	Lócus H		
	2	1	0
2	a	b	c
1	d	e	f
0	g	h	i

Lócus G	Lócus H	
	H1	H2
G1	A=4a+2b+2d+e	B=4c+2b+df+e
G2	C=4g+2h+2d+e	D=4i+2h+2f+e

Figura 4.3: No modelo de interação lócus x lócus, Plink estima a OR (Odds Ratio) como log (AD/BC).

O teste Z-score é aplicado segundo a equação 4.2 onde a variância V é estimada da forma: $\frac{1}{A}+\frac{1}{B}+\frac{1}{C}+\frac{1}{D}$

$$Teste_{Z-score} = \frac{[OR(Casos)-OR(Controles)]^2}{V(Casos)+V(Controles)} \quad [4.2]$$

Outros módulos proporcionados pelo programa são gestão dos dados, estatísticas para controle de qualidade dos dados, detecção de estratificação da população, teste de associação básico, teste de haplótipo e preditores multimarcadores, análise de CNV (em fase de teste), meta-análise e outros testes básicos. O programa está disponível no site web [http://pngu.mgh.harvard.edu/~purcell/plink/].

Tools for efficient epistasis detection in GWAS [Zhang X. et al, 2011]

As ferramentas apresentadas por Zhang et al., são três métodos que exploram alguma propriedade de algum teste estatístico usado para mitigar problemas de testes múltiplos que limitam a maioria dos algoritmos exaustivos. Eles provêm um conjunto de métodos, FastANOVA, COE e TEAM que consideram um número linear de testes para desse modo evitar a penalidade dos testes múltiplos. COE é uma generalização do anterior FastANOVA, incluindo teste chi-quadrado e a razão de verossimilhança.

O primeiro programa é FastANOVA [Zhang X. et al, 2008], utiliza um limite superior do teste ANOVA (teste de análise de variância) de dois lócus para podar o espaço de busca. O limite superior é expresso como a soma de dois termos. O primeiro termo é baseado no teste ANOVA de um único SNP. O segundo termo é baseado no genótipo de pares de SNPs independente de permutações. Esta propriedade permite indexar pares de SNP numa matriz com base na relação entre os genótipos de SNPs. Devido a o número de entradas na matriz ser limitado pelo número de indivíduos no estudo, muitos pares de SNPs compartilharam uma entrada comum. Além disso, pode ser demonstrado que todos os pares de SNP indexados pela mesma entrada têm exatamente o mesmo limite superior. Portanto, pode-se calcular o limite superior para um grupo de pares de SNP. Outra propriedade importante é que a estrutura de indexação só precisa ser construída uma vez e pode ser reutilizado para todos os dados permutados. Utilizando o limite superior e a estrutura de indexação,

FastANOVA necessita apenas executar o teste ANOVA em um pequeno número de pares de SNPs sem o risco de perder qualquer par significativo.

O segundo programa, COE [Zhang X. et al., 2010], utiliza otimização convexa. Ele se baseia em que uma grande variedade de testes estatísticos, tais como o teste qui-quadrado, o teste de razão de probabilidade (também conhecido como teste-G), e testes com base em entropia são todos testes com funções convexas de frequências observadas em Tabelas de contingência. Uma vez que o valor máximo de uma função convexa é atingido nos vértices do seu domínio convexo, por restrições sobre as frequências observadas nas Tabelas de contingência, pode-se determinar o domínio da função convexa e obter o seu valor máximo. Este valor máximo é utilizado como o limite superior no teste estatístico para filtrar pares de SNPs insignificantes. COE é aplicável a todos os testes que são convexos.

O terceiro programa, TEAM [Zhang X. et al., 2010], foi desenvolvido para superar as limitações dos anteriores que foram projetados para estudos com genótipos homozigotos e tamanhos de amostra relativamente pequenos. TEAM (do inglês, *Tree-based Epistasis Association Mapping*), é um algoritmo exaustivo que utiliza o mínimo de uma árvore de expansão (do inglês, *minimum spaning tree*) para atualizar de forma incremental as Tabelas de contingência para testes epistáticos sem verificar todos os indivíduos. Ele suporta qualquer teste estatístico baseado em Tabelas de contingência, e permite tanto o calculo da taxa de erro baseada em família como a taxa de controle de descoberta de falsos positivos.

TEAM computa exaustivamente todas as interações possíveis de 2-lócus usando um teste de permutação. Testes de permutação são geralmente mais precisos que métodos de ajuste direto, como por exemplo a correção de Bonferroni, para identificar interações epistáticas significantes, pero com um alto custo computacional. Se dois SNPs tem os mesmos genótipos na maioria das amostras, o calculo de suas Tabelas de contingência pode ser compartilhado considerando apenas aquelas amostras com diferentes genótipos.

TEAM utiliza uma árvore de expansão mínima para maximizar a computação das Tabelas de contingência compartilhadas, reduzindo assim o custo computacional. Na árvore, um nó representa um SNP e as arestas denotam o número de amostras com diferentes genótipos entre os SNPs conectados. Esta árvore, torna mais rápida a computação que os métodos de força bruta em uma ordem de magnitude. Então,

pode-se obter os valores exatos dos testes investigando a árvore de expansão mínima sem verificar todos os indivíduos.

4.2 MÉTODOS DE BUSCA NÃO EXAUSTIVA

Este tipo de métodos realiza uma busca parcial das possíveis associações (de k marcadores com certo fenótipo) para completar o processamento de forma relativamente rápida. Apesar de eficientes e rápidos comparados com os métodos exaustivos, estes métodos muitas vezes dependem do acaso para selecionar SNPs que exercem influência sobre a doença. Não é possível saber se eles conseguiram identificar ou alcançar a correta solução para um conjunto de dados específico. A medida que os conjuntos de dados crescem em número de SNPs, as chances de encontrar os dados corretos diminuem devido ao crescimento do espaço de busca.Os algoritmos exaustivos podem ser classificados segundo a estratégia utilizada para a redução do espaço de busca em métodos estocásticos e em métodos gulosos.

Os métodos estocásticos realizam uma investigação probabilística do espaço de busca. Alguns começam com um modelo composto por um conjunto aleatório de SNPs e tentam melhorar a sua precisão de classificação, enquanto outros usam um subconjunto pequeno que foram previamente selecionados incorporando conhecimento experto sobre os dados.

Os métodos de busca gulosa utilizam algoritmos que buscam uma solução para o problema tomando decisões que levam a um novo ótimo local a cada passo da execução. Com isso, espera-se que ao final da busca seja alcançado o ótimo global sem que seja necessário analisar todas as situações possíveis.

Algumas ferramentas que se baseiam em métodos estocásticos são:

AntEpiSeeker [**WANG Y. et al, 2010**]

Este método é um algoritmo derivado do ACO [DORIGO & GAMBARDELLA, 1997] e apresenta duas etapas. Na primeira, ele usa um teste qui-quadrado para verificar associação entre interação k-lócus e o fenótipo, e inicialmente, nenhuma suposição sobre a interação é feita no AntEpiSeeker.

A probabilidade de uma formiga adicionar o SNP k em seu caminho (ex. uma k-lócus interação) na iteração i é definida como $p_k(i) = \tau_k(i) / \sum_{j=1}^{N} \tau_j(i)$, onde $\tau_k(i)$ é o feromônio. O feromônio é atualizado de acordo a $\tau_k(i+1) = (1-\rho) \cdot \tau_k(i) + 0.1 \cdot \sum_{j=1}^{J} \chi_k^j(i)$, onde ρ é o coeficiente de evaporação, *J* é o número de interações k-lócus contendo k SNP na iteração i, e $\chi_k^j(i)$ é o valor qui-quadrado da interação *j*. Na segunda etapa, AntEpiSeeker conduz uma busca exaustiva de interações dentro do conjunto de SNPs altamente suspeitos, e dentro do conjunto reduzido de SNPs com níveis de feromônios no topo do ranking.

BEAM [Zhang Y, Liu JS., 2007]

Este método usa um modelo Bayesiano com um algoritmo *Metropolis-Hasting* para partição dos marcadores em três grupos: um grupo G_0 contém marcadores não ligados à doença, um grupo G_1 inclui marcadores que contribuem independentemente à doença, e um grupo G_2 que está composto de marcadores que influenciam conjuntamente à doença. Após a fase de partição, a verificação da significância dos SNPs candidatos é realizada com a utilização do teste estatístico B.

Em BEAM, existem duas probabilidades a priori que necesitam ser pré-determinadas: a probabilidade de cada marcador pertencer a G_1 e de pertencer a G_2. Primeiro, uma aplicação de regressão logística seleciona a fração ε mais significativa de SNPs em função de seus efeitos marginais, e posteriormente verifica-se todas as interações 2-lócus desses SNPs utilizando regressão logística com teste de razão de verossimilhança.

Epiforest [Jiang R. et al, 2009]

É uma abordagem baseada em *random forest* para a detecção de interações epistáticas em estudos caso-controle. Primeiro, um algoritmo *random forest* é executado com todos os SNPs para obter a importância de Gini de cada SNP e, em seguida, é realizada uma seleção de características utilizando um algoritmo *sliding window sequencial forward* (SWSFS) selecionando um subconjunto de SNPs. Todas as interações possíveis são enumeradas para este subconjunto obtido como resultado do algoritmo SWSFS. Depois deste passo um pequeno subconjunto de candidatos

SNPs, aqueles que têm a contribuição mais significante para a discriminação entre casos contra e controles, é selecionado.

Na segunda etapa, um procedimento hierárquico é adotado para declarar que a significância estatística dos SNPs candidatos estão associados com o risco de doença. Em análises de 1 lócus, é aplicado o teste estatístico B para cada SNP candidato, e então reporta-se todos os SNPs cujos p-valores são menores que um certo α depois de aplicar correção de Bonferroni para L testes. Em testes de interação de 2-lócus, se aplica o teste estatístico β para todos as interações possíveis de 2-lócus, e reporta-se as interações com p-valor menor que certo α depois da correção de Bonferroni para L(L-1)/2 testes. Se ambos SNPs investigados no teste de interação já foram reportados no teste de 1 lócus, então o teste de sua interação não será realizado.

Epimode [Tang W. et al, 2009]

EpiMODE (do inglês, *epistatic MOdule DEtection*) é uma generalização do método BEAM. Este método introduz módulos epistáticos para descrever o efeito independente de um único lócus ou efeitos interativos de múltiplos loci sobre as doenças. A base da ferramenta epiMODE é a definição de "módulos epistáticos" como uma pequena unidade genética que, independentemente influencia o risco de doença. Com base nesta noção, achados de SNPs que verificam epistasis são atribuídos a módulos epistáticos. Esta atribuição é feita calculando a probabilidade dos dados observados dado um padrão determinado de partição de SNP usando um modelo Bayesiano e, em seguida, obtendo a probabilidade posterior de um SNP pertencente a cada módulo epistático.

Uma estratégia "*Gibbs sampling*" de amostragem com salto reversível de cadeia Markov e Monte Carlo (RJ-MCMC) é empregado para obter a probabilidade a posterior, e para facilitar a detecção dos módulos epistáticos. Finalmente, epiMODE reordena os testes de hipótese para filtrar os módulos epistáticos significativos.

GENN [Motsinger et al, 2008]

Grammatical evolution neural networks (GENN) é um software que utiliza uma abordagem de redes neurais em conjunto com algoritmos de computação evolucionaria para otimizar os parâmetros de entrada, arquitetura e pesos de uma rede

neural para melhorar o poder de identificar interações gene-gene. A técnica de evolução gramatical separa o genótipo do fenótipo no processo evolucionário e permite maior diversidade genética dentro da população que outro algoritmo evolucionário. Em GENN a gramática permite definir múltiplas conexões entre os nódos selecionados pelo algoritmo. Também o número de conexões permite que redes neurais mais complexas possam evoluir e por tanto incrementar seu poder de detecção de associação.

INTERSNP [Herold et al., 2009]

Este método seleciona SNPs para análise de interação utilizando informação a priori, tal como dados próprios de relevância genética/biológica, localização genômica, classe ou função. Fontes de informação para definir estratégias significativas podem ser evidências estatísticas de associação e relevância genética/biológica (locação genômica, classe funcional ou informação de pathway). O software inclui módulos baseados em regressão logística também como modelos log-linear para análise de múltiplos SNPs, simulações de Monte-Carlo para avaliar a significância genômica, análise de associação de pathway e análise de haplótipos, entre outros.

MECCPM – SNP [Miller et al., 2009]

Usa um modelo de máxima entropia (MECPM) acoplado com um modelo de busca estruturada. MECPM explicita interações que conferem poder preditivo do fenótipo. Ele identifica um subconjunto de marcadores permitindo k formas diferentes de interações entre estes marcadores.

MECCPM constrói um fenótipo a posteriori, baseado no principio de máxima entropia, codificando restrições do modelo que correspondem a interações 1 a 1. Permite codificação para modelo dominante ou recessivo para cada lócus em uma interação candidata. Busca interações utilizando uma heurística gulosa avaliando candidatos até a quinta ordem. Utiliza o critério de informação Bayesiana (BIC) como medida na estratégia de seleção.

MegaSNPHunter [Wan X. et al, 2009]

MegaSNPHunter recebe como dados de entrada os genótipos de casos e controles e produz uma lista ordenada de interações multi-SNP. O método funciona da seguinte maneira: o genoma inteiro é dividido em vários curtos subgenomas que cobrem uma área genômica de possíveis efeitos haplótipos. Para cada um destes subgenomas MegaSNPHunter constrói um classificador baseado em árvore investigando as interações multi-SNP, e seguidamente mede a importância dos SNPs com base em suas contribuições no classificador. O método mantém os SNPs relativamente mais importantes e permite que haja competição de uns com os outros em cada nível. A competição termina quando o número de SNPs selecionados é menor do que o tamanho do subgenoma. Finalmente, MegaSNPHunter provê extratos e relatórios valiosos das interações multi-SNP.

PBEAM [**Zhang Y. & Liu J., 2007**]

É uma versão paralela do BEAM (Bayesian Epistasis Mapping Association). Ele usa cadeia de Markov com algoritmo Monte Carlo, *Markov Chain Monte Carlo* (MCMC), para procurar efeitos de um único marcador e efeitos de interação de múltiplos *loci*. O algoritmo BEAM tem dois componentes essenciais: uma ferramenta de inferência Bayesiana para epistasia, implementada via MCMC, e um teste estatístico para avaliar a significância estatística. Como resultado desta análise, BEAM proporciona uma probabilidade a posteriori para cada marcador ou epistasia estar associado com a doença. Ele classifica os SNPs em três tipos: SNPs associados com a doença, SNPs contribuindo para a suscetibilidade à doença de forma independente e SNPs que influenciam o risco de doença em conjunto com outro SNP.

PLINK (fast epistasis) [Purcell et al., 2007]

No procedimento de busca exaustiva de *Plink*, uma busca simultânea é conduzida sobre todos os loci. O procedimento de busca não exaustiva de *Plink*, chamado "fast-epistasis", identifica um conjunto de loci que atingem um nível de significância convincente na avaliação do teste de associação de 1 lócus. Posteriormente examina todas as possíveis interações dos loci do conjunto selecionado.

Algoritmo

a) Definir o nível de significância α;
b) Aplicar o teste de associação para um único lócus e selecionar S o conjunto de loci cuja significância é maior a α;
c) Avaliar todos os pares (2 loci) do conjunto S aplicando a equação 4.1;
d) Estimar a OR. Para estimar a OR no modelo de interação de lócus x lócus, *Plink* preenche uma Tabela com as frequências alélicas da combinação dos loci e posteriormente colapsa essa Tabela 3x3 em uma Tabela 2x2 como é mostrada na Figura 4.3;
e) Calcular Z score como na equação 4.2;
f) Aplicar correção para testes múltiplos. Rejeitar todos os pares de loci com probabilidade ajustada $p(x_1,x_2) > \alpha$.

SNPHarvester [Yang C. et al, 2009]

Este é um método de busca gulosa e pode selecionar um conjunto de grupos significativos de SNP dentre centenas de milhares de forma eficiente. Estes grupos selecionados de SNPs podem então ser analisados por outros métodos. SNPHarvester é uma ferramenta útil porque a maioria das ferramentas de procura de interações epistáticas não consegue lidar com a grande quantidade de dados nos estudos GWA, portanto, eles precisam reduzir o conjunto de dados. SNPHarvester reduz eficientemente o número de SNPs e permite a aplicação direta de ferramentas estatísticas existentes na detecção de interação. SNPHarvester é uma ferramenta intermediária que usa genótipos de estudos GWA e proporciona grupos de SNPs, que devem ser analisados por programas como o MDR.

O método basicamente busca dentro de um grande conjunto de dados de SNPs, conjuntos ou grupos de SNPs que melhor explicam a existência de interações, e utiliza modelos de regressão logística sobre esses grupos para identificar as interações significantes entre os SNPs. Inicialmente o algoritmo SNPHarvester faz uma busca sobre todos os L SNPs do conjunto de dados, detectando os SNPs que sejam significantes. Os SNPs são considerados significantes com base no valor obtido pelo teste de qui-quadrado com dois graus de liberdade, após correção de Bonferroni.

Todos os SNPs que se mostrem significantes são removidos. O método tem como objetivo encontrar as interações epistáticas, ou seja, SNPs que apresentam efeitos conjuntos. Aqueles SNPs com efeitos principais são descartados da análise. Assim, para um valor fixo k, definido como o número de interações que serão testadas pelo método, onde $k \leq \ln_3 N_d - 1$, e N_d é o número de casos.

Assim, utilizando os SNPs sem efeitos principais são gerados múltiplos paths através de um algoritmo de busca local denominado *PathSeeker*. O algoritmo PathSeeker calcula a pontuação de cada path, de tal forma que percorre todo conjunto de SNPs que não estão no path, verificando se a troca de um dos SNPs do path, pelo SNP do conjunto aumenta a pontuação. Caso a pontuação aumente, o SNP é trocado, caso contrário o path permanece igual, ou seja com os mesmos SNPs. Dessa forma, o algoritmo PathSeeker atualiza apenas um SNP por vez em um path. Ao mesmo tempo em que se verifica a pontuação do path também é verificado se o valor excede um determinado limiar T. Caso ultrapasse, o path é adicionado a uma coleção M composta por grupos de SNPs que serão analisados em um pós-processamento. No pós-processamento, os SNPs selecionados em um path são analisados utilizando uma regressão logística penalizada que indicará as melhores interações encontradas.

SNPRuler [Wan X. et al., 2010]

SNPRuler usa uma abordagem de aprendizado baseado em regras para detectar interações epistáticas. A aprendizagem de regras é utilizada para inferir interações onde cada interação epistática implicitamente contém algumas regras preditivas. Descobrir e avaliar regras é muito mais fácil e mais rápido do que encontrar e avaliar as interações. O algoritmo de aprendizagem utilizado procura identificar as regras para inferir possíveis interações epistáticas. Apesar de uma regra preditiva não pode garantir a existência de interação epistática entre os SNPs na regra, esta abordagem reduz o número de possíveis interações de forma a aproveitar aquelas estatisticamente significativas.

Com o objetivo de encontrar as melhores regras preditivas, uma medida de relevância U(·) é utilizada para ordenar as regras que contenham interações verdadeiras. A partir da medida da regra preditiva U(·), um limite superior é definido para evitar a expansão desnecessária de uma determinada regra, evitando uma busca exaustiva ou estocástica das interações. Após obtidas as melhores regras utilizando a

medida U(·), o algoritmo então constrói uma árvore de busca para cada regra selecionada, onde cada nodo representa um SNP e cada ramo que liga os nódos representa uma possível interação. Posteriormente, um método de busca em profundidade (do inglês, *depth-first transversal*) gera e avalia as possíveis interações utilizando a estatística qui-quadrado ajustada pelo uso da correção de Bonferroni. Ao final o algoritmo exibe em sua saída uma lista de interações ordenadas através da estatística qui-quadrado.

4.3 COMPARAÇÃO DOS MÉTODOS DE BUSCA

Como o tamanho dos conjuntos de dados disponíveis excede meio milhão de marcadores, é evidente que uma pesquisa exaustiva ou de força bruta das interações de SNP-SNP apresentará dificuldades computacionais e estatísticas que poderão afetar sua viabilidade. Alguns destes métodos, como FastANOVA, TEAM, EpiSNP e FastEpistasis fazem uso da computação de alto desempenho para superar a complexidade computacional, mas tropeçam na necessidade do controle de testes múltiplos, levando assim a resultados enviesados. Outros métodos conseguem trabalhar bem com um conjunto de dados de tamanho pequeno, como MDR, que tem melhor desempenho que métodos que utilizam regressão logística.

No entanto, alguns dos métodos discutidos neste capítulo podem ser adaptados para tirar vantagem da informação biológica. Os métodos estocásticos como Epiforest, MegaSNPHunter, Epimode, BEAM, GENN selecionam iterativamente um pequeno número de locos e realizam um teste completo para epistasia. Esta estratégia baseia-se na sorte para selecionar loci que interagem em pelo menos uma iteração. Outros como MECCPM, SNPRuler, InterSNP, SNPHarvester e AntiEpiSeeker realizam uma busca parcial das interações utilizando critérios heurísticos para filtrar a quantidade de SNPs no estudo. O sucesso desta estratégia depende da natureza das iterações presentes no conjunto de dados: as interações epistáticas puras sem efeitos principais são susceptíveis de serem desperdiçadas.

Ainda assim, o exame incompleto de todas as possibilidades traz consigo outro conjunto de problemas. Como se pode ter certeza de que as interações que escolhemos ignorar são desinteressantes? Interações puras realmente existem em genética? Quantas iterações são necessárias em uma busca aleatória para chegar a

uma conclusão razoável? É difícil ou impossível de responder satisfatoriamente a essas perguntas. No próximo capítulo será descrita a proposta de uma metodologia que pode ajudar a superar estas dificuldades, fornecendo ferramentas úteis à execução dos experimentos e análises de GWAS.

A Tabela 4.1 mostra uma comparação dos métodos descritos considerando critérios como nome do algoritmo, numero de interações suportadas, tamanho amostral para 2 loci, tipo de teste aplicado e estratégia de busca empregada denotada como 'E' para busca exaustiva e 'NE' para busca não exaustiva.

Algoritmo	**Nro de interações**	**T. amostra (para 2 loci)**	**Tipo de teste aplicado**	**Estratégia de busca**
FastANOVA	2 loci	~100,000	Teste ANOVA	E
COE	2 loci	~100,000	Testes estatísticos convexos como Chi quadrado, razão de máxima verossimilhança, Informação mutua e teste Cochran-Armitage	E
TEAM	2 loci	~500,000	Arvore de expansão mínima baseada em testes estatísticos convexos (como em COE)	E
MDR	k loci	~10,000	Data mining	E
BOOST	2 loci	~500,000	Máxima Verossimilhança	E
EpiSNP	2 loci	~500,000	Modelo Kempthorne	E
FastEpistasis	2 loci	~500,000	Regressão logística	E
PLINK	2 loci	~500,000	Regressão logística	E/NE
AntEpiSeeker	k loci	~100,000	Chi-quadrado/ACO	NE
SNPRuler	k loci	~100,000	Aprendizado baseado em regras	NE
InterSNP	2 loci	~300,000	Regressão logística	NE
MECCPM	k-loci	~300,000	Critério de informação bayesiana (BIC)	NE
SNPHarvester	k-loci	~500,000	Regressão logística penalizada	NE
BEAM	2 loci	~500,000	Modelo bayesiano	NE
Epiforest	2 loci	~100,000	Random Forest	NE
GENN	2 loci	~100	Neural Network	NE
MIGA-2L	2 loci	~500,000	Algoritmo Genético	NE

Tabela 4.1: Comparação de alguns métodos utilizados para detecção de interação de SNPs.

A Tabela 4.2 apresenta vantagens e desvantagens de alguns dos métodos descritos neste capitulo. Os principais critérios considerados foram poder de detecção, tamanho amostral, método de validação para evitar resultados espúrios (falsos positivos) entre outros.

Algoritmo	Vantagens	Desvantagens
FastANOVA, COE	Disponibiliza vários testes estatísticos convexos; Disponível para uso.	Não considera genótipos heterozigotos; Tamanho amostral pequeno; Carece de validação, o que poderia resultar em resultados enviesados.
TEAM	Disponibiliza vários testes estatísticos convexos; Disponível para uso.	Carece de validação, o que poderia resultar em resultados enviesados.
MDR	Bom poder para detectar associação; Não assume um modelo genético apriori; Utiliza modelo de validação cruzada com fator 10; Bem avaliado e entendido por vários grupos de pesquisa; Disponível para uso.	Algoritmo intratável para amostras de escala genômica; Tem problemas para detectar associações na presencia de lócus com heterogeneidade.
FastEpistasis	Fornece processamento paralelo do modulo epistasis de Plink; Escala linearmente com o numero de processadores considerado.	Erro no calculo da Variância; Não realiza estimação dos genótipos devido a dados incompletos; Problemas para detectar associações em um conjunto de genótipos sem efeito marginal.
PLINK (exaustivo)	Bom poder para detector associação assumindo certo tipo de modelo genético sobre os dados; Amplamente difundido e disponível para uso.	O teste exaustivo é computacionalmente custoso. Problemas na detecção de associação para modelos com efeito marginal fraco.
Plink (não exaustivo)	Tratável para dados de escala genômica; Método simple e facil de implementar; Disponível para uso.	Detecção de associação em ausência de dados com efeito marginal fraco é perdida devido a uma busca incompleta do espaço das possíveis associações.
SNPRuler	Seu algoritmo baseado em aprendizado de regras fornece fácil interpretação; Não assume uma distribuição apriori sobre os dados; Proporciona uma lista de interações classificadas por significância.	Não pode detectar interações epistáticas contendo regras conjuntas; Não considera modelos de heterogeneidade genética (ex. efeito marginal fraco);. Não realiza validação para evitar resultados espúrios para reduzir os falsos positivos.
SNPHarvester	Complexidade de busca linear; Fornece a possibilidade de remover SNPs com significante efeito marginal para a detecção correta de interações epistáticas.	A remoção de SNPs com efeitos marginais significativos limita a possibilidade de identificar todos resultados das interações epistáticas; A seleção aleatória do conjunto inicial dos SNPs utilizada pelo algoritmo PathSeeker pode limitar a detecção de associações importantes.
BEAM	Permite incorporar conhecimento experto utilizando uma distribuição a priori sobre os dados; Bom poder de detecção de associação em modelos de interação com MAF baixo; Disponibilidade para uso.	Problemas para detectar associações sobre dados de genótipos sem efeito principal; Tempo de execução lento em comparação com outros métodos.
Epiforest	Random Forest são rapidos para construir; Bom poder de detecção de interações com efeito epistático puro; Suporta vários formatos de arquivos.	Problemas para detectar interações com pequeno ou nenhum efeito marginal; Utiliza uma votação consensual que limita a lista de loci de suscetibilidade com o fenótipo estudado.
GENN	Capacidade de aprender sobre um determinado conjunto de dados e fazer previsões sobre os mesmos, onde o resultado da doença é desconhecida; Software disponível para uso.	Precisa de ajustes sobre os dados; Factível somente para um pequeno conjunto de dados.

Tabela 4.2a: Vantagens e desvantagens de métodos para detecção de associação (2 loci).

Algoritmo	Vantagens	Desvantagens
MIGA-2L	Utiliza validação cruzada estratificada de fator 10 para evitar resultados espúrios; Pode ser utilizado para dados de escala genômica; Não assume nenhuma distribuição apriori sobre os dados; Bom poder de detecção de interação sobre vários modelos genéticos.	A escolha dos parâmetros para a execução pode afetar no desempenho do algoritmo; Não fornece uso de fenótipos contínuos; Não realiza estimação dos genótipos devido a dados incompletos.

Tabela 4.2b: Vantagens e desvantagens do algoritmo *MIGA-2L* para detecção de associação (2 loci). *MIGA-2L* será descrito no proximo capitulo.

4.4 SÍNTESE DO CAPÍTULO

Este capítulo teve como finalidade fazer uma revisão geral sobre as ferramentas utilizadas para investigar efeitos de interação epistática em GWAS. Estes procedimentos foram vistos de uma maneira panorâmica, procurando enfatizar a técnica de abordagem empregada por cada um deles. Nos últimos anos muitos métodos foram propostos, a fim de resumir alguns deles e ajudar no seu reconhecimento, estes métodos foram agrupados segundo o tipo de busca empregada para a identificação de uma interação epistática. Foram incluídos métodos que utilizam uma busca exaustiva e os que utilizam busca não exaustiva. De todo o grupo, pode-se destacar o ***Plink*** por ser uma ferramenta robusta e uma das mais difundidas e utilizadas em estudos que envolvem GWAS. O modulo 'fast-epistasis' de ***Plink*** foi escolhido para realizar um estudo comparativo com o algoritmo *MIGA-2L*, que foi desenvolvido neste trabalho de pesquisa para investigar associação de SNPs com doenças. No final deste capitulo foram apresentadas Tabelas comparativas dos métodos citados considerando vários critérios de desempenho.

CAPÍTULO 5: METODOLOGIA PROPOSTA

Este trabalho teve como enfoque o desenvolvimento de uma metodologia para a descoberta de marcadores genéticos (SNPs) de doenças abarcando desde o pré-processamento dos dados até a identificação dos haplótipos que manifestam risco de desenvolvimento da doença estudada. A ideia básica é que dado um conjunto de SNPs de indivíduos casos e controles, a metodologia consiga descobrir um subconjunto destes relacionados com a doença em estudo. A metodologia proposta pode ser esquematizada em 6 passos, como é mostrado na Figura 5.1.

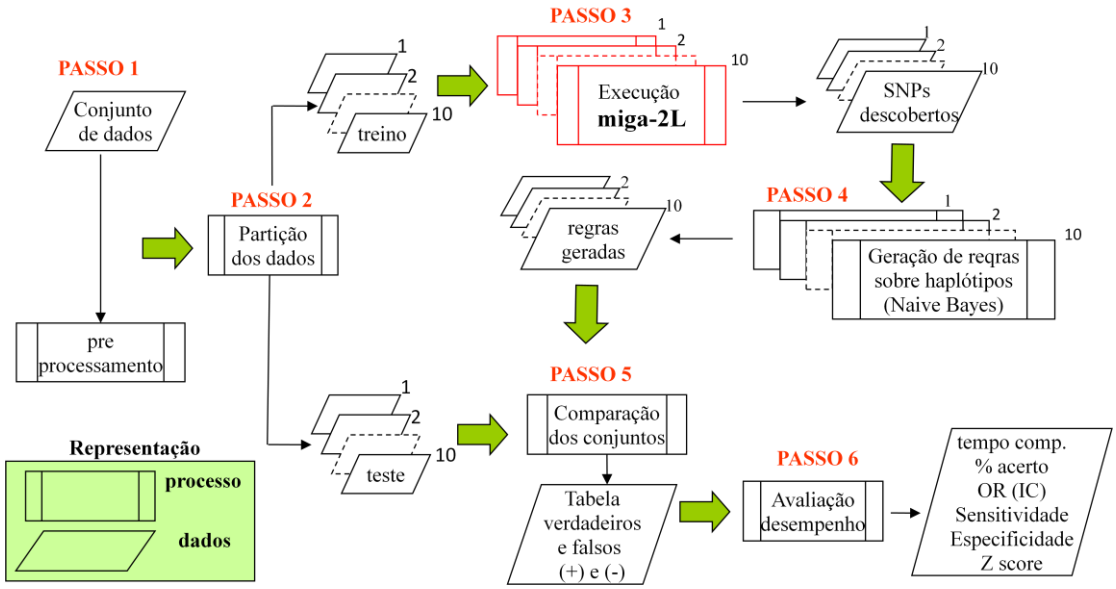


Figura 5.1: Visão geral da metodologia proposta, esquematizada em 6 passos.

Em resumo, a metodologia considera um conjunto de SNPs de indivíduos casos e controles e realiza uma avaliação da qualidade dos dados para seguidamente particionar o conjunto completo em subconjuntos de treino e teste utilizando a estratégia de validação cruzada. Posteriormente, cada conjunto de treino será submetido com a execução de um algoritmo genético para descobrir os SNPs que apresentam susceptibilidade com a doença em estudo. Depois da seleção dos SNPs será realizada uma classificação para determinar os haplótipos associados com a doença. Finalmente será realizada a avaliação do desempenho da metodologia.

MOTOR DE GERENCIAMENTO DO WORKFLOW: QUIRON

A metodologia proposta faz uso de um Sistema de Gerenciamento de Workflow (SGWf) para a execução e gestão de todas as etapas mostradas na Figura 5.1. Os SGWf permitem automatizar uma sequencia de ações, atividades ou tarefas na execução de um experimento científico, permitem realizar um controle de cada etapa do mesmo e aportam as ferramentas necessárias para seu controle ou gestão do fluxo de trabalho. Ele se caracteriza, principalmente, pela adequada integração com sistemas de informação atuais: banco de dados, gestão documental, mensagens, etc., permitindo a ampliação de um processo simples à integração de vários processos inter-relacionados. SGWf é um software que suporta modelagem e execução de workflows científicos, coletando a proveniência dos dados durante o desenho e execução do workflow. No entanto, experimentos de grande-escala requerem frequentemente o uso de um ambiente de computação paralela. Poucos SGWf estão prontos para execução paralela nos atuais ambientes de computação de alto desempenho. Assim, para permitir o uso eficiente desses recursos, foi utilizado Quiron [Ogasawara et al., 2011], um motor de workflow científico de dados centralizados que executa, em paralelo, aplicações científicas.

O Quiron implementa um gerenciador para dirigir a execução do workflow. Basicamente, Quiron coordena a execução paralela do workflow, atribuindo conjuntos de parâmetros de entrada diferentes aos nós de computação. Quiron usa uma abordagem de dados centralizados usando álgebra de workflow científico para gerenciar a execução paralela do workflow eficientemente. A álgebra padroniza o consumo e produção de dados e também abre horizontes para otimização do workflow. Quiron utiliza uma linguagem declarativa (XML) para definir os workflows e transformá-los em uma expressão algébrica permitindo a otimização automática do workflow, ele também estabelece um plano de execução otimizado paralelo para workflow.

A execução paralela de um workflow apresenta várias dificuldades para a coleta de dados de proveniência, porque estes dados também são distribuídos em todo o ambiente de computação de alto desempenho em diferentes nós de um cluster ou mesmo em diferentes máquinas virtuais em um ambiente de nuvem. No banco de dados gerado pelo Quiron, apenas os dados de proveniência são armazenados, tais como metadados e os vários resultados extraídos. Os dados de aplicativos

intermediários, tais como arquivos enormes e outros resultados de computação complexos, são apenas referenciados no banco de dados. Estes arquivos são mantidos na área de armazenamento da aplicação. Proveniência é essencial para experiências científicas e de engenharia e garante que o experimento possa ser reproduzido sobre condições diferentes. Quiron requer software adicional, como PostgresSQL, Java e bibliotecas adicionais como MPJ[Carpenter et al., 2000] e HSQLDB [Simpson and Toussi, 2007]. Estes são softwares de código aberto que podem estar disponíveis em centros de Computação de alto desempenho. Quiron está disponível no cluster do Centro de Computação de Alto Desempenho da Universidade Federal do Rio de Janeiro para qualquer projeto ou cientista que deseja usá-lo.

A Figura 5.2 apresenta uma visão simplificada de como Quiron trabalha em um cluster de computação de alto desempenho em ambiente paralelo onde A, B e C representam as atividades do workflow.

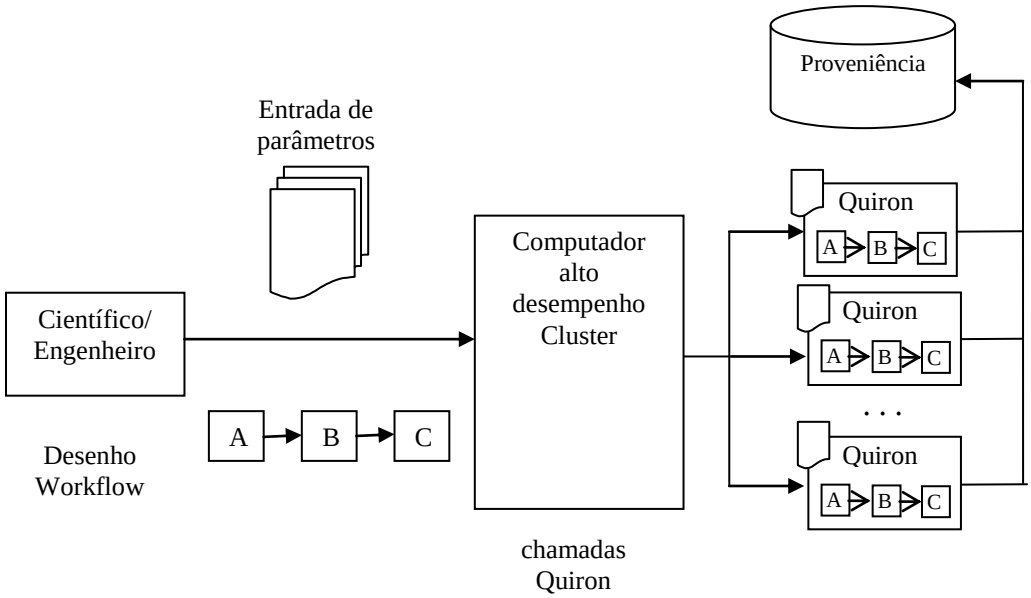


Figura 5.2: Arquitetura de execução de Quiron

5.1 PASSO 1: PRÉ-PROCESSAMENTO DO CONJUNTO DE DADOS

O conjunto de dados utilizado segue o formato padrão de estudos de associação que consideram grupos de indivíduos casos e controles. Estes dados podem ser representados como uma matriz onde as linhas representam os indivíduos casos e controles e as colunas representam os marcadores genéticos ou SNPs que serão analisados no estudo. Uma coluna adicional identifica se o indivíduo é caso ou controle.

Existem varias nomenclaturas para definir um SNP, para nosso propósito consideramos cada SNP como a informação do valor do alelo que é formado por {A>C, A>G, A>T, C>G, C>T, G>T}, o símbolo ">" representa alternativa, isto significa a ocorrência de mais de um alelo em um lócus, onde pelo menos dois alelos aparecem com frequência > 1% na população. A informação dada pelo SNP em certo lócus é chamada genótipo. Como os seres humanos são diplóides, isto é, possuem duas cópias de cada cromossomo, podem acontecer os seguintes casos: os dois cromossomos contêm o mesmo alelo que é o mais presente na população (*homozigoto dominante*), ambos contêm o mesmo alelo que é o mais raro na população (*homozigoto recessivo*) e um cromossomo possui o alelo mais comum enquanto outro possui o mais raro (*heterozigoto*).

Ind	SNP_1	SNP_2	SNP_3	SNP_4	SNP_5	Classe
G_1	1	1	2	1	1	1
G_2	1	1	2	0	1	1
G_3	2	1	2	1	2	1
G_4	2	1	2	1	2	1
G_5	1	1	0	1	2	0
G_6	1	1	0	1	0	0
G_7	2	1	0	0	0	0
G_8	2	2	0	1	1	0

Tabela 5.1: Tabela de *6* indivíduos casos e controles com seus genótipos correspondentes em 5 SNPs. A coluna ressaltada indica um exemplo da variabilidade da frequência do genótipo entre casos e controles.

Para evitar confusão na terminologia utilizada nas próximas secções, uma o termo “amostra” será utilizado para descrever os genótipos correspondentes a um certo indivíduo, representada por uma linha na Tabela 5.1, o tipo de marcador utilizado neste trabalho de tese é o SNP, representado por uma coluna na Tabela 5.1 e seu valor é dado pelo seu genótipo, representado por uma célula na Tabela 5.1.

A Tabela 5.1 mostra um exemplo fictício de uma sequencia de genótipos. Nela estão retratados os indivíduos com seus respectivos genótipos para cada SNP. Para fins computacionais os possíveis valores assumidos por cada SNP, *homozigoto dominante*, *homozigoto recessivo* e *heterozigoto* estão representados com 0, 2 e 1 respectivamente. A última coluna, isto é, a informação de caso-controle, é representada com 1 para casos e 0 para controles.

5.1.1 ANÁLISE DE CONTROLE DE QUALIDADE

Com o objetivo de reduzir o número de associações com falsos positivos e falsos negativos, é importante realizar uma avaliação da qualidade dos dados que foram obtidos da coleta de amostras de indivíduos casos e controles, selecionados no desenho do estudo e no processo de determinação de genótipos. Potenciais vieses podem ser introduzidos nestes processos precedentes.

No entanto, realizar um controle de qualidade em dados de escala genômica é uma tarefa complicada devido ao tamanho dos dados e aos critérios que devem ser tidos em conta para uma avaliação adequada. Critérios de controle de qualidade são subjetivos e variam de um estudo a outro. Os filtros para selecionar amostras e marcadores para remoção não devem ser tão rigorosos a fim de não remover a maioria dos dados analisados, porém devem eliminar marcadores de pior qualidade.

Na tentativa de remover falsas associações positivas, minimizando o erro no processo, é preciso fazer um controle de qualidade que proporcione flexibilidade na avaliação de cada critério a fim de remover indivíduos ou marcadores com taxas de erro muito elevadas. Já que muitos milhares de casos e controles foram genotipados para maximizar o poder de detectar associação, a remoção de um punhado de indivíduos deve ter pouco efeito sobre o estudo. No entanto, cada marcador removido em um estudo é potencialmente uma associação com a doença negligenciada e, assim, o impacto da remoção de um marcador é potencialmente maior do que a remoção de

um indivíduo. A remoção de uma pequena porcentagem destes não deve diminuir consideravelmente o desempenho do estudo já que técnicas de imputação do genótipo podem ser utilizadas para recuperar estes marcadores.

Por este motivo, uma boa pratica é considerar primeiro um controle de qualidade sobre as amostras (análise por linha) para seguidamente conduzir a avaliação sobre os marcadores (análise por coluna). Deste modo, cuidamos de não remover erroneamente marcadores devido a um subconjunto de amostras mal genotipadas, mas o estudo continua susceptível à remoção errônea de amostras com base em um subconjunto de marcadores mal genotipados.

Com a finalidade de realizar uma abordagem cuidadosa para evitar a remoção desnecessária de amostras e marcadores, neste trabalho utilizamos Quiron, um sistema de workflow paralelo, para a identificação de marcadores e amostras que devem ser removidas antes da realização da análise de associação. O modelo proposto consegue analisar dados de escala genômica e integrar programas científicos robustos para interagir com o usuário mostrando gráficos que ajudam na inspeção visual e na escolha de parâmetros adequados para realizar o controle de qualidade. Nas seções seguintes são descritos os critérios de controle de qualidade considerados, os quais foram agrupados por critérios de qualidade por amostra e critérios de qualidade por marcador.

5.1.2 CRITÉRIOS DE CONTROLE DE QUALIDADE

A metodologia utilizada para atribuir qualidade aos dados primeiro realiza uma avaliação da qualidade das amostras e posteriormente à remoção das amostras de baixa qualidade, realiza uma avaliação da qualidade dos marcadores (SNPs), identificando aqueles que serão excluídos do estudo. Estes controles foram implementados utilizando softwares científicos de livre acesso e comprovada robustez para análise GWAS (Seção 5.1.4).

Critérios para avaliação das amostras

Neste processo cuidamos de não remover erroneamente marcadores devido a um subconjunto de indivíduos mal genotipados, por isso tivemos em conta 4 critérios de qualidade que se descrevem a continuação.

1. Determinação da qualidade do genótipo (qualidade de DNA baixa).

A qualidade do DNA é medida segundo a taxa de falha de determinação do genótipo e a taxa de heterozigosidade de cada amostra. Amostras com baixa qualidade de DNA ou concentração estão, frequentemente, abaixo da média de taxas de determinação (*call rates*) e precisão do genótipo. Assim as amostras com mais que 3-7% de genótipos com falha ou ausentes foram selecionadas para remoção.

2. Contaminação da amostra de DNA ou endogamia e identificação de discordantes (outliers) de heterozigosidade através de autossomos.

A distribuição da heterozigosidade média (excluindo os cromossomos sexuais) em todos os indivíduos deve ser inspecionada para identificar indivíduos com um número excessivo ou reduzido de genótipos heterozigotos, isto pode ser indicativo de contaminação por DNA ou endogamia, respectivamente.

Por essa razão é avaliada a diferença da distribuição de heterozigocidade em homens e mulheres. Não deve haver heterozigocidade no sexo masculino, senão espera-se algum erro de genotipagem. As discrepâncias nas informações sobre gênero pode refletir erros nos dados, mistura da amostra e/ou informação de sexo inconsistente com cromossomos sexuais.

3. Duplicação ou parentesco baseado sobre *identidade-por-estado* (do inglês *identity-by-state*).

As amostras devem ser independentes, isto é, o parentesco máximo entre qualquer par de indivíduos é inferior a um parente de segundo grau. Se parentes de primeiro ou de segundo grau estão presentes, um viés pode ser introduzido para o estudo porque os genótipos dentro das famílias vão ser sobre representados, e, assim, a amostra pode não ser mais um claro reflexo das frequências alélicas na população total. A métrica utilizada para identificar indivíduos duplicados e relacionados é IBS (*identity by state*) a qual é calculada para cada par de indivíduos com base da proporção média de alelos compartilhados em comum em SNPs genotipados (excluindo os cromossomos sexuais).

A média da população IBS irá variar dependendo da frequência do alelo dos marcadores genotipados dentro dessa população. Indivíduos relacionados irão compartilhar mais alelos IBS do que o esperado por acaso, com o grau de partilha adicional proporcional ao grau de parentesco.

Considerando *M* marcadores, o IBS entre o *i-ésimo* e *j-ésimo* indivíduo é dado pela equação 5.1.

$$IBS = 1 - \left(\frac{1}{2} * M\right) * \sum_k \left|G_{ik} - G_{jk}\right| \qquad [5.1]$$

onde, G_{ik} denota a quantidade do alelo menor (em nosso caso 0) levado pelo *i-ésimo* indivíduo no SNP *k.*

Amostras idênticas irão compartilhar IBS perto de 100% (permitindo assim erros de genotipagem). Indivíduos relacionados (aparentados) irão compartilhar IBS maior do que indivíduos não aparentados.

4. Avaliação de incompatibilidades com informação externa (mistura de amostras) e ascendência da população remota (confusão devido à estrutura da população)

Em estudos genéticos a principal fonte de confusão é a estratificação da população, em que as diferenças genotípicas entre casos e controles são geradas por causa de origens diferentes da população ao invés de qualquer efeito sobre o risco de doença [Cardon&Palmer, 2003; Campbell et al., 2005]. No esforço para remover ou reduzir o efeito da estratificação da população é usado análise de componentes principais (PCA) [Patterson & Reich, 2006], este método permite identificar indivíduos com diferenças ancestrais em grande escala. O modelo PCA é construído usando dados de genótipos de genomas de populações ancestrais conhecidas (ex. Europa (CEU), Ásia (CHB+JPT) e África (YRI)), estes dados são obtidos do HapMap Internacional Consortium [2003]. Devido às diferenças genéticas em grande escala entre estes 3 grupos ancestrais, os primeiros dois componentes são suficientes para agrupar separadamente indivíduos destas 3 populações.

Critérios de avaliação dos marcadores.

O impacto da remoção de marcadores (SNPs) no estudo pode causar resultados espúrios, já que podemos remover um marcador potencialmente associado com a doença ou podemos deixar como parte do estudo um marcador mal genotipado, introduzindo informação errônea. Para superar estes inconvenientes tivemos em conta quatro critérios de qualidade que se descrevem a continuação.

1. Identificação de SNPs com uma excessiva ausência de genótipos.

Dados brutos vindos da genotipagem trazem uma porção de dados faltantes. O critério utilizado nestes casos é realizar a remoção de SNPs, chamamos de subótimos, com uma taxa de determinação do genótipo menor que 97% como sugerido pelo protocolo de controle de qualidade publicado por um grupo de pesquisadores da Welcome Trust Consortium [Anderson et al., 2010].

2. Identificação de SNPs demonstrando um significante desvio desde o equilíbrio de Hardy-Weinberg (HWE).

Removemos SNPs que mostram desvio significante o **HWE,** que podem ser indicativos de um erro de determinação de genótipo. No entanto, desvios do HWE poderiam também indicar seleção. Uma amostra caso pode mostrar desvios do HWE em *loci* associados com a doença, e, obviamente, seria contra-produtivo remover esses *loci* de novas investigações. Portanto, somente as amostras de controle devem ser utilizadas no teste para desvios de HWE. O limiar de significância para que SNPs permaneçam em equilíbrio de Hardy-Weinberg tem variado muito de um estudo a outro (p-valores variam entre 0,001 e 5,7 × 10-7)[The Wellcome Trust, 2007]. Neste estudo, SNPs com p-valor<0.00001 em controles foram removidos.

3. Identificação de SNPs com uma diferencia significante na taxa de genótipos ausentes entre casos e controles.

Em estudos onde os casos e/ou controles foram obtidos de várias fontes diferentes, é aconselhável testar diferenças significativas na taxa de determinação do genótipo, frequência do alelo e frequência do genótipo entre esses vários grupos para assegurar que é possível tratar o combinado conjunto caso ou controle como um grupo homogêneo. Neste estudo foram removidos SNPs com uma diferença significante (p-valor<0.00001) de taxa de genótipos ausentes entre casos e controles, como sugerido em [Anderson et al. 2010].

4. Identificação de SNPs com MAF muito baixo.

Tipicamente são utilizados uma taxa de frequência de alelo menor, MAF, entre 1% a 2% é aplicado, mas estudos com tamanho de amostra pequeno poderiam necessitar uma proporção mais alta. Neste estudo, SNPs com MAF<0.01 foram removidos [Anderson et al., 2010].

No entanto, mesmo após um rigoroso controle de qualidade de SNP, erros de genotipagem podem ainda persistirem. Verificar manualmente gráficos é a melhor maneira de garantir que a determinação de genótipos seja robusta e, portanto, é essencial que todos os SNPs associados com o estado da doença sejam inspecionados manualmente antes de escolher SNPs para estudos de seguimento (*follow-up genotyping*). Na Figura 5.3 se mostra o fluxo do processo.

5.1.3 WORKFLOW PARALELO PARA CONTROLE DE QUALIDADE DOS DADOS

Neste estudo foram considerados todos os critérios descritos na secção anterior. O workflow foi desenhado para aproveitar o paralelismo intrínseco dos processos que estão envolvidos no QC. A Figura 5.3 apresenta um esquema do fluxo de processos desenhado para a implementação do workflow.

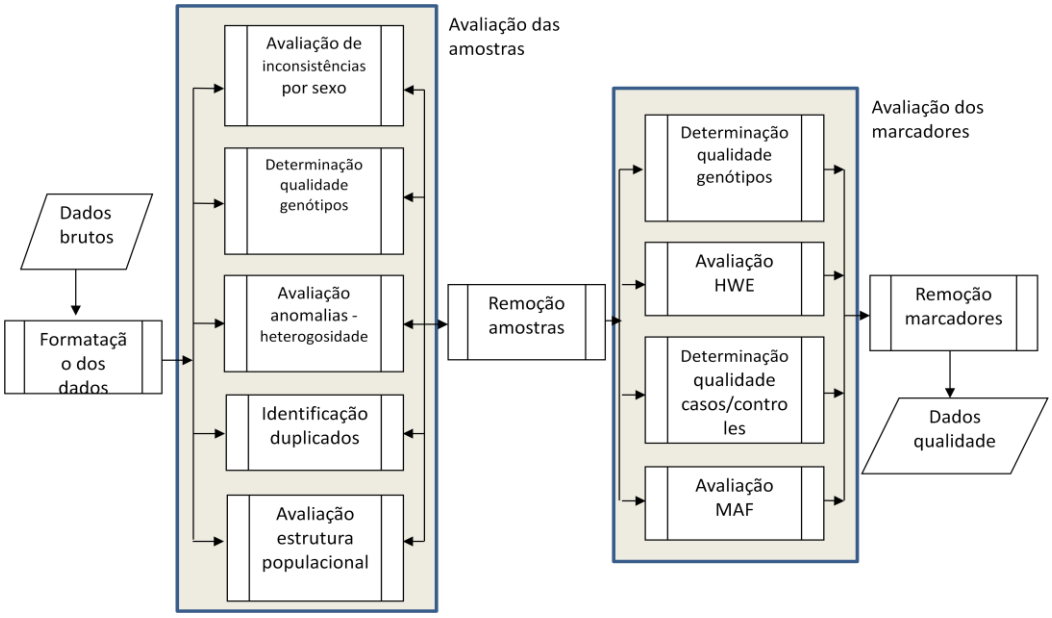


Figura 5.3: Fluxo do processo de controle de qualidade dos dados. Note que o workflow consiste de uma sequencia de passos concatenados (conectados), que segue o paradigma de fluxo, onde cada passo segue o precedente.

Cada processo foi implementado utilizando programas computacionais que foram integrados com algoritmos especialmente desenvolvidos para o workflow desenhado. A Figura 5.4 exibe o workflow conceitual. Em alguns pontos da execução é possível fazer uma parada para analisar os relatórios e gráficos que são fornecidos durante a execução. Nestas análises pode-se mudar algum parâmetro e seguidamente

realizar a re-execução do workflow. Também, pode-se fazer a escolha de novos parâmetros que irão alimentar a atividade seguinte.

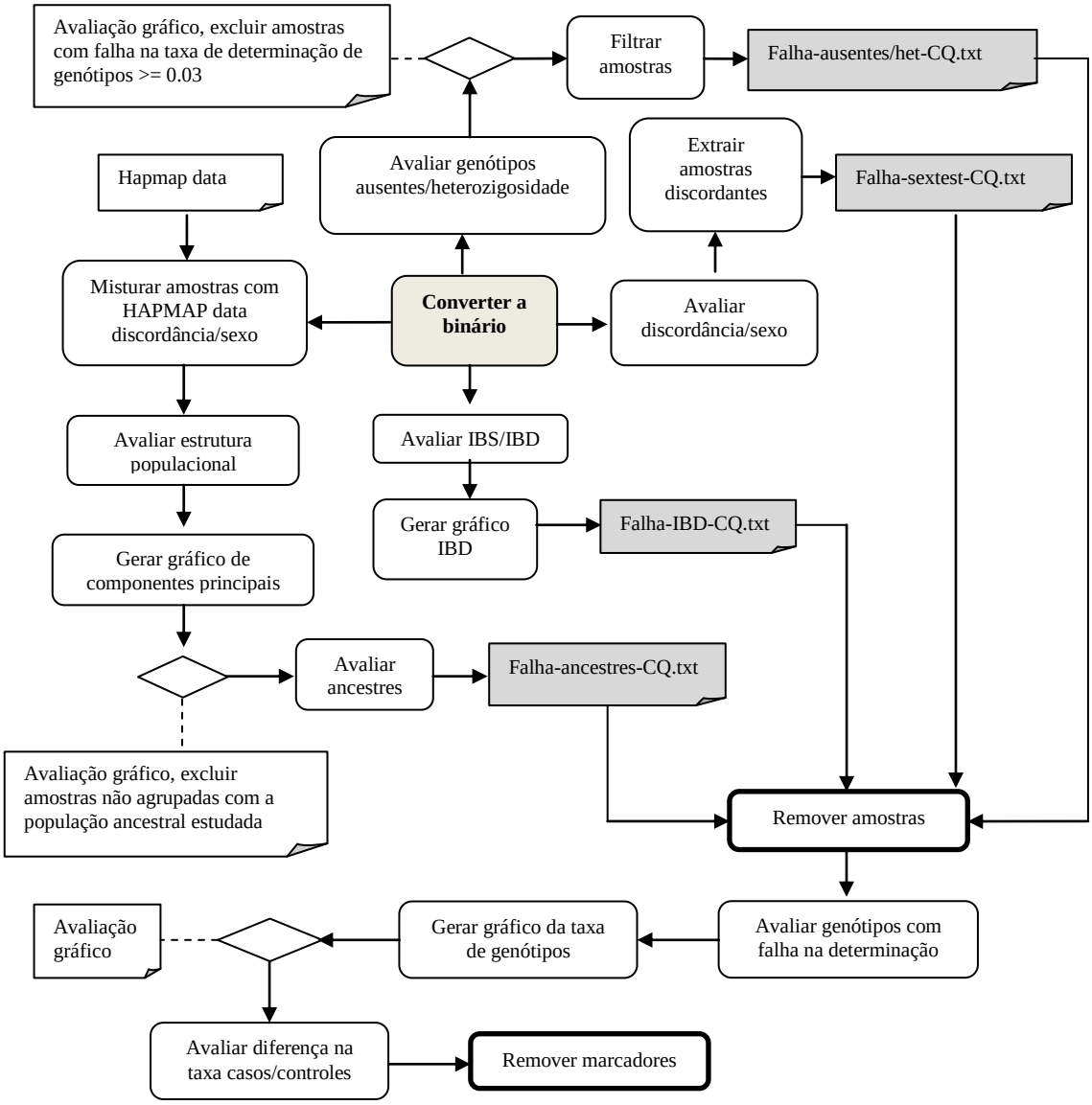


Figura 5.4: Workflow conceitual para o pré processamento dos dados.

5.1.4 PROGRAMAS COMPUTACIONAIS UTILIZADOS

Os softwares científicos que foram integrados no workflow Quiron são de livre acesso e de comprovada robustez para análise GWAS, a continuação listamos eles:

- **PLINK** um software para análise de GWAS. http://pngu.mgh.harvard.edu/~purcell/plink/
- **SMARTPCA**.pl software para PCA.

  http://genepath.med.harvard.ude/~reich/Software.htm

- **R** ambiente estatístico para análise de dados e gráficos. http://cran-r-project.org
- **Perl** é um linguagem de programação interpretada, é especialmente versátil no processamento de cadeias (strings), manipulação de texto e na análise de padrões, implementado através de expressões regulares.

5.2 PASSO 2: PARTIÇÃO DOS DADOS

Para realizar a avaliação do desempenho do modelo proposto, os dados originais são divididos em subconjuntos de treino e teste. Estes subconjuntos deverão ser independentes e balanceados, com 50% de seus dados correspondentes a casos e 50% correspondentes a controles. Os dados de treino são disponibilizados para ser analisados pelo algoritmo genético que realiza a identificação e seleção dos SNPs mais significativos. Posteriormente é feita a classificação dos haplótipos que sugerem uma potencial associação com a doença estudada. Finalmente, é feita a avaliação do desempenho do algoritmo comparando estes resultados com os dados de teste.

Cabe ressaltar que os modelos obtidos a partir dos dados de treino só serão considerados "bons", do ponto de vista da acurácia preditiva, se ele classificar corretamente uma alta porcentagem das instancias (exemplos) dos dados de teste. Em outras palavras, esses dados devem representar um conhecimento que possa ser generalizado para os dados de teste, que não foram utilizados durante o treinamento.

Esta divisão reduz o tamanho dos dados analisados, considerando 2/3 dos dados para treino e 1/3 para teste, sendo está uma proporção comunmente utilizada e sugerida no estudo realizado em [Kohavi R., 1995]. Como a acurácia dos resultados é estimada baseada em uma única partição dos dados (teste), que não é muito significante do ponto de vista estatístico foi utilizada a estratégia de validação cruzada estratificada de k partições (k-fold cross-validation).

Na validação cruzada, primeiro todos os dados são aleatoriamente divididos em k mutuamente exclusivas partições do mesmo tamanho, onde k é um parâmetro definido pelo usuário. Nós usamos um valor de k=10, produzindo 10 procedimentos de validação cruzada. Esta escolha foi baseada no estudo descrito por Kohavi [Kohavi, 1995] onde mostrou que para conjuntos de dados do mundo real semelhantes aos utilizados nesta tese, o melhor método a ser usado para a seleção do modelo é a validação cruzada estratificada em dez vezes. As partições realizadas são do mesmo tamanho onde 50% dos dados são casos e 50% controles, mantendo assim cada partição balanceada.

Então o algoritmo é executado 10 vezes, e na *i-ésima* execução, onde i=1,2,..,10, a *i-ésima* partição será usada como subconjunto de teste e as 9 restantes serão combinadas e usadas como subconjunto de treinamento para essa execução. A Figura 5.5 mostra um exemplo da partição do conjunto de dados, as partições de cor escura representam o conjunto de treino e a partição de cor clara representa o conjunto de teste para cada simulação.

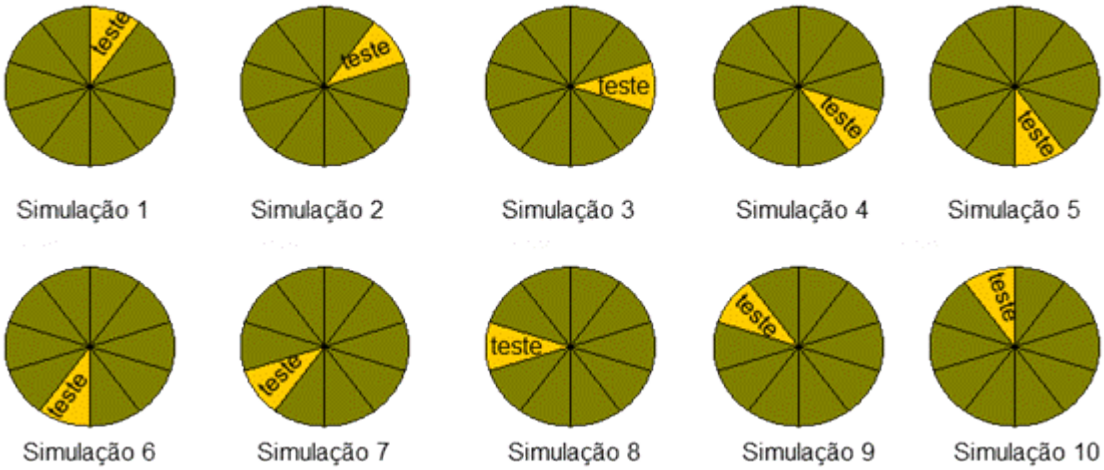


Figura 5.5: A validação cruzada é repetida 10 vezes permutando todos os subconjuntos. Cada partição é usada uma vez para teste e exatamente *9* vezes para treino.

5.3 PASSO 3: EXECUÇÃO DO *MIGA-2L* PARA DESCOBERTA DE SNPS ASSOCIADOS A DOENÇAS

Com a finalidade de apresentar o algoritmo *MIGA-2L*, uma sequencia de genótipos ***g*** correspondente a *m* SNPs, foi representada como $\mathbf{G}=\{g_1, g_2, ..., g_m\}$, onde $g_i \in \{0, 1, 2\}$. Foi utilizado o valor 0 e 1 para representar os alelos homozigotos e 2 para representar os alelos heterozigotos e $0 \leq i \leq m$.

$$\boldsymbol{g_i} = \begin{cases} 0\text{: dois alelos do } \textit{i-ésimo} \text{ SNP são homozigotos dominantes} \\ 1\text{: dois alelos do } \textit{i-ésimo} \text{ SNP são homozigotos recessivos} \\ 2\text{: dois alelos do } \textit{i-ésimo} \text{ SNP são heterozigotos} \end{cases}$$

Inicialmente consideramos um conjunto de dados de sequências de genótipos sobre *m* SNPs de N indivíduos que queremos analisar. Esta amostra se divide em dois grupos de indivíduos, aqueles que estão relacionados com certa doença, chamados de casos, e aqueles indivíduos não relacionados com a doença, chamados controles. Então, a representação da *k-ésima* amostra pertencente a um indivíduo será:

$$\mathbf{G}_k = g_1 \# g_2 \# ... \# g_m \# C, \text{ onde } C=\{0,1\}$$

onde, C=0 corresponde aos indivíduos controles e C=1 corresponde aos indivíduos casos. O símbolo "#" representa concatenação.

Nosso objetivo é encontrar uma combinação par de SNPs, $S = \{ g_i, g_j \}$, onde $0 \leq i,j \leq m$ e $i \neq j$, o qual consiste em selecionar o par de SNPs mais significativos do conjunto de dados de genótipos considerado. Estes deverão prover informação preditiva sobre a doença que está sendo estudada. Na próxima seção se descrevem os principais algoritmos envolvidos nesta etapa, que são o algoritmo de seleção e o algoritmo de classificação.

5.3.1 ALGORITMO *MIGA-2L*

O objetivo do algoritmo consiste em, dado um conjunto de dados de genótipos relacionados a certa doença, descobrir os pares de SNPs que são mais informativos no conjunto. Como estratégia deste processo de busca, um Algoritmo Genético (AG) foi

aplicado. Os AGs demonstraram ser adequados na otimização de problemas complexos como é o caso da tarefa que queremos resolver. A principal razão de nossa escolha é que os AGs são capazes de explorar os efeitos das interações entre SNPs, sem pressupor conhecimento a priori do modelo genético que possui a doença estudada, enquanto que outras metodologias poderiam ignora-lhas devido ao desconhecimento de um modelo válido adequado.

A maioria das abordagens computacionais de doenças comuns apresentadas no capítulo anterior, estão interessadas em considerar todas as possíveis combinações de interações (combinações de dois, de três, etc.). No entanto, como foi apontado na introdução, uma busca exaustiva pode ser extremadamente demorada. (ex. para 21 SNPs existem $\binom{21}{2} = 210$ possíveis combinações pares, $\binom{21}{3} = 1330$ combinações triplas, $\binom{21}{4} = 5985$ combinações quádruplas, etc.), o número de testes que deverá ser feito leva a uma computação intensiva.

Os AGs são métodos de busca robustos e flexíveis, que tendem a lidar bem com interações entre variáveis, devido a sua natureza de busca global. Desta forma, intuitivamente eles podem ser facilmente adaptados para tratar um alto grau de interação entre SNPs. Os AGs, através de um processo aleatório como seleção natural (sobrevivência do mais apto), mutação e cruzamento, investigam só um subconjunto destas possíveis interações. Contudo, havendo descoberto uma interação importante, ele é capaz de preservar este padrão em futuras gerações [Congdon C.B., 1995; Packard N.H., 1990; Freitas, 2001; Carvalho, 2005].

Os AGs foram desenvolvidos por John Holland que em 1975 publicou o seu livro "Adaptation in Natural and Artificial Systems" [Goldberg D., 1989]. Os componentes principais são o esquema de codificação, inicialização da população, função de avaliação (do inglês, *fitness function*), seleção, operador de cruzamento e o operador de mutação. O fluxograma do algoritmo proposto para a tarefa de descoberta dos SNPs que indicam associação com a doença é mostrado na Figura 5.6.

Descrição do Algoritmo Genético

O AG cria inicialmente, de forma aleatória, uma população ***P*** de tamanho ***NPOP***. Cada elemento desta população tem o comprimento ***m***, igual ao número de SNPs que é considerado no estudo. O algoritmo avalia cada elemento da população ***P,***

escolhendo de forma aleatória duas posições entre 1 e *m*. Estas posições correspondem às localizações de dois SNPs no arquivo de Genótipos **G**. Então, é computada a informação mútua [Cover & Thomas, 2006] desses dois SNPs no conjunto de dados de Genótipos **G**. Este valor será chamado ***fitness*** do elemento. Posteriormente, os operadores de mutação e cruzamento são aplicados aos elementos da população selecionados. Logo após a aplicação dos operadores genéticos, como o tamanho da população ***P*** pode crescer, são selecionados ***NPOP*** melhores elementos, segundo seu ***fitness***, para formar parte da nova população. Este processo é repetido até que um certo número de gerações seja alcançado. À continuação se descreve com mais detalhe cada componente do algoritmo genético.

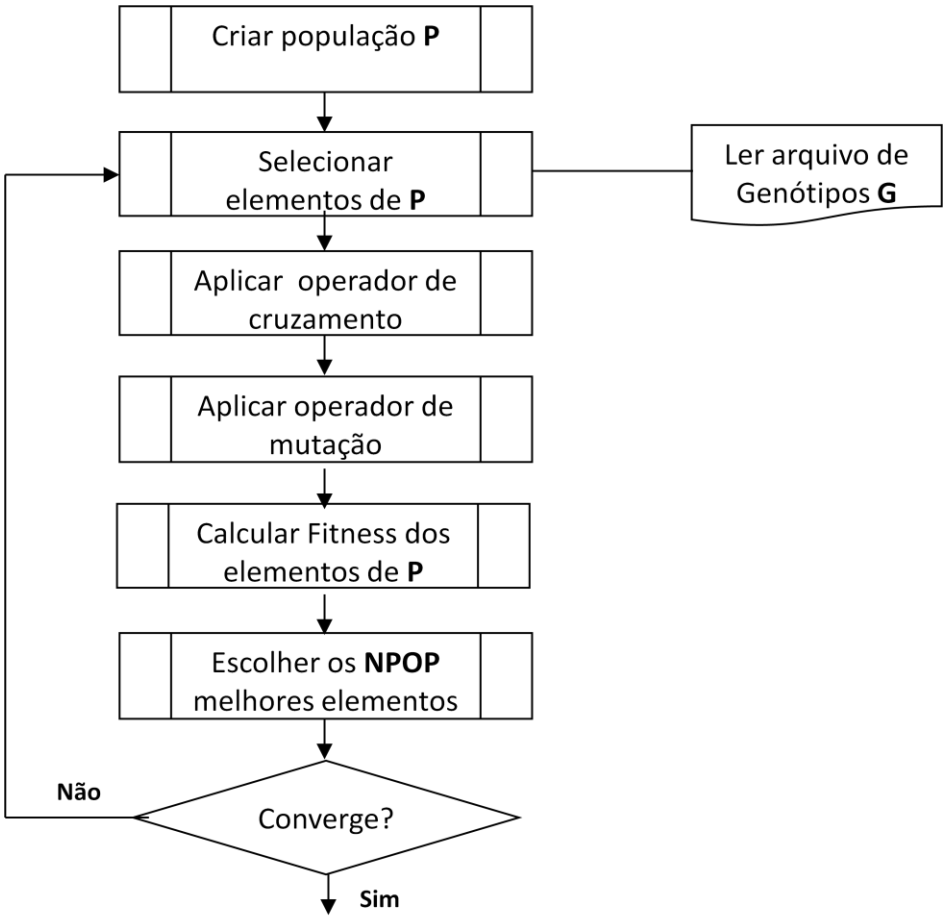


Figura 5.6 Fluxograma do algoritmo *MIGA-2L* (Mutual Information Genetic Algorithm – 2 loci) [Goldberg D., 1989]

5.3.1.1 Esquema de codificação

Fundamental para a estrutura do AG é o esquema de codificação. Nesta implementação, o método de codificação binária foi utilizado para representar cada elemento da população **P** do AG, que tem *m* bits de comprimento, igual à quantidade total de SNPs sendo considerados no estudo. O método de codificação binária utilizado pode ser descrito para o *i-ésimo* elemento da população **P** como:

$$E_i = S_{i1} \# S_{i2} \# ... \# S_{im},$$

onde $0 \leq i \leq NPOP$, $S_{ij} = \{0,1\}$, $0 \leq j \leq m$, *NPOP* é o tamanho da população **P,** e S_{ij} = 1 significa que o *j-ésimo* SNP sobre o *i-ésimo* elemento foi selecionado ou ativado.

Por exemplo, neste esquema de codificação um elemento representado por E_i = 1 # 0 # 0 # 0 # 1, descreve os SNP_1 e SNP_5 como ativos, os SNPs restantes (SNP_2, SNP_3 e SNP_4) são mascarados, representados como inativos e não serão considerados no processo de avaliação. A Figura 5.7 corresponde ao exemplo.

SNP_1	SNP_2	SNP_3	SNP_4	**SNP_5**
1	0	0	0	1

Figura 5.7 Representação do *i-ésimo* elemento da população AG de comprimento 5, indicando os SNP_1 e SNP_5 como ativos.

5.3.1.2 População inicial P

O modelo de inicialização foi produzido atribuindo o valor 1 a duas posições escolhidas de forma aleatória em cada elemento da população **P**. Desta forma cada elemento de **P** possui dois bits igual a 1 (ativos) e os restantes igual a 0.

De aqui em diante a população do AG será chamada **P** para diferenciar do conjunto de dados de genótipos que será chamado **G,** que utilizado na avaliação de cada elemento de **P**.

5.3.1.3 Função de avaliação (*Fitness function*)

A função de avaliação é um dos mais importantes parâmetros em um AG. Ela é usada para determinar quais elementos serão selecionados durante a operação de

seleção. Nesta implementação foi utilizada uma abordagem baseada na teoria da informação [Cover and Thomas, 2006] aplicada como medida biológica para investigar duas variáveis. Esta abordagem foi descrita no *Apêndice*.

5.3.1.4 Representação das variáveis aleatórias como medida biológica

O objetivo perseguido é quantificar a informação que, por exemplo, dois SNPs provêm sobre um fenótipo C, tal como uma doença em particular. Baseado na teoria da informação, como descrito no *Apêndice*, esse "grau de informação" pode ser obtido definindo esses dois SNPs e o fenótipo C como as variáveis aleatórias. Dois marcadores di-alélicos, SNP_1 e SNP_2, possuem 9 combinações de genótipos como se mostra na Tabela 5.2. Então, a função de informação mútua, definida como $I(SNP_1;SNP_2;C)$, reduz os dados 9-dimensionais para uma variável de 1-dimensão.

		SNP_1		
		AA	Aa	aa
SNP_2	BB	AABB	AaBB	aaBB
	Bb	AABb	AaBb	aaBb
	bb	AAbb	Aabb	aabb

Tabela 5.2: As 9 combinações possíveis de dois marcadores (SNPs)

Basicamente, a função da Informação Mutual, $I(SNP_1;SNP_2;C)$, proporciona uma forma de medir o grau de informação que os SNP_1 e SNP_2 têm em presença da doença C. Neste contexto, a ideia é formular um teste de hipótese, onde a hipótese nula avalia se a doença e os SNP_1 e SNP_2 são independentes, ou seja se $I(SNP_1;SNP_2) = I(SNP_1;SNP_2;C)$ ou de outra forma se $I(SNP_1;SNP_2;C) - I(SNP_1;SNP_2) = 0$. Partindo deste razoamento, pode-se avaliar a associação entre a doença C e os marcadores SNP_1 e SNP_2 calculando se sua diferencia é diferente a zero.

Os dados de genótipos, para indivíduos casos e controles, providos como dados de entrada serão utilizados para calcular as probabilidades da frequência relativa de cada genótipo. Para o propósito da descrição do algoritmo este conjunto de dados será chamado ***G***. O dados de genótipos correspondentes a indivíduos

relacionados com a doença, ou seja casos, será chamado ***D***. O *i-ésimo* SNP é descrito pelo genótipo g_i e pode assumir 3 valores diferentes {0,1,2} como foi citado anteriormente.

Usando a definição de entropia de uma variável aleatória, pode-se definir a entropia $H(g_i)$ em **G** e a entropia condicional $H(g_i|C)$ em **D** nas equações 5.2 e 5.3.

$$H(g_i) = \sum_{g_i} p(g_i) log_2 p(g_i) \quad [5.2]$$

$$H(g_i|C) = -\sum_{g_i} p(g_i|C=1) log_2 p(g_i|C=1) \quad [5.3]$$

Considerando dois SNPs, g_i e g_j, define-se sua entropia conjunta $H(g_i,g_j)$ em G e a entropia condicional $H(g_i,g_j|C)$ em D nas equações 5.4 e 5.5.

$$H(g_i, g_j) = -\sum_{g_i}\sum_{g_j} p(g_i, g_j) log_2 p(g_i, g_j) \quad [5.4]$$

$$H(g_i, g_j|C) = -\sum_{g_i}\sum_{g_j} p(g_i, g_j|C=1) log_2 p(g_i, g_j|C=1) \quad [5.5]$$

A informação mútua dos SNPs g_i e g_j é definida na equação 5.6 como:

$$I(g_i,g_j) = H(g_i) + H(g_j) - H(g_i,g_j) \quad [5.6]$$

Na população de genótipos D, indivíduos com a doença (casos), a informação mútua dos SNPs g_i e g_j é definido na equação 5.7 como:

$$I(g_i,g_j|C) = H(g_i|C) + H(g_j|C) - H(g_i,g_j|C) \quad [5.7]$$

Para quaisquer dois SNPs g_i e g_j, $I(g_i,g_j) \geq 0$ e $I(g_i,g_j)=0$ se e somente se g_i e g_j são independentes. Para quaisquer dois SNPs g_i e g_j, $I(g_i,g_j|C) \geq 0$ e $I(g_i,g_j|C) = 0$ se e somente se g_i e g_j são condicionalmente independentes dado o conhecimento da doença C.

Daqui, o grau de informação que dois SNPs, g_i e g_j, podem ter ao respeito de um fenótipo C é definido pela diferença da informação mútua dos dois SNPs na presença da doença (C=1) e a informação mútua dos dois SNPs na população de genótipos geral (conjunto G). Esta diferença será chamada de ***Ganho de informação*** e será calculada pela equação 5.8.

$$\mathbf{GI} = I(g_i,g_j|C) - I(g_i,g_j) \quad [5.8]$$

Finalmente, a função de avaliação ou ***fitness*** é formulada como:

$$\textbf{Maximo GI}(g_i, g_j, C) \qquad [5.9]$$

5.3.1.5 Operador de seleção.

Nesta implementação foi usado o método da roleta, que é comumente utilizado e simples de aplicar. Basicamente, a seleção aplica o seguinte mecanismo: cada elemento da população **P** é associado com uma fatia sobre uma roda virtual. Um setor cobre uma área maior na roleta quando o correspondente elemento tem um valor de função de *fitness* alto, enquanto um valor baixo é representado por um setor menor.

5.3.1.6 Operador de cruzamento.

Depois do processo de seleção o operador de cruzamento é aplicado. Neste processo um ponto de corte é escolhido de forma aleatória para cruzar dois elementos selecionados da população **P**. Os bits a partir deste ponto de corte são trocados entre esses dois elementos produzindo novos elementos. Na Figura 5.8 se mostra um cruzamento que produz dois novos elementos.

Caso 1: dois bits ativos

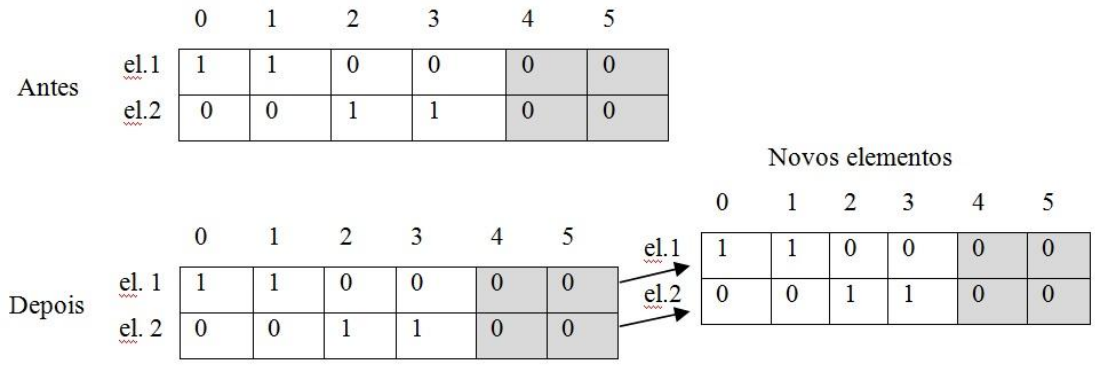


Figura 5.8. Cruzamento de dois elementos de tamanho 5 e ponto de corte igual a 4. Os bits trocados são representados pelo sombreamento. Neste caso o ponto de corte produz dois novos elementos com 2 bits ativos.

Segundo o ponto de corte escolhido, o número de bits ativos em cada elemento gerado no cruzamento pode variar de 0 a 4. Então, de acordo à quantidade de bits ativos em cada elemento depois do cruzamento, 3 casos podem ser

identificados: no caso 1 (Figura 5.8), os dois elementos possuem 2 bits ativos; no caso 2, um elemento fica com 1 bit ativo e o outro fica com 3 bits ativos; e no caso 3, um elemento fica com 4 bits ativos e o outro sem bits ativos.

A codificação implementada só considera elementos com 2 bits ativos já que o algoritmo avalia a interação entre dois SNPs. Então, para que os elementos da população **P** permaneçam com 2 bits ativos, aplicamos certas regras dependendo do caso identificado. No caso 1, mostrado na Figura 5.8, os dois elementos gerados possuem 2 bits ativos e não representam nenhuma mudança com respeito aos elementos selecionados antes do cruzamento e, portanto, esses dois elementos permanecem na população **P**. No caso 2 (um elemento com 3 bits ativos e o outro com 1 bit ativo), Figura 5.9, o elemento com 3 bits ativos é desdobrado em três novos elementos os quais resultam da combinação par desses 3 bits. Os novos elementos serão inseridos na população **P,** e aquele elemento que ficou com 1 bit ativo será descartado já que não cumpre os requisitos da codificação do algoritmo.

Caso 2: Depois do cruzamento obtém-se um elemento com três bits ativos e outro elemento com 1 bit ativo

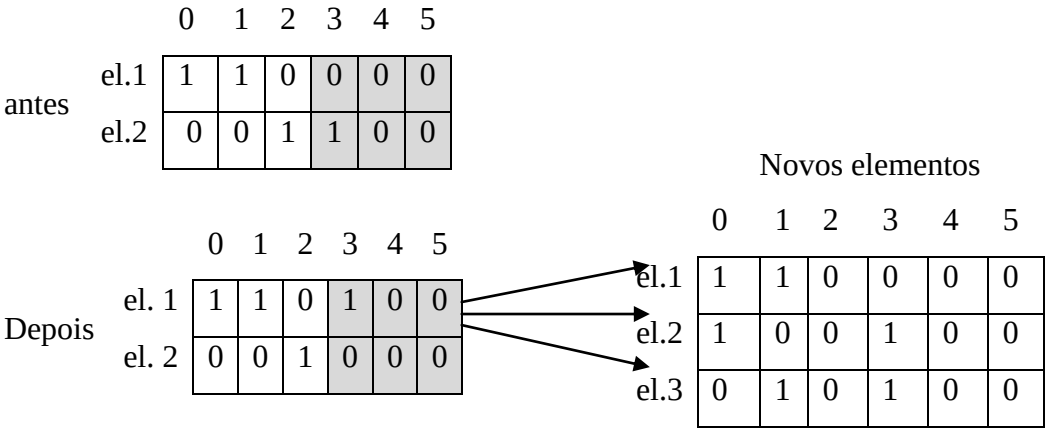


Figura 5.9. Cruzamento com ponto de corte igual a 3. Os bits trocados são representados pelo sombreamento. Neste caso o ponto de corte produz três novos elementos com 2 bits ativos.

No caso 3 (um elemento com 4 bits ativos e o outro sem bits ativos), Figura 5.10, o elemento que fica sem bits ativos é descartado pela mesma regra do caso 2. O outro elemento fica com quatro bits ativos, então aplicamos o desdobramento da

mesma forma que na regra do caso 2 que gera seis novos elementos que resultam da combinação par dos 4 bits ativos.

Caso 3: Depois do cruzamento, obtém-se um elemento com quatro bits ativos e outro com 1 bit ativo

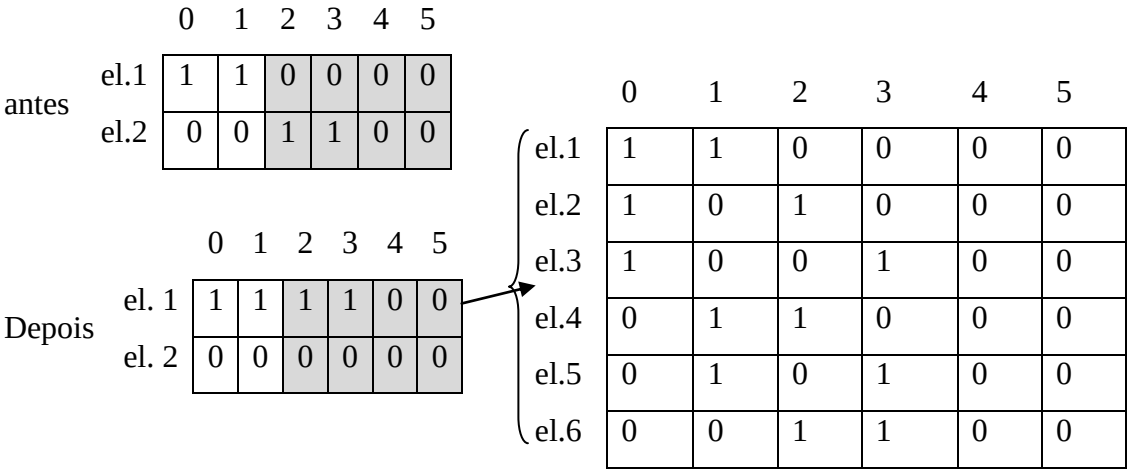


Figura 5.10. Cruzamento com ponto de corte igual a 2. Os bits trocados são representados pelo sombreamento. Neste caso o ponto de corte produz seis novos elementos com 2 bits ativos.

5.3.1.7 Operador de Mutação

Depois do cruzamento é aplicado o operador de mutação. Neste processo, para cada elemento selecionado da população **P,** é escolhido um ponto de forma aleatória cujo bit está zero e o ativamos a 1. Este processo produz dois novos elementos, que resultam da combinação par desse novo bit ativo com aqueles dois que já estavam ativos, Figura 5.11.

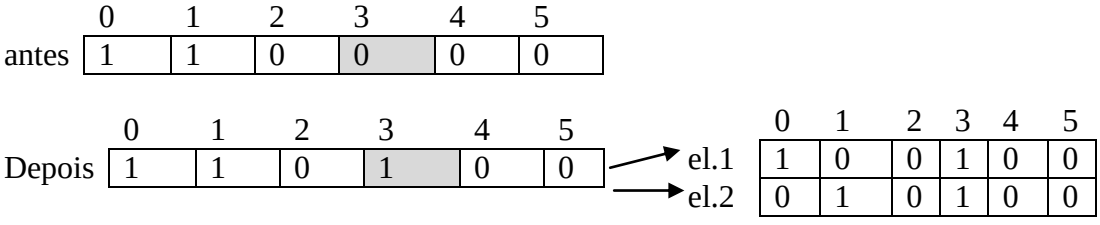


Figura 5.11. O ponto escolhido igual a 3 é representado pelo sombreamento. A mutação produz 2 novos elementos com 2 bits ativos.

5.3.1.8 Manter o melhores

A aplicação dos operadores genéticos de cruzamento e mutação gera novos elementos que são incorporados à população **P** aumentando seu tamanho inicial. O processo de manter os melhores escolhe os melhores NPOP elementos baseado na função de *fitness* e descarta os elementos restantes. Desta forma, o tamanho da população mantém-se a mesma em cada geração.

5.3.1.9 Critério de parada

Todo o processo se repete até a convergência da solução. Ao finalizar, a combinação de SNPs que teve o maior valor na sua função de avaliação é escolhida como a mais significativa da amostra de genótipos que está sendo estudada.

5.4 PASSO 4: CLASSIFICAÇÃO DOS HAPLÓTIPOS EM CASOS OU CONTROLES

Uma vez descobertos os SNPs que sugerem associação com a doença, precisamos determinar quais haplótipos apresentam susceptibilidade à doença e quais indicam proteção à doença. Para expressar essa informação, agruparemos os haplótipos obtidos da combinação de dois SNPs, g_i e g_j, em duas classes. A primeira classe, chamada *"caso"*, e a segunda classe, chamada *"controle"*. Então uma regra que classifica um haplótipo $\boldsymbol{h_k}$, associada com o risco à doença, será do tipo:

SE "h_k" **ENTÃO** "caso"

Esta regra indica que os SNPs g_i e g_j, com haplótipos "h_k" (Tabela 5.3) são classificados como casos, ou seja, estes haplótipos demonstram uma associação positiva, ou de risco com a doença estudada. Aquele haplótipo que apresenta proteção será do tipo **SE** "h_k" **ENTÃO** "controle".

Haplotipo(h_k)	00	01	02	10	11	12	20	21	22
$\boldsymbol{g_i}$	0	0	0	1	1	1	2	2	2
$\boldsymbol{g_j}$	0	1	2	0	1	2	0	1	2

Tabela 5.3: haplótipos gerados da combinação de dois genótipos.

Um haplótipo é uma combinação dos genótipos g_i e g_j,. A Tabela 5.3 mostra que para dois SNPs existe em total 9 haplótipos.

O algoritmo utilizado para a obtenção das regras é um classificador Bayesiano simples, chamado *Naive Bayes* [Duda, Hart, Stork, 2001]. Uma breve descrição pode ser encontrada no *Apêndice*.

No capítulo seguinte serão mostrados exemplos desta implementação com dados simulados e reais. Na Figura 5.12 se apresenta o fluxo do processo do passo 3 da metodologia.

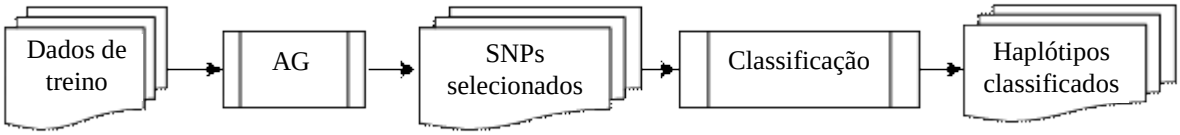


Figura 5.12: Fluxo do processo para a descobrir os SNPs de risco e classificar os haplótipos que apresentam associação com a doença estudada.

5.5 PASSO 5: COMPARAÇÃO DO CONJUNTO DE REGRAS E O CONJUNTO DE TESTE.

Cada execução completa do AG gera um conjunto de regras e tem também associado um conjunto de teste, como foi descrito no passo 2. Como são 10 execuções do algoritmo, isto significa que teremos 10 conjuntos de regras e 10 conjuntos de teste. Então nesta etapa são comparados estes dois conjuntos calculando as instancias do conjunto de teste que são cobertos pelo conjunto de regras.

Neste processo todas as instancias do conjunto de teste serão verificadas, para isso os seguintes dados são calculados para preencher a Tabela chamada de contingência:

VP (Verdadeiros positivos) = total de instancias casos no conjunto de teste cobertos pelas regras casos do conjunto de regras;

VN (Verdadeiros negativos) = total de instancias controles no conjunto de teste cobertos pelas regras controles do conjunto de regras;

FP (Falsos positivos) = total de instancias controles no conjunto de teste cobertos pelas regras casos do conjunto de regras;

FN (Falsos negativos) = total de instancias casos no conjunto de teste cobertos pelas regras controles do conjunto de regras.

Uma matriz de confusão contém os valores dos verdadeiros e falsos positivos e negativos, medidas que são habitualmente utilizadas na prática médica para auxiliar na avaliação da qualidade de um teste de diagnóstico. Assim, ao solicitar um teste de diagnóstico o médico se vê diante de quatro possibilidades: o exame resultar positivo na presença da doença (verdadeiro-positivo), positivo na sua ausência (falso-positivo), negativo na ausência da doença (verdadeiro-negativo) e negativo na ausência da doença (falso-negativo). Onde positivo é sinônimo de anormal e negativo sinônimo de normal.

Finalmente teremos 10 Tabelas de contingência que serão calculadas comparando os 10 conjuntos de regras com seus conjuntos de teste correspondente. Na Tabela 5.4 é mostrada uma Tabela de contingência padrão para falsos e verdadeiros positivos e negativos.

		Conjunto Teste	
		$Teste_{casos}$	$Teste_{controles}$
Conjunto Regras	$Regras_{casos}$	VP	FP
	$Regras_{controles}$	FN	VN

Tabela 5.4: Tabela de falsos e verdadeiros positivos e negativos

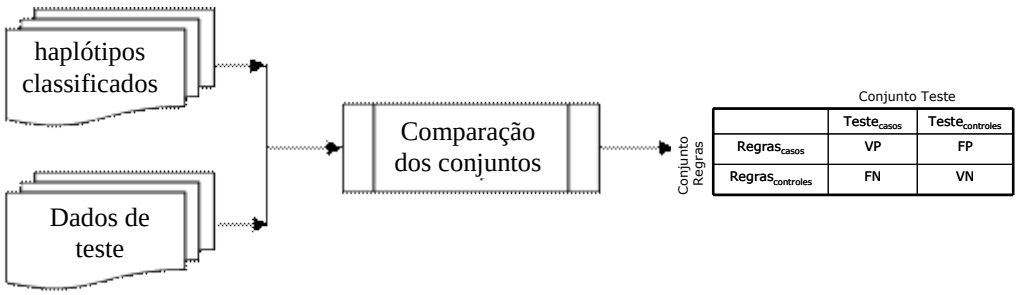


Figura 5.13: Fluxo do processo para gerar a Tabela de verdadeiros e falsos positivos e negativos.

5.6 PASSO 6: AVALIAÇÃO DE DESEMPENHO DA METODOLOGIA

Para avaliar o desempenho da metodologia, em cada execução será estimado o percentual de acurácia do conjunto de regras sobre o conjunto de teste, ou seja, o número de instancias no conjunto de teste que o conjunto de regras cobre. Este percentual é dado pela proporção entre os verdadeiros positivos e negativos em relação a todos os resultados possíveis.

$$acurácia = \frac{VP+VN}{VP+FN+FP+VN}$$

No final, a acurácia média estimada é simplesmente a média aritmética das 10 taxas de acurácia obtidas em cada execução.

$$acurácia\ média\ = \frac{\left(\sum_{i=1}^{10} \%acerto_i\right)}{10}$$

Também serão considerados como medida de desempenho três propriedades muito importantes em epidemiologia que medem a qualidade de um teste de diagnóstico. Estas propriedades são: Sensibilidade, Especificidade e Razão de chances.

Sensibilidade (S): é definido como a proporção de verdadeiros positivos entre todos os doentes (casos).

$$S = \frac{VP}{VP+FN}$$

Especificidade (E) : é definido como a proporção de verdadeiros negativos entre todos os sadios (controles).

$$E = \frac{VN}{FP+VN}$$

Já que existe um contrabalanço entre estas duas propriedades, de tal modo que quando uma delas aumenta a outra diminui e vice-versa, utilizaremos a soma das duas [Medronho R., 2009].

$$S + E = \frac{VP}{VP+FN} + \frac{VN}{FP+VN}$$

Curvas ROC

Uma forma eficiente de demonstrar a relação normalmente antagônica entre a sensibilidade e a especificidade dos estudos que apresentam resultados contínuos são as Curvas de Características de Operação do Receptor (Curvas ROC- Receiver Operating Characteristic). A Curva ROC é uma ferramenta poderosa para medir e especificar problemas no desempenho do diagnóstico em medicina por permitir estudar a variação da sensibilidade e especificidade para diferentes valores de corte.

A Curva ROC é um gráfico construído a partir dos valores estimados de sensibilidade (ou taxa de verdadeiros positivos) versus taxa de falsos positivos.

Razão de chances (Odds ratio)

A razão de chances é um bom estimador do risco relativo (RR) de ocorrência da doença nos expostos em relação aos não expostos, sempre que a prevalência da doença estudada nos não expostos seja igual ou menor que 5%.

Por exemplo, para uma Tabela 2x2 como a Tabela 5.5 o *Odds-ratio (OR)* é definido pela equação 5.10:

		Casos	Controles
Exposição fator risco	Sim	a	b
	Não	c	d

Tabela 5.5: Tabela de contingência 2x2. Contagem das frequências.

O *odds-ratio* é calculado pela equação

$$OR = \frac{axd}{bxc} \qquad [5.10]$$

"A odds-ratio (OR) é definida como a probabilidade de que um evento ocorra dividido pela probabilidade de que ele não ocorra"

A OR varia entre 0 a infinito, sendo o valor 1 indicador de não associação entre exposição e doença. Valores menores que 1 sugerem proteção da exposição e valores maiores que 1 sugerem um efeito deletério da exposição. Quanto mais distante de 1 para cima ou para baixo, mais forte é a associação.

A palavra "sugerem", utilizada no parágrafo anterior foi proposital, já que os dados sob estudo são amostrais. Assim, há que se considerar que as estimativas observadas podem refletir meras flutuações amostrais do verdadeiro efeito da exposição à doença. O verdadeiro efeito nunca será conhecido, mas pode-se dispor de uma "boa" estimativa dele quando se tem uma amostra representativa da população de referência.

Teste de significância

Uma vez calculada a OR, é preciso estimar se as OR obtidas são significantes. Para isto é necessário calcular seu erro padrão (SE do inglês *standard error*) e seu intervalo de confiança (por exemplo de 95%), para verificar se o 1 esta contido nesse intervalo. Se o extremo inferior deste exceder o valor de 1, se pode considerar igual que um teste de significação estatística.

Em estudos de GWAS, onde consideramos frequências alélicas, o cálculo da OR pode ser aproximado como medida estatística assumindo algum modelo epistático para a construção de uma Tabela de 2x2. Neste experimento foi assumido um modelo multiplicativo. Então é realizada a contagem dos alelos casos e controles, como se mostra na Tabela 5.6.

		Locus2		
		BB	Bb	bb
Locus1	AA	a	b	c
	Aa	d	e	f
	aa	g	h	i

Tabela 5.6: Contagem das frequências de genótipos nos loci 1 e 2.

Especificamente contamos os alelos independentes (A,a,B,b) observados em dois *loci* nos indivíduos casos (cujos valores são representados pelas letras a,b,c,d,e,f,g,h,i). De forma similar, se realiza a contagem para os controles.

Posteriormente, a Tabela 5.6 é colapsada em uma Tabela 2x2, seguindo a lógica seguinte: primeiro conta-se os alelos em um locus, por ex. Locus 2 considerando B condicional sobre o genótipo A, representado como uma Tabela 3x2, onde cada cela é calculada como se mostra na Tabela 5.7.

		Locus2	
		B	B
Locus1	AA	2a+b	2c+b
	Aa	2d+e	2f+e
	aa	2g+h	2i+h

Tabela 5.7: Tabela 3x2 de frequências condicionais do genótipo B sobre o genótipo A.

Esta Tabela é de novo colapsada em uma Tabela 2x2 como a Tabela 5.8

	B	b
A	C=4a+2b+2d+e	D=4c+2b+2f+e
a	E=4g+2h+2d+e	F=4i+2h+2f+e

Tabela 5.8: Tabela de contingência 2x2

Com os dados desta Tabela, agora é possível calcular a OR=CF/ED entre os loci A e B e seu SE para casos e controles em forma separada.

Pode se assumir que os dados seguem uma distribuição normal pelo qual podemos escolher o teste estatístico **Z** *score*. O teste estatístico Z é obtido calculando a diferencia entre as OR em casos e controles, segundo a equação 5.11:

$$\mathbf{Z} = (\log(R) - \log(S)) \ / \ \mathrm{sqrt}(SE(R) + SE(S)) \qquad [5.11]$$

onde, R e S são as *odds-ratio* para casos e controles respectivamente e SE é o erro padrão da OR.

Intervalo de confiança

Para o cálculo do intervalo de confiança de 95%, seguimos o seguinte procedimento:

1. calcular o logaritmo neperiano (ln) da OR (logarítmicos dos números naturais);
2. calcular o erro padrão (SE) do ln OR com a seguinte fórmula:

$$\mathrm{SE} = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

   onde a,b,c e d são os valores contidos nas celas da Tabela de contingência 2x2 (Tabela 5.8);

3. multiplicar o SE (ln OR) por Z score (Z=1,96 para 95%);
4. o resultado obtido em 3, se soma e se resta do ln OR;
5. para os valores obtidos, calcula-se o antilogarítmo neperiano e assim, se obtém o limite inferior e superior do intervalo de confiança. Se este intervalo não incluir o valor 1, pode-se considerar uma significância estatística equivalente a um nível de 5%.

5.7 SÍNTESE DO CAPÍTULO

Neste capítulo foi apresentada a metodologia utilizada para o descobrimento de marcadores genéticos de doenças. O processo começa com o tratamento dos dados brutos vindos da genotipagem que são submetidos a um processo de controle de qualidade seguindo um conjunto de critérios que foram detalhados no passo 2. Posteriormente, foi descrito o processo de particionamento dos dados que passaram o controle de qualidade. Seguidamente, cada partição dos dados é utilizada na execução do algoritmo *MIGA-2L* para identificar as interações Snp-Snp mais relevantes. Finalmente, foram descritas as medidas de desempenho utilizadas para avaliar o algoritmo.

No capítulo seguinte são apresentados os experimentos realizados com dados gerados através de simulações e com dados reais da população caucasiana.

CAPÍTULO 6: EXPERIMENTOS COMPUTACIONAIS

Neste capítulo são apresentados os resultados dos experimentos computacionais realizados em conjuntos de dados simulados e com dados reais. Tipicamente, os dados podem vir de observações do mundo real, com conhecidas associações, ou de dados simulados, gerados segundo modelos de dados que apresentam efeitos epistáticos em conhecidos loci funcionais. Dados reais são preferidos sobre os simulados, já que os modelos utilizados para gerar as simulações poderiam não representar da forma precisa processos biológicos complexos que envolvem as doenças humanas. Lastimosamente, se tem poucas referencias sobre interações epistáticas que foram descobertas e replicadas. Este é o motivo pelo qual se deve recorrer a simulações para avaliar o desempenho de algoritmos para GWAS.

Todos os experimentos foram realizados utilizando um computador Altix ICE 8400, com 128 CPUs Intel Xeon (640 cores) e 64 nós de processamento. Este computador possui memória distribuída de até 1.28 TBytes, funcionando com um sistema operacional Suse Linux Enterprise Server (SLES) e SGI Performance Suite. Foram utilizados compiladores Intel e GNU (C/C++) com suporte OpenMP. Este computador é parte do Núcleo de atendimento em computação de alto desempenho da Coppe/UFRJ (http://www.nacad.ufrj.br/)

6.1 PRÉ-PROCESSAMENTO DO CONJUNTO DE DADOS

Para avaliar o desempenho desta etapa da metodologia, foi utilizado um conjunto de dados de genótipos de 317,503 marcadores correspondentes a SNPs sobre 2,000 indivíduos, onde 1,023 são casos e 977 são controles. Destes, 997 correspondem a homens e 1,003 correspondem a mulheres. Este conjunto de dados contem 11,440 haplótipos de heterozigotos e 3,286 genótipos de SNPs ausentes ou não determinados. O conjunto de dados foi submetido aos critérios de controle de qualidade descrito no capítulo 5, referido como passo 1 da metodologia proposta.

Os dados brutos de genótipos foram coletados utilizando a plataforma de genotipagem Affymetrix [http://www.affymetrix.com/estore/] e inicialmente registrados em formato Chiamo [MARCHINI et al, 2007]. Posteriormente, foram convertidos a um formato de arquivos padrão .ped e .map. Estes formatos de arquivos

são utilizados pelo programa *Plink* e tornaram-se formatos padrão em GWAS (Tabela 6.1). O arquivo .ped contem os valores dos genótipos e informação referente às amostras. O arquivo .map contem a descrição de cada marcador (SNP) descrito em quatro variáveis. Finalmente, estes arquivos foram convertidos a um formato binário para otimizar seu tratamento.

.ped	.map
Família ID Individuo ID Pai ID Mãe ID Sexo(1=homem; 2=mulher; outro=desconhecido) Fenótipo	cromossomo (1-22, X, Y or 0 if unplaced) rs# ou snp identificador distancia Genética (morgans) posição em pares de base (bp units)

Tabela 6.1 Informação contida nos arquivos .ped e .map

Depois da formatação dos arquivos do conjunto de dados, começou a verificação da qualidade das amostras. Assim avaliou-se a discordância das amostras através dos cromossomos autossomos, calculando a heterozigosidade média em cada amostra. No desenho do workflow (Figura 5.4) esta atividade é indicada como 'Avaliar discordância/sexo' e gera um arquivo de texto com as amostras que não passaram na avaliação.

Em outra atividade, chamada 'Avaliar genótipos ausentes/heterozigosidade', compara-se a taxa de determinação de genótipos com a taxa de distribuição de genótipos heterozigotos. Esta atividade gera um gráfico onde se observa a densidade de genótipos nas amostras (Figura 6.1a). Cada ponto no gráfico denota uma amostra, onde o eixo X corresponde ao valor da taxa genótipos ausentes e o eixo Y corresponde ao valor da taxa de heterozigosidade. Depois de analisar o gráfico da Figura 6.1a foi feita a remoção das amostras cuja taxa de falha de genótipos foi acima do valor 0.03.

Também foi realizada a análise de duplicação ou parentesco das amostras utilizando o indicador IBS, identidade por estado, que é calculado para cada par de amostras com base na proporção de alelos compartilhados em comum. Esta atividade, chamada no workflow como 'Avaliar IBS/IBD', gera um gráfico onde cada ponto representa um par de amostras com valores das probabilidades de IBD, Pr(IBD=1)=0

no eixo X, e Pr(IBD=0)=0 no eixo Y. O IBD=1 significa que duas amostras compartilham 100% de seus alelos, o que pode indicar que estas amostras estão duplicadas ou que possuem algum nível de parentesco. A Figura 6.1b indica que existe um par de irmãos no canto inferior esquerdo, onde Pr(IBD=0)=0 e Pr(IBD=1)=0. Isto significa que este par de amostras compartilham 2 alelos idênticos por descendência em cada lócus do genoma. Isto pode ser uma indicação de amostra duplicada ou um conjunto de gêmeos idênticos. Os pontos no quadrante inferior direito do gráfico sugerem algumas amostras relacionadas, provavelmente em segundo ou terceiro grau de parentesco. O gráfico da Figura 6.1b dá uma pista sobre a escolha do limiar para corrigir potenciais problemas de identidade das amostras.

Do mesmo modo, foi realizada a avaliação de incompatibilidades devido à mistura das amostras e/ou confusão devido à estrutura da população. Aqui foi utilizado o programa SMARTPCA/EIGENSOFT para realizar o cálculo de componentes principais. O conjunto de dados estudado foi cruzado com 3 grupos de populações de referencia, população europeia (CEU), população asiática (CHB + JPT) e população africana (YRI). Estes dados foram obtidos do consórcio Hapmap [http://hapmap.ncbi.nlm.nih.gov/]. A Figura 6.1c mostra o gráfico de 2 componentes principais. O cruzamento resulto em 11 casos e 19 controles que não ficaram agrupados em nenhum grupo populacional, estas amostras foram removidas do estudo.

Posteriormente à remoção de amostras que não passaram o controle de qualidade, se realizou a análise de qualidade por marcador. Deste modo, foi verificada a qualidade dos genótipos calculando a taxa de falha na determinação do genótipo. Neste processo foi gerado um histograma (Figura 6.1d) para mostrar a taxa de falha de genótipos por amostra. Assim, aqueles SNPs acima de 3% na sua taxa foram removidos da análise devido ao excesso na taxa de falha de genótipos. Também nesta etapa, foram removidos os marcadores cuja frequência de alelo menor (MAF) foi menor que 0.01 e cujo p-valor de HWE foi menor que 0.00001.

Finalmente foram removidas 56 amostras e 3623 marcadores. O tempo computacional empregado pelo workflow de controle de qualidade foi aproximadamente de 4 horas, enquanto que o tempo empregado no *Plink* foi de aproximadamente 12 horas. A Tabela 6.2 apresenta um resumo do tamanho do conjunto de dados utilizado comparando-o antes e depois do pré-processamento.

Tamanho do conjunto de dados	Antes do pré processamento	Depois do pré processamento
Marcadores (SNPs)	317,503	313,880
Casos	1,023	1,000
Controles	977	944
Homens	997	966
Mulheres	1,003	978

Tabela 6.2 Tabela comparativa do tamanho do conjunto de dados antes e depois do pré processamento

A utilização do Quiron, assim como o uso dos softwares escolhidos otimizou o tempo computacional, ajudando a diminuir o tempo de processamento, já que várias atividades do workflow podem ser processadas em paralelo. Como foi mencionado no capítulo anterior, não existe um consenso sobre os critérios a serem considerados no controle de qualidade dos dados GWAS. Por esse motivo, o uso de um workflow provê maior liberdade e flexibilidade na escolha desses critérios, assim como dos programas científicos utilizados.

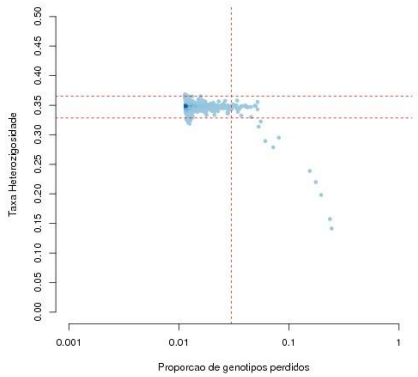


Figura 6.1a A densidade da amostra é indicada pelo sombreamento e as linhas tracejadas denotam o limiar a ser utilizado como corte. O valor escolhido como limiar foi de 0.03.

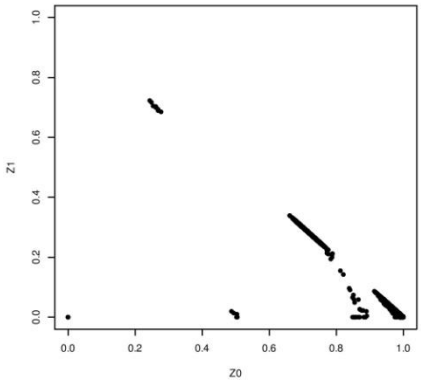


Figura 6.1b Identificação de amostras duplicadas. O gráfico identifica todos os pares de indivíduos com um IBD > 0.185. Aquelas amostras com IBD acima deste valor foram removidas.

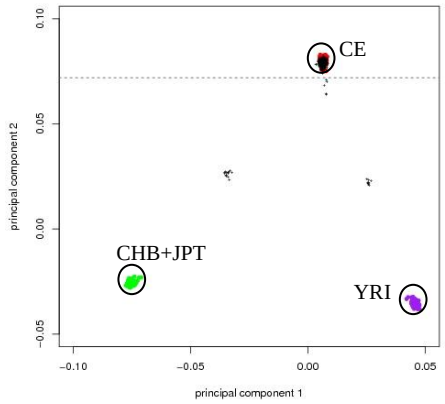


Figura 6.1c. Agrupamento ancestral baseado em amostras de referencia de Hapmap3: CEU (população européia), CHB+JPT (população asiática) e YRI (população africana). Cada ponto no gráfico corresponde a uma amostra. Aquelas que estão fora dos 3 grupos (em círculos) apresentam problemas de estrutura populacional e foram removidas do estudo.

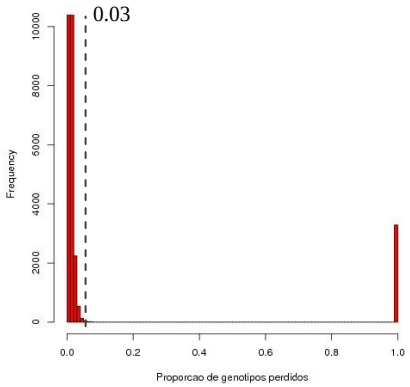


Figura 6.1d Proporção de falha de genótipos que permaneceram no conjunto de dados depois da remoção das amostras que falharam no Controle de qualidade. A linha tracejada indica o limiar escolhido de 3% para a remoção dos marcadores.

6.2 DESCOBERTA DE SNPS ASSOCIADOS À DOENÇAS UTILIZANDO *MIGA-2L*

Para compreender as causas subjacentes de doenças de características complexas, muitas vezes é necessário considerar os efeitos genéticos conjuntos também conhecido como epistasis através de todo o genoma. O conceito de epistasis discutido no capítulo 3, é geralmente definido como a interação entre diferentes genes. Aqui utilizaremos o algoritmo *MIGA-2L*, considerando a definição de epistasis estatística, para descrever o fenômeno biológico que um lócus com efeito sobre o fenótipo dependente de outro lócus. Desta forma, a análise matemática de epistasis fica mais simples.

Com o fim de avaliar o desempenho do algoritmo *MIGA-2L* proposto para detectar interação SNP-SNP, um estudo comparativo foi realizado com o modulo 'fast-epistasis' fornecido pelo programa *Plink.* Foram utilizados conjuntos de dados sintéticos de referência e conjunto de dados reias de 5 doenças comuns. O algoritmo *MIGA-2L* foi descrito no capítulo 5 e desenvolvido neste trabalho de tese. O nome foi

definido devido à abreviação, do inglês, de *mútual information genetic algorithm* para 2 loci. O programa *Plink* é utilizado amplamente pela comunidade científica em GWAS. O algoritmo utilizado no modulo 'fast-epistasis', de *Plink*, foi descrito no capitulo 4 como um método de busca não exaustiva. Na comparação também foram utilizadas medidas de desempenho epidemiológicas tanto quanto computacionais descritas no capítulo 5.

Os parâmetros utilizados em cada algoritmo foram:

***MIGA-2L* :** Os parâmetros a serem considerados são número de gerações, tamanho da população do algoritmo genético, taxa de cruzamento e taxa de mutação. Também pode se escolher o fator de validação cruzada;

***Plink*:** Para cada par de SNPs, *PLINK* aplica regressão logística sobre o modelo $P \sim \beta0 + \beta1 * Snp1 + \beta2 * Snp2 + \beta3 * Snp1 * Snp2 + \epsilon$, onde P é o fenótipo estudado. Neste estudo foi considerado o parâmetro "epistasis" do programa.

6.2.1 EXPERIMENTOS COM DADOS SIMULADOS

Um modelo de dados de interação epistática descreve como o efeito da combinação de genótipos (ex. SNPs) influencia no risco à doença. A maioria dos GWAS assumem a suposição de que o risco inerente à doença pode aumentar (ou diminuir) segundo a frequência dos alelos. Por isso associações de doenças são muitas vezes conceituadas em duas dimensões: frequência do alelo a qual é determinada pela penetrância e o tamanho do efeito que é estimado sobre a base da definição da OR (Odds Ratio) da doença. Neste contexto, foram simulados dois tipos de modelos de dados de interação epistática: com efeito forte chamado efeito principal e com efeito fraco ou sem efeito.

Um modelo epistático com efeito principal á aquele que descreve SNPs com efeito individual moderado ou grande sobre a doença e, um modelo epistático com efeito fraco ou sem efeito principal é aquele que descreve SNPs com pouco ou nenhum efeito individual, mas que apresentam forte influencia quando estão atuando em conjunto. Um exemplo foi descrito no capitulo 3 na Figura 3.3.

Conjunto de dados considerando loci com efeito principal

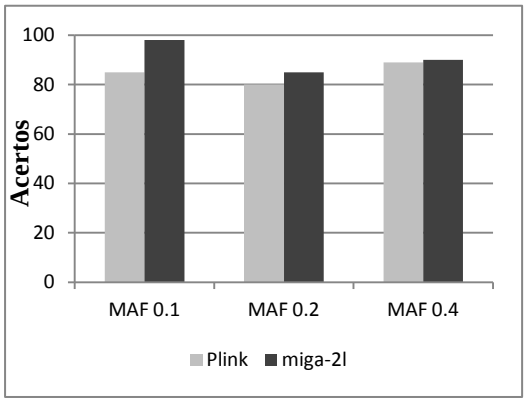
O conjunto de dados considerando loci com efeito principal foi gerado por Xai Wan et al. (2010). Nestas simulações, quatro modelos epistáticos foram considerados (Tabela 6.3). Cada modelo foi divido em três grupos segundo sua herdabilidade e seu MAF; cada grupo consta de 100 arquivos. Foi considerada uma prevalência da doença na população igual a 0.1. O modelo 1 é um modelo multiplicativo [MARCHINI et al, 2005]. O modelo 2 é um modelo epistático [NEUMAN & RICE, 1992], que foi usado para descrever doença lateral [LEVY J. and NAGYLAKI T., 1992] e a cor do suíno [LERNER 1968]. O modelo 3 é um modelo clássico epistático [FRANKEL & SCHORK, 1996; LI & REICH, 2000]. O modelo 4 é conhecido como o modelo XOR.

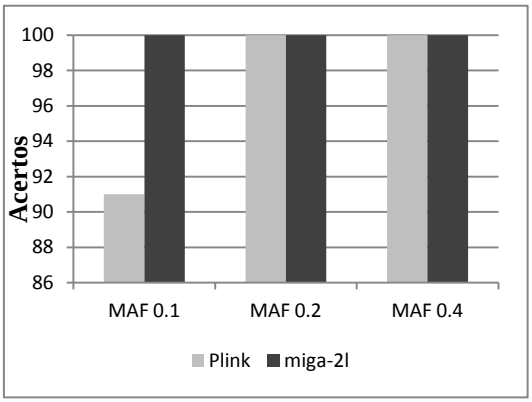
Nestas simulações, os valores escolhidos para herdabilidade foram h^2=0.03 para o modelo 1 e h^2 = 0,02 para modelos 2, 3 e 4. Todos os genótipos foram gerados baseados no principio de Hardy-Weinberg (HWE). Os valores de MAF considerados foram: 0.1, 0.2 e 0.4 para os quatro modelos. Com este esquema foram gerados 12 grupos de 100 arquivos de dados de genótipos cada um. Cada conjunto de dados contem 1000 SNPs e 1600 indivíduos considerando um desenho caso-controle balanceado (800 casos e 800 controles). Na Tabela 6.3 se apresentam os parâmetros utilizados na geração de cada conjunto de dados. Por conveniência a descrição da geração destes dados são listados no *Apêndice*.

Modelo	Conjunto	Herdabilidade	MAF
1	1	0.03	0.1
	2	0.03	0.2
	3	0.03	0.4
2	4	0.02	0.1
	5	0.02	0.2
	6	0.02	0.4
3	7	0.02	0.1
	8	0.02	0.2
	9	0.02	0.4
4	10	0.02	0.1
	11	0.02	0.2
	12	0.02	0.4

Tabela 6.3 Taxa de herdabilidade e MAF utilizada na simulação de 12 conjuntos de dados com efeito principal.

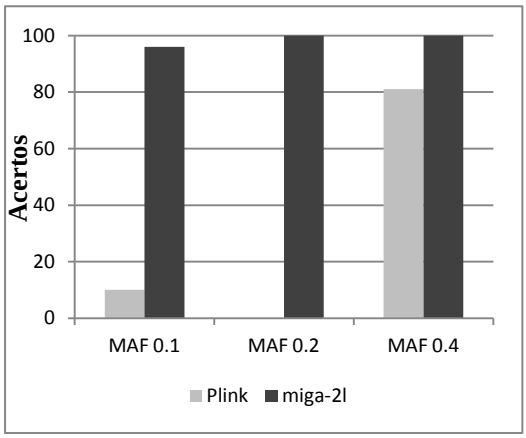
Cada arquivo de dados de genótipos possui dois SNPs funcionais que interagem com o efeito principal. O objetivo do experimento é que os algoritmos identifiquem esses SNPs funcionais. Assim, foi computado para cada grupo as vezes que cada algoritmo fez a identificação correta. As Figura s 6.2a e 6.2b exibem os resultados obtidos nos experimentos realizado com *Plink* e *MIGA-2L*. O eixo vertical indica a quantidade de acertos que é definido como a proporção dos 100 conjuntos de dados onde a interação de SNPs funcionais é identificada. A ausência de barra significa nenhum acerto.

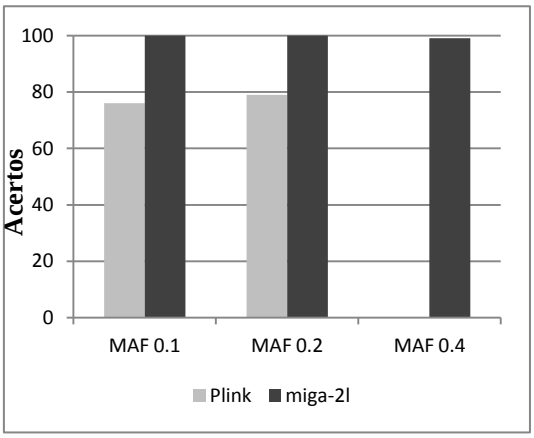




Modelo 1 (Multiplicativo): $h^2 = 0.03$ **Modelo 2 (epistasis)**: $h^2 = 0.02$

Figura 6.2a: Neste dois modelos *MIGA-2L* supera na porcentagem de acertos a *Plink*. Quando a frequência do alelo menor é pequena MAF= 0.1 *Plink* tem problemas para identificar os SNPs funcionais.





Modelo 3 (epistasis clássico): $h^2 = 0.02$ **Modelo 4 (XOR)**: $h^2 = 0.02$

Figura 6.2b: No modelo 3 pode se notar que *Plink* não tem nenhum acerto com MAF=0.2. Igualmente no modelo 4, com MAF=0.4. Fato que confirma uma forte dependência de *Plink* com a frequência alélica em modelos genéticos heterogêneos.

Em todos os modelos com efeito principal, *MIGA-2L* supera em numero de acertos a *Plink*. A opção de 'fast-epistasis' de Plink seleciona SNPs com efeito principal marginal de um único SNP, ignorando SNPs que poderiam ter um efeito de interação forte em conjunto com outros SNPs. Enquanto, *MIGA-2L* consegue capturar a interação SNP-SNP sem ser confundido pelo efeito principal de um único lócus.

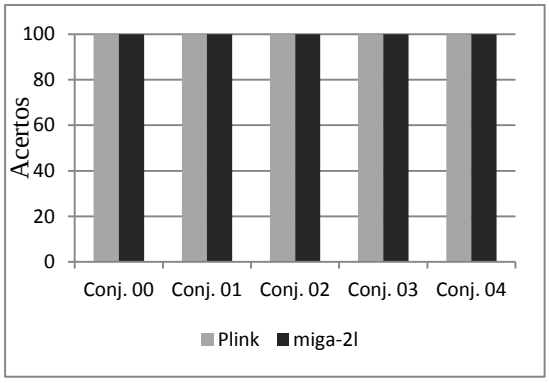
Conjunto de dados com loci sem efeito principal

Para abordar um conjunto de dados simulados sem efeito principal, utilizamos os dados gerados por Velez et al. (2007). Estas simulações contemplam 70 modelos epistáticos, construídos a partir de funções de penetrância e diferentes combinações nas taxas de herdabilidade e MAF (Apêndice). Estes modelos são listados na Tabela 6.4, onde cada conjunto simulado esta composto por um total de 100 conjuntos de dados de genótipos. Cada conjunto de dado contêm 1000 SNPs e 1600 indivíduos, considerando um desenho balanceado.

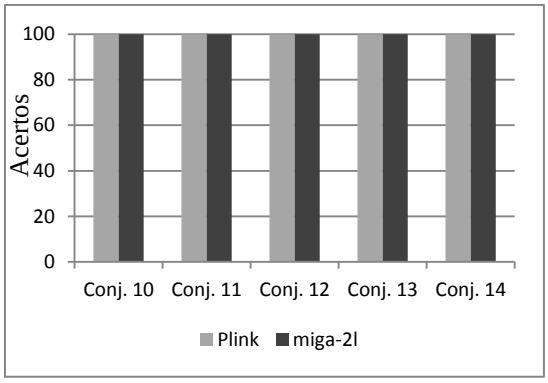
Modelo	Conjuntos	Herdabilidade	MAF
Modelo 1	00-04	0.4	0.2
Modelo 2	05-09	0.4	0.4
Modelo 3	10-14	0.3	0.2
Modelo 4	15-19	0.3	0.4
Modelo 5	20-24	0.2	0.2
Modelo 6	25-29	0.2	0.4
Modelo 7	30-34	0.1	0.2
Modelo 8	35-39	0.1	0.4
Modelo 9	40-44	0.05	0.2
Modelo 10	45-49	0.05	0.4
Modelo 11	50-54	0.025	0.2
Modelo 12	55-59	0.025	0.4
Modelo 13	60-64	0.01	0.2
Modelo 14	65-69	0.01	0.4

Tabela 6.4 Taxa de herdabilidade e MAF utilizada na simulação de 70 conjuntos de dados sem efeito principal.

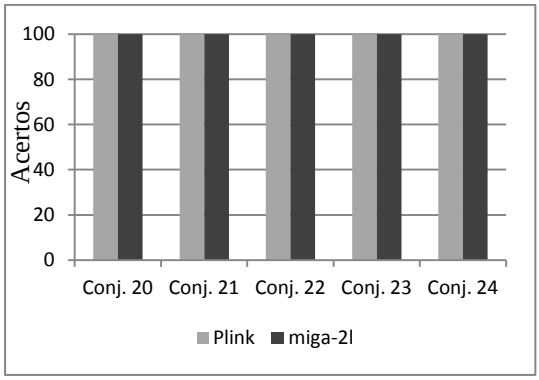
Neste experimento, foram usados todos os 70 modelos puros epistáticos sem efeito principal para comparar o desempenho de *MIGA-2L* e *Plink*. A herdabilidade h^2 controla a variação fenotípica de estes 70 modelos, atribuindo valores desde 0.01 até 0.4. O MAF varia de 0.2 a 0.4. Os resultados comparativos para os 70 modelos são mostrados nas Figuras 6.3a até 6.3e. Também, como no experimento anterior, pode se observar uma porcentagem superior de acertos do *MIGA-2L* sobre *Plink*. Os detalhes dos parâmetros para a geração destes 70 modelos epistáticos são listados no Apêndice deste documento de tese.



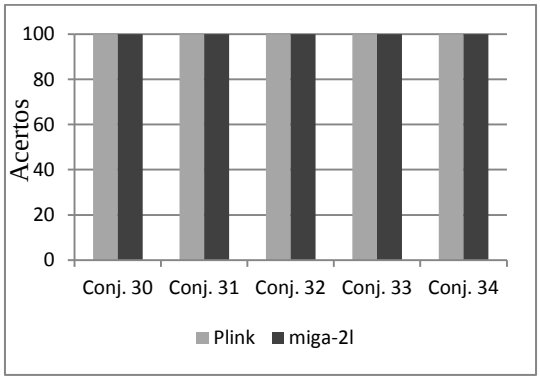
Modelo 1: h^2 0.4 e MAF 0.2



Modelo 3: h^2 0.3 e MAF 0.2

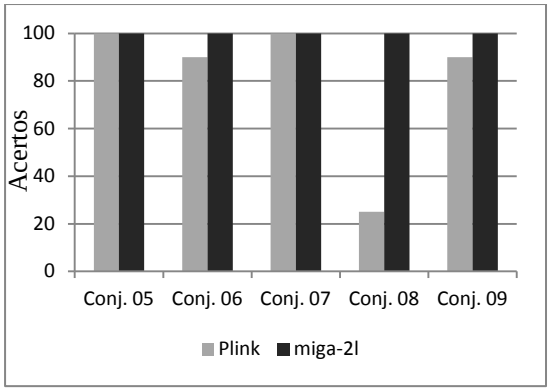


Modelo 5: h^2 0.2 e MAF 0.2

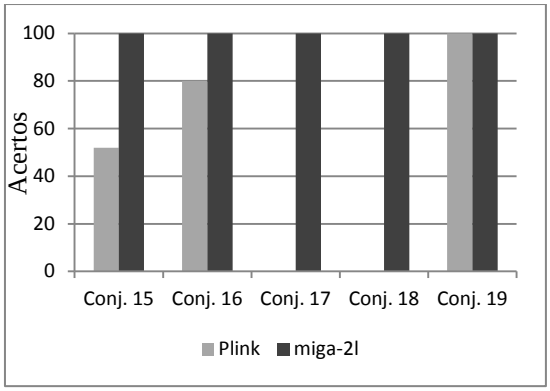


Modelo 7: h^2 0.1 e MAF 0.2

Figura 6.3a: Os gráficos mostram os resultados obtidos quando a frequência do alelo menor MAF=0.2. Os dois algoritmos conseguem obter 100% de acertos, ou seja conseguem descobrir os SNPs funcionais que interagem nestes modelos genéticos.



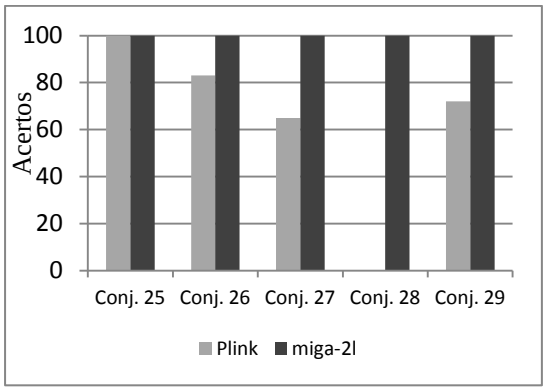
Modelo 2: h^2 0.4 e MAF 0.4



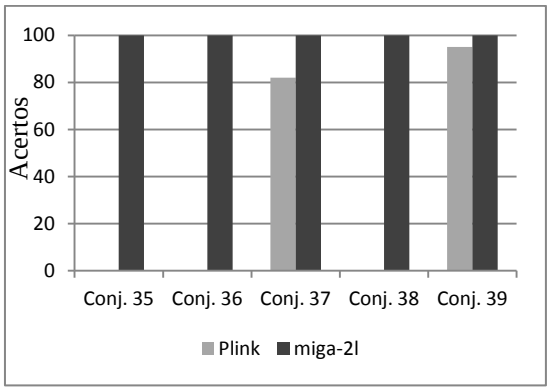
Modelo 4: h^2 0.3 e MAF 0.4

Figura 6.3b: Taxa de acerto obtido nos Modelos 2 e 4 para MAF=0.4. *MIGA-2L* supera a *Plink* no numero de acertos.

No modelo 2 *Plink* tem uma porcentagem de acerto superior a 90% com exceção do conjunto 8. No modelo 4 *Plink* não tem nenhum acerto nos conjuntos 17 e 18. Isto deve se que as frequências alélicas destes modelos genéticos possuem uma alta heterogeneidade genética fazendo que *Plink* ignore aqueles SNPs com baixo efeito marginal individual. As frequências alélicas são estimadas considerando diferentes valores de penetrância em combinação com o MAF e a herdabilidade h^2.

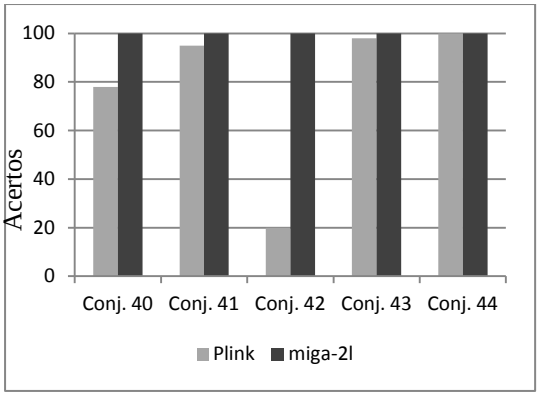


Modelo 6: h^2 0.2 e MAF 0.4

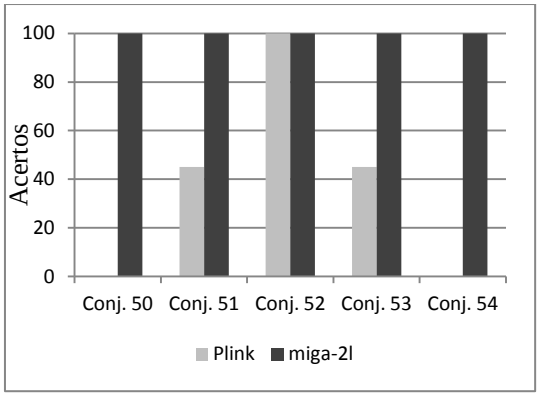


Modelo 8: h^2 0.1 e MAF 0.4

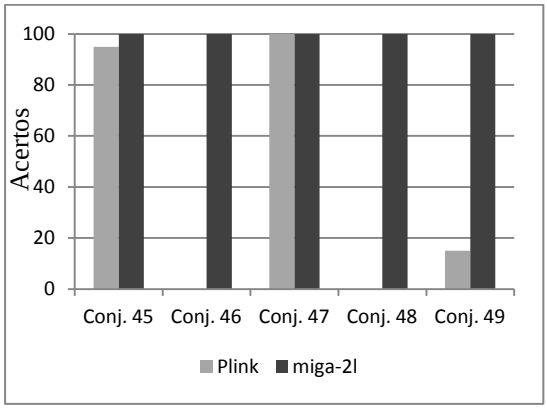
Figura 6.3c: Nos modelos 6 e 8 *MIGA-2L* continua com 100% de acertos enquanto *Plink* apresenta nenhum acerto em alguns conjuntos de dados. No modelo 8 *Plink* diminui sua taxa de acerto porque fica afetado com a diminuição da taxa de herdabilidade.



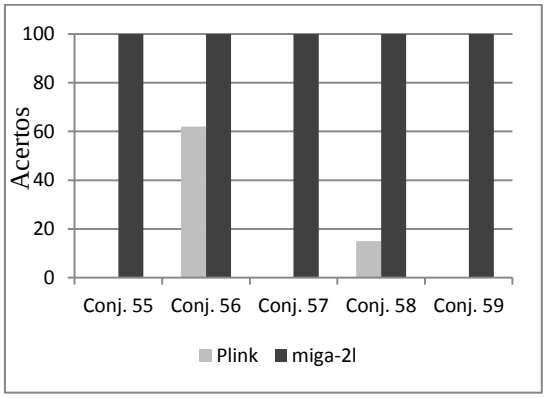
Modelo 9: h^2 0.05 e MAF 0.2



Modelo 11: h^2 0.025 e MAF 0.2

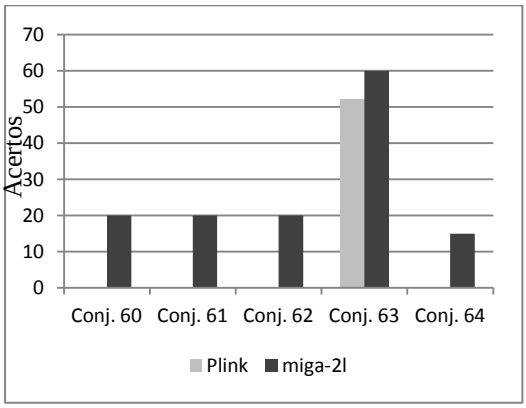


Modelo 10: h^2 0.05 e MAF 0.4

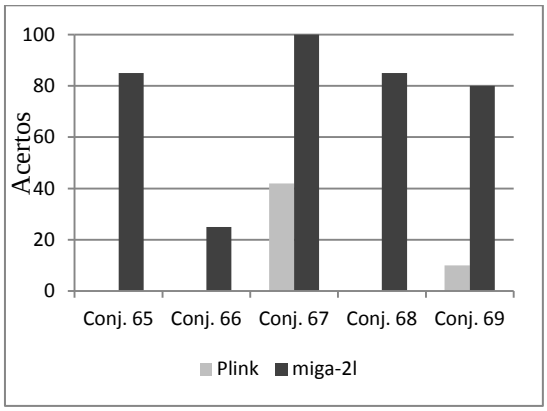


Modelo 12: h^2 0.025 e MAF 0.4

Figura 6.3d: Estes modelos seguem o padrão dos modelos anteriores mostrados na Figura 6.3c.



Modelo 13: h^2 0.01 e MAF 0.2



Modelo 14: h^2 0.01 e MAF 0.4

Figura 6.3e: Nestes dois modelos *MIGA-2L* apresenta uma diminuição na porcentagem de acerto devido a influencia da taxa de herdabilidade h^2.

Plink e *MIGA-2L* apresentam problemas para identificar os SNPs funcionais quando a taxa de herdabilidade diminui e a frequência de alelo menor aumenta MAF=0.4. Como pode se ver nos modelos 13 e 14, mostrados na Figura 6.3e, onde conjuntos de dados têm nenhum acerto ou uma porcentagem muito baixa de acertos.

Isto deve-se a que o coeficiente h^2 tem influenciado a taxa de penetrância em valores muito baixos, dificultando a identificação da proporção da variação no fenótipo que é atribuível ao genótipo. Nestes casos de herdabilidade e penetrância baixas, fatores ambientais deveriam ser considerados para ajudar uma melhor identificação dos SNPs de risco.

Relação entre o modulo 'fast-epistasis' de *Plink* e *MIGA-2L*

O algoritmo *MIGA-2L* abrange um espaço de modelos de dados maior que *Plink* já que não precisa assumir um modelo genético especifico. A diferencia principal entre *MIGA-2L* e *Plink* é a forma em que eles avaliam o efeito da interação de SNPs com a doença. Plink realiza uma avaliação alelo x alelo, onde três categorias de genótipos são colapsadas em duas categorias enquanto *MIGA-2L* realiza a avaliação genótipo x genótipo. Figura 6.4

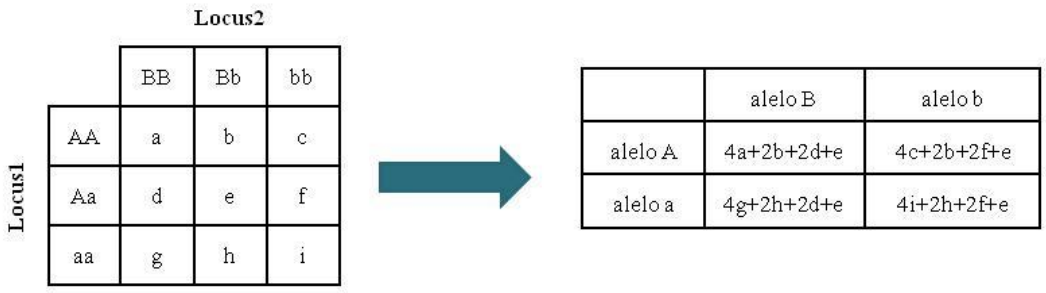


Figura 6.4: *Plink* assume que os dados seguem um modelo aditivo para dessa forma colapsar as 3 categorias de genótipos em uma Tabela de contingencia de alelos 2x2.

6.2.2 EXPERIMENTOS SOBRE DADOS REAIS: DIABETES TIPO I

Nesta seção são apresentados os resultados dos experimentos realizados com o conjunto de dados de diabetes tipo 1 (T1D). Este conjunto de dados foi proporcionado pela Wellcome Trust Case Control Consortium (WTCCC). Eles foram gerados usando um chip affymetrix de 500K para coletar os genótipos de 500,000 marcadores.

6.2.2.1 Pré-processamento do conjunto de dados de diabetes tipo 1 (T1D)

O conjunto de dados originado do projeto WTCCC1, contém 2,000 amostras de pacientes com T1D e 3,004 amostras de controles, dos quais 1,504 foram coletadas de uma coorte de britânicos nascidos em 1958 e 1,500 controles adicionais cujas amostras procedem do Serviço nacional de sangue do Reino Unido.

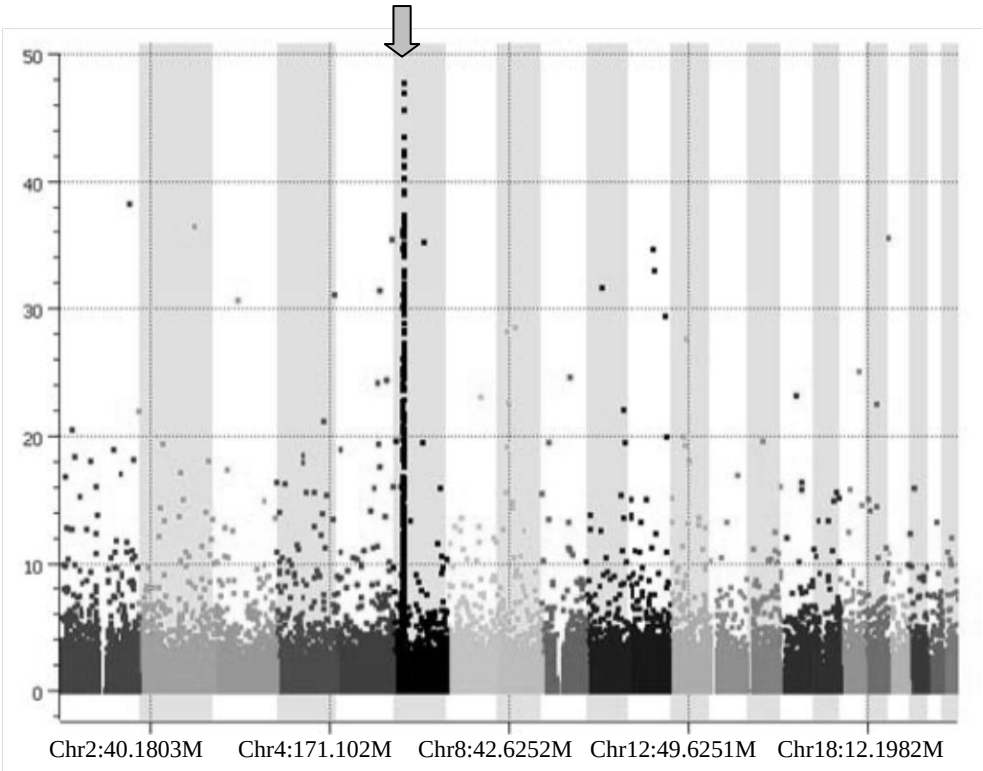


Figura 6.5: Gráfico de Manhattan correspondente a T1D. A seta assinala a região localizada no cromossomo 6.

Este conjunto de dados foi submetido ao controle de qualidade descrito no capítulo anterior. Uma análise inicial dos dados foi feita utilizando o gráfico de Manhattan (Figura 6.5). Este gráfico fornece uma forma de visualizar cada SNPs através dos cromossomos humanos posicionando-o como um ponto no gráfico, onde a altura corresponde à força de associação desse SNP com a doença expressada pelo $\log_{10}$ do p-valor obtido utilizando a estatística Chi-quadrado. Os pontos localizados acima de certo limiar, por exemplo um p-valor > 10E-5, poderiam ser considerados como altamente suspeitos, ou seja eles podem ser SNPs apresentando associação com a doença estudada ou poderiam ser SNPs com erros de genotipagem. No gráfico de Manhatan da Figura 6.5 pode-se observar um pico que esta localizado no cromossomo 6. Este cromossomo é altamente polimórfico e, por essa razão, foi estudado com mais detalhe neste trabalho.

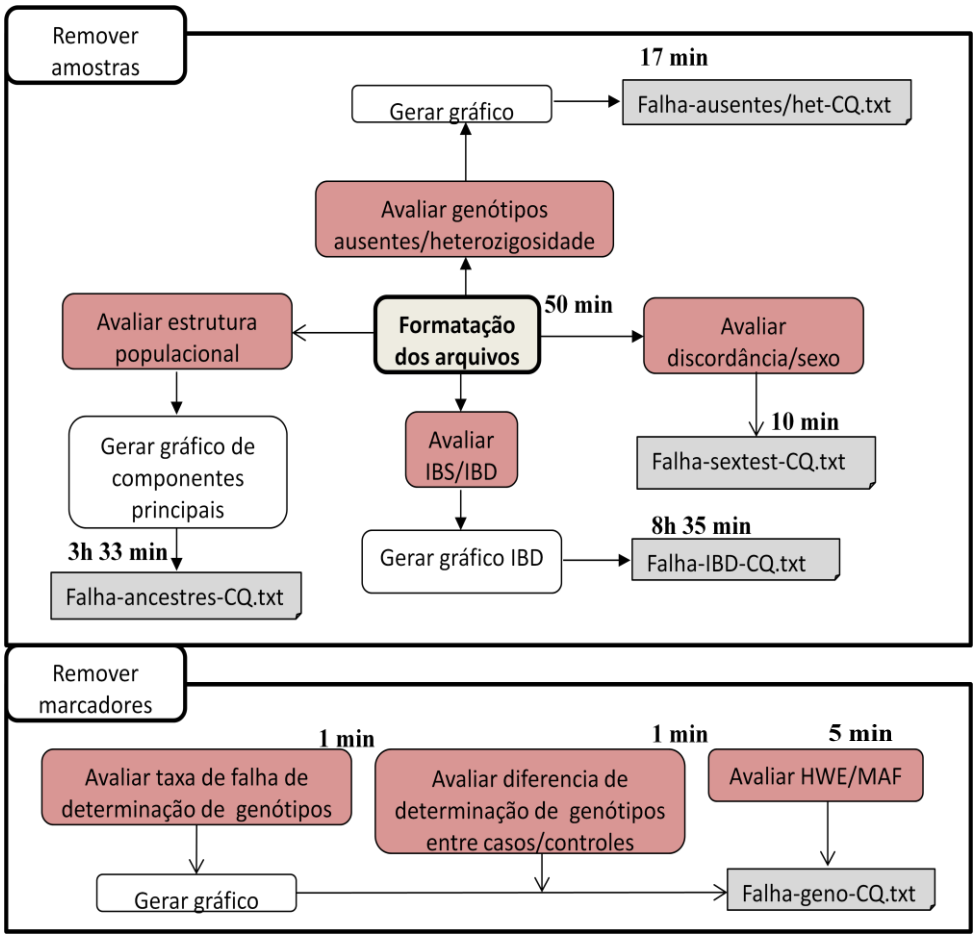


Figura 6.6: Tempo computacional obtido em cada tarefa do pré-processam ento dos dados. Os tempos foram computados utilizando o motor de Workflow Quiron.

Como foi descrito no capítulo anterior, o controle de qualidade foi realizado primeiro avaliando a qualidade das amostras e posteriormente a qualidade dos marcadores. Nesta análise a quantidade de SNPs que passaram o controle de qualidade foi de 352,538 SNPs dos 500,000 SNPs originais. A Figura 6.6 apresenta o workflow conceitual a que foi submetido utilizando o motor de workflow Quiron, com os tempos computacionais por cada tarefa, assim como o tempo total empregado para remoção das amostras e para remoção dos marcadores que não passaram no controle de qualidade.

As Figura s 6.7a, 6.7b e 6.7c mostram alguns gráficos gerados durante o controle de qualidade das amostras. Estes gráficos servem para ajudar na escolha de um certo valor (ou limiar) para filtrar as amostras com falha na qualidade.

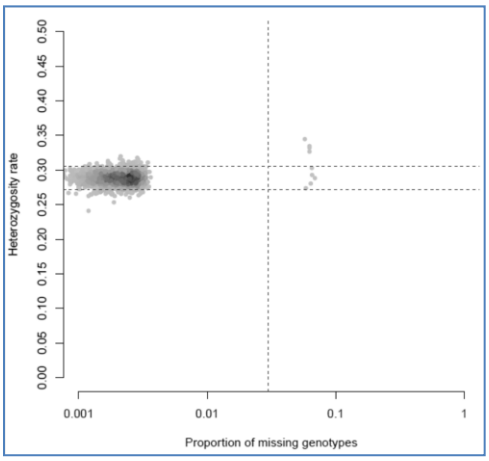


Figura 6.7a: Relação entre a taxa de heterozigosidade e a proporção de genótipos ausentes no conjunto de dados de T1D

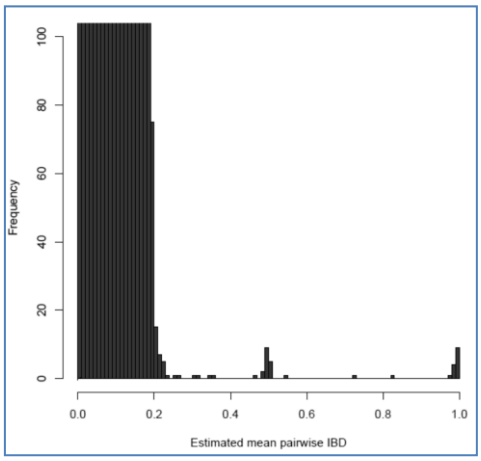


Figura 6.7b: Indivíduos duplicados e relacionados no conjunto de T1D. Note que o IBD >0.25 indica os indivíduos a serem removidos da amostra.

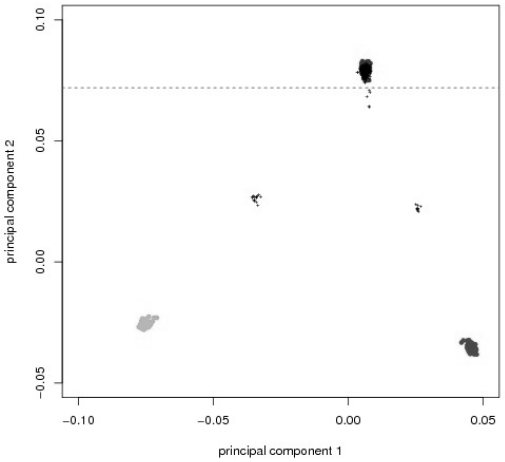


Figura 6.7c Resultados da análise de componentes principais para as amostras casos e controles. Esta análise mostra a diferencia na frequência dos alelos nos milhares de marcadores, indicando diferenças étnicas.

Tempo Computacional

Minimizar o viés potencial e erro no resultado de GWAS depende da qualidade global dos dados. Neste contexto, o grande número de SNPs genotipados em estudos GWA caso-controle constitui um grande desafio computacional devido aos critérios envolvidos no controle de qualidade dos dados. O que torna o processo operacionalmente intensivo e desafiador. A Tabela 6.5 apresenta os tempos empregados por cada atividade.

Nome da tarefa	**Tempo**
Avaliação das amostras	
Formatar arquivo a binário/ avaliar discordância segundo sexo	60 min
Avaliar taxa genótipos ausentes vs. heterozigosidade	17 min
Avaliar indivíduos duplicados e relacionados	8 hs 35 min
Avaliar estrutura populacional	2 hs 15 min
Remover amostras falhadas	1 min
Avaliação dos marcadores	
Avaliar taxa de genótipos ausentes	1 min
Avaliar taxa de genótipos ausentes entre casos e controles	1 min
Avaliar HWE e MAF	5 min
Remover marcadores que não passaram a avaliação	1 min
Tempo total empregado usando Quiron	8 hs 35 min

Tabela 6.5 Tempo computacional empregado no pré processamento do conjunto de dados de Diabetes Tipo 1.

Análise do cromossomo 6 de T1D.

O cromossomo 6 é altamente polimórfico, contendo a região MHC com uma alta densidade de genes. Estudos prévios reportam forte associação de um único lócus entre genes pertencentes a região MHC e T1D [Noble and Erlich, 2012], tal como HLA-DQB1 e HLA-DRB1. Porém ainda não está claro qual e como muitos loci dentro da região MHC e fora dela determinam susceptibilidade a T1D.

Interações sem efeitos principais significantes podem prover informação adicional para ajudar a entender *loci* associados a doenças. Em muitos estudos SNPs envolvidos neste tipo de interações são frequentemente excluídos da análise. Nesta seção são mostrados as interações descobertas pelo algoritmo *MIGA-2L* no cromossomo 6 de T1D. Posteriormente, uma inspeção dessas interações foi realizada para entender sua funcionalidade na doença.

O tamanho da amostra analisada foi de 35,865 SNPs sobre um total de 4,612 indivíduos, sendo 2,646 controles e 1,966 casos. A Figura 6.8 mostra o gráfico de Manhattam para o cromossomo 6, onde se pode observar alguns picos de associação de 1 único lócus.

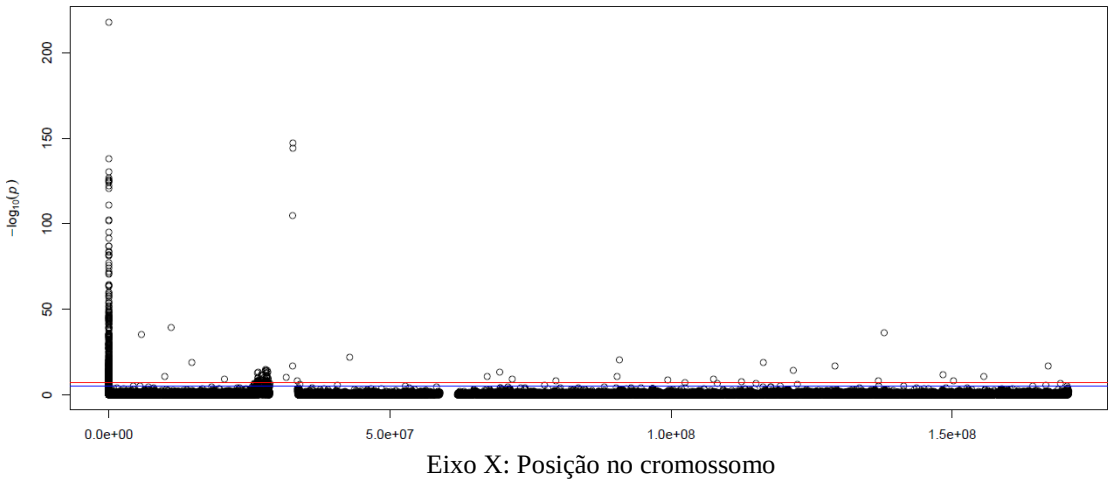


Figura 6.8: Gráfico de Manhattan do cromossomo 6 da amostra T1D do projeto WTCCC1. Cada ponto é um SNP em uma localização do cromossomo. O eixo Y representa a força de associação de cada SNP com a doença obtido com um teste −log10 de Chi-quadrado/P value. As associações mais fortes formam picos e SNPs correlacionados mostram o mesmo sinal. O gráfico de Manhattan com pontos acima de certo limiar (ex. a linha solida indicando um P-valor < $5x10^{-7}$) deveria ser visto como altamente suspeito.

Outro tipo de gráfico muito interessante é Q-Q plot, este gráfico mostra no eixo X a distribuição esperada de genótipos usando o teste estatístico Chi-quadrado, comparado a sua distribuição observada no eixo Y. A Figura 6.8a apresenta o Q-Q plot para os dados pertencentes ao cromossomo 6 de T1D. Note que a linha tracejada serve para indicar a tendência que os dados deveriam seguir. Desvios da linha podem sugerir associação desses dados com a doença estudada.

Na Figura 6.9a se observam muitos SNPs que não seguem a linha tracejada. Isto deve-se a que a região HLA localizado no cromossomo 6 é uma significante e grande região que está impactando no estudo de associação. A Figura 6.9b mostra um melhor ajuste dos dados, onde os SNPs correspondentes à região HLA foram removidos para constatar o impacto que essa região está causando no estudo de

associação. Com isto, pode-se estimar que resultados interessantes se encontram escondidos nos dados e que certamente uma análise gráfico pode estar excluindo interações significantes como aquelas sem efeito principal.

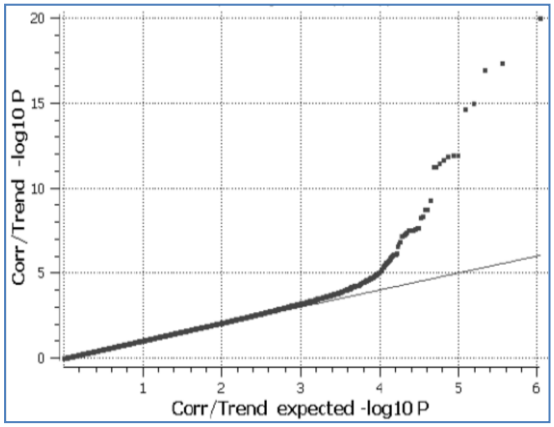


Figura 6.9a: Gráfico Q-Q dos valores observados e valores esperados da computação da associação utilizando um teste estatístico chi-quadrado.

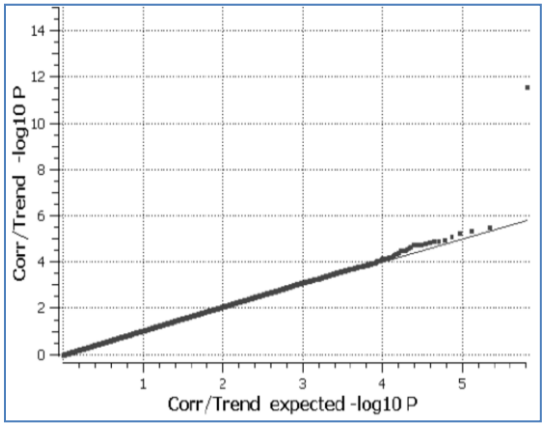


Figura 6.9b: Gráfico Q-Q obtido com dados de genótipos de SNPs não localizados na região HLA.

6.2.2.2 Execução do *MIGA-2L* para descoberta de interação SNP-SNP.

O conjunto de dados do cromossomo 6 de T1D foi submetido para ser analisado pelo algoritmo *MIGA-2L* e investigar as interações SNP-SNP. Previamente foi realizada a partição do conjunto de dados, utilizando a estratégia de validação cruzada com k=10. Gerando desta forma 10 conjuntos balanceados de dados de treino e 10 conjuntos balanceados de dados de teste.

Os resultados encontrados pelo algoritmo *MIGA-2L* são mostrados na Tabela 6.6. A Tabela mostra as interações identificadas que tiveram o maior nível de significância estatística aplicando o teste estatístico Z score. Todas as interações têm um p-valor < 0.0001. Estes SNPs apresentam uma associação com a doença estudada nos indivíduos portadores dessas interações. Uma avaliação mais informativa pode ser feita efetuando uma análise detalhada das variantes dos haplótipos que compõem as interações descobertas. Desta forma, pode-se estimar quais haplótipos incrementam o risco de desenvolvimento da doença nos indivíduos portadores dessas interações comparados com os não portadores. Esta análise, por cada interação SNP-SNP, e incluindo seus 9 haplótipos possíveis, é mostrada na Tabela 6.7. Esta Tabela apresenta

o computo das odds ratio (OR) e o intervalo de confiança computado sobre cada haplótipo das interações listadas na Tabela 6.6.

SNP - Locus1	SNP - Locus2	Z score
rs2147653	rs6455867	45.742
rs2147653	rs6932546	46.133
rs2147653	rs2982908	44.841
rs2147653	rs6937679	43.611
rs2147653	rs2206256	49.904

Tabela 6.6 Interações encontradas pelo *MIGA-2L* no cromossomo 6 em T1D. A primeira e segunda coluna apresentam os SNPs identificados segundo a referencia do banco de dados "dbSNP" cujo formato é SNP ID número ("rs#"; "refSNP cluster").

	rs2147653-rs6937679		rs2147653-rs6455867		rs2147653-rs6932546		rs2147653-rs2982908		rs2147653-rs2206256	
Hap	OR	IC(95%)	OR	IC(95%)	OR	IC(95%)	OR	IC(95%)	OR	IC(95%)
AABB	1.07	0.94-1.22	1.06	0.93-1.21	1.06	0.93-1.21	1.06	0.93-1.21	1.05	0.92-1.20
AABb	0.78	0.59-1.03	0.79	0.6-1.04	0.79	0.6-1.04	0.77	0.58-1.02	0.77	0.61-0.97
AAbb	0.5	0.16-1.56	0.5	0.16-1.56	0.5	0.16-1.56	0.58	0.21-1.63	0.53	0.22-1.25
AaBB	0.6	0.17-1.73	Na	Na	Na	Na	Na	Na	Na	Na
AaBb	0.98	0.87-1.09	0.98	0.87-1.09	0.98	0.87-1.09	0.98	0.87-1.10	1	0.89-1.12
Aabb	1.04	0.8-1.35	1.03	0.79-1.35	1.02	0.78-1.34	1.02	0.78-1.34	1.07	0.84-1.36
aaBB	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na
aaBb	0.05	0.01-0.21	0.03	0.00-0.19	0.03	0.00-0.19	0.03	0.00-0.19	0.03	0.00-0.18
aabb	**1.17**	**1.01-1.34**	**1.17**	**1.02-1.34**	**1.17**	**1.02-1.35**	**1.17**	**1.02-1.35**	**1.17**	**1.01-1.35**

Tabela 6.7 Valores das Odds-ratio e seu intervalo de confiança de cada haplótipo das interações SNP-SNP identificadas no cromossomo 6 para T1D. O haplótipo recessivo aabb (resaltado na Tabela) apresenta uma OR > 1, indicando uma associação positiva com T1D. Os alelos recessivos são as formas mais graves de mutação já que perderam o sítio de reconhecimento que seria utilizado na transcrição de proteínas.

Como era esperado, a maioria dos estudos GWA para fenótipos discretos, apresentam valores de OR detectáveis entre 1,1 e 1,3 [Goldstein D.B, 2009]. Assim, na Tabela 6.7 o haplótipo recessivo mostra uma OR de 1,17 que pode-ser interpretado como que os indivíduos portadores de este haplótipo apresentam um incremento no

risco de desenvolvimento da doença entre 15% e 20% comparados com os não portadores.

Tempo computacional

O tempo total empregado por *MIGA-2L* na investigação de interação SNP-SNP foi de 55 minutos utilizando o workflow Quiron. O programa *Plink* demorou 42 horas para investigar a interação de 2 loci utilizando a opção - -epistasis (busca exaustiva) e 12 horas utilizando a opção - -fast-epistasis (busca aleatória), sobre o mesmo conjunto de dados.

6.3 ANÁLISE FUNCIONAL

O algoritmo *MIGA-2L* revelou que existem SNPs associados significativamente com T1D no cromossomo 6. Os SNPs da Tabela 6.6 pertencem à região do gene PACRG2 cuja relevância fisiológica ou patofisiológica é desconhecida. Este é um gene co-regulador da Parkina, uma proteína produzida pelo gene PARK2. O promotor do gene da Parkina é um promotor bidireccional, regulando a transcrição da Parkina e do gene upstream antissense [Andrew B. West et al. 2003] que possui 5 exões e um comprimento total de 0.6 Mb.

A Figura 6.10 apresenta a via de interação gênica envolvida com o gene PACRG e PARK2. Outros genes envolvidos com PACRG são o SYT11 gene regulador da insulina e o SLC11A1, o qual apresenta associação com T1D.

Mutações no gene da Parkina (PARK2) surgem com elevada frequência, sendo observadas em cerca de 50% dos casos de doença de Parkinson hereditária [Betarbet et al. 2005]. Quando o gene se encontra mutado manifesta-se um fenótipo autossômico recessivo juvenil da doença de Parkinson. Além disso, alguns estudos apontaram uma associação do PARK2 e SLC11A1 (anteriormente chamado NRAMP1) com a hanseníase [SOUZA & PEREIRA, 2007].

Neste contexto, pode-se interpretar que o gene PACRG contém um sitio de ligação para um fator regulador de transcrição de proteínas relacionadas a T1D e que leva a pessoas portadoras do haplótipo recessivo à uma falha na transcrição devido à perda desse sítio de reconhecimento. Esse fato aumenta o risco no desenvolvimento

de diabetes tipo 1 assim como de doenças como Parkinson e hanseníase. Tem que ser mencionado aqui que fatores ambientais também influenciam e tem que ser considerados, mas estes achados podem ser referidos adicionalmente como uma explicação alternativa para a etiologia da T1D na população do Reino Unido.

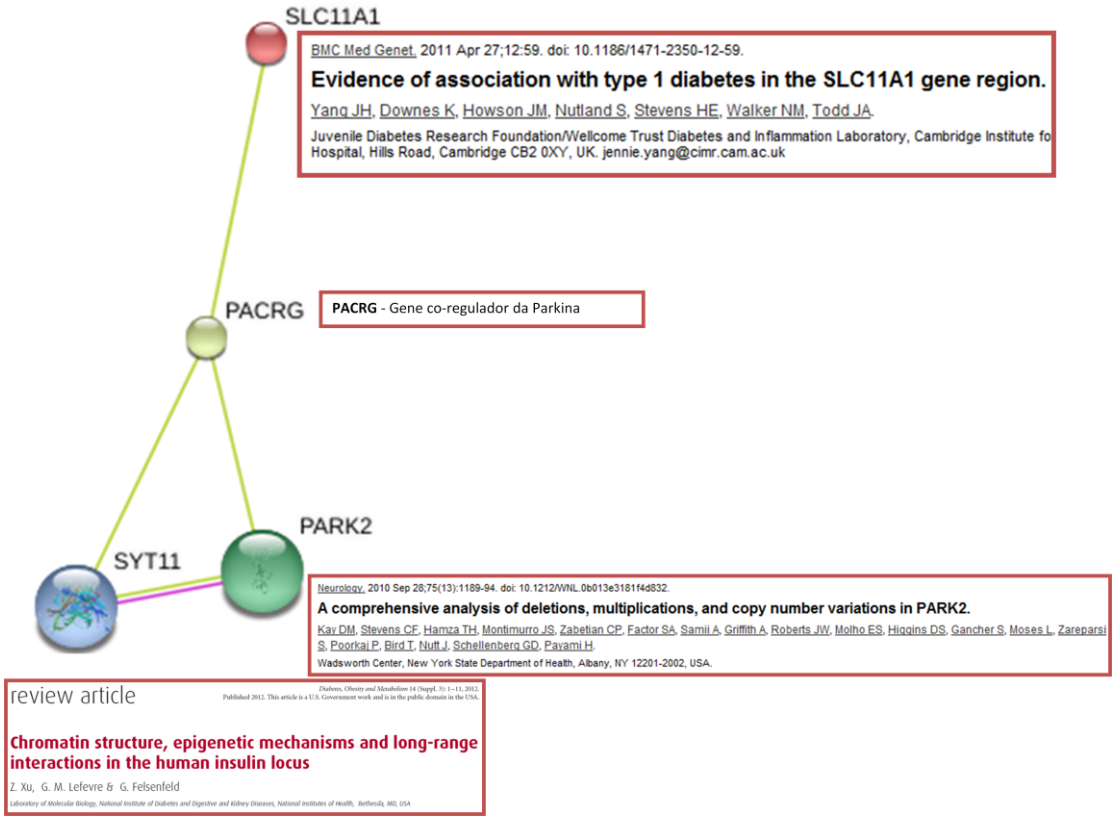


Figura 6.10 Via de interação gênica do gene PACRG o qual interage genes como SYT11, PARK2 e SLC11A1. Estudos mostram que estes genes estão associados a T1D e outras doenças como Parkinson e hanseníase [Fonte: Raquel Barbosa – INCA/Rio de Janeiro].

6.4 Experimentos sobre 4 conjuntos de dados da WTCCC1 no Cromossomo 6

A metodologia proposta neste trabalho de tese também foi aplicada para análise de outros 4 conjuntos de dados do projeto da WTCCC1. Este projeto inclui além da T1D, 4 doenças humanas comuns, tais como, doença da artéria coronária (CAD), hipertensão (HT), diabetes tipo 2 (T2D) e desordem bipolar (BD). O pré-processamento dos dados foi realizado da mesma forma como na análise de diabetes

tipo1 (T1D). A Tabela 6.8 mostra o número de SNPs que restaram depois do pré-processamento dos dados sobre o cromossomo 6.

As Tabelas 6.9a até 6.9d listam as associações SNP-SNP mais significantes encontradas por *MIGA-2L* no cromossomo 6. Estes resultados foram obtidos considerando p-valores < 0.0001 utilizando o teste estatístico Z score.

Diabetes tipo 2 (T2D)	**Hipertensão (HT)**	**Desordem bipolar (BD)**	**Doença da artéria coronária (CAD)**
27,396 SNPs	27,387 SNPs	27,392	27,395

Tabela 6.8 Quantidade de marcadores considerados no GWAS para T2D, HT, BD e CAD.

6.4.1 Doença de artéria coronária (CAD)

A doença arterial coronariana é o estreitamento dos pequenos vasos sanguíneos que fornecem sangue e oxigênio ao coração. Ela também é chamada de doença cardíaca coronária.

O algoritmo *MIGA-2L* encontrou associação de CAD com os SNPs mostrados na Tabela 6.9a. Estes SNPs pertencem aos genes OR2H1, OR2H2, RPS17P1, UBD, PPP1R11, ZNRD1-AS1, TRIM15, TRIM31, TRIM40.

SNP1	**SNP2**
rs734960	rs9504552
rs9504552	rs7739310
rs2517595	rs2523995
rs734960	rs539703
rs2394401	rs2523995

Tabela 6.9a: Interações SNP-SNP identificadas em CAD.

6.4.2 Hipertensão (HT)

A Hipertensão arterial é uma doença crônica determinada por elevados níveis de pressão sanguínea nas artérias, o que faz com que o coração tenha que exercer um esforço maior do que o normal para fazer circular o sangue através dos vasos sanguíneos. Os SNPs achados pelo *MIGA-2L* estão localizados nos genes LOC441179 e LOC100422263.

SNP1	SNP2
rs554158	rs6454756
rs554158	rs6934594
rs554158	rs3798867
rs554158	rs9456734
rs554158	rs9493450
rs554158	rs211218

Tabela 6.9b: Interações SNP-SNP identificadas em HT.

6.4.3 Desordem bipolar (BD)

O desordem bipolar é uma forma de transtorno de humor caracterizado pela variação extrema do humor entre uma fase maníaca ou hipomaníaca1 , que são estágios diferentes pela gradação dos seus sintomas, hiperatividade física e mental, e uma fase de depressão, inibição, lentidão para conceber e realizar ideias, e ansiedade ou tristeza.

Os SNPs associados com BD descobertos pelo *MIGA-2L* se encontram nas regiões dos genes HLA-F e FOXQ1.

SNP1	SNP2
rs2747436	rs29254
rs2438083	rs977674
rs2438083	rs977673
rs1729549	rs1190806

Tabela 6.9c: Interações SNP-SNP identificadas em BD.

6.4.4 Diabetes Tipo 2 (TD2)

A diabetes mellitus tipo 2 é um distúrbio metabólico caracterizado pelo elevado nível de glicose no sangue no âmbito da resistência à insulina e pela insuficiência relativa de insulina. Distingue-se da diabetes mellitus tipo 1, na qual se verifica a deficiência completa de insulina devido à destruição dos ilhéus de Langerhans no pâncreas. Os SNPs associados a TD2 pertencem às regiões dos genes RBBBP4P3, CLVS2, ATXN1 e HIVEP2.

SNP1	**SNP2**
rs11758386	rs683831
rs311210	rs683831
rs4314501	rs683831
rs7756217	rs683831
rs236949	rs683831

Tabela 6.9d: Interações SNP-SNP identificadas em TD2.

6.5 Síntese do capítulo

A interpretação biológica de dados GWA é um grande desafio. Associações são frequentemente encontradas em SNPs localizados em genes desertos ou a uma distância significativa de um gene conhecido. Uma estratégia frequentemente utilizada por geneticistas é restringir a análise a SNPs localizados dentro de 10kb de um gene conhecido. Porém, esta não é uma boa prática já que uma fração importante do genoma está sendo ignorado. Essa estratégia facilita a interpretação biológica e ajuda a realizar um enfoque sobre o que o cientista entende melhor.

Qualquer que seja a estratégia empregada pelo cientista em GWAS, este capítulo foi apresentado para demonstrar que a metodologia proposta neste trabalho de tese fornece ferramentas válidas e rápidas para a análise de dados em escala genômica. A metodologia foi testada tanto com dados simulados correspondentes a 82 modelos diferentes de doença como também com dados reais de 5 doenças comuns.

CAPÍTULO 7: CONCLUSÃO E TRABALHOS FUTUROS

O problema GWAS consiste principalmente em descobrir quais são os marcadores genéticos que são relevantes em relação com a doença estudada, utilizando um conjunto de dados de genótipos de indivíduos em escala genômica. Na busca de uma metodologia eficiente para realizar esta análise foi encontrado que os métodos de data mining mostram-se adequados para o tratamento destes dados já que as etapas mais importantes e informativas são as de pre-processamento e seleção.

No âmbito que este trabalho de tese foi desenvolvido, visando oferecer uma metodologia livre de erros metodológicos e fornecer de forma rápida e simples a obtenção dos resultados, que são apresentados com medidas de desempenho computacionais tanto como epidemiológicas, demonstrando sua aplicabilidade em dados reais como a identificação de padrões que distinguem doenças como diabetes tipo 1.

7.1 Contribuição da tese

Esta proposta de tese teve como enfoque os estudos GWA. Nesse contexto, a maior contribuição deste trabalho é a apresentação de uma metodologia para fornecer de forma simples e rápida a análise de dados de genótipos em escala genômica, abrangendo desde o controle de qualidade dos dados brutos, vindos da genotipagem, seguindo pela análise de associação, até chegar à classificação do risco dos haplótipos correspondentes aos SNPs selecionados. Foi demonstrado no capítulo anterior que a metodologia proposta é aplicável para a análise de conjuntos de dados reais de genótipos (SNPs) em escala genômica assim como para a análise de vários modelos epistáticos.

Estudos GWA tem vários problemas e limitações que podem ser atendidos por meio do controle de qualidade adequado e bom desenho do estudo. A obtenção dos dados de genótipos de tamanho suficiente para análise deixou de ser um problema graças às tecnologias de genotipagem cada vez mais precisas e accessíveis que permitem sequenciar o genoma completo de vários indivíduos. Por isso, uma vez superadas questões como um desenho de estudo bem definido de grupos de casos e controles e um tamanho de amostra suficiente, o principal desafio é o tratamento desses dados de alta dimensionalidade.

A proposta apresentada neste trabalho toma proveito das vantagens de um motor de workflow para auxiliar em todo o processo de GWAS. Sua utilização fornece integração, sequencialidade e interação com o usuário fornecendo informação necessária para facilitar a tomada de decisões. Cada etapa da metodologia toma especial cuidado de forma a evitar ou minimizar os erros que possam enviesar os resultados do estudo. O primeiro passo da metodologia considera múltiplos critérios que envolvem certa complexidade. Estes critérios foram abordados utilizando uma estratégia de otimização para evitar a eliminação de dados de genótipos e/ou amostras que possam ser relevantes para a identificação da associação. Estes passos críticos são fundamentais para o sucesso de um estudo de caso-controle e são necessários antes do teste de associação. Além disso, os programas e algoritmos selecionados para este primeiro passo são amigáveis ao usuário, amplamente difundidos na comunidade científica e computacionalmente eficientes. O processo de controle de qualidade dos dados demora menos de 2 horas para uma amostra de 317,503 SNPs e 2,000 indivíduos.

À continuação, segue o teste de associação onde o objetivo é a identificação e/ou seleção dos marcadores que se associam com a doença estudada. Esta é a etapa mais desafiante do ponto de vista computacional. O algoritmo proposto baseado na teoria da informação é auxiliado por um algoritmo genético que utiliza máscaras de grupos de SNPs para otimizar a busca e identificar os pares de SNPs que mostram uma associação relevante. Utilizando esta estratégia, o algoritmo *MIGA-2L* pode realizar a análise de associação em escala genômica 60 vezes mais rápido que *Plink*; o método mais difundido para GWAS e considerado "o estado da arte". O algoritmo genético, graças a sua natureza, consegue convergir rapidamente não precisando realizar uma busca exaustiva de todas as interações possíveis.

O algoritmo *MIGA-2L* foi desenvolvido para testar epistasia estatística em estudos caso-controle com fenótipos binários, mas é facilmente escalável para testar interações de maior grau, para agregar a utilização de outras funções de avaliação e para tratar dados de fenótipos com múltiplas variáveis. No entanto, a extensão do método proposto para lidar com dados fenotípicos com valor contínuo não é simples, a menos que os valores do fenótipo possam ser dicotomizados apropriadamente.

7.2 Trabalhos Futuros

Como foi explicado anteriormente, o enfoque deste estudo GWA atende a desenhos caso-controle onde o fenótipo é representado como uma variável binária. Certas melhoras podem ser feitas à metodologia descrita neste trabalho de forma a estender para outros tipos de estudos, como aqueles baseados em famílias e estudos de genes fazendo as considerações apropriadas sobre os dados.

Assim, a metodologia proposta não pode ser aplicada a GWAS envolvendo fenótipos representados como uma variável contínua a menos que os fenótipos contínuos possam ser dicotomizados. O mesmo se aplica aos marcadores genéticos que são tratados como variáveis categóricas (ex. SNP com valores ternários).

Covariáveis ambientais e variáveis de mistura genética poderiam ser quantitativas ou ordinais (ex. água, solo, pressão, temperatura, oxigeno etc.). Estender esta abordagem para permitir covariáveis mais gerais pode ser considerado em um trabalho futuro. Existem muitas formas naturais de realizar o tratamento de medidas contínuas, por exemplo, se a covariável foi discreta ou pode ser discretizada o método proposto pode ser estendido fazendo um tratamento direto. Senão, então outro tipo de pré-processamento deve ser feito para adaptá-la, como por exemplo utilizar regressão logística para ajustar a covariável.

Além disso, não foi investigado o tratamento dos dados de genótipos faltantes e seu efeito sobre o estudo. É muito comum a falta de genótipos em GWAS. Em geral, o mecanismo exato subjacente da falta de dados é desconhecido para os investigadores. Atualmente, o pressuposto de aletoriedade é feito para explicar a falta de genótipos (ou seja, se assume que genótipos e alelos diferentes estão faltando com a mesma probabilidade). No entanto, poucos estudos têm examinado a magnitude dos efeitos quando esta hipótese simplificadora é violada.

Como foi mencionado no capítulo anterior, a interpretação biológica dos achados continua sendo um grande desafio. Neste contexto, primeiro precisa-se identificar a variante causal. Em seguida, são necessárias provas experimentais para demonstrar o efeito molecular da variante sobre o gene e a doença/fenótipo. Todas estes validações funcionais baseadas em laboratório são altamente dependentes sobre o tipo de variantes, genes e doenças. A tendência atual e promissora é executar ferramentas genômicas adicionais de alto desempenho (por exemplo, considerando

arranjos de expression gênica de todo o genoma sobre tecidos relevantes), em paralelo com os estudos GWA, a fim de facilitar a interpretação biológica.

Finalmente, a metodologia proposta teve o propósito de ser o mais flexivel possível, de forma a conseguir em trabalhos futuros adicionar outros módulos que ajudem na análise de dados genomicos. Nesta visão é que a metodologia utiliza um motor de workflow de características flexíveis e escaláveis.

REFERÊNCIAS BIBLIOGRÁFICAS

ALMGREN P., BENDAHL P.O., BENGTSSON H., HOSSJER O. AND PERFEKT R., 2003, *Statistic in Genetics*. Lund University, Lund Institute of Technology, Centre for Mathematical Sciences, Mathematical Statistic.

ANDERSON C. A., PETTERSSON F.H., CLARKE G.M., CARDON L.R., MORRIS A.P., and ZONDERVAN K.T., 2010, "Data quality control in genetic case-control association studies". *Nat Protoc.*; vol.5, no. 9, pp. 1564–1573.

ARDLIE, K. G., KRUGLYAK, L., SEIELSTAD, M., 2002, "Patterns of linkage disequilibrium in the human genome", *Nat Rev Genet*, v.3, n.4, pp.299–309.

BALDING, D. J., 2006, "A tutorial on statistical methods for population association studies", *Nat Rev Genet*, v.7, n.10, pp. 781-791.

BATESON W, 1910, "Mendels principles of heredity", *Molecular and General Genetics MGG*, 3: 108–109.

BETARBET, R., SHERER, T.B. & GREENAMYRE, J TIMOTHY, 2005. "Ubiquitin-proteasome system and Parkinson's diseases". *Experimental neurology*, 191 Suppl , pp.S17-27.

BREIMAN L., 2001, "Random forests", *Machine Learning*, 45:5–32.

CANTOR RM, LANGE K, SINSHEIMER JS, 2010, "Prioritizing GWAS results: A review of statistical methods and recommendations for their application", *Am J Hum Genet.*, v.86, n.1, pp.6-22.

CARDON L.R. AND BELL J.I., 2001, "Association study designs for complex diseases", *Nature Reviews in Genetics*, v.2, pp. 91-99.

CARVALHO, D., *Árvore de decisão/Algoritmo genético para tratar o problema pequenos disjuntos em classificação de dados*. Tese de D.Sc., COPPE/UFRJ, Rio de Janeiro, Brasil, 2005.

CHEN X, LIU CT, ZHANG M, ZHANG H., 2007, "A forest-based approach to identifying gene and gene gene interactions", *Proc.Natl.Acad.Sci.*, v. 104, n. 49, pp.19199-19203.

CLARK, T.G., DE LORIO, M., GRIFFITHS, R.G., FARRALL, M., 2005, "Finding Association in Dense Genetic Maps: A Genetics Algorithm Approach". *Human Heredity*, v.60, pp. 97–108.

CLARK, T.G., DE LORIO, M., GRIFFITHS, R.G., 2008, "An Evolutionary Algorithm to Find Associations in Dense Genetic Maps", *IEEE Transactions on Evolutionary Computation*, v.12, n.3, pp. 297–306.

CONGDON, C.B., 1995, *A Comparison of Genetic Algorithms and other Machine Learning Systems on a Complex Classification Task from Common Disease*. Ph.D. Thesis in Computer Science, University of Michigan, Michigan, USA.

CORDELL H.J., "Epistasis: What it Means, What it Doesn't Mean, and Statistical Methods to Detect it in Humans", 2002, *Human Molecular Genetics*, vol. 11, no. 20, pp. 2463-2468.

CORDELL H.J., "Detecting Gene-Gene Interactions that Underlie Human Diseases", 2009, *Nature Rev. Genetics*, vol. 10, no. 6, pp. 392-404.

COVER T., THOMAS J., July 2006, *Elements of the Information theory*, 2nd edition, John Wiley & Sons, Inc.

CULVERHOUSE R, SUAREZ BK, LIN J, REICH T, 2002, "A perspective on epistasis: limits of models displaying no main effect", *Am J Hum Genet.* 70(2):461-71.

DEWAN A., LIU M., HARTMAN S., ZHANG S.S., LIU D.T., ZHAO C., TAM P.O., CHAN W.M., LAM D.S., SNYDER M. HTRA1 promoter polymorphism in wet age-related macular degeneration. Science. 2006;314:989–992.

DORIGO M, GAMBARDELLA LM: Ant colonies for the travelling salesman problem. Biosystems 1997, 43(2):73-81.

DUDA R.O., HART P.E., STORK D.G., 2001, *Pattern Classification*, 2nd edition, John Wiley & Sons, Inc.

ESTER, M., 1996, "A density-based algorithm for discovering clusters in large spatial databases with noise", In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, AAAI Press, Portland, OR, pp. 226–231.

ERICHSEN, H. C. AND CHANOCK, S. J., 2004, "SNPs in cancer research and treatment", *British Journal of Cancer*, v.90, pp. 747–751.

FAGERHOLM E, AHLQVIST E, FORSBLOM C, SANDHOLM N, SYREENI A, PARKKONEN M, MCKNIGHT AJ, TARNOW L, MAXWELL AP, PARVING HH, GROOP L, GROOP PH; Finn Diane Study Group, "SNP in the genome-wide association study hotspot on chromosome 9p21 confers susceptibility to diabetic nephropathy in type 1 diabetes", *Diabetologia*. 2012 Sep;55(9):2386-93.

FISHER R A, 1918, "The correlation between relatives on the supposition of Mendelian inheritance", *Transactions of the Royal Society of Edingurgh*, 52:399-433.

FONTANAROSA J., YANG D., "A Block-Based Evolutionary Optimization Strategy to Investigate Gene-Gene Interactions in Genetic Association Studies", Bioinformatics and Biomedicine Workshops (BIBMW), 2010 IEEE International Conference, 18 Dec. 2010, vol. 330-335.

FRANKEL W.N. and SCHORK N.J., "Who's afraid of epistasis?", 1996, *Nature genetics*, v.14, n.4: pp.371–373.

FREITAS, A.A., 2001, "Understanding the Crucial Role of Attribute Interaction in Data Mining", *Artificial Intelligence Review*, v.16, n.3, p.177-199.

FREITAS, A.A., 2002, *Data Mining and Knowledge Discovery with Evolutionary Algorithms*. New York, Springer-Verlag.

GHOUSSAINI M, FLETCHER O, MICHAILIDOU K, TURNBULL C, SCHMIDT MK, DICKS E, DENNIS J, WANG Q, HUMPHREYS MK, LUCCARINI C et al., 2012, "Genome-wide association analysis identifies three new breast cancer susceptibility loci", *Nat Genet.*, v.44(3): p. 312–318.

GOLDBERG, DAVID E. Genetic Algorithms in Search, Optimization, and Machine Learning. EUA: Addison-Wesley, 1989.

GOLDSTEIN D.B. Common genetic variation and human traits. N. Engl. J. Med. 2009;360:1696–1698.

GREENE CS, WHITE BC, MOORE JH., 2008, "Ant Colony Optimization for Genome-Wide Genetic Analysis", *Lect Notes Comput Sci.*, v.5217, pp. 37-47.

GREENE CS, WHITE BC, MOORE JH., 2009, May 18, "Sensible Initialization Using Expert Knowledge for Genome-Wide Analysis of Epistasis Using Genetic Programming", *Genet Evol Comput Conf.*, pp.1289-1296.

HAGENAUER, J., DAWY, Z., GOEBEL, B., HANUS, P., MUELLER, J.C., 2004. Genomic analysis using methods from information theory. IEEE Information Theory Workshop (ITW 2004), 55-59.

HINDORFF LA, MACARTHUR J (European Bioinformatics Institute), MORALES J (European Bioinformatics Institute), JUNKINS HA, HALL PN, KLEMM AK, and MANOLIO TA. A Catalog of Published Genome-Wide Association Studies. Available at: Available at: www.genome.gov/gwastudies [date of access].

HEROLD C, STEFFENS M, BROCKSCHMIDT FF, BAUR MP, BECKER T., 2009, "INTERSNP: genome-wide interaction analysis guided by a priori information", *Bioinformatics*, v.25 , pp. 3275–3281.

HOFFEE, P. A., 2000, "Genética Médica Molecular", Guanabara Koogan Edição, Oxford, Blackwell Science Limited.

IOANNIDIS J.P., NTZANI E.E., TRIKALINOS T.A. AND CONTOPOULOS- IOANNIDIS D.G., 2001, "Replication validity of genetic association studies", *Nature Genetics*, v.29, pp. 306-309.

JIANG R, TANG W, WU X, FU W., 2009 , "A random forest approach to the detection of epistatic interactions in case-control studies", *BMC Bioinformatics*, v.10, Suppl 1, S65.

JOHNSON, A.; O'DONNELL, C., 2009, "An open access database of genome-wide association results", *BMC medical genetics*, v. 10, n. 6.

JUYAL G, PRASAD P, SENAPATI S, MIDHA V, SOOD A, AMRE D, JUYAL RC, BK T, "An investigation of genome-wide studies reported susceptibility loci for ulcerative colitis shows limited replication in north Indians", *PLoS One*, 2011 Jan 31;6(1):e16565.

KEMPTHORNE O., 1954, "The correlation between relatives in a random mating population", *Proc R Soc Lond B Biol Sci.*, v.143, n.910, pp.102-13.

KEMPTHORNE O, 1968, "The correlation between relatives on the supposition of mendelian inheritance", *American Journal of Human Genetics*, 20: 402.

KLEIN R.J., ZEISS C., CHEW E.Y., TSAI J.Y., SACKLER R.S., HAYNES C., HENNING A.K., SANGIOVANNI J.P., MANE S.M., MAYNE S.T., 2005, "Complement factor H polymorphism in age-related macular degeneration", *Science*, v.308: pp.385–389.

KOHAVI R., 1995, "A study of cross-validation and bootstrap for accuracy estimation and model selection", In: *Proceedings of the 14th international joint conference on Artificial intelligence*, pp. 1137-1143, San Francisco, CA, USA.

KOOPERBERG C., RUCZINSKI I., 2005, "Identifying Interacting SNPs using Monte Carlo Logic Regression", *Genetic Epidemiology*, v.28, n.2, pp. 157-70.

KURREEMAN FA, STAHL EA, OKADA Y, LIAO K, DIOGO D, RAYCHAUDHURI S, FREUDENBERG J, KOCHI Y, et al, 2012, "Use of a multiethnic approach to identify rheumatoid- arthritis-susceptibility loci, 1p36 and 17q12", *Am J Hum Genet.*, v.90(3):pp.524-32.

LEVY J. and NAGYLAKI T., 1972, "A model for the genetics of handedness", *Genetics*, v. 72, n.1: pp.117–128.

LERNER I.M., Heredity, Evolution, and Society. W.H. Freeman, San Francisco, 1968.

LI W. and REICH J., 2000, "A complete enumeration and classification of two-locus disease models", *Human Heredity*, v.50: pp. 334–349.

LI NN, CHANG XL, MAO XY, ZHANG JH, ZHAO DM, TAN EK, PENG R., 2012, "GWAS-linked GAK locus in Parkinson's disease in Han Chinese and meta-analysis", *Hum Genet.*, v.131(7): pp.1089-93.

LU S, XIE Y, LIN K, LI S, ZHOU Y, MA P, LV Z, ZHOU X, 2012, "Genome-Wide Association Studies-derived susceptibility loci in Type 2 Diabetes: confirmation in a Chinese population", *Clin Invest Med.*, v. 35(5):E327.

MA L., H.B. RUNESHA, D. DVORKIN, J.R. GARBE, AND Y. DA., 2008, "Parallel and serial computing tools for testing single-locus and epistatic SNP effects

of quantitative traits in genome-wide association studies", *BMC Bioinformatics*, v.9, pp.315.

MANOLIO T.A., 2010, "Genome wide association studies and assessment of the risk of disease", N Engl J Med 2010 Jul 8;363(2):166-76, *N Engl J Med*, v.363, n.2, pp.166-76.

MARCHINI J. et al., 2005, "Genome-wide strategies for detecting multiple loci that influence complex diseases", *Nature Genetics*, 37:413–417.

MARCHINI J., HOWIE B., 2010, "Genotype imputation for genome-wide association studies", *Nature Genetics*, 11: 499-511.

MARCHINI J., HOWIE B., S. MYERS, G. MCVEAN and P. DONNELLY, 2007, "A new multipoint method for genome-wide association studies via imputation of genotypes", *Nature Genetics*, 39 : 906-913.

MEDRONHO R., BLOCH K.V., LUIZ R.R., WERNECK G.L., 2009, *Epidemiologia*. 2ed. São Paulo, Atheneu.

MILLER D.J., ZHANG Y., YU G., LIU Y., CHEN L., LANGEFELD C.D., HERRINGTON D, WANG Y., 2009, "An algorithm for learning maximum entropy probability models of disease risk that efficiently searches and sparingly encodes multilocus genomic interactions", *Bioinformatics*, v. 25, n.19, pp.2478-2485.

MILLSTEIN J, CONTI DV, GILLILAND FD, GAUDERMAN WJ., 2006, "A testing framework for identifying susceptibility genes in the presence of epistasis", *Am.J.Hum.Genet.*, v. 78, n.1, pp.15-27.

MOORE JH, HAHN LW, RITCHIE MD, THORNTON TA, WHITE BC., 2004, "Routine Discovery of Complex Genetic Models using Genetic Algorithms", *Appl Soft Comput*, v. 4, n.1, pp. 79-86.

MOORE JH. et al., 2006, "A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility", *J Theor Biol.*, pp.241:252–61.

MOORE JH, ASSELBERGS FW, WILLIAMS SM., 2010, "Bioinformatics challenges for genome-wide association studies", *Bioinformatics*, v.26, n.4, pp.445-455.

MOTSINGER-REIF AA, FANELLI TJ, DAVIS AC, RITCHIE MD., 2008, "Power of grammatical evolution neural networks to detect gene-gene interactions in the presence of error", *BMC Res Notes.*, v. 1, n.65, pp. 1-8.

MPICH."http://www.mcs.anl.gov/research/projects/mpich2/overview", visited June, 2011.

NEALE B, FERREIRA M, MEDLAND S, POSTHUMA D. Statistical in genetics: Gene Mapping through Linkage and Association. Taylor & Francis, 2008 US.

NEUMAN R.J. and RICE J.P., 1992 , "Two-locus models of disease", *Genetic Epidemiology*, vol. 9, n.5:pp. 347–365.

NOBLE J. A. and ERLICH H.A., 2012, "Genetics of Type 1 Diabetes", *Cold Spring Harb Perspect Med.*, v. 2, n.1: a007732.

PACKARD N.H., 1990, "A Genetic learning algorithm for the analysis of complex data", *Complex Sustem*, v.4, pp.543-572.

PANKRATZ N. D., WOJCIESZEK J., FOROUD T., 2007, "Parkinson Disease Overview", Gene Reviews, www.ncbi.nlm.nih.gov. Reference Type: Internet Communication.

PARK MY, HASTIE T., 2008, "Penalized logistic regression for detecting gene interactions", *Biostatistics*, v.9, n.1, pp. 30-50.

PARKES M et al., 2007, "Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility", *Nat Genet.*, 39(7):pp. 830-2.

PHILLIPS P.C., "Epistasis the Essential Role of Gene Interactions in the Structure and Evolution of Genetic Systems", 2008, *Nature Rev. Genetics*, vol. 9, no. 11, pp. 855-867.

PURCELL S, NEALE B, TODD-BROWN K, THOMAS L, FERREIRA MA, BENDER D, MALLER J, SKLAR P, DE BAKKER PI, DALY MJ, SHAM PC., 2007, "PLINK: a tool set for whole-genome association and population-based linkage analyses", *Am. J. Hum. Genet*., v.81, pp. 559–575.

RISCH, N. and MERIKANGAS, K., 1996, "The future of genetic studies of complex human diseases", *Science* **273**: 1516-1517.

RITCHIE MD, HAHN LW, ROODI N, BAILEY LR, DUPONT WD, PARL FF, MOORE JH., 2001, "Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer", *Am.J.Hum.Genet*., v.69, n.1, pp.138-147.

SCHÜPBACH T., XENARIOS I., BERGMANN S., KAPUR K., 2010, "FastEpistasis: A high performance computing solution for quantitative trait epistasis", *Bioinformatics*, v.26, n.11, pp.1468-9.

SLADEK R, ROCHELEAU G, RUNG J., 2007, "A genome-wide association study identifies novel risk loci for type 2 diabetes", *Nature*, v.445, n. 7130, pp. 881–885.

SHAH, S., KUSIAK, A., 2004, "Data mining and genetic algorithm based gene/SNP selection", *Artificial Intelligence in Medicine*, v. 31, pp.183–196

SHAM P., 1998, *Statistic in human genetics*, Arnold Applications of Statistic Series, London, Arnold.

SPENCER CC, SU Z, DONNELLY P, MARCHINI J., "Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip", *PLoS Genet*, v. 5:e1000477-e1000477.

SOUZA VNB, PEREIRA AC., 2007, "Genética humana na susceptibilidade à hanseníase", *Hansen Int*.;32(1): 81-93.

TANG W, WU X, JIANG R, LI Y., 2009, "Epistatic module detection for case-control studies: a Bayesian model with a Gibbs sampling strategy", *PLoS Genet*, v.5, n.5. e1000464.

TOMLINSON I. P. M., CARVAJAL-CARMONA L. G., DOBBINS S. E., TENESA A., JONES A.M. et al, 2011, "Multiple Common Susceptibility Variants near BMP Pathway Loci GREM1, BMP4, and BMP2 Explain Part of the Missing Heritability of Colorectal Cancer", *PLoS Genet.*, v. 7(6): e1002105.

VELEZ DR, WHITE BC, MOTSINGER AA, BUSH WS, RITCHIE MD, WILLIAMS SM, MOORE JH, 2007, "A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction", *Genet Epidemiol.*, v.31, n.4, pp.306-315.

VISSCHER PM, BROWN MA, MCCARTHY MI, YANG J, 2012, “Five years of GWAS discovery”, *Am J Hum Genet.*, 90(1):7-24.

WAN X, YANG C, YANG Q, XUE H, TANG NL, YU W., 2009, “MegaSNPHunter: a learning approach to detect disease predisposition SNPs and high level interactions in genome wide association studies”, *BMC Bioinformatics*, v.10, n.13.

WAN X, YANG C, YANG Q, XUE H, TANG NL, YU W, 2010, “Predictive rule inference for epistatic interaction detection in genome-wide association studies”, *Bioinformatics*, v.26, n.1, pp.30-37.

WAN X, YANG C, YANG Q, XUE H, FAN X, TANG NL, YU W, 2010, “BOOST: A Fast Approach to Detecting Gene-Gene Interactions in Genome-wide Case-Control Studies”, *The American Journal of Human Genetics*, v.87, n.3, pp. 325-340.

WANG Y, LIU X, ROBBINS K, REKAYA R, 2010, “AntEpiSeeker: detecting epistatic interactions for case-control studies using a two-stage ant colony optimization algorithm”, *BMC Res Notes*, vol. 3:117.

WELLCOME TRUST CASE CONTROL CONSORTIUM, 2007, “Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls”, *Nature*, v.447, n. 7145, pp. 661–78.

WEST, ANDREW B. et al., 2003. “Identification of a Novel Gene Linked to Parkin via a Bi-directional Promoter”. *Journal of Molecular Biology*, 326(1), pp.11-19.

WRAY, N.R., 2005, “Allele frequencies and the r2 measure of linkage disequilibrium: impact on design and interpretation of association studies.”, *Twin Research and Human Genetics*, v. 8, pp. 87-94.

YANG, C., WAN, X., YANG, Q., XUE, H., YU, W., 2010, “Identifying main effects and epistatic interactions from large-scale snp data via adaptive group lasso”, *BMC Bioinformatics* , v.11, Suppl 1, S18.

YANG C, HE Z, WAN X, YANG Q, XUE H, YU W, 2009, “SNPHarvester: a filtering-based approach for detecting epistatic interactions in genome-wide association studies”, *Bioinformatics*, v.25, n.4, pp.504-511.

ZHANG X, ZOU F, WANG W. “FastANOVA: an Efficient Algorithm for Genome-Wide Association Study”, *Proc. KDD*, 2008. pp. 821–829.

ZHANG X, HUANG S, ZOU F, WANG W., 2010, “TEAM: efficient two-locus epistasis tests in human genome-wide association study”, *Bioinformatics*, v.26, n.12, pp. i217-27.

ZHANG Y, LIU JS., 2007, “Bayesian inference of epistatic interactions in case-control studies”, *Nat.Genet.*, v.39, n. 9, pp.1167-1173.

ZHANG X, PAN F, XIE Y, ZOU F, WANG W., 2011, “Tools for efficient epistasis detection in genome-wide association study”, *Source Code Biol Med.*, v.6, n.1(Jan), pp.1.

ZHANG X, HUANG S, ZOU F, WANG W., 2011, “COE: a General Approach for Efficient Genome-Wide Two-Locus Epistatic Test in Disease Association Study”, *Journal of Computational Biology*. 2010;17(3):401–415.

ZHENG T, WANG H, LO SH., 2006, “Backward genotype-trait association (BGTA)-based dissection of complex traits in case-control designs”, *Hum.Hered.*, v.62, n.4, pp.196-212.

APÊNDICE

A1. Informação Mutua de duas variáveis aleatórias

A informação mútua de duas variáveis aleatórias é a quantidade de dependência mútua das duas variáveis aleatórias. Formalmente, a informação mútua de duas variáveis aleatórias discretas X e Y pode ser definido como:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right) \quad [1]$$

onde, *p(x,y)* é a função de distribuição da probabilidade conjunta de *X* e *Y*, e *p(x)* e *p(y)* são as funções de distribuição das probabilidades marginais de *X* e *Y*, respectivamente.

A informação mútua pode equivalentemente ser expressada em termos da entropia de uma variável aleatória. Intuitivamente, ela mede a informação que *X* e *Y* compartilham, ou seja, ela mede o quanto conhecer uma dessas variáveis reduz a incerteza sobre a outra.

Se a incerteza de uma variável aleatória X é medida por sua entropia H(X), calculada pela equação $-\sum_x p(x) log_2 p(x)$ (definida por Shannon em 1948); então a incerteza de uma variável aleatória X dado o conhecimento de outra variável aleatória Y é medida por sua entropia condicional H(X|Y); e a incerteza de um par de variáveis aleatórias X,Y é medida pela entropia conjunta H(X,Y), calculada pela equação $-\sum_x \sum_y p(x,y) log_2 p(x,y)$. Estas quantidades são relacionadas na equação 2.

$$H(X,Y) = H(X) + H(X|Y) = H(Y) + H(Y|X) \quad [2]$$

Aqui, a Informação mútua é definida como o grau de incerteza em X, menos a quantidade de incerteza em X que permanece depois que Y é conhecido, o que é equivalente a dizer "a quantidade de incerteza em X que é removida por conhecer Y". O mesmo equivale para Y, como se mostra na equação 3

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \quad [3]$$

Da equação anterior, I(X;Y) pode-se expressar como

$$I(X;Y) = [H(X) + H(Y)] - H(X,Y) \quad [4]$$

Portanto, o significado físico de I(X;Y) é "a redução da incerteza de X devido ao conhecimento de Y" (ou vice-versa) e esta relação pode ser representada em um

diagrama de Venn (Figura 1a) na qual a entropia das variáveis H(X) e H(Y) são representadas por dois conjuntos sobrepostos. A entropia das duas variáveis é representada pela união de estes conjuntos, e a informação mútua de X e Y é representada por sua interseção. Também a entropia condicional está representada no diagrama que é indicada pela subtração dos conjuntos, de modo que, por exemplo, o conjunto representado por H(X|Y) resulta de subtrair o conjunto representando H(Y) do conjunto representando por H(X).

A informação mútua pode ser condicionada sobre o conhecimento de outra variável aleatória incluindo esta condição sobre todos os termos da definição. Por exemplo, dada uma terceira variável aleatória Z, a informação mútua I(X;Y|Z) é igual a H(X|Z) - H(X|Y,Z), isto significa "a redução da incerteza de X devido a Y (ou vice-versa), quando Z é dado". O diagrama de Venn da Figura 1b mostra que o conjunto representando I(X;Y|Z) resulta de subtrair o conjunto representando H(Z) desde o conjunto representado por I(X;Y).

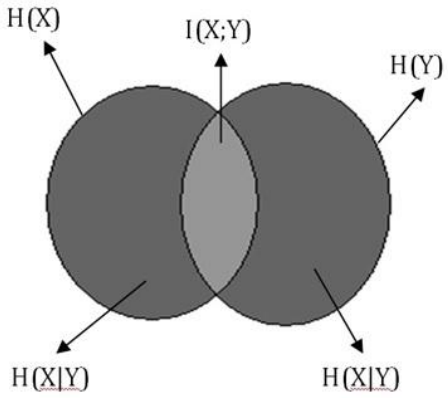


Figura 1a: Informação mutua de duas variáveis.

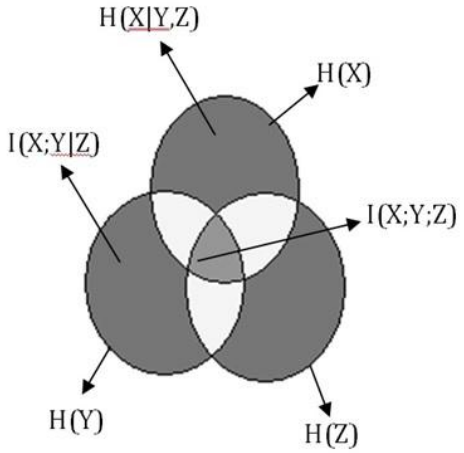


Figura 1b: Informação mutua de três variáveis.

A informação mútua referente a três variáveis X, Y, Z pode ser definida pela equação 5 [McGill, 1954; Watanabe, 1960; Cover and Thomas, 2006]

$$I(X;Y;Z) = [H(X|Z) + H(Y|Z)] - H(X,Y \mid Z) \quad [5]$$

Das definições acima pode se concluir a seguinte equação:

$$I(X;Y;Z) = I(X;Y) - I(X;Y|Z) \quad [6]$$

Isto significa "a redução da informação mútua comum para duas variáveis devido ao conhecimento de uma terceira variável".

A2. Classificador Naive Bayes.

Ele forma parte do grupo de classificadores Bayesianos. Eles são classificadores estatísticos que classificam um objeto numa determinada classe baseando-se na probabilidade deste objeto pertencer a esta classe. O classificador *Naive Bayes* supõe como hipótese que o efeito do valor de um atributo não-classe é independente dos valores dos outros atributos. Isto é, o valor de um atributo não influencia o valor dos outros. Esta hipótese tem como objetivo facilitar os cálculos envolvidos na tarefa de classificação.

O Naive Bayes é aplicado da seguinte maneira:

O procedimento consiste em contar o número de aparições de cada haplótipo, da Tabela 5.3, chamado h_k onde $1 \leq k \leq 9$, agrupados por C_n, onde n={casos,controles}. Seguidamente, calculamos a probabilidade de h_k de pertencer a C_{casos} ou $C_{controles}$. Esta probabilidade $P[C_n|h_k]$ também é chamada *probabilidade posterior*, e pode ser calculada utilizando o teorema de Bayes.

$$P[C_n|h_k] = \frac{P[h_k|C_n] \ * P[C_n]}{P[h_k]}$$

As probabilidades $P[h_k|C_n]$ podem ser calculadas a partir das amostras da seguinte maneira:

$$P[h_k|C_n] = \frac{numero\ de\ h_k\ classificadas\ em\ C_n}{numero\ de\ C_n}$$

A probabilidade posterior é a probabilidade de que h_k pertença à classe C_n e a probabilidade a priori $P[C_n]$ é a probabilidade total da classe C_n.

O critério de classificação é chamado de Máximo a *posteriori*, ou MAP, no qual se classifica para a classe C_n com o Máximo das probabilidades *posteriori* das duas classes,

Classe (C_n) = $Máximo_n\{P[C_n|h_k]\} = máximo_n\{f(h_k|C_n)p(C_n)\}$

Daqui, a função discriminante é definida como igual a sua probabilidade posterior:

$$d_n(h_k) = f(h_k|n)p(C_n)$$

A classe é aquela que maximiza o discriminante. Então uma vez obtido o Máximo das probabilidades a posteriori criamos a regra do tipo SE h_k ENTÃO C_n.

A3. Conjunto de dados com efeito principal

Foram gerados 4 classes de modelo de dados. O primeiro modelo, também chamado modelo multiplicativo [MARCHINI et al., 2005], o segundo modelo chamado de modelo epistático [NEWMAN & RICE, 1992], o terceiro modelo é um clássico modelo epistático [LEVY & NAGYLAKI, 1972; LERNER & FREEMAN, 1968] e o quarto é descrito como modelo XOR (OR exclusivo). A geração dos arquivos de genótipos pertencentes a cada modelo foi baseada considerando Tabelas de Odds Ratio como se mostra a continuação.

Seja $\mathrm{p}(\mathrm{D}|\mathrm{G_i})$ a probabilidade de que um indivíduo seja afetado pela doença dado seu genótipo $\mathrm{G_i}$ (ex. penetrância de $\mathrm{G_i}$), e seja $\mathrm{p}(\overline{\mathrm{D}}|\mathrm{G_i})$ a probabilidade de que um indivíduo não seja afetado dado seu genótipo $\mathrm{G_i}$. Baseado na definição da Odds Ratio da doença,

$$\mathrm{ODD_{G_i}} = \frac{\mathrm{p}(\mathrm{D}|\mathrm{G_i})}{\mathrm{p}(\overline{\mathrm{D}}|\mathrm{G_i})} = \frac{\mathrm{p}(\mathrm{D}|\mathrm{G_i})}{1 - \mathrm{p}(\overline{\mathrm{D}}|\mathrm{G_i})}$$

A penetrância $\mathrm{p}(\mathrm{D}|\mathrm{G_i})$ do genótipo pode ser calculado pela equação,

$$\mathrm{p}(\mathrm{D}|\mathrm{G_i}) = \frac{\mathrm{ODD_{G_i}}}{1 + \mathrm{ODD_{G_i}}}$$

A prevalência $\mathrm{p}(\mathrm{D})$ e a herdabilidade genética h^2 são estimadas pelas equações seguintes,

$$\mathrm{p}(\mathrm{D}) = \sum_{\mathrm{i}} \mathrm{p}(\mathrm{D}|\mathrm{G_i})\mathrm{p}(\mathrm{G_i})$$

$$\mathrm{h}^2 = \frac{\sum_{\mathrm{i}} \big(\mathrm{p}(\mathrm{D}|\mathrm{G_i}) - \mathrm{p}(\mathrm{D})\big)^2 \mathrm{p}(\mathrm{G_i})}{\mathrm{p}(\mathrm{D})(1 - \mathrm{p}(\mathrm{D}))}$$

Na simulação, a prevalência $p(D)$ e a herdabilidade h^2 são controladas pelos parâmetros α e θ da Tabela A1. Primeiro os valores da prevalência e herdabilidade são determinados para depois resolver numericamente os parâmetros (α e θ) baseados nas equações correspondentes.

Por exemplo, seja $p(D) = 0.1$ e h^2=0.03 no modelo 1. Então α= 0.09989 e θ=3.4481 para uma frequência de alelo menor (MAF) = 0.1.

Modelo 1	BB	Bb	bb
AA	A	α	α
Aa	α	α (1+θ)	α (1+θ)2
AA	α	α (1+θ)	α (1+θ)4

Modelo 2	BB	Bb	bb
AA	A	α (1+θ)	α (1+θ)
Aa	α (1+θ)	α	α
aa	α (1+θ)	α	α

Modelo 3	BB	Bb	bb
AA	A	α	α (1+θ)
Aa	α	α (1+θ)	α
aa	α (1+θ)	α (1+θ)	α

Modelo 4	BB	Bb	bb
AA	A	α (1+θ)	α
Aa	α (1+θ)	α	α (1+θ)
aa	α	α (1+θ)	α

Tabela A1. Os parâmetros α e θ controlam a prevalência da doença *p(D)* e a herdabilidade h^2

A4. Conjunto de dados sem efeito principal.

Estes modelos sem efeito principal são discutidos amplamente em [CULVERHOUSE et al., 2002; VELEZ et al., 2007]. Nesta tese foram utilizados 70 modelos epistáticos sem efeito principal os quais são listados na próxima página. A herdabilidade h^2 controla a variação fenotípica destes 70 modelos, com valores desde 0.01 até 0.4 e o MAF abrange valores desde 0.2 até 0.4.

Tabela A2: Valores de penetrância dos 70 modelos epistáticos com seus parâmetros de herdabilidade h^2 e MAF respectivos.

h^2=0.4, MAF=0.2				h^2=0.4, MAF=0.4			
Conjunto 00	AA	Aa	aa	Conjunto 05	AA	Aa	aa
BB	0.486	0.960	0.538	BB	0.077	0.656	0.880
Bb	0.947	0.004	0.811	Bb	0.892	0.235	0.312
bb	0.640	0.606	0.909	bb	0.174	0.842	0.106
h^2=0.4, MAF=0.2				h^2=0.4, MAF=0.4			
Conjunto 01	AA	Aa	aa	Conjunto 06	AA	Aa	aa
BB	0.469	0.956	0.697	BB	0.895	0.323	0.161
Bb	0.945	0.019	0.585	Bb	0.068	0.728	0.806
bb	0.786	0.407	0.013	bb	0.925	0.233	0.362
h^2=0.4, MAF=0.2				h^2=0.4, MAF=0.4			
Conjunto 02	AA	Aa	aa	Conjunto 07	AA	Aa	aa
BB	0.498	0.954	0.786	BB	0.805	0.251	0.085
Bb	0.978	0.038	0.428	Bb	0.002	0.668	0.638
bb	0.590	0.821	0.380	bb	0.830	0.079	0.542
h^2=0.4, MAF=0.2				h^2=0.4, MAF=0.4			
Conjunto 03	AA	Aa	aa	Conjunto 08	AA	Aa	aa
BB	0.505	0.988	0.624	BB	0.307	0.682	0.958
Bb	0.945	0.085	0.807	Bb	0.997	0.390	0.281
bb	0.969	0.116	0.159	bb	0.012	0.990	0.698
h^2=0.4, MAF=0.2				h^2=0.4, MAF=0.4			
Conjunto 04	AA	Aa	aa	Conjunto 09	AA	Aa	aa
BB	0.486	0.963	0.512	BB	0.083	0.891	0.037
Bb	0.941	0.006	0.899	Bb	0.619	0.271	0.691
bb	0.691	0.541	0.614	bb	0.853	0.079	0.742

h^2=0.3, MAF=0.2				h^2=0.3, MAF=0.4			
Conjunto 10	AA	Aa	aa	Conjunto 15	AA	Aa	aa
BB	0.500	0.926	0.615	BB	0.891	0.362	0.480
Bb	0.895	0.131	0.647	Bb	0.213	0.829	0.601
bb	0.858	0.160	0.999	bb	0.925	0.267	0.685
h^2=0.3, MAF=0.2				h^2=0.3, MAF=0.4			
Conjunto 11	AA	Aa	aa	Conjunto 16	AA	Aa	aa
BB	0.413	0.851	0.535	BB	0.077	0.689	0.417
Bb	0.831	0.008	0.580	Bb	0.763	0.150	0.491
bb	0.692	0.268	0.736	bb	0.196	0.657	0.247
h^2=0.3, MAF=0.2				h^2=0.3, MAF=0.4			
Conjunto 12	AA	Aa	aa	Conjunto 17	AA	Aa	aa
BB	0.455	0.848	0.897	BB	0.132	0.793	0.274
Bb	0.890	0.088	0.016	Bb	0.799	0.213	0.514
bb	0.562	0.686	0.467	bb	0.255	0.528	0.793
h^2=0.3, MAF=0.2				h^2=0.3, MAF=0.4			
Conjunto 13	AA	Aa	aa	Conjunto 18	AA	Aa	aa
BB	0.609	0.980	0.980	BB	0.611	0.104	0.759
Bb	0.993	0.300	0.275	Bb	0.180	0.674	0.019
bb	0.876	0.483	0.683	bb	0.532	0.189	0.681
h^2=0.3, MAF=0.2				h^2=0.3, MAF=0.4			
Conjunto 14	AA	Aa	aa	Conjunto 19	AA	Aa	aa
BB	0.446	0.844	0.774	BB	0.091	0.827	0.863
Bb	0.879	0.044	0.233	Bb	0.869	0.393	0.415
bb	0.492	0.796	0.410	bb	0.738	0.508	0.363

h^2=0.2, MAF=0.2				h^2=0.2, MAF=0.4			
Conjunto 20	AA	Aa	aa	Conjunto 25	AA	Aa	aa
BB	0.428	0.757	0.812	BB	0.356	0.891	0.809
Bb	0.788	0.132	0.044	Bb	0.955	0.508	0.611
bb	0.559	0.548	0.373	bb	0.617	0.755	0.630
h^2=0.2, MAF=0.2				h^2=0.2, MAF=0.4			
Conjunto 21	AA	Aa	aa	Conjunto 26	AA	Aa	aa
BB	0.507	0.842	0.605	BB	0.086	0.536	0.641
Bb	0.845	0.162	0.629	Bb	0.677	0.275	0.096
bb	0.581	0.678	0.729	bb	0.219	0.413	0.712
h^2=0.2, MAF=0.2				h^2=0.2, MAF=0.4			
Conjunto 22	AA	Aa	aa	Conjunto 27	AA	Aa	aa
BB	0.577	0.247	0.428	BB	0.855	0.339	0.772
Bb	0.227	0.928	0.578	Bb	0.513	0.651	0.607
bb	0.586	0.262	0.158	bb	0.250	0.999	0.154
h^2=0.2, MAF=0.2				h^2=0.2, MAF=0.4			
Conjunto 23	AA	Aa	aa	Conjunto 28	AA	Aa	aa
BB	0.340	0.637	0.654	BB	0.506	0.838	0.024
Bb	0.689	0.017	0.041	Bb	0.603	0.454	0.957
bb	0.242	0.866	0.403	bb	0.729	0.427	0.753
h^2=0.2, MAF=0.2				h^2=0.2, MAF=0.4			
Conjunto 24	AA	Aa	aa	Conjunto 29	AA	Aa	aa
BB	0.387	0.726	0.734	BB	0.393	0.764	0.664
Bb	0.749	0.090	0.034	Bb	0.850	0.398	0.733
bb	0.551	0.401	0.724	bb	0.406	0.927	0.147

h^2=0.1, MAF=0.2				h^2=0.1, MAF=0.4			
Conjunto 30	AA	Aa	aa	Conjunto 35	AA	Aa	aa
BB	0.463	0.703	0.431	BB	0.137	0.484	0.187
Bb	0.653	0.277	0.806	Bb	0.482	0.166	0.365
bb	0.830	0.008	0.129	bb	0.193	0.361	0.430
h^2=0.1, MAF=0.2				h^2=0.1, MAF=0.4			
Conjunto 31	AA	Aa	aa	Conjunto 36	AA	Aa	aa
BB	0.319	0.507	0.569	BB	0.469	0.198	0.754
Bb	0.553	0.105	0.045	Bb	0.337	0.502	0.141
bb	0.203	0.777	0.280	bb	0.339	0.453	0.285
h^2=0.1, MAF=0.2				h^2=0.1, MAF=0.4			
Conjunto 32	AA	Aa	aa	Conjunto 37	AA	Aa	aa
BB	0.627	0.393	0.335	BB	0.478	0.311	0.864
Bb	0.396	0.779	0.953	Bb	0.387	0.579	0.263
bb	0.174	0.842	0.106	bb	0.634	0.436	0.138
h^2=0.1, MAF=0.2				h^2=0.1, MAF=0.4			
Conjunto 33	AA	Aa	aa	Conjunto 38	AA	Aa	aa
BB	0.297	0.540	0.441	BB	0.068	0.299	0.017
Bb	0.541	0.072	0.278	Bb	0.289	0.044	0.285
bb	0.434	0.293	0.228	bb	0.048	0.262	0.174
h^2=0.1, MAF=0.2				h^2=0.1, MAF=0.4			
Conjunto 34	AA	Aa	aa	Conjunto 39	AA	Aa	aa
BB	0.332	0.562	0.573	BB	0.539	0.120	0.258
Bb	0.583	0.112	0.147	Bb	0.165	0.378	0.325
bb	0.399	0.496	0.033	bb	0.123	0.426	0.276

h^2=0.05, MAF=0.2				h^2=0.05, MAF=0.4			
Conjunto 40	AA	Aa	aa	Conjunto 45	AA	Aa	aa
BB	0.492	0.664	0.481	BB	0.002	0.155	0.214
Bb	0.642	0.330	0.746	Bb	0.199	0.071	0.022
bb	0.656	0.396	0.000	bb	0.081	0.122	0.135
h^2=0.05, MAF=0.2				h^2=0.05, MAF=0.4			
Conjunto 41	AA	Aa	aa	Conjunto 46	AA	Aa	aa
BB	0.499	0.639	0.765	BB	0.188	0.020	0.171
Bb	0.666	0.389	0.083	Bb	0.032	0.174	0.059
bb	0.543	0.527	0.953	bb	0.134	0.087	0.092
h^2=0.05, MAF=0.2				h^2=0.05, MAF=0.4			
Conjunto 42	AA	Aa	aa	Conjunto 47	AA	Aa	aa
BB	0.212	0.350	0.116	BB	0.005	0.179	0.251
Bb	0.336	0.054	0.495	Bb	0.211	0.100	0.026
bb	0.227	0.273	0.495	bb	0.156	0.098	0.156
h^2=0.05, MAF=0.2				h^2=0.05, MAF=0.4			
Conjunto 43	AA	Aa	aa	Conjunto 48	AA	Aa	aa
BB	0.805	0.683	0.638	BB	0.174	0.321	0.154
Bb	0.657	0.936	0.989	Bb	0.223	0.254	0.245
bb	0.850	0.564	0.866	bb	0.448	0.025	0.424
h^2=0.05, MAF=0.2				h^2=0.05, MAF=0.4			
Conjunto 44	AA	Aa	aa	Conjunto 49	AA	Aa	aa
BB	0.638	0.488	0.383	BB	0.098	0.219	0.302
Bb	0.464	0.765	0.957	Bb	0.302	0.126	0.121
bb	0.580	0.562	0.719	bb	0.053	0.308	0.136

h^2=0.025, MAF=0.2				h^2=0.025, MAF=0.4			
Conjunto 50	AA	Aa	aa	Conjunto 55	AA	Aa	aa
BB	0.495	0.415	0.657	BB	0.166	0.165	0.128
Bb	0.429	0.616	0.121	Bb	0.114	0.199	0.143
bb	0.552	0.331	0.419	bb	0.281	0.028	0.281
h^2=0.025, MAF=0.2				h^2=0.025, MAF=0.4			
Conjunto 51	AA	Aa	aa	Conjunto 56	AA	Aa	aa
BB	0.592	0.691	0.743	BB	0.108	0.006	0.080
Bb	0.712	0.493	0.419	Bb	0.026	0.079	0.046
bb	0.580	0.746	0.504	bb	0.021	0.090	0.025
h^2=0.025, MAF=0.2				h^2=0.025, MAF=0.4			
Conjunto 52	AA	Aa	aa	Conjunto 57	AA	Aa	aa
BB	0.108	0.194	0.186	BB	0.006	0.094	0.008
Bb	0.196	0.037	0.045	Bb	0.079	0.016	0.076
bb	0.172	0.073	0.130	bb	0.052	0.043	0.057
h^2=0.025, MAF=0.2				h^2=0.025, MAF=0.4			
Conjunto 53	AA	Aa	aa	Conjunto 58	AA	Aa	aa
BB	0.112	0.186	0.128	BB	0.199	0.072	0.168
Bb	0.193	0.024	0.138	Bb	0.086	0.187	0.076
bb	0.079	0.236	0.251	bb	0.125	0.108	0.226
h^2=0.025, MAF=0.2				h^2=0.025, MAF=0.4			
Conjunto 54	AA	Aa	aa	Conjunto 59	AA	Aa	aa
BB	0.272	0.192	0.185	BB	0.165	0.096	0.262
Bb	0.172	0.367	0.390	Bb	0.166	0.151	0.091
bb	0.345	0.069	0.005	bb	0.050	0.250	0.056

h^2=0.01, MAF=0.2				h^2=0.01, MAF=0.4			
Conjunto 60	AA	Aa	aa	Conjunto 65	AA	Aa	aa
BB	0.247	0.301	0.205	BB	0.103	0.063	0.124
Bb	0.300	0.173	0.378	Bb	0.098	0.086	0.069
bb	0.215	0.357	0.268	bb	0.021	0.147	0.059
h^2=0.01, MAF=0.2				h^2=0.01, MAF=0.4			
Conjunto 61	AA	Aa	aa	Conjunto 66	AA	Aa	aa
BB	0.222	0.276	0.141	BB	0.185	0.291	0.234
Bb	0.259	0.169	0.401	Bb	0.286	0.201	0.277
bb	0.278	0.128	0.420	bb	0.249	0.266	0.166
h^2=0.01, MAF=0.2				h^2=0.01, MAF=0.4			
Conjunto 62	AA	Aa	aa	Conjunto 67	AA	Aa	aa
BB	0.260	0.221	0.201	BB	0.073	0.042	0.015
Bb	0.204	0.315	0.348	Bb	0.024	0.064	0.059
bb	0.339	0.074	0.128	bb	0.068	0.019	0.095
h^2=0.01, MAF=0.2				h^2=0.01, MAF=0.4			
Conjunto 63	AA	Aa	aa	Conjunto 68	AA	Aa	aa
BB	0.139	0.188	0.221	BB	0.046	0.127	0.069
Bb	0.190	0.111	0.020	Bb	0.115	0.067	0.097
bb	0.206	0.051	0.253	bb	0.107	0.069	0.108
h^2=0.01, MAF=0.2				h^2=0.01, MAF=0.4			
Conjunto 64	AA	Aa	aa	Conjunto 69	AA	Aa	aa
BB	0.558	0.616	0.674	BB	0.095	0.122	0.127
Bb	0.632	0.499	0.418	Bb	0.097	0.129	0.010
bb	0.546	0.674	0.395	bb	0.201	0.044	0.122