Universidade Federal de Minas Gerais

Escola de Veterinária

Vivian Paula Silva Felipe

Efeito da imputação de genótipos sobre a predição de características quantitativas de ratos utilizando marcadores genéticos

Belo Horizonte

2013

Vivian Paula Silva Felipe

Efeito da imputação de genótipos sobre a predição de características quantitativas de ratos utilizando marcadores genéticos

Tese apresentada ao Programa de Pós-Graduação em Zootecnia da Escola de Veterinária da Universidade Federal de Minas Gerais como requisito parcial para Obtenção do grau de Doutor em Zootecnia

Área de concentração: Genética e Melhoramento Animal

Prof. Orientador: Martinho de Almeida e Silva

Belo Horizonte

2013

Tese defendida e aprovada em 18/03/2013 pela Comissão examinadora composta pelos seguintes membros:

Prof. Martinho de Almeida e Silva (Orientador)

Prof. José Aurélio Garcia Bergmann

Prof. Idalmo Garcia Pereira

Prof. Fernando Enrique Madalena

Prof. Aldrin Vieira Pires

"*Essentially, all models are wrong, but some are useful*"

George E. P. Box, 1987

AGRADECIMENTOS

Aos meus pais , Roberto e Luci Mary, irmãos e família (incluindo Boninho, Bebezão e Nisinha) por todo carinho e apoio incondicionais, sem eles nada seria possível.

Ao Rocha, meu amor, pelo companheirismo e apoio em todos os momentos.

Ao meu querido orientador Prof. Martinho de Almeida e Silva pelo incentivo, entusiasmo e dedicação.

Aos meus co-orientadores Prof. Guilherme J.M. Rosa e Prof. Fabyano Fonseca pelo auxílio no desenvolvimento deste projeto. Em especial ao Prof. Guilherme Rosa por me abrir as portas de seu laboratório, o que foi fundamental para o meu crescimento pessoal e profissional.

Aos colegas da UFMG que me acompanharam durante toda a minha evolução. Em especial Raphael Wenceslau, Bruno Valente, Luciana Salles de Freitas, Glaucyana Gouvêa dos Santos e Natascha Almeida. E ao "grupo das codornas" Fabiana Ferreira, Rodrigo Godinho, Sirlene Lázaro, Vera Ferreira, Arthur Fernandes, Flaviana Miranda, Laila Alvarenga , entre outros que passaram por lá e pelo grupo de estudos...

Aos colegas da UW que fizeram todos os momentos fora de minha "casa" muito mais agradáveis e proveitosos.

Aos membros da banca por aceitarem participar desta importante etapa da minha vida.

Ao CNPq pela bolsa de estudos.

SUMÁRIO

CAPÍTULO 1

Resumo ____ **09**

Abstract ____ **09**

Introdução ____ **09**

Revisão de Literatura ____ **12**

1. Seleção Genômica ... 12
   1.1. Genética molecular e de populações: conceitos básicos ... 12
   1.2. WGR: Modelos e dimensionalidade ... 15
      1.2.1. Quadrados mínimos e GBLUP ... 17
      1.2.2. Modelos bayesianos: o alfabeto e o LASSO ... 21
      1.2.3. Modelos semi e não-paramétricos ... 26
         1.2.3.1. RKHS ... 26
         1.2.3.2. Redes Neurais Artificiais ... 27
            1.2.3.2.1. Histórico ... 27
            1.2.3.2.2. RNA: Definições e estrutura ... 28
            1.2.3.2.3. RNA: Treinamento e aprendizado ... 32
            1.2.3.2.4. Regularização de Redes Neurais Artificiais ... 36
            1.2.3.2.5. Aplicações no melhoramento genético ... 38
2. Imputação de genótipos ... 41
   2.1. Metodologias para imputação ... 41
      2.1.1. Modelos com cadeias de Markov ocultas ... 41
         - fastPHASE ... 43
         - MACH ... 44
         - IMPUTE v1 e v2 ... 44
         - Beagle ... 45
      2.1.2. Utilização de informações de parentesco ... 47
   2.2. Fatores que afetam a acurácia de imputação ... 49
      2.2.1. Software utilizado ... 49
      2.2.2. Tamanho da amostra ... 50
      2.2.3. Densidade dos marcadores ... 51
      2.2.4. Conexidade ... 52
      2.2.5. Frequência alélica do marcador ... 52
      2.2.6. Estrutura da população ... 53

2.3. Resultados da aplicação da imputação de genótipos em trabalhos de pesquisa... 53

Referências Bibliográficas _____ 55

CAPÍTULO 2

Efeito da imputação de genótipos sobre a predição de características quantitativas de ratos utilizando marcadores genéticos _____ **61**

Resumo 61

Abstract 61

Introdução 61

Material e Métodos 64

Resultados 71

Discussão 78

Conclusões 81

Referências Bibliográficas 81

LISTA DE TABELAS (Capítulo 1)

Tabela 1 - Acurácia para cada característica e modelo, correlação média sem validação cruzada para cada modelo, e QME para cada modelo (Fonte: Heslot et al., 2012) 40

LISTA DE TABELAS (Capítulo 2)

Tabela 1 - Distribuição de indivíduos para cada arquivo analisado 65
Tabela 2 - Acurácia de imputação e taxa de erro p/ 90, 75 e 50% de genótipos mascarados.. 71
Tabela 3 - Correlações entre peso corporal predito e mensurado utilizando LASSO Bayesiano (LB), Reproducing Kernel Hilbert Spaces (RKHS) e Redes Neurais Artificiais com Regularização Bayesiana para todas as taxas de mascaramento e distribuição de famílias para dados de teste 75
Tabela 4 - Correlações entre o índice de massa corporal predito e mensurado utilizando LASSO Bayesiano (LB), Reproducing Kernel Hilbert Spaces (RKHS) e Redes Neurais Artificiais com Regularização Bayesiana para todas as taxas de mascaramento e distribuição de famílias para dados de teste 76
Tabela 5 - Quadrado médio do erro de predição para análise do peso corporal utilizando o LASSO bayesiano (LB), Reproducing Kernel Hilbert Spaces (RKHS) e Redes Neurais Artificiais com Regularização Bayesiana (RNARB) de acordo com a distribuição de famílias entre dados de treinamento e dados de teste e taxa de mascaramento de genótipos 77
Tabela 6 - Quadrado médio do erro de predição para análise do índice de massa corporal utilizando o LASSO bayesiano (LB), Reproducing Kernel Hilbert Spaces (RKHS) e Redes Neurais Artificiais com Regularização Bayesiana (RNARB) de acordo com a distribuição de famílias entre dados de treinamento e dados de teste e taxa de mascaramento de genótipos ...78

LISTA DE FIGURAS (Capítulo 1)

Figura 1. O ancestral (1) deixa descendentes (2) e, a cada geração, o cromossomo dos ancestrais é quebrado por recombinação até que em uma última geração apenas um pequeno segmento do cromossomo idêntico ao do ancestral comum é conservado (3)..... 14
Figura 2. SNP 15
Figura 3. Priori para o efeito dos marcadores com média 0 e variância 1 25
Figura 4. Arquitetura de uma RNA com duas camadas contendo 5 neurons no hidden layer e um neuron no output layer 30
Figura 5. Ilustração do cálculo de δ_j para a unidade oculta j pela retropropagação dos $\delta's$ vindo das unidades k para as quais j envia conexões 33
Figura 6. Gradient descent em RNA. Generalização da aplicação da regra delta, em que a atualização dos pesos se dá em direção ao decréscimo da função de erro 34
Figura 7. Algoritmos para estimativa de parâmetros em RNA 35
Figura 8. Ilustração do early stopping 37
Figura 9. Correlações entre o fenótipo observado e fenótipo predito para modelo linear e RNA com diferentes arquiteturas 38
Figura 10. Correlações entre predições e observações para treinamento (r_{train}), teste (r_{test}), validação ($r_{validation}$) e total ($r_{y-\hat{y}}$) para arquiteturas linear e até 7 neurônios 39
Figura 11. Parâmetros probabilísticos de um HMM (Wikipédia) 42
Figura 12. Representação gráfica de HMM para imputação de genótipos 43
Figura 13. Esquema geral de como a imputação funciona 45
Figura 14. Ilustração ressaltando as principais diferenças entre modelos baseados em Li e Stephens (2003), a base de MACH, IMPUTE e fastPHASE, e o modelo de Browning (Browning, 2006), a base do Beagle 46
Figura 15. Imputação de genótipos para uma amostra de indivíduos aparentados 48
Figura 16. Acurácia de imputação em função do tamanho da amostra para diferentes metodologias (Browning e Browning, 2011) 51

LISTA DE FIGURAS (Capítulo 2)

Figura 1. Arquitetura de uma RNA com duas camadas contendo 5 neurônios na camada oculta e um neurônio na camada de saída..... 70
Figura 2. Correlações entre PC predito e mensurado entre famílias (A) e dentro de famílias (B) para o LASSO bayesiano (LB), Reproducing Kernel Hilbert Spaces (RKHS) e Redes Neurais Artificiais com Regularização Bayesiana (RNARB) para arquivos de teste contendo marcadores imputados e não-imputados (painel reduzido) contendo diferentes densidade de genótipos identificados..... 73
Figura 3. Correlações entre IMC predito e mensurado entre famílias (A) e dentro de famílias (B) para o LASSO bayesiano (LB), Reproducing Kernel Hilbert Spaces (RKHS) e Redes Neurais Artificiais com Regularização Bayesiana (RNARB) para arquivos de teste contendo marcadores imputados e não-imputados (painel reduzido) contendo diferentes densidade de genótipos identificados..... 74

CAPÍTULO 1

Resumo: Objetivou-se neste primeiro capítulo, definir conceitos básicos em genética quantitatica e molecular e revisar metodologias aplicadas para seleção genômica. Primeiramente modelos utilizados para a predição de efeitos de marcadores são descritos dentre eles regressões Bayesianas (BayesA, BayesB, Bayesian LASSO, entre outros), regressões semi-paramétricas (*Reproducing Kernel Hilbert Spaces*) e métodos não paramétricos (Redes Neurais com Regularização Bayesiana). Quanto aos métodos semi e não-paramétricos, eles têm sido propostos para seleção genômica com a vantagem de capturar efeitos não-lineares entre marcadores, o que seria impossível ajustando modelos lineares. Outra ferramenta também descrita é a imputação de genótipos, utilizada para preencher dados faltantes, unir bancos de dados prevenientes de diferentes painéis de marcadores ou mesmo aumentar a quantidade de SNP's contidos nos painéis.

Palavras-chave: Efeito de substituição, imputação, regressão, seleção genômica

Abstract: *The objective in this first chapter was to define basic concepts in quantitative and molecular genetics and to review methodologies that have been applied for genome-enabled prediction. First, models for markers effects prediction were described as Bayesian regressions (BayesA, BayesB, Bayesian LASSO, and others), semi-parametric regression (Reproducing Kernel Hilbert Spaces) and non-parametric methods (Bayesian Regularized Neural Networks). Regarding the semi and non-parametric methods, they have been proposed for genome-enabled prediction having the advantage of capturing non-linear effects between markers, which would be impossible fitting linear models. Another tool also described in this review is the genotype imputation that have been applied for fill missing data from the lab, merge data sets from different chips and even increase the number of SNP's contained in the chips.*

Keywords: Genomic selection, imputation, regression, substitution effect

INTRODUÇÃO

Na chamada "era genômica", criaram-se expectativas quanto às aplicações da engenharia genética que revolucionariam a maneira como programas de seleção seriam conduzidos por meio de clonagem, transgênicos, *Quantitative trait locos* (QTL), e Seleção Assistida por Marcadores (MAS) em substituição aos métodos tradicionais de avaliação genética, porém estes são pouco utilizados atualmente e têm contribuição baixíssima para o ganho genético em programas de seleção (Dekkers, 2004).

Entretanto, recentemente, a disponibilidade massiva de informações genômicas individuais de muitos animais despertou a idéia de criação de modelos funcionais que utilizem este tipo de informação para melhorar a predição do valor genético dos animais de características economicamente importantes. Ao contrário do que se desejava antes, que era a identificação de QTL's fazendo a localização de áreas específicas do genoma com genes favoráveis e usar marcadores que fossem capazes de identificar indivíduos geneticamente superiores, a seleção genômica, termo apresentado pela primeira vez por Meuwissen et al. (2001), objetiva a utilização de toda informação de polimorfismos (*Single Nucleotide Polymorphisms* - SNPs) que estejam em desequilíbrio de ligação com o QTL, inserindo-a como covariáveis no modelo preditivo como informação extra à avaliação genética tradicional. Como principais vantagens da seleção genômica citam-se o aumento na acurácia do valor genético predito e também a melhora na consistência da informação dos coeficientes de Wright da matriz de parentesco (Hayes et al., 2007).

A seleção genômica, também referida em artigos científicos como predição assistida por marcadores ("*genome-enabled prediction*") ou regressão com informação genômica ampla ("*whole genome regression*" – WGR), é tema relativamente recente no melhoramento genético de animais e plantas. Esta metodologia tem sido aplicada tanto para seleção de indivíduos geneticamente superiores (via predição de de valores genéticos utilizando dados de marcadores ao longo do genoma) quanto para decisões de manejo baseadas em fenótipos preditos (Meuwissen et al., 2001, Goddard e Hayes, 2007).

Diversos modelos de seleção genômica tem sido propostos que vão desde o método dos quadrados mínimos e BLUP (*Best Linear Unbiased Predictor*) até modelos de regressão com inferência bayesiana e semi ou não-paramétricos. E alguns destes serão discutidos ao longo desta monografia. É importante ressaltar que atualmente existe vasta literatura internacional abordando temáticas específicas dentro do tópico desta revisão, porém alguns trabalhos pioneiros ou de grande contribuição teórica serão utilizados para esta revisão como: Meuwissen et al, 2001, Goddard e Hayes, 2007, Gianola et al., 2009, de los Campos et al., 2012.

Portanto, é importante que se tenha senso crítico para a escolha do modelo a ser utilizado para seleção genômica, já que ainda não existe um modelo universalmente melhor para todas as situações. Além disso, a implementação da predição utilizando informações do genoma traz um número grande de questões como: Qual a maneira mais eficiente de incorporar a informação do genoma para predição? O quanto se ganha em habilidade de predição utilizando chips de maior densidade? A pré-seleção de marcadores melhora a predição? Quais indivíduos serão fenotipados? Quais e quantos indivíduos serão genotipados? WGR ou GWAS (Genome wide assisted selection? (Ober et al., 2012).

Ainda dentro do tópico seleção genômica, várias ferramentas têm sido propostas, e uma de grande importância é a imputação de genótipos. Na estatística, a palavra imputação se refere à substituição de algum valor para o dado perdido (Wikipédia). A imputação do genótipo é um termo usado para descrever o processo de predição ou imputação de genótipos que não são diretamente identificados na amostra de indivíduos (Marchini e Howie, 2010). Esta estratégia tem sido proposta para predição assistida por marcadores (*genome-enabled prediction*) para imputar painéis de SNPs de baixa densidade para alta densidade a fim de melhorar a acurácia e estimar efeito para marcadores usando todos os indivíduos disponíveis ao mesmo tempo (Weigel et al., 2010, Mulder et al., 2011). As vantagens da imputação irão depender de sua acurácia (Weigel et al., 2010), estrutura da população (Weigel et al., 2010, Dassonneville et al., 2011) e da arquitetura genética da característica avaliada (Moser et al., 2010).

Dentre as utilidades da imputação de SNPs não genotipados citam-se o preenchimento de dados não sequenciados pela plataforma de genotipagem para que se possa completar a matriz X (seleção genômica ou estudos de associação pangenômico – GWAS), combinação de banco de dados de painéis de SNP diferentes, imputar de chips de baixa densidade para alta densidade ou até mesmo sequência completa. A imputação é usada também para tirar vantagem do desequilíbrio de ligação na população para melhorar a probabilidade de identificar (*call*) corretamente os genótipos dos dados de sequência (Hayes, 2011). Existem vários métodos e softwares disponíveis para imputação, alguns considerando estrutura de famílias, outros apenas a informação de haplótipo obtida na população referência. Uma etapa importante, e diferenciada entre metodologias, para imputar os genótipos é a formação de haplótipos (denominada *phasing*). As suas aplicações não se restrigem apenas à imputação mas também auxiliam no entendimento do papel da variação genética e doenças, identificação de genótipos em microarranjo (microarray) e dados de sequência, inferir histórico demográfico de uma população, inferir pontos de recombinação, detectar mutação recorrente e assinaturas de seleção e modelar regulação cis da expressão gênica (Browning e Browning, 2011).

Diversos trabalhos de pesquisa têm mostrado que imputar os SNPs não genotipados e, combinando esta informação com os genótipos observados, causa aumento do poder estatístico para estudos de GWAS facilitando, assim, a detecção de mutações causais raras e aumento da acurácia/habilidade de predição para estimativa dos GEBV (valores genéticos genômicos).

Portanto, neste primeiro capítulo, objetivou-se definir conceitos básicos em genética de populações e genética molecular e revisar metodologias implementadas para seleção genômica como modelos de predição e imputação de genótipos.

REVISÃO DE LITERATURA

1. Seleção genômica

1.1.Genética molecular e de populações – Conceitos básicos

Antes de definir metodologias para seleção genômica é importante fazer uma pequena revisão em conceitos básicos da genética clássica de populações e avaliação do impacto dos métodos de seleção convencional aplicadas no aumento do potencial genético de produção.

O principal objetivo de um programa de seleção é causar melhora genética de uma população e, por conseqüência, aumento do lucro. O progresso genético anual pode ser avaliado com base na seguinte fórmula:

$$\Delta G = \frac{i\sigma_g r}{IG}$$

em que, *i* é a intensidade de seleção, σ_g é o desvio-padrão genético e *r* a acurácia de seleção, sendo *IG* o intervalo de geração. Portanto, qualquer ferramenta que aumente algum dos parâmetros do numerador ou cause diminuição no intervalo de geração trará maior resposta à seleção. Entretanto, alguns destes não são tão facilmente manipuláveis como o σ_g (inerente à população) e a intensidade de seleção, que quando muito alta, pode causar um atraso ainda maior no IG, em razão do atraso na seleção de características expressas depois da puberdade (Garrick, notas de aula). Sendo assim, o maior ganho anual pode ser gerado, principalmente, tanto pelo aumento no valor da acurácia de seleção quanto pela diminuição no intervalo de geração.

Outro conceito importante, um pouco esquecido em metodologias tradicionais de avaliação genética é o desequilíbrio de ligação (LD – r^2 ou D') (Curiosidade: o termo mais adequado seria desequilíbrio de fase gamética, pois se refere à fase haplóide do cromossomo) recebeu grande atenção em genética de populações no passado (Hill, 1974, Lewontin, 1988) e tem se tornado um tópico de grande importância para a pesquisa em razão da disponibilidade de dados genômicos. Conceitualmente, este termo se refere à associação estatística entre alelos em locos distintos (que inclusive podem estar ou não no mesmo cromossomo) em uma mesma população, ou seja, mede a habilidade de um alelo agir como marcador que pode ser utilizado para predizer algum outro alelo em particular. O LD ocorre pelo fato da possibilidade de existência de ligação física entre locos genômicos, que faz com que sejam herdados em conjunto, ou seja, pode haver desequilíbrio entre locos não ligados, porém é mais comum estar

presente (e persistente) entre locos ligados. O LD pode ser influenciado por vários fatores como a distância física (em centimorgans ou pares de base) entre os locos, seleção, taxa de recombinação, mutação, acasalamentos preferenciais e estrutura histórica da população (existência ou não de "*bottlenecks"*). Por exemplo, o LD em populações de bovinos é bem maior do que em humanos, em razão da maneira como esses grupos foram historicamente formados. Se alelos de dois locos estão associados por alguma ligação física, o equilíbrio (independência) é eventualmente atingido assintoticamente com o passar das gerações de acasalamento ao acaso (Gianola et al., 2012).

Na utilização de marcadores, a aplicação da genética molecular no melhoramento animal se baseia na capacidade de genotipagem de indivíduos para locos genéticos específicos. Neste contexto, segundo Dekkers (2004), três tipos de polimorfismos genéticos nos locos podem ser identificados: 1) marcadores diretos: locos que fazem codificação para a mutação funcional (mutações causais); 2) marcadores LD: locos em desequilíbrio de ligação com a codificação da mutação funcional em toda a população; 3) marcadores em equilíbrio de ligação (LE): locos em equilíbrio de ligação com a codificação da mutação funcional em populações divergentes. Estes três tipos de marcadores diferem não só pelo método de detecção como também pela sua aplicação em programas de seleção. Marcadores diretos e marcadores LD podem ser utilizados para seleção de genótipos entre populações em razão da associação entre genótipo e fenótipo. O mapeamento de desequilíbrio de ligação (LD) do QTL explora bem os níveis de associação entre marcadores e QTL. Estas associações aparecem porque existem pequenos segmentos do cromossomo na população que têm origem no mesmo ancestral comum. Estes segmentos de cromossomo, que levam ao mesmo ancestral comum sem intervenção de recombinação, carregarão marcadores alélicos idênticos ou haplótipos e, se existe algum QTL neste mesmo segmento, ele também será carregado (Hayes, 2007).

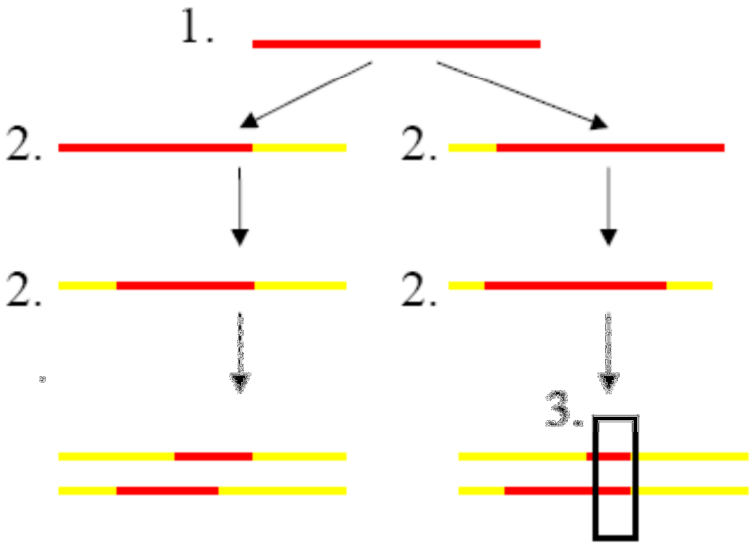

Figura 1. O ancestral (1) deixa descendentes (2) e, a cada geração, o cromossomo dos ancestrais é quebrado por recombinação até que em uma última geração apenas um pequeno segmento do cromossomo idêntico ao do ancestral comum é conservado (3). (Fonte: Hayes, 2007)

O cálculo clássico do LD, proposto por Lewontin e Kojima (1960), reflete a diferença entre as frequências esperadas dos haplótipos sob a pressuposição de independência, e a frequências observadas dos haplótipos. Entretanto, em razão da grande atenção que este tópico tem recebido, novas maneiras de expressar esta dependência entre alelos tem sido propostas (Mangin et al., 2011, Gianola et al., 2012, Morota et al., 2012).

É importante salientar que o LD é continuamente quebrado pela recombinação, e muda com o tempo, de acordo com a seguinte fórmula: Dt=Do(1-c)t, em que *c* é a taxa de recombinação e *t* o tempo. Sendo assim, a ligação marcador-QTL não é eterna, demonstrando que mesmo que a informação genômica traga grandes benefícios para o melhoramento, o conhecimento a respeito dos efeitos de SNPs deverá ser continuamente atualizado (claro que com base nas informações fenotípicas mensuradas).

Dentro da genética molecular é importante revisitar alguns conceitos como marcador molecular ou genético, que nada mais é que um gene ou fragmento do DNA em um certo ponto do genoma em um cromossomo, que pode ser utilizado para identificar indivíduos ou espécies (Wikipédia). Exitem vários tipos de marcadores, porém nesta revisão será abordado apenas o SNPem razão da sua extensa aplicação em WGR.

O SNP é uma variação na sequência de DNA que ocorre quando um único nucleotídeo – A, T, C ou G – no genoma difere entre indivíduos de uma espécie (Wikipédia). Ou seja, o marcador

SNP pode ser definido como a diferença em uma única base nucleotídica de mesma posição em diferentes animais, podendo causar mudança ou não (mutação silenciosa) na codificação protéica. Como exemplo, tem-se a sequência de bases para dois animais, representada na Figura 2. :

Figura 2. SNP Animal 1. AACCTAAAAC

Animal 2. AACTTAAAAC

Neste caso, existe diferença nas bases (ou alelos) C e T entre os animais. O estudo envolvendo este tipo de marcador permite a avaliação de polimorfismos ao longo do genoma responsáveis por variação da característica observável em diferentes animais.

Outro termo comum é haplótipo, que designa um conjunto de SNPs que possuem associação estatística, formando um bloco. Esta associação é relevante visto que a identificação de apenas alguns SNPs em uma sequência pode revelar outros SNPs não genotipados, se os possíveis haplótipos são conhecidos, o que tem sido a base para programas que fazem imputação do genótipo.

1.2. WGR – Modelos e dimensionalidade

A fim de superar as dificuldades encontradas pelos métodos de avaliação genética com informação do genoma, Meuwissen et al. (2001) propuseram uma variante da seleção assistida por marcadores denominada “seleção genômica”. A grande diferença deste método de seleção é a utilização de marcadores que cubram o genoma de maneira ampla para que se possa explicar o máximo possível da variância genética pelos marcadores. O modelo geral é representado por:

$$y = \mu 1_n + \sum_i X_i g_i + e$$

em que y representa fenótipo, μ é a média, X_i é a matriz de delineamento que aloca animais ao efeito do haplótipo, SNP ou segmento de cromossomo contendo informação dos marcadores e g_i o vetor dos seus respectivos efeitos, e e representando o resíduo.

O mais importante desta ferramenta de seleção é que tem a propriedade desejável de estimar o efeito dos segmentos de cromossomo de forma simultânea, que evita problemas com superestimação dos efeitos do QTL em decorrência dos testes múltiplos (Hayes, 2007).

Segundo Hayes (2007), a implementação da seleção genômica, conceitualmente, corre em dois passos:

1) Estimativa dos efeitos dos segmentos de cromossomo na população referência;
2) Predição dos valores genéticos genômicos (GEBV's – *"Genomic estimated breeding values"*) para animais que não são da população referência.

Para predição dos GEBV's é feito o somatório dos efeitos de cada segmento do cromossomo ao longo de todo o genoma:

$$GEBV = \sum_i^n X_i \hat{g}_i$$

Em que n é o número de segmentos de cromossomo, X_i é a matriz de delineamento que aloca animais ao efeito do haplótipo no segmento i; e $\hat{g}_i$ é o vetor de efeitos dos haplótipos dentro do segmento de cromossomo i.

A grande dificuldade está no primeiro passo, pois em razão da disponibilização de painéis de SNP de alta densidade, o número de marcadores (p) pode exceder em muito o número de dados (n), e o ajuste de regressões deste tipo requer a utilização de algum tipo de seleção de variáveis ou procedimento de estimativa com regularização/encurtamento (*shrinkage)* ou a combinação de ambos (de los Campos, 2012). Os modelos de predição podem ser divididos em duas classes: aqueles que utilizam sistema de equações linear e o não-linear, sendo que no primeiro todos os marcadores contribuem igualmente para a variância genética (ou seja, não existem genes de efeito maior), por outro lado, o segundo assume que a distribuição a priori dos marcadores ou o efeito dos QTL's não é normal. Claro que a variância genética pode não ser igual para todos os segmentos de cromossomo ou marcadores porque, por exemplo, genes de efeito maior podem existir em alguns cromossomos. Em ambas as metodologias o vetor de informações é modelado como função linear dos efeitos não conhecidos, entretanto as soluções para os efeitos no contexto da predição não linear são funções não-lineares do vetor de dados (VanRaden, 2008).

1.2.1. Quadrados mínimos e GBLUP

Meuwissen et al. (2001) apresentam o método dos quadrados mínimos e BLUP para inferência dos valores genéticos a partir de informações genômicas. No primeiro, o efeito dos haplótipos (SNPs - gi) é considerado fixo enquanto que no segundo o haplótipo tem efeito aleatório, com distribuição normal e variância homogênea para diferentes segmentos de cromossomo. O seguinte modelo é adotado por estas abordagens:

$y = 1\mu + \sum_i Xg_i + e$

onde somatório $\sum_i$ é feito ao longo dos segmentos do genoma (cM). E a obtenção das predições de g_i é feita também pela solução das equações de modelos mistos no caso do BLUP.

Para o método dos quadrados mínimos, é impossível, por exemplo, estimar o efeito de 50.000 haplótipos com apenas 200 observações fenotípicas simultaneamente, portanto faz-se necessária a utilização de método de seleção de segmentos de cromossomo ("*stepwise"*) que tenham contribuição significativa para a variação da característica. Neste caso, em razão do limiar imposto, somente QTL's com maior efeito são detectados e considerados no modelo de forma que a variância genética total não é capturada pelos marcadores (Meuwissen et al., 2001).

A solução para os quadrados mínimos ordinários (OLS) de g é dada por:

$$\hat{g}_{OLS} \underset{min}{=} \sum_i (y_i - \sum_j x_{ij} g_j)^2$$

onde $\sum_i (y_i - \sum_j x_{ij} g_j)^2$ é a soma de quadrados do resíduo. O estimador é então dado por $\hat{g}_{OLS} = [X'X]^{-1}X'y$.

Já o quadrado médio do erro do estimador (QME) é : $QME(\hat{\theta}) = E\left[(\theta - \hat{\theta})^2\right]$ onde θ é o valor verdadeiro do parâmetro e $\hat{\theta}$ é o estimador, o qual é função dos dados. A esperança do QME na fórmula é dado com respeito a todas as amostras dos dados. Normalmente X é tratado como fixo e a esperança é dada somente com respeito às possíveis realizações de y dado X.

O QME pode ser decomposto em dois componentes: $QME(\hat{\theta}) = [\theta - E(\theta)]^2 + Var(\hat{\theta})$, onde $[\theta - E(\theta)]$ e $Var(\hat{\theta})$ são o viés e a variância do estimador, respectivamente.

A esperança da estimativa OLS para os coeficientes de regressão em [1] é:

$$E\lfloor\hat{g}_{OLS}|X\rfloor = [X'X]^{-1}X'E[y]$$
$$= [X'X]^{-1}X'E[Xg + e]$$
$$= [X'X]^{-1}X'Xg + [X'X]^{-1}X'E[e]$$
$$= g + [X'X]^{-1}X'E[e]$$

se o modelo [1] está correto, $E[e] = 0$, portanto: $E\lfloor\hat{g}_{OLS}|X\rfloor = g$. Em palavras, se o modelo linear apresentado em [1] é aplicado, OLS gera estimativa não viesada dos coeficientes de regressão. O termo $Var(\hat{\theta})$, no contexto frequentista, é a medida de incerteza e reflete a variabilidade do estimador em amostragens repetidas. Entretanto, quando p>n a estimativa OLS não é única por $X'X$ ser singular. Portanto, para seleção genômica outros modelos são necessários.

O RRBLUP (Random Regression BLUP) pode ser utilizado como maneira de confrontar o problema p>>n, penalizando estimativas dos coeficientes de regressão. A principal ideia nestes modelos com penalização é reduzir o QME pela redução da variância do estimador, mesmo que introduza algum viés. Considerando o mesma equação [1], as estimativas do RRBLUP possuem forma similar ao OLS, tendo adicionada uma constante à diagonal da matriz de coeficientes, como segue:

$$[X'X + \lambda D]\hat{g}_{RR} = X'y$$

onde λ é a constante e D é a diagonal da matrix com zero na sua primeira entrada da diagonal (a fim de evitar encurtamento da estimativa do intercepto e 1's no restante da diagonal). Quando λ é igual a zero, a solução é OLS. Inserindo a constante às entradas da diagonal da matriz de coeficientes, torna a mesma não-singular e encurta as estimativas dos coeficientes de regressão em direção a zero. Este artifício introduz viés, porém reduz a variância das estimativas, como já foi dito anteriormente; no caso p>>n isto pode reduzir estimativas do QME e pode gerar predições mais acuradas. Outros estimadores que aplicam penalização serão discutidos no avançar desta revisão.

Derivando, no contexto bayesiano, a esperança condicional do efeito dos marcadores e dos valores genômicos, tem-se o melhor preditor no sentido do QME.

Verossimilhança: $[y|g, \sigma_e^2] \sim N(Xg, I\sigma_e^2)$

Priori: $[g|\sigma_g^2] \sim N(0, D^{-1}\sigma_g^2)$

A distribuição a posteriori de g é normal multivariada com média igual à solução do seguinte sistema: $[X'X + \lambda D]\hat{g}_{RR} = X'y$; estas são as esquações da regressão de cumieira e também do BLUP de g dado y. Ressaltando que a razão $\frac{\sigma_e^2}{\sigma_g^2}$ é equivalente ao λ no RR. Se o modelo é totalmente Bayesiano pode-se atribuir prioris a cada um desses parâmetros de variância, o que permite que sejam inferidos dos mesmo dados utilizados para a estimativa de efeito dos marcadores.

A esperança condicional dos valores genômicos dado y é:

$$E[Xg|y, \sigma_e^2\ ] = XE[g|y, \sigma_e^2]$$
$$= XX'[XX + \lambda I]^{-1}y$$
$$= [I + \lambda G^{-1}]^{-1}y$$

onde G=XX'. Este é o chamado GBLUP dos valores genômicos.

A formação da matriz de parentesco genômico (G , definição no trabalho de VanRaden (2008) que pode gerar confundimento com a matriz de variância e covariância genética aditiva em soluções de equações de modelos mistos) pode ser feita de três maneiras:

A primeira, proposta por Legarra e Miztal (2008), inicialmente é necessária a obtenção da matriz M que especifica quais alelos marcadores cada indivíduo herdou e tem dimensão n x m (em que n se refere ao número de indivíduos e m ao número de marcadores). Equações podem incluir informação do marcador fazendo-se M*M' obtendo-se matriz nxn em que os elementos da diagonal são o número de locos homozigotos para cada indivíduo e fora da diagonal dá a medida do número de alelos compartilhados por indivíduos aparentados, considerando que M tem valores -1, 0 e 1 correspondente ao homozigoto, heterozigoto e outro homozigoto. Já para M'*M (matriz mxm) a diagonal informa o número de indivíduos homozigotos para cada loco e fora da diagonal tem-se a medida do número de vezes que alelos em locos distintos foram herdados pelo mesmo indivíduo.

A matriz P, que também é necessária no processo de obtenção de G, contém as freqüências alélicas expressas como a diferença entre a freqüência de um alelo e 0,5, posteriormente multiplicado por 2 sendo, então, a coluna *i* de P igual a 2(pi – 0,5) (nesta caso a coluna se refere ao loco).

A subtração da matriz M pela matriz P, nos dá a matriz Z que coloca os valores médios dos efeitos dos alelos em zero. Esta subtração dá mais crédito para alelos raros do que para alelos comuns no cálculo do parentesco genômico (VanRaden, 2008). Também considera que a endogamia é mais alta se o indivíduo é homozigoto para alelos raros.

A primeira forma de obtenção de G é, então, feita por:

$$G = \frac{ZZ'}{2\sum p_i(1-p_i)}$$

O denominador $2\sum p_i(1-p_i)$ coloca G na mesma escala de A. Para conhecer o grau de endogamia basta pegar o elemento G_{kk} e subtrair 1. Para saber o parentesco entre dois indivíduos utiliza-se a mesma fórmula da correlação neste caso expresso como $\frac{G_{jk}}{\sqrt{G_{jj}G_{kk}}}$.

A segunda forma de calcular a matriz G foi proposta em humanos (Amin et al. 2007) é feita pela ponderação dos marcadores pela transposta de suas variâncias esperadas. A fórmula é expressa como: $G = ZDZ';\ onde\ D\ é\ diagonal\ com\ D_{ii} = \frac{1}{m[2p_i(1-p_i)]}$.

O terceiro método para obter G não requere as freqüências alélicas e, ao invés disto, ajusta pela homozigose média pela regressão de MM' sobre A (que é o valor esperado de G) para obter G utilizando o seguinte modelo:

$$MM' = g_0 11' + g_1 A + E,$$

onde g_0 é o intercepto e g_1 é a inclinação. A matriz E inclui a diferença entre as frações esperadas de DNA em comum mais a medida de erro já que a sequência completa de DNA não está disponível e sim, apenas uma amostra de marcadores genotipados.

Revertendo a equação para o cálculo, tem-se:

$$\frac{MM' - g_0 11'}{g_1} = G$$

E o sistema de equações, pode ser representado por:

$$\begin{bmatrix} n^2 & \sum_j \sum_k A_{jk} \\ \sum_j \sum_k A_{jk} & \sum_j \sum_k A_{jk}^2 \end{bmatrix} \begin{bmatrix} g_0 \\ g_1 \end{bmatrix} = \begin{bmatrix} \sum_j \sum_k (MM')_{jk} \\ \sum_j \sum_k (MM')_{jk} A_{jk} \end{bmatrix}$$

A matriz G é positiva semidefinida com os dois primeiros métodos, mas pode ser singular se o número de locos é limitado ou se dois indivíduos possuem genótipos idênticos.

Um ponto importante, abordado por Forni et al. (2011), é que a estimativa dos parâmetros pode ser viesada se os coeficientes de parentesco genômico estão em escala diferente do que os coeficientes de parentesco baseado no pedigree e mostra que um escalonamento pode ser feito utilizando a freqüência alélica observada e o re-escalonamento da matriz de parentesco genômico para se obter diagonal média igual a 1.

Também existe um método de se obter a matriz G melhorada ou G^* por meio de média ponderada, em que:

$$G^*=wG+(1-w)A$$

Sendo que w = A#2 (cada elemento de A elevado ao quadrado) (normalmente w é igual a 1). No caso em que o número de marcadores é limitado e A é não singular.

$$w = \frac{0{,}05^2}{(0{,}05^2 + \frac{0{,}125}{m})}$$

em que m se refere ao número de marcadores.

VanRaden (2008) compara os três métodos citados acima e conclui que o terceiro foi o melhor, porém as condições testadas não consideravam que os alelos na população base pode não estar em equilibrio. O autor também mostra que o tempo para geração desta matriz é aumentado pelo número de marcadores multiplicado pelo número de touros ao quadrado e foi de aproximadamente 10 segundos por touro.

Forni et al. (2011) avaliaram diferentes maneiras de se formar a matriz G para o "*single-step GBLUP"*: 1) Utiliza frequências alélicas iguais a 0,5 (G05); 2) Frequência alélica igual à média da menor frequência (GMF); 3) Frequência alélica observada (GOF); 4) Matriz G considerando determinação aleatória das frequências alélicas (GOF*) e; 5) Matriz normalizada com diagonal igual a 1 (GN). Os autores concluem que as acurácias dos valores genéticos com informação do genoma foram inflacionados exceto para 5.

1.2.2. Modelos bayesianos: o Alfabeto e o LASSO

Na inferência bayesiana, a probabilidade é usada para quantificar a incerteza ou conhecimento a respeito dos possíveis valores de parâmetros desconhecidos. Em sua essência, probabilidades *a priori* quantificam a incerteza a respeitos dos parâmetros antes dos dados serem analisados, até que os parâmetros são relacionados aos dados por meio do modelo ou verossimilhança, que é a densidade probabilidade condicional dos dados *dados* os parâmetros. A *priori* e a verossimilhança são combinadas utilizando o teorema de Bayes para obtenção das probabilidades *a posteriori*, que são as probabilidades condicionais dos parâmetros dados os dados. Os métodos bayesianos têm se mostrado muito úteis, eficientes e flexíveis para a análise de dados genômicos, porém com algumas inconveniências para implementação computacional (Kärkkäinen e Sillampaa, 2012).

Meuwissen et al. (2001) apresentam dois modelos hierárquicos, chamados Bayes A e Bayes B, que são amplamente discutidos em artigos de pesquisa envolvendo animais e plantas. O método Bayes A trata do mesmo modelo representado em [1] para o método dos quadrados mínimos e GBLUP, porém como regressão bayesiana considerando distribuição normal como

distribuição a priori dos efeitos de haplótipo. Adicionalmente a variância σ^2_{gi} dos diferentes segmentos de cromossomo pode ser diferente, isto é, a priori considera-se que muitos QTLs têm efeito pequeno e alguns têm efeito maior. A distribuição a priori deste parâmetro é uma qui-quadrado invertida com parâmetro de escala p(σ^2_{gi}) ~ χ^{-2} (v,S) em que *S* é o parâmetro de escala e *v* é o número de graus de liberdade. Esta é uma escolha conveniente porque quando a informação da distribuição a priori é combinada com os dados a resultante é uma distribuição a posteriori também de qui-quadrado invertido com parâmetro de escala (Wang et al., 1993). O problema desta metodologia é o mesmo das anteriores em que não é possível atribuir valores nulos para efeitos de segmento de cromossomo. O modelo linear proposto é:

$$y_i = \mu + \sum_{j=1}^{p} x_{ij} b_j + e_i \ ,$$

em que i=1,2,...,n e j=1,2,..,m, sendo x_{ij} o código para o genótipo do SNP j (x=-1,0,1) e b_j o efeito aditivo de substituição do SNP j. Lembrando que n<<p. A distribuição condicional dos dados é dada por:

$$y_i|\mu, x_i, b, \sigma_e^2 \sim N\left(\mu + \sum_{j=1}^{p} x_{ij} b_j \, , \sigma_e^2\right)$$

Tendo como prioris:

$\mu \sim uniform$;

$\sigma_e^2 \sim v_e S_e^2 \chi_{v_e}^{-2}$;

$b_j \sim N\left(0, \sigma_{b_j}^2\right); j = 1,2, \dots, p$;

$\sigma_{b_j}^2 \sim vS^2\chi_v^{-2}$ para todo j.

Como proposto por Meuwissen et al. (2001)

$b_j|\sigma_j^2 \sim N\left(0, \sigma_j^2\right)$ sendo j=1,2,...,p

$\sigma_j^2|v, S^2 \sim vS^2\chi_v^{-2}$ (Notando que cada SNP tem uma variância)

A priori para efeito de marcador é uma distribuição de t com escala e graus de liberdade conhecidos. E segundo Gianola et al. (2009), marginalmente, todos os marcadores possuem a mesma variância, que pode ser demonstrado pela integração a densidade normal ao longo de σ_j^2:

$$p(b_j|v,S^2) = \int_0^\infty N(0,\sigma_j^2)\, p(\sigma_j^2|v,S^2)\sigma_j^2$$

$$\propto \int_0^\infty (\sigma_j^2)^{-\left(\frac{1+v+2}{2}\right)} exp\left(-\frac{\sigma_{b_j}^2 + vS^2}{2\sigma_j^2}\right) d\sigma_j^2$$

$$\propto \left(1+\frac{\sigma_j^2}{vS^2}\right)^{-\left(\frac{v+1}{2}\right)} \rightarrow t(0,v,S^2)$$

Apesar de alguns problemas apontados por Gianola (notas de aula) como influência forte da priori, impossibilidade de aprendizado a respeito da variância dos efeitos de marcador, o Bayes A pode fornecer boa habilidade de predição em cenários em que o modelo aditivo é predominante.

Outra forma bayesiana alternativa para seleção genômica seria o BayesB que tem as seguintes pressuposições:

$$b_j|\sigma_j^2 \sim \begin{cases} ponto\ de\ alta\ densidade\ em\ uma\ constante\ k\ se\ \sigma_j^2 = 0 \\ N(0,\sigma_j^2)\ se\ \sigma_j^2 > 0 \end{cases}$$

$$\sigma_j^2|\pi = \begin{cases} 0\ com\ probabilidade\ \pi \\ vS^2\chi_v^{-2}\ com\ probabilidade\ 1-\pi \end{cases}$$

Meuwissen et al. (2001) assumem que π é conhecido. Lembrando que se a variância a priori é 0, significa certeza absoluta (Gianola, notas de aula). Portanto, neste caso, a priori é uma mistura de um ponto de alta densidade em 0 e uam distribuição de t.

Outros modelos tem sido propostos como o Bayes-C0 (Fernando e Garrick, 2009),é semelhante ao BLUP, pois considera a mesma variância para todos os SNP's. E o Bayes-Cπ que é extensão do modelo anterior sendo uma mistura de modelos em que permite estimar simultaneamente partes dos dados com diferentes valores, no caso do Bayes C0 o termo π é igual a 0. E por aí vai… A diferença entre todos estes Bayes-() está na priori. Em teoria, o que melhor se aproximasse da explicação da realidade, ou seja, como o genótipo gera o fenótipo, seria o "escolhido".

Outro modelo bastante abordado na literatura é o LASSO (*Least Absolute Shrinkage and Selection Operator*; TIBSHIRANI, 1996) bayesiano proposto para seleção genômica por de los Campos et al. (2009) que permite forma de regularização ainda mais adequada para situações em que a maioria dos marcadores apresenta efeito igual a 0. Tibshirani (1996) propôs este método de regressão que combina seleção de variáveis e encurtamento baseado em regularização ao mesmo tempo. Neste modelo, uma penalidade proporcional à norma dos

coeficientes de regressão é adicionada à fórmula de otimização, permitindo que a seleção de variáveis e encurtamento ocorram simultaneamente. A versão bayesiana do LASSO foi proposta por Park e Casella (2008) que descreveram a implentação no amostrador de Gibbs. O problema de otimização pode ser visto como:

$$\min_{\beta}\left\{\sum(y_i - x'\beta)^2 + \lambda\sum_j \|\beta\|_j^2\right\},$$

onde a primeira parte se refere à solução dos quadrados mínimos (já apresentada anteriormente) e a segunda é o fator de penalização λ (maior que zero). O maior valor de λ indica maior encurtamento e alguns β's podem ser até zerados.

Na interpretação Bayesiana do LASSO, a solução pode ser vista como a moda *a posteriori* de um modelo Bayesiano com verossimilhança gaussiana, $p(y|\beta,\sigma_e^2) = \prod_{i=1}^{n} N(y_i|x_i'\beta,\sigma_e^2)$ e a priori de β que é o produto de p densidades independentes, com média zero e exponenciais duplas; dado por $p(\beta|\lambda) = \prod_{j=1}^{p}\left(exp\left(-\left(\frac{\beta_j^2}{2\sigma_\beta^2}\right)/\sqrt{2\pi\sigma_\beta^2}\right)\right)$ em que o componente σ_β^2 é comum a todos os coeficientes de regressão (de los Campos et al., 2009). O parâmetro λ, também chamado de parâmetro de suavização, controla o balanço entre a qualidade de ajuste e a complexidade do modelo. Este tem papel central no modelo, a medida que seu valor se aproxima de zero, a solução se aproxima da solução dos quadrados mínimos; enquanto que um alto valor proporciona encurtamento mais forte e priori mais precisa para β.

A distribuição exponencial dupla (DE) (ou Laplace) possui representação hierárquica conveniente como mistura de densidades Gaussianas escalonadas (Rosa et al., 2003). De acordo com Park e Casella (2008):

$$\beta_j \sim DE(\beta_j \mid \lambda) = \frac{\lambda}{2}e^{-\lambda|\beta_j|}$$

$$= \int_0^\infty \left[\frac{1}{\sqrt{2\pi\sigma_j^2}}e^{-(\beta_j^2/2\sigma_j^2)}\right]\left[\frac{\lambda^2}{2}e^{-\lambda^2/2\sigma_j^2}\right]d\sigma_j^2,$$

e a representação completa do modelo, seguindo de los Campos et al. (2009):

Verossimilhança: $p(y|\beta,\sigma_\varepsilon^2) \sim \prod_{i=1}^{n} N(y_i|x_i'\beta,\sigma_\varepsilon^2)$

$$\text{Priori:} \quad \begin{cases} p(\beta, \sigma_\varepsilon^2, \tau^2, \lambda^2) = p(\beta \mid \sigma_\varepsilon^2, \tau^2)\, p(\sigma_\varepsilon^2)\, p(\tau^2 \mid \lambda)\, p(\lambda^2) \\ = \left[\prod_{j=1}^{p} N(\beta_j \mid 0, \tau^2 \sigma_\varepsilon^2) \right] \chi^{-2}(\sigma_\varepsilon^2 \mid g.l., S) \\ x \left[\prod_{j=1}^{p} \exp(\tau_j^2 \mid \lambda) \right] G(\lambda^2 \mid \alpha_1, \alpha_2) \end{cases}$$

onde, $N(y_i \mid x_i'\beta, \sigma_\varepsilon^2)$ e $N(\beta_j \mid 0, \tau^2\sigma_\varepsilon^2)$ são as densidades normais com médias $x_i'\beta$ e 0, e variâncias σ_ε^2 e $\tau^2\sigma_\varepsilon^2$ atribuídas ao fenótipo e ao efeito dos marcadores, respectivamente; distribuição qui-quadrada invertida com parâmetro de escala $\chi^{-2}(\sigma_\varepsilon^2 \mid g.l., S)$ com g.l. graus de liberdade e S como parâmetro de escala é utilizada para o resíduo; uma distribuição exponencial dupla $\exp(\tau_j^2 \mid \lambda)$ indexada por λ e $G(\lambda^2 \mid \alpha_1, \alpha_2)$ é a distribuição gama com parâmetro de forma α_1 e parâmetro de taxa α_2. Este modelo possui a vantagem de permitir maior densidade para marcadores com efeito zero o que é mais biologicamente factível (Figura).

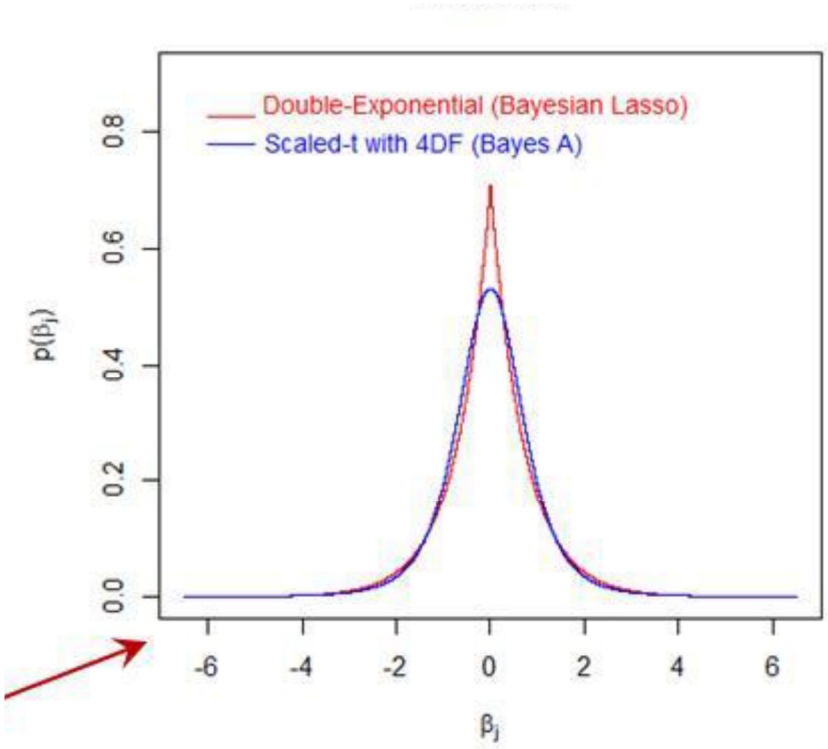

Figura 3. Priori para o efeito dos marcadores com média 0 e variância 1 (de los Campos et al., 2012)

1.2.3. WGR – Modelos semi e não-paramétricos

O interesse em modelos não paramétricos para a predição de características complexas tem aumentado. Literatura em Espaço de Hilbert reproduzido por Kernel (*Reproducing Kernel Hilbert Spaces – RKHS*) (de los Campos et al., 2009a, de los Campos et al., 2010, Gianola et al., 2008), funções de base radial (*Radial Basis Functions – RBF*) (Long et al., 2010, González-Camacho et al., 2012) e redes neurais (*Artificial Neural Network – ANN*) (Okut et al., 2011, Gianola et al., 2011) pode ser encontrada. Gianola et al. (2011) justificam que essas regressões não-paramétricas podem fornecer *insights* sobre a arquitetura genética pois podem capturar não-linearidades, o que seria impossível com regressões bayesianas lineares.

1.2.3.1.RKHS

A teoria de RKHS foi desenvolvida por Aronszajn (1950) e tem sido aplicada em vários ramos da estatística e aprendizado de máquinas por muitos anos, a base teórica é dada em Wahba (1990). Esta regressão semi-paramétrica dos fenótipos sobre os genótipos foi proposta por Gianola et al. (2006) e Gianola e Kaam (2008). O RKHS tem a propriedade de possuir um espaço infinito de funções para busca da dependência entre o *input* e a variável *target* definido pela medida de distância utilizada (neste caso, o tipo de kernel) sem nenhuma premissa adicional a respeito da relação fenótipo-marcador. O método consiste da combinação do modelo aditivo clássico com uma função desconhecida dos marcadores, que pode ser inferida não-parametricamente, e tem o potencial de capturar possíveis interações sem a necessidade de modelá-las explicitamente (Gianola et al., 2006). Para mapear a relação entre *input* (genótipos) e os *targets* (fenótipos), uma coleção de funções no espaço de Hilbert (diz-se $f \in H$) é utilizada na qual um elemento, $\hat{f}$, é escolhido baseado em algum critério (soma de quadrado do resíduo penalizada ou densidade *a posteriori*, por exemplo) (de los Campos et al., 2010). O problema de otimização para obtenção das estimativas no RKHS é similar à descrita para o BL, que é:

$$\hat{f} = \min_{f \in H}\{l(f, y) + \lambda \|f\|_H^2\},$$

onde $l(f, y)$ é a função de perda referente a alguma medida de qualidade de ajuste; e $\|f\|_H^2$ é a norma de f ao quadrado referente à complexidade do modelo; e λ controla o balanço entre qualidade de ajuste e complexidade do modelo.

De acordo com o teorema de Moore-Aronszajn, cada RKHS é associado a um único kernel positivo definido. Em RKHS os marcadores são utilizados para construir a matriz de covariância que mede a distância de genótipos em que $Cov(g_i, g_j) \sim K(x_i, x_j)$ em que x_i e x_j são os vetores contendo os genótipos para os indivíduos i e j, e $K(.,.)$ é o *Reproducing Kernel* (RK) relacionado com a função positiva definida (de los Campos et al., 2010).

A matriz de *kernel* pode ser definida de várias formas, a maneira mais comum é utilizar o kernel Gaussiano em razão da sua facilidade de implementação (Gianola, notas de aula), como $K(x_i, x_j) = exp\{-h \times d_{ij}\}$ em que h é o parâmetros de largura de banda e d_{ij} é a matriz de distância Euclidiana, dada por $d_{ij} = \sum_{k=1}^{p}(x_{ik} - x_{jk})^2$. A escolha de h é feita por comparação de modelos e deve considerar a distribuição observada em d_{ij}. O parâmetro de largura de banda é livre e pode ter forte influência sobre a predição, já que um valor muito baixo causa perda de informação (suavização excessiva, tendendo a uma distribuição uniforme) e muito alto gera dados irrelevantes. Uma forma automática de se escolhar o kernel é utilizando o média de kernéis (*kernel averaging- KA)* baseada na mediana amostral de d_{ij} . Crossa et al. (2010) descrevem uma maneira de KA em que $h = a \times q_{0.5}^{-1}$ em que a foi igualado a -5, -1 e -1/5 e $q_{0.5}$ é a mediana amostral de d_{ij}.

1.2.3.2.Redes Neurais Artificiais

Outro metodologia utilizada para predição utilizando dados genômicos são as Redes Neurais Artificiais (RNA). Mais especificamente sua implementação bayesiana com regularização tem sido aplicada (*Bayesian Regularized Neural Networks – BRANN*). A BRANN é uma rede *feed-forward* que envolve uma máxima *a posteriori* em que o regularizador pode ser visto como o logaritmo da disttribuição *a priori* (Bishop, 2006). Portanto, este modelo atribui uma distribuição de probabilidade para os pesos (*weights)* e vieses para que as predições sejam feitas em um contexto Bayesiano. Detalhes são apresentados por Mackay (1992).

1.2.3.2.1. Histórico

A RNA é um modelo matemático que tenta simular a estrutura e funcionalidades das redes neurais biológicas. McCulloch e Pitts (1943) são reconhecidos como os autores da primeira rede neural artificial. Mais tarde, grande avanço nesta área veio com a publicação do livro "The organization of the behavior" por Donald Hebb em 1949, que reforçou a teoria desenvolvida por McCulloch-Pitts sobre os neurônios e como eles funcionavam. Hebb aborda vários pontos importantes para a evolução das RNA, como a "fortificação" das redes cada vez que são usadas, o que ilustra a importância do treinamento das redes.

Na década de 50 e 60, vários pesquisadores começam a trabalhar com o *perceptron* (algoritmo de classificação supervisionada desenvolvido por Rosenblatt em 1958), que foi a primeira aplicação prática da rede neural artificial. E, em 1969, Minsky e Papert mostram que os perceptrons são limitados (o clássico problema XOR), o que faz com que a pesquisa em RNA enfraqueça.

Apenas no início da década de 80, pesquisadores renovam o interesse para a pesquisa em redes neurais. Neste período foi desenvolvida a teoria essencial para

implementação das redes neurais artificiais como o processamento em paralelo e backpropagation para atualização das ponderações (Rumelhart e McClelland, 1986) .

1.2.3.2.2. Redes Neurais Artificiais: Definições e Estrutura

O *neurônio* artificial é a unidade básica para a construção das RNA. A informação que chega dos dados de entrada é ponderada e processada por funções de ativação ao longo da estrutura da RNA até gerar a saída. O modelo básico de uma rede neural pode ser descrito como uma série de transformações, onde primeiro são contruídas M combinações lineares das variáveis de entrada x_1 até x_D que tomam a seguinte forma (Bishop, 2006):

$$a_j = \sum_{i=1}^{D} w_{ji}^{(1)} x_i + w_{j0}^{(1)} \quad [1]$$

onde j=1,...,M e 1 indica que os parâmetros estão na primeira camada da rede, em que $w_{ji}^{(1)}$ são os pesos (weights) e $w_{j0}^{(1)}$ são os vieses e a_j são as ativações, que posteriormente serão transformadas alguma função de ativação não-linear $h(.)$ fornecendo

$$z_j = h(a_j). \quad [2]$$

As funções $h(.)$ são geralmente sigmóides, e z_j são as saídas das chamadas unidades ocultas (*hidden nodes*). Estes valores são novamente combinados linearmente para gerar as ativações para as unidades de saída (que fornecem a saída)

$$a_k = \sum_{j=1}^{M} w_{kj}^{(2)} z_j + w_{k0}^{(2)} \quad [3]$$

onde k=1,...,M e K é o número total de saídas. Esta transformação corresponde à segunda camada da rede, e novamente $w_{k0}^{(2)}$ são os vieses. Finalmente as ativações da unidade de saída são transformadas utilizando a função de ativação apropriada para gerar as saídas da rede y_k. A escolha da função de ativação é determinada pela natureza dos dados e a distribuição assumida para as variáveis alvo. De acordo com Beale et al. (2011) a utilização de uma função de ativação tangente-sigmóide (por ser diferenciável, que é uma característica essencial a toda RNA) seguida de uma linear permite a aproximação de qualquer $f(x, y)$.

Os vários estágios apresentados acima podem ser combinados para visão da função geral da RNA:

$$y_k(x, w) = \sigma\left(\sum_{j=1}^{M} w_{kj}^{(2)} h\left(\sum_{i=1}^{D} w_{ji}^{(1)} x_i + w_{j0}^{(1)}\right) + w_{k0}^{(2)}\right) \quad [4]$$

Portanto o modelo de uma rede neural é simplesmente uma função não-linear de um grupo de variáveis de entrada $\{x_i\}$ para um grupo de variáveis alvo $\{y_k\}$ controlada por um vetor *w* de parâmetros ajustáveis.

Esta função também pode ser representada na forma de diagrama (Figura 4), porém não pode ser interpretado como modelo gráfico probabilístico em razão de seus nós internos representarem variáveis determinísticas. Segundo Bishop (2006) o processo de avaliação pode ser interpretado como a propagação avante (*forward propagation*) da informação através da rede, por isso RNA com essa topologia são chamadas de redes neurais *feed-forward*.

A arquitetura de uma rede é definida previamente às análises, e a sua conformação ideal é feita via comparação de modelos. Podem ser definidas 2 ou mais camadas, e diferentes quantidades de neurônios na camada oculta. Normalmente, para a análise de dados genômicos, tem-se uma quantidade muito grande de informações de genótipo para cada animal, que entrarão como input e serão processadas primeiramente nos neurônios do *hidden layer* aos quais estão interconectados e posteriormente no neurônio do *output layer*. Portanto, como a natureza dos dados utilizados para seleção genômica proporciona quantidade de parâmetros muito grande (p=(núm. de inputs x núm. de neurônios no hidden layer)+(2 x núm. de neurônios no hidden layer)+ (núm. de neurônios no output layer)) deve-se escolhar rede com arquitetura mais compacta, com apenas um *hidden layer* e poucos neurônios nesta camada.

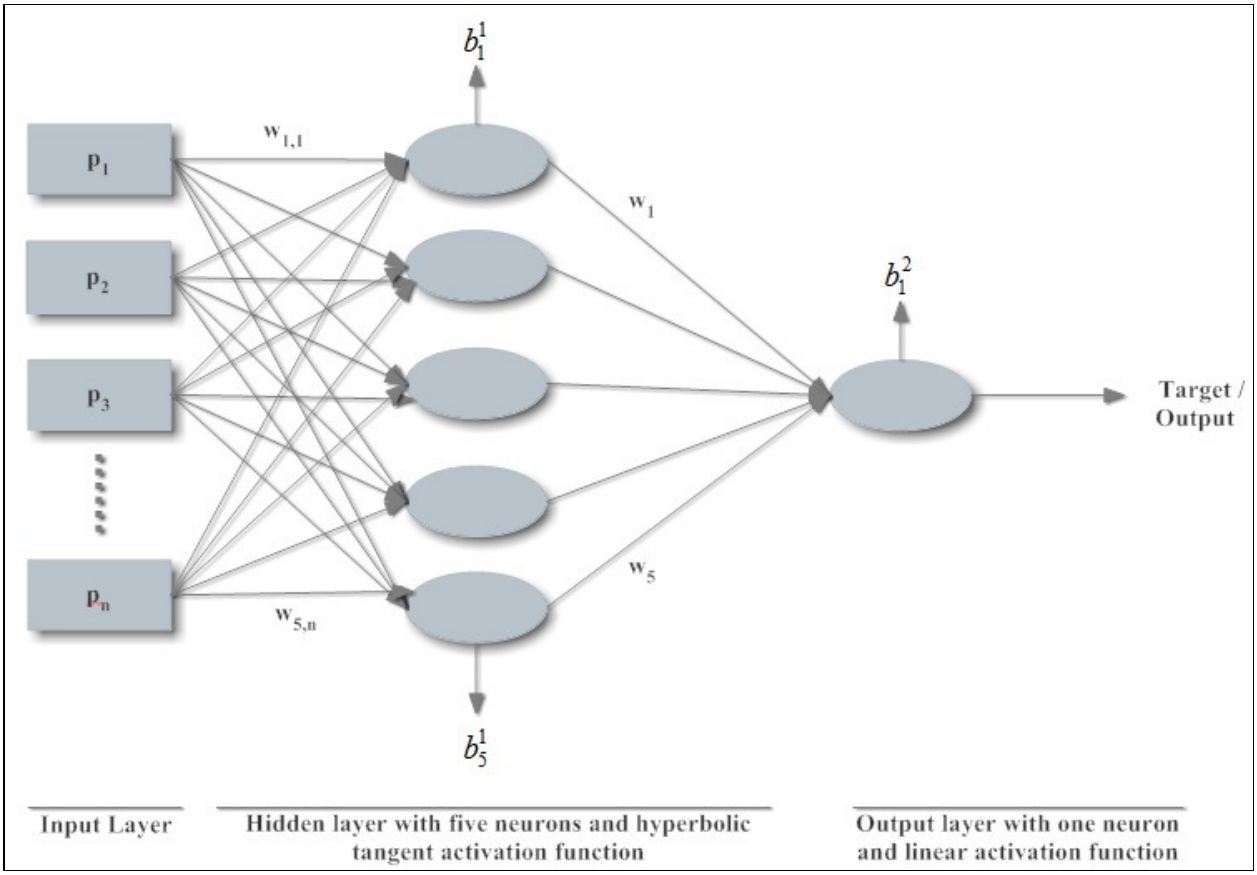

Figura 4. Arquitetura de uma RNA com duas camadas contendo cinco neurônios no hidden layer e um neurônio no output layer. Os p_i são as entradas para cada SNP de cada animal, sendo o número de linhas igual ao número de SNPs (o processamento de cada animal é feito em paralelo); os $w_{j,i}$ são as ponderações onde j se refere à identificação do neurônio no hidden layer e i identifica o SNP; b_j^k são os vieses onde j é a identificação do neurônio e k é a identificação da camada.

Para as RNA, existem 3 paradigmas de aprendizado (Suzuki, 2011):

- Aprendizado supervisionado: técnica de aprendizado de máquinas que estabelece parâmetros para uma RNA utilizando dados de treinamento. Os dados de treinamento são constituídos por entradas válidas (genótipos, por exemplo) e correspondentes saídas desejadas (fenótipos ou valores genéticos, por exemplo) representados como vetores.
- Aprendizado não-supervisionado: técnica de aprendizado de máquinas que estabelece parâmetros para uma RNA baseado nos dados e uma função de custo que deve ser minimizada. Esse tipo de aprendizado é aplicado para problemas

envolvendo estimação como modelagem estatística, compressão, filtragem e análise de agrupamento.

- Aprendizado por reforço: nesta técnica de aprendizado de máquinas, os dados geralmente não são fornecidos, mas gerados a partir de interações com o ambiente. Para cada entrada apresentada, é produzida uma indicação (reforço) sobre a adequação das saídas correspondentes produzidas pela rede (Wikipédia).

Treinamento ou Aprendizagem é o processo pelo qual as ponderações (*weights*) são modificadas à luz dos dados enquanto a rede tenta produzir a saída desejada (Okut et al., 2011). E, após o treinamento, a rede pode ser usada para a predição de saídas não-observadas. Este processo de ajuste dos pesos é realizado pelo algoritmo de aprendizagem, responsável por armazenar na rede o conhecimento observado obtido por meio de exemplos. Na literatura são apresentados vários algoritmos de aprendizagem, dentre eles o *backpropagation* que é o mais utilizado (Hajmeer, 2000).

Haykin et al. (1999) definem quatro tipos de regra de aprendizagem:

- Correção do erro: aplicado para o treinamento supervisionado, onde o erro (valor da saída – valor observado) vai sendo diminuído com o avanço no número de iterações pelo re-ajuste dos pesos.
- Hebbiana: Baseado em Hebb (1949) que afirma que "se dois neurônios em ambos os lados de uma sinapse são ativados sincronicamente, então a força desta sinapse é seletivamente aumentada". Portanto, nesta regra o ajuste dos pesos se baseia na atividade do *neurônio*.
- Boltzmann: é um procedimento de aprendizagem não-supervisionado para modelar uma distribuição de probabilidade.
- Competitiva: neste modelo de aprendizagem, os neurônios competem entre si e somente o que apresentar maior similaridade com o pradrão da entrada será ativado. Os neurônios próximos ao neurônio ativado terão seus pesos re-ajustados.

Portanto, três decisões devem ser tomadas para construção de uma RNA: 1) Arquitetura da rede; 2) Algoritmo de aprendizagem; e 3) Função de ativação. Todas elas podendo influenciar a função de mapeamento $f(x, y)$. Como já citado anteriormente, para seleção genômica, é preferível uma RNA com arquitetura menos complexa (apenas um *hidden layer*) que utilize função de ativação sigmóide (neurônio do *hidden layer*) seguida de uma linear (neurônio do *output layer*) e aprendizado baseado na atualização dos pesos com base na avaliação do erro. Preferivelmente, faz-se uma pré-seleção de SNPs e/ou regularização da RNA para evitar problemas de sobreajuste e para melhor generalização dos resultados obtidos.

1.2.3.2.3. Redes Neurais Artificiais: Treinamento e aprendizado

A maior parte dos algoritmos de treinamento aplica processo iterativo para minimização de uma função de erro, com ajuste dos pesos feitos de forma sequencial (Bishop, 2006). Em cada passo, duas etapas podem ser distintas: 1ª) Derivadas da função de erro com respeito aos pesos é avaliada (a retropropagação – *backpropagation* – possui método computacionalmente eficiente para avaliar essas derivadas) e retropropagada através da rede; 2ª) As derivadas são usadas para o ajuste dos pesos (a técnica mais simples, *gradient descent*, foi proposta originalmente por Rumelhart em 1986).

A retropropagação, como já foi apontado, é um método comum para treinamento de RNA. É um método de aprendizado supervisionado, e é uma generalização da regra delta (regra de aprendizado para atualização dos pesos – *gradient descent*) (Wikipédia). Portanto, este método requer dados para entradas e saídas para formação do arquivo de treinamento e que a função de ativação usada pelos *neurônios* artificiais seja diferenciável.

Voltando à fórmula apresentada em [1] (desconsiderando o parâmetro de viés): $a_j = \sum_i w_{ji} z_i$ (somatório das entradas ponderadas), onde z_i é a ativação da unidade ou entrada, que envia a conexão para a unidade *j*, e w_{ji} é o peso associado a esta conexão. Esta soma é então transformada por uma função de ativação não-linear $h(.)$ para gerar a ativação z_j para a unidade j na forma

$$z_j = h(a_j) \qquad [5]$$

em que a unidade *j* pode ser o output.

Considerando agora a avaliação da derivada de E_n ($soma\ de\ quadrado\ do\ erro\ dada\ por\ \frac{1}{2}\sum_k (y_{nk} - t_{nk})^2\ onde\ y_{nk} = y_k(x_n, w)$) com respeito ao peso w_{ji}, percebe-se que o primeiro depende do segundo apenas via somatório das entradas a_j para a unidade *j* (Bishop, 2006).

$$\frac{\partial E_n}{\partial w_{ji}} = (y_{nj} - t_{nj}) x_{ni} \qquad [6]$$

Aplicando a regra da cadeia:

$$\frac{\partial E_n}{\partial w_{ji}} = \frac{\partial E_n}{\partial a_j} \frac{\partial a_j}{\partial w_{ji}} \qquad [7]$$

em que $\delta_j \equiv \frac{\partial E_n}{\partial a_j}$ [8], onde $\delta' s$ são referenciados como erros. A partir de [1] obtem-se:

$$\frac{\partial a_j}{\partial w_{ji}} = z_i \qquad [9]$$

Então, fazendo as substituições, tem-se:

$$\frac{\partial E_n}{\partial w_{ji}} = \delta_j z_i \qquad [10]$$

Mostrando que a derivada é obtida pela multiplicação do valor δ para a unidade de saída no final da ponderação pelo valor de z da unidade de entrada no final da ponderação. Portanto, para avaliar as derivadas, é necessário calcular apenas δ_j para cada unidade oculta e de saída na rede e então aplicar a fórmula [10]. Para as unidades de saída tem-se:

$$\delta_k = y_k - t_k \, . \qquad [11]$$

E para avaliar os $\delta's$ para as unidades ocultas, faz-se uso novamente da regra da cadeia para as derivadas parciais,

$$\delta_j \equiv \frac{\partial E_n}{\partial a_j} = \sum_k \frac{\partial E_n}{\partial a_k} \frac{\partial a_k}{\partial a_j} \qquad [12]$$

Onde o somatório é feito para todas as unidades k para as quais a unidade j envia conexões (Figura 2). Substituindo-se a definição de δ dada por [8] em [12], tem-se a fórmula de retropropagação:

$$\delta_j = h'(a_j) \sum_k w_{kj} \delta_k \qquad [13]$$

Que mostra que o valor de δ para uma unidade oculta pode ser obtido pela retropropagação dos $\delta's$ das unidades acima na rede.

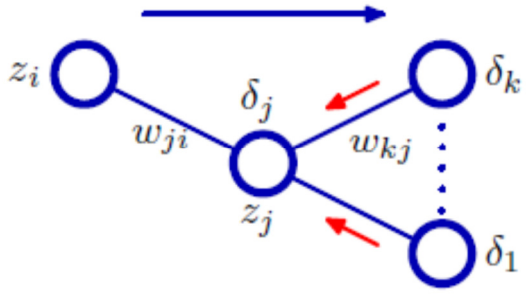

Figura 5. Ilustração do cálculo de δ_j para a unidade oculta j pela retropropagação dos $\delta's$ vindo das unidades k para as quais j envia conexões. A seta azul denota a direção do fluxo da informação durante a propagação avante, e as setas vermelhas indicam a retropropagação da informação de erro (Bishop, 2006).

Resumindo, as etapas de retropropagação são:

1. Propagação avantes: transformação das entradas para encontrar as ativações das unidades ocultas e de saída;
2. Avaliar todos os δ_k para as unidade de saída utilizando [11];
3. Retropropagar $\delta's$ utilizando [13] para obter δ_j para cada unidade oculta da rede;
4. Usar [10] para avaliar as derivadas requeridas.

Para atualização dos pesos da rede pode ser aplicado método de *gradient descent* (Figura 6), em que é dado um pequeno "passo" em direção ao gradiente negativo a fim de minimizar o erro, com base em:

$$w^{(\tau+1)} = w^{(\tau)} - \eta \nabla E(w^{(\tau)}) \qquad [14]$$

onde o parâmetro $\eta > 0$ é conhecido como taxa de aprendizado, que define o tamanho do passo de atualização (Bishop, 2006). Após a atualização, o gradiente é reavaliado para o novo vetor de pesos e o processo é repetido. A função de erro, neste caso, é definida com respeito ao arquivo de treinamento, portanto, cada passo requer que o arquivo inteiro seja processado antes de se avaliar ∇E.

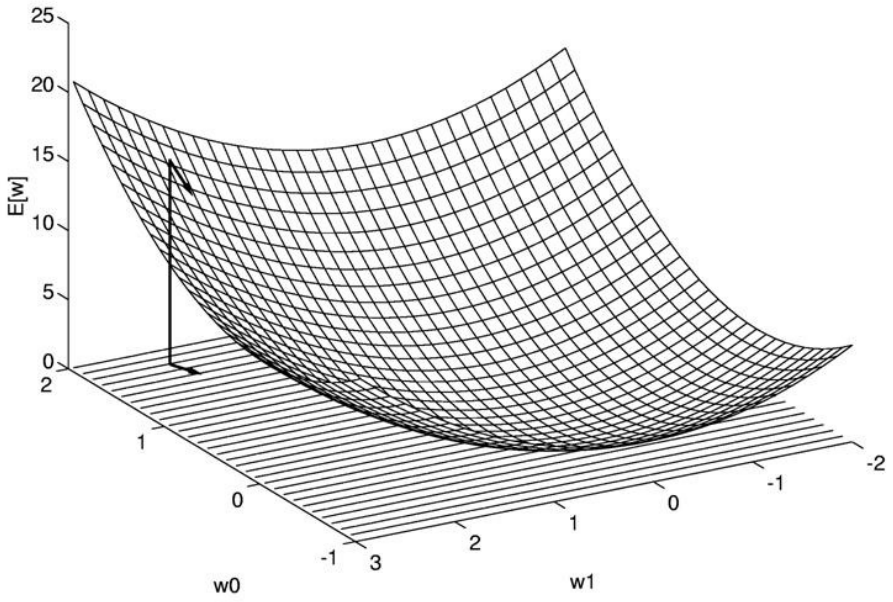

Figura 6. *Gradient descent* em RNA. Generalização da aplicação da regra delta, em que a atualização dos pesos se dá em direção ao decréscimo da função de erro.

Existem várias críticas ao método de gradient descent (Bishop e Nabney, 2008) devido a baixa taxa de convergência. Para isso métodos mais eficientes dora propostos como

métodos *quasi-Newton* que são mais robustos e mais rápidos (Nocedal e Wright, 1999). Diferentemente do gradient descent, estes algoritmos possuem a propriedade de que a função de erro sempre decresce a cada iteração a não ser que o vetor de pesos tenha atingido um mínimo local ou global (Bishop, 2006) (Figuras 7).

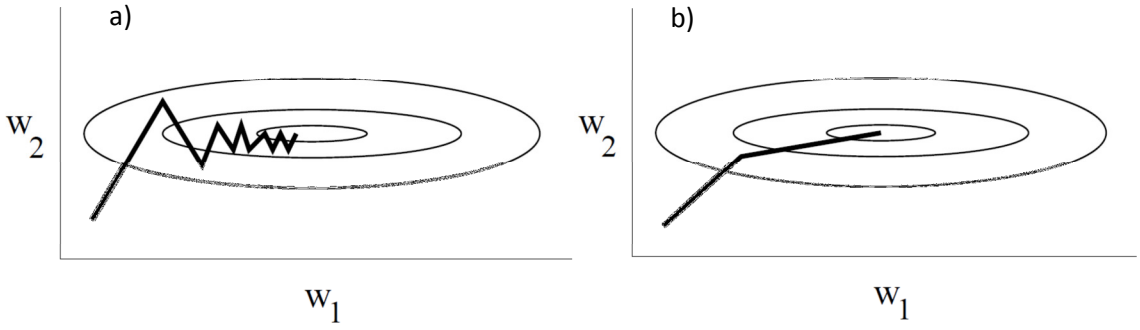

Figura 7. Algoritmos para estimativa de parâmetros em RNA. A) *Gradient descent:* problema de convergência, pois a diminuição da função de erro não é garantida a cada iteração . B) Método Newton: melhora visível na convergência. (Material disponível em http://www.robots.ox.ac.uk/~sjrob/Teaching/OxPR/l5.pdf)

Um método de otimização muito utilizado e com grande potencial de aplicação para aprendizado supervisionado é o algoritmo de Levenberg-Marquardt (LM), que envolve técnica iterativa para localizar o mínimo de uma função que é expressa como a soma de quadrados de funções não-lineares. O algoritmo LM pode ser visto como uma combinação dos métodos *gradient descent* e Gauss-Newton. Quando a solução está longe do valor correto, o algoritmo se comporta como o método de gradient descent (lento, mas com convergência garantida); enquanto que quando a solução está próxima do valor correto, se torna o método de Gauss-Newton (Lourakis, 2005). Este algoritmo tem implementação no *nnet toolbox* no programa MATLAB.

As RNA também possuem implementação bayesiana. Mackay (1995) foi o primeiro a apresentar a incorporação do paradigma de aprendizado bayesiano às redes neurais artificiais. O método apresentado pelo autor é baseado em uma aproximação Gaussiana da distribuição a posteriori dos pesos. A retropropagação bayesiana ajusta os parâmetros do weight decay (que será apresentado no próximo tópico) automaticamente para seus valores alvo, o que melhora a generalização da rede (Thodberg, 1995). Segundo Neal (1995), a aplicação dos princípios bayesianos para modelos de redes neurais aparentemente parecem complicados, dado que *a priori* para os parâmetros da rede não possui nenhuma relação com o conhecimento *a priori*, e a

integração da *posteriori* pode ser computacionalmente inexequível. O primeiro problema foi solucionado pela convergência de *prioris* a processos Gaussianos e o segundo usando métodos de MCMC (Markov chain Monte Carlo).

1.2.3.2.4. Regularização de Redes Neurais Artificiais

O conceito de regularização, na área de aprendizado de máquinas, se refere à introdução de informação adicional para evitar sobreajuste, que pode ser justificado pela Navalha de Occam. Esta informação é geralmente uma penalidade aplicada para a complexidade do modelo pela imposição de limites à norma do vetor espacial (neste caso o vetor contendo os pesos e vieses).

Um método "rústico" de regularização seria a escolha do número de *neurônios* no *hidden layer*. Um número menor, implica em maior simplicidade do modelo.

Outra maneira de reduzir o sobreajuste, é a adição de um termo de penalidade proporcional à soma de quadrados dos pesos e vieses à função de erro, resultando em um procedimento de estimação de máxima verossimilhança penalizada (Neal, 1995). Esta modificação é conhecida como *weight decay*, em razão do favorecimento de pesos com valores menores. O problema está na determinação da magnitude da penalidade que deve ser aplicada ao peso (necessidade de balanço entre sobreajuste e "sub"ajuste), que pode ser feito via comparação de modelos ou validação cruzada. Este regularizador tem a forma:

$$M(w) = \beta E_D + \alpha E_W$$

Onde $E_W = \frac{1}{2}\sum_i w_i^2$. Este termo adicional favorece valores pequenos de w e diminui a tendência de o modelo sobreajustar os dados de treinamento (Mackay, 1995).

Como proposto por Mackay (1995), o aprendizado das redes neurais pode incorporar o paradigma de inferência bayesiana e propõe a seguinte função objetiva:

$$P(w|D,\alpha,\beta,H) = \frac{P(D|w,\beta,H)P(w|\alpha,H)}{P(D|\alpha,\beta,H)}$$

em que D é o vetor de dados de entrada para treinamento, w se refere aos pesos da rede, H é a arquitetura da rede e α e β são os parâmetros de regularização. Seguindo o Teorema de Bayes, $P(D|w,\beta,H)$ é a função de verossimilhança, $P(w|\alpha,H)$ é a priori para os pesos para a arquitetura escolhida e $P(D|\alpha,\beta,H)$ é o fator de normalização. Mackay (1995) apresenta aproximação para a *posteriori* $P(w|D,\alpha,\beta,H)$ como uma Gaussiana.

Os parâmetros α e β determinam a complexidade do modelo (arquitetura da rede, forma da *priori* atribuída aos parâmetros, forma de β no modelo). Diferentes valores para os hiperparâmetros α e β definem diferentes sub-modelos, e para inferir estes valores, basta aplicar as regras de teoria de probabilidade (Mackay, 1995):

$$P(\alpha, \beta|D, H) = \frac{P(D|\alpha, \beta, H)P(\alpha, \beta|H)}{P(D|H)}$$

Em resumo, este método apresentado por Mackay, os pesos e os vieses da rede são tratados como variáveis aleatórias com distribuições especificadas. Os parâmetros de regularização são relacionados às variâncias desconhecidas associadas a estas distribuições.

Uma outra maneira alternativa de regularização é o *early stopping,* nesta técnica os dados disponíveis são divididos em três grupos: treinamento, validação e teste. O primeiro é usado para cômputo do gradiente e atualização dos pesos e vieses e o segundo para monitorar o decréscimo da função de erro. Quando os dados começam a ser sobreajustados, o erro para o arquivo de validação começa a aumentar (Figura 8).

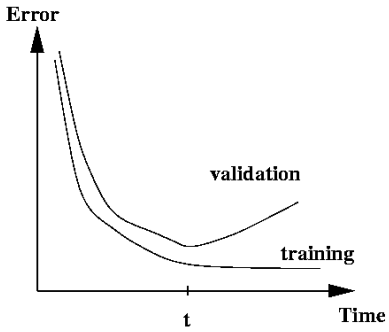

Figura 8. Ilustração do early stopping. Treinamento pára no tempo t, quando o erro para o arquivo de validação começ a aumentar.

A parada no processo de treinamento evita o "sobre"treinamento da rede, e com isso, diminuindo assim a complexidade do modelo (o número efetivo de parâmetros na rede aumenta com o passar das iterações) (Bishop, 2006). Entretanto diversos autores relatam superioridade da regularização bayesiana das RNA sobre o early stopping (Duon e Liong, 2004).

A vantagem da regularização bayesiana é que apenas com o arquivo de treinamento é possível inferir os parâmetros e hiperparâmetros (Mackay, 1995).

1.2.3.2.5. Aplicações no melhoramento genético

Dentro da área de genética e melhoramento animal, a aplicação das RNA é tema relativamente recente, principalmente se revisado para a literatura relacionada com seleção genômica.

Gianola et al. (2011) foram um dos primeiros a publicarem estudo envolvendo a utilização de RNA para a predição de características complexas. Os autores justificam que as RNA têm o potencial de acomodar relações complexas entre os dados de entrada e a variável resposta e por isso, são candidatas interessantes para a análise de características complexas afetadas por formas desconhecidas de interação entre genes. Primeiramente, os autores apresentam a visão do modelo infinitesimal de Fisher como uma RNA. E avaliam modelos de *Bayesian Regularized Artificial Neural Networks* (BRANN) para a predição de gordura, leite e rendimento de proteína para vacas usando como entrada informações do parentesco baseado na matriz de pedigree e genômica. Na figura 9 são apresentado os resultados obtidos para habilidade de predição, evidenciando que as RNA (diversas arquiteturas – um *hidden layer* com diferentes números de neurônios) forneceram melhores predições para o fenótipo que o modelo linear (*Bayesian Ridge Regression*), a utilização da informação de marcadores foi melhor para predição. Os autores concluem que RNA pode ser útil para a predição de características complexas utilizando dados genômicos de alta dimensão, situação em que o número de parâmetros a se estimar é maior do que o número de informações da amostra.

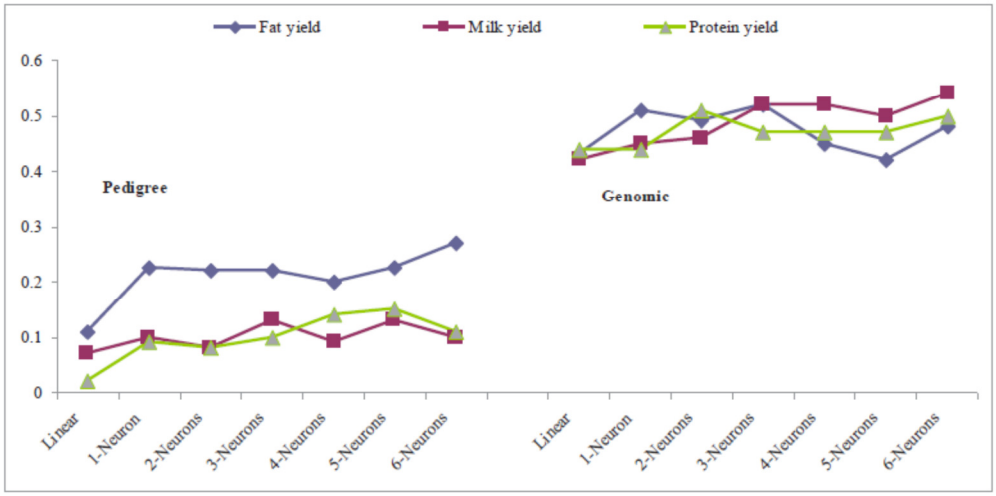

Figura 9. Correlações entre o fenótipo observado e fenótipo predito para modelo linear e RNA com diferentes arquiteturas (Gianola et al., 2011).

Um aspecto interessante de se utilizar a informação do grau de relacionamento entre indivíduos como entrada da rede restringe a quantidade de inputs para *n* (número de indivíduos na matriz A ou G), já que a utilização das informações dos marcadores diretamente envolve modelo com quantidade muito maior de pesos, que pode não ser computacionalmente viável.

Em um outro trabalho com a utilização de BRANN para a predição do índice de massa corporal (IMC) em ratos Okut et al. (2011) comparam o efeito sobre a qualidade de predição utilizando redes com diferentes arquiteturas. As entradas foram os genótipos para 798 SNPs (previamente selecionados por efeito) e a variável alvo considerada foi mensurações para índice de massa corporal pré-corrigido para os efeitos fixos. As redes testadas foram linear (1 neurônio no hidden layer com função de ativação linear) ou com 1, 2, 3, 4, 5, 6, ou 7 neurônios no hidden layer. Alguns resultados podem ser vistos na Figura 10. Os autores chegaram à conclusão que a rede contendo 5 neurônios na camada oculta fornecia melhores prediçoes, inclusive melhor que o modelo linear proposto. Um resultado interessante deste trabalho é que comparando o resultado para correlação entre predições e observações de IMC com os resultados obtidos por de los Campos (2009) utilizando os mesmos dados porém com o arquivo completo de genótipos (~11000 SNPs) e um modelo linear e (*Bayesian LASSO*) tem-se correlações semelhantes. O que mostra o potencial das RNA para a predição mesmo com o número limitado de entradas.

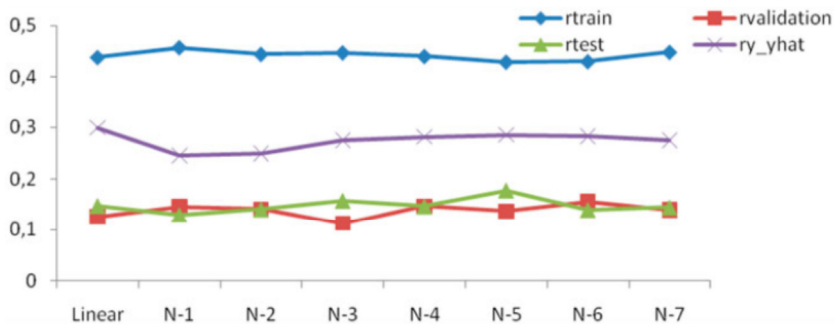

Figura 10. Correlações entre predições e observações para treinamento (r_{train}), teste (r_{test}), validação ($r_{validation}$) e total ($r_{y-\hat{y}}$) para arquiteturas linear e até 7 neurônios. (Okut et al., 2011)

Heslot et al. (2012) comparam diversas metodologias, dentre elas algumas envolvendo aprendizado de máquinas, para predição utilizando dados genômicos em plantas (milho, trigo e cevada). Os autores encontraram níveis de acurácia semelhantes entre vários modelos, entretanto o nível de sobreajuste variou muito assim como o tempo computacional e a distribuição para efeito dos marcadores. Os resultados para redes neurais mostram desempenho pobre em termos de quadrado médio do erro, mas que não foi refletido para a qualidade de predição (0,55 com validação cruzada e 0,85 sem validação cruzada), que foi comparável à acurácia alcançada em quase todos os outros modelos (Tabela 1). O esperado é que RNAs melhorassem a captação de sinais dos dados de genótipos e alcançassem, no mínimo, a mesma acurácia de modelos lineares. Entretanto, na prática, isto parece não ocorrer, uma explicação seria a tendência de sobreajuste das RNA ou tipos de algoritmos utilizados não serem eficientes.

Tabela 1. Acurácia para cada característica e modelo, correlação média sem validação cruzada para cada modelo, e QME para cada modelo (Fonte: Heslot et al., 2012).

Dataset†	Trait‡	RR-BLUP§	BL	Elastic net	wBSR	BayesCπ	E-Bayes	RKHS	SVM	RF	NNET
Barley 1	Yield	0.53	0.55	0.52	0.53	0.53	0.53	0.6	0.43	0.56	0.51
Barley CAP	Betaglucan	0.57	0.57	0.57	0.57	0.57	0.57	0.6	0.35	0.55	0.54
Bay × Sha (Bay-0 × Shahdara)	FLOSD	0.82	0.82	0.83	0.83	0.82	0.82	0.83	0.8	0.85	0.82
	DM10	0.63	0.63	0.63	0.64	0.63	0.63	0.64	0.56	0.57	0.56
	DM3	0.4	0.39	0.40	0.4	0.39	0.4	0.41	0.33	0.38	0.35
Panel maize	Moisture	0.75	0.75	0.75	0.76	0.75	0.73	0.79	0.45	0.73	0.73
	Yield	0.63	0.63	0.61	0.63	0.63	0.59	0.64	0.32	0.6	0.59
Diallel maize	Moisture	0.74	0.74	0.72	0.73	0.74	0.73	0.75	0.56	0.61	0.72
	Yield	0.52	0.52	0.49	0.51	0.52	0.51	0.5	0.29	0.49	0.48
Wheat CIMMYT	YLD1	0.51	0.5	0.46	0.48	0.51	0.49	0.59	0.36	0.52	0.54
	YLD2	0.5	0.49	0.45	0.5	0.5	0.46	0.52	0.36	0.43	0.51
	YLD4	0.38	0.37	0.35	0.36	0.38	0.36	0.43	0.32	0.38	0.43
	YLD5	0.44	0.47	0.42	0.47	0.44	0.39	0.52	0.27	0.46	0.44
Wheat Cornell	Yield	0.36	0.35	0.37	0.37	0.34	0.26	0.28	0.22	0.36	0.36
	Height	0.45	0.44	0.41	0.44	0.44	0.41	0.55	0.37	0.46	0.45
Wheat diallel	Height	0.64	0.66	0.68	0.67	0.66	0.67	0.73	0.51	0.62	0.67
	TKW	0.6	0.57	0.59	0.6	0.59	0.59	0.68	0.41	0.54	0.65
	Yield	0.53	0.52	0.51	0.52	0.53	0.51	0.58	0.39	0.52	0.57
Average accuracy (cross-validated)		0.56	0.56	0.54	0.56	0.55	0.54	0.59	0.41	0.54	0.55
Average non-cross-validated correlation		0.77	0.79	0.75	0.77	0.77	0.93	0.99	0.89	0.76	0.85
Average MSE		0.67	0.67	0.69	0.68	0.68	0.76	0.64	1.36	0.72	10.54

RR-BLUP: Random regression BLUP; BL: Bayesian LASSO; wBSR: Bayesian Shrinkage Regression; RKHS: Reporducing Kernel Hilbert Spaces; SVM: Support vector machine; RF: Random forests; NNET: Neural networks.

Ventura et al. (2012) utilizam RNA para a predição do valor genético do peso de bovinos de corte aos 205 dias de idade. Os autores mostram outros trabalhos semelhantes (Neves, 2007, Meireles, 2005) que indicam o potencial das RNAs na avaliação genética em bovino de leite e corte. Neste trabalho foi utilizada RNA do tipo perceptron multicamada e algoritmo de Levenberg-Marquadt e como entradas foram inseridas as variáveis idade da mãe ao parto, estação, região, peso aos 205 dias de idade e como saída o valor genético (obtido por meio do BLUP) para a característica em questão. Resultados mostraram correlacões de 0,68 (1994-1995) e 0,74 (1991-1993) entre as estimativas do BLUP e da RNA no arquivo de validação. Entretanto, este modelo não seria recomendável para avaliações genéticas em razão da sua dependência da metodologia BLUP pela incapacidade de incorporação da matriz de relacionamento entre animais. Além do mais, esta forma de utilização da RNA não faz uso de todo seu potencial de aproximação já que restringe aos efeitos captados pelo BLUP que é apenas aditivo.

2. Imputação

2.1.Metodologias para Imputação

Vários métodos de imputação têm sido propostos e são implementados em programas como fastPHASE (Scheet e Stephens, 2006), Beagle (Browning e Browning, 2009), MACH (Li et al., 2010), IMPUTEv2 (Howie et al., 2009), findhap (VanRaden et al., 2011), Fimpute (Sargolzaei et al., 2010) entre outros. Estas metodologias imputam os genótipos desconhecidos baseados na recontrução de haplótipos com base na informação de LD entre os SNPs. Todos eles utilizam diferentes métodos para reconstruir o haplótipo o que leva a diferenças na acurácia dos genótipos estimados e tempo computacional para imputação.

2.1.1. Modelos com Cadeias de Markov Ocultas (Hidden Markov Models)

Se todos os haplótipos dos indivíduos ao longo do genoma e quais SNPs estão associados a cada haplótipo único na população fossem conhecidos, seria possível inferir ou imputar genótipos carregados para um indivíduo em qualquer loco. Entretanto, na prática, não se conhecem os haplótipos verdadeiros que cada indivíduo carrega. Para este problema foram sugeridos os Modelos de Cadeia de Markov Oculta (*Hidden Markov Models – HMM*). Os HMM são uma classe de modelos matematicamente elegantes e computacionalmente exequíveis em que os dados observados são gerados por um processo não observado de Markov (Browning e Browning, 2011). Um processo de Markov é um processo probabilístico em que a distribuição dos estados futuros dependem somente do estado atual e não de estados anteriores.

Em HMM, o estado oculto (hidden state) - os haplótipos verdadeiros na população - geram as observações, que são os genótipos. HMM têm sido amplamente utilizados para estimar a probabilidade de um indivíduo carregar um genótipo em um SNP em particular, dadas as informações para este mesmo indivíduos para outros SNPs e o resto da população (Hayes, 2011).

Num modelo de Markov tradicional os estados (*states*) podem ser diretamente observados, ao contrário dos HMM, no qual os estados são ocultos e resultados estocásticos destes estados são observados (Rehnberg, 2009). Como os estados possuem distribuição de probabilidade sobre os seus resultados, as observações destes resultados trazem, portanto, informação sobre os próprios estados desconhecidos. A representação gráfica de um HMM apresentada na Figura 11.

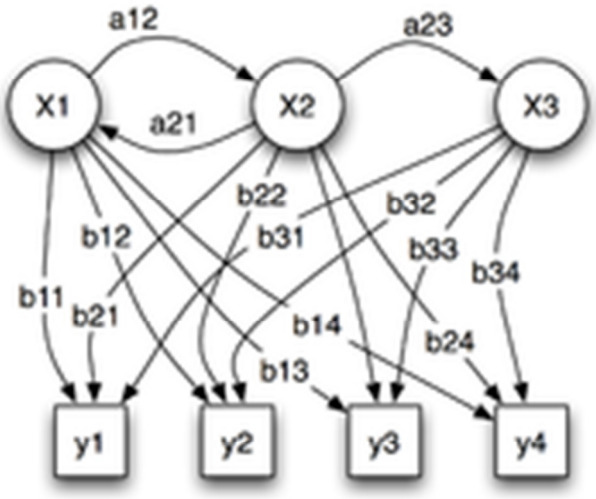

Figura 11. Parâmetros probabilísticos de um HMM (Wikipédia)
X = {X1, X2, X3} – estados ocultos (*hidden states*)
y={y1, y2, y3, y4} – possíveis observações de X
{a_{ij}} – probabilidades de transição
{b_{ij}} – probabilidades dos resultados para cada estado (x_i tendo o resultado y_i)

Na aplicação desta metodologia para imputação de genótipos, os hidden states representam os haplótipos da amostra de referência. As probabilidades de transição determinam a forma com que os hidden states podem mudar de uma posição do cromossomo para outra, ou seja, descreve a probabilidade de se mover de um haplótipo para outro (em razão de recombinação ou mutação, por exemplo); e as probabilidades de emissão unem estados não observados aos dados observados, portanto são a probabilidade de observar um genótipo carregado pelo indivíduo para um haplótipo em particular. Neste caso, a observação são os genótipos (observados ou não). Estes modelos reconhecem que novos haplótipos são derivados de velhos haplótipos pelo processo de mutação e recombinação (Browning e Browning, 2011). Portanto, a inferência do possível genótipo é impossível de ser realizada deterministicamente. O problema na imputação é derivar probabilidades para o genótipo, dados os hidden states, genótipos esparsos, taxas de recombinação e outros parâmetros populacionais.

Então, para dados genômicos o HMM pode ser representado conforme apresentado na Figura 12.

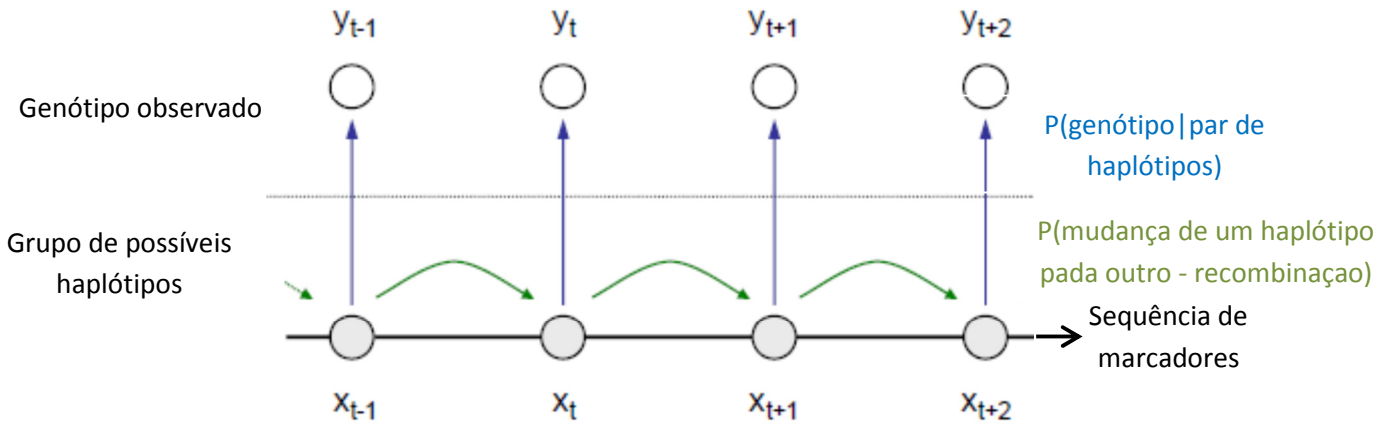

Figura 12. Representação gráfica de HMM para imputação de genótipos (Adaptado de Nothnagel, 2010).

A diferença entre métodos está nas pressuposições a respeito dos hidden states e a forma com que as probabilidades de transição, probabilidades de emissão e probabilidades iniciais são derivadas.

fastPHASE. Este programa foi apresentado por Scheet e Stephens (2009). O método usa a observação de que haplótipos tendem a se agrupar em conjuntos de haplótipos relacionados ou similares. Segundo Marchini e Howie (2010) o modelo especifica um grupo de K estados não observados ou conjunto que são designados para representar os haplótipos comuns. O *k*-ésimo conjunto recebe o peso α_{kl} que denota a proporção que este grupo de haplótipos é observado, com:

$$\sum_k \alpha_{kl} = 1$$

Cada conjunto possui também uma frequência associada (θ_{kl}) para o alelo 1 em cada ponto, ou seja, a frequência alélica dentro de cada grupo. O dados de genótipo para cada indivíduo é então modelado como um HMM :

$$P(G_i|\alpha, \theta, r) = \sum_z P(G_i|Z_i, \theta)P(Z_i|\alpha, r)$$

em que *r* são parâmetros que controlam transições entre estados para cada SNP e Z são os hidden states. Na equação, $P(G_i|Z_i, \theta)$ modela quão prováveis são os genótipos observados

dados os haplótipos e $P(Z_i|\alpha, r)$ modela padrões de mudança entre estados, mas estes estados representam grupos (conjunto de haplótipos quase idênticos) ao invés de haplótipos referência.

A verossimilhança do genótipo G_i é:

$$L(G, H|\alpha, \theta, r) = \prod P(G_i|\alpha, \theta, r) \prod P(H_i|\alpha, \theta, r)$$

Para ajuste do modelo é utilizado algoritmo de maximização da esperança (algoritmo EM) e os genótipos são imputados condicionalmente às estimativas dos parâmetros.

Os autores do programa sugerem a definição do número de grupos K=20 (que é fixo, diferente do Beagle). Como vantagem deste método tem-se o agrupamento de hidden states e por consequência menor tempo para cálculo dos parâmetros.

MACH. Este programa é semelhante ao fastPHASE em alguns aspectos, entretanto utiliza os próprios haplótipos como hidden states e não grupos. Este método funciona pela atualização sucessiva da fase do genótipo de cada indivíduo condicional às estimativas atuais do haplótipo de todas as outras amostras (Marchini e Howie, 2010). Os indivíduos são removidos do grupo de haplótipos de referência um a um e, após várias iterações, as estimativas dos haplótipos convergem para uma solução ótima. O modelo é:

$$P(G_i|D_{-i}, \theta, \eta) = \sum_Z P(G_i|Z, \theta, \eta) P(Z|D_{-i}, \theta)$$

onde D_{-i} é o grupo de haplótipos estimados para todos os indivíduos, exceto i, Z corresponde aos hidden states do HMM, e η é um parâmetro de erro que controla quão similar G_i é dos haplótipos e θ é o parâmetro de "crossover"que controla a transição entre os hidden states.

IMPUTE v1 e v2. O método de v1 é baseado em HMM para o vetor de genótipos de cada indivíduo, G_i, condicional a H (haplótipos na população referência), e um conjunto de parâmetros. O modelo pode ser escrito como:

$$P(G_i|H, \theta, \rho) = \sum_z P(G_i|Z, \theta), P(Z|H, \rho)$$

em que $Z=\{Z_1,...,Z_L\}$ com $Z_j=\{Z_{j1},Z_{j2}\}$ e $Z_{jk}=\{1,...,N\}$. O Z_j pode ser designado para o par de haplótipos do painel de referência para o SNP j que está sendo copiado para formar o vetor de genótipos. O termo $P(Z|H, \rho)$ modela como o par de haplótipos copiados muda ao longo da sequência e é definido por uma cadeia de Markov em que a mudança entre estados depende da estimativa do mapa de recombinação (ρ) pelo genoma. O termo $P(G_i|Z, \theta)$ permite que o vetor do genótipo observado seja diferente dos genótipos determinados pelo haplótipo e é controlado pelo parâmetro de mutação θ (Marchini e Howie, 2010).

O IMPUTE v2 é similar ao previamente descrito. Primeiramente, SNPs são divididos em dois grupos: T (genotipado na amostra e no painel de referência) e U (genotipado apenas no painel de referência). Na primeira fase são estimados os haplótipos para SNPs em T, utilizando IMPUTE v1. E, posteriomente, imputar os SNPs em U condicionalmente aos haplótipos estimados. O passo de imputação é haplóide (haplótipos da amostra são comparados diretamente com haplótipos da população referência), e, portanto, é mais rápido com comparado com a mesma metodologia porém diplóide (compara genótipos).

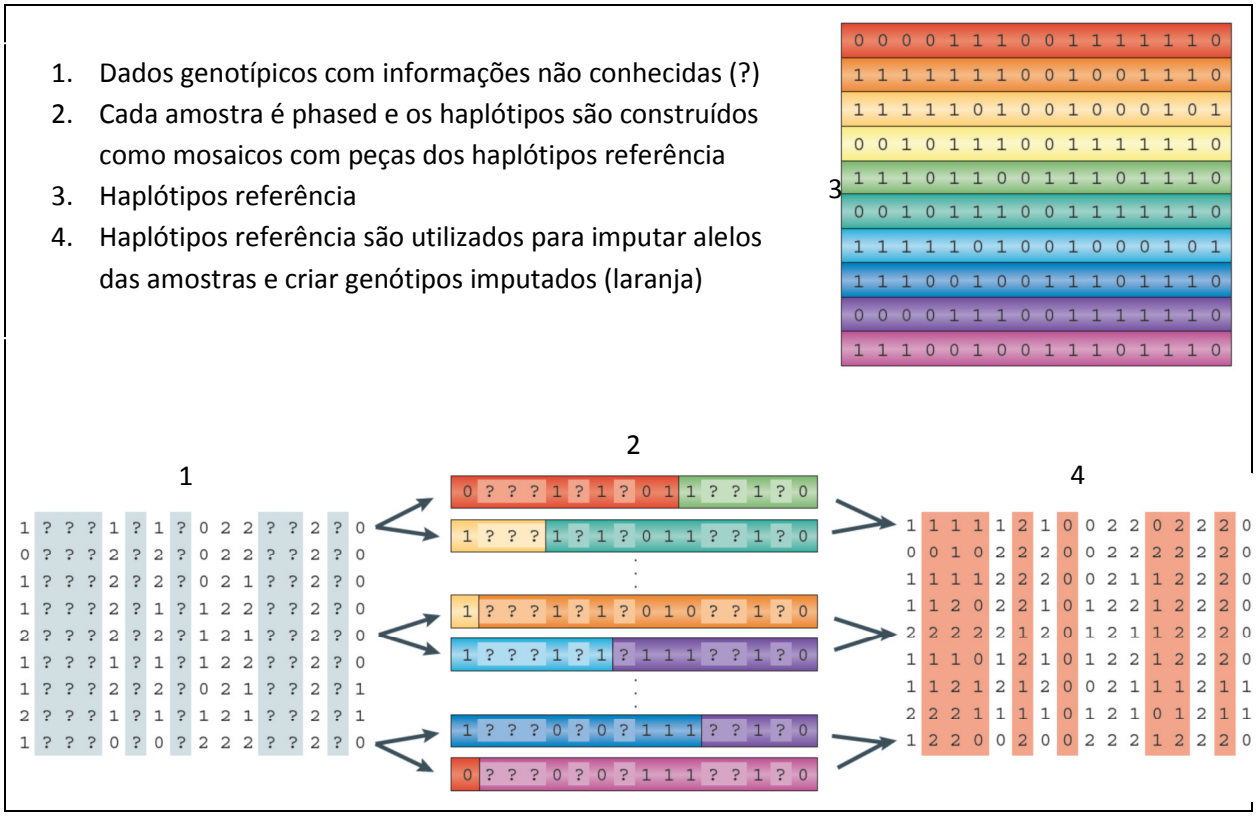

Figura 13. Esquema geral de como a imputação funciona (Adaptado de Marchini e Howie, 2010).

Beagle. Beagle utiliza um método diferente dos programas citados anteriormente para definir os *hidden states*. Como o fastPHASE, faz agrupamento de haplótipos por semelhança, porém a forma como estes grupos são montados é diferente, já que o K (número de grupos) pode variar. Beagle também é baseado em HMM e agrupa os haplótipos em cada loco, e o agrupamento se adapta à quantidade de informação disponível com isso o número de grupos aumenta globalmente com o tamanho da amostra e localmente com o LD (Browning e Browning, 2011). Este modelo é tido como mais parcimonioso pois, além de possuir menor número de hidden states em razão do agrupamento de haplótipos, o modelo considera apenas

uma pequena parcela de todas as possíveis transições entre estados em uma posição e estados da próxima posição (o protocolo de Li e Stephens utilizado pelo MACH e IMPUTEv2 permite todas as transições possíveis) (Figura 4– Browning, 2008).

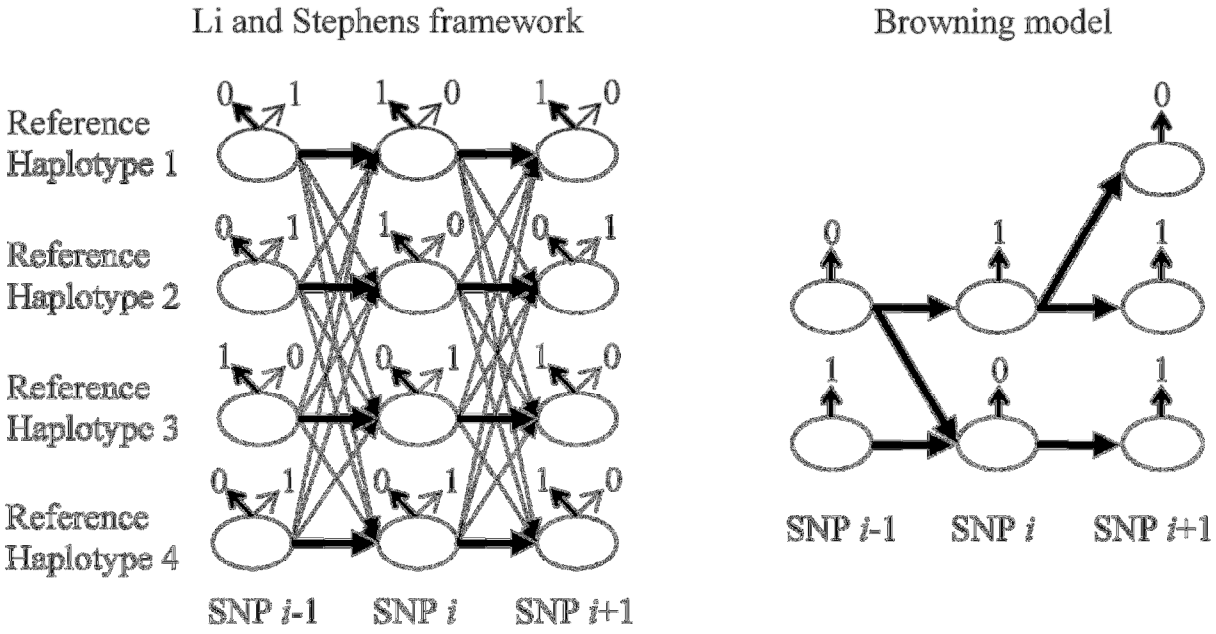

Figura 14. Ilustração ressaltando as principais diferenças entre modelos baseados em Li e Stephens (2003), a base de MACH, IMPUTE e fastPHASE, e o modelo de Browning (Browning, 2006), a base do Beagle. O exemplo mostra modelos cobrindo 3 marcadores (SNPs i-1, i e i+1). As elipses são os *hidden states*. Para o sistema de Li e Stephens, estes estados são definidos pelos haplótipos referência, enquanto que para o modelo de Browning os estados são grupos de haplótipos localizados. O modelo de Browning tende a ter menor número de estados para qualquer marcador. E o número de estados pode variar de marcador para marcador para o modelo de Browning enquanto de para Li e Stephens ele é fixo. As setas entre os *hidden states* de um SNP para outro são as transições de HMM. Para Li e Stephens, transições com maior probabilidade *a priori* (presentes no haplótipos de referência) são representadas pelas setas em negrito, enquanto que as setas mais finas representam alguma recombinação histórica. Para o modelo de Browning, existem no máximo k transicões saindo de um estado para outro, onde K é o número de alelos do próximo marcador (no caso de SNPs, é igual a 2), o que ajuda a manter o modelo mais parcimonioso. As setas saindo da parte superior das elipses são possíveis emissões do HMM, ou seja, os alelos observados. Para Li e Stephens, emissões com maior probabilidade a priori (alelos nos haplótipos de referência) são representadas com as setas em negrito, enquanto que as setas finas representam mutações para outros alelos. Os haplótipos referência são 011, 010, 101 e 001. Para o modelo de Browning, existe apenas uma emissão possível de cada estado (Browning, 2008).

No Beagle, parâmetros como recombinação e mutação não estão explícitos no modelo, mas são capturados implicitamente pela informação contida na amostra. Por outro lado, tem-se a representação do modelo por grupos de haplótipos e as possíveis transições entre eles, juntamente com as frequências observadas destas transições (Browning, 2008) (Figura 14). No caso de grandes amostras este tipo de modelagem é conveniente pois reduz o espaço de busca do algoritmo e a estimativa de parâmetros como recombinação se tornam menos importantes.

O modelo de Browning determina um HMM, e as setas representadas no DAG (directed acyclic graph) são os estados. Para especificar o HMM, é necessário também especificar as probabilidades de emissão, as probabilidade de estado inicial e as probabilidades de transmissão. Cada estado (seta) emite com probabilidade 1 o alelo que marca a seta. Portanto, o estado determina um único alelo observado, mas o alelo observado para um marcador não necessariamente determina o estado, pois setas com pais (elipse à esquerda) distintos, no mesmo nível do gráfico, podem ser identificadas com o mesmo alelo (Browning, 2008). As probabilidades de estado inicial e de transição são obtidas por meio de contagem das setas.

2.1.2. Utilização de informações de parentesco

Todos os modelos até agora apresentados não consideram a informação de parentesco para a realização da imputação de genótipos. Se o pedigree entre população de referência e amostra é conhecido, essa informação pode ser adicionada ao modelo para imputação a fim de melhorar sua acurácia. Neste caso, são fornecidas informação de duplas (pai-filho) ou trios (pai-mãe-filho) e os genótipos são tranformados em haplótipos simplesmente contando alelos que ocorrem em conjunto nestes indivíduos (co-segregação alélica) (Hayes, 2007), isto é, os alelos idênticos por descendência em diferentes pontos em um mesmo cromossomo estarão em um único haplótipo no filho e um único haplótipo no pai, assumindo que nenhuma recombinação ocorreu (Figura 15). O problema neste caso é inferir o haplótipo quando, por exemplo, ambos pai e filho (ou pai, mãe e filho, se a informação for para trios) são heterozigotos para algum determinado SNP, para esta situação, utiliza-se também a informação usando frequêcia dos haplótipos na população. E, como os indivíduos aparentados irão compartilhar haplótipos mais longos (região de IBD), os descendentes podem ser genotipados com mais espaçamento entre marcadores.

Alguns softwares apresentados sofreram modificações para receberem este tipo de informação (Druet e Georges, 2010), como o Beagle e o fastPHASE.

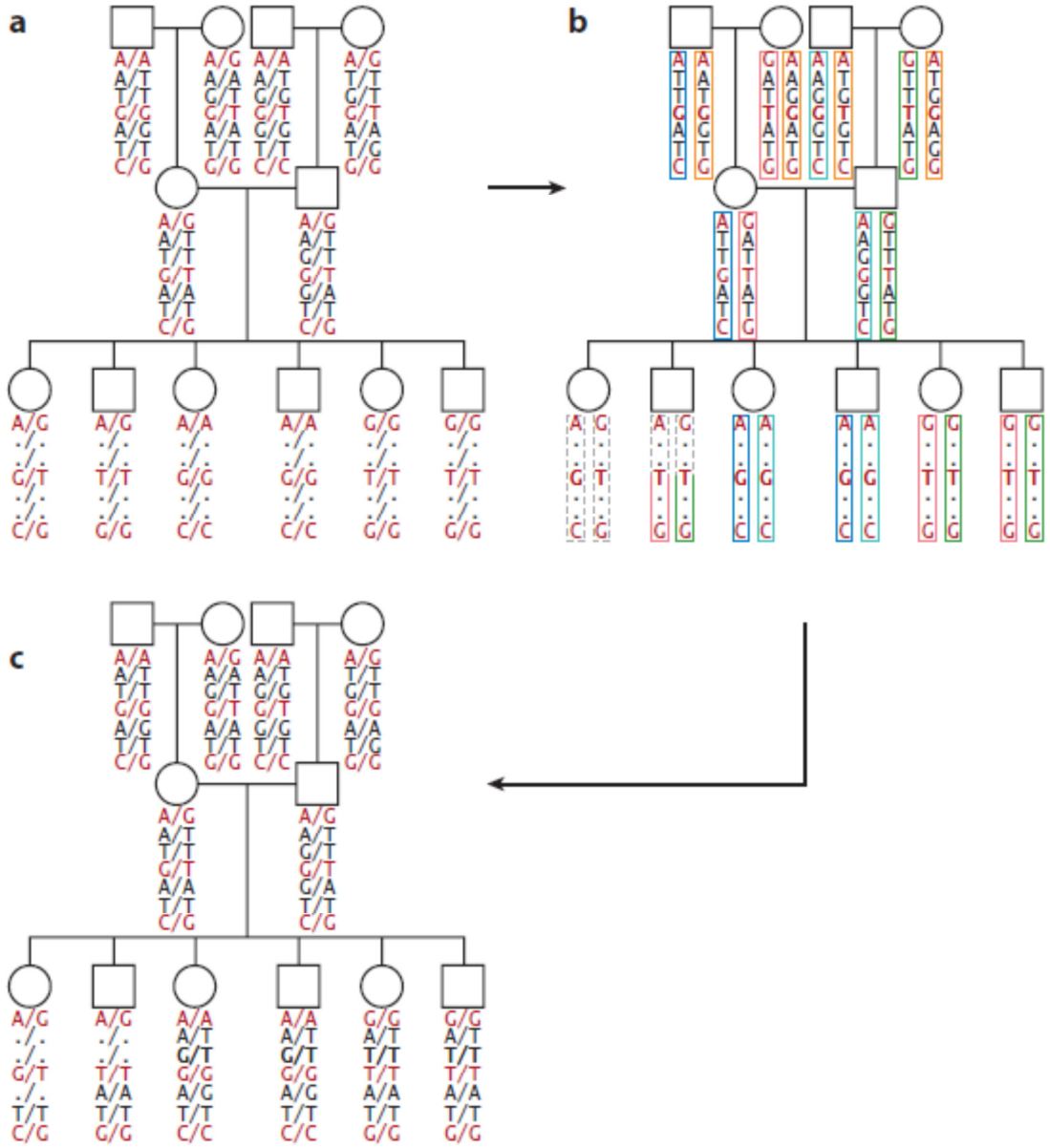

Figura 15. Imputação de genótipos para uma amostra de indivíduos aparentados. (a) Os dados observados – genótipos para um grupo de marcadores. Neste caso, parte destes marcadores foram genotipados para todos os indivíduos (vermelho), enquanto que o restante foi identificado apenas para alguns indivíduos (preto – para as duas primeiras gerações do pedigree); (b) O processo de inferir a informação de identidade por descendência (IBD) observando os marcadores disponíveis para todos os indivíduos. Cada segmento de IBD que aparece em mais de um indivíduo recebe uma mesma cor. Por exemplo, o segmento marcado em azul é compartilhado entre o primeiro indivíduo na geração de avós e o primeiro indivíduo na geração de pais, e indivíduos 3 e 4 da última geração; (c) Genótipos observados e

informação de IBD combinadas para preencher os genótipos desconhecidos na última geração. (Li et al., 2009).

Kong et al. (2008), baseado na informação de genótipos de animais aparentados, sugeriram o método chamado "long range phasing", o ponto chave desta metodologia é a observação de que se animais possuem genótipos homozigotos não-conflitantes ao longo de locos consecutivos, eles possuem pelo menos um longo haplótipo em comum, e a probabilidade de que este haplótipo tenha origem em um ancestral comum é alta. Uma aplicação útil deste método é para imputar genótipos densos em indivíduos que possuem poucos genótipos identificados a partir da informação de genótipos de parentes (Daetwyler et al., 2011, Hickey, 2011).

2.2. Fatores que afetam a acurácia da imputação

Vários fatores podem afetar a acurácia de imputação, dentre eles:

2.2.1. Software utilizado

Existem vários trabalhos que comparam a acurácia de imputação de genótipos utilizando vários dos softwares disponíveis (Pei et al., 2008, Nho et al., 2011, Sun et al., 2012), e, em grande parte deles, a acurácia entre metodologias é semelhante. Entretanto, dependendo da estrutura do banco de dados, alguns softwares devem ser preferidos. Por exemplo, no caso de banco de dados grandes (mais de 1000 indivíduos) é preferível a utilização de métodos que façam agrupamento de haplótipos, como o Beagle, pois o tempo de análise é menor e a acurácia relatada em diversos trabalhos está acima de 90%.

Browning e Browning (2009) compararam a acurácia de imputação do programa Beagle 3.0 e IMPUTE 0.5.0 com painéis de referência de 60, 300 e 600 indivíduos (em humanos) e 188 indivíduos na amostra e mostram que a diferença entre estes dois métodos diminui como o aumento do tamanho do painel de referência. Os resultados para correlação entre frequências alélicas reais e estimadas foram: 0,9902 (BEAGLE) e 0,9917 (IMPUTE) com o painel de 60 indivíduos, 0,9975 (BEAGLE) e 0,9976 (IMPUTE) com o painel de referência de 300 indivíduos, e 0,9984 (BEAGLE) e 0,9982 (IMPUTE) para o painel de referência com 600 indivíduos. Por outro lado, o tempo computacional para imputar dados em uma região de 5mb no cromossomo 1 na amostra de 188 indivíduos para painéis de referência contendo 300, 600 e 1200 indivíduos para o BEAGLE foram 2,7min, 5,5min e 12min, respectivamente e para o IMPUTE foram 60,3min, 220,2min e 829,6min.

Para gado de leite, Jonhston e Kistemaker (2011) compararam primeiramente os softwares BEAGLE e MACH para imputar genótipos do cromossomo 1 de animais da raça Pardo Suiça e mostram que ambos possuem alta acurácia, porém o MACH foi duas vezes mais lento. Os autores também mostraram que o esquema para imputação onde primeiro usa-se o FImpute (Sargolzaei, 2010) que utiliza informação do pedigree e depois o Beagle para imputar chips de 3k para 50k foi o de maior sucesso.

Sun et al. (2012) compararam o desempenho de 6 softwares para imputação (Beagle, IMPUTE, fastPHASE, Alphaimpute, findhap e Fimpute) para imputar genótipos em 29 cromossomos de 5k para 50k SNPs em bovinos de corte (3078 animais de raça Angus). Os autores obtiveram valores de acurácia entre 0,8677 e 0,9858 e as mais altas foram para os programas Beagle e FImpute. Também apresentaram um método para usar todos os resultados dos métodos de imputação em conjunto, a fim de resolver inconsistências entre as diferentes metodologias, e puderam alcançar melhor acurácia do que o Beagle.

2.2.2. Tamanho da amostra

Se todos os outros fatores são mantidos constantes, quanto maior o tamanho da amostra, maior a acurácia da inferência dos haplótipos e, por consequência, maior a acurácia de imputação. Sendo que algumas metodologias parecem ser mais sensíveis que outras à mudança do tamanho amostral (Figura 16). E, de acordo com Browning e Browning (2011), aumentar o tamanho do painel de referência é uma maneira simples de se aumentar a acurácia da imputação.

Alguns trabalhos em humanos mostram que a combinação de painéis de referência melhora os resultados da imputação (Nho et al., 2011). Zhang et al. (2011) estudando o efeito dos diferentes tamanhos do painel de referência sobre a qualidade da imputação de genótipos em humanos concluem que um tamanho ideal para o painel de referência é crucial, entretanto o tamanho da amostra (indivíduos com genótipos desconhecidos) tem pouco efeito sobre a qualidade da predição dos genótipos. O que é consistente com os resultados de Huang et al. (2009) que relataram independência condicional da imputação sobre o tamanho amostral dado o painel de referência.

Por outro lado, Huang et al. (2012) relatam que o tamanho em si do painel de referência não é importante, já que um grande número de indivíduos não necessariamente significa grande quantidade de informação. Os autores mostram a possibilidade de se utilizar painéis de referência menores e informativos para a amostra em estudo (gado Hereford) com a utilização de, por exemplo, painéis com informação de várias populações e gerações.

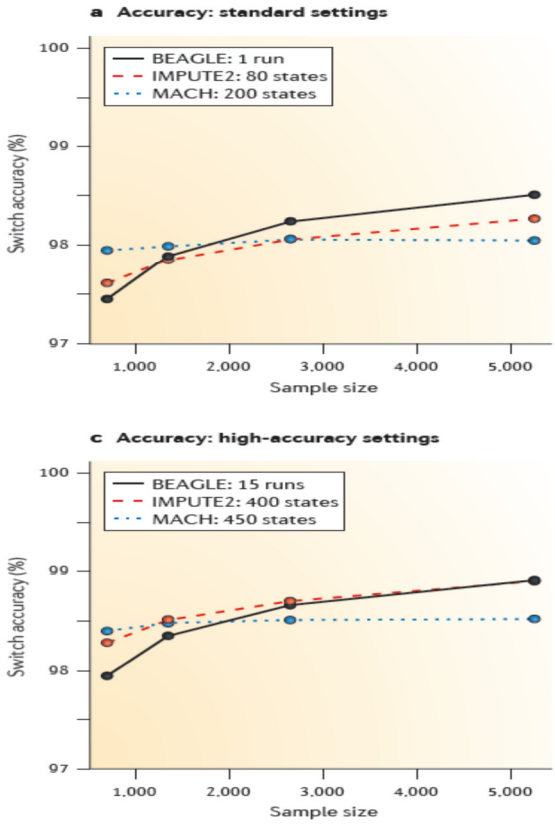

Figura 16. Acurácia da imputação em função do tamanho da amostra para diferentes metodologias (Browning e Browning, 2011).

2.2.3. Densidade dos marcadores

Outro fator importante a ser considerado para a predição de genótipos não observados é a proporção entre SNPs desconhecidos/SNPs sequenciados e presentes no painel de referência. Diversos estudos tentam buscar qual deve ser a quantidade mínina de marcadores presentes na população amostral para obtenção de boa qualidade dos genótipos imputados.

Weigel et al. (2010) compararam o fastPHASE 1.2 e IMPUTE 2.0 para a predição de genótipos em gado Jersey com painel de referência contendo 2542 animais e 43385 SNPs para painéis com diferentes densidades. Os resultados deste estudo mostraram que a acurácia de imputação pode ser modesta (correlação <0.8) quando são utilizados menos de 1000 SNPs, enquanto que para plataformas com média densidade (2000 a 4000 SNPs) podem fornecer boa acurácia (correlações entre 0.8 a 0.95).

Larmer et al. (2010) apresentaram resultados de imputação de genótipos em populações de gado de leite com painéis de baixa a média densidade (6k e 50k respectivamente) para o chip HD (700k) de bovinos. E, a imputação de 6k para HD apresentou menor acurácia (correlações entre 0,88 e 0,97) que a imputação de 50k para HD (correlações entre 0,95 e 0,99), mostrando que a informação contida no chip de 50k foi melhor para recuperar a informação de haplótipos formadas para a população referência sequenciada usando o chip HD. Esta diferença foi maior em raças que possuiam menor tamanho do painel de referência (Ayrshire e Guernsey).

Marcadores igualmente espaçados ao longo do genoma também favorem o processo de imputação, fornecendo maior acurácia. Isto porque os SNPs que flanqueiam a identificação dos haplótipos estarão bem ditribuídos e representados (levando em consideração que se tenha quantidade mínima de marcadores sequenciados na população da amostra). A Illumina lançou um chip de 6k com marcadores igualmentes espaçados e maior densidade de marcadores nas extremidades para genotipagem de animais com intenção de imputação de genótipos para 50k ou mais.

2.2.4. Conexidade entre indivíduos que formam o painel de referência e para imputação

O conhecimento do parentesco entre indivíduos, se utilizado em conjunto com a informação da frequência de haplótipos (trios de pai-mão-filho, por exemplo), resulta em predições mais acuradas dos genótipos desconhecidos se comparado com o método que utiliza apenas informação de indivíduos não relacionados (Marchini et al., 2006). Porém, Browning e Browning (2011) mostram que, mesmo se indivíduos de parentesco próximo são tratados como indivíduos não aparentados, os seus haplótipos serão estimados com mais acurácia do que se fossem estimados para indivíduos verdadeiramente não aparentados.

Weigel et al. (2011), em um outro trabalho com imputação de genótipo para gado Jersey, mostram que é possível obter melhor acurácia de predição do genótipo se 50% dos touros são colocados no painel de referência.

2.2.5. Frequência alélica do marcador

Uma outra razão para se utilizar um grande número de indivíduos na população referência é assegurar que alelos raros sejam capturados, e possam ser acuradamente imputados nos indivíduos da amostra (Hayes, 2011).

De acordo com Marchini e Howie (2010) a taxa de erro da imputação aumenta com a diminuição da frequência alélica mínima. Isto ocorre em razão da dificuldade de se

identificar essas variantes raras em haplótipos, ou usá-las para flanquear a sequência do haplótico no qual está contido. Para melhorar a predição é necessário a utilização de parentes (trios, por exemplo).

2.2.6. Estrutura da população

Em razão dos fatores citados acima, a estrutura da população também afeta a acurácia de imputação. Populações diferentes irão possuir diferente número de animais no painel de referência (raças menos representativas possuem menor número, por exemplo), tamanho efetivo (se endogamia é alta, menor a frequência alélica para alguns marcadores), histórico na formação da população (a presença de "gargalos" aumenta o desequilíbrio de ligação), entre outros.

Em ovelhas, por exemplo, Hayes et al. (2011) reportaram acurácias baixas de imputação em nas raças Poll Dorset, White Suffolk e Border Leicesters (80% de correlação), Merino (71%). Isso se deve ao fato de que o nível de desequilíbrio é mais baixo nessas populações dado o maior tamanho efetivo destas população, se comparado com gado de leite, por exemplo.

2.3. Resultados da aplicação da imputação em trabalhos de pesquisa

O resultado da acurácia da imputação em si é um tema interessante, entretanto não é o resultado final utilizado em programas de seleção. Portanto, conhecer o impacto da imputação de genótipos sobre a qualidade da predição de valores genômicos é essencial para saber a real utilidade desta ferramenta. As seguintes perguntas devem ser feitas: a qualidade da predição de GEBVs com chips 3k é inferior ao de 50k, sendo este último realmente sequenciado ou imputado a partir de 3k? O quanto da habilidade de predição um chip de 50k imputado consegue captar de um chip de 50k com genótipos diretamente sequenciados? A resposta para a primeira pergunta parece ser simples, mas pode não ser tão trivial dado o problema de dimensionalidade enfrentado por modelos para seleção genômica.

Em gado de leite, resultados positivos da imputação de painéis de marcadores de baixa densidade para 50k SNPs tem sido descrito por vários autores. Weigel et al. (2010) avaliaram a acurácia da imputação de 43385 SNPs em gado Jersey quando 1, 2, 5, 10, 20, 40, ou 80% destes locos foram genotipados na população alvo (candidatos à seleção). Ambos IMPUTE 2.0 e fastPHASE foram utilizados, e os autores obtiveram baixa acurácia (<0,8) quando menos de 1000 SNPs foram usados, entretanto, quando 4000 SNPs foram usados a acurácia de imputação foi boa (0,95). Em um estudo posterior, Weigel et al. (2011) avaliaram o efeito da imputação sobre a acurácia das estimativas dos valores genéticos genômicos (GEBV). Eles concluíram que se a população de candidatos à seleção fosse genotipada com pelo menos 3000

SNPs imputando para 43000, a GEBV foi predita com 95% da acurácia captada pelos 43000 SNPs. Eles também demonstraram que utilizando genótipos imputados resultou em GEBV aproximadamente 5% mais acurada do que usando 3000 SNPs sem imputação.

Mulder et al. (2011) avaliaram a habilidade de predição de GEBV quando genótipos para alguns SNPs foram imputados ao invés de serem diretamente sequenciados. Os autores utilizaram dados de 9378 animais da raça Holandesa com chips contendo 384, 3000 ou 6000 SNPs selecionados com base na frequência alélica mínima (MAF) ou pela posição para serem imputados para 50k com DAGPHASE e outros programas. Neste trabalho, a taxa de erro de imputação foi alta para o chip de 384 SNPs, enquanto que a acurácia foi maior para os chips de 3k e 6k SNPs. Com 3k SNPs eles obtiveram aumento entre 84 e 90% da acurácia dos GEBV preditos para o chip de 50k comparado com a avaliação tradicional (apenas com o pedigree). Dassonneville et al. (2011) em um estudo semelhante em gado Holandês concluíram que a inputação de chips de 3k para 50k pode ser uma alternativa útil para a pré-seleção de animais jovens, e para a genotipagem massiva de fêmeas da população, já que a perda em acurácia de GEBV usando genótipos imputados ao invés de chips de 50k foi de apenas 0.02.

Berry e Kearney (2011) avaliaram o efeito da imputação de genótipos de chips de 3k disponíveis comercialmente a chips de 50k (51602 SNPs) em população de Holandês-Friesian utilizando o programa Beagle. Após imputação, usaram os dados completos e os dados imputados para predição de valores e nenhuma ou apenas pequena diferença entre as predições foi encontrada. Os autores concluem que esta ferramenta pode ser aplicada em programas de seleção genômica a fim de reduzir o custo.

Outra aplicação interessante da imputação de genótipos foi apresentada por Druet et al. (2010), onde duas populações, cada uma sequenciada para diferentes painéis de marcadores contendo aproximadamente 28000 SNPs, sendo aproximadamente 9000 coincidentes, tiveram a imputação de seus genótipos para 60k com taxas de erro muito baixas. O que seria uma estratégia muito útil para estudos envolvendo meta análise.

Em suínos e aves grande número de índividuos em famílias de irmãos completos é comum. De acordo com Hayes (2011), em populações como essas, uma estratégia de imputação possível seria genotipar os pais com chips de maior densidade (60k, por exemplo), e a progênie com painéis de marcadores de menor densidade (384, por exemplo). Dado o número limitado de recombinações que ocorre entre pais e progênie, este número pequeno de marcadores já seria suficiente para determinar o haplótipo da progênie dados os haplótipos dos pais, o que possibilita alta qualidade da imputação. Além disso, modelos que aproveitam a informação de parentesco, como apresentado por Habier et al. (2010), podem ser muito úteis para estas populações.

Em ovelhas, a acurácia da imputação é mais baixa do que em outras espécies (Hayes et al., 2011), portanto a sua aplicabilidade em programas de seleção deve ser avaliada com mais cautela. A maior taxa de erro de genotipagem pode causar impacto negativo sobre a predição de GEBVs para estes animais.

REFERÊNCIAS BIBLIOGRÁFICAS

1. ARONSZAJN N. Theory of reproducing kernels. Trans Am Math Soc, v.686, p.337-404, 1950.
2. BASHEER, I.A., HAJMEER, M. Artificial Neural Networks: fundamentals, computing, design, and application. Journal of Microbiological Methods, v.43, p.3-31, 2000.
3. BERRY DP, KEARNEY JF. Imputation of genotypes from low- to high-density genotyping platforms and implications for genomic selection. Animal, v.5(8), p.1162-1169, 2011.
4. BISHOP C.M. Pattern Recognition and Machine Learning. Singapura: Springer, 2006.
5. BISHOP, C.M., e Nabney, I.T. Pattern Recognition and Machine Learning: A Matlab companion. Springer, 2008.
6. BROWNING BL, BROWNING SR. A unified approach to genotype imputation and haplotype phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet, v.84, p.210-223, 2009.
7. BROWNING, S.R. Missing data Imputation and haplotype phase inference for genome-wide association studies. Hum Genet, v.124(5), p.439-450, 2008.
8. BROWNING, S.R. Multilocus Mapping Using Variable-length Markov Chains. American Journal of Human Genetics, v.78, p.903-913, 2006.
9. BROWNING, S.R., BROWNING, B.L. Haplotype phasing: existing methods and new developments. Nature Reviews Genetics, v.12, p.703-714, 2011.
10. CROSSA J, DE LOS CAMPOS G, PÉREZ P, GIANOLA D, BURGUEÑO J, ARAUS JL, MAKUMBI D, SINGH RP, DREISIGACKER S, YAN J, ARIEF V, BANZINGER M, BRAUN H. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics, v.186, p.713-724, 2010.
11. DAETWYLER, H.D., WIGGANS, G.R., HAYES, B.J., WOOLLIAMS, J.A., GODDARD M.E. Imputation of missing genotypes from sparse to high density using long-range phasing. Genetics, v.189, p.317-327, 2011.
12. DASSONNEVILLE, R., BRØDUM, R.F., DRUET, T. et al. Effect of imputing markers from low-density chip on the reliability of genomic breeding values in Holstein populations. J Dairy Sci, v.94, p.3679-3686, 2011.
13. DE LOS CAMPOS, G., GIANOLA, D., ROSA, G.J.M., WEIGEL, K.A., CROSSA, J. Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genet Res, v.92, p.295-308, 2010.

14. DE LOS CAMPOS, G., GIANOLA, D., ROSA, G.J.M. Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. J Anim Sci, v.87, p.1883-1887, 2009.
15. DE LOS CAMPOS, G., NAYA, H., GIANOLA, D., CROSSA, J. et al. Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics, v.182, p.375-385, 2009a.
16. DE LOS CAMPOS, G., HICKEY, J.M., PONG-WONG, R. et al. Whole Genome Regression and Prediction Methods Applied to Plant amd Animal Breeding. Genetics, 2012.
17. DEKKERS, J.C. Commercial application of marker and gene-assisted selection in livestock: strategies and lessons. J. Anim. Sci., v.82 E-Suppl, E313–E328. 2004.
18. DOAN, C.D., LIONG, S.Y. Generalization for Multilayer Neural Network: Bayesian Regularization or Early Stopping. In: Proceedings of Asia Pacific Association of Hidrology and Water Resources 2nd Conference. 2004.
19. DRUET, T. E GEORGES, M. A hidden Markov model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait loco fine mapping. Genetics, v.184, p.789–798, 2010.
20. DRUET, T., SCHROOTEN, C. E DE ROOS, A.P.W. Imputation of genotypes from different single nucleotide polymorphism panels in dairy cattle. Journal of Dairy Science, v.93, p.5443–5454, 2010.
21. FERNANDO, R.L., GARRICK, D.J. GenSel - User manual for a portfolio of genomic selection related analyses. Animal Breeding and Genetics, Iowa State University, Ames, 2009. http://taurus.ansci.iastate.edu/gensel
22. FORNI, S., AGUILAR, I., MIZTAL, I. Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. Genetics selection evolution. v.43, 2011.
23. GIANOLA, D., OKUT, H., WEIGEL, K.A., ROSA, G.J.M. Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. BMC Genet, v.12, p.87, 2011.
24. GIANOLA, D., VAN KAAM, J.B.C.H.M. Reproducing Kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics, v.178, p.2289-2303, 2008.
25. GIANOLA, D., DE LOS CAMPOS, G., HILL, W.G., MANFREDI, E., FERNANDO, R. Additive genetic variability and the bayesian alphabet. Genetics, v.183, p.347-63, 2009.
26. GIANOLA, D., MANFREDI, E., SIMIANER, H. On measures of association among genetic variables. Animal Genetics, v.43, p.19-35, 2012.
27. GIANOLA, D., FERNANDO, R.L., STELLA, A. Genomic-Assisted prediction of genetic value with semiparametric procedures. Genetics, v.173, p.1761-1776, 2006.
28. GODDARD, M.E., HAYES, B.J.. Review: Genomic selection. J. Anim. Breed. Genet., v. 124, p. 323-330, 2007.

29. GONZÁLEZ-CAMACHO, J.M., DE LOS CAMPOS, G., PÉREZ, P. et al. Genome-enabled prediction of genetic values using radial basis function neural networks. Theor Appl Genet, 2012, doi 10.1007/s00122-012-1868-9.
30. HABIER, D., TETENS, J., SEEFRIED, F., LICHTNER, P., THALLER, G. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genet. Sel. Evol., v.42, n.5, 2010.
31. HAYES, B.J., CHAMBERLAIN, A.C., MCPARTLAN, H. et al. Accuracy of marker assisted selection with single markers and marker haplotypes in cattle. Genome. Res., v.89(4), p.215-20, 2007.
32. HAYES, B.J. BEN HAYES COURSE NOTES. Toulouse (12-16 de setembro), 2011.
33. HAYES, B.J. QTL Mapping, MAS and Genomic Selection. Iowa State University: short course. 118p. 2007.
34. HAYES, B.J., BOWMAN, P.J., DAETWYLER, H.D. et al. Accuracy of genotype imputation in sheep breeds. Animal Genetics, v.43, p.72-80, 2011.
35. HAYKIN, S. Redes Neurais, Princípios e prática. Bookman. 1999.
36. HEBB, D.O. The organization of behavior: A neuropsychological theory. New York: Wiley, 1949.
37. HESLOT, N.H., YANG, H., SORRELS, M.E., JANNINK, J. Genomic selection in plant breeding: A comparison of models. Crop Science, v.52, p.146-160, 2012.
38. HICKEY, J.M., KINGHORN, B.P., TIER, B., Wilson, J.F. et al. A combined long-range phasing and long haplotype imputation method to impute phase for SNP genotypes. Genet Sel Evol, v.43, n.12, 2011.
39. HILL, W G. Estimation of linkage disequilibrium in randomly mating populations. Heredity, v.33, p.229-239, 1974.
40. HOWIE, B.N., DONNELLY, P., MARCHINI, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genetics, v.5(6), e1000529, 2009
41. HUANG, L., LI, Y., SINGLETON, A.B. et al. Genotype-Imputation accuracy across Worldwide Human populations. Am J Hum Genet, v.84(2), p.235-250, 2009.
42. HUANG, Y., HICKEY, J.M., CLEVELAND, M.A., MALTECCA, C. Assessment of alternative genotyping strategies to maximize imputation accuracy at minimal cost. Genetics Selection Evolution, v.44, n.25, 2012.
43. JOHNSTON, J., KISTEMAKER, G., SULLIVAN, P.G. Comparison of different imputation methods. Interbull Bulletin 44. Stavanger, Norway. Agosto 2011.
44. KÄRKKÄINEN, H.P., SILLANPÄÄ, M.J. Robustness of Bayesian multilocus association models to cryptic relatedness. Ann Hum Genet, v.76, p.510-23, 2012.
45. KONG, A., MASSON, G., FRIGGE, M.L. et al. Detection of sharing by descent, long-range phasing and haplotype imputation. Nat Genet, v.40, p.1068-1075, 2008.
46. LARMER, S., SARGOLZAEI, M., VENTURA, R., SCHENKEL, F. Imputation accuracy from low to high density using within and across breed reference populations in Holstein. Guernsey e Ayrshire cattle. Anais... GEBMAR2012. Disponível em: http://www.cdn.ca/Articles/GEBMAR2012/Imputation%20accuracy%20from%20low%20to%20high%20density%20-%20Larmer.pdf.

47. LEWONTIN, R.C. On measures of gametic disequilibrium. Genetics, v.120, p.849–852, 1984.
48. LI, N., STEPHENS, M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics, v.165, p.2213-2233, 2003.
49. LI, Y., WILLER, C., SANNA, S., ABECASIS, G. Genotype imputation. Annu Rev Genom Human Genet, v.10, p.387-406, 2009.
50. LONG, N., GIANOLA, D., ROSA, G.J.M. et al. Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality in broilers. J. Anim. Breed. Genet. v.124, p.377–389, 2007.
51. LOURAKIS, M.I.A. A Brief Description of the Levenberg-Marquadt Algorithm Implemented. 2005.
52. MACKAY, D.J.C. Bayesian interpolation. Neural Computation, v.4, p.415-447, 1992.
53. MACKAY, D.J.C. Information Theory, Inference and Learning Algorithms. Cambridge: Cambridge University Press, 2008.
54. MACKAY, D.J.C. Bayesian Methods for Neural Networks: Theory and Applications. Course Notes for Neural Networks Summer School. 1995.
55. MANGIN, B., SIBERCHICOT, A., NICOLAS, S. et al. Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. Heredity, v.108, p.285-91, 2011.
56. MARCHINI, J., HOWIE, B. Genotype imputation for genome-wide association studies. Nat Rev Genet, v.11, p.499–511, 2010.
57. Marchini, J., CUTLER, D., PATTERSON, N. et al. A comparison of phasing algorithms for trios and unrelated individuals. Am. J. Hum. Genet., v.78, p.437-450, 2006.
58. MCCULLOCH, W.S., PITTS, W. A logical calculus of the ideas immanent in neurônio activity. Bull. Math. Biophys., v.5, p.115-133, 1943.
59. MEUWISSEN, T. H. E., HAYES, B.J. E GODDARD, M.E.. Prediction of total genetic value using genomewide dense marker maps. Genetics, v.157, p.1819–1829, 2001.
60. MINSKY, M., PAPERT, S. Perceptrons. MIT press, Cambridge, MA, 1969.
61. MOROTA, G., VALENTE, B.D., ROSA, G.J.M. et al. An assessment of linkage disequilibrium in Holstein cattle using a Bayesian network. J Anim Breed Genet, v.129, p.474-87, 2012.
62. MOSER G, KHATKAR MS, HAYES BJ, RAADSMA HW: Accuracy of direct genomic values in Holstein bulss and cows using subsets of SNP markers. Genet Sel Evol, v.42, n.37, 2010.
63. MULDER, H.A., CALUS, M.P.L., DRUET, T., SCHROOTEN, C. Imputation of genotypes with low-density chips and its effect on reliability of direct genomic values in Dutch Holstein cattle. J Dairy Sci, v.95, p.876-889, 2011.
64. NEAL, R. Bayesian Learning for Neural Networks. PhD Thesis. Toronto, 1995.
65. NHO, K., SHEN, S., SWAMINATHAN, S. ET AL. The effect of reference panels and software tools on genotype imputation. AMIA Annu Symp Proc, v.2011, p.1013-8, 2011
66. NOCEDAL, J., WRIGHT, S.J. Numerical Optimization. Springer Verlag. 1999.

67. NOTHNAGEL, M.. Apresentação de slides: Genotype Imputation. 2010. Disponível em: http://www.uni-kiel.de/medinfo/lehre/vorlesungen/genepi/folien/genepi_20.pdf
68. OBER, U., AYROLES, J.F., STONE, E.A. ET AL. Using Whole-Genome sequence data to predict quantitative trait phenotypes in Drosophila melanogaster. PLoS Genet, v.8(5), e1002685. doi:10.1371/journal.pgen.1002685, 2012.
69. OKUT, H., GIANOLA, D., ROSA, G.J.M., WEIGEL, K.A. Prediction of body mass index in mice using dense molecular markers and a regularized neural network. Genet Res Camb, v.93, p.189-201, 2011.
70. PARK, T., CASELLA, G. The Bayesian LASSO. J Am Stat Assoc, v.103, p.681-686, 2008.
71. PEI, Y, LI, J., ZHANG, L., et al. Analyses and comparison of accuracy of different genotype imputation methods. PLoS ONE, v.3(10), e3551, 2008.
72. REHNBERG, E. Evaluation of a multipoint method for imputing genotypes using HapMap III. Exammensarbete 2009. Disponível em: http://www2.math.su.se/matstat/reports/serieb/2009/rep5/report.pdf.
73. ROSA, G.J.M., PADOVANI, C.R., GIANOLA, D. Robust Linear Mixed Models with Normal/Independent Distributions and Bayesian MCMC Implementation. Biometrical Journal, v.5, p.573-590, 2003.
74. ROSENBLATT, F. The Perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review, v.65, p.386-408, 1958.
75. RUMELHART, D., MCCLELLAND, J. Parallel distributed processing, vol I. MIT press: Cambridge. 1986
76. SARGOLZAEI M., F. SCHENKEL, J. CHESNAIS. Comparison between two methods of imputation (AIPL vs Boviteq/CGIL), for 6,246 genotypes provided by AIPL. In: Dairy Cattle Breeding and Genetics Committee Meeting, October 5, 2010, University of Guelph, ON, Canada. 2010.
77. SCHEET, P., STEPHENS, M. A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet, v.78, p.629-644, 2006.
78. SUN, C., WU, X., WEIGEL, K. et al. An ensemble-based approach to imputation of moderate-density genotypes for genomic selection with application to Angus cattle. Genetics Research, v.94, p.133-150, 2012.
79. SUZUKI, H. Artificial neural networks: methodological advances and biomedical applications. Croatia: Intech. 2011.
80. THODBERG, H. Improving Generalization of neural networks through pruning. International Journal of Neural Systems, v.1, p.317-326, 1990.
81. TIBSHIRANI, R. Regression shrinkage and selection via the LASSO. J R Stat Soc B, v.58, p.267-288, 1996.
82. VAN RADEN, P.M., O'CONNELL, J.R., WIGGANS, G.R., WEIGEL, K.A. Genomic evaluation with many more genotypes. Genet. Sel. Evol., v.43, n.10, 2011.
83. VANRADEN, P.M. Efficient methods to compute genomic predictions. J. Dairy Sci.. v. 91, p.4414-4423, 2008.

84. VENTURA, R.R., SILVA, M.A., MEDEIROS, T.H. et al. Uso de redes neurais artificiais na predição de valores genéticos para peso aos 205 dias em bovinos da raça Tabapuã. Arq. Bras. Med. Vet. Zootec, v.64, p.411-418, 2012.
85. WAHBA, G. Spline models for observational data. Soc Ind Appl Math, 1990.
86. WANG, C. S., RUTLEDGE, J.J., GIANOLA, D., Marginal inferences about variance components in a mixed linear model using Gibbs sampling. Genet. Sel. Evol., v.25, p.41–62, 1993.
87. WEIGEL, K.A., DE LOS CAMPOS, G., VAZQUEZ, A.I. et al. Accuracy of direct genomic values derived from imputed single nucleotide polymorphism genotypes in Jersey cattle. J Dairy Sci, v.93, p.5423-5435, 2011.
88. WEIGEL, K.A., VAN TASSELL, C.P., O'CONNELL, J.R. et al. Prediction of unobserved single nucleotide polymorphism genotypes of Jersey cattle using reference panels and population-based imputation algorithms. J Dairy Sci, v.93, p.2229-2238, 2010.
89. ZHANG, B., ZHI, D., ZHANG, K., GAO, G. et al. Practical consideration of genotype imputation: Sample size, window size, reference choice, and untyped rate. Stat Interface, v.4(3), p.339-352, 2011.

CAPÍTULO 2

Efeito da imputação de genótipos sobre a predição de características quantitativas de ratos utilizando marcadores genéticos

Resumo: Objetivou-se avaliar o efeito da imputação de genótipos sobre a acurácia de predição de fenótipos. A hipótese é que todo o sinal genético e informação disponível está contida completamente no banco de dados contendo genótipos observados. Um banco de dados de ratos disponível na internet contendo informações de 1809 SNPs igualmente espaçados ao longo do genoma e 1881 indivíduos foi utilizado. As características consideradas foram peso corporal e índice de massa corporal. E, a partir do arquivo completo de genótipos, apenas 201, 453 ou 905 dos SNPs foram selecionados como genotipados (taxas de mascaramento de 90, 75 e 50%, respectivamente), sendo o restante imputado pelo software Beagle. Em seguida, foram ajustados o LASSO bayesiano (LB), Reproducing Kernel Hilbert Spaces (RKHS) e Redes Neurais Artificiais com Regularização Bayesiana (RNARB) utilizando parte dos SNPs e o painel completo com e sem imputação de genótipos. O método RKHS obteve a melhor acurácia de predição. A imputação de genótipos pareceu ter o mesmo efeito sobre a eficiência do BL e RKHS, enquanto que a RNARB resultou em predições mais sensíveis aos erros de imputação. Em cenários em que a acurácia de imputação de genótipos foi boa e taxas de mascaramento de genótipo de 75 e 50%, a imputação não foi muito vantajosa. Entretanto, quando a informação do genótipo era esparsa (90% dos genótipos mascarados), a imputação de genótipos foi capaz de trazer informação a respeito de marcadores importantes e melhorou, assim, a habilidade de predição do modelo. Os resultados obtidos mostram que nem sempre a imputação de genótipos desconhecidos é vantajosa para a predição de fenótipos. O ganho em se imputar genótipos dependerá da conexidade entre dados de treinamento e candidatos à seleção, herdabilidade da característica, número de marcadores disponíveis no painel original e o método utilizados para predizer o efeito dos marcadores.

Palavras chave: imputação, modelo não-paramétrico, seleção genômica

[Effect of Genotype Imputation on Genome-Enabled Prediction of Quantitative Traits in mice]

Abstract: *The goal was to evaluate the effect of genotype imputation on phenotypes prediction accuracy. The hypothesis underlying this work was that all genetic signal and information*

available in a data set is contained entirely in the observed genotypes. A publicly available dataset on mice was used, with information of 1,809 SNPs equally spaced along the genome of 1,881 animals. The traits considered were body weight and body mass index. And, from the full set of SNPs, only 201 (90% masking rate), 453(75% masking rate) or 905 (50% masking rate) were selected as the genotyped SNPs, with the remaining marker imputed using the Beagle software. Then, Bayesian Lasso (BL), Reproducing Kernel Hilbert Spaces (RKHS) and Bayesian Regularized Artificial Neural Networks (BRANN) were fitted using the subsets and the full panel of SNPs with and without genotype imputation. RKHS method showed the best predictive accuracy. Genotype imputation seemed to have the same effect on efficiency of BL and RKHS, whereas BRANN resulted in more sensible predictions due to imputation error. In scenarios which genotype imputation accuracy was good and masking rates of 75% and 50%, the genotype imputation did not bring great benefit. However, when genotype information was sparse (90% masking), imputation of genotypes brought information about important markers and improved predictive ability. The obtained results show that not always the imputation of unknown genotypes is advantageous for phenotypic prediction. The gain of imputing genotypes will depend on the connectedness between reference population and selection candidates, heritability of the trait, number of markers available in the original panel, and the method used to predict marker effects.

Keywords: *genomic selection, imputation, non-parametric model*

INTRODUÇÃO

A seleção genômica é tema recente nas áreas de produção animal e de plantas e tem sido aplicada para predição de valores genéticos para fins de seleção (Meuwissen et al., 2001 e Goddard e Hayes, 2007) e também para decisões de manejo baseadas no fenótipo predito (Lee et al., 2008, Weigel et al., 2010). Uma estratégia comum nesta área é a imputação de marcadores em chips de baixa densidade para maiores densidades a fim de melhorar a acurácia preditiva dos modelos e estimar efeito de substituição do alelo de todos os indivíduos ao mesmo tempo (Weigel et al., 2010, Mulder et al., 2011). Alternativamente, alguns autores propõem o uso da informação de co-segregação proveniente de chips de menor densidade contendo marcadores igualmente espaçados ou selecionados por efeito para captar sinais de alelos SNPs em alta densidade (Habier et al., 2009). E, como apresentado por Weigel et al. (2009), um painel de baixa densidade contendo SNPs selecionados pode fornecer grande parte da acurácia de uma avaliação utilizando painel de alta densidade.

A vantagem em imputar genótipos irá depender da acurácia de imputação (Weigel et al., 2010), estrutura da população (Weigel et al., 2010, Dassonneville et al., 2011) e arquitetura genética da característica alvo (Moser et al., 2010). Esta ferramenta pode ser útil para revelar genótipos não contidos nos *chips* ou não identificados pelas plataformas de sequenciamento e agrupar indivídios genotipados com *chips* diferentes e possui a vantagem econômica para obtenção de marcadores em alta densidade a partir de *chips* de baixo custo. Vários estudos mostram que os métodos e softwares disponíveis fornecem boa acurácia de imputação de genótipos (Mulder et al., 2011, Browning e Browning, 2011 e Calus et al., 2011). Portanto, esta metodologia tem sido considerada uma boa alternativa para reduzir o custo de genotipagem de animais e é sugerida para aplicações comeciais como a triagem de touros jovens para serem incluídos no teste de progênie e vacas (Weigel et al., 2010).

Com respeito aos modelos para análise de dados genômicos, o interesse em métodos semi e não-paramétricos para predição de características complexas têm aumentado, dentre eles o regressões Reproducing Kernel Hilbert Spaces (RKHS) sobre marcadores (Gianola e Kaam, 2008, de los Campos et al., 2009 e de los Campos et al., 2010), funções de base radial (Long et al., 2010, González-Camacho et al., 2012), e redes neurais artificiais (Okut et al., 2011, Gianola et al., 2011). Gianola et al. (2011) justificam a utilização de tais modelos pois essas regressões não-paramétricas possuem a capacidade de capturar interações complexas e não-linearidades entre marcadores, o que é impossível utilizando regressões lineares com inferência Bayesiana.

O objetivo-se com o trabalho explorar se modelos mais elaborados, como métodos semi e não-paramétricos, poderiam capturar sinais de *chips* de baixa densidade sem a necessidade de imputar genótipos. Já que a seleção genômica, ao contrário dos estudos de associação utilizando informação genômica ampla (genome-wide association studies – GWAS), não objetiva encontrar regiões específicas do genoma que afetam as características avaliadas, mas somente predizer valores genômicos ou fenótipos. Por exemplo, Okut et al. (2011) utilizaram redes neurais artificiais com regularização bayesiana (BRANN) incluindo 798 SNPs para predizer o índice de massa corporal (IMC) em ratos e obtiveram uma correlação global entre o fenótipo observado e predito variando entre 0,25 e 0,5. Resultados semelhantes foram obtidos por de los Campos et al., (2009) utilizando o LASSO bayesiano, porém incluindo 10.946 SNPs. Objetivamente, neste trabalho, o efeito da imputação sobre a habilidade de predição de características complexas em ratos utilizando métodos paramétrico, semi e não-paramétricos foi avaliado.

MATERIAL E MÉTODOS

Dados

Um banco de dados em ratos disponível ao público (http://gscan.well.ox.ac.uk/) foi utilizado. A descrição completa da população e outros detalhes se encontram em Mott et al. (2006), Valdar et al. (2006a), Valdar et al. (2006b). Estes dados também foram utilizados por outros estudos dentro do tópico seleção genômica para ajuste de regressão bayesiana (Legarra et al., 2008 e de los Campos et al., 2009).

Somente animais com informação fenotípica e genotípica (call rate maior que 95%) foram considerados para análise. Locos com frequência alélica mínima menor que 0,05, *call rate* menor que 95% ou fora no equilíbrio de Hardy-Weinberg foram descartados do arquivo de dados. As duas características, peso corporal às 10 semanas de idade (PC) e índice de massa corporal (IMC) foram pré-corrigidas pelo ajuste do seguinte modelo linear misto:

$$y = X\beta + Zu + Wc + e,$$

em que y é o vetor das observações para o fenótipo mensurado (PC ou IMC); β é o vetor de efeitos fixos para idade, sexo e densidade da gaiola; $\hat{u}$ é o vetor dos efeitos genéticos aditivos; c é o vetor para efeito de gaiola; X, Z and W são as matrizes de incidência para os efeitos fixos, genético aleatório e efeito aleatório de gaiola; e e é o vetor de resíduos aleatórios com distribuição multivariada normal $e \sim N(0, I\sigma_e^2)$. Admite-se que os efeitos aleatórios genético aditivo e de gaiola eram independentes com distribuições $u \sim N(0, A\sigma_u^2)$ e $c \sim N(0, I_c\sigma_c^2)$, respectivamente, em que A é a matriz dos coeficientes de parentesco e I_c é matriz identidade em que 531 é o número de gaiolas. Como discutido por Legarra et al. (2008), existe confundimento entre efeito de família e efeito de gaiola já que a maioria dos indivíduos alocados na mesma gaiola são irmãos completos, portanto é possível que o efeito genético aditivo esteja subestimado. Entretanto, é pertinente assumir que isto impactará de maneira semelhante a habilidade de predição dos diferentes modelos considerados neste estudo. Portanto, a variável alvo, após correção, é $\hat{t} = Z\hat{u} + \hat{e}$, que presumidamente inclui todos os tipos de efeitos genéticos (aditivo, dominância, epistasia) assim com o efeitos ambientais não explicados pelo modelo misto descrito acima.

Do arquivo de genótipos, 1809, 905, 453 e 201 SNPs igualmente espaçados foram selecionados de 10.348 SNPs. No total, 1881 e 1823 indivíduos foram considerados para análise para PC e IMC, respectivamente. Para cada caso, aproximadamente 2/3 dos indivíduos foram designados como população de treinamento (ou população referência) e 1/3 como teste ou candidatos à seleção (Tabela 1), para comparação de modelos via validação cruzada. Os

arquivos de treinamento e teste foram divididos considerando a informação de família de duas maneiras: entre famílias,em que algumas famílias foram utilizados para treinamento dos modelos e outras para o teste, e dentro de famílias, em que indivíduos de uma mesma família foram distribuídos entre treinamento e teste. Então, a imputação de genótipos e a predição de fenótipos foi realizada para todos os cenários considerando característica altamente e medianamente herdável (PC e IMC, respectivamente), arquivos contendo painéis reduzidos (201, 453 ou 905 SNPs) e com imputação de genótipos, distribuição de famílias entre os arquivos de treinamento e teste e os diferentes modelos preditivos.

Tabela 1. Distribuição de indivíduos por arquivo analisado

Característica[(1)]	Entre famílias		Dentro de famílias		No. total de indivíduos
	Arquivo de treinamento	Arquivo de teste	Arquivo de treinamento	Arquivo de teste	
PC	1200	681	1200	681	1881
IMC	1165	658	1161	662	1823

[(1)]PC: Peso corporal na 10ª semana de idade; IMC: índice de massa corporal.

Imputação de genótipos

Os arquivos para teste (candidatos à seleção) contendo 201, 453 e 905 SNPs foram imputados para 1809 SNPs utilizando o software Beagle (Browning e Browning, 2009), que utiliza modelo de agrupamento localizado de haplótipos. Este software é baseado em moldelos de cadeia de Markov oculta que agrupa os haplótipos em cada loco e o agrupamento se adapta à quantidade de informação disponível de forma que o número de grupos aumenta globalmente com o tamanho da amostra e localmente com aumento do desequilíbrio de ligação (Browning e Browning, 2011). A imputação de genótipos foi realizada em ambos os cenário de distribuição de famílias entre a população referência e a de candidatos à seleção ignorando a informação de parentesco. A acurácia da imputação foi avaliada comparando o arquivo de dados completo, sem mascaramento de genótipos, e o arquivo com genótipos mascarados e posteriormente imputados. O número de acertos foi então dividido pelo número total de SNPs, dando a porcentagem de genótipos imputados corretamente.

LASSO bayesiano (LB)

Tibshirani (1996) propôs um método de regressão, conhecido como *Least Angle Shrinkage Selection Operator* (LASSO), que combina seleção de variáveis e encurtamento baseado em regularização ao mesmo tempo. Neste modelo, um termo de penalidade, proporcional à norma do vetor contendo os coeficientes de regressão, é adicionado à fórmula do problema de

otimização, permitindo a seleção de variáveis e o encurtamento dos coeficientes de forma simultânea. O problema de otimização é descrito por:

$$\min_{\beta}\{\sum (y_i - x_i'\beta)^2 + \lambda \sum_j \|\beta\|_j^2\},$$

em que a primeira parte se refere à solução de quadrados mínimos e a segunda é o fator de penalização λ, com valor esperado maior que 0. Um λ mais alto proporciona meior encurtamento e alguns β's podem ser até zerados (seleção de variável).

A versão bayesiana do LASSO foi proposta por Park e Casella (2008) que descreveram a implementação para o amostrador de Gibbs deste método. Em sua interpretação bayesiana, o LASSO pode ser visto como a moda a posteriori de um modelo Bayesiano com verossimilhança gaussina, $p(y|\beta,\sigma_e^2) = \prod_{i=1}^{n} N(y_i|x_i'\beta,\sigma_e^2)$ e a priori para β é o produto de p densidades exponenciais duplas independentes e com média igual a zero (de los Campos et al., 2009).

A distribuição exponencial dupla (ou Laplace) possui representação hierárquica conveniente como mistura de densidades Gaussianas escaladas (Rosa et al., 2003). De acordo com Park e Casella (2008):

$$\beta_j \sim DE(\beta_j \mid \lambda) = \frac{\lambda}{2} e^{-\lambda|\beta_j|}$$

$$= \int_0^{\infty} \left[\frac{1}{\sqrt{2\pi\sigma_j^2}} e^{-(\beta_j^2/2\sigma_j^2)} \right] \left[\frac{\lambda^2}{2} e^{-\lambda^2/2\sigma_j^2} \right] d\sigma_j^2,$$

e as *prioris* para os parâmetros, seguindo de los Campos et al. (2009), são:

Priori:

$$p(\beta,\sigma_\varepsilon^2,\tau^2,\lambda^2) = p(\beta|\sigma_\varepsilon^2,\tau^2)\,p(\sigma_\varepsilon^2)\,p(\tau^2|\lambda)\,p(\lambda^2)$$

$$= \left[\prod_{j=1}^{p} N(\beta_j|0,\tau^2\sigma_\varepsilon^2) \right] \chi^{-2}(\sigma_\varepsilon^2 \mid d.f.,S)$$

$$x \left[\prod_{j=1}^{p} \exp(\tau_j^2 \mid \lambda) \right] G(\lambda^2 \mid \alpha_1,\alpha_2)$$

em que, $N(y|x_i'\beta,\sigma_\varepsilon^2)$ e $N(\beta_j|0,\tau^2\sigma_\varepsilon^2)$ são as densidades normais atribuídas com médias $x_i'\beta$ e 0, e variâncias σ_ε^2 e $\tau^2\sigma_\varepsilon^2$ para o fenótipo e efeitos dos marcadores, respectivamente. Uma distribuição qui-quadrada invertida com parâmetro de escala $\chi^{-2}(\sigma_\varepsilon^2 \mid g.l.,S)$ com *g.l.* graus de liberdade e *S* como parâmetro de escala é usada nesta parametrização. Uma distribuição exponencial dupla $\exp(\tau_j^2 \mid \lambda)$ indexada por λ e $G(\lambda^2 \mid \alpha_1,\alpha_2)$ é a distribuição Gamma com parâmetro de forma α_1 e taxa α_2. O parâmetro λ, também conhecido como parâmetro de suavização, tem papel central no modelo pois controla o balanço entre qualidade de ajuste e

complexidade do modelo ajustado. A medida que seu valor se aproxima de 0, a solução se aproxima da solução de quadrados mínimos; enquanto que valores mais altos fornecem *priori* mais estreita para β e maior encurtamento das estimativas. Comparado à regressão bayesiana, este modelo possui a vantagem de permitir maior densidade para marcadores com efeito zero, o que pode ser mais plausível biologicamente.

Este modelo foi treinado em todos os cenários considerados (diferentes densidades dos *chips* com e sem imputação e distribuição de famílias entre arquivos de treinamento e teste). Inferências se basearam em 70.000 amostras após burn-in das 5.000 primeiras amostras. Os parâmetros da distribuição *a priori* foram: $S_{\varepsilon} = d.f._{\varepsilon} = S_{u} = d.f._{u} = 1$ e $p(\lambda^2) \propto Gamma(1.2, 1e-5)$. O pacote BLR (Pérez et al., 2010), desenvolvido para o software R, foi utilizado para as análises. Os modelos treinados foram então utilizados para predizer fenótipos da população de teste, e a habilidade de predição foi avaliada pela correlação entre o fenótipo mensurado e predito e o quadrado médio do erro de predição (QMEP).

Reproducing Kernel Hilbert Space (RKHS)

A teoria do RKHS foi desenvolvida por Aronszajn (1950) e tem sido aplicada em diversas áreas da estatística e aprendizado de máquinas (como Support vector machines) por vários anos; a base teórica é fornecida por Wahba (1990). Este modelo semi-paramétrico foi proposto por Gianola, Fernando e Stella (2006) e Gianola e Kaam (2008) no contexto de regressão de fenótipos sobre os genótipos.

O RKHS apresenta a propriedade de possuir um espaço infinito de funções para busca da dependência entre a variável de entra e as variáveis alvo que é definido pela medida de distância utilizada (neste caso o tipo de kernel) sem nenhuma necessidade de pressuposições adicionais a respeito da relação fenótipo-genótipo. O método consiste da combinação do modelo aditivo clássico com uma função desconhecida dos marcadores, que é inferida não-parametricamente, e tem o potencial de capturar interações complexas sem modelá-las explicitamente (Gianola et al., 2006). Para mapear a função entre entradas (genótipos) e variáveis alvo (fenótipos), uma coleção de funções definidas no espaço de Hilbert ($f \in H$) é usada, da qual um elemento $\hat{f}$ é escolhido com base em algum critério (quadrado médio do erro penalizado ou densidade *a posteriori*, por exemplo) (de los Campos et al., 2010). O problema de otimização para obtenção das estimativas de RKHS é similar ao descrito para LB, que é:

$$\hat{f} = \arg\min_{f \in H} \{ l(f, y) + \lambda \| f \|_H^2 \},$$

em que $l(f, y)$ é a função de perda representando a medida de qualidade de ajuste; $\|f\|_H^2$ é o quadrado da norma de f que está relacionado com a complexidade do modelo; e λ controla o balanço entre ambos os termos apresentados anteriormente.

De acordo com o teorema de Moore-Aronsajn, cada RKHS é associado a um único kernel positivo definido. No RKHS, os marcadores são utilizados para construir uma matriz de covariância que mede as distâncias entre os genótipos, em que $Cov(g_i, g_{i'}) \sim K(x_i, x_{i'})$, com x_i e $x_{i'}$ representando vetores contendo os genótipos para os indivíduos *i* e *i'* , e $K(.,.)$ é o Reproducing Kernel (RK) relacionado à função positiva definida (de los Campos, 2010).

Neste trabalho, a matriz de kernel (K) foi definina como Kernel Gaussiano $K(x_i, x_{i'}) = \exp\{-h \times d_{ii'}\}$, em que h é o parâmetro de largura de banda e $d_{ii'} = \sum_{k=1}^{p} (x_{ik} - x_{i'k})^2$ representa cada elemento da matriz de distância Euclidiana ao quadrado. A escolha de *h* é feita via comparação de modelos e deve considerar a distribuição observada em $d_{ii'}$. Neste estudo, foi realizada a média de kérneis a fim de "automatizar" a escolha do kernel baseado na mediana amostral de $d_{ii'}$, como descrito por Crossa et al. (2010). Portanto, $h = a \times q_{0,5}^{-1}$ em que *a* foi igualado a -5, -1 e -1/5, e $q_{0,5}$ é a mediana amostral de $d_{ii'}$, para os três kérneis utilizados. Neste modelo, os valores genéticos foram a soma de três componentes $g = f_1 + f_2 + f_3$, sendo $p(f_1, f_2, f_3 | \sigma^2_{\alpha,1}, \sigma^2_{\alpha,2}, \sigma^2_{\alpha,3}) = N(f_1 | 0, K_1 \sigma^2_{\alpha,1}) N(f_2 | 0, K_2 \sigma^2_{\alpha,2}) N(f_3 | 0, K_3 \sigma^2_{\alpha,3})$. A parâmetros de variância destes componentes foram considerados desconhecidos e receberam distribuições *a priori* como qui-quadrada invertida com parâmetro de escala idênticas e independentes com graus de liberdade e parâmetros de escala iguais a $df=5$ e $S = (\text{var}(y) / 2 \times (df - 2))$, respectivamente. Amostras da distribuição *a posteriori* foram obtidas por meio do amostrador de Gibbs (de los Campos et al., 2010). Inferências foram baseadas em 45.000 amostras após o burn-in de 5000 amostras.

Redes Neurais Artificiais com Regularização Bayesiana (RNARB)

Uma RNARB é uma rede feed-forward que envolve estimativa da máxima a posteriori em que o regularizador pode ser visto como o logaritmo de uma distribuição *a priori* dos parâmetros (Bishop, 1996). Portanto, este modelo atribui uma distribuição de probabilidade aos pesos e vieses da rede, dessa forma as predições são realizadas no contexto bayesiano o que melhora a generalização dos resultados. Maiores detalhes são apresentados por Mackay (1995).

Uma simples rede feed-forward utiliza pesos e vieses iniciais e transforma a informação fornecida na entrada (neste caso o código para os genótipos) através de cada neurônio de conexão presente na camada oculta (também conhecida como "caixinha preta") utilizando uma

função de ativação. A informação é então enviada para o neurônio na camada de saída e novamente tranformado utilizando outra função de ativação, gerando, assim, a saída (variável alvo). Posteriormente, os resultados são retropropagados para atualizar os pesos e vieses por meio de derivadas. Portanto, nenhuma pressuposição com respeito à relação entre genótipo (entrada) e fenótipo (saída) é feita neste tipo de modelo. Após o treinamento, as saídas são calculadas como:

$$\hat{t}_i = g\{\sum_{k=1}^{s} w_k f(\sum_{j=1}^{R} w_{kj} p_j + b_k^1) + b^2\},$$

em que $\hat{t}_i$ é o fenótipo estimado; g e f são as funções de ativação das camadas de saída e oculta, respectivamente; w_k e w_{kj} são os pesos saindo dos neurônios da camada oculta e enviando informações para a camada de saída e da entrada para a camada oculta, respectivamente; e b_k^1 e b^2 são os vieses.

Treinamento é o processo pelo qual os pesos são modificados à luz dos dados enquanto a rede tenta gerar o resultado ótimo (Okut et al., 2011). E, após o treinamento, a rede pode ser utilizada para predizer fenótipos desconhecidos a partir das informações contidas nos *chips* de DNA.

Na RNARB, adicionalmente a medida típica de performance dada pela soma de quadrado do erro, uma penalidade é imposta a pesos de valor alto a fim de se obter um mapeamento mais suave. A função objetiva é:

$$f = \beta E_D(D|w,M) + \alpha E_w(w|M),$$

onde $E_D(D|w,M)$ é a soma de quadrados dos erros de predição (D: dados, w:pesos, M:arquitetura), $E_w(w|M)$, conhecido como weight decay, é a soma de quadrado dos pesos da rede, e α e β são os parâmetros de regularização, que controlam o balanço entre qualidade de ajuste e suavização.

A distribuição a posteriori de w dados α, β, D e M é (Mackay, 2008):

$$P(w|D,\alpha,\beta,M) = \frac{P(D|w,\beta,M)P(w|\alpha,M)}{P(D|\alpha,\beta,M)}$$

Seguindo o teorema de Bayes, $P(D|w,\beta,M)$ é a função de verossimilhança, $P(w|\alpha,M)$ é a *priori* para os pesos considerando a arquitetura escolhida para a rede, e $P(D|\alpha,\beta,M)$ é o fator de normalização.

Para evitar sobreajuste, a arquitetura da rede e o número de *epochs* (iterações) foram previamente testados. Uma rede contendo 5 neurônios na camada oculta e função de ativação

tangente sigmóide e 1 neurônio na camada de saída com função de ativação linear foi utilizada (Figura 1). O número de iterações foi de 30. Para remover a influência dos valores iniciais, os resultados foram a média de 20 análises repetidas. Para melhorar a generalização, early stopping foi testado para regularização da rede porém a regularização bayesiana mostrou melhores resultados. O software MATLAB (Demuth et al., 2009) foi utilizado para as análises. Assim como LB e RKHS, a habilidade de predição foi avaliada com base na correlação entre fenótipos predito e mensurado e quadrado médio do erro de predição.

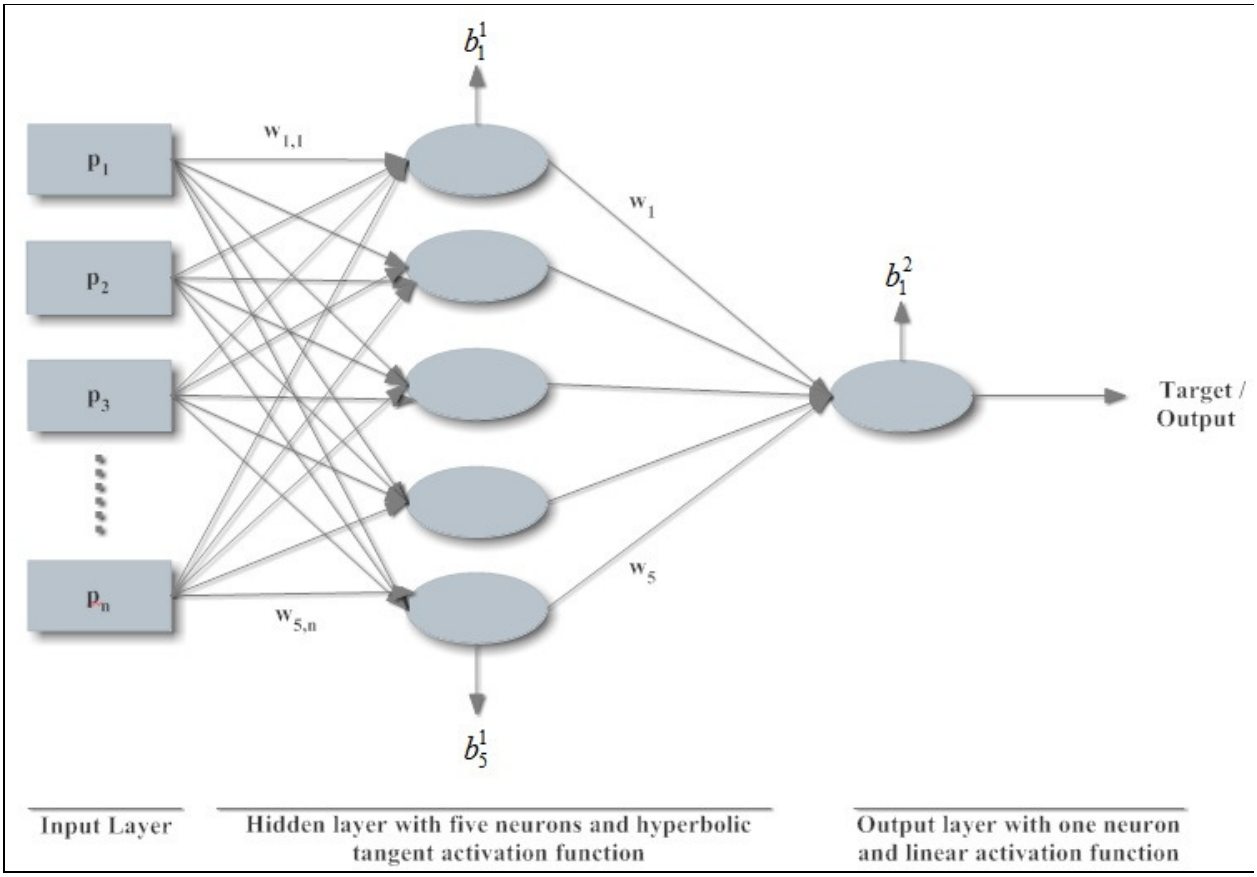

Figura 1. Arquitetura de uma RNA com duas camadas contendo 5 neurônios na camada oculta e um neurônio na camada de saída. Os p_i são as entradas para cada animal em que o número de colunas é igual ao número de SNPs; os $w_{j,i}$ são os pesos em que j é o indicador da camada oculta e i identifica o animal; b_l^k são os vieses em que l identifica o neurônio e k identifica a camada.

RESULTADOS

Acurácia de imputação

Foi verificada boa acurácia de imputação de genótipos em todos os cenários testados (Tabela 2). A menor acurácia de imputação foi obtida na situação em que aproximadamente 90% dos genótipos foram mascarados e o painel de referência continha indivíduos não relacionados ao indivíduos no arquivo para imputação (75% de acurácia). Pode ser observado que, mesmo que o Beagle não utilize diretamente a informação do parentesco, a maior conexidade entre indivíduos no panel de referência e no arquivo contendo genótipos não identificados trouxe melhora na acurácia de imputação. A explicação pode ser a similaridade de padrão de desequilíbrio de ligação, o que facilita a identificação de haplótipos, entre o arquivo a ser imputado e o painel de referência que serve como base para imputação dos genótipos desconhecidos. O erro mais comum foi o confundimento entre heterozigoto e homozigoto para o alelo de maior frequência (em média 65% do total de erros).

Tabela 2. Acurácia de imputação e taxa de erro para 90, 75 e 50% de genótipos mascarados

	90%		75%		50%	
	Entre famílias	Dentro de famílias	Entre famílias	Dentro de famílias	Entre famílias	Dentro de famílias
Accurácia	75%	79%	91%	93,8%	97,2%	98,1%
Erro 0↔1[a]	21%	22%	24%	27%	27%	21%
Erro 1↔2[b]	67%	68%	67%	67%	64%	56%
Erro 0↔2[c]	12%	1%	9%	6%	9%	13%

[a] Erro pela troca entre 0 e 1; [b] Erro pela troca entre 1 e 2; [c] Erro pela troca entre 0 e 2

Habilidade de predição considerando taxa de mascaramento de genótipos, método e informação de família

Os resultados para correlação entre fenótipos predito e mensurado (PC e IMC) no arquivo de teste estão descritos nas Figuras 2 e 3 e Tabelas 3 e 4. A distribuição de famílias entre dados de treinamento (referência) e dados para teste (candidatos à seleção) afetou a habilidade de predição de todos os modelos ajustados, maior conexidade entre estes dois arquivos forneceu melhor acurácia de predição, como esperado. Na Figura 3 verifica-se maior correlação média entre fenótipos predito e mensurado para as estimativas dentro de famílias. Portanto, a informação proveniente de indivíduos com parentesco próximo para estimativa de efeitos dos SNPs foi vantajosa para a predição de novos fenótipos. Resultados similares foram apresentados por Legarra et al. (2008) e os autores concluíram que o arquivo de dados dentro de famílias é fonte mais acurada de informação do que entre famílias para a avaliação

genômica ampla e seleção. Portanto, esta informação é um ponto importante para delineamento de programas de seleção.

Como esperado, houve maior habilidade de predição para a característica PC comparada à IMC, já que a última possui menor herdabilidade. O melhor método de predição neste estudo foi o RKHS utilizando média de kérneis, e o pior foi a RNARB, provavelmente em razão de problemas com sobreajuste.

O efeito da imputação de genótipos sobre a habilidade de predição dos modelos ajustados dependeu da estrutura dos dados, herdabilidade da característica e método. Nas Figuras 2 e 3 verifica-se que a imputação pareceu ser mais vantajosa para a predição de fenótipos quando a conexidade entre dados de treinamento e teste é baixa, e a característica possuía herdabilidade mais alta (PC, no caso), especialmente para os modelos BL e RKHS. Para IMC este benefício foi menos evidente e, em alguns cenários, a imputação de genótipos deteriorou a habilidade de predição quando a acurácia de imputação foi baixa. Com respeito aos métodos, a imputação teve efeito semelhante sobre a eficiência do LB e RKHS, enquanto que a RNARB resultou em predições mais sensíveis aos erros de imputação. Para outros cenários, com boa acurácia de imputação e taxas de mascaramento de 75% e 50%, a imputação de genótipos não trouxe muita informação, o que pode ser observados nas Figuras 2B, 3A e 3B. Entretanto, quando a informação genotípica era esparsa (201 SNPs), a imputação trouxe informação de marcadores importantes para melhorar a predição de fenótipos (Figuras 2A e 3B).

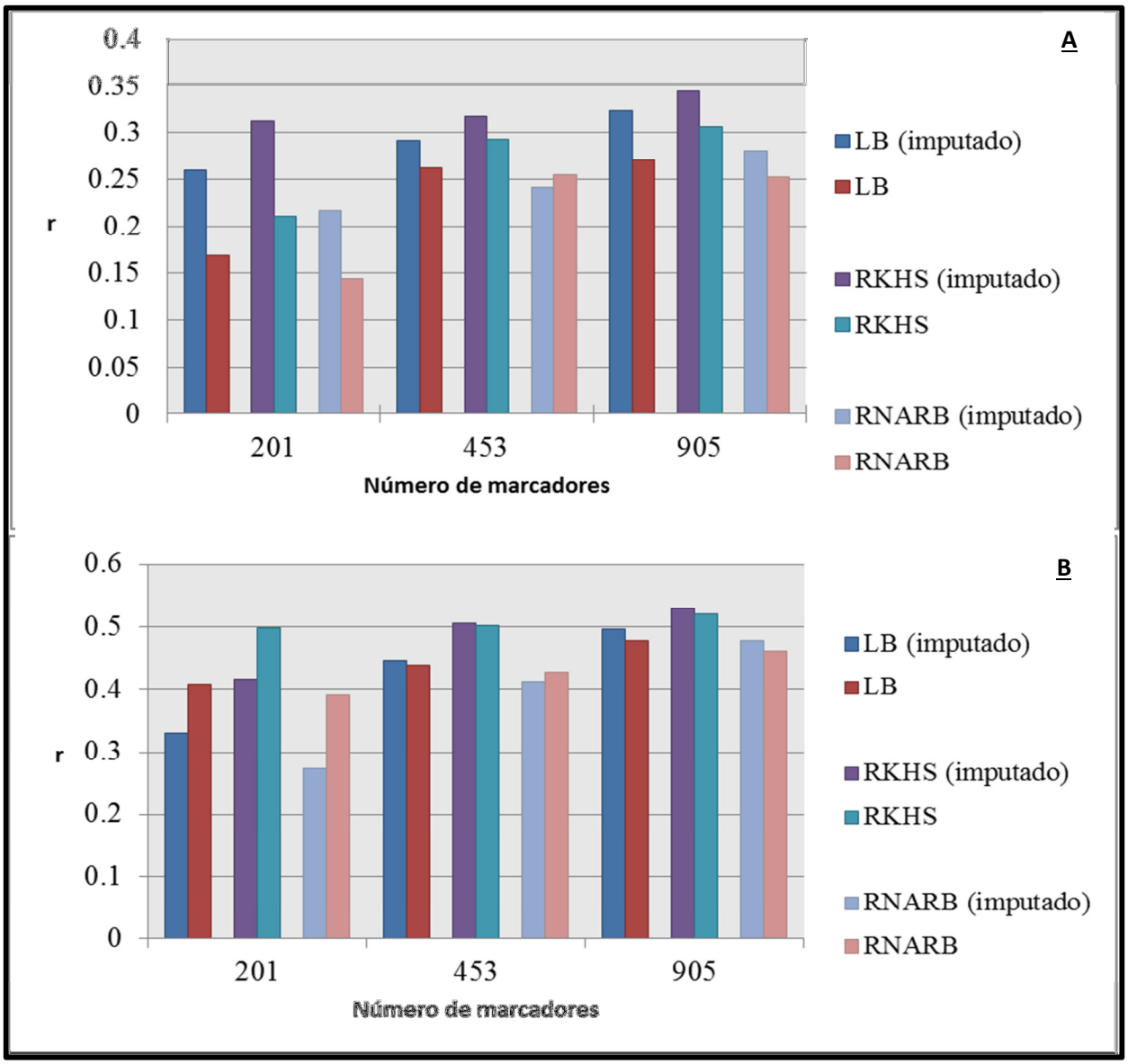

Figura 2. Correlações entre PC predito e mensurado entre famílias (A) e dentro de famílias (B) para o LASSO bayesiano (LB), Reproducing Kernel Hilbert Spaces (RKHS) e Redes Neurais Artificiais com Regularização Bayesiana (RNARB) para arquivos de teste contendo marcadores imputados e não-imputados (painel reduzido) contendo diferentes densidade de genótipos identificados.

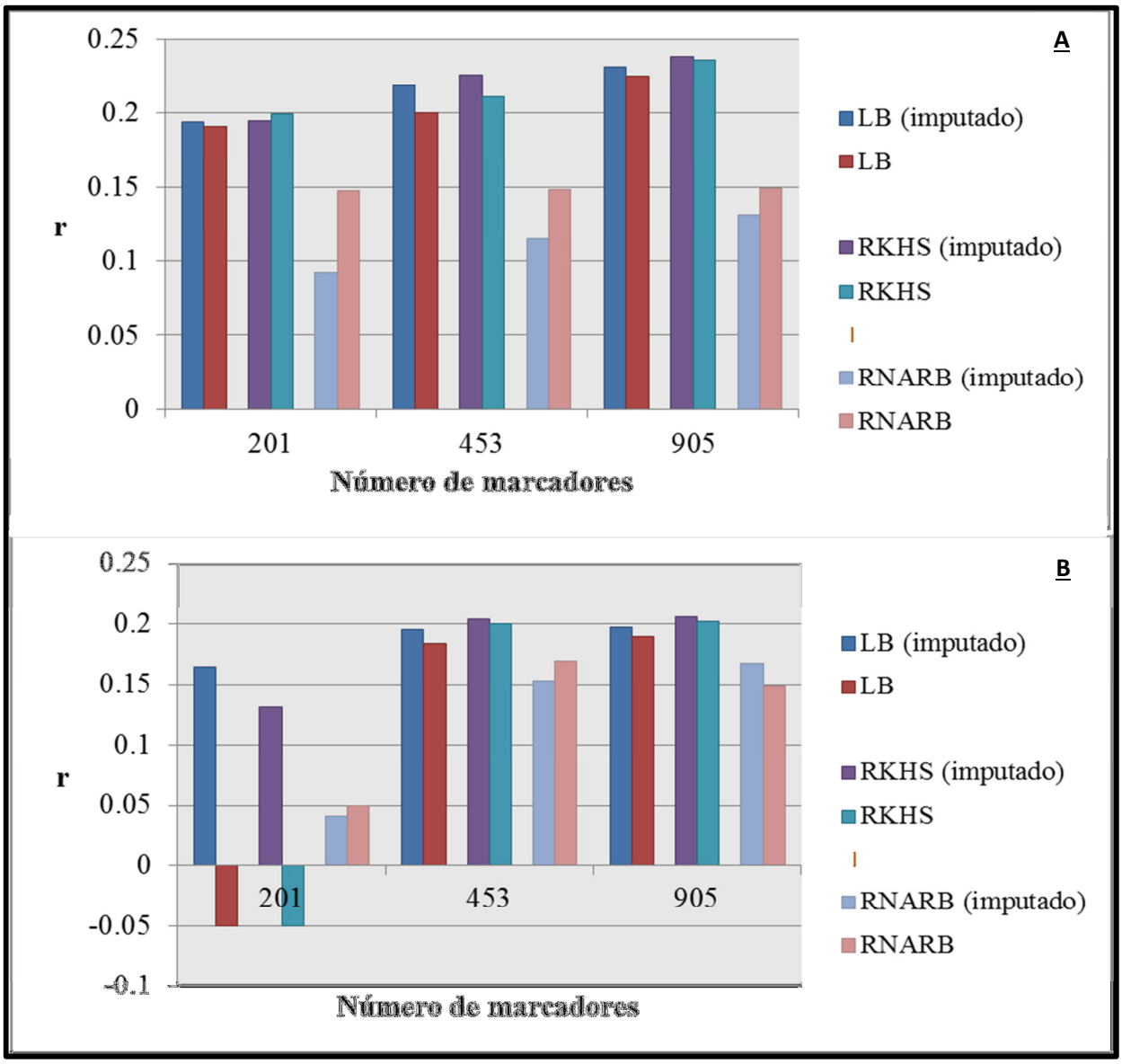

Figura 3. Correlações entre IMC predito e mensurado entre famílias (A) e dentro de famílias (B) para o LASSO bayesiano (LB), Reproducing Kernel Hilbert Spaces (RKHS) e Redes Neurais Artificiais com Regularização Bayesiana (RNARB) para arquivos de teste contendo marcadores imputados e não-imputados (painel reduzido) contendo diferentes densidade de genótipos identificados.

Tabela 3. Correlações entre peso corporal predito e mensurado utilizando LASSO Bayesiano (LB), Reproducing Kernel Hilbert Spaces (RKHS) e Redes Neurais Artificiais com Regularização Bayesiana (RNARB) para todas as taxas de mascaramento e distribuição de famílias para dados de teste

90% dos genótipos mascarados						
Modelo	**Entre famílias**			**Dentro de famílias**		
	1809	1809i[a]	201	1809	1809i[a]	201
LB	0.347	0.259	0.169	0.500	0.330	0.407
RKHS	0.347	0.312	0.210	0.527	0.417	0.499
RNARB	0.330	0.217	0.144	0.490	0.274	0.392
75% dos genótipos mascarados						
Modelo	**Entre famílias**			**Dentro de famílias**		
	1809	1809i[b]	453	1809	1809i[b]	453
LB	0.343	0.291	0.262	0.499	0.447	0.440
RKHS	0.348	0.317	0.293	0.528	0.506	0.501
RNARB	0.320	0.241	0.255	0.492	0.414	0.428
50% dos genótipos mascarados						
Modelo	**Entre famílias**			**Dentro de famílias**		
	1809	1809i[c]	905	1809	1809i[c]	905
LB	0.342	0.324	0.271	0.499	0.496	0.477
RKHS	0.343	0.345	0.306	0.530	0.530	0.520
RNARB	0.320	0.281	0.252	0.492	0.478	0.461

[a] Imputados a partir de 201 SNPs;

[b] Imputados a partir de 453 SNPs;

[c] Imputados a partir de 905 SNPs.

Tabela 4. Correlações entre o índice de massa corporal predito e mensurado utilizando LASSO Bayesiano (LB), Reproducing Kernel Hilbert Spaces (RKHS) e Redes Neurais Artificiais com Regularização Bayesiana (RNARB) para todas as taxas de mascaramento e distribuição de famílias para dados de teste

90% dos genótipos mascarados						
Modelo	**Entre famílias**			**Dentro de famílias**		
	1809	1809i[a]	201	1809	1809i[a]	201
LB	0.2277	0.1935	0.1909	0.1989	0.1645	-0.047
RKHS	0.2380	0.1950	0.1990	0.2080	0.1320	-0.054
RNARB	0.1628	0.0923	0.1472	0.1630	0.0410	0.054
75% dos genótipos mascarados						
Modelo	**Entre famílias**			**Dentro de famílias**		
	1809	1809i[b]	453	1809	1809i[b]	453
LB	0.2289	0.2190	0.1996	0.2006	0.1958	0.1837
RKHS	0.2380	0.2260	0.2110	0.2080	0.2040	0.2000
RNARB	0.1180	0.1150	0.1450	0.1720	0.1540	0.1700
50% dos genótipos mascarados						
Modelo	**Entre famílias**			**Dentro de famílias**		
	1809	1809i[c]	905	1809	1809i[c]	905
LB	0.227	0.231	0.225	0.199	0.197	0.1895
RKHS	0.238	0.238	0.236	0.207	0.206	0.202
RNARB	0.118	0.131	0.149	0.172	0.168	0.149

[a] Imputados a partir de 201 SNPs;

[b] Imputados a partir de 453 SNPs;

[c] Imputados a partir de 905 SNPs.

Quadrado médio do erro de predição (QMEP)

Os resultados para QMEP estão descritos nas Tabelas 5 e 6. Para PC, os menores valores de QMEP foram para predições realizadas dentro de famílias com dados genômicos completos (1809 SNPs) o que corrobora resultados obtidos para habilidade de predição previamente descritos. Em geral, maiores taxas de mascaramento de genótipos resultaram em maior QMEP para PC, e dados contendo genótipos imputados forneceram melhor qualidade de ajuste comparado aos dados originais sem imputação. Enquanto que o QMEP observado para IMC não se modifica com as diferentes taxas de mascaramento ou imputação de genótipos tanto para o LB quando para o RKHS.

A RNARB obteve maiores valores para QMEP, o que já era esperado dados os resultados de correlação apresentados.

Tabela 5. Quadrado médio do erro de predição para análise do peso corporal utilizando o LASSO bayesiano (LB), Reproducing Kernel Hilbert Spaces (RKHS) e Redes Neurais Artificiais com Regularização Bayesiana (RNARB) de acordo com a distribuição de famílias entre dados de treinamento e dados de teste e taxa de mascaramento de genótipos

90% dos genótipos mascarados						
Modelo	**Entre famílias**			**Dentro de famílias**		
	1809	1809i[a]	201	1809	1809i[a]	201
LB	5.03	5.32	5.67	4.18	4.99	4.71
RKHS	4.92	5.2	5.36	4.15	4.75	4.66
RNARB	5.36	5.52	5.54	5.26	5.4	5.52
75% dos genótipos mascarados						
Modelo	**Entre famílias**			**Dentro de famílias**		
	1809	1809i[b]	453	1809	1809i[b]	453
LB	5.05	5.25	5.44	4.18	4.45	4.52
RKHS	4.92	5.04	5.11	4.13	4.23	4.21
RNARB	5.38	5.44	5.44	5.26	5.32	5.33
50% dos genótipos mascarados						
Modelo	**Entre famílias**			**Dentro de famílias**		
	1809	1809i[c]	905	1809	1809i[c]	905
LB	5.06	5.12	5.49	4.18	4.19	4.32
RKHS	4.94	4.94	5.01	4.06	4.08	4.12
RNARB	5.2	5.24	5.44	5.26	5.27	5.28

[a] Imputados a partir de 201 SNPs;

[b] Imputados a partir de 453 SNPs;

[c] Imputados a partir de 905 SNPs.

Tabela 6. Quadrado médio do erro de predição para análise do índice de massa corporal utilizando o LASSO bayesiano (LB), Reproducing Kernel Hilbert Spaces (RKHS) e Redes Neurais Artificiais com Regularização Bayesiana (RNARB) de acordo com a distribuição de famílias entre dados de treinamento e dados de teste e taxa de mascaramento de genótipos

90% dos genótipos mascarados						
Modelo	**Entre famílias**			**Dentro de famílias**		
	1809	1809i[a]	201	1809	1809i[a]	201
BL	0.002	0.002	0.002	0.002	0.002	0.002
RKHS	0.002	0.002	0.002	0.002	0.002	0.002
NN	0.013	0.014	0.010	0.042	0.036	0.024
75% dos genótipos mascarados						
Modelo	**Entre famílias**			**Dentro de famílias**		
	1809	1809i[b]	453	1809	1809i[b]	453
BL	0.002	0.002	0.002	0.002	0.002	0.002
RKHS	0.002	0.002	0.002	0.002	0.002	0.002
NN	0.021	0.023	0.015	0.044	0.045	0.041
50% dos genótipos mascarados						
Modelo	**Entre famílias**			**Dentro de famílias**		
	1809	1809i[c]	905	1809	1809i[c]	905
BL	0.002	0.002	0.002	0.002	0.002	0.002
RKHS	0.002	0.002	0.002	0.002	0.002	0.002
NN	0.021	0.023	0.016	0.040	0.040	0.047

[a] Imputados a partir de 201 SNPs;

[b] Imputados a partir de 453 SNPs;

[c] Imputados a partir de 905 SNPs.

DISCUSSÃO

Recentemente, alguns estudos têm investigado a habilidade de predição de modelos utilizando apenas uma parte dos SNPs com e sem imputação (Weigel et al., 2010, Mulder et al., 2011, Daetwyler et al., 2011, Berry e Kearney, 2011). Em geral, a habilidade de predição melhorou com a aplicação da imputação de genótipos, o que faz com que vários pesquisadores recomendem esta ferramenta a fim de reduzir custos em programas de seleção genômica. Entretanto, todos os estudos envolvendo imputação de genótipos para seleção genômica, consideraram apenas modelos lineares, como regressão de cumieira, LASSO bayesiano ou GBLUP (Weigel et al., 2010, Mulder et al., 2011, Dassonneville et al., 2011), que são

especificamento adaptados para captar sinais genéticos aditivos mas não aptos a capturar efeitos genéticos não-aditivos como dominância e epistasia.

A hipótese por trás deste trabalho é que todo o sinal genético e informação disponível no banco de dados está inteiramente contida nos genótipos observados e, portanto, nada é adicionado pela utilização da imputação de genótipos além da conveniência e facilidade de implementação de modelos preditivos utilizando os softwares disponíveis. Neste contexto, se métodos mais elaborados, capazes de modelar efeitos genéticos não-aditivos, são adotados o efeito da imputação pode ser desprezível. Com isso, o objetivo deste trabalho foi investigar o efeito da imputação de genótipos sobre a habilidade de predição de diferentes modelos (especificamente o LB, RKHS e RNARB) utilizando diferentes densidades de SNPs.

Pelos resultados obtidos verifica-se que nem sempre a imputação de genótipos desconhecidos é vantajosa para a predição de fenótipos. O ganho de se imputar genótipos dependerá da conexidade entre a população referência e candidatos à seleção, herdabilidade da característica, número de marcadores disponíveis no painel original, e o métodos utilizado para predição dos efeitos dos marcadores.Weigel et al. (2010) investigaram o efeito da imputação de uma chip de baixa densidade para o chip de 50k sobre a acurácia dos valores genômicos diretos em gado Jersey por meio do LB, e os autores encontraram que a imputação de genótipos melhorou a habilidade de predição do modelo em situações em que a acurácia de imputação era alta; caso contrário o painel reduzido contendo número original de SNPs deveria ser preferido. No mesmo contexto, Mulder et al. (2011) mostraram que em razão da magnitude de erros de imputação, o "ruído" adicionado pode ser maior que o benefício da utilização de SNPs para a predição de valores genômicos. Portanto, apenas SNPs com alta acurácia de imputação terão efeito positivo sobre a confiabilidade do valor genômico direto. Neste estudo, verifica-se que se a acurácia de imputação é baixa, o modelo contendo apenas os genótipos observados fornece melhor predição que o modelos com adição de genótipos imputados. A correlação entre o PC predito e mensurado dentro de familias utilizando o banco de dados completos com 1809 SNPs, dados completos com parte dos dados imputados e dados reduzidos (201 SNPs) foi 0,52, 0,42 e 0,50 utilizando o RKHS, indicando que a imputação não trouxe nenhuma informação adicional ao modelo. Para os outros cenários com diferentes taxas de mascaramento o arquivo de dados contendo genótipos imputados resultou, em média, em correlação 4% mais alta entre fenótipo predito e mensurado. Para o IMC, o arquivo de dados reduzido forneceu 89% da habilidade de predição do arquivo de dados completo contendo genótipos imputados e 78% da habilidade preditiva do banco de dados completo com genótipos observados, em média para todos os cenários testados. Portanto, em casos de características de herdabilidade mediana, a imputação de genótipos pode ser útil para a predição de fenótipos.

Comparando correlações para distribuição entre e dentro de famílias entre os arquivos de treinamento e teste, a imputação de genótipos pareceu ser mais efetiva na melhora da acurácia de predição quando a conexidade entre estes dois arquivos é fraca. Outros estudos que

exploraram o efeito da informação dentro de famílias e entre famílias (Legarra, 2008) também indicaram a necessidade de genotipar e fenotipar indivíduos de parentesco próximo a fim de melhorar a habilidade de predição.

Com respeito aos modelos, era esperado que modelos não-paramétricos fornecessem menor diferença entre resultados para dados completos imputados e painel reduzido, dado que estes modelos possuem a habilidade de capturar efeitos não-lineares entre preditores e respostas and aprender sobre formas funcionais complexas (Gianola et al., 2011). Entretanto, os resultados deste estudo indicam que o efeito da imputação de genótipos foi semelhante para qualidade de predição do LB e RKHS. Uma exceção foi o modelo de RNABR, que não foi capaz de lidar com erros de imputação e resultou em piores predições para o arquivo completo contendo genótipos imputados. Com isso, resultados indicaram que a qualidade dos dados imputados é fator fundamental a ser considerado quando se utiliza RNARB para predição de fenótipos. A imputação a partir de 905 marcadores para o painel completo (1809 marcadores) melhorou discretamente a predição pelo método RNARB, já que o taxa de erro de identificação de genótipos foi baixa para este arquivo.

Outra discussão, que vai além do escopo deste trabalho, é sobre o uso de chips contendo SNPs igualmente espaçados para seleção genômica. A principal vantagem destes chips é que eles dispensam a necessidade de chips de baixa densidade para cada característica especificamente, e conferem confiabilidade dos valores genômicos similar a dos chips contem SNPs selecionados por efeito (Moser et al., 2010). Comparando os resultados obtido com resultados apresentados na literatura sobre seleção genômica utilizando o mesmo arquivo de dados em ratos, pode ser visto que nenhuma mudança importante na habilidade de predição é observada quando se utiliza o painel completo de SNPs. Por exemplo, de los Campos et al. (2009) utilizou 10.946 SNPs no modelo LB e observaram correlação de 0,306 entre valores fenotípicos e predições para o IMC em ratos. Neste estudo, foi obtido quase 95% desta acurácia utilizando o mesmo método, porém apenas incluindo 1809 SNPs igualmente espaçados. Adicionalmente, Okut et al. (2011) reportaram correlação entre predições e observações no arquivo de treinamento de 0,18 para IMC utilizando RNARB e 798 SNPs pré-selecionados por efeito. Neste trabalho, foi obtido correlação de 0,15 para o mesmo modelos utilizando 905 marcadores igualmente espaçados, o que sugere que a RNARB funciona melhor utilizando marcadores selecionados com maior efeito.

Resultados similares foram observados em termos de QMEP. Maiores erros de imputação levaram a maiores valores de QMEP, fazendo com que resultados de modelos contendo o painel reduzido fossem melhores que de modelos inserindo marcadores imputados.

CONCLUSÕES

Resultados deste estudo mostram que a imputação nem sempre ajuda a melhorar a acurácia de predição de modelos paramétricos, semi e não-paramétricos. Para o peso corporal, característica de herdabilidade alta, a imputação de genótipos melhorou a habilidade de predição quando a conexidade entre os painéis de refeência e candidatos à seleção era baixa. Para índice de massa corporal, a aplicação da imputação de genótipos foi mais vantajosa especialmente quando o arquivo de genótipos era esparso (201 SNPs) utilizando BL e RKHS. Em outros cenários, a imputação apenas causou melhora discreta ou até mesmo piorou a habilidade de predição em situações em que se obteve baixa acurácia de identificação de genótipos. A Rede Neural Artificial com Regularização Bayesiana parece ser mais sensível a erros de imputação; portanto a utilização de genótipos imputados neste modelo deve ser avaliado com cautela.

REFERÊNCIAS BIBLIOGRÁFICAS

3. ARONSZAJN, N. Theory of reproducing kernels. *Trans Am Math Soc,*v.686, p.337-404, 1950.
4. BERRY, D.P., KEARNEY, J.F. Imputation of genotypes from low- to high-density genotyping platforms and implications for genomic selection. *Animal,* v.5(8), p.1162-1169, 2011.
5. BISHOP, CM: Pattern Recognition and Machine Learning. Singapore: Springer, 2006.
6. BROWNING, B.L., BROWNING, S.R. A unified approach to genotype imputation and haplotype phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet*, v.84, p.210-223, 2009.
7. BROWNING, S.R., BROWNING, B.L. Haplotype phasing: existing methods and new developments. *Nature Rev Genet*, v.12, p.703-714, 2011.
8. CALUS, M.P.L., VEERKAMP, R.F., MULDER, H.A. Imputation of missing single nucleotide polymorphism genotypes using multivariate mixed model framework. *J Anim Sci* , v.89, p.2042-2049, 2011.
9. CROSSA, J., DE LOS CAMPOS, G, PÉREZ, P., GIANOLA, D., BURGUEÑO, J., ARAUS, J.L., MAKUMBI, D., SINGH, R.P., DREISIGACKER, S., YAN, J., ARIEF, V., BANZINGER, M., BRAUN, H. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics*, v.186, p.713-724, 2010.
10. DAETWYLER, H.D., WIGGANS, G.R., HAYES, B.J., WOOLLIAMS, J.A., GODDARD, M.E. Imputation of missing genotypes from sparse to high density using long-range phasing. *Genetics*, v.189, p.317-327, 2011.

11. DASSONNEVILLE, R., BRØDUM, R.F., DRUET, T., FRITZ, S., GUILLAUME, F., GULDBRANDTSEN, B., LUND, M.S., DUCROCQ, V., SU, G. Effect of imputing markers from low-density chip on the reliability of genomic breeding values in Holstein populations. *J Dairy Sci*, v.94, p.3679-3686, 2011.
12. DE LOS CAMPOS, G., GIANOLA, D., ROSA, G.J.M., WEIGEL, K.A., CROSSA, J. Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. *Genet Res,* v.92, p.295-308, 2010.
13. DE LOS CAMPOS, G., GIANOLA, D., ROSA, G.J.M. Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. *J Anim Sci*, v.87, p.1883-1887, 2009.
14. DE LOS CAMPOS, G., NAYA, H., GIANOLA, D., CROSSA, J., LEGARRA, A., MANFREDI, E., WEIGEL, K.A., COTES, J.M. Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics*, v.182, p.375-385, 2009.
15. DEMUTH, H., BEALE, M., HAGAN, M. Neural Network Toolbox 6 User's Guide. The MathWorks, Inc 2009.
16. GIANOLA, D., FERNANDO, R.L., STELLA, A. Genomic assisted prediction of genetic value with semi-parametric procedures. *Genetics,* v.173, p.1761-1776, 2006.
17. GIANOLA, D., OKUT, H., WEIGEL, K.A., ROSA, G.J.M. Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. *BMC Genet*, v.12, n.87, 2011.
18. GIANOLA D., VAN KAAM, J.B.C.H.M. Reproducing Kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. *Genetics,* v.178, p.2289-2303, 2008.
19. GODDARD, M.E., HAYES, B.J. Genomic Selection. *J Anim Breed Genet,* v.124, p.323-330, 2007.
20. GONZÁLEZ-CAMACHO, J.M., DE LOS CAMPOS, G., PÉREZ, P., GIANOLA, D., CAIRNS, J.E., MAHUKU, G., BABU, R., CROSSA, J. Genome-enabled prediction of genetic values using radial basis function neural networks. *Theor Appl Genet*, doi 10.1007/s00122-012-1868-9. Accessed in 07/02/2012.
21. HABIER, D., FERNANDO, R.L., DEKKERS, C.M. Genomic selection using low-density marker panels. *Genetics*, v.182, p.343-353, 2009.
22. LEE, S.H., VAN DER WERF, J.H.J., HAYES, B.J., GODDARD, M.E., VISSCHER, P.M. Predicting unobserved phenotypes for complex traits from whole-genome SNP data. *PLoS Genet*, v.4(10), versão eletrônica:100023, 2008.
23. LEGARRA, A., ROBERT-GRANIÉ, C., MANFREDI, E., ELSEN, J. Performance of genomic selection in mice. *Genetics*, v.180, p.611-618, 2008.
24. LONG, N., GIANOLA, D., ROSA, G.J.M., WEIGEL, K.A., KRANIS, A., GONZÁLEZ-RECIO, O. Radial basis function regression methods for predicting quantitative traits using SNP markers. *Genet Res*, v.92(3), p.209-225, 2010.

25. MACKAY, D.J.C. Bayesian interpolation. *Neural Computation*, v.4, p.415-447, 1992.
26. MACKAY, D.J.C. Information Theory, Inference and Learning Algorithms. Cambridge: Cambridge University Press 2008.
27. MEUWISSEN, T.H., HAYES, B.J., GODDARD, M.E. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, v.157, p.1819-1829, 2001.
28. MOSER, G., KHATKAR, M.S., HAYES, B.J., RAADSMA, H.W. Accuracy of direct genomic values in Holstein bulss and cows using subsets of SNP markers. *Genet Sel Evol*, v.42, n.37, 2010.
29. MOTT, R. Finding the molecular basis of complex genetic variation in humans and mice. *Phil Trans R Soc B*, v.361, p.393-402, 2006.
30. MULDER, H.A., CALUS, M.P.L., DRUET, T., SCHROOTEN, C. Imputation of genotypes with low-density chips and its effect on reliability of direct genomic values in Dutch Holstein cattle. *J Dairy Sci,* v.95, p.876-889, 2011.
31. OKUT, H., GIANOLA, D., ROSA, G.J.M., WEIGEL, K.A. Prediction of body mass index in mice using dense molecular markers and a regularized neural network. *Genet Res Camb,* v.93, p.189-201, 2011.
32. PARK, T., CASELLA, G. The Bayesian LASSO. *J Am Stat Assoc*, v.103, p.681-686, 2008.
33. PÉREZ, P., DE LOS CAMPOS, G., CROSSA, J., GIANOLA, D. Genomic-enabled prediction based on molecular markers and pedigree using Bayesian Linear Regression package in R. *The Plant Genome*, v.3, p.106-116, 2010.
34. ROSA, G.J.M., PADOVANI, C.R., GIANOLA, D. Robust Linear Mixed Models with Normal/Independent Distributions and Bayesian MCMC Implementation. *Biometrical Journal*, v.5, p.573-590, 2003.
35. TIBSHIRANI, R. Regression shrinkage and selection via the LASSO. J R Stat Soc B, v.58, p.267-288, 1996.
36. VALDAR, W., SOLBERG, L.C., GAUGUIER, D., BURNETT, S., KLENERMAN, P. Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nature Genet*, v.38, p.879–887, 2006.
37. VALDAR, W., SOLBERG, L.C., GAUGUIER, D., COOKSON, W.O., RAWLINS, J.N.P. Genetic and environmental effects on complex traits in mice. *Genetics*, v.174, p.959–984, 2006.
38. WAHBA, G. Spline models for observational data. *Soc Ind Appl Math*, 1990.
39. WEIGEL, K.A., DE LOS CAMPOS, G., GONZÁLEZ-RECIO, O., NAYA, H., WU, X.L., LONG, N., ROSA, G.J.M., GIANOLA, D. Predictive ability of direct genomic values for lifetime net merit of Holstein sires using selected subsets of single nucleotide polymorphism markers. *J Dairy Sci*, v.92, p.5248-5257, 2009.
40. WEIGEL, K.A., DE LOS CAMPOS, G., VAZQUEZ, A.I., ROSA, G.J.M., GIANOLA, D., VAN TASSELL, C.P. Accuracy of direct genomic values derived from imputed

single nucleotide polymorphism genotypes in Jersey cattle. *J Dairy Sci*, v.93, p.5423-5435, 2010.