DESAFIOS BIOMÉTRICOS NO MELHORAMENTO GENÉTICO

1ª Edição

Gen Melhor

Grupo de Estudos em Genética e Melhoramento

40 ANOS
PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA E MELHORAMENTO
1976 2016
Universidade Federal de Viçosa

UFV 1926 · 2016

Willian Hytalo Ludke
Andréa Carla Bastos Andrade
Leonardo Volpato
Dênia Pires de Almeida
Isadora Cristina Martins de Oliveira
José Teodoro de Paiva
Michele Jorge da Silva
Murilo Viotto Del Conte
Thaís Cristina Silva
Vinícius Costa Almeida
Vitor Batista Pinto

Desafios Biométricos no Melhoramento Genético

1ª edição

Universidade Federal de Viçosa
Viçosa – MG
2017

Dados Internacionais de Catalogação na Publicação (CIP)

(eDOC BRASIL, Belo Horizonte/MG)

D441

Desafios biométricos no melhoramento genético / Willian Hytalo Ludke... [et al.]. – Viçosa (MG): GenMelhor, 2017.
166 p. : il. ; 2.807 kbytes

Formato: PDF
Requisitos do sistema: Adobe Acrobat Reader.
Modo de Acesso: World Wide Web
Inclui bibliografia.
ISBN 978-85-93566-00-4

1. Genética e melhoramento. I Andrade, Andréa Carla Bsatos. II.Volpato, Leonardo. III. Almeida, Dênia, Pires. IV. Oliveira, Isadora Cristina Martins de. V. Paiva, José Teodoro de. VI. Silva, Michele Jorge da. VII. Del Conte, Murilo Viotto. VIII. Silva, Thaís Cristina. IX. Almeida, Vinícius Costa. X. Pinto, Vitor Batista. XI. Título.

CDD-576

PREFÁCIO

A pesquisa científica tem como característica *a busca pelo conhecimento*, dispondo-se de diferentes ferramentas, que vão desde as criteriosas avaliações, até aos métodos de inferência, que permitem a conclusão de um pensamento / ideia sobre uma determinada área em estudo. Por sua vez, esse conhecimento tem a necessidade de constante propagação, sendo o *livro* uma forma completa de fazê-la com o intuito de formação, condução e conclusão de novas ideias.

Este livro reuniu diferentes informações científicas com importantes abordagens sobre os desafios biométricos na área de *Genética e Melhoramento Animal e Vegetal*. Aqui, o processo de atualização visa contribuir com os anseios das comunidades acadêmico-científica e profissional-científica.

Espera-se que o leitor aproveite ao máximo deste amplo leque de informações, que foi construído com a contribuição de *grandes cientistas* na área da biometria e que se torna uma *grande mostra de temas* que vem sendo abordados e discutidos. De certo, o conteúdo que o espera nas próximas páginas irá norteá-lo quanto às tendências da pesquisa para ciência.

Andréa Carla Bastos Andrade
Coordenadora Geral do GenMelhor

SUMÁRIO

Biometria aplicada ao melhoramento de espécies anuais e seus desafios

Ivan Schuster[1], *Freddy Mora*[2]

Introdução

Uma grande quantidade de métodos biométricos está disponível para o uso nos programas de melhoramento de plantas anuais. Alguns são bastante conhecidos, e sua descrição e aplicação é facilmente encontrada na literatura, na forma de livros didáticos. Outros métodos são mais desafiadores, mas seu uso tem se intensificado nos programas de melhoramento de plantas anuais, em virtude da necessidade de incremento dos ganhos genéticos obtidos atualmente. Neste trabalho foi priorizada a apresentação dos métodos biométricos menos convencionais, mas cujo uso começa a se intensificar nos programas de melhoramento de plantas anuais. Todos estes métodos, embora mais desafiadores do que os métodos comumente utilizados no período anterior à última década, têm contribuído de forma significativa para reduzir o tempo necessário para o desenvolvimento de novos produtos, e aumentar o ganho genético.

Ganho Genético

O objetivo de todo programa de melhoramento é o ganho genético, normalmente referido nos modelos biométricos como ganho por seleção. Para que possa haver ganhos genéticos, é necessário que haja variabilidade genética. Assim, pode-se gerar recombinantes, selecionar as melhores combinações genéticas, e consequentemente obter ganhos genéticos.

No entanto, o melhorista normalmente observa o fenótipo, e não o genótipo. O fenótipo observado resulta da expressão de genes do organismo (genótipo) com a influência de fatores ambientais (ambiente) e das interações entre eles. Neste sentido, a variação fenotípica pode ser representada por:

V (Fenótipo) = V (Genótipo + Ambiente)

V (Fenótipo) = V (Genótipo) + V (Ambiente) + 2 COV (Genótipo, Ambiente)

Ou, em termos dos parâmetros da população:

$$\sigma_F^2 = \sigma_G^2 + \sigma_A^2 + 2\sigma_{GA}$$

[1]Engenheiro Agrônomo, M.S., D.S. e Pesquisador da Dow Agrosciences. E-mail: ivanschuster.ivan@gmail.com.
[2]Engenheiro Florestal, M.S., D.S. e Professor da Universidade de Talca. E-mail: morapoblete@gmail.com.

Em que σ_F^2 é a variância fenotípica, ou a variância observada, σ_G^2 é a variância genotípica, ou a variância devida aos efeitos genéticos, σ_A^2 é a variância devida aos efeitos ambientais e σ_{GA}^2 é a variância devida à interação entre os efeitos genéticos e ambientais, sendo σ_{GA} a covariância entre os efeitos do genótipo e do ambiente.

A variância genotípica (σ_G^2) pode ser escrita em termos da variância devida aos efeitos aditivos (σ_a^2), variância devida aos efeitos de dominância (σ_d^2) e variância devida às interações epistáticas (σ_i^2), de maneira que:

$$\sigma_F^2 = \sigma_a^2 + \sigma_d^2 + \sigma_i^2 + \sigma_A^2 + 2\sigma_{GA}$$

Para estimar a proporção da variabilidade total que é devida às causas genéticas e ambientais, deve-se padronizar todos os parâmetros por σ_F^2:

$$\frac{\sigma_F^2}{\sigma_F^2} = \frac{\sigma_G^2}{\sigma_F^2} + \frac{\sigma_A^2 + 2\sigma_{GA}}{\sigma_F^2} = \frac{\sigma_a^2}{\sigma_F^2} + \frac{\sigma_d^2 + \sigma_i^2 + \sigma_A^2 + 2\sigma_{GA}}{\sigma_F^2}$$

Ou,

$$1 = H^2 + z = h^2 + z'$$

Em que: $H^2 = \sigma_G^2 / \sigma_F^2$ e $h^2 = \sigma_a^2 / \sigma_F^2$ correspondem às herdabilidades no sentido amplo e restrito, respectivamente. Herdabilidades são definidas em termos analíticos como a proporção da variação fenotípica que é atribuível à variação genética entre os indivíduos de uma população. z é a proporção da variação fenotípica atribuível aos efeitos ambientais, e z' a proporção da variação fenotípica atribuível aos efeitos genéticos não aditivos mais o efeito ambiental, considerando $\sigma_{GA} = 0$. Para as espécies em que o melhoramento precisa obter linhagens homozigotas, como nas espécies autógamas ou nas espécies alógamas em que são produzidas linhagens endogâmicas para a obtenção de híbridos, os efeitos genéticos não aditivos não são considerados, pois não estão presentes nas linhagens endogâmicas homozigotas. A herdabilidade no sentido restrito expressa a relação entre a variância genética útil ao melhoramento de linhagens endogâmicas, e a variância fenotípica, sendo:

$$h^2 = 1 - z'$$

O melhoramento genético consiste em selecionar, em uma população que possua variabilidade genética, os melhores indivíduos, que irão compor uma nova população, a população selecionada (ou população melhorada). A diferença ente as médias da população selecionada ($\overline{X_S}$) e a população original ($\overline{X_P}$) é o diferencial de seleção (*DS*)

(figura 1). Uma vez que a seleção é realizada a partir de avaliação fenotípica, o *DS* deve ser ponderado pela herdabilidade para representar o ganho genético.

$GS = DSh^2$

Em que *GS* representa tanto ganho por seleção como ganho genético.

O *DS* depende, por sua vez, da intensidade de seleção (*i*). Maior intensidade de seleção, resulta em menor número de indivíduos selecionados, maior média dos indivíduos selecionados, e maior *DS* (Figura 1). A intensidade de seleção pode ser expressa em unidades de desvio padrão.

$i = DS/\sigma$

e

$DS = i\sigma$

O *GS*, ou ganho genético, pode ser expresso então por:

$GS = \ ih^2\sigma_F$

$GS = ihh\sigma_F$

$GS = ih\frac{\sigma_a}{\sigma_F}\sigma_F$

$GS = ih\sigma_a$

Em que *h* é definido como acurácia (raiz quadrada da herdabilidade) e a variância genética está representada pela raiz quadrada da variância aditiva (σ_a), ou desvio genético aditivo. Esta é a equação central do melhoramento genético, e pode ser definida como a equação do melhorista (Lush 1937).

Acurácia, variância genética e intensidade de seleção definem o ganho genético por ciclo de seleção. No melhoramento genético, o objetivo é maximizar o ganho genético por unidade de tempo. Assim, ganho genético deve levar em consideração também o tempo necessário para a criação e desenvolvimento de novas linhagens, híbridos ou variedades, em anos (*n*).

$GS = ih\sigma_a/n$

A diminuição no tempo necessário para a criação e desenvolvimento de novos produtos (linhagens, híbridos ou variedades), resulta em aumento do ganho genético.

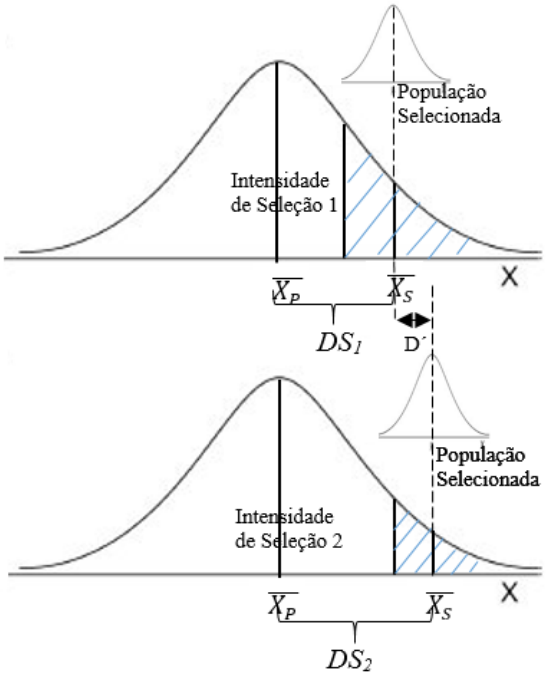


Figura 1. Considerando uma população que apresenta variabilidade para a característica de interesse, ao selecionar-se os melhores indivíduos, a média da população selecionada ($\overline{X_s}$) será maior do que a média da população original ($\overline{X_p}$). A diferença entre a média da população selecionada e da média da população original é o diferencial de seleção (*DS*). Ao aplicar-se uma intensidade de seleção maior (Intensidade de Seleção 2), o diferencial de seleção é maior do que quando se aplica uma intensidade de seleção menor (Intensidade de Seleção 1). D´ é a diferença no *DS* devida à diferença na intensidade de seleção.

Biometria no melhoramento de espécies anuais: Desafios e oportunidades

O aumento do ganho genético é obtido pelo aumento na intensidade de seleção, na herdabilidade (expressa pela acurácia – *h*), e na variância genética (expressa pelo desvio aditivo - σ_a), e pela diminuição no número de anos necessários para a obtenção dos novos produtos (*n*). De uma forma simplificada, os métodos biométricos utilizados no melhoramento objetivam otimizar cada um dos componentes da equação do ganho genético.

Intensidade de Seleção

A Intensidade de Seleção (*i*) é o *DS* padronizado. Considerando-se distribuição normal da característica sob seleção, a intensidade de seleção é o *DS* expresso em unidades de desvio padrão ($i=DS/\sigma$). Na figura 2, a área achureada corresponde a proporção (*P*) dos indivíduos selecionados, e a diferença entre a $\overline{X_S}$ e $\overline{X_P}$ é o *DS*. *T* é o valor do truncamento, ou seja, é o ponto de corte para a seleção de uma proporção *P* de indivíduos. Este valor não é conhecido, antes que a proporção *P* de indivíduos tenha sido selecionada.

Considerando distribuição normal, a média da proporção *P* de indivíduos selecionados é obtida integrando-se:

$$P = \int_T^{\infty} f(x)dx$$

$$\overline{X_S} = \frac{1}{p}\int_T^{\infty} x\, f(x)dx$$

Sendo que sob distribuição normal:

$$f(x) = \frac{1}{\sqrt{(2\pi)}\sigma} e^{\frac{-(X-\overline{X})^2}{2\sigma^2}}$$

Considerando:

$DS=\overline{X_S}$ - $\overline{X_P}$

Pode-se deduzir:

$$\therefore \quad DS = \frac{z}{p}\sigma$$

Em que z é a ordenada correspondente à altura da linha *T* quando esta corta a curva da distribuição normal padronizada (Figura 2).

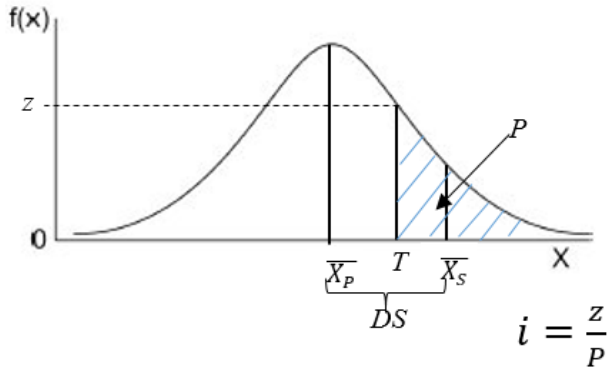
Considerando-se que ($i = DS/\sigma$):

$$i = \frac{\frac{z}{p}\sigma}{\sigma}$$

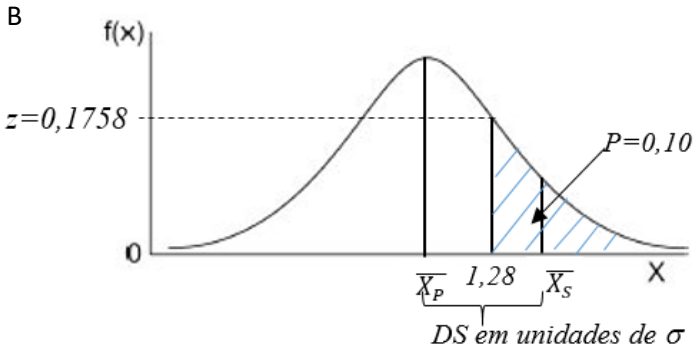
$$i = \frac{z}{p}$$

$$DS = i\sigma$$

A



$$i = \frac{z}{P}$$



$$i = \frac{z}{P} = \frac{0{,}1758}{0{,}10} = 1{,}758$$

Figura 2. Estimação do índice de seleção (*i*), considerando-se distribuição normal. A) Distribuição normal e seleção de uma proporção P de indivíduos na extremidade direita da distribuição. B) Distribuição normal padronizada, e estimação da intensidade de seleção (*i*) para a seleção de 10% dos indivíduos superiores da população.

Portanto, para obter o índice de seleção, é necessário obter o valor de *z* (ordenada da curva normal, no ponto onde *T* corta a linha da curva). Considerando a distribuição normal padrão (μ=0 e σ^2=1), pode-se utilizar a tabela da área sob a curva normal padronizada e a tabela que contém as ordenadas da curva normal padronizada (encontradas em livros de estatística) para obter o valor de *z*. O valor de *z* é a ordenada correspondente ao valor *T* (Figura 2).

Por exemplo, para selecionar de 10% dos indivíduos com maior (ou com menor) valor, o valor de *T* na distribuição normal padronizada é 1,28 (1,28σ maior do que a média da população). O valor da ordenada *z* correspondente a 1,28σ na distribuição normal padronizada é 0,1758. Dividindo-se o valor de *z* (0,1758) pela proporção *P* de indivíduos selecionados (0,1), tem-se a intensidade de seleção para a seleção de 10% dos indivíduos superiores (1,758) (Figura 2).

Valores de intensidade de seleção em função da proporção de indivíduos selecionados podem ser obtidas em tabelas (Tabela 1).

Uma vez que $GS = ih\sigma_a/n$, o aumento na intensidade de seleção resultará em aumento no ganho genético. Consideremos um programa de melhoramento que selecione, em média, 10% dos indivíduos superiores. A intensidade de seleção aplicada é de 1,76 (Tabela 1). Consideremos que este programa de melhoramento tenha como objetivo aumentar em 50% o ganho genético em relação aos ganhos genéticos obtidos atualmente. Isso significa que, se os ganhos genéticos atuais estão em 2% ao ano, o objetivo é obter ganhos genéticos de 3% ao ano. Se este objetivo tiver que ser atingido apenas aumentando

a intensidade de seleção, a intensidade de seleção deverá passar de 1,75 para 2,625 (1,75x1,5). Na tabela 1, observa-se que a intensidade de seleção 2,70 significa selecionar 1% dos indivíduos. Para atingir seu objetivo, este programa de melhoramento deverá selecionar pouco mais de 1% dos indivíduos, ao invés dos 10% selecionados atualmente. Para manter o mesmo número final de indivíduos selecionados, será necessário aumentar em quase 10 vezes o tamanho das populações de seleção.

Caso a população inicial não seja aumentada na proporção definida pelo incremento na intensidade de seleção, o número de indivíduos selecionados será menor. Com isso, terá todas as implicações relacionadas a populações pequenas, como pequeno tamanho efetivo, deriva genética e redução da variabilidade genética. A longo prazo, aumentar a intensidade de seleção sem aumentar adequadamente o tamanho das populações pode levar a diminuição dos ganhos genéticos. É fácil concluir, portanto, que o ganho genético não pode ser aumentado aumentando-se exclusivamente a intensidade de seleção.

Tabela 1. Valores de intensidade de seleção em função da proporção de indivíduos selecionados.

% de Indivíduos Selecionados (P)	Intensidade de seleção (i)
0,90	0,20
0,80	0,35
0,70	0,50
0,60	0,64
0,50	0,80
0,40	0,97
0,30	1,16
0,20	1,40
0,10	1,76
0,05	2,08
0,04	2,16
0,03	2,27
0,02	2,44
0,01	2,70

Acurácia

A acurácia é definida como a raiz quadrada da herdabilidade ($\sqrt{h^2} = h$). Assim, aumentar a acurácia significa aumentar a herdabilidade. Consideremos o objetivo de aumentar em 50% o ganho genético em relação aos ganhos genéticos atuais, como definido acima. Se este objetivo tiver que ser obtidos aumentando-se apenas a acurácia, será necessário aumentar a herdabilidade em 2,25 vezes ($1{,}5\sqrt{h^2} = \sqrt{2{,}25h^2}$). Aumento

de herdabilidade é obtido pela redução dos efeitos não genéticos (diminuição da variância ambiental). Para isso, melhores desenhos experimentais, escolha de áreas uniformidades, avaliações fenotípicas mais precisas, seleção com base em valores genéticos e não em valores fenotípicos, etc., devem ser utilizadas.

Também pode-se aumentar a herdabilidade, aumentando-se a variância genética, pelo aumento das recombinações. Para que a maior variabilidade genética possa ser explorada no melhoramento, é necessário aumentar o tamanho das populações de melhoramento, para permitir a ocorrência do maior número possível de recombinações. Mesmo assim, também não se pode esperar aumentar significativamente os ganhos genéticos aumentando-se somente a herdabilidade.

Variância Genética

O componente σ_a do *GS* é o desvio aditivo, ou desvio genético herdável. Considerando-se novamente o objetivo de aumento do ganho genético de 50% em relação ao ganho genético atualmente obtido, a variância genética herdável (σ_a^2) também deve ser aumentada em 2,25 vezes ($\sqrt{2{,}25\sigma_a^2} = 1{,}5\sigma_a$). Diferentemente dos parâmetros anteriores, o aumento da variância genética aditiva por si só não resulta necessariamente em aumento do ganho genético. Aumento de variabilidade genética pode ser conseguido com introdução de germoplasma externo. Dependendo da origem do germoplasma introduzido (germoplasma exótico, por exemplo), a maior parte da variabilidade introduzida pode não ser útil para o ganho genético. Variabilidade que que não resulte em ganhos genéticos, é variabilidade não útil.

A utilização de germoplasma não adaptado para gerar mais recombinações nas populações de melhoramento pode gerar populações com maior variância genética. Mas se esta variabilidade não gerar indivíduos no extremo direito da curva de desempenho, este aumento na variância genética não se traduz em aumento do ganho genético (Figura 3).

Por este motivo, aumento de variabilidade genética deve ser entendido como aumento da variabilidade genética útil, aquela que contribui para aumentar a frequência de indivíduos na extremidade direita da distribuição normal, movendo esta extremidade mais à direita. Métodos biométricos, associados com métodos moleculares, podem identificar apenas a porção útil da variabilidade obtida a partir de germoplasma externo (exótico ou não), e métodos de melhoramento específicos (ou de pré-melhoramento)

podem ser utilizados para introduzir apenas a porção útil da variabilidade no germoplasma elite.

Para que a maior variabilidade genética possa ser explorada adequadamente, é necessário aumento significativo no tamanho das populações, o que pode inviabilizar o processo de seleção. Aumentar somente a variabilidade genética isoladamente também não permite ganhos genéticos significativos.

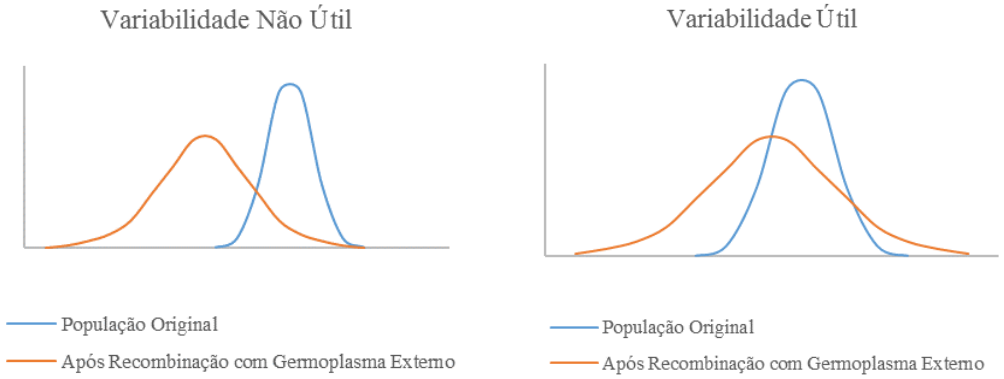


Figura 3. A) Variabilidade não útil: Após a introdução de germoplasma externo e recombinação com o germoplasma original, houve aumento de variabilidade, porém isso não resultou em aumento do desempenho, pois o limite +direito da curva continua o mesmo de antes da introdução da variabilidade. B) Variabilidade útil: Após a introdução de germoplasma externo e recombinação com o germoplasma original, houve aumento de variabilidade, redução da média da população, porém apareceram indivíduos com melhor desempenho do que os melhores indivíduos da população original (mais à direita da curva).

Tempo necessário para o desenvolvimento de novos produtos

O ganho genético pode ser aumentado também pela reciclagem mais rápida da "nova genética". Os novos produtos desenvolvidos pelos programas de melhoramento, com maior valor genético, são a principal fonte de germoplasma para ser utilizada na geração das novas populações de melhoramento (respeitando-se a manutenção da variabilidade genética útil).

Quanto mais cedo a "nova genética" for utilizada para novas recombinações, maior será o ganho genético. Considerando-se o objetivo de aumentar em 50% os ganhos genéticos atuais em um programa de melhoramento, ajustando-se apenas o tempo de lançamento de reciclagem dos novos produtos, o tempo atual precisa ser reduzido para 2/3, ou seja, se este programa recicla sua "nova genética" em seis anos, deverá fazê-lo em 4 anos.

$$1{,}5GS = \frac{1{,}5(ih\sigma_a)}{n} = \frac{ih\sigma_a}{\frac{n}{1{,}5}} = \frac{ih\sigma_a}{0{,}67n}$$

Uma vez que reduzir em 1/3 o tempo de obtenção ou reciclagem dos novos produtos nem sempre é possível, o aumento do ganho genético também não pode ser baseado apenas em diminuição neste tempo.

Melhorar apenas um componente da equação do melhorista, isoladamente, resulta em ganhos genéticos menores. Ganhos genéticos significativos podem ser obtidos otimizando-se cada um dos componentes da equação do melhorista: a) aumentando-se a intensidade de seleção, b) aumentando-se a acurácia da estimação (ou predição) do valor genético, tanto a partir de dados fenotípicos quanto moleculares, c) explorando-se melhor a variabilidade genética útil de cada espécie, e d) reduzindo-se o tempo entre gerações de melhoramento e até mesmo reduzindo-se o número de gerações de seleção, para reduzir o ciclo de reciclagem das linhagens, tanto autógamas ou alógamas.

Biometria no melhoramento de espécies anuais: Aplicações

Dependendo do modo de reprodução da espécie, tanto os métodos utilizados no melhoramento genético quanto o produto obtido pode ser diferente (variedade ou híbrido). Melhoramento de variedades autógamas, melhoramento de linhagens alógamas e melhoramento de híbridos são realizados utilizando-se métodos específicos para cada um destes. No entanto, independentemente dos métodos utilizados, os ganhos genéticos respondem a mesma equação:

$GS = ih\sigma_a/n$

Nas figuras 4 e 5 estão representadas, de maneira generalizada, as etapas do melhoramento genético de plantas alógamas (Figura 4) e autógamas (Figura 5). Ao lado de cada etapa dos programas de melhoramento estão identificadas as possibilidades de análises biométricas que podem ser utilizadas para aumentar o ganho genético.

Biometria na Caracterização do Germoplasma Base

A primeira oportunidade de uso da biometria no melhoramento de plantas está na caracterização da variabilidade genética existente no germoplasma base. Com algumas pequenas variações, a caracterização do germoplasma base pode utilizar os mesmos métodos tanto para plantas alógamas como para plantas autógamas.

O conhecimento da variabilidade genética do germoplasma base é importante para quantificar a variância genética disponível para recombinação e para classificar o germoplasma disponível. No entanto, variabilidade genética que reduza a média das populações não é variabilidade genética útil (Figura 3). Para quantificar a variância

genética útil é necessário conhecer, além da variabilidade genética, o valor genético dos indivíduos que compõe o germoplasma base.

A introdução de variabilidade genética a partir de germoplasma exótico, ou não adaptado, é desejável, desde que possam trazer valor ao germoplasma adaptado. No entanto, a utilização direta deste germoplasma não adaptado pode reintroduzir alelos indesejáveis que já haviam sido eliminados pelo melhoramento. Por esse motivo, o germoplasma não adaptado deve passar por alguns ciclos de melhoramento paralelos antes de ser utilizado no melhoramento de linhagens (tanto alógamas quanto autógamas).

Uma vez que o germoplasma base elite será utilizado para gerar recombinantes, é importante conhecer o valor de cada indivíduo como genitor, ou seja, a sua capacidade de produzir progênies superiores. Por isso, estimativas de capacidade geral de combinação (CGC) também são realizadas no germoplasma base.

Em plantas alógamas é importante agrupar o germoplasma base em grupos heteróticos (GH), para que as recombinações com o objetivo de desenvolvimento de novas linhagens possam ser realizadas preferencialmente dentro do mesmo grupo heterótico, e assim manter a identidade dos grupos heteróticos e possibilitar maior heterose nos cruzamentos de linhagens de diferentes grupos heteróticos, no momento da criação de híbridos.

Embora em plantas autógamas o germoplasma não precise ser classificado em grupos heteróticos, é importante utilizar algum tipo de classificação, para que possam ser geradas populações que, além de elevada média, também apresentam elevada variância genética.

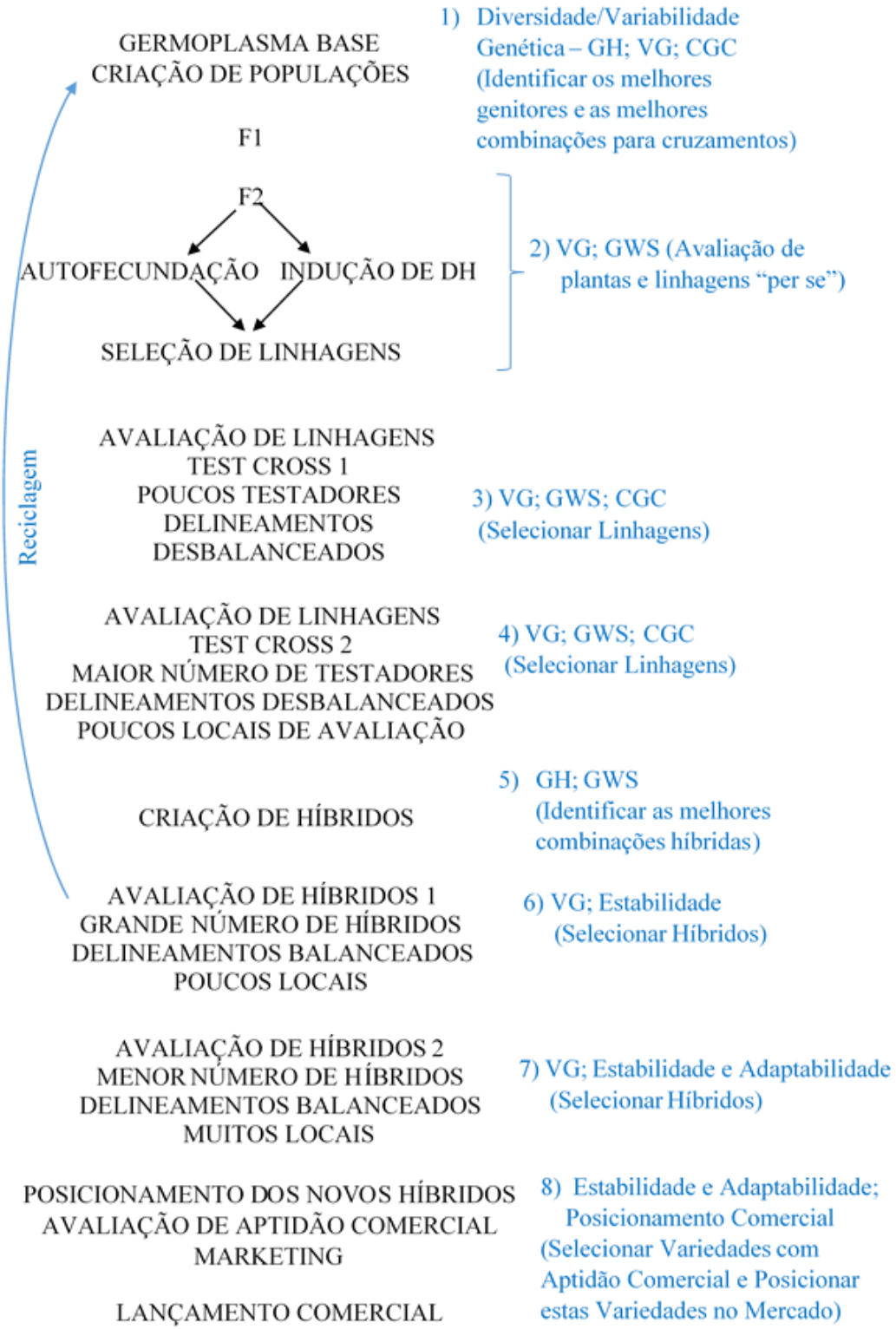


Figura 4. Oportunidades da aplicação dos métodos biometricos no melhoramento de espécies alógamas, com o objetivo de criação de híbridos. As etapas do melhoramento estão apresentadas à esquerda, e as análises biométricas aplicáveis, à direita. Entre parêntesis está descrito o objetivo das análises biométricas em cada fase do programa de melhoramento. GH, Grupo Heterótico; VG, Valor Genético; CGC, Capacidade Geral de Combinação; GWS, *Genomewide Selection* (Seleção Genômica Ampla).

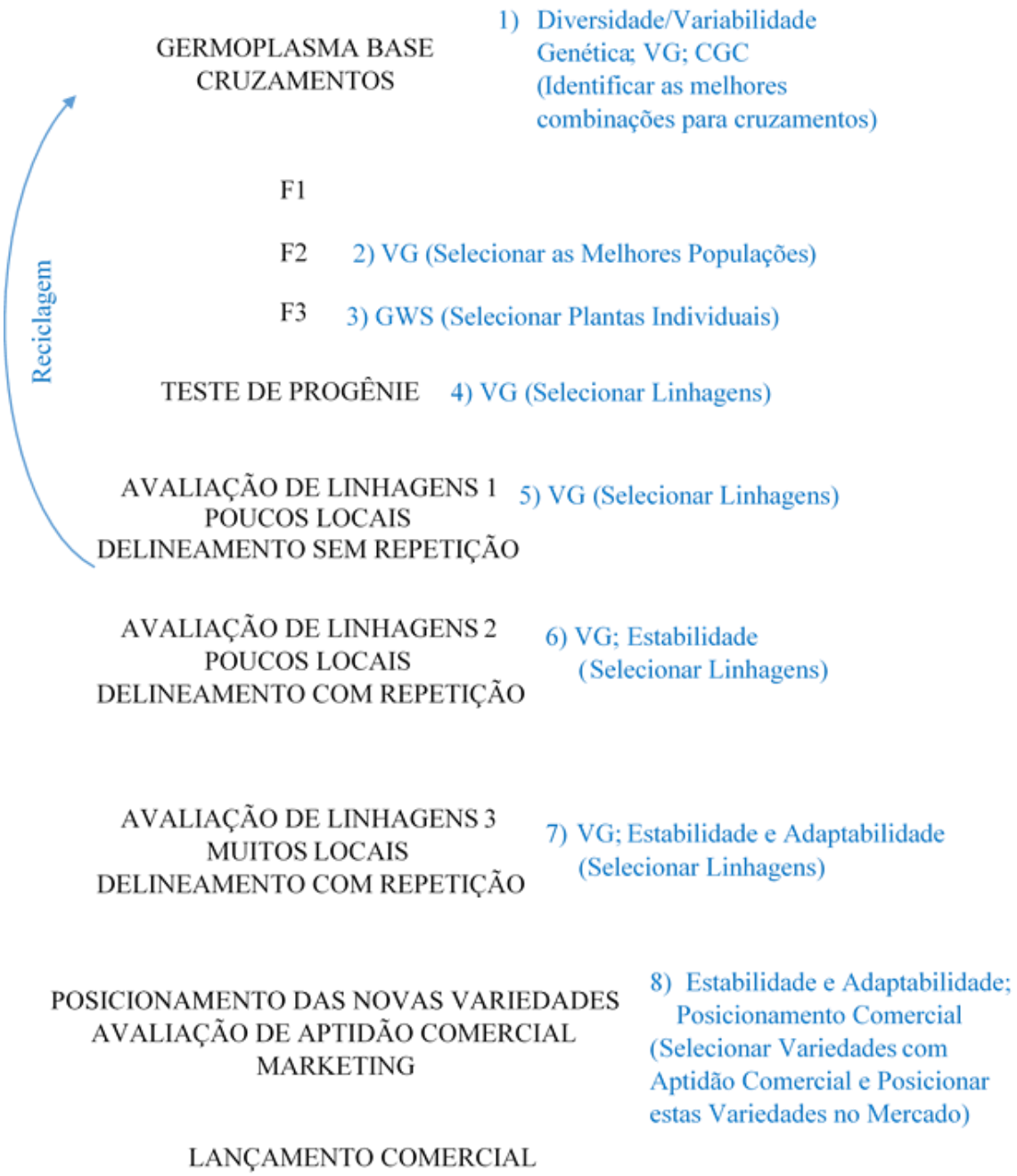


Figura 5. Oportunidades da aplicação de biometria no melhoramento de espécies autógamas com o objetivo de desenvolvimento de variedades. As etapas do melhoramento estão apresentadas à esquerda, e as análises biométricas aplicáveis, à direita. Entre parêntesis está descrito o objetivo de cada fase do programa de melhoramento. VG, Valor Genético; CGC, Capacidade Geral de Combinação; GWS, *Genomewide Selection* (Seleção Genômica Ampla).

Predição de Cruzamentos para Populações de Melhoramento

Utilizando-se o valor genético e o genótipo dos parentais para um grande número de marcadores moleculares, pode-se predizer quais cruzamentos podem ser realizados para aumentar a probabilidade de gerar linhagens superiores. Nesta fase, pode-se utilizar modelos que consigam predizer a média e variância das populações derivadas de qualquer

possível cruzamento. Isso possibilita ao melhorista ampliar muito sua base de variabilidade, criando um grande número de cruzamentos, testando os cruzamentos "in silico" inicialmente, e então realizar biologicamente somente os cruzamentos com maior probabilidade de gerar as melhores populações. Além disso, pode-se definir o tamanho adequado de cada população F2, em função da predição da variância esperada em cada população.

O desenvolvimento de modelos de seleção genômica ampla (*Genomewide Selection* – GWS), tem por objetivo predizer o valor genético dos indivíduos, com base em genotipagem de milhares de SNPs. Modelos de predição são construídos em populações de estudo, utilizando-se dados de genótipo de marcadores e dados fenotípicos.

Para a construção dos modelos de predição de híbridos ou variedades, não é necessário conhecer quais regiões genômicas estão envolvidas na determinação do valor genético predito dos indivíduos. No entanto, o desafio da predição das melhores combinações genotípicas está em identificar as regiões genômicas que contribuam diferentemente com o aumento do valor genético, e a importância de cada uma destas regiões. Essa informação poderá ser utilizada para identificar quais linhagens deverão ser cruzadas preferencialmente, para gerar populações com a maior probabilidade de conter os recombinantes contendo o maior número possível de regiões associadas ao aumento do valor genético dos indivíduos. Poderá ser feita predição inclusive de recombinações múltiplas, que permitam reunir e um indivíduo, regiões genômicas presentes em diversas linhagens. Além disso, no caso da necessidade de recombinação genômica (crossing over) entre regiões que contribuam de forma positiva ligadas a regiões que contribuam de forma negativa para o valor genético, pode-se estimar o tamanho da população necessária para obtenção dos recombinantes mais raros.

O custo de genotipagem tem barateado de tal modo que a genotipagem por sequenciamento (*Genotyping by Sequencing*- GBS) é atualmente uma das plataformas mais econômicas de genotipagem. Ou seja, ao invés de utilizar marcadores moleculares específicos, em plataformas de genotipagem de SNPs, pode-se obter a sequência genômica completa dos indivíduos, e não apenas os genótipos nas regiões cobertas pelos marcadores contidos em um kit de SNPs.

A identificação do valor genético de cada região genômica pode possibilitar o melhoramento dirigido para um genótipo específico, o "genótipo ideal". Análogo ao melhoramento por ideótipo, poderá ser possível realizar o melhoramento por genótipo. Ou, seja, pode-se "desenhar" o genótipo ideal, e programar ciclos de recombinação que

combinem regiões favoráveis atualmente presentes em um grande número de linhagens (adaptadas e não adaptadas). Pode-se recombinar as linhagens que contenham as regiões desejadas, duas a duas, selecionar os recombinantes desejados e recombinar novamente com outras linhagens que possuam outras regiões de interesse, até a obtenção do "genótipo ideal", análogo a montagem de um quebra-cabeças.

Obviamente o genótipo ideal será diferente para cada área de adaptação, e também estará em constante evolução, à medida que novas regiões genômicas que contribuam positivamente para o valor genético sejam identificadas, novas fontes de germoplasma sejam utilizadas, e epistasias do tipo aditivo-aditivo sejam identificadas a partir dos novos recombinantes. O desenvolvimento de tais modelos, e a capacidade de processamento desta enorme quantidade de dados ainda são um desafio para a aplicação da biometria com este objetivo.

Seleção Precoce de Populações Segregantes

A seleção precoce de populações F2, para eliminação das populações com menor potencial de gerar descendência superior, pode aumentar a eficiência, e consequentemente a acurácia nas etapas seguintes dos programas de melhoramento, por permitir focar os esforços em populações mais promissoras. As sementes das populações F2 podem ser divididas em sub-amostras, que constituem as repetições para as análises de rendimento na geração F2, e modelos de predição do valor genético podem ser utilizados para predizer o valor de cada população. Populações de maior valor podem ser prtiorizadas para as etapas seguintes do melhoramento, eliminando-se as demais. Com isso, um maior número de populações pode ser criado, explorando de forma mais ampla a variabilidade genética disponível, no germoplasma base, e otimizando os recursos dos programas de melhoramento nas gerações seguintes do melhoramento, pela eliminação precoce das populações com menor potencial de gerar linhagens superiores. RML BLUP ou inferência Bayesiana podem ser utilizados para predição do valor genético de cada população, utilizando-se como matriz de parentesco, a genealogia, quando disponível, ou dados de similaridade genética obtidos por marcadores moleculares.

Seleção de Plantas Individuais

Tanto no melhoramento de plantas alógamas quanto no melhoramento de plantas autógamas, a seleção de plantas individuais e autofecundação (natural no caso de plantas autógamas) é o início da etapa de desenvolvimento de linhagens endogâmicas. Em milho,

plantas individuais também podem ser selecionadas para indução de haploidia, ao invés de autofecundação. Modelos preditivos do valor genético de cada planta individual, com base na genotipagem de alta densidade destas plantas, podem ser usados nesta fase, para selecionar as plantas com base no genótipo (GWS). Modelos preditivos podem ser usados também para selecionar plantas duplo-haplóides na primeira geração após a duplicação cromossômica (H_0), ou até mesmo nas plântulas haplóides antes do procedimento de duplicação cromossômica, dependendo da estratégia do programa de melhoramento, e do custo/benefício.

A seleção visual de plantas individuais apresenta herdabilidade muito baixa, e este é o momento em que o uso de seleção genômica ampla nos programas de melhoramento pode gerar os maiores benefícios. A seleção genômica baseada em um modelo preditivo de alta eficácia pode identificar mais eficientemente as plantas com maior potencial de gerar as melhores linhagens endogâmicas nesta fase de seleção de plantas individuais, até mesmo sem a necessidade de avaliação fenotípica. Por este motivo, esta geração pode ser conduzida na contra estação de cultivo da espécie, possibilitando mais de uma geração de seleção por ano, reduzindo-se o intervalo de tempo até a reciclagem das novas linhagens, e consequentemente permitindo que seja obtido maior ganho genético.

Seleção de Linhagens e de Híbridos

Em plantas autógamas, as etapas seguintes do programa de melhoramento visam a seleção das melhores linhagens, até o avanço comercial destas linhagens como novas cultivares. A seleção de linhagens, ao longo dos anos de avaliação, pode ser realizada com base em médias fenotípicas, mas a probabilidade de identificar as melhores linhagens aumenta se a seleção for realizada com base no seu valor genotípico, que pode ser predito utilizando-se modelos lineares mistos, inferência Bayesiana, entre outros. A medida que o número de locais e anos de avaliação aumenta, estimativas de adaptabilidade e de estabilidade também serão obtidas, para identificar futuras variedades que possuam adaptação ampla, e também identificar as regiões ideias para cultivo das novas variedades com adaptação específica.

Em plantas alógamas, além do desenvolvimento das linhagens endogâmicas, é necessário ainda identificar as melhores combinações híbridas possíveis com estas linhagens selecionadas. Nesta fase, a estimação da CGC das linhagens é realizada por meio de cruzamentos teste (*test cross*). Novamente, a estimação da CGC com base em dados de valor genotípico, ao invés de valor fenotípico, aumenta a chance de identificar

corretamente o valor de CGC das linhagens. Além disso, os resultados obtidos nos cruzaentos teste utilizando-se testadores de cada grupo heterótico são utilizados para classificação das novas linhagens nos grupos heteróticos.

As linhagens identificadas com elevado valor “per se” (valor genotípico) e alta CGC são então utilizadas para a confecção dos novos híbridos para avaliação de desempenho, respeitando-se o grupo heterótico das linhagens.

A classificação das novas linhagens nos grupos heteróticos pode ser feita também através de análise de similaridade genética, utilizando-se genotipagem de alta densidade. E os novos híbridos a serem testados podem também ser preditos através de modelos de seleção genômica ampla. Desta forma, pode-se utilizar um grande número de linhagens, e produzir, “in silico”, literalmente milhões de híbridos, se for necessário. Esta grande quantidade de híbridos pode então ser avaliada também em “experimentos virtuais”, onde serão identificados os híbridos que possuam maior probabilidade de obter desempenho superior. Apenas estes híbridos de melhor desempenho nos experimentos virtuais precisam ser produzidos biologicamente e avaliados em avaliação de desempenho a campo.

Como exemplo, consideremos um programa de melhoramento que tenha capacidade de avaliar mil híbridos. O melhorista pode criar o número de híbridos que quiser (2mil, 5mil, etc.) e avaliar estes híbridos em experimentos virtuais, e selecionar, com base na simulação “in silico” os mil híbridos com maior potencial de desempenho superior. A atividade de avaliação de híbridos a campo será a mesma, com o mesmo tamanho. Mas o melhorista terá partido de uma base muito maior. Na primeira geração de avaliação de híbridos (geração virtual), um número muito maior de híbridos foi avaliado, de forma que os híbridos em teste de campo possuem probabilidade maior de apresentarem melhor desempenho, comparando-se a criação de apenas mil híbridos e avaliação direta a campo.

A seleção dos híbridos, ao longo dos anos de avaliação a campo também pode ser realizada com base em dados fenotípicos ou genotípicos, com preferência para a análise de dados genotípicos, ou valor genético. Assim como na avaliação de linhagens autógamas, a medida que o número de locais e anos de avaliação dos híbridos aumenta, estimativas de adaptabilidade e de estabilidade também serão obtidas. Assim, é possível identificar híbridos de adaptação ampla, e também identificar as regiões ideias para cultivo dos híbridos com adaptação específica.

Métodos Biométricos Utilizados no Melhoramento de Plantas Anuais

Nas sessões seguintes serão apresentados alguns exemplos de métodos biométricos que são, ou podem ser, utilizados no melhoramento de plantas anuais, para a estimação da variabilidade genética do germoplasma, variância genética das características de interesse, herdabilidade e predição do valor genético. Estes métodos, tem por objetivo otimizar os componentes da equação do ganho de seleção: aumentar a intensidade de seleção, a acurácia (ou herdabilidade) e o desvio genético aditivo (ou variância genética aditiva), e reduzir o tempo para a reciclagem das novas linhagens. A utilização destes métodos na rotina do melhoramento ainda é um desafio para muitos programas, mas vencer este desafio e utilizar médos de predição do valor genético, e associar análise de dados fenotípicos e moleculares é essencial para poder obter incrementso no ganho genético.

Análise de diversidade genética e classificação do germoplasma

A variabilidade genética é essencial no germoplasma base, para que possam ser obtidas as recombinações desejadas nas populações de melhoramento. Para conhecer a variabilidade disponível, é necessário quantificar a variabilidade genética disponível, e também classificar o germoplasma em função desta variabilidade, através de análises de agrupamento.

A análise da variabilidade genética pode ser realizada com dados morfológicos, quantitativos ou multicategóricos, ou com dados moleculares, tanto de proteína como de DNA. A análise de diversidade genética a nível de DNA avalia a variabilidade genética a nível de genoma no germoplasma de interesse, e não é influenciada pela expressão diferencial das características em função do ambiente em que a coleção de germoplasma é avaliada. Além disso, a medida que se incluem novas amostras no germoplasma, apenas estas novas amostras precisam ser genotipadas com mesmo conjunto de marcadores moleculares anteriormente utilizado para classificar o germoplasma existente. Os dados de marcadores moleculares do germoplasma introduzido são então incluídos no banco de dados moleculares do germoplasma completo, e o germoplasma introduzido pode facilmente ser classificado em relação ao germoplasma já existente. Ao musar características morfológicas quantitativas para classificar o germoplasma, todo o germoplasma deve ser avaliado novamente, nas mesmas condições ambientais, para poder comparar o germoplasma introduzido com o germoplasma já existente.

Análise de diversidade genética, ou de similaridade genética, normalmente utilizam o conceito de identidade por estado (IBS – *Identity by State*). Ou seja, dois indivíduos que possuem o mesmo alelo em determinado loco são considerados idênticos neste loco, independentemente se esta identidade foi obtida por descendência ou por eventos independentes de mutações. A expressão básica da similaridade genética (*S*) é:

$$S = IBS = \frac{Ac}{At}$$

Em que *Ac* é o número de alelos comuns, e *At* o número total de alelos avaliados. A dissimilaridade, por sua vez, é expressa por:

$$D = 1 - IBS$$

Similaridade ou dissimilaridade genética obtidos por IBS variam de zero a 1. Em conjuntos de populações em que há variabilidade dentro das populações é necessário incluir a estimativa de endogamia na estimação de similaridade genética. No entanto, no melhoramento de espécies que objetivam a obtenção de linhagens homozigotas, assume-se que a endogamia das linhagens e variedades que compõe o germoplasma base tenha endogamia igual a um, e a estimação simples da IBS com base na proporção de alelos comuns em relação ao número total de alelos é adequada.

Os dados de similaridade genética são utilizados então para o agrupamento dos indivíduos, ou classificação do germoplasma. Alguns métodos comumente utilizados para classificação do germoplasma são método de otimização, como o método de Tocher, métodos hierárquicos, que são expressos na forma de dendrogramas, e análise de componentes principais, que são expressos em gráficos de dispersão com duas ou três dimensões. Além disso, classificação com base em modelos também pode ser utilizada.

O método otimização de Tocher é um método de agrupamento, cuja pressuposição básica é que as distâncias dentro dos grupos sejam menores do que as distâncias entre grupos. Inicialmente define-se o valor máximo de distância que será permitida dentro dos grupos, ou seja, o limite de distância para permitir a entrada de um novo indivíduo no grupo. Esta distância limite para permitir a entrada de um novo indivíduo no grupo, normalmente é definida como a máxima distância entre as mínimas distâncias. Lista-se os valores das menores distâncias para todos os indivíduos, e o maior valor entre estas menores distâncias será o ponto de corte para a entrada de cada novo indivíduo no grupo. Considerando a existência de um grupo contendo os indivíduos *i* e *j*, ao ingressar o indivíduo *k* neste grupo, a distância dentro do grupo será:

$$d(i,j)k = [d(i,k) + d(j,k)]/n$$

Em que *d(i,k)* e *d(j,k)* são as distâncias entre os pares de indivíduos *i* e *k*, *j* e *k*, e *n* é o número de distâncias que estão sendo somadas no numerador. Analogamente:

$d(i,j,k)l = [d(i,l) + d(j,l) + d(k,l)]/n$

Quando a distância dentro do grupo ultrapassa o limite (ponto de corte) para a entrada no grupo, o grupo é finalizado apenas com os indivíduos que mantenham a distância dentro do grupo menor do que o limite, e os indivíduos restantes são submetidos a um novo ciclo de avaliações de grupos, até que todos os indivíduos que possam pertencer a algum grupo tenham sido agrupados.

Na análise de agrupamento por método hierárquico, como vizinho mais próximo, vizinho mais distante, UPGMA, Ward, entre outros, o objetivo é a obtenção de uma árvore e suas ramificações (dendrograma), sem necessariamente estar preocupado com o número de grupos a ser formado, embora os grupos possam ser identificados pelas ramificações do dendrograma.

O método de agrupamento hierárquico mais utilizado é o UPGMA (*Unwheighted Pair Group Mean Average* ou agrupamento pelas médias não ponderadas das distâncias). Considerando a distância entre os itens *i* e *j*, sendo $i \neq j$, o agrupamento pelo método UPGMA utiliza o seguinte algoritmo:

a. Primeiro grupo: $d(i,j) = mín\ d(i,j)$,

Ou seja, o par de indivíduos que tiver a menor distância inicia a formação do dendrograma, e a distância que ligará *i* e *j* no dendrograma será a distância genética entre estes dois indivíduos. Uma nova matriz de distâncias é construída então, eliminando-se os indivíduos *i* e *j*, e incluindo-se o grupo *(i,j)*, com as distâncias calculadas em relação a cada indivíduo restante.

b. Distância de um indivíduo *k* ao grupo *(i,j)*:

$d(i,j)k = 1/2\ [d(i,k) + d(j,k)]$

Sendo $k \neq i \neq j$. *k* será o indivíduo que apresentar menor valor para esta distância. *k* será ligado ao grupo *i* e *j*, no dendrograma, a uma distância projetada no eixo das ordenadas igual a distância calculada acima. O algoritmo segue, incluindo novos indivíduos ao grupo previamente formado, sendo que os indivíduos com as menores distâncias serão escolhidos para entrada, a cada ciclo. O denominador para a estimação das distâncias entre o grupo já formado, e o novo entrante é sempre igual ao número de termos sendo somados dentro do parêntesis. A cada etapa, a matriz de distâncias é recalculada, retirando-se da matriz os indivíduos já incluidos no agrupamento, e incluindo os grupos já formados, com as respectivas distâncias em relação aos indivíduos restantes.

c. $d(i,j,k)l = 1/3\ [d(i,l) + d(j,l) + d(j,l)]$

$i \neq j \neq k \neq l$.

O algoritmo segue, até que todos os indivíduos tenham sido inseridos no dendrograma. As distâncias entre dois grupos (*i*,*j*) e (*k*,*l*) são estimadas por:

$d(i,j)(k,l) = 1/4\ [d(i,k) + d(j,k) + d(i,l) + d(j,l)]$

Sendo $k \neq i \neq j \neq l$.

A relação existente entre as linhagens e variedades que compõe o germoplasma base é expressa na foram de dendrograma (Figura 6). Embora o objetivo da construção de dendrogramas seja avaliar a relação existente entre os indivíduos, pode-se utilizar a figura para definir a formação de grupos de indivíduos. Os dendrogramas podem ser obtidos tanto a partir de dados de similaridade quanto de dissimilaridade genética. O entanto, dendrogramas obtidos por dissimilaridade genética são mais simples de interpretar, uma vez que linhas mais longas significam maior distância genética (ou dissimilaridade), e as linhas mais curtas menor distância.

A definição dos grupos pode ser obtida de forma visual, realizando-se cortes nas regiões do dendrograma que apresentem grandes distâncias, ou seja, onde houver uma mudança abrupta na ramificação. Métodos estatísticos também podem ser usados para a definição dos grupos. Khattree e Naik (2000) apresentaram três métodos para definir a formação dos grupos. O Desvio Padrão Médio (DP_w) é a raiz quadrada da variância dentro do grupo. Considerando que o novo grupo formado tenha as variâncias S_{w1}^2 , S_{w2}^2 , ..., S_{wv}^2 para *v* variáveis, temos:

$$DP_w \sqrt{\frac{1}{v}\sum_{j=1}^{v} S_{x1}^2}$$

Quanto menor o valor de DP_w mais similares os indivíduos dentro do grupo. Outro critério para a definição de grupos é o coeficiente de determinação (R^2) para o grupo formado:

$$R_w^2 = 1 - \frac{SQD_w}{SQT}$$

Em que *SQT* é a soma de quadrados total corrigida de todos os indivíduos, somada sobre todas as variáveis, e SQD_w é a soma de quadrados corrigida, sobre todas as variáveis, considerando apenas os indivíduos dentro do grupo. Valores altos de R_w^2 definem a formação dos grupos. A medida que o valor de R_w^2 diminui, a consistência do grupo formado também diminui.

Os critérios de DP_w, R_w^2 e R^2 parcial são aplicados a cada ponto de fusão do dendrograma, e com o conjunto dos critérios, em todos os pontos de fusão, defini-se os pontos de corte para a formação dos grupos.

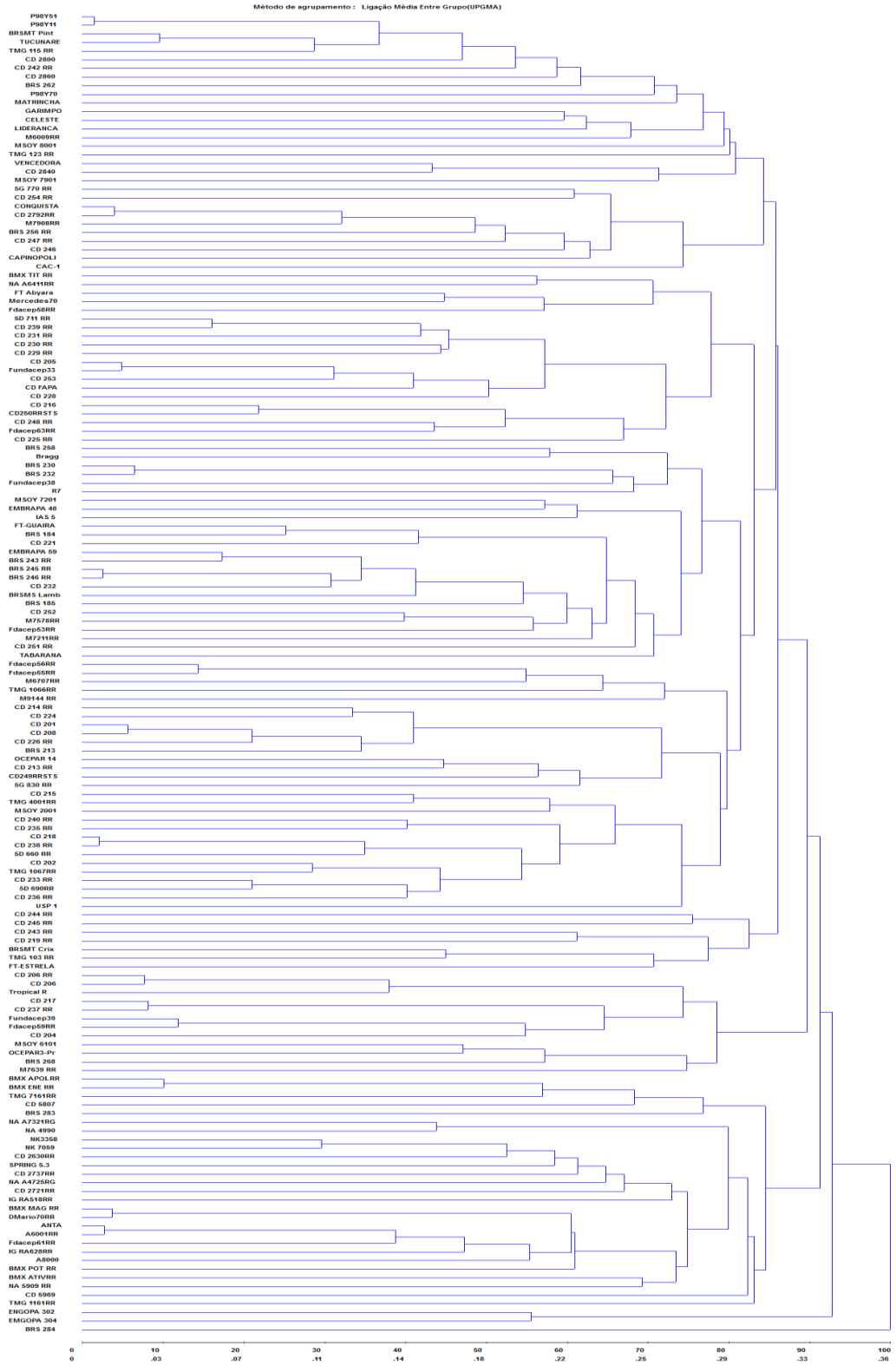


Figura 6. Dendrograma obtido pelo método UPGMA a partir das distâncias genéticas entre 153 cultivares de soja (Oliveira, 2014).

Outro critério, denominado de R^2 parcial mede a redução no valor de R^2 a cada entrada de um novo indivíduo no grupo. Valores pequenos de redução no R^2 indicam pequena perda de homogeneidade do grupo, e o novo indivíduo é admitido no grupo.

Para avaliar se o dendrograma representa a estrutura genética da população, utiliza-se a correlação cofenética. Correlação cofenética (r_p) é obtida pela correlação simples (correlação de Pearson) entre as dissimilaridades genéticas obtidas na avaliação da população, e as distâncias calculadas no dendrograma.

$$r_p = \frac{Cov(x,y)}{\sqrt{\sigma_x^2 \sigma_y^2}}$$

Em que *Cov(x,y)* é a covariância das distâncias originais e as distâncias calculadas a partir do dendrograma, σ_x^2 é a variância das distâncias originais, e σ_y^2 é a variância das distâncias calculadas a partir do dendrograma. Quanto maior a correlação cofenética, melhor o dendrograma representa a estrutura da população. Dendrogramas com correlação cofenética baixa devem ser avaliados com restrições, pois podem não representar relacionamento reais entre os individuos.

O agrupamento do germoplasma baseado em distâncias genéticas é a forma mais comum de representar o relacionamento entre os indivíduos da população, e seu agrupamento. Mas análises de agrupamento podem ser obtidas também por métodos baseados em componentes principais e em modelos, que não utilizam as distâncias genéticas entre os indivíduos para realizar os agrupamentos.

Componentes principais são combinações lineares das variáveis originais (fenotípicas ou moleculares). O número de componentes principais é igual ao número de variáveis, porém os componentes principais são estimados de modo a acumularem o máximo da informação (variância) nos primeiros componentes de serem independentes entre si.

A partir de um conjunto de dados de *v* variáveis (x_{i1}, x_{i2}, ..., x_{iv}), obtém-se um novo conjunto de dados (Y_{i1}, Y_{i2}, ..., Y_{iv}), sendo que:

$Yi1 = a_1x_{i1} + a_2x_{i2} + ... + a_vx_{iv}$

$Yi2 = b_1x_{i1} + b_2x_{i2} + ... + b_vx_{iv}$

Sendo:

$$\sum_j a_j^2 = \sum_j b_j^2 = 1$$

$$\sum_j a_j b_j = 0$$

Esta última restrição é necessária para que os componentes sejam não correlacionados.

A estimação dos componentes principais visa obter o vetor de componentes *a* que maximize a variância de Y_{i1}, em seguida estimar o vetor de componentes *b* que maximize a variância restante em Y_{i2}, e assim por diante. Assim, entre todos os componentes principais, Y_{i1} deve apresentar a maior variância, Y_{i2} a segunda maior variância, até o vetor do último componente, que deve ter a menor variância.

A variância de Y_{i1} é obtida por:

$$V(Y_{i1}) = V(Y_1) = \sum_j a_{jj}^2 r_{jj} \sum_{j \neq} \sum_k a_j a_k r_{jk} = \sum_j \sum_k a_j a_k r_{jk}$$

Em que r_{jk} é o elemento da linha j na coluna k de R, $r_{jj} = 1$ e R é a matriz de covariâncias ou a matriz de correlações fenotípicas entre as variáveis. De forma matricial: $V(Y1) = a'Ra = \lambda_1$

Em que a' é um vetor $1 \times v$ de elementos de a_j ($j = 1, 2, ..., v$). O objetivo da estimação é obter o vetor *a* de forma que a variância de Y_1 seja máxima, sob a restrição no conjunto de soluções de que $a'a = 1$. λ_1 é denominado de autovalor de Y_1, e *a* o seu autovetor associado.

Para estimação do vetor *b*, além da restrição $b'b = 1$, aplica-se também ao conjunto de resultados a restrição $b'a = a'b = 0$. Com isso, além do vetor *b* conferir o máximo da variância restante a Y_2, ainda garante que Y_1 e Y_2 sejam não correlacionados. O mesmo procedimento é adotado na estimação dos demais componentes principais. Os vetores a, b, c,.. contém os escores dos componentes principais Y_1, Y_2, Y_3,.., e os valores dos componentes principais podem ser plotados em gráficos bi ou tri-dimensionais (Figura 7).

Componentes principais são aplicáveis em análises de agrupamento genético quando conseguem reter em seus dois ou três primeiros componentes, pelo menos 80% da variação total. Com isso, a dispersão dos indivíduos em gráficos bi ou tri-dimensionais, utilizando-se os escores dos autovetores dos primeiros componentes principais, permite a avaliação visual da relação existente entre os indivíduos, e da formação de grupos. Caso a variância dos primeiros componentes principais seja menor, a dispersão gráfica dos indivíduos poderá não representar a relação real entre os mesmos.

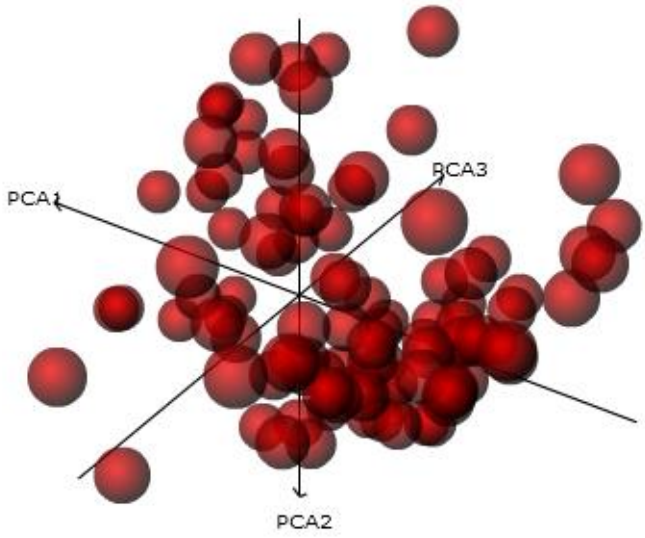


Figura 7. Dispersão gráfica de 72 linhagens de milho, utilizando-se os escores dos três primeiros componentes principais. Para melhor visualização das distâncias entre as linhagens, é necessário utilizar um programa que permita rotação da figura, para poder observar de diferentes ângulos.

Pritchard et al (2000) apresentaram um método baseado em modelo, em que a estrutura da população é estimada ao assumir que as observações de cada grupo são obtidas aleatoriamente de algum modelo paramétrico. As inferências para os parâmetros correspondentes a cada grupo são então feitas em conjunto com a inferência para a associação do cluster para cada indivíduo, usando algum método estatístico padrão (por exemplo, máxima verossimilhança ou métodos Bayesianos).

O modelo desenvolvido para avaliação da estrutura de populações natrais, pode ser aplicado em populações de germoplasma base de melhoramento genético, ou coleções de germoplasma, para verificar a estrutura da coleção de germoplasma. Considerando a avaliação de uma população de germoplasma de um programa de melhoramento, o termo grupo utilizado a seguir representa o termo populações ou subpopulações, nos estudos de populações naturais.

O modelo considera que cada grupo (subpopulação) é definido pelas frequências alélicas do grupo. Consideremos que X é a matriz que contém os genótipos de todos os indivíduos, Z a matriz (desconhecida) que contém o grupo de origem de cada indivíduo, e P a matriz que contém as frequências alélicas (desconhecidas) de caga grupo. Considerando equilíbrio de Hardy-Weinberg dentro dos grupos, e completo equilíbrio de ligação entre os locos dentro da população, cada alelo, em cada loco, em cada indivíduo,

é tomado (amostrado) independentemente a partir de uma distribuição de frequência específica, e define a distribuição de probabilidade *Pr(X|Z,P)*.

A inferência sobre as matrizes *Z* e *P* pode ser feita especificando-se modelos a priori, a partir de inferência Bayesiana: *Pr(Z)* e *Pr(P)*. Considerando a matriz conhecida de genótipos da população (*X*), o conhecimento sobre *Z* e *P* é dado então pela distribuição posterior:

$Pr(Z,P|X) \propto Pr(Z)Pr(P)Pr(X|Z,P)$

Esta distribuição não é fácil de calcular com exatidão, mas é possível obter uma amostragem aproximada ($Z^{(1)}P^{(1)}$), ($Z^{(2)}$,$P^{(2)}$),...($Z^{(M)}$, $P^{(M)}$), para a *Pr(Z,P|X)*, usando cadeias de Markov Monte Carlo(MCMC – *Markov chain Monte Carlo*). As inferências sobre *Z* e *P* são então realizadas pelo resumo das estatísticas obtidas nestas amostragens.

Considerando que nas populações de germoplasma base utilizadas no melhoramento genético, as populações não são compostas por grupos completamente puros, ou seja, indivíduos de um determinado grupo possuem partes do genoma que caracteriza outro grupo, introduz-se um vetor *Q*, que contém as proporções do genoma cada indivíduo que pertencem a cada grupo, e a distribuição utilizada é:

Pr(Z,P,Q|X)

Os elementos de *X*, *Z*, *P* e *Q* são:

$\left(x_l^{(i,1)}, x_l^{(i,2)}\right)$ = genótipo do indivíduo *i* no loco *l*, sendo *i* = 1, 2, ... *N* e *l* = 1,2, ... *L*.

$z_l^{(i,a)}$= grupo de origem do alelo $x_l^{(i,a)}$

p_{klj} = frequência do alelo *j* no loco *l* no grupo *k*, sendo k = 1, 2, ... *K* e *j* = 1, 2, ... J_l é o número de alelos distintos observados no loco *l*.

$q_k^{(i)}$= proporção do genoma do indivíduo *i* que pertence ao grupo *k*.

O interesse para o estudo da estrutura da população está na estimativa do número *K* de grupos e o vetor *Q* de proporções do genoma de cada indivíduo que pertence a cada k_i grupo.

$$\Pr\left(x_l^{(i,a)} = j \middle| Z, P, Q\right) = p_{zl^{(i,a)}lj}$$

$$\Pr\left(z_l^{(i,a)} = k \middle| P, Q\right) = q_k^{(i)}$$

Sendo que $p_{zl^{(i,a)}lj}$ é a frequência do alelo *j* no loco *l* na população de origem do alelo $x_l^{(i,a)}$.

A resolução do modelo para a obtenção de *Z*, *P* e *Q* é feita por método iterativo, utilizando-se o algoritmo MCMC. Iniciando com um valor inicial de $Z^{(0)}$, realizam-se *M* iterações seguindo as etapas abaixo, para *m=1, 2, ..., M*.

Passo 1: amostrar $P^{(m)}$, $Q^{(m)}$ de Pr(P, $Q|X$, $Z^{(m-1)}$)

Passo 2: Amostrar $Z^{(m)}$ de Pr($Z|X$, $P^{(m-1)}$, $Q^{(m-1)}$)

Passo 3: atualizar a matriz Q

A obtenção da o vetor $Z^{(0)}$ para iniciar o processo de iteração pode ser feita considerando-se que a probabilidade do alelo $x_l^{(i,a)}$ ser proveniente do grupo *k* é a mesma para todo *k*.

$\Pr(z_l^{(i,a)} = k) = 1/K$

Sendo *K* maiúsculo o número total de grupos.

O passo 1 consiste em estimar as frequências alélicas de cada grupo e a proporção do genoma de cada indivíduo proveniente de cada grupo, assumindo que o grupo de origem de cada cópia de cada alelo em cada loco é conhecida (*Z*).

O passo 2 consiste em estimar a o grupo de origem de cada alelo em cada loco (*P*), assumindo que as frequências alélicas em cada grupo e a proporção do genoma de cada indivíduo proveniente de cada grupo (*Q*) são conhecidos. Após sucessivas iterações, são obtidas as amostras ($Z^{(m)}$,$P^{(m)}$,$Q^{(m)}$), ($Z^{(m+c)}$,$P^{(m+c)}$,$Q^{(m+c)}$), ($Z^{(m+2c)}$,$P^{(m+2c)}$,$Q^{(m+2c)}$),.... Estas são amostras aproximadamente independentes de *Pr(Z,P,Q|X)*, sendo *c* o tamanho dos intervalos entre as amostra.

A definição do número mais provável de grupos é feita simulando *K* grupos, e estimando-se a deviança Bayesina (*D*) das simulações de cada valor de *k* simulado.

D (Z, P, Q) = -2 log Pr(X|Z, P, Q).

O maior valor para *log Pr(X|Z, P, Q)* identifica o tamanho mais adequado de K. Este modelo está implementado no programa computacional STRUCTURE (http://pritch.bsd.uchicago.edu), e foi estendido posteriormente para considerar desequilíbrio de ligação e correlação entre frequências alélicas (Falush et al., 2003), implementado no mesmo programa. Gao et al. (2007) fizeram uma nova extensão do modelo, permitindo acomodar diferentes níveis de endogamia nas populações, podendo desta forma dispensar a pressuposição de equilíbrio de Hardy-Weinberg. Este modelo está implantado no programa computacional InSctruct (http://cbsuapps.tc.cornell.edu/InStruct.aspx).

Na apresentação dos resultados de estrutura da população cada grupo é representado por uma cor, e cada indivíduo é representaod por um gráfico de barras, contendo a proporção de segmentos das diversas cores, de acordo com a participação de cada grupo no seu genoma (Figura 8).

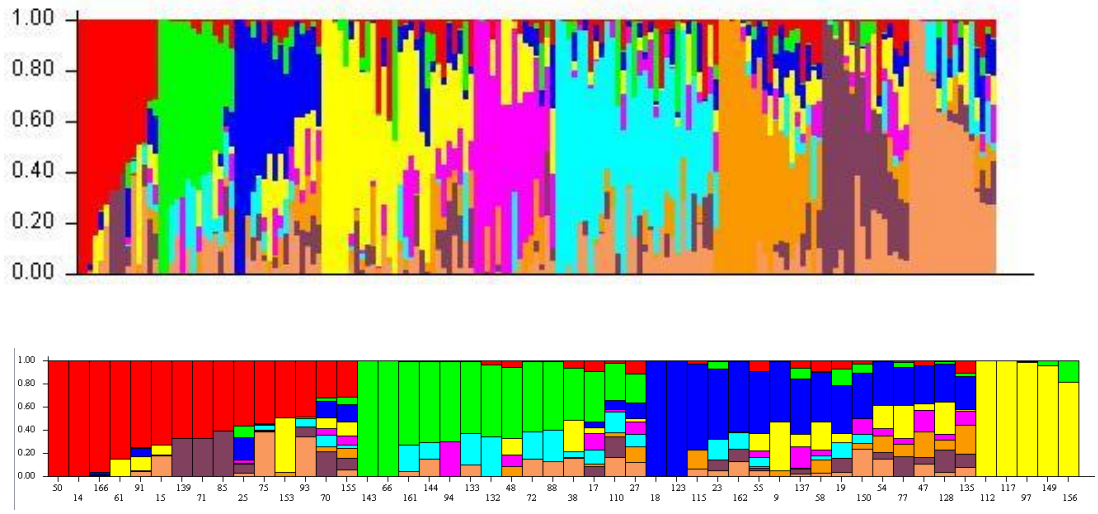


Figura 8. Representação gráfica da estrutura de uma população de variedades de soja. No painel superior, uma representação geral da estrutura de toda a população. No painel inferior, o detalhe da composição genética de cada indivíduo, considerando os indivíduos mais a esquerda do painel superior. Cada indivíduo é representado por uma barra vertical em ambos os painéis. No painel inferior, as barras estão claramente identificadas. Cada cor representa um grupo distinto.

Estimação de Variância e Herdabilidade

Método ANOVA

Tradicionalmente, os componentes de variância são estimados usando os estimadores da análise de variância (ANOVA), os quais são obtidos igualando os quadrados médios observados e esperados a partir de uma análise de variância convencional e resolvendo as equações resultantes (Henderson 1953). Searle (1995) exemplificou os estimadores de variância via método ANOVA. Neste caso, considere o um delineamento inteiramente casualisado (ou ANOVA de um fator) de *a* grupos (exemplo: genótipos) com *n* observações cada. O modelo para a informação *j* no grupo *i* (y_{ij}) é representado como:

$$y_{ij} = \mu + g_i + e_{ij}$$

Para *i*=1, 2, ..., *a*, e *j*=1, 2, ..., *n*; em que: μ representa a média geral (ou intercepto), g_i é o efeito do grupo *i* ou genótipo *i*, e e_{ij} o termo do erro aleatório. As pressuposições do modelo de componentes de variância são:

$E(g_i) = 0, E(e_{ij}) = 0 \quad \forall\ i$ e j,

$V(g_i) = \sigma_g^2, V(e_{ij}) = \sigma_e^2 \quad \forall\ i$ e j,

$COV(g_i, g_{i'}) = 0 \quad \forall\ i \neq i'$,

$COV(e_{ij}, e_{i'j'}) = 0$, exceto quando $i = i'$ e $j = j'$, que corresponde a variância.

Com esse modelo tem-se as somas de quadrados e a esperança de quadrados médios da tabela de análise de variância:

FV	GL	SQ	QM	EQM
Entre genótipos	$a-1$	$SQG = n\sum_{i=1}^{a}(\bar{y}_{i..} - \bar{y}_{..})^2$	$QMG = \frac{SQG}{a-1}$	$E(QMG) = n\sigma_g^2 + \sigma_e^2$
Dentro	$a(n-1)$	$SQD = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.})^2$	$QMD = \frac{SQD}{a(n-1)}$	$E(QMD) = \sigma_e^2$
Total	$an-1$	$SQT = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2$		

FV: fonte de variação; GL: graus de liberdade; SQ: soma de quadrados; QM: quadrados médios; EQM: esperança de quadrados médios.

Igualando os termos dos quadrados médios com seu valor esperado (esperança matemática) tem-se:

$QMG = n\hat{\sigma}_g^2 + \hat{\sigma}_e^2$

$QMD = \hat{\sigma}_e^2$

Neste exemplo os estimadores dos componentes de variâncias obtidos pelo método ANOVA são:

$\hat{\sigma}_e^2 = QMD$, $\hat{\sigma}_g^2 = (QMG\text{-}QMD)/n$

As propriedades deste método clássico, foram resumidas por Searle (1995). Uma desvantagem importante, é que a estimativa da variância genotípica pode ser negativa quando as diferenças entre genótipos são muito pequenas (não significativas) e/ou quando o coeficiente de variação experimental é grande (baixa precisão). Quando um componente de variância genotípica é estimado com valor negativo, diversas pesquisas consideram esse componente como zero ($\hat{\sigma}_g^2 = 0$). Apesar de suas deficiências, o método ANOVA tem sido muito utilizado em experimentos genéticos com dados balanceados em várias

espécies. No entanto, a estimação das variâncias utilizando-se modelos lineares mistos apresenta acurácia bastante superior.

Predição de Valor Genotípico dos Indivíduos

Modelos lineares mistos

Dois dos componentes do ganho genético são a acurácia da determinação do valor genotípico dos indivíduos, e da intensidade de seleção. Modelos lineares mistos e inferência bayesiana podem predizer valor genético com elevada acurácia, permitindo aumentar a intensidade de seleção, sem receio de deixar de selecionar os melhores indivíduos.

Em termos simples, um modelo misto é um modelo estatístico que contém efeitos fixos e efeitos aleatórios. Efeitos aleatórios são de particular importância no melhoramento genético, visto que o pesquisador está interessado em realizar inferências além dos valores particulares da variável independente utilizada em um estudo, geralmente genótipos (famílias, clones, cultivares, etc.). Efeitos fixos podem ser pensados em termos de diferenças (desvios em relação a uma média, por exemplo), enquanto efeitos aleatórios são definidos por uma distribuição e não por diferenças. Portanto, conclusões sobre efeitos aleatórios devem ser expressas em termos de variância.

No melhoramento genético de espécies anuais, o modelo linear misto tem sido usado na estimação de parâmetros genéticos (Freitas et al. 2013, Torres et al. 2015), no mapeamento por associação genôica (Zhang et al. 2010, Famoso et al. 2011, Mora et al. 2015), e na identificação de locos de características quantitativas (QTL) em populações biparentais (Bocianowski e Nowosad, 2015, Prado et al. 2014, Malosetti et al. 2011), dentre outras aplicações. O modelo linear misto pode ser escrito na forma matricial como:

$$y = X\beta + Za + \varepsilon$$

Em que: y é o vetor (dimensão $n\ x\ 1$) de observações, ou seja, a variável resposta observável; β é o vetor (dimensão $p\ x\ 1$) de parâmetros desconhecidos contendo os efeitos fixos; a é o vetor (dimensão $q\ x\ 1$) de efeitos aleatórios; ε é o vetor ($n\ x\ 1$) dos efeitos residuais aleatórios. $X(n\ x\ p)$ e $Z\ (n\ x\ q)$ são as matrizes de incidência para β e a, respectivamente.

Completando a definição do modelo, e admitindo normalidade (N), tem-se que:

$$\begin{bmatrix} y \\ a \\ \varepsilon \end{bmatrix} \sim N\left(\begin{bmatrix} X\beta \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} ZGZ' + R & ZG & R \\ ZG' & G & \varphi \\ R & \varphi & R \end{bmatrix} \right)$$

Em que: *G* é a matriz de variância-covariância genética; *R* é a matriz de variância-covariância residual (usualmente $R = I\sigma_e^2$, em que *I* é a matriz identidade, e σ_e^2 é a variância residual, e assume-se que os resíduos não são correlacionados). *0* é um vetor nulo, e φ é uma matriz nula. Quando as informações de parentesco (pedigree) estão disponíveis, $G = A\sigma_a^2$, em que *A* é a matriz de parentesco associado com as plantas individuais, σ_a^2 é a variância aditiva e *a*, no modelo, corresponde ao vetor dos efeitos genéticos aditivos individuais. A matriz de parentesco pode ser obtida também por marcadores moleculares. Quando as informações de parentesco ou pedigree são desconhecidas, isto é, o objetivo é avaliar apenas o efeito de genótipos (ex. cultivares, famílias, híbridos, etc.) e o efeito de genótipo é considerado aleatório, tem-se $G = I\sigma_G^2$, em que *I* é a matriz identidade, e σ_G^2 é a variância genotípica. Neste caso, deve-se assumir que os genótipos não são correlacionados [covariância genotípica: *COV(G) = 0*].

Melhor predição linear não-viesada

No modelo misto descrito anteriormente, uma vez que que *a* é um vetor de efeitos aleatórios e não um parâmetro fixo, é correto falar de predição de *a* ao invés de estimação de *a*. A predição de *a* pode ser obtida derivando-se a função de verossimilhança em relação a *a*. Para a obtenção das equações de modelos mistos, a função de verossimilhança deve ser derivada conjuntamente em relação a *a* e em relação β (vetor de efeitos fixos). Assim, considerando distribuição normal dos resíduos e dos efeitos aleatórios, tem-se a função densidade de probabilidade (Martins et al. 1995):

$$f(y/X\beta, a, R, G) = f(y, \theta)$$

Para uma amostra (y1, y2, y3..., yn), dada a função de densidade de probabilidade conjunta, vista como função do vetor de parâmetros desconhecidos θ, a função de verossimilhança (*L*) é expressa como:

$$L(\theta, y_1, y_2 ..., y_n) = f(y_1, y_2 ..., y_n, \theta)$$

$$L = \frac{1}{(2\pi)^{n/2}|R|^{1/2}} e^{-1/2[y-X\beta-Za)`R^{-1}(y-X\beta-Za)]} \cdot \frac{1}{(2\pi)^{n/2}|G|^{1/2}} e^{-1/2[a`G^{-1}a]}$$

Os estimadores que maximizam a função de verossimilhança, ou seja, que estimam ou predizem os parâmetros com a máxima probabilidade, são obtidos pela derivação dessa função de verossimilhança em relação a β e a. Uma vez que derivar uma função produtório é muito trabalhoso, utiliza-se a função logaritmo da função de verossimilhança, que é uma função somatório, de derivação mais simples. Os valores de máxima probabilidade da função de verossimilhança e da função logaritmo da verossimilhança, ocorrem nos mesmos valores estimados (ou preditos) dos parâmetros. A função logaritmo da função de verossimilhança é denominada função suporte. A função suporte para a função de verossimilhança de modelos mistos apresentada acima é (Martins et al. 1995):

$$L' = (y - X\beta - Za)'R^{-1}(y - X\beta - Za) + a'G^{-1}a$$
$$= (y'R^{-1} - \beta'X'R^{-1} - a'Z'R^{-1})(y - X\beta - Za) + a'G^{-1}a$$

Diferenciando L' em relação a β e a, e igualando as derivadas a zero, obtém-se a função escore para β e a:

$$\frac{\partial L}{\partial \beta'} = X'R^{-1}X\beta + X'R^{-1}Za - X'R^{-1}y = 0$$

$$\frac{\partial L}{\partial a'} = Z'R^{-1}X\beta + Z'R^{-1}Za + G^{-1}a - Z'R^{-1}y = 0$$

Resolvendo, obtém-se as equações de modelos mistos:

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + G^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} X'R^{-1}y \\ Z'R^{-1}y \end{bmatrix}$$

Isto é, o melhor estimador linear não-viesado de β (ou BLUE, do inglês; *Best Linear Unbiased Estimator*) e a melhor predição linear não-viesada de a (ou BLUP, do inglês: *Best Linear Unbiased Predictor*). Henderson (1973) demonstrou que pode ser estimado β, e predito a, simultaneamente. As propriedades do BLUP e BLUE são, resumidamente (Martins et al. 1995):

- Melhor: é melhor, porque tem mínima variância do erro de predição/estimação (*minimum mean squared error: MSE*) ou seja: $MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$

- Linear: Significa que as predições ou estimações são funções lineares das observações fenotípicas.

- Não-viesado: Significa que o valor esperado da estimativa é igual a seu verdadeiro valor, ou seja: $E(\hat{\theta}) = \theta$

O método BLUP também é chamado de método o procedimento de predição, que é muito popular entre os estatísticos clássicos (ou frequentistas) e tem sido extensivamente usado na análise genética de animais e plantas.

Método de estimação de componentes de variância: REML

No geral, em experimentos com dados desbalanceados os métodos de máxima verossimilhança (os quais requerem algoritmos iterativos mais complexos), são preferidos. Patterson e Thompson (1971) introduziram o método da máxima verossimilhança residual (ou restrita) (REML), que inclui um ajuste dos graus de liberdade utilizados na estimação dos efeitos fixos do modelo linear misto. Esse método tem tido bastante relevância em todas as etapas de análise envolvidas no melhoramento genético de plantas. Piepho e Möhring (2006), por exemplo, analisaram a utilidade dos modelos lineares mistos em etapas avançadas de seleção de cultivares, como em ensaios de valor de cultivo e uso (VCU), os quais consideram ensaios conduzidos em vários ambientes, demonstrando que o método REML é preferível ao da máxima verossimilhança, na estimação dos componentes de variância.

O método REML foi proposto na análise de blocos incompletos com tamanhos de bloco não necessariamente iguais. O método consiste em maximizar a verossimilhança, não de todos os dados, mas sim de um conjunto de contrastes do erro. Portanto, no método REML, a função de verossimilhança é dividida em duas partes independentes, uma referente aos efeitos fixos e outra referente aos efeitos aleatórios. O seguinte modelo misto geral é aplicado, com *g* sendo o vetor dos efeitos aleatórios de genótipos (exemplo: linhagens):

$y = X\beta + Zg + \varepsilon$

Sendo a variância do vetor de observações $V(y) = V = ZGZ' + R$, e assumindo distribuição normal multivariada, com matriz de variância genotípica $G = I\sigma_g^2$, e matriz dos resíduos não correlacionados $R = I\sigma_e^2$. No método REML um conjunto de contrastes ortogonais da forma $K'y$ é usado na função de verossimilhança apenas da parte aleatória (Patterson e Thompson 1971), em que $K'X = 0$, de tal modo que o logaritmo natural da função de verossimilhança é:

$$lnL(K'y) = L_1 = constante - \frac{1}{2} ln|K'VK| - \frac{1}{2} y'K(K'VK)^{-1}K'y$$

Em que a constante é $r(K')ln(2\pi)$, e não depende dos componentes de variância e, podendo ser ignorada. $r(K')$ é o rank da matriz K', sendo $r(K') = N - r(X)$. As seguintes igualdades são obtidas (McCulloch e Searle 2001):

$$ln|K'VK| = ln|V| + ln|X'V^{-1}X|$$

$$y'K(K'VK)^{-1}K'y = (y - X\hat{\beta})'V^{-1}(y - X\hat{\beta})$$

De tal forma que a função L_1 pode ser reescrita como:

$$L_2 = -\frac{1}{2}ln|V| - \frac{1}{2}ln|X'VX| - \frac{1}{2}(y - X\hat{\beta})'V^{-1}(y - X\hat{\beta})$$

Podemos deixar em evidencia as matrizes de variância *R* e *G*, de tal forma que a seguinte igualdade é cumprida (Meyer e Smith 1996):

$$L_3 = -\frac{1}{2}ln|R| - \frac{1}{2}ln|G| - \frac{1}{2}ln|C| - \frac{1}{2}y'Py$$

Sendo *C* a matriz dos coeficientes das equações do modelo misto, descrito anteriormente, e *P* é uma matriz de projeção, isto é:

$$C = \begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + G^{-1} \end{bmatrix} \; ; \; P = V^{-1} - V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}$$

As derivadas da função de verossimilhança podem então ser obtidas por meio da diferenciação de alguma das funções suporte (Meyer e Smith 1996; Martins 1995), sendo a mais comum L_3. Entretanto, de acordo com Van Vleck e Boldman (1993) os termos em L_3 difíceis de computar são $ln|C|$ (logaritmo do determinante da matriz dos coeficientes das equações do modelo misto), e $y'Py = y'R^{-1}y - \hat{\beta}'X'R^{-1}y - \hat{a}Z'R^{-1}y$ (soma de quadrados residual generalizada). Outra opção para maximizar a função de verossimilhança, portanto, é utilizar algoritmos livres de derivada ou DF-REML (do inglês *derivative-free* REML). Smith e Graser (1986) e Graser et al. (1987) introduziram esse procedimento que maximiza a verossimilhança usando um procedimento em que o logaritmo da verossimilhança é calculado por combinações de estimativas dos parâmetros até que a combinação que maximiza a probabilidade é encontrada. A partir desse momento, vários algoritmos DF foram propostos, como por exemplo Meyer (1989) e Boldman e Van Vleck (1991).

Inferência dos efeitos aleatórios ou dos componentes de variância

No modelo ANOVA a inferência de qualquer efeito (fixo ou aleatório) é realizada em base no teste de *F*. Nos modelos lineares mistos, o teste LRT (*Likelihood Ratio Test*- teste da razão de verossimilhança) é o teste natural para comparar dois modelos mistos cujos parâmetros de interesse (exemplo: efeito aleatório do genótipo, ou variância genotípica) foram estimados pelo método da máxima verossimilhança. Em princípio, deseja-se testar a hipótese nula de que um efeito aleatório tem variância igual a zero (exemplo: $H_0 : \sigma_g^2 = 0$). Para isso, comparam-se dois modelos: um modelo completo e um modelo reduzido (ou nulo). O modelo completopode ser, por exemplo:

$$y = X\beta + Zg + \varepsilon$$

E o modelo reduzido, que não inclui o termo aleatório cuja hipótese de nulidade se quer testar é:

$$y = X\beta + \varepsilon'$$

Em que:

$$\varepsilon' = Zg + \varepsilon$$

Se o efeito do genótipo (*g*) for não significativo (ou seja $H_0 : \sigma_g^2 = 0$), então $\varepsilon' = \varepsilon$. A diferença entre os logaritmos de verossimilhança de ambos os modelos, multiplicada por dois gera o teste estatístico *LRT*. No exemplo:

$$LRT = 2 \cdot \left[logL(X\beta + Zg + \varepsilon) - logL(X\beta + \varepsilon') \right]$$

Que é idêntica a razão:

$$LRT = 2 \cdot log\left[\frac{L(X\beta + Zg + \varepsilon)}{L(X\beta + \varepsilon')} \right]$$

Teoricamente *LRT* é uma variável aleatória com distribuição Qui-quadrado com graus de liberdade igual à diferença no número de parâmetros entre os dois modelos, no exemplo, um grau de liberdade ($LRT \sim \chi^2_{p-q}$, em que *p* e *q* são os números de parâmetros do modelo completo e reduzido, respectivamente). É importante ressaltar que as verossimilhanças $L(X\beta + Zg + \varepsilon)$ e $L(X\beta + \varepsilon')$ deveriam ser restritas (variâncias estimadas pelo REML) e, portanto, o teste *LRT* será restrito (RLRT, *restricted likelihood ratio test* – teste da razão de verossimilhança restrito) (Rodovalho et al. 2014; Arnhold et al. 2009).

Uma alternativa ao uso do teste de razão de verossimilhança é adotar os chamados critérios de informação, sendo os mais comuns, por exemplo, o Critério de

Informação de Akaike (AIC; Akaike 1974) e o Critério de Informação de Schwarz ou Bayesiano (SIC ou BIC; Schwarz 1978). Os critérios de informação foram desenvolvidos com base na teoria da informação; um ramo da matemática aplicada, relacionada à quantificação (o processo de contagem e medição) de informações. A diferença em relação ao método *LRT* (que compara modelos aninhados), é que os critérios de informação podem ser aplicados na comparação de modelos tanto aninhados como não aninhados. *AIC* e *BIC* se definem como:

$$AIC = -2logL + 2p$$

$$BIC = -2logL + plog(n)$$

Em que log *L* é o logaritmo da verossimilhança maximizada de um determinado modelo, *p* é o número de parâmetros do modelo, *n* o tamanho amostral. Assim como no *LRT*, o logaritmo da verossimilhança será restrito se os componentes de variância foram estimados pelo procedimento REML. O *BIC* é mais conservador que o *AIC* já que penaliza pelo tamanho amostral; a menos que o tamanho da amostra *n* seja extremamente pequeno, e portanto selecionará modelos mais simples que contenham menos parâmetros a serem estimados.

O modelo escolhido será aquele que minimiza o valor de determinado critério de informação (*AIC* ou *BIC*). Para o exemplo utilizado acima, o modelo completo (que inclui o efeito do genótipo, ou seja $y = X\beta + Zg + \varepsilon$) será escolhido se tiver o menor valor do critério de informação comparado com o valor do modelo reduzido (sem informações do genótipo). Neath e Cavanaugh (2012) propuseram os limites de tais diferenças (particularmente no caso das diferenças serem pequenas). Para o exemplo considerado, seja *BIC*(***y1***) o valor do *BIC* para o modelo reduzido $y1 = X\beta + \varepsilon'$, e *BIC*(***y2***) o valor do *BIC* para o modelo completo $y2 = X\beta + Zg + \varepsilon$, de tal modo que a diferença entre os valores de *BIC* para os modelos ***y1*** e ***y2*** seja representada por $\Delta_{12} = BIC(y1) - BIC(y2)$ (assumindo que o modelo ***y2*** tem um menor valor de *BIC*). Os níveis de evidência fornecidos pelo delta *BIC* (Δ_{12}) são:

Δ_{12}	Evidencia contra o modelo ***y1*** (modelo reduzido), a favor de ***y2***
0-2	Não existe evidencia suficiente a favor do modelo ***y2***
3-6	Evidência positiva contra o modelo ***y1***
7-10	Fortes evidências contra o modelo ***y1***
>10	Muito forte evidência contra o modelo ***y1***

O *BIC* é chamado de "bayesiano" por sua relação matemática com o fator Bayes (BF), que é um teste bayesiano clássico de comparação entre modelos:

$$BF \approx exp\left\{-\frac{1}{2}\Delta_{12}\right\}$$

Em que Δ_{12} é a diferença entre os valores de *BIC* para os modelos ***y1*** e ***y2***, descritos anteriormente, isto é $\Delta_{12} = BIC(y1) - BIC(y2)$.

Inferência dos efeitos fixos

Embora estejamos interessados em utilizar os métodos de máxima verossimilhança (de preferência restrita) na estimação dos componentes de variância, muitas vezes o pesquisador precisa realizar inferências dos efeitos fixos β (vetor de efeitos fixos). Quando os parâmetros de componentes de variância são estimados usando máxima verossimilhança (ML), dois modelos aninhados com diferentes estruturas de efeitos fixos, porem com a mesma estrutura de variância, podem ser comparados usando o teste da razão da verossimilhança (Gumedze e Dunne 2011). O teste RLRT (cujos parâmetros foram estimados pela máxima verossimilhança restrita) não pode ser usado para as inferências dos efeitos fixos do modelo misto, já que o método REML usa apenas a verossimilhança da parte aleatória, desconsiderando a parte dos efeitos fixos. Uma alternativa é utilizar o teste estatístico de Wald, em que os testes de hipóteses e intervalos de confiança para β são construídos usando uma matriz L (combinação linear de β) de dimensão $(l \; x \; p)$. Portanto, testa-se a hipótese $H_0: L\beta = 0$, por meio do estatístico:

$$F = \frac{1}{l}(\hat{\beta} - \beta)'L(L'\hat{\Phi}L)^{-1}L'(\hat{\beta} - \beta)$$

Em que $\hat{\Phi}$ é uma matriz de covariância ajustada para $\hat{\beta}$. Kenward e Roger (1997) propuseram o teste estatístico de Wald escalado (*scaled* Wald) para inferência dos efeitos fixos em pequenas amostras usando o método REML. Esse método basicamente aproxima a distribuição F usando um escalar λ ($0 < \lambda < 1$):

$$F^* = \lambda F = \frac{\lambda}{l}(\hat{\beta} - \beta)'L(L'\hat{\Phi}L)^{-1}L'(\hat{\beta} - \beta)$$

Spilke et al. (2005) demonstraram que o método Kenward-Roger fornece melhor controle do erro tipo I no teste de hipótese e intervalos de confiança dos efeitos fixos, em modelos mistos de experimentos em blocos com dados perdidos.

Inferência Bayesiana

O teorema de Bayes é baseado em sua observação: "Considerando o número de vezes que um resultado eventual tem ocorrido e o número de vezes que não tem ocorrido nas mesmas circunstancias, e denotando por θ a probabilidade, desconhecida para o observador, de ocorrência de dito resultado numa prova individual, deseja-se encontrar um método pelo qual poderíamos obter alguma conclusão de dita probabilidade" (p.298; Bayes 1763). Considera-se θ uma variável aleatória com distribuição de probabilidades "*a priori*" conhecida, a partir da qual é possível a caracterização das propriedades e a definição da distribuição de probabilidades da variável condicionada (θ/Yn) baseada em um conjunto finito de observações.

A regra, axioma ou teorema de Bayes é muito simples, e se deriva de maneira imediata a partir da definição de probabilidade condicional. Tendo-se dois sucessos A e B (onde A e B são ambos sucessos possíveis, ou seja, com probabilidade não nula), então a probabilidade condicional de A dado B:

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

Analogamente,

$$P(B/A) = \frac{P(A \cap B)}{P(A)}$$

Isolando $P(A \cap B)$:

$$P(A \cap B) = P(B/A) \cdot P(A)$$

Substituindo

$$P(A/B) = \frac{P(B/A) \cdot P(A)}{P(B)}$$

Analogamente considerando parâmetros e dados:

$$P(\theta/y) = \frac{P(y/\theta) \cdot P(\theta)}{P(y)}$$

$$P(\theta/y) \propto P(y/\theta) \cdot P(\theta)$$

Em que: θ é o parâmetro (variável aleatória). Temos uma probabilidade a priore [$P(\theta)$] subjetiva em relação a θ, a qual atualizamos, e $P(y) \sim 1$.

Modelagem estatística Bayesiana

Os métodos Bayesianos vêm sendo utilizados desde a década dos 80, no melhoramento animal, como alternativa ao método de estimação/predição REML/BLUP, na estimação dos componentes de (co) variância e parâmetros genéticos (Van Tassel e Van Vleck 1996; Resende 2002). As análises das distribuições a posteriori dos parâmetros de interesse envolvem o cálculo de integrais complexas, as quais são impossíveis de se resolver pela via computacional usual. Devido a isso, os métodos de Monte Carlo via Cadeias de Markov (*Markov Chain Monte Carlo*, MCMC) são utilizados para resolver as integrações numéricas e permitir a análise das distribuições a posteriori (Blasco, 2001). Diversos trabalhos no melhoramento vegetal (Cappa e Cantet, 2006; Arriagada et al., 2012), têm enfatizado as diferenças entre o enfoque Bayesiano e Frequentista (abordagem clássica) para a obtenção dos parâmetros genéticos, os quais têm utilizado diversos algoritmos MCMC, como por exemplo: o amostrador de Gibbs, algoritmo de Cadeias de Independência (*Independence Chain*) e Salto Reversível (*Reversible Jump* – MCMC), dentre outros. Esses métodos MCMC permitem a análise de complexas superfícies de verossimilhança e o cálculo de distribuições a posteriori Bayesianas (Gonçalves-Vidigal et al. 2008). O algoritmo de Gibbs é o método MCMC mais utilizado na avaliação genética via inferência Bayesiana, nos programas de melhoramento animal e de plantas (Blasco, 2001; Mora e Serra 2014; Cappa e Cantet, 2006; Van Tasell e Van Vleck, 1996).

Um aspecto importante da inferência Bayesiana, tem a ver com a conjugação ou modelos de prioris conjugadas (Ehlers 2003). Modelos conjugados são escassos, mas é usual encontrar em modelos simples, e se caracteriza porque podemos encontrar a distribuição a posteriori exata. Quando a distribuição posterior é da mesma família do que o prior, temos a denominada conjugação (*conjugacy*).

Suponha que temos uma família de meios-irmãos de milho com algum grau de resistência a uma determinada doença. Testamos 205 plantas da família durante uma determinada safra. Naquela safra 163 plantas foram resistentes. Suponha que essa família resiste à determinada doença com probabilidade π. O objetivo então é estimar π. Temos 205 observações Bernoulli ou uma observação Y, em que: $Y \sim \text{Binomial}(n, \pi)$. As seguintes pressuposições são assumidas: a) cada progênie (planta) é um ensaio Bernoulli, b) a família tem a mesma probabilidade de ser resistente em cada progênie, c) as respostas de resistência são independentes. Podemos usar a distribuição beta como um prior para π, já que esta suporta [0,1],

$$p(\pi / y) \propto p(y / \pi) \cdot p(\pi)$$
$$= \text{Binomial}(n, \pi) \cdot \text{Beta}(\alpha, \beta)$$
$$= \binom{n}{y} \pi^{y} (1-\pi)^{(n-y)} \cdot \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \pi^{(\alpha-1)} (1-\pi)^{(\beta-1)}$$

$$p(\pi / y) \propto \pi^{y} (1-\pi)^{(n-y)} \cdot \pi^{(\alpha-1)} (1-\pi)^{(\beta-1)}$$

$$p(\pi / y) \propto \pi^{y+\alpha-1} (1-\pi)^{n-y+\beta-1}$$

A distribuição a posteriori é simplesmente uma distribuição Beta $(y+\alpha, n-y+\beta)$, ou seja, é da mesma família do que o prior (*conjugacy*).

Quando temos um modelo de não conjugação, torna-se difícil encontrar a distribuição a posteriori dos parâmetros de interesse. Um exemplo de um modelo de não-conjugação é o modelo de avaliação genética de Henderson. Como exemplo, consideremos o seguinte modelo multi-característica não conjugado, num delineamento de blocos ao acaso (Gonçalves-Vidigal et al. 2008):

$$y_i = X_i \beta_i + Z_i \mu_i + W_i p_i + \varepsilon_i$$

Em que y_i é o vetor das respostas observadas para cada característica considerada na analise (i = 1, 2, ...n); X_i, Z_i e W_i (i=1, 2, ...n) são as matrizes de incidência dos efeitos de blocos, genotípicos e da parcela; β_i (i=1, 2, ...n) é o vetor dos efeitos de blocos; μ_i (i=1, 2, ...n) é o vetor de efeitos genotípicos; p_i (i=1, 2, ...n) é o vetor de efeito da parcela; e ε_i (i=1, 2, ...n) é o vetor dos efeitos de erros. Para a análise multi-característica, as seguintes matrizes de variância-covariância (simétricas) são definidas:

Matriz de variância-covariância genotípica:

$$\hat{G} = \hat{G}_0 \otimes I \text{ , e } \hat{G}_0 = \begin{bmatrix} \hat{\sigma}^2_{g_1} & \hat{\sigma}_{g_1 g_2} & \cdots & \hat{\sigma}_{g_1 g_n} \\ & \hat{\sigma}^2_{g_2} & \cdots & \hat{\sigma}_{g_2 g_n} \\ & & \ddots & \vdots \\ & & & \hat{\sigma}^2_{g_n} \end{bmatrix},$$

Matriz de variância-covariância da parcela:

$$\hat{P} = \hat{P}_0 \otimes I \text{ , e } \hat{P}_0 = \begin{bmatrix} \hat{\sigma}^2_{p_1} & \hat{\sigma}_{p_1 p_2} & \cdots & \hat{\sigma}_{p_1 p_n} \\ & \hat{\sigma}^2_{p_2} & \cdots & \hat{\sigma}_{p_2 p_n} \\ & & \ddots & \vdots \\ & & & \hat{\sigma}^2_{p_n} \end{bmatrix},$$

Matriz de variância-covariância dos resíduos:

$$\hat{R} = \hat{R}_0 \otimes I \text{ , e } \hat{R}_0 = \begin{bmatrix} \hat{\sigma}^2_{\varepsilon_1} & \hat{\sigma}_{\varepsilon_1\varepsilon_2} & \cdots & \hat{\sigma}_{\varepsilon_1\varepsilon_n} \\ & \hat{\sigma}^2_{\varepsilon_2} & \cdots & \hat{\sigma}_{\varepsilon_2\varepsilon_n} \\ & & \ddots & \vdots \\ & & & \hat{\sigma}^2_{\varepsilon_n} \end{bmatrix},$$

Em que $\otimes$ representa o produto de Kronecker. Para avaliação/seleção baseada em plantas individuais (Rodovalho et al. 2008; Gonçalves-Vidigal et al. 2008), tem-se: $\hat{G} = \hat{G}_0 \otimes A$, em que *A* é a matriz de parentesco, e $\hat{G}_0$ contem estrutura de variância-covariância aditiva ($\hat{\sigma}^2_{a_i}$ e $\hat{\sigma}_{a_i a_j}$).

As distribuições a priori mais comumente assumidas para os efeitos de blocos, fenótipos, genótipos e parcela são as seguintes:

$f(\beta) \sim$ Uniforme (*flat prior*),

$f(y \mid \beta,\mu,p,G_0 \otimes I,P_0 \otimes I) \sim \text{Normal}(X\beta + Z\mu + Wp, R_0 \otimes I)$,

$f(\mu \mid G_0 \otimes I) \sim \text{Normal}(0, G_0 \otimes I)$,

$f(p \mid P_0 \otimes I) \sim \text{Normal}(0, P_0 \otimes I)$.

Diversas distribuições a priori são assumidas para os componentes de (co) variância (matrizes de variância-covariância). Alguns exemplos são: ~ Gama Inversa (Wolfinger e Kass 2000; Rodovalho et al. 2014), ~ Wishart invertida (Van Tassell e Van Vleck, 1996; Mora e Serra 2014; Gonçalves-Vidigal et al. 2008), ou ~ Qui-quadrado invertida (Van Tassell e Van Vleck, 1996).

Na análise multi-característica (modelo do exemplo anterior) é sugerida a distribuição a priori Wishart invertida (IW) para G_0, P_0 e R_0 (Van Tassell and Van Vleck, 1996). Ignorando todos os termos que não envolvem o parâmetro de interesse, as densidades condicionais a priori para os componentes de variância-covariância genotípica, parcela e residual são:

$$f(G_0/G^*,q_a) \propto |G|^{-\frac{q_a+m_g+1}{2}} \exp\left(-\frac{1}{2}\text{tr}\left(G^{*-1}G^{-1}\right)\right),$$

$$f(P_0/P^*,q_p) \propto |P|^{-\frac{q_p+m_p+1}{2}} \exp\left(-\frac{1}{2}tr\left(P^{*-1}P^{-1}\right)\right),$$

$$f(R_0/R^*,q_e) \propto |R|^{-\frac{q_e+m_e+1}{2}} \exp\left(-\frac{1}{2}tr\left(R^{*-1}R^{-1}\right)\right).$$

Em que q_a, q_p e q_e são os graus de liberdade das distribuições, equivalentes aos graus de credibilidade Bayesianos; *G**, *P** e *R**, são as matrizes de parâmetros escala; m_g, m_p e m_e são as ordens das matrizes G_0, P_0 e R_0, respectivamente.

A distribuição sintetizada a posteriori é:

$$f(\beta,u,p,G_0 \otimes I,P_0 \otimes I,R_0 \otimes I \mid y,q_a,G^*,q_p,P^*,q_e,R^*) \propto$$
$$f(y \mid \beta,u,p,G_0 \otimes I,P_0 \otimes I) \cdot f(G_0/G^*,q_a) \cdot f(P_0/P^*,q_p) \cdot$$
$$f(R_0/R^*,q_e) \cdot f(\beta) \cdot f(u|G_0 \otimes I) \cdot f(p|P_0 \otimes I)$$

Para examinar a significância da estatística do efeito genotípico ($H_0 : \sigma_g^2 = 0$), na linha Bayesiana, pode se empregar o Factor Bayes (Kass e Raftery 1995), ou também pode ser usado o Critério de Informação de *Deviance* (DIC) como definido por Spiegelhalter et al. (2002):

$$DIC = \overline{D} + pD$$

Em que $\overline{D}$ é uma medida Bayesiana de ajuste de um determinado modelo, e definida como a expectativa posterior da *deviance* ($\overline{D} = E_{\theta/y}\left[-2\ln f(y/\theta)\right]$); *pD* é o número efetivo de parâmetros do modelo (mede a complexidade do modelo). *DIC* é concebido como uma generalização do Critério de Informação de Akaike, descrito anteriormente. Assim como os critérios de informação *AIC* e *BIC*, um valor de *DIC* menor indica um melhor ajuste para o conjunto de dados.

Biometria aplicada a dados genômicos no melhoramento de plantas anuais.

Mapeamento Genético e Seleção Assistida por Marcadores Moleculares

O ganho genético é dependente da intensidade de seleção, acurácia, desvios genéticos aditivos e do número de gerações. A seleção assistida por marcadores moleculares (SAM) tanto de características de herança simples quanto para características quantitativas, pode contribuir para aumentar a intensidade de seleção, uma vez que a seleção é realizada pela avaliação genotípica e não fenotípica. A SAM também pode contribuir com o aumento da acurácia, uma vez que é realizada diretamente a nível de DNA, e não é afetada pela interação genótipo x ambiente, e pode ser aplicada em qualquer fase do desenvolvimento da planta. A SAM também pode aumentar o desvio genético aditivo, uma vez que o efeito aditivo é quantificado no mapeamento de QTLs e a seleção pode ser realizada diretamente sobre o valor aditivo da característica. Por fim, a SAM pode diminuir o intervalo de tempo para a obtenção de novas linhagens ou cultivares, pois pode-se utilizar seleção na contra-estação de cultivo, quando a seleção fenotípica não

pode ser realizada. Com isso, pode-se realizar maior número de gerações de seleção por ano.

Um benefício adicional para o uso da SAM é a redução do custo e do tempo para a seleção, dependendo da complexidade da característica sob seleção e dos métodos de seleção fenotípica utilizados.

A primeira etapa para a utilização da SAM nos programas de melhoramento, é o mapeamento dos genes ou QTLs que controlam a característica de interesse. Mapeamento genético pode ser realizado tanto em populações estruturadas (desenvolvidas pelo cruzamento entre indivíduos contrastantes para a característica de interesse), quanto por populações não estruturadas.

O mapeamento genético de populações estruturadas baseia-se na frequência de recombinação entre locos segregantes nas populações derivadas de cruzamentos biparentais. Diversos tipos de populações de mapeamento podem ser utilizadas, como populações F2, populações derivadas de F2 por autofecundação, RILs, duplo-haploides e retrocruzamento (Schuster e Cruz, 2008). Independentemente do tipo de população utilizada, as estimativas de frequência de recombinação remetem as frequências de recombinação ocorridas nos gametas dos indivíduos F1.

Marcadores moleculares tem distribuição binomial em uma população segregante. Para a estimação da frequência de recombinação, é necessário considerar as frequências de pares de marcadores, e neste caso, existem mais de duas combinações possíveis (genótipos para dois marcadores). Sendo assim, utiliza-se a distribuição multinomial.

A estimação das frequências de recombinação é realizada pelo método da máxima verossimilhança, considerando-se a distribuição multinomial dos pares de marcadores moleculares. A distribuição multinomial possui a seguinte função densidade de probabilidade:

$$f(x) = \frac{N!}{n_1!n_2!...n_n!} p_1^{n1} p_2^{n2} ... p_n^{nn}$$

Em que N é o tamanho da população, p_i são as probabilidades de ocorrência de cada uma das n possíveis classes. A função de máxima verossimilhança para a distribuição multinomial é:

$$L(p_i \mid x_i) = \lambda\, p_1^{n1} p_2^{n2} \cdots p_n^{nn}$$

Em que:

$$\lambda = \frac{N!}{n_1! n_2! \ldots n_n!}$$

A resolução desta função é simplificada pela sua derivação em função da frequência de recombinação, permitindo converter os produtórios em somatórios, facilitando assim a estimação das frequências de recombinação. A função derivada é denominada de função suporte.

$$\ell(p_i \mid x_i) = \ln(\lambda p_1^{n1} p_2^{n2} \cdots p_n^{nn})$$

$$\ell(p_i \mid x_i) = \ln \lambda + n_1 \ln p_1 + n_2 \ln p_2 + \cdots + n_n \ln p_n$$

Igualando-se a função suporte a zero, obtém-se a função escore, que é a função utilizada para a estimação dos parâmetros, neste caso, a frequência de recombinação.

Para o caso do mapeamento genético de uma população F2 utilizando-se marcadores moleculares codominantes, as classes de probabilidades associadas a cada genótipo estão apresentadas na tabela 2.

Tabela 2. Classes genotípicas, números observados e frequências esperadas, dos marcadores em dois locos (A e B) em uma população F2, sendo que A e B podem ser tanto marcadores moleculares quanto genes.

Genótipo	Número observado	Frequência Esperada
AABB	n_1	$p_1 = ¼ (1\text{-}r)^2$
AABb	n_2	$p_2 = ½ r (1\text{-}r)$
AAbb	n_3	$p_3 = ¼ r^2$
AaBB	n_4	$p_4 = ½ r (1\text{-}r)$
AaBb	n_5	$p_5 = ½ (1\text{-}r)^2 + ½ r^2$
Aabb	n_6	$p_6 = ½ r(1\text{-}r)$
aaBB	n_7	$p_7 = ¼ r^2$
aaBb	n_8	$p_8 = ½ r (1\text{-}r)$
aabb	n_9	$p_9 = ¼ (1\text{-}r)^2$

Na tabela 2 pode-se verificar que:

$$p_1 = p_9 = \alpha_1 = \frac{1}{4}(1-r)^2$$

$$p_2 = p_4 = p_6 = p_8 = \alpha_2 = \frac{1}{2} r(1-r)$$

$$p_3 = p_7 = \alpha_3 = \frac{1}{4} r^2$$

$p_5 = \alpha_4 = \frac{1}{2}(1-r)^2 + \frac{1}{2}r^2$

Considerando ainda:

$n_a = n_1 + n_9$

$n_b = n_2 + n_4 + n_6 + n_8$

$n_c = n_3 + n_7$

$n_d = n_5$

$N = n_a+n_b+n_c+n_d$

A função de máxima verossimilhança para a estimação da frequência de recombinação entre os locos A e B é:

$$L(r \mid n_i) = \lambda\left[\frac{1}{4}(1-r)^2\right]^{n_a}\left[\frac{1}{2}r(1-r)\right]^{n_b}\left[\frac{1}{4}r^2\right]^{n_c}\left[\frac{1}{2}(1-r)^2 + \frac{1}{2}r^2\right]^{n_d}$$

E a função suporte é:

$$\frac{\partial \ell(r \mid n_i)}{\partial r} = \frac{(n_b + 2n_c) - 2r(n_a + n_b + n_c)}{r(1-r)} + \frac{2(2r-1)n_d}{(1-r)^2 + r^2}$$

A função escore é obtida igualando-se a função suporte a zero:

$$\frac{(n_b + 2n_c) - 2r(n_a + n_b + n_c)}{r(1-r)} + \frac{2(2r-1)n_d}{(1-r)^2 + r^2} = 0$$

$$(nb + 2nc) - 2(na + 2nb + 3nc + nd)\hat{r} + 2(2na + 3nb + 4nc + 3nd)\hat{r}^2 - 4N\hat{r}^3 = 0$$

Resolvendo-se a função escore acima, obtém-se a estimativa da frequência de recombinação (*r*).

Frequências de recombinação não são aditivas, uma vez que, se ocorrerem dois eventos (ou qualquer outro número par) de *crossing over* no intervalo entre dois locos, o resultado obtido não resulta em recombinação, uma vez que os alelos que se encontravam orginalmente em cis, voltam para este estado cis após número par de eventos de *crossing over*. Para estimar o número de eventos de crossing over a partir das frequências de recombinação, utilizam-se as funções de mapeamento. As mais comuns são as funções de mapeamento de Haldane e de Morgan (Schuster e Cruz, 2008).

Função de Haldane: $mH = -\frac{1}{2}\ln(1-2r)$

Função de Kosambi: $mK - \frac{1}{4}\ln\left(\frac{1-2r}{1+2r}\right)$

As funções de mapeamento estimam as unidades de mapa em Morgan. Multiplicando-se as unidades de mapa em Morgan por 100, obtém-se as unidades em centiMorgan (cM).

Os mapas genéticos de ligação, além de mapearem locos de herança simples (genes únicos), e identificar os marcadores moleculares úteis para seleção destes genes, ainda servem de base para o mapeamento de QTLs (*Quantitative Trait Loci*, ou locos controladores de características quantitativas). Os métodos de mapeamento de QTLs em populações derivadas de cruzamentos biparentais avaliam os intervalos entre marcadores adjacentes. Por isso, necessitam das informações do mapa de ligação dos marcadores.

Mapeamento de QTLs

O mapeamento de QTLs em populações estruturadas utilizando a estratégia de mapeamento de intervalos (IC – *Interval Mapping*), foi inicialmente proposto por Lander e Botstain (1989). Mais tarde, Zeng (1994) propôs o mapeamento por intervalos composto (CIM – *Composite Interval Mapping*), que permite aumento na precisão do mapeamento de QTLs (Schuster e Cruz, 2008). Por ser mais completo, consideraremos apenas o CIM.

Para realização do mapeamento de intervalos, é necessário conhecer as freqüências de recombinação entre os marcadores adjacentes, ao longo de todo o genoma da espécie. Quanto mais saturado o mapa genético, ou seja, quanto mais marcadores estiverem mapeados, e à distâncias menores (intervalos menores entre marcadores) maior a probabilidade de identificar QTLs e sua localização no mapa genético.

Considerando o cruzamento entre duas linhagens homozigotas com genótipos M_1QM_2/m_1qm_2, para os marcadores M_1 e M_2 e para o QTL Q, a figura 9 ilustra o arranjo genotípico da geração F1, e as notações utilizadas.

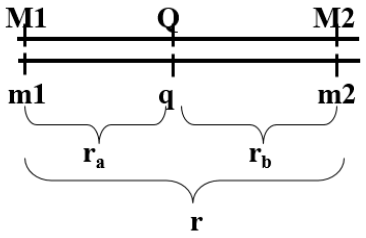


Figura 9. Esquema de um intervalo entre dois marcadores (M_1 e M_2), contendo um QTL a uma frequência de recombinação r_a do marcador M_1 e r_b do marcador M_2. A freqüência de recombinação entre os dois marcadores (r) é obtida pelo mapa de ligação.

No arranjo da figura 9, a frequência de recombinação r é obtida do mapa genético. A frequência de recombinação r_a, entre a posição testada para a presença do QTL e o

marcador M_1, recebe valores que vão de zero a r, com intervalos definidos pelo pesquisador, mas geralmente de 0,01 (1% de recombinação). O valor de r_b é estimado a partir de r_a, para completar o intervalo r. Os testes de hipótese são realizados a cada valor de r_a simulado.

Os modelos de mapeamento de QTLs por intervalos (tanto IM quanto CIM) visam testar a hipótese de que exista um QTL no intervalo entre cada par de marcadores, a cada posição r_a simulada dentro destes intervalos. O modelo de CIM utiliza, além da informação dos genótipos dos marcadores M1 e M2 que estão flanqueando o intervalo que está sendo avaliado, também os genótipos de outros marcadores, em outros cromossomos, que também estejam associados à característica que está sendo avaliada (Figura 10).

A estimação dos parâmetros em cada posição de r_a pode ser realizada por máxima verossimilhança ou por regressão linear, sendo que quando os intervalos são pequenos, o que é desejado para o mapeamento de QTLs, os dois modelos apresentam resultados muito similares, sendo que o modelo de regrassão linear é de resolução muito mais simples.

Haley e Knott (1992) apresentaram um modelo de mapeamento de intervalos baseados em modelo de regressão (simples e múltipla). O modelo de regressão para mapeamento por intervalos de QTLs em populações F2 é:

$Yj = \mu + ax^* + dz^* + ej$

Em que:

Yj é o valor da característica Y no indivíduo j.

μ é a média da característica na população.

a é o efeito aditivo do loco que está sendo estudado, sobre a característica.

d é o efeito de dominância do loco que está sendo estudado, sobre a característica.

*x** e *z** são as variáveis condicionadoras e dependem dos genótipos dos marcadores flanqueando o QTL, no indivíduo *j* (Tabela 3).

Este modelo pode ser facilmente ajustado para o mapeamento por intervalos composto, fazendo:

$$Y_j = [\mu + ax^* + dy^*] + \sum_{k \neq i,i+1}^{c} b_k x_{kj} + e_j$$

Em que:

Y_j é o valor da característica Y no indivíduo j.

μ é a média da característica Y na população.

a é o efeito aditivo do loco que está sendo estudado, sobre a característica Y.

d é o efeito de dominância do loco que está sendo estudado, sobre a característica Y.

*x** e *z** são as variáveis condicionadoras e dependem dos genótipos dos marcadores flanqueando o QTL, no indivíduo *j* (Tabela 3).

$\sum_{k \neq i,i+1}^{c} b_k x_{kj}$ é o somatório dos efeitos dos marcadores utilizados como cofatores no mapeamento por intervalo composto;

k identifica o loco marcador;

j identifica o indivíduo sendo considerado;

c é o número de marcadores utilizados como cofatores.

A estimação é feita utilizando-se quadrados mínimos. A i-ésima linha da matriz X_{ra} é dada por *1*; *x*(*M_i*,* r_a*)*; *z*(*M_i*,* r_a*);C1; C2;…Cn*, sendo C os cofatores. Os valores de x* e z* são obtidos da tabela 3. Para os cofatores utilizam-se os códigos dos marcadores: -1 para um dos homozigotos, 1 para o outro homozigoto, e zero para os heterozigotos. Com estes valores codificados para os genótipos dos marcadores, obtém-se uma estimativa do efeito aditivo de cada marcador (aa=-a, Aa=0 e AA=+a).

Matrizes Y, X e β utilizadas no modelo de mapeamento por intervalos composto possuem o seguinte formato:

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} \qquad X_{ra} = \begin{bmatrix} 1 & x_1^* & z_1^* & C1_1 & ... & Cn_1 \\ 1 & x_2^* & z_2^* & C1_2 & ... & Cn_2 \\ 1 & x_3^* & z_3^* & C1_3 & ... & Cn_3 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_n^* & z_n^* & C1_n & ... & Cn_n \end{bmatrix}$$

$$\beta_{r_a} = (X_{r_a}' X_{r_a})^{-1} X_{r_a}' Y = \begin{bmatrix} \mu \\ a \\ d \\ C1 \\ ... \\ Cn \end{bmatrix}$$

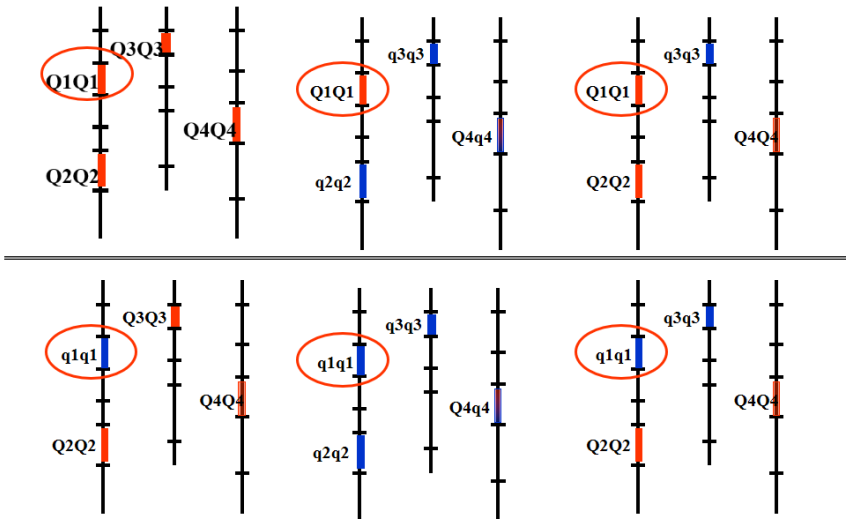


Figura 10. Ao utilizar marcadores moleculares flanqueando o intervalo contendo o QTL 1, os indivíduos foram fixados em função do genótipo para o QTL1. Mas em cada grupo contendo os genótipos alternativos do QTL1, existem outros QTLs segregando, o que aumenta o ruído na detecção do QTL1. Ao utilizar marcadores associados aos demais QTLs como cofatores no modelo de mapeamento de QTLs, os efeitos destes outros QTLs são fixados, reduzindo-se o ruído, e aumentando a precisão e o poder de detecção do QTL1.

A regressão é computada para cada valor de r_a entre cada par de marcadores. Valores de r_a significativos indicam presença de um QTL na posição do r_a específico. O teste de significância para a presença de um QTL em cada posição de r_a é a razão de verossimilhança (LRT – *Likelihood Ratio Test*), que em mapeamento de QTLs é denominado de LR.

Assumindo que os fenótipos têm distribuição normal em cada genótipo do QTL, se o QTL está completamente ligado a um dos marcadores (r_a=0 ou r_a=r) o resíduo da regressão acima é normalmente distribuído. Neste caso a estimativa da regressão é também uma estimativa de máxima verossimilhança e o teste LR pode ser expresso como:

$$LR = n \ln \frac{SQR_{(reduzido)}}{SQR_{(completo)}}$$

em que:

$SQR_{(reduzido)}$ = Soma de quadrados do resíduo do modelo reduzido

$SQR_{(completo)}$ = Soma de quadrados do resíduo do modelo completo.

Os modelos completos e reduzidos a serem testados estão apresentados na tabela 4.

Tabela 3. Esperança dos efeitos genotípicos médios para um QTL no intervalo entre os marcadores A e B, para todos os possíveis genótipos dos marcadores flanqueando o QTL, em uma população F2, em função da posição do QTL dentro do intervalo, quando não se considera a ocorrência de dupla recombinação.

	Probabilidade condicional[#]			Variáveis condicionadoras	
Genótipo	QQ	Qq	qq	(a)x*	(d)z*
AABB	1	0	0	1	0
AABb	1-p	p	0	1-p	P
AAbb	$(1-p)^2$	2p(1-p)	p^2	1-2p	2p(1-p)
AaBB	p	1-p	0	p	1-p
AaBb	cp(1-p)	1-2cp(1-p)	cp(1-p)	0	1-2cp(1-p)
Aabb	0	1-p	p	-p	1-p
aaBB	p^2	2p(1-p)	$(1-p)^2$	-(1-2p)	2p(1-p)
aaBb	0	p	1-p	-(1-p)	P
aabb	0	0	1	-1	0

p=ra/r ; (1-p)= rb/r; q=r2/[r2+(1-r)2]

probabilidade de ocorrência de cada um dos genótipos do eventual QTL no intervalo entre os marcadores A e B, dado os genótipos dos marcadores em cada linha.

Tabela 4. Modelos completos e reduzido para a estimativa de LR em diferentes populações de mapeamento. Para F2 o número de graus de liberdade é 2, e para as demais populações, o número de graus de liberdade é 1.

Modelo	População	CIM
Completo	F_2	$Y_j = [\mu + ax^*_{ij} + dz^*_{ij}] + \sum_{k \neq i,i+1}^{c} b_k x_{kj} + e_j$
	RC, RIL e DH	$Y_j = [\mu + gx^*_{ij}] + \sum_{k \neq i,i+1}^{c} b_k x_{kj} + e_j$
Reduzido	F_2, RC, RIL e DH	$Y_j = \mu + \sum_{k \neq i,i+1}^{c} b_k x_{kj} + e_j$

O valor crítico para declarar a existência de um QTL em uma determinada posição r_a pode ser obtido por um teste permutações. Para realização deste teste, indexam-se os indivíduos de 1 até n. Os valores fenotípicos das características são permutados entre os indivíduos, e estes dados permutados são submetidos a análise para a presença de QTLs. Os resultados desta análise são armazenados, e o procedimento (permutação, análise de QTLs e armazenamento dos resultados) é repetido N vezes. Com a permutação dos valores fenotípicos entre os indivíduos, qualquer associação entre marcadores e valores fenotípicos é destruída. Os resultados obtidos nestas análises correspondem aos resultados esperados sob a hipótese nula. Ao final deste procedimento, teremos armazenados os resultados das N análises com os dados permutados. Estes valores são ordenados por ordem de magnitude.

O valor crítico de LR para um nível α de significância é o valor de magnitude N(1-α), ou seja, após 1.000 permutações o valor crítico para um nível de significância de 5% é o 950º valor ordenado.

São necessárias 1.000 permutações para declarar um nível de significância de 5% que seja estável, e cerca de 10.000 permutações para um nível de significância de 1% estável (Churchill e Doerge, 1994). A figura 11 ilustra o mapeamento de um QTL de resistência ao nematoide de cisto da soja no GL G da soja.

Mapeamento de QTLs por Análise de Associação

O mapeamento de QTLs por análise de associação, ou por estudo da associação genômica ampla (GWAS – *Genome-wide Association Study*) detecta e localiza QTLs baseado na intensidade da correlação entre marcadores moleculares e as características fenotípicas. A associação de um marcador com um QTL ocorre quando ambos estão no mesmo bloco de desequilíbrio de ligação.

O desequilíbrio de ligação é definido como a associação não ao acaso de alelos em locos diferentes do mesmo cromossomo. Se considerarmos que um determinado loco possui os alelos A1 e A2 com frequências de pA1 e pA2, e um segundo loco possui os alelos B1 e B2 em frequências de pB1 e pB2, então, no equilíbrio, mesmo que os locos estejam ligados, a frequência esperada de um determinado haplótipo é o produto das frequências dos alelos constituintes deste haplótipo. No equilíbrio, pA1B1=pA1pB1, sendo que A e B podem ser, por exemplo, locos de marcadores moleculares e de QTLs. Uma vez que o tamanho dos blocos de desequilíbrio de ligação é pequeno, apenas correlações entre QTLs e marcadores proximamente ligados devem permanecer, facilitando o mapeamento mais fino.

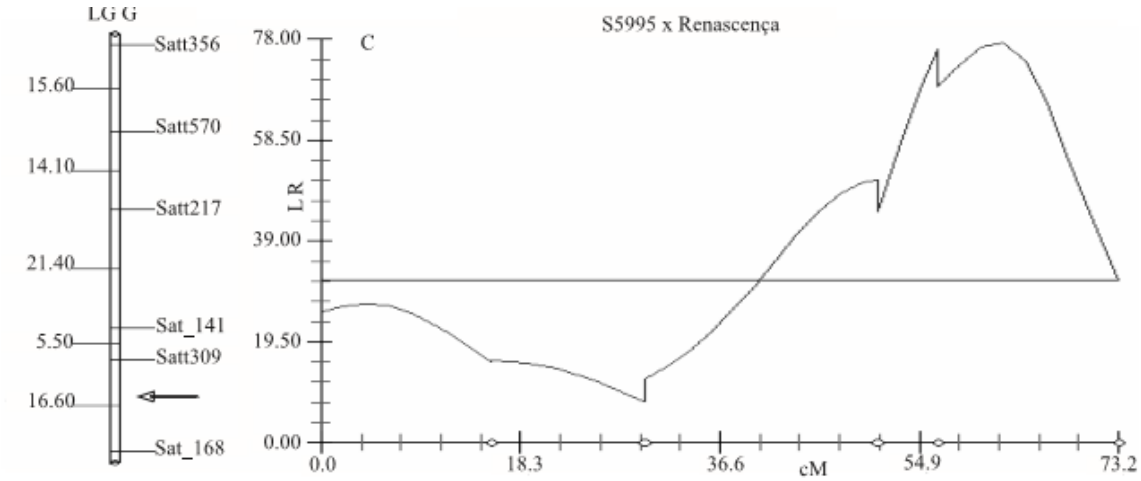


Figura 11. Identificação de um QTL para resistência ao Nematóide de Cisto da Soja. O mapa de ligação está apresentado à esquerda, e o mapeamento de QTL a direita. A linha horizontal representa o ponto de corte para a significância a 1% de probabilidade, obtido pelo teste de permutação. A parte da curva de probabilidade que ultrapassa a linha de corte indica presença de um QTL nesta região. A posição mais provável do QTL está indicada por uma seta no mapa de ligação à esquerda. Fonte: Silva et al. (2007).

Para o mapeamento por associação genômica, não há necessidade de preparar uma população de mapeamento. Esse tipo de mapeamento utiliza uma coleção de cultivares, ou linhagens de populações de melhoramento, para detectar blocos em desequilíbrio de ligação. Para identificar os blocos de desequilíbrio de ligação, é necessário utilizar grande densidade de marcadores.

Vários métodos de associação podem ser utilizados para detectar associação de marcadores moleculares com QTLs. Podem ser utilizados, por exemplo, análise de correlação entre os escores dos marcadores e o fenótipo da característica até modelos mais robustos, como os modelos lineares (mistos ou generalizados), modelos derivados de BLUP e modelos derivados do método Bayesiano.

Para evitar associações falso-positivas em função da existência de sub-grpos estruturados dentro da população de estudo, uma análise de estrutura de população deve ser realizada, e a matriz *Q*, contendo as informações da estrutura da população (como visto no tópico de Análise de diversidade genética e classificação do germoplasma) deve ser incorporada no modelo. A análise de associação dos dados genotípicos com o fenótipo utilizando método de Modelos Lineares Mistos tem o seguinte modelo:

$Y = X\beta + Zu + e$

em que *y* é o vetor (*n x 1*) de observações fenotípicas observadas, β é um vetor contendo os efeitos fixos, incluído os marcadores, a estrutura da população (Q) e o intercepto, *X* é a matriz *Q* contendo a estrutura da população, *u* é um vetor de efeitos genéticos aditivos

aleatórios dos múltiplos QTLs no background dos genótipos (a ser estimado), considerando a coancestralidade dos indivíduos; Z $(n\ x\ q)$ é uma matriz de incidência para *u*, e *e* é o vetor de efeitos residuais aleatórios. O vetor *u* tem distribuição normal, com média zero e variância $K\sigma_a^2$ e *e* tem distribuição normal com média zero e variância $I\sigma_e^2$. *K* é a matriz (*n x n*) de parentesco obtida por marcadores moleculares pelo método IBS (*Identity by Similarity*, ou identidade por semelhança). A matriz de parentesco (*K*) é calculada pela proporção de SNPs idênticos entre pares de indivíduos.

A hipótese nula para cada marcador molecular é que não haja associação com o fenótipo. Esta hipótese é testada para cada marcador individual, por algum teste estatístico, como teste de t ou teste de F, por exemplo. As probabilidades associadas a cada marcador geralmente são expressas através da dispersão gráfica de –log10(P), através de gráficos denominados de Manhatan Plot (Figura 12), para melhor visualização da associação marcador-fenótipo. O nível de significância para a associação SNP/fenótipo, como em qualquer estudo, depende das probabilidades de erro tipo I e erro tipo II admitidas. Em geral, níveis de significância iguais ou inferiores a 0,1% são utilizados. Nível de significância de 0,1% significa valor de -log10(P) igual a 3. Para visualização dos resultados, os valores de -log10(P) são plotados para cada marcador em um gráfico Manhatan Plot (Figura 12).

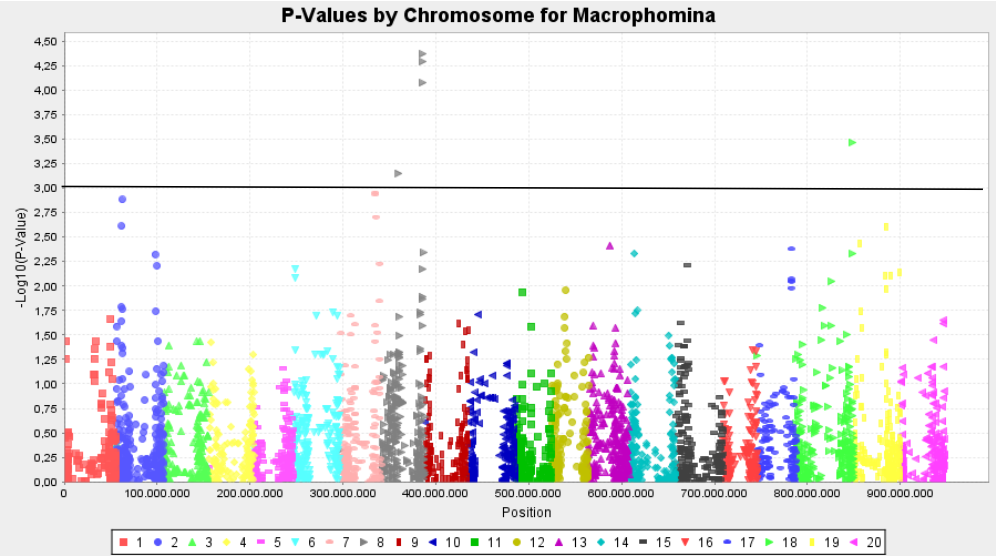


Figura 12. Manhatan Plot. Plotagem dos valores de *-Log10(P)*, em que *P* foi obtido pela análise de associação por Modelos Lineares Múltiplos. Cada cor e forma representa um cromossomo. A linha horizontal representa o ponto de corte para 0,1% de probabilidade [*-Log10*(0,001)=3]. Pontos acima da linha de corte indicam região genômica associada com a característica de interesse.

Por utilizar uma coleção de cultivares ou linhagens, o mapeamento por associação genômica apresenta diversas vantagens sobre o mapeamento utilizando cruzamentos bi-parentais: uma ampla variação genética é representada por diversos haplótipos, não sendo os marcadores ou QTLs limitados apenas aos alelos ou haplótipos presentes nos dois genitores; o mapeamento de associação tem uma maior resolução por acumular todas as meioses ao longo das gerações; os dados fenotípicos que caracterizam cada cultivar, inclusive dados já obtidos anteriormente, podem ser utilizados, não havendo a necessidade de testá-los com populações especiais de mapeamento, diminuindo o tempo para a realização do mapeamento de QTLs (Buckler IV e Thornsberry, 2002; Sorkheh et al., 2008)

Mapeamento de QTLs por recombinação, em populações estruturadas, e por associação, em coleções de germoplasma, não são exclusivos. A utilização de ambos os métodos possibilita maior precisão nos resultados de mapeamento, especialmente quando alelos raros estão presentes, pois a associação de alelos raros com uma característica fenotípica não é precisamente detectada no mapeamento por associação. Neste caso, mapeamento de QTLs em populações estruturadas é mais indicado, pois em uma população derivada de cruzamento entre linhagens endogâmicas, a frequência de todos os alelos segregantes é 0,5.

Seleção genômica ampla: predição do valor genético total

O melhoramento assistido por informações genômicas pode ser baseado na seleção assistida por marcadores (MAS; *marker-assisted selection*) ou na seleção genômica (GS, *Genome Selection* ou GWS, *Genome-wide Selection*) (Meuwissen et al. 2001; Crossa et al. 2014). De acordo com Lui et al. (2016), a eficiência comparada entre MAS e GWS depende da arquitetura genética subjacente da característica sob seleção. Na MAS, os valores genotípicos dos indivíduos são preditos baseado nos efeitos de marcadores selecionados, e é mais efetivo em características controladas por poucos QTL cada um controlando uma proporção relativamente grande da variação fenotípica. Por outra parte, GWS é preferível para características complexas que são controlados por um número grande de genes e, sobre a qual, os fatores ambientais são muito relevantes. A seleção genômica pode aumentar a taxa de ganho genético, aumentando a precisão dos valores genéticos estimados, e reduzindo o intervalo de gerações de melhoramento, com uma melhor utilização dos recursos genéticos disponíveis (Daetwyler et al. 2013).

No método GWS, desenvolve-se um modelo de predição genômica, incluindo a predição simultânea dos efeitos genéticos de milhares de marcadores de DNA, sem a utilização de testes de significância para as marcas individuais.

A acurácia da predição do modelo de GWS depende de quatro parâmetros (Hayes et al. 2009; Grattapaglia 2014): a. A extensão do desequilíbrio de ligação entre marcadores e QTL, b. O tamanho populacional ou número de indivíduos com fenótipos e genótipos na população de treinamento, a partir da qual são estimados os efeitos dos marcadores, c. A herdabilidade da característica estudada, e d. A distribuição dos efeitos dos QTL.

Uma vez que método GWS consiste na seleção simultânea para dezenas de milhares de marcadores, cobrindo todo o genoma, é esperado que todos os alelos de interesse estejam em desequilíbrio de ligação com um ou mais marcadores genotipados, e seus efeitos serão capturados nos modelos preditivos. O sucesso da GWS como método de MAS no melhoramento genético depende da extensão dos blocos de desequilíbrio de ligação nas populações avaliadas. Populações de melhoramento em que poucos indivíduos parentais foram utilizados como população base para a geração da variabilidade são mais indicadas para a utilização deste método.

Valores genéticos genômicos (GEBV) e modelos de predição

Na seleção genômica os efeitos de marcadores que tenham uma cobertura ampla do genoma (ex. SNP) podem ser estimados baseados nos seguintes procedimentos: RR–BLUP (Random Regression – BLUP; Meuwissen et al. 2001), BRR (Bayesian ridge regression), BayesA, BayesB, BayesC$_\Pi$ ou BayesLASSO (Bayesian Least Absolute Shrinkage and Selection Operator). Em todos os casos, as informações genotípicas podem ser ajustadas usando o seguinte modelo padrão:

$$y_C = 1\mu + Zm + \varepsilon$$

Em que y_C é o vetor de fenótipos deregredido dos valores genéticos preditos, μ é a média geral, *1* é um vetor de uns, *m* é o vetor dos efeitos de marcadores (aleatórios do ponto de vista dos modelos mistos), ε é o vetor dos efeitos residuais, e *Z* é a matriz de incidência de *m*.

A diferença entre BayesLASSO e os outros métodos Bayesianos é a pressuposição a priori para os efeitos dos marcadores. Por exemplo, BayesLASSO assume que os efeitos de todos os marcadores seguem uma distribuição dupla exponencial:

$$p(q) = \prod_{j=1}^{k} 1/2\,\lambda \exp(-\lambda|q_j|)$$

Em que *q* é o vetor dos efeitos dos marcadores; *k* é o número de marcadores; λ é um parâmetro taxa (*rate parameter*) que é amostrado a partir da distribuição uniforme.

O modelo BayesA é equivalente ao modelo de regressão *shrinkage* Bayesiano (BSR) (George et al. 1993), em que todos os marcadores são incluídos como co-variáveis no modelos de predição, e as estimativas dos efeitos dos marcadores são reduzidas ao assumir uma distribuição normal com media *0* e variância especifica para cada marcador, como distribuições a priori dos efeitos dos marcadores.

BayesB é considerado uma variante do método Bayesiano de seleção de variáveis de busca estocástica (BSSVS, *Bayesian stochastic search variable selection*; Xu 1989). Uma distribuição normal com media *0* e variância especifica para cada marcador é assumida como efeito dos marcadores a priori em BayesB (tal como em BayesA) se os marcadores são incluídos no modelo com probabilidade 1-π e, caso contrário, a média e a variância são fixadas em *0* com probabilidade π (Hayashi e Iwata 2013).

Os modelos Bayesianos de predição são ajustados usando os procedimentos estatísticos que extraem amostras a partir da densidade posterior usando alguma variante dos métodos MCMC, como por exemplo, o algoritmo de amostragem de Gibbs com atualização escalar. Comparativamente, a qualidade do modelo Bayesiano pode ser checada usando o critério de informação de *deviance* (DIC).

Em GWS, cada planta individual (ou linha endogâmica, por exemplo) deverá ter seu EBV genômico (GEBV, *Genomic Estimated Breeding Value* – Valor Genônico Estimado de Melhoramento), o qual é obtido multiplicando a matriz de incidência *Z* (do marcador) pelo vetor dos efeitos dos marcadores estimados, de acordo com a seguinte expressão (Resende et al. 2012b):

$$\hat{g}_j = \sum_{i}^{n} Z_{ij} \cdot \hat{m}_i$$

Em que *i* é o alelo específico do marcador *i* sobre o indivíduo *j*, e *n* é o número total de marcadores.

Para a execução deuma programa de GWS, é necessário inicialmente desenvolver o modelo em uma população de treinamento. Esta população de treinamento é dividida em duas partes. Uma parte dos indivíduos são utilizados para o desenvolvimento do modelo de predição (população de descoberta), e outra parte é utilizada para a validação

do modelo população de validação). A figura 13 ilustra as etapas para o desenvolvimento e implantação da GWS em um programa de melhoramento de plantas anuais.

O ideal é que as populações de treinamento e de validação sejam populações independentes. Modelos mais robustos são obtidos quando se dispõe de dados de mais locais e de mais de um ano. Os modelos preditivos podem ser obtidos com os dados de avaliaçãode um ou mais anos, e a validação pode ser realizada no ano subsequente ao desenvolvimento do modelo. Além disso, a medida que o modelo é utilizado, e os híbridos ou variedades selecionadas pelo modelo são avaliados a campo, os resultados destas avaliações podem retroalimentar o modelo, tornando-o cada vez mais preciso.

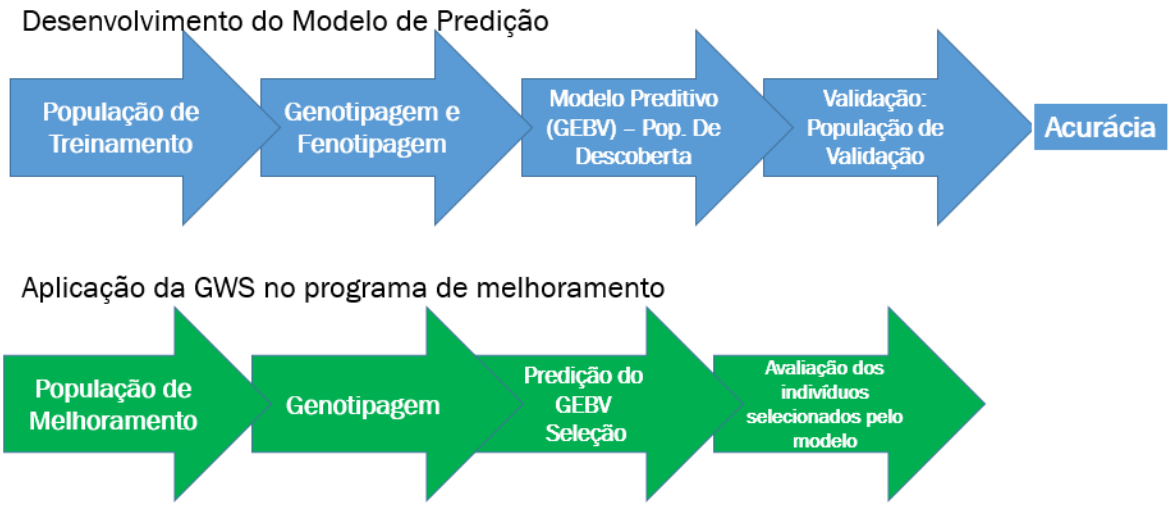


Figura 13. Ilustração esquemática do desenvolvimento de um modelo de predição e aplicação em um programa de melhoramento de plantas.

O objetivo da utilização de GWS nos programas de melhoramento não é diminuir a quantidade de híbridos oulinhagens a serem avaliados a campo. O objetivo é aumentar a base de seleção. Um programa de melhoramento que consiga conduzir N híbridos ou linhagens por ano, pode sintetizar “in silico”, 2N ou mais híbridos ou linhagens, e usar o modelo para selecionar os N que possuam maior EBV Genômico, e assim, aumentar a probabilidade de obter produtos superiores nas avaliações de campo.

Considerações Finais

Na última década, o melhoramento genético de plantas anuais, especialmente as comodities como soja e milho, passaram por grandes transformações, em função da tecnologia disponível para ser aplicada no melhoramento. Desde o uso de equipamentos modernos de plantio e colheita, novas ferramentas de análise de dados, e incorporação de

biotecnologia, as novas tecnologias permitiram modificar programas de melhoramento que permaneciam inalterados por décadas. A exigência por produtos cada vez mais produtivos e completos em termos de sanidade, qualidade e estabilidade, e a grade concorrência no mercado de sementes, fez com que tivessem que ser desenvolvidos métodos de "melhoramento acelerarado". Métodos biométricos conhecidos há tempos, e outros desenvolvidos recentemente, podem agora ser utilizados em rotinas de programasde melhoramento graças a evolução dos equipamentos de informática, capazer de analizar quantidades enormes de dados. A associação de modelos biométricos que tenham capacidade de predizer o valor genético dos indivíduos com elevada acurácia, com as novas tecnologias disponíveis aos melhoristas, são essenciais para atingir os objetivos de ganhos genéticos crescentes.

Referências

BOCIANOWSKI, J.; NOWOSAD, K. Mixed linear model approaches in mapping QTLs with epistatic effects by a simulation study. **Euphytica**. v.202. p.459-467, 2015.

BUCKER IV, E.S.; THORNSBERRY, J.M. Plant molecular diversity and application to genomics. **Current Opinion in Plant Biology.** v.5, p.107-111, 2002.

CROSSA, J.; PEREZ, P.; HICKEY, J.; BURGUEÑO, J.; ORNELLA, L.; CERÓN-ROJAS, J.; ZHANG, X.; DREISIGACKER, S.; BABU, R.; LI, Y.; BONNETT, D.; MATHEWS, K. Genomic prediction in CIMMYT maize and wheat breeding programs. **Heredity. v.**112, p.48–60, 2014.

DAETWYLER, H.D.; CALUS, M.P.L.; PONG-WONG, R.; De Los CAMPOS, G.; HICKEY J.M. Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. **Genetics**. v.193, n.2 p.347-365, 2013.

DOERGE, R.W.; CHURCHILL, G.A. Issues in genetics mapping of quantitative trait loci. In: LOWER, R. (Ed.). **ASHS/CSSA Joint Plant Breeding Symposium on Analysis of Molecular Marker Data.** Corvallis: Oregon State University, p.27-32, 1994.

FAMOSO, A.N.; ZHAO, K.; CLARK, R.T.; TUNG, C.W.; WRIGHT, M.H.; BUSTAMANTE, C.; KOCHIAN, L.V.; McCOUCH, S.R. Genetic Architecture of Aluminum Tolerance in Rice (Oryza *sativa*) Determined through Genome-Wide Association Analysis and QTL Mapping. **PLoS Genet**. v.7, n.8, e1002221, 2011.

FREITAS, I.L.J.; AMARAL JUNIOR, A.T.; VIANA, A.P.; PENA, G.F.; CABRAL, P.S.; VITTORAZZI, C.; SILVA, T.R.C. Ganho genético avaliado com índices de seleção

e com REML/Blup em milho-pipoca. **Pesquisa agropecuária brasileira.** v.48, n.11, p.1464-1471, 2013.

GEORGE, E.I.; McCULLOGH, R.E. Variable selection via Gibbs sampling. **Journal of American Statistic Association.** v.88, p.881–889, 1993.

GRATTAPAGLIA, D. Breeding forest trees by genomic selection: current progress and the way forward. Chapter 26. In: **Advances in genomics of plant genetic resources.** Tuberosa, R, A Graner, E Frison (Eds.). New York: Springer; p.652-682, 2014.

HALEY, C.S.; KNOTT, S.A. A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. **Heredity.** V. 69, P.315–324, 1992.

HAYASHI, T.; IWATA, H. A Bayesian method and its variational approximation for prediction of genomic breeding values in multiple traits. **BMC Bioinformatics**. v.14, p.34, 2013.

HAYES, B.J.; BOWMAN, P.J.; CHAMBERLAIN, A.J.; GODDARD, M.E. Invited review: Genomic selection in dairy cattle: progress and challenges. **Journal of Dairy Scices.** v.92, n.2, p.433-43, 2009.

HENDERSON, C.R. Best linear unbiased estimation and prediction under a selection model. **Biometrics.** v.31, n.2, p423–447, 1975.

HENDERSON, C.R. Estimation of variance and covariance componentes. **Biometrics**. v.17, p.226-52, 1953.

KHATTREE, R.; NAIK, D.N. **Multivariate data reduction and discrimination with SAS Software.** Cary, NC, SAS institute Inc., 2000. 338p.

LANDER, E.S.; D. BOTSTEIN, 1989. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. **Genetics,** v.121, p. 185–199, 1989.

LIU G.; ZHAO, Y.; GOWDA, M.; LONGIN, C.F.H.; REIF, J.C.; METTE M.F. Predicting hybrid performances for quality traits through genomic-assisted approaches in Central European wheat. **PLoS ONE**. v.11, n.7, e0158635, 2016.

LUSH, J. L. **Animal Breeding Plans**. Ames, Iowa: Iowa State Press, 1937.

MALOSETTI, M.; Van EEUWIJK, F.A.; BOER, M.P.; CASAS, A.M.; ELÍA, M.; MORALEJO, M.; BHAT, P.R.; RAMSAY, L.; MOLINA-CANO, J.L. Gene and QTL detection in a three-way barley cross under selection by a mixed model with kinship information using SNPs. **Theoretical and Applied Genetics.** v.122, p.1605-1616. 2011.

MARTINS, E.N. **Desenvolvimento de uma estratégia computacional para a seleção de coelhos usando a melhor predição linear não-viesada.** 117f. Tese D.S. Universidade Federal de Viçosa, Viçosa, 1995.

MEUWISSEN, T.H.; HAYES, B.J.; GODDARD M.E. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**. v.157, n.4, p.1819-29, 2001.

MORA, F.; CASTILLO, D.; LADO, B.; MATUS, I.; POLAND, L;. BELZILE, F.; Von ZITZEWITZ, J.; Del POZO, A. Genome-wide association mapping of agronomic traits and carbon isotope discrimination in a worldwide germplasm collection of spring wheat using SNP markers. **Molecular Breeding**. v.35, p.69, 2015.

OLIVEIRA, M.B. **Análise de desequilíbrio de ligação e diversidade genética em soja utilizando marcadores moleculares**. 56 f. Tese D.S. Universidade Estadual de Maringá, Maringá, 2014.

PIEPHO, H.P.; MÖHRING. J. Selection in Cultivar Trials—Is It Ignorable?. **Crop** Sciences. v.46, p.192-201, 2006.

PRADO, S.A.; LÓPEZ, C.G.; SENIOR, M.L.; BORRÁS, L. The genetic architecture of maize (Zea mays L.) kernel weight determination. **G3**. v.4, n.9, p.1611-1621, 2014.

RESENDE, M.D.V.; RESENDE Jr, M.F.R.; SANSALONI, C.P.; PETROLI, C.D.; MISSIAGGIA, A.A.; AGUIAR, A.M.; ABAD, J.M.; TAKAHASHI, E.K.; ROSADO, A.M.; FARIA, D.A.; PAPPAS Jr, G.J.; KILIAN, A.; GRATTAPAGLIA, D. Genomic selection for growth and wood quality in Eucalyptus: capturing the missing heritability and accelerating breeding for complex traits in forest trees. **New Phytologist**. v.194, p.116–128, 2012b.

RESENDE, M.D.V.; SILVA, F.F.; LOPES P.S.; AZEVEDO, C.F. **Seleção Genômica Ampla (GWS) via Modelos Mistos (REML/BLUP), Inferência Bayesiana (MCMC), Regressão Aleatória Multivariada (RRM) e Estatística Espacial.** Viçosa, Universidade Federal de Viçosa/Departamento de Estatística. 2012a. 291 p.

SCHUSTER, I.; CRUZ, C.D. **Estatística genômica aplicada a populações derivadas de cruzamentos controlados.** Viçosa, Ed. UFV. 2004. 568p.

SEARLE R.S. An overview of variance component estimation. **Metrika**. v.42, p.215–230, 1995.

SILVA, M.F.; SCHUSTER, I.; SILVA, J.F.V.; FERREIRA, A.; BARROS, E.G.; MOREIRA, M.A. Validation of microsatellite markers for assisted selection of soybean resistance to cyst nematode races 3 and 14. **Pesquisa agropecuária brasileira**. v.42, n.8, p.1143-1150, 2007.

SORKHEH, K.; MALYSHEVA-OTTO, L.V.; WIRTHENSOHN, M.G.; TARKESH-ESFAHANI, S.; MARTÍNEZ-GÓMEZ, P. Linkage disequilibrium, genetic association

mapping and gene localization in crop plants. **Genetics and Molecular Biology**. v.31, n.4, p.805-814, 2008.

TORRES, F.E.; TEODORO, P.E.; SAGRILO, E.; CECCON, G.; CORREA A.M. Interação genótipo x ambiente em genótipos de feijão-caupi semiprostrado via modelos mistos. **Bragantia.**, v. 74, n. 3, p.255-260, 2015.

XU, S. Estimating polygenic effects using makers of the entire genome. **Genetics**. v.163, p.789–891, 1989.

ZAPATA-VALENZUELA, J.; HASBUN R.Z. Mejoramiento genético forestal acelerado mediante selección genómica. **Bosque.** v.32, n.3, p209-213, 2011.

ZENG, Z.B. Precision mapping of quantitative trait loci. **Genetics,** v. 136, p.1457-1468, 1994.

ZHANG, Z.; ERSOZ, E.; LAI, C.Q.; TODHUNTER, R.J.; TIWARI, H.K.; GORE, M.A.; BRADBURY, P.J.; YU, J.; ARNETT, D.K.; ORDOVAS, J.M.; BUCKLER, E.S. Mixed linear model approach adapted for genome-wide association studies. **Nature Genetics.** v.42, n.4, p.355–360, 2010.

Atualidades da Biometria no Melhoramento de Plantas Perenes

Marcos Deon Vilela de Resende[1]; Fabyano Fonseca e Silva[2]; Camila Ferreira Azevedo[3]; Itamara Bomfim Gois[4].

Introdução

O melhoramento genético de plantas perenes é fortemente calcado no uso da Biometria. As atividades e temas essenciais em um programa dessas espécies são apresentados a seguir: definição do objetivo (caráter objetivo ou função objetivo) do melhoramento; definição dos critérios de seleção (caracteres usados na seleção); estudo do controle genético dos caracteres (herdabilidades, repetibilidade, correlações genéticas e grau de dominância); escolha do germoplasma (populações, espécies, genitores, clones); experimentação (escolha do delineamento experimental, número de repetições, número de plantas por parcela, avaliação de covariáveis para controle estatístico); escolha dos métodos de seleção (individual, família, dentro de família, otimizado); escolha dos delineamentos de cruzamento (polinização aberta, policruzamentos, pares simples, dialélicos, fatoriais); escolha de estratégias de melhoramento (seleção recorrente intrapopulacional ou seleção recorrente recíproca); aumento da velocidade do melhoramento (seleção precoce, seleção genômica); conhecimento do sistema reprodutivo e escolha do sistema de propagação para plantios; escolha do tamanho (Ne) e estrutura de populações para melhoramento de curto e longo prazo.

A Biometria desempenha papel fundamental em todas essas atividades. De especial importância é a estimação ou predição dos valores genéticos individuais com alta precisão e acurácia. Isto norteia todas as demais atividades. Assim, a modelagem genética e estatística das principais características de interesse do melhoramento merece um tratamento mais detalhado conforme realizado a seguir. Os problemas genéticos e ferramentas estatísticas para solução são: (i) predição de observações futuras: ***predição*** da realização de variáveis aleatórias; (ii) controle genético dos caracteres (herdabilidades, repetibilidade e correlações) e detecção de genes: ***inferência*** estatística; (iii) alocação de

[1]Estatístico e Engenheiro Agrônomo, M.S., D.S. e Pesquisador da Embrapa Floresta. E-mail: marcos.deon@gmail.com
[2]Zootecnista, M.S., D.S. e Professor da Universidade Federal de Viçosa. E-mail: fabyanofonseca@ufv.br
[3]Matemática, M.S., D.S. e Professora da Universidade Federal de Viçosa. E-mail: camila.azevedo@ufv.br
[4]Engenheira Florestal, M.S., e D.S. E-mail: itamarafloresta@gmail.com

cruzamentos: teoria da ***decisão*** estatística, obtida deterministicamente via programação linear; (iv) análise da expressão gênica: ***reconhecimento de padrões***, obtida via análises multivariadas de agrupamento.

Mas é importante relatar que o melhoramento em si, não é só isso. Ramalho et al. (2012) definem o melhoramento genético como arte, ciência, capacidade gerencial e aspecto comercial.

Modelos estatísticos e métodos de estimação

Atualmente existem três abordagens estatísticas principais que são empregadas na estimação e inferência estatística, em termos de método de estimação: quadrados mínimos (LS) ou método generalizado dos momentos; máxima verossimilhança residual (REML) e estimação Bayesiana (BAYES).

Genericamente, as variáveis associadas às características genéticas são classificadas como fixas ou aleatórias. Variáveis fixas são denotadas "parâmetros" e nenhuma suposição é realizada sobre suas distribuições. As variáveis aleatórias são assumidas como amostradas de uma distribuição de probabilidade com parâmetros conhecidos. Estimativas de parâmetros pelos métodos de máxima verossimilhança e bayesiano devem situar-se no espaço paramétrico, pois se fora desse espaço tais estimativas tem verossimilhança zero. Isto não é garantido pelos estimadores de quadrados mínimos. Para variáveis contínuas, a verossimilhança é computada como a densidade estatística da distribuição condicional à amostra. A densidade estatística f(y), para uma variável contínua y, é definida como a ordenada da função distribuição para um dado valor de y.

O método LS não permite considerar a correlação entre níveis ou efeitos pertencentes aos vários fatores, por exemplo, não considera a correlação entre níveis dos efeitos do fator tratamentos. Por outro lado, o REML e o BAYES permitem considerar essas correlações. Em termos do fator tratamentos, quando o mesmo tem uma conotação genética (comparação de indivíduos, por exemplo), a matriz de correlação entre os níveis dos efeitos do fator pode ser não correlacionada (diagonal D, podendo ser uma identidade I no caso de efeitos aleatórios ou, uma matriz nula no caso de efeitos fixos) ou correlacionada dada por dois tipos de informação: genealógica (matriz de correlação A) e genômica (matriz de correlação G). Em termos do melhoramento de animais e plantas, a seleção genética pode ser realizada via: (i) seleção fenômica (valores genéticos preditos com base em genealogia e fenótipos); (ii) seleção genômica (valores genéticos preditos

com base em genótipos marcadores e fenótipos); (iii) seleção geno-fenômica (valores genéticos preditos com base em genótipos marcadores, fenótipos e genealogia).

Adicionalmente, uma estrutura multivariada ou dela derivada (longitudinal, espacial, curvilínea) pode ser imposta tanto aos fatores de tratamentos ($I \otimes H$; $A \otimes H$; $G \otimes H$), quanto a outros efeitos aleatórios como os resíduos ($I \otimes H$), em que H é uma matriz que descreve a estrutura de correlação.

A combinação desses três métodos de estimação com os três tipos de matriz de correlação (assumindo distribuição normal multivariada das observações individuais) fornece os sete tipos gerais de abordagem estatística utilizados na análise genética, conforme o Quadro 1. E dentro dessas abordagens podem ser ajustados os seguintes modelos: Univariados, Multivariados, Longitudinais, Espaciais, Curvilíneos, de Competição, Categóricos, de Sobrevivência.

Quadro 1. Modelos Estatísticos e Métodos de Estimação.

Matriz de Correlação/ Método de Estimação	**Correlação D**	**Correlação A**	**Correlação G**	**Modelos**
LS	D-LS	-	-	Univariados, Multivariados
REML	D-REML	A-REML	G-REML	Univariados, Multivariados, Longitudinais, Espaciais, Curvilineos, Competição, Categóricos, Generalizados, Sobrevivência
BAYES	D-BAYES	A-BAYES	G-BAYES	Univariados, Multivariados, Longitudinais, Espaciais, Curvilineos, Competição, Categóricos, Generalizados, Sobrevivência

De maneira geral, uma análise estatística completa envolve as seguintes seis atividades: a estimação de componentes de média, a estimação de componentes de variância, a realização de testes de hipóteses, a inferência sobre a acurácia (confiabilidade), viés, e precisão da estimação/predição. Um resumo é apresentado no Quadro 2. Por exemplo, no contexto dos modelos mistos, a realização dessas atividades envolve a predição BLUP, a estimação REML, a análise de deviance, o cômputo da acurácia de predição e da variância do erro de predição, respectivamente. No

procedimento REML/BLUP o viés é assumido como nulo, pois estes estimadores/ preditores pertencem à classe dos melhores estimadores/preditores lineares não viesados (BLUE/BLUP).

Quadro 2. Métodos e Medidas de Ajuste de Modelos.

Método de Estimação	Componentes de Médias	Componentes de Variância	Testes de Hipótese	Precisão	Viés	Confia-bilidade	Acurá-cia
LS	GLS	EQM	ANOVA-F	SED	Nulo	r^2	$(1-1/F)^{1/2}$
REML	BLUE, BLUP	REML	Deviance-LRT	SEP	Nulo	r^2_{gg}	r_{gg}
BAYES	MAP	MAP	HPD	STD	-	-	1-1/t

MAP: máximo a posteriori; EQM: esperança de quadrados médios; HPD: highest posterior density; SED: erro padrão da diferença entre tratamentos; SEP= raiz(PEV); STD: desvio padrão da estimativa; F: estatística F de Snedecor; t: estatística t de Student.

Nos artigos científicos, os resultados gerados pelas seis atividades de uma análise estatística completa devem ser discutidos. Na área de genética o seguinte roteiro pode ser seguido:

Testes de Hipóteses: inferência sobre a significância da variabilidade genética (Vg).
Componentes de Variância e suas proporções: inferência sobre o controle genético (alto, moderado e baixo) ou herdabilidades e correlações entre caracteres, coeficientes de variação.
Componentes de Médias: valores genéticos e ganhos genéticos.
Precisão: PEV (variância do erro de predição); relação PEV/Vg (com espaço paramétrico entre 0 e 1).
Acurácia: com espaço paramétrico entre 0 e 1 e classificação conforme Resende e Duarte (2007).
Viés: dado pela regressão de y em $\hat{y}$ ($\beta(y/\hat{y})$) em que $\beta=1$ é o ideal e indica não viés.

Modelos de avaliação genética

Modelos gerais de avaliação genética com aplicação ao ***melhoramento de plantas*** são apresentados a seguir. Os procedimentos aplicáveis são similares dentro de grupos, de acordo com a seguinte classificação (Resende et al., 2014).

(a) **Plantas perenes alógamas** (eucalipto, pinus, café canéfora, cana-de açúcar, fruteiras, forrageiras): *estimação baseada simultaneamente na herdabilidade, repetibilidade, estrutura de covariância e curvas de crescimento; genealogia variável.*

(b) **Plantas perenes autógamas** (café arábica, pêssego, damasco, coco anão, limão, leucena): *estimação baseada simultaneamente na herdabilidade, repetibilidade, estrutura de covariância e curvas de crescimento; genealogia fixa a partir de cada F2 até F6.*

(c) **Plantas perenes de reprodução assexuada** (cana-de açúcar, seringueira, laranjeira, braquiária, panicum): *estimação baseada simultaneamente na herdabilidade no sentido amplo, repetibilidade, estrutura de covariância e curvas de crescimento; genealogia constante através das fases clonais.*

(d) **Plantas anuais alógamas** (milho, pipoca, girassol, brócolis, cenoura): *estimação baseada na herdabilidade; genealogia variável.*

(e) **Plantas anuais autógamas** (feijão, soja, arroz, trigo, aveia): *estimação baseada na herdabilidade; genealogia fixa a partir de cada F2 até F6.*

(f) **Plantas de ciclos anuais e de reprodução assexuada** (mandioca, batata): *estimação baseada na herdabilidade no sentido amplo; genealogia constante através das fases clonais.*

Uma classificação aproximada dos grupos de plantas quanto aos modelos de avaliação genética em função da estrutura experimental de campo e modelos de avaliação genética é apresentada a seguir:

a) **Culturas Anuais e Olerícolas (Arroz, Milho, Soja, Feijão, Trigo, Aveia, Cevada, Sorgo, Algodão, Batata, Mandioca):** dados analisados com uma observação por parcela e inferência ao nível do efeito de tratamentos genéticos (linhagens, híbridos, clones, famílias, cultivares, acessos).

b) **Forrageiras e Cana-de-Açúcar:** dados analisados com uma observação por

parcela e medidas repetidas, inferência ao nível do efeito de tratamentos genéticos (linhagens, híbridos, clones, famílias, cultivares, acessos).

c) **Espécies Florestais:** dados tomados ao nível de plantas individuais sem medidas repetidas, inferência ao nível de Indivíduos, genitores e clones.

d) **Fruteiras, Palmáceas (Açaí, Coco, Dendê, Pupunha, Tâmara), Estimulantes (Café, Cacau, Guaraná, Erva-mate), Seringueira:** dados tomados ao nível de plantas individuais com medidas repetidas, inferência ao nível de Indivíduos, genitores e clones.

A complexidade dos modelos, a dificuldade na avaliação genética e o grau de desbalanceamento aumentam de (a) para (d), ou seja, a complexidade aumenta na seguinte seqüência: Culturas Anuais e Olerícolas; Forrageiras e Cana-de-Açúcar; Espécies Florestais; Frutíferas. Verifica-se que as culturas perenes são as que mais demandam do Biometrista, pois, geralmente requerem abordagens multivariadas em função das medidas repetidas no tempo.

No Quadro 3 são apresentadas as principais modelagens aplicadas a Dados Multivariados.

Quadro 3. Modelagens Aplicadas a Dados Multivariados

Dados Multivariados	Modelagem
Múltiplos Caracteres	Multivariada Não Estruturada; Componentes Principais; Fatores Latentes
Múltiplos Ambientes	Componentes Principais Centrados (AMMI); Fatores Latentes Analíticos Mistos (FAMM); Regressão Aleatória via Normas de Reação (Curvilíneo)
Medidas Repetidas	Simetria Composta; Autoregressivos; Curvilíneos (Regressão Aleatória via Polinômios de Legendre, via Polinômios Segmentados (Spline) do tipo B; via Séries de Fourier)
Dados Espaciais	Modelos Autoregressivos Separáveis AR1 x AR1; Modelo Geoestatístico Exponencial; Modelo Geoestatístico Esférico
Séries Temporais	Modelos de Médias Móveis; Modelos Autoregressivos; Modelos ARMA; Modelos ARIMA; Filtro de Kalman (BLUP), Séries de Fourier

Tipos de Cultivares e o Alvo do Melhoramento

No melhoramento de plantas, os principais tipos de cultivares que podem ser gerados são apresentados no quadro a seguir.

Quadro 4. Tipos de cultivares de acordo com as categorias de espécies

Tipos de Cultivares	Sistema Reprodutivo das Espécies	Exemplos
Clones	Alógamas, Autógamas, Assexuadas	Perenes Alógamas (cana, eucalipto, guaraná, seringueira, cacau, café canéfora, frutíferas, forrageiras, etc); Perene Autógama (pessegueiro); Perenes de Ciclo Anual (mandioca, batata).
Linhagens	Autógamas	Autógamas Anuais (arroz, trigo, aveia, feijão, soja, alface, pimenta); Perene Autógama (café arábica).
Híbridos Simples de Linhagens	Alógamas, Autógamas	Alógamas Anuais (milho, girassol, cebola, cenoura, brássicas, abóboras); Autógamas Anuais (tomate, pimentão, fumo).

Os três tipos de cultivares visam capturar os genótipos com as duas características mais importantes para o sistema produtivo (a produtividade sustentável e a homogeneidade dos produtos), dadas as condicionantes biológicas (sistema reprodutivo e de propagação) de cada espécie. Dessa forma, as estratégias de melhoramento (as quais geram os dados a serem analisados estatisticamente) são similares para as espécies dentro de cada um desses três grupos. Clones e híbridos de linhagens capturam tanto os efeitos genéticos aditivos quanto os de dominância (que, sob divergência genética, propiciam heterose). Linhagens capturam apenas os efeitos genéticos aditivos. Os programas de melhoramento visando linhagens e híbridos de linhagens são similares (hibridação entre linhagens, seguida por condução de gerações endogâmicas até a seleção de novas linhagens). No entanto, no caso quando o alvo são híbridos, o processo não para aí. Prossegue-se com o cruzamento entre as linhagens visando à identificação dos híbridos superiores. Com o advento dos duplo-haplóides a etapa de condução de gerações endogâmicas está sendo suprimida e as linhagens obtidas (até então sem nenhum teste de campo) tem os seus cruzamentos preditos como o auxílio da fenômica e da genômica.

Seleção genômica

A seleção genômica ampla (GWS) ou seleção genômica (GS) foi proposta em 2001 como uma forma de aumentar a eficiência e acelerar o melhoramento genético. A GWS enfatiza a predição simultânea (sem o uso de testes de significância para marcas individuais) dos efeitos genéticos de milhares de marcadores genéticos de DNA (SNP, DArT, Microssatélites) dispersos em todo o genoma de um organismo, de forma a capturar os efeitos de todos os locos (tanto de pequenos quanto de grandes efeitos) e explicar toda a variação genética de um caráter quantitativo.

Um método ideal para GWS deve contemplar três atributos: (i) **acomodar a arquitetura genética** do caráter em termos de genes de pequenos e grandes efeitos e suas distribuições; (ii) realizar a **regularização** do processo de estimação em presença de multicolinearidade e grande número de marcadores, usando para isso estimadores do tipo *shrinkage*; (iii) realizar a **seleção de covariáveis** (marcadores) que afetam a característica em análise. O problema principal da GWS é a estimação de um grande número de efeitos a partir de um limitado número de observações e também as colinearidades advindas do desequilíbrio de ligação entre os marcadores. Os estimadores do tipo *shrinkage* lidam adequadamente com isso, tratando os efeitos de marcadores como variáveis aleatórias e estimando-os simultaneamente (Resende, 2007; Resende et al., 2008). Os principais métodos para a GWS podem ser divididos em três grandes classes: regressão explícita, implícita e com redução dimensional. Na primeira classe, destacam-se os métodos RR-BLUP, LASSO (*Least Absolute Shrinkage and Selection Operator*), Rede Elástica (*Elastic Net* – EN), BayesA e BayesB, dentre outros. Na classe de regressão implícita, citam-se os métodos de redes neurais, RKHS (*Reproducing Kernel Hilbert Spaces*, que é um método semi-paramétricos e regressão kernel não paramétrica via modelos aditivos generalizados. Dentre os métodos de regressão com redução dimensional, destacam-se o de componentes independentes, quadrados mínimos parciais e de componentes principais.

Os métodos de regressão explícita são divididos em dois grupos: (i) métodos de estimação penalizada (RR-BLUP, LASSO, EN, RR-BLUP-Het); (ii) métodos de estimação bayesiana (BayesA, BayesB, Fast BayesB, BayesCπ, BayesDπ, BayesR, BayesRS, BLASSO, IBLASSO e outros).

Se os efeitos de marcadores são tomados como fixos, não é possível considerar as covariâncias entre os efeitos de marcadores. Com alta densidade de marcadores, mais de um marcador estará em desequilíbrio de ligação com um QTL segregante. Isto resultará

em covariância entre efeitos de marcadores. A maioria dos marcadores não terão efeito algum sobre um caráter. Assim, os efeitos estimados desses marcadores vazios serão falsos. Este problema é maior no caso em que os marcadores são considerados de efeitos fixos, pois nesse caso, esses pseudos efeitos não serão regressados em direção a zero.

Um refinamento da seleção genômica pode ser conseguido pelo uso de QTNs em lugar de SNPs. A evolução da tecnologia genômica é previsível e a mutação causal de uma variação genética em nível de nucleotídeo (QTN) poderá ser acessada em um futuro próximo. Assim, a seleção genômica poderá ser aperfeiçoada pelo uso direto dos QTNs em lugar dos SNPs. O uso dos QTNs trará as seguintes vantagens: (i) a GWS não dependerá do desequilíbrio de ligação pois o QTN será acessado diretamente e não via marcadores; isto aumentará a durabilidade da predição genômica, a qual será útil também a longo prazo; (ii) a predição genômica poderá ter validade (transferibilidade) através de diferentes populações e espécies em um mesmo gênero; (iii) a predição genômica usará QTNs específicos para cada caráter, ao contrário do G-BLUP via SNPs, o qual usa a mesma matriz de parentesco G para todas as características; (iv) os índices de seleção multicaracterísticos ponderarão diretamente os QTNs e não as características fenotípicas; (v) a seleção genômica poderá usar um menor número de gerações (apenas as últimas) para a composição da matriz G, isto trará maior ganho genético e menor massa de dados a serem processados; (vi) as frequências alélicas dos QTNs serão acessadas diretamente e não via desequilíbrio de ligação com SNPs.

Modelos Hierárquicos Generalizados e Quase-Verossimilhança Estendida: Unificação da Análise Estatística em Biometria e Genética

Tradicionalmente, três classes de abordagens são utilizadas para Inferência: Frequentista, Fisheriana (Verossimilhança) e Bayesiana. Segundo essas abordagens, os intervalos de confiança para variáveis aleatórias não observáveis podem ser: Fiducial-Fisheriano (intervalo fixo, para incógnita fixa), Frequentista (intervalo aleatório, para incógnita fixa), Bayesiano (intervalo fixo, para incógnita aleatória). A Verossimilhança-H (hierárquica) surgiu como uma forma de unificação das três classes de abordagens. Essa idéia será discorrida a seguir.

Os modelos lineares mistos para variáveis com distribuição normal foram desenvolvidos por Henderson e são implementados via predição BLUP e estimação de componentes de variância por REML desenvolvido por Patterson e Thompson. Os modelos lineares generalizados (GLM) foram desenvolvidos por Nelder e Wedderburner

em 1972 para lidar com variáveis descontínuas. Em 1976, Lee e Nelder estenderam a abordagem BLUP para uma classe ampla de modelos estatísticos com efeitos aleatórios, denominada modelos lineares generalizados hierárquicos (HGLM).

Nos GLM assume-se que os resíduos podem não apresentar distribuição normal, mas, os demais efeitos aleatórios do modelo seguem a distribuição normal. Entretanto, essa suposição nem sempre é adequada. Um exemplo é a situação em que os dados seguem distribuição de Poisson e a função de ligação especificada para os resíduos é a logarítmica. Nesse caso, uma suposição mais apropriada para os demais fatores aleatórios é uma distribuição gama com função de ligação logarítmica. Modelos em que uma distribuição de probabilidade e uma função de ligação podem ser especificadas para cada fator aleatório do modelo pertencem à classe dos HGLM. Como os fatores aleatórios nem sempre são de classificação hierárquica, uma denominação alternativa para os HGLM é modelos lineares mistos generalizados estratificados. Um preditor BLUP para HGLM's foi apresentado por Lee e Ha em 2010. Para HGLM's não normais o BLUP linear pode não ser eficiente. Os autores apresentaram uma combinação do BLUP com modelos Tweedie de dispersão baseados em distribuição exponencial.

Gianola e Fernando, em 1986, propuseram a estimação Bayesiana para modelos de avaliação genética. Além das distribuições (normais) adotadas para os efeitos aleatórios (g) no modelo linear clássico e para a verossimilhança do vetor de observações (y), a abordagem bayesiana requer atribuições para as distribuições a priori dos efeitos fixos e componentes de variância. A atribuição de distribuições a priori não informativas ou uniformes para os efeitos fixos e componentes de variância é uma forma de caracterizar um conhecimento a priori vago sobre os referidos efeitos e componentes. Assim, a estimação dos efeitos fixos e de efeitos aleatórios do modelo frequentista, por meio da abordagem Bayesiana, pode ser realizada desde que sejam atribuídas prioris não informativas para os efeitos fixos, prioris normais para os efeitos aleatórios e verossimilhança normal para o vetor de observações. Utilizando-se distribuições a priori não informativas para os efeitos fixos e componentes de variância, as modas das distribuições marginais a posteriori dos componentes de variância correspondem às estimativas obtidas por REML.

Por outro lado, a estimação frequentista de modelos Bayesianos pode ser realizada via HGLM, com vantagens computacionais (menor tempo; critério trivial de convergência). Os HGLM podem ser ajustados usando suas verossimilhanças hierárquicas (h-L), que é uma extensão da verossimilhança conjunta usada por Henderson

e consiste de uma densidade conjunta para as observações e efeitos aleatórios. As estimativas de efeitos fixos e aleatórios são derivadas da maximização da h-L e produzem extensões diretas das equações de modelos mistos de Henderson. Os componentes de variância são estimados pela maximização do perfil da h-L ajustada, que é uma extensão direta do REML. Dessa forma, os HGLM estendem a teoria familiar do BLUP empregado no melhoramento animal para uma classe mais ampla de modelos.

Explorando a natureza hierárquica da h-L, modelos para os componentes de variância dos parâmetros de dispersão podem ser adicionados um por um. Uma classe ampla de distribuições pode ser usada para modelar tanto a variável resposta quanto os efeitos aleatórios, fato que aumenta a flexibilidade da modelagem. Exemplos são: o caso da variável resposta com distribuição de Poisson, com distribuição gama para os efeitos aleatórios; os modelos de fragilidade para análise de sobrevivência, que permitem lidar com heterogeneidade pela inclusão de efeitos aleatórios para variância residual e modelos para alisamento dos dados usando efeitos aleatórios. A h-L pode ser usada também na derivação de ferramentas para seleção de modelos. O AIC condicional da h-L é na verdade equivalente ao Critério de Informação da Deviance (DIC) aplicada na estatística bayesiana. Existe um pacote HGLM no R que pode ser aplicado aos modelos usados no melhoramento animal. Também o pacote bigRR usa o pacote HGLM para a predição genômica. Uma aplicação interessante é ajustar a MAF (frequência do alelo mais raro) como covariável (o ponto de corte para inclusão de uma marca na análise pode ser dado por $MAF = 1 / (2N)^{1/2}$; isto advém do desvio padrão de uma proporção, dado por $(pq)^{1/2} / (2N)^{1/2}$, em que N é o número de indivíduos genotipados, significando que quanto menor N maior deve ser a MAF adotada para que o efeito da marca seja estimado com precisão).

Métodos bayesianos para realizar a seleção de covariáveis na predição genômica e detecção de QTL foram desenvolvidos. Entretanto é também possível usar a abordagem frequentista HGLM com vantagens computacionais. Nessa abordagem um modelo de mistura gama escalada, para distribuição dos efeitos pode ser usado e propicia uma família inteira de verossimilhanças penalizadas, incluindo a abordagem LASSO. Essa família de modelo pode ser ajustada usando algoritmo de quadrados mínimos ponderados iterativos. Múltiplos testes podem ser vistos como um problema de predição múltipla associada à hipótese nula verdadeira ou falsa e estas predições podem ser feitas pelo ajuste de efeitos aleatórios. Assim, a estimação de efeitos aleatórios sobre várias suposições de distribuição é de grande interesse em várias áreas e pode ser feita via HGLM (Resende, 2015).

Genética Quântica: Física Estatística na Genética de Populações e Quantitativa

As palavras ***Quântica*** e ***Quantitativa*** estão associadas à quantidade e quantificação. A ***Mecânica Estatística*** é uma área da ***Física Estatística*** ou ***Física Quântica*** que apresenta importantes contribuições para a Estatística e Biometria. Como exemplo podem ser citados os conceitos de ***Entropia***, ***Negentropia***, ***Espaço de Hilbert***, ***Informação de Kullback-Leibler***, ***Teorema H de Boltzmann*** e ***Distribuição de Gibbs*** para o algoritmo de MCMC (Amostragem para Simulação Estocástica). Josiah Willard Gibbs (1839-1903) foi o primeiro grande físico teórico dos Estados Unidos e cunhou o termo "Mecânica Estatística".

Em Física, a **entropia** é uma grandeza termodinâmica que mede o grau de irreversibilidade de um sistema, geralmente associada à denominação de desordem de um sistema termodinâmico. É medida na unidade [J/K] (joules por kelvin). De acordo com a segunda lei da termodinâmica, trabalho pode ser completamente convertido em calor ou energia, mas energia térmica não pode ser completamente convertida em trabalho. Com a entropia procura-se mensurar a parcela de energia que não pode mais ser transformada em trabalho em transformações termodinâmicas à determinada temperatura. A parcela de energia interna de um sistema em seu equilíbrio termodinâmico que não pode mais ser convertida em trabalho à temperatura de equilíbrio, pode ser determinada pelo produto da entropia S pela temperatura absoluta T do sistema no respectivo estado.

Em mecânica estatística clássica, o teorema-H, introduzido por Ludwig Boltzmann em 1872, descreve a tendência de aumento na quantidade H, a qual representa a entropia da termodinâmica. O teorema-H está associado à segunda lei da termodinâmica, fundamentada nos processos irreversíveis. A medida de informação entrópica S_H propicia uma solução exata sob equilíbrio estatístico. A distribuição que maximiza a entropia, sujeita à restrições, é a distribuição de Boltzmann.

A entropia refere-se à medida de incerteza ou imprevisibilidade de uma variável aleatória ou de um sistema físico. É um conceito usado na ***Teoria de Informação*** e quantifica o valor esperado de informação, sendo, portanto, equivalente ao ***conteúdo de informação***. Para uma variável com Distribuição Normal a entropia é dada por $\frac{1}{2}\ln(2\pi e\sigma^2)$. O termo ***negentropia*** refere-se à ordem ou entropia negativa. O princípio da máxima entropia tem sido usado para derivar as expressões matemáticas da genética de

populações dos caracteres quantitativos sob equilíbrio, mutação, seleção, migração (fluxo gênico) e deriva genética.

Em um modelo matemático geral para o equilíbrio genético em populações, baseado no princípio da máxima entropia, foi provado que a distribuição de probabilidade da máxima entropia foi equivalente à Lei do Equilíbrio de Hardy-Weinberg. Mostrou-se também que a população atinge o equilíbrio genético quando a entropia dos genótipos na população atinge o máximo valor possível. Isto indica que cruzamentos aleatórios na população é um processo irreversível, o qual aumenta a entropia genotípica da população, ao passo que a endogamia e seleção a diminui. No melhoramento genético a seleção e/ou endogamia são usadas para diminuir a entropia da população e o cruzamento para aumenta-la. Assim, o melhoramento atua por meio da regulação da entropia da população.

Em 2012, pela primeira vez, demonstrações teóricas detalhadas foram realizadas usando ***Teoria da Informação*** via entropias das distribuições de frequências alélicas para caracterizar mutação, seleção, deriva genética e fluxo gênico. Também o desequilíbrio de ligação (r^2) foi abordado e estabelecidas as relações entre r^2 e I, sendo I a ***Informação Mútua***.

A compreensão da evolução das frequências alélicas de uma população pode ser interpretada como um processo biológico cujos mecanismos são exatamente correspondentes às medidas teóricas de informação e as técnicas da teoria de informação podem agregar e simplificar a análise teórica de alguns processos evolutivos (e de melhoramento via seleção) experimentados pelas populações. As conexões entre genética de populações e teoria de informação são feitas por meio do uso de várias quantidades baseadas em informação, as quais serão definidas a seguir.

Teoria da Informação e da Comunicação

A primeira e mais famosa teoria de informação é o conceito de entropia de Shannon. Para uma distribuição de uma variável aleatória com n diferentes estados com probabilidade P(i) em que $\sum_{i=1}^{n} P(i) = 1$, a entropia S é definida como: $S = -\sum_{i=1}^{n} P(i) \log P(i)$. O valor de S sempre situa entre o mínimo de zero para a distribuição trivial em que um evento ocorre com probabilidade 1 e o máximo de $S = \log n$, para a distribuição uniforme através dos n estados. Essa simples definição de entropia advém do fato de que a entropia tem várias ordens dependendo dos graus de liberdade definidos na distribuição. Shannon foi o precursor da comunicação digital. Ele estabeleceu os fundamentos matemáticos da

Teoria de Comunicação, criando uma definição precisa para o vago conceito de informação.

Três quantidades de informação úteis na genética quantitativa de populações

O método estatístico de Máxima Entropia (MAXENT) é uma alternativa aos métodos de Máxima Verossimilhança e Bayesianos. Uma comparação entre os principais métodos de análise espectral é apresentada no quadro a seguir.

Método	**Distribuição conhecida para os resíduos**	**Informação *a priori* para as quantidades sendo estimadas**
Máxima verossimilhança	Sim	Não
Máxima entropia	Não	Sim
Bayesiano	Sim	Sim
Auto-regressivos	Caso especial de máxima entropia	Caso especial de máxima entropia

A escolha do método depende de em qual classe de problemas pretende-se usar o método. A estimativa espectral de máxima entropia é, para um particular tipo de dados, idêntica na forma analítica com aquele de modelo auto-regressivo.

Existem três quantidades de informação úteis na genética quantitativa de populações: (i) Entropia (S); (ii) Divergência de Kullback-Leibler (D); (iii) Informação Mútua (I). As relações entre essas quantidades e conceitos de genética de populações são apresentados na Quadro 5.

Quadro 5. Analogias entre quantidades da Teoria da Informação e conceitos de Genética de Populações.

Medida	**Equação**	**Significância**
Entropia S_0	$S_0 = \log n$	Máxima entropia de um modelo com n alelos.
Entropia S_1	$S_1 = -\sum_{i=1}^{n} P(i) \log P(i)$	Entropia de frequências alélicas; essencial na definição da deriva genética em função do tempo.
Entropia S_2	$S_2 = -\sum_{i=1}^{n} \sum_{j=1}^{n} P(i,j) \log P(i,j)$	Entropia do loco baseada em frequências genotípicas (par de alelos); essencial na medida de mudança genotípica.
Divergência D	$D(f,g) = \sum_{i=1}^{n} f(i) \log \frac{f(i)}{g(i)}$	Uso no modelo de deriva genética.
Informação Mútua I	$I = \sum_{i=1}^{n} \sum_{j=1}^{n} P(i,j) \log \frac{P(i,j)}{P(i)P(j)}$	Uso no modelo de seleção e cruzamentos não aleatórios ou preferenciais.

Entropia (S)

A Teoria da Informação foi desenvolvida por Shannon por volta de 1940 e propiciou uma revolução nas comunicações que culminou com a criação da internet. A visão revolucionária de entropia e informação permitiu a quantificação das taxas de transferência de informação. Mais tarde os artigos de Edwin Jaines mostraram que a mecânica estatística do equilíbrio pode ser derivada pelo uso da teoria da informação e suposições de máxima entropia. Recentemente, mostrou-se que existem grandes ligações entre a teoria de informação e a genética quantitativa de populações.

Divergência de Kullback-Leibler (D)

Além da medida de entropia existem duas outras quantidades úteis: (i) a divergência de Kullback-Leibler (D); (ii) informação mútua (I). D mede a diferença entre duas distribuições de probabilidade. Para duas distribuições, f e g, a D de f para g é definida por $D(f,g)=\sum_{i=1}^{n} f(i)\log\frac{f(i)}{g(i)}$, em que D ≥ 0. Entretanto, D não é uma distância métrica, pois, não é simétrica em relação às distâncias entre as distribuições e $D(f,g)\neq D(g,f)$. Outra forma de expressar D é $D(f,g)=S_X(f,g)-S(f)$, em que S_X é a entropia cruzada representada por $S_X=-\sum_{i=1}^{n} f(i)\log g(i)$. A distância D é essencial na discussão de deriva genética (Quadro 5).

A distância entre distribuições estatísticas pode ser medida pela métrica de Kullback-Leibler (KL). Em inferência bayesiana o grau de aprendizagem estatística pode ser quantificado pela KL e também pela distância de Hellinger entre as densidades *a priori* e *a posteriori* dos efeitos do modelo. Distribuições *a priori* com menores distâncias para *a posteriori* podem ser escolhidas, usando o conceito de máxima entropia.

Informação Mútua (I)

A Informação Mútua (I) entre duas variáveis aleatórias, i e j, é a representação da entropia de uma variável que pode ser derivada da entropia de outra variável em uma distribuição. Shannon foi o primeiro a usá-la. Estatisticamente, I é dado por: $I=\sum_{i=1}^{n}\sum_{j=1}^{n} P(i,j)\log\frac{P(i,j)}{P(i)P(j)}$. Outra forma de expressar I é via $I=S(i)+S(j)-S(i,j)=S(i)-S(i\mid j)=S(j)-S(j\mid i)$. A informação mútua pode ser usada para representar os efeitos da seleção e cruzamentos não aleatórios na população.

Informação mútua e Desequilíbrio de Ligação

Na literatura, novas medidas de desequilíbrio de ligação foram propostas usando entropia S, divergência D e I. O desequilíbrio r^2 pode ser derivado da informação mútua. Sob aproximação linear $I \approx r^2$ é uma medida em nível de dois locos. Para um modelo com dois locos e dois pares de alelos A, a, B, b, a entropia esperada de segunda ordem do sistema é representada como $S(A,B) = S_A + S_B - r^2$. A mudança na entropia devida à seleção pode ser aproximada como $\Delta S = I - I' = r^2 - I'$. Assim, uma vez que r^2 é sempre positivo, qualquer nível de desequilíbrio de ligação afeta a taxa de redução da diversidade genética, pela diminuição na mudança do decréscimo da entropia causada pela seleção.

Capacidade de canal e efeitos da seleção

Capacidade de canal é a taxa máxima na qual um sinal composto por símbolos, com uma probabilidade para cada símbolo, pode ser transmitido. Para uma população, um canal pode ser representado pelos efeitos combinados de cruzamento e adaptação dos descendentes. A reprodução sexual permite maior capacidade de canal do que a assexual. A capacidade de canal (C) de uma população diploide pode ser representada por: $C = 2Ne(-p\log p - q\log q + \mu\log\mu + (1-\mu)\log(1-\mu)$ ou $C = 2Ne(S - S_m)$, em que Ne é o tamanho efetivo e p e q são as frequências alélicas. Usando aproximações, esta equação pode ser representada por $C = 2Ne(S - 2\mu) = 2Ne(h - 2\mu)$. Assim, a capacidade de canal é dada como função da entropia das frequências alélicas menos duas vezes a taxa de mutação (μ). Dessa forma, a diversidade genética aumenta a capacidade de canal ao passo que a mutação a reduz. Sob Equilíbrio de Hardy-Weinberg, $C = 2NeS$. Sendo a heterozigose dada por $h = 4Ne\mu$, tem-se que $C = 2Ne(4Ne\mu - 2\mu)$. De forma aproximada $C = 8Ne^2\mu$. A seleção natural reduz a capacidade de canal (Resende, 2015).

Considerações finais

O presente texto abordou o estado da arte da Biometria aplicada ao melhoramento de plantas perenes. Modelos estatísticos e métodos de estimação aplicados nesta categoria de plantas são abordados, destacando-se os métodos de Máxima Verossimilhança Restrita, Bayesianos e de Máxima Entropia. Introduziu-se também o método de estimação denominado Verossimilhança Hierárquica. A Seleção Genômica foi descrita como uma ferramenta essencial no aumento da eficiência seletiva e da velocidade do melhoramento

via aumento da acurácia e ganho por unidade de tempo por meio da seleção precoce. Finalmente, foram relatadas possíveis contribuições advindas da Física Estatística em prol da Genética de Populações e Quantitativa.

Referências

JAYNES, E. T. Information theory and statistical mechanics. ***Phys. Rev.*, 106,** 620. 1957.

JAYNES, E. T. Where do we stand on Maximum Entropy. In: ***The Maximum Entropy Formalism***, R. D. Levine and M. Tribus (eds.), M. I. T. Press, Cambridge, MA, 1979, p. 15.

LEE, Y.; NELDER, J.A.; PAWITAN, Y. Generalized Linear Models with Random Effects. *Chapman e Hall/ CRC*. 2006.

LINDLEY, D.V.; SMITH. A.F.M. Bayes estimates for the linear model. **Journal of the Royal Statistical Society, Series B**, v. 34, p.1-41, 1972.

LYNCH, M.; WALSH, B. **Genetics and analysis of quantitative traits**. Sunderland: Sinauer Associates, Inc., 1997. 980 p.

PAWITAN, Y. **In All Likelihood**. Statistical Modelling and Inference Using Likelihood. Oxford University Press, 2001.

RAMALHO, M. A. P. et al. Aplicações da genética quantitativa no melhoramento de plantas autógamas. **Lavras: UFLA**, 2012.

RESENDE, M.D.V. **Genética Biométrica e Estatística no Melhoramento de Plantas Perenes**. Brasília: Embrapa Informação Tecnológica, 2002. 975p.

RESENDE, M. D. V. Seleção genômica ampla (GWS) e modelos lineares mistos. In: **Matemática e Estatística na Análise de Experimentos e no Melhoramento Genético**. Colombo: Embrapa Florestas, 2007. p. 517-534.

RESENDE, M.D.V. de; **Genética Quantitativa e de Populações.** Visconde do Rio Branco: Suprema, 2015. 463p .

RESENDE, M. D. V. de; DUARTE, J. B. Precisão e controle de qualidade em experimentos de avaliação de cultivares. **Pesquisa Agropecuária Tropical**, v. 37, n. 3, p. 182-194, 2007.

RESENDE, M. D. V. de; LOPES, P. S.; SILVA, R.L.; PIRES, I.E. Seleção genômica ampla (GWS) e maximização da eficiência do melhoramento genético. **Pesquisa Florestal Brasileira**, v. 56, p. 63-78, 2008.

RESENDE, M.D.V. de; SILVA, F. F. E. ; AZEVEDO, C. F. . **Estatística Matemática, Biométrica e Computacional**. Visconde do Rio Branco: Suprema, 2014. 881p .

SORENSEN, D.; GIANOLA, D. **Likelihood, Bayesian and MCMC methods in quantitative genetics**. New York: Springer Verlag, 2002. 740 p.

Biometria aplicada ao melhoramento animal e seus desafios

Paulo Sávio Lopes[1], Hinayah Rojas de Oliveira[2], Vinicius Silva Junqueira[3], Renata Veroneze, Delvan[4] Alves da Silva[5], Fabyano Fonseca e Silva[6]

Introdução

A eficiência dos programas de melhoramento genético depende da acurácia com que os indivíduos submetidos à seleção são avaliados. Para maximizar o progresso genético, é importante a correta classificação destes com base no valor genético predito.

Diversos métodos de avaliação genética de animais têm sido utilizados ao longo dos anos na escolha dos indivíduos candidatos à seleção. O índice de seleção foi utilizado por muitas décadas na seleção de animais. Posteriormente, a metodologia de modelos mistos, que é utilizada na obtenção do melhor preditor linear não-viesado (BLUP – *Best Linear Unbiased Predictor*) dos valores genéticos dos animais, substituiu o índice de seleção. Recentemente, a seleção genômica passou a ser utilizada na obtenção dos GBLUP (BLUP Genômico) dos animais.

As metodologias de predição dos valores genéticos dos animais pressupõem o conhecimento prévio dos componentes de variâncias e covariâncias das características envolvidas no processo de seleção. Como estes não são conhecidos, o que se faz é estimá-los a partir dos próprios dados.

Nesse contexto, diversas metodologias foram propostas para estimação desses componentes de (co)variância, entre as quais se destacam os métodos de Henderson, a máxima verossimilhança restrita (REML – *Restricted Maximum Likelihood*) e os métodos Bayesianos. Os métodos de Henderson foram utilizados até década de 90, quando então foram substituídos pelo REML, que desde então, vem sendo usado no Melhoramento Animal, em razão de suas propriedades de estimador, da disponibilidade de vários

[1]Zootecnista, M.S., D.S. e Professor da Universidade Federal de Viçosa. E-mail: plopes@ufv.br.
[2]Zootecnista, M.S. e Doutoranda da Universidade Federal de Viçosa. E-mail: hynayah@gmail.com.
[3]Médico Veterinário, M.S. e Doutorando da Universidade Federal de Viçosa. E-mail: junqueiravinicius@hotmail.com
[4]Zootecnista, M.S., D.S. e Pesquisadora da Empresa Brasileira de Pesquisa Agropecuária de Minas Gerais. E-mail: veronezerenata@gmail.com
[5]Zootecnista, M.S. e Doutorando da Universidade Federal de Viçosa. E-mail: delvanalves@hotmail.com
[6]Zootecnista, M.S., D.S. e Professor da Universidade Federal de Viçosa. E-mail: fabyanofonseca@ufv.br

softwares e da facilidade de uso. Recentemente, as abordagens Bayesianas têm se destacado, por serem mais flexíveis e adequadas ao uso em grandes bancos de dados.

Ressalta-se, ainda, a grande gama de modelos que têm sido utilizados na avaliação genética das diversas espécies de animais de produção. As equações de modelos mistos de Henderson, mediante uso do modelo animal, permitem a inclusão da informação completa de família por meio da matriz de parentesco; a comparação de indivíduos de diferentes níveis de efeitos fixos; a avaliação simultânea de reprodutores, fêmeas e progênie; a avaliação de indivíduos sem observações, com observações perdidas ou com observações em apenas algumas características; bem como a avaliação de características múltiplas, medidas repetidas, efeito materno, efeitos não-aditivos e interação genótipo x ambiente. As informações de marcadores moleculares também podem ser introduzidas na avaliação genética, mediante utilização da matriz de parentesco genômico nas equações de modelos mistos em substituição à matriz de parentesco tradicional de pedigree.

Índice de Seleção

O índice de seleção, inicialmente desenvolvido por Smith (1936), para plantas, e posteriormente por Hazel (1943), para animais, caracteriza-se por combinar todas as características em apenas um número (índice) para cada animal.

O valor genotípico ou genético total pode ser obtido pela soma dos valores genotípicos ou genéticos das características ponderadas pelos respectivos valores econômicos, dado por

$$\mathbf{H = a_1 g_1 + a_2 g_2 + \cdots + a_n g_n}$$

em que $\mathbf{H}$ é o agregado genotípico ou valor genético total; $\mathbf{a_i}$, o valor econômico da característica i; e $\mathbf{g_i}$, o valor genotípico ou genético da característica i.

Nesse caso, a seleção é feita por um índice que é definido como uma função linear dos valores das observações medidas nos indivíduos (valores fenotípicos), ou seja,

$$\mathbf{I = b_1 X_1 + b_2 X_2 + \cdots + b_n X_n}$$

em que $\mathbf{I}$ é o índice de seleção; $\mathbf{X_i}$, valor observado da característica i no animal (valor fenotípico); e $\mathbf{b_i}$, o valor de ponderação da característica i.

Para obtenção dos valores de ponderação (**b**) do índice de seleção são necessários os conhecimentos dos valores econômicos e das variâncias e covariâncias genéticas e fenotípicas das características.

Os valores de ponderação são obtidos mediante o seguinte sistema de equações:

$$\begin{bmatrix} \sigma^2_{X_1} & \sigma_{X_1X_2} & \cdots & \sigma_{X_1X_n} \\ \sigma_{X_1X_2} & \sigma^2_{X_2} & \cdots & \sigma_{X_2X_n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{X_1X_n} & \sigma_{X_2X_n} & \cdots & \sigma^2_{X_n} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} = \begin{bmatrix} \sigma^2_{a_1} & \sigma_{a_1a_2} & \cdots & \sigma_{a_1a_n} \\ \sigma_{a_1a_2} & \sigma^2_{a_2} & \cdots & \sigma_{a_2a_n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{a_1a_n} & \sigma_{a_2a_n} & \cdots & \sigma^2_{a_n} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}$$

ou ainda,

$$\mathbf{Pb = Ga}$$

dessa maneira, **b** é obtido por

$$\mathbf{b = P^{-1}Ga}$$

em que

P é a matriz de variâncias e covariâncias fenotípicas das características;

G é a matriz de variâncias e covariâncias genéticas das características;

b é o vetor de valores de ponderação das características;

a é o vetor de valores econômicos das características.

Apesar da sua eficiência e facilidade de uso, o índice de seleção foi substituído a várias décadas pela metodologia de modelos mistos na avaliação genética de animais.

Modelos Mistos

A metodologia de modelos mistos (MMM) é utilizada na obtenção do melhor preditor linear não-viesado (BLUP – *Best Linear Unbiased Predictor*) dos valores genéticos dos animais. Consiste na obtenção dos valores genéticos preditos dos animais, tomados como aleatórios, ajustando-se os dados, concomitantemente, aos efeitos fixos e ao número desigual de informações nas subclasses.

Um aspecto relevante das equações de modelos mistos de Henderson é a possibilidade de utilização da matriz de parentesco (Pedigree ou Genômica) e do modelo animal, em que a equação que descreve uma observação contém um termo referente ao valor genético do animal que a produziu.

Matriz de Parentesco

Henderson (1976a) e Quaas (1976) definem a matriz dos numeradores dos coeficientes de parentesco - "*The numerator relationship matrix* - **A**" como uma matriz com as seguintes propriedades:

- **A** é simétrica;

- $\mathbf{a_{ii} = 1 + f_i}$, em que $\mathbf{a_{ii}}$ é o i-ésimo elemento da diagonal de $\mathbf{A}$, e $\mathbf{f_i}$, o coeficiente de endogamia do i-ésimo indivíduo (Wright);
- $\mathbf{a_{ij}}$ é o ij-ésimo elemento fora da diagonal de $\mathbf{A}$, que é o numerador do coeficiente de parentesco de Wright, entre o indivíduo i e o indivíduo j, ou o relacionamento genético aditivo entre o indivíduo i e o indivíduo j;
- $\mathbf{r_{ij} = \frac{a_{ij}}{\sqrt{a_{ii}a_{jj}}}}$ é o coeficiente de parentesco de Wright entre o i-ésimo e o j-ésimo indivíduo.

O uso das Equações de Modelos Mistos de Henderson (Henderson, 1963 e 1973) na avaliação genética animal envolve o cálculo da inversa de **A**. Quando se considera um grande número de animais, os métodos usuais de inversão de matrizes são impraticáveis na obtenção de $\mathbf{A^{-1}}$. Para contornar essas dificuldades, Henderson (1976a) e Quaas (1976) desenvolveram algoritmos para obtenção direta de $\mathbf{A}$ e de $\mathbf{A^{-1}}$. Esses algoritmos tornaram viável o uso do modelo animal em grande volume de dados na avaliação genética animal.

Modelo Animal

No modelo animal, a equação para uma observação contém um termo referente ao valor genético do individuo que produziu a observação. Dessa maneira, este modelo permite a inclusão de animais com observações perdidas ou com observações em apenas algumas características, e avaliação simultânea de reprodutores, fêmeas e progênie.

O modelo animal utilizado na avaliação genética ao incluir apenas o efeito genético aditivo é representado por

$$\mathbf{y = X\beta + Za + e}$$

em que $\mathbf{y}$ é o vetor de observações dos animais; $\mathbf{X}$, a matriz de incidência de efeitos fixos; β , o vetor de efeitos fixos; $\mathbf{Z}$, a matriz de incidência dos valores genéticos dos animais; $\mathbf{a}$, o vetor de dos valores genéticos dos animais; e $\mathbf{e}$, o vetor de erros aleatórios.

Admitindo-se que $\mathbf{y}$, $\mathbf{a}$ e $\mathbf{e}$ tenham distribuição normal multivariada,

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{a} \\ \mathbf{e} \end{bmatrix} \sim \mathbf{NMV}\left[\begin{bmatrix} \mathbf{X\beta} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{ZGZ'+R} & \mathbf{ZG} & \mathbf{R} \\ \mathbf{GZ'} & \mathbf{G} & \phi \\ \mathbf{R} & \phi & \mathbf{R} \end{bmatrix}\right]$$

em que $\mathbf{R} = \mathbf{I} \otimes \mathbf{R_0}$ ou $\mathbf{R} = \mathbf{R_0} \otimes \mathbf{I}$, sendo $\mathbf{I}$, matriz identidade; $\mathbf{R_0}$, matriz de variâncias e covariâncias residuais entre as características; $\mathbf{G} = \mathbf{A} \otimes \mathbf{G_0}$ ou $\mathbf{G} = \mathbf{G_0} \otimes \mathbf{A}$ sendo $\mathbf{A}$, matriz dos numeradores dos coeficientes de parentesco de Wright; e $\mathbf{G_0}$, matriz de variâncias e covariâncias genéticas aditivas entre as características; o sistema de equações de modelos mistos de Henderson (1963, 1973 e 1984) é dado por

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}^0 \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}$$

Quando se tratar de avaliação de uma única característica, o sistema de equações de modelos mistos pode ser representado por

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z}+\mathbf{A}^{-1}\alpha \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}^0 \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}$$

em que α é a razão entre a variância genética aditiva (σ^2_A) e a residual (σ^2_e) da característica.

Modelo Animal Reduzido

Para contornar o inconveniente da manipulação de matrizes de ordem muito grande, Quaas e Pollak (1980) derivaram o modelo animal reduzido, em que as equações dos indivíduos que não possuem progênie ($\mathbf{n}$ = não-pais) são absorvidas nas equações dos indivíduos que possuem progênie ($\mathbf{p}$ = pais + mães). Assim, o sistema de equações de modelos mistos é reduzido de forma que são obtidas inicialmente as soluções para os efeitos fixos ($\boldsymbol{\beta}^0$) e para os valores genéticos preditos dos pais ($\hat{\mathbf{a}}_\mathbf{p}$). Numa segunda etapa são obtidas as soluções para os valores genéticos preditos dos indivíduos não-pais ($\hat{\mathbf{a}}_\mathbf{n}$).

Ao utilizar a relação entre pais e filhos, dadas por $\mathbf{a} = \frac{1}{2}\mathbf{P}\mathbf{a} + \varphi$, em que $\mathbf{a}$ é o vetor de valores genéticos dos indivíduos, $\mathbf{P}$, uma matriz que relaciona os pais com as progênies e φ, um vetor dos efeitos de amostragem mendeliana, o modelo animal reduzido é dado por (Quaas e Pollak, 1980; Martins et al., 1997)

$$\begin{bmatrix} \mathbf{y_p} \\ \mathbf{y_n} \end{bmatrix} = \begin{bmatrix} \mathbf{X_p} \\ \mathbf{X_n} \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{Z_p} \\ \frac{1}{2}\mathbf{Z_n}\mathbf{P_{np}} \end{bmatrix} + \begin{bmatrix} \mathbf{e_p} \\ \mathbf{Z_n}\varphi_\mathbf{n} + \mathbf{e_n} \end{bmatrix}$$

em que $\mathbf{y_p}$ é o vetor de observações dos indivíduos pais (aqueles que têm progênie), $\mathbf{y_n}$, o de indivíduos não-pais (aqueles que não têm progênie), $\boldsymbol{\beta}$, o de efeitos fixos, $\mathbf{e_p}$ e $\mathbf{e_n}$

, vetores de resíduos e φ, o de efeitos de amostragem mendeliana, $\mathbf{X_p}$ e $\mathbf{X_n}$, são matrizes de incidência de observações dos indivíduos pais e não-pais, respectivamente, $\mathbf{Z_p}$ e $\mathbf{Z_n}$, de incidência de valores genéticos dos indivíduos pais e não-pais, respectivamente.

O sistema de equações de modelos mistos do modelo animal reduzido é dado por

$$\begin{bmatrix} \mathbf{X_p'R_p^{-1}X_p + X_nR_n^{-1}X_n} & \mathbf{X_p'R_p^{-1}Z_p} & \mathbf{X_n'R_n^{-1}Z_n} \\ \mathbf{Z_p'R_p^{-1}X_p} & \mathbf{Z_p'R_p^{-1}Z_p + A^{pp} \otimes G_0^{-1}} & \mathbf{A^{pn} \otimes G_0^{-1}} \\ \mathbf{Z_n'R_n^{-1}X_n} & \mathbf{A^{np} \otimes G_0^{-1}} & \mathbf{Z_n'R_n^{-1}Z_n + A^{nn} \otimes G_0^{-1}} \end{bmatrix} \begin{bmatrix} \beta^0 \\ \hat{a}_p \\ \hat{a}_n \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{X_p'R_p^{-1}y_p + X_n'R_n^{-1}y_n} \\ \mathbf{Z_p'R^{-1}y_p} \\ \mathbf{Z_n'R_n^{-1}y_n} \end{bmatrix}$$

em que $\mathbf{R_p = I_p \otimes R_0}$, $\mathbf{R_n = I_n \otimes R_0}$, sendo , $\mathbf{I_p}$ e $\mathbf{I_n}$, matrizes identidades de ordem p e n respectivamente; $\mathbf{R_0}$, matriz de variâncias e covariâncias residuais entre as características; $\mathbf{G_0}$, matriz de variâncias e covariâncias genéticas aditivas entre as características ; e

$$\mathbf{A^{-1}} = \begin{bmatrix} \mathbf{A^{pp}} & \mathbf{A^{pn}} \\ \mathbf{A^{np}} & \mathbf{A^{nn}} \end{bmatrix} = \begin{bmatrix} \mathbf{A_{pp}} & \mathbf{A_{pn}} \\ \mathbf{A_{np}} & \mathbf{A_{nn}} \end{bmatrix}^{-1}$$

a inversa da matriz dos numeradores dos coeficientes de parentesco de Wright; sendo p, os indivíduos pais e n **n**, os não-pais.

Martins et al. (1997) demonstraram que após a absorção da terceira equação na primeira e na segunda equações, este sistema se reduz a

$$\begin{bmatrix} \mathbf{X'R_m^{-1}X} & \mathbf{X'R_m^{-1}Z_m} \\ \mathbf{Z_m'R_m^{-1}X} & \mathbf{Z_m'R_m^{-1}Z_m + A_{pp}^{-1} \otimes G_0^{-1}} \end{bmatrix} \begin{bmatrix} \beta^0 \\ \hat{a}_p \end{bmatrix} = \begin{bmatrix} \mathbf{X'R_m^{-1}y} \\ \mathbf{Z_m'R_m^{-1}y} \end{bmatrix}$$

em que

$$\mathbf{R_m^{-1}} = \begin{bmatrix} \mathbf{R_p} & \phi \\ \phi & \mathbf{R_{n*}^{-1}} \end{bmatrix}, \ \mathbf{Z_m} = \begin{bmatrix} \mathbf{Z_p} \\ \mathbf{\frac{1}{2} Z_n P_{np} \otimes I_0} \end{bmatrix}$$

sendo $\mathbf{R_n^* = Z_n D_n \otimes G_0 Z_n' + R_n}$; $\mathbf{D_n}$ é uma matriz diagonal e $\mathbf{I_0}$ é uma matriz identidade das mesmas dimensões de $\mathbf{G_0}$.

Modelo com efeito materno e efeito de ambiente permanente materno

Em mamíferos, algumas características são influenciadas pelo efeito genético do próprio indivíduo (efeito genético direto), pelo efeito proporcionado pela mãe durante a gestação e na fase de amamentação (efeito genético materno) e pelos efeitos de ambiente.

O modelo utilizado na avaliação genética de características com influência materna é dado por

$$\mathbf{y = X\beta + Z_1a + Z_2m + Z_3pm + e}$$

em que $\mathbf{m}$ é o vetor de efeitos genéticos maternos, $\mathbf{pm}$, o vetor de efeitos de ambientes permanentes maternos, sendo os demais termos definidos anteriormente.

O sistema de equações de modelos mistos é dado por

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z_1} & \mathbf{X'Z_2} & \mathbf{X'Z_3} \\ \mathbf{Z_1'X} & \mathbf{Z_1'Z_1 + A^{-1}\alpha_1} & \mathbf{Z_1'Z_2 + A^{-1}\alpha_{12}} & \mathbf{Z_1'Z_3} \\ \mathbf{Z_2'X} & \mathbf{Z_2'Z_1 + A^{-1}\alpha_{21}} & \mathbf{Z_2'Z_2 + A^{-1}\alpha_2} & \mathbf{Z_2'Z_3} \\ \mathbf{Z_3'X} & \mathbf{Z_3'Z_1} & \mathbf{Z_3'Z_2} & \mathbf{Z_3'Z_3 + I\alpha_3} \end{bmatrix} \begin{bmatrix} \beta^0 \\ \hat{\mathbf{a}} \\ \hat{\mathbf{m}} \\ \hat{\mathbf{pm}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z_1'y} \\ \mathbf{Z_2'y} \\ \mathbf{Z_3'y} \end{bmatrix}$$

em que

$$\begin{bmatrix} \alpha_1 & \alpha_{12} \\ \alpha_{21} & \alpha_2 \end{bmatrix} = \begin{bmatrix} \sigma_a^2 & \sigma_{am} \\ \sigma_{am} & \sigma_m^2 \end{bmatrix}^{-1} \sigma_e^2, \text{ e } \alpha_3 = \frac{\sigma_e^2}{\sigma_{pm}^2}$$

sendo σ_a^2, a variância genética aditiva direta, σ_m^2, a variância genética aditiva materna, σ_{pm}^2, a variância de ambiente permanente materno, σ_e^2, a variância residual, e σ_{am}, a covariância genética aditiva direta e materna.

Modelos com Medidas Repetidas

Algumas características de interesse zootécnico podem ser mensuradas repetidas vezes ao longo da vida de um indivíduo, como é o caso da produção de leite e do peso corporal do animal. Para estas características, ditas características longitudinais, é possível predizer o valor genético dos indivíduos em tempos específicos, considerando o efeito de ambiente permanente (efeitos não genéticos que podem influenciar de forma permanente a expressão da característica ao longo da vida produtiva do animal). Considerar o efeito de ambiente permanente nas avaliações de características longitudinais permite, na maioria dos casos, predizer os valores genéticos de forma mais acurada.

Dentre as alternativas existentes que permitem predizer os efeitos aleatórios via BLUP e modelar a estrutura de correlação entre as medidas repetidas estão os modelos de repetibilidade e de regressão aleatória.

Modelo de Repetibilidade

Os modelos de repetibilidade constituem uma alternativa simples para avaliar características mensuradas repetidamente ao longo da vida produtiva do animal, visto que estes modelos assumem, para um mesmo indivíduo, que a correlação genética entre todas as medidas é igual a uma unidade (Mrode, 2014). Assim, considera-se que a mesma característica, governada pelos mesmos genes, é medida ao longo do tempo.

Diversos trabalhos considerando o modelo de repetibilidade para analisar características longitudinais como produção de leite estão disponíveis na literatura (Meyer et al., 1994, Suzuki e Van Vleck, 1994, Ptak e Schaeffer, 1993). Os modelos de repetibilidade têm sido também utilizados em estudos de seleção e associação genômica para resistência a carrapatos em bovinos (Cardoso et al., 2015, Mapholi et al., 2016).

O modelo animal unicaracterístico de repetibilidade pode ser representado como segue

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Za} + \mathbf{Wp} + \mathbf{e}$$

em que **y** é o vetor de observações, **β** é o vetor de efeitos fixos, **a, p** e **e** são os vetores de efeitos aleatórios genéticos aditivos, de ambiente permanente e residuais, respectivamente. **X, Z** e **W** são as matrizes de incidência dos referidos efeitos. Além disso, considera-se que $\mathbf{a} \sim N(0, \mathbf{A}\sigma_a^2)$, $\mathbf{p} \sim N(0, \mathbf{I}\sigma_p^2)$ e $\mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$, em que **A** é a matriz dos numeradores dos coeficientes de parentesco de Wright; **I** é a matriz identidade cuja ordem é igual ao número de animais com observações; σ_a^2, σ_p^2 e σ_e^2 são os componentes de variância genético aditivo, de ambiente permanente e residual, respectivamente.

O sistema de equações de modelos mistos de Henderson (Henderson, 1963, Henderson Jr, 1982) que permite a predição dos valores genéticos aditivos e dos efeitos de ambiente permanente, considerando uma característica, é dado por:

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z} & \mathbf{X'W} \\ \mathbf{Z'X} & \mathbf{Z'Z} + \mathbf{A}^{-1}\alpha_1 & \mathbf{Z'W} \\ \mathbf{W'X} & \mathbf{W'Z} & \mathbf{W'W} + \mathbf{I}\alpha_2 \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{a}} \\ \hat{\mathbf{p}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \\ \mathbf{W'y} \end{bmatrix}$$

em que $\alpha_1 = \frac{\sigma_e^2}{\sigma_a^2}$, $\alpha_2 = \frac{\sigma_e^2}{\sigma_p^2}$ e as demais simbologias foram definidas anteriormente.

Além de pressupor que a correlação genética entre todas as medidas é igual a unidade, a variância fenotípica de todas as medidas e a correlação de ambiente permanente entre todos os pares de medidas são consideradas iguais nos modelos de repetibilidade. Assim, o valor genético aditivo e o efeito de ambiente permanente preditos para cada animal são os mesmos para todas as idades (Mrode, 2014).

Contudo, na prática, é esperado que os genes que governem a característica não sejam os mesmos em todas as idades, podendo um determinado animal ser melhor que outro em uma idade e não o ser em outra. Desta forma, como as pressuposições de homogeneidade de variâncias e igualdade de correlação não são realistas para grande parte das características longitudinais de interesse, é recomendada a utilização dos modelos de regressão aleatória.

Modelo de Regressão Aleatória

Dados longitudinais como os de produção de leite no dia do controle necessitam de um tratamento estatístico especial, uma vez que o padrão de covariâncias entre as medidas repetidas é bem estruturado. Assim, os modelos de regressão aleatória têm sido uma alternativa para analisar dados de tais características, visto que tem demonstrado maior acurácia na predição dos valores genéticos do que os modelos de repetibilidade e permitem selecionar animais com base em toda a curva de lactação, não apenas em tempos específicos (Andonov et al., 2013, Menéndez-Buxadera et al., 2010, Schaeffer and Dekkers, 1994).

Os modelos de regressão aleatória foram inicialmente propostos por Henderson Jr (Henderson Jr, 1982) e Laird e Ware (Laird and Ware, 1982) como uma alternativa para corrigir o problema de superparametrização das análises multicaracterísticas, principalmente em situações em que há grande número de mensurações, visto que o objetivo da regressão aleatória é estimar os coeficientes da regressão que expliquem a variação da característica em função do tempo. Ou seja, ao utilizar modelos de regressão as predições não são realizadas a cada tempo especificamente, mas sim para os coeficientes da regressão (Schaeffer and Dekkers, 1994, Schaeffer, 2004).

Desta forma, as diferenças entre indivíduos são modeladas a partir dos coeficientes da regressão que representam o comportamento dos efeitos aleatórios genéticos aditivos e de ambiente permanente, que são obtidos como desvio de uma curva média da população (curva fixa), e as variações não explicadas pela regressão formam o resíduo. As estimativas dos componentes de covariância atribuídos aos coeficientes de

regressão aleatória permitem estimar a covariância entre quaisquer valores da variável independente para o efeito aleatório modelado, através da função de covariância. A função de covariância descreve toda a estrutura de covariância do efeito estudado para o intervalo considerado na variável independente. Maiores detalhes podem ser encontrados em Meyer e Hill (1997) e Schaeffer (2004).

Dentre as funções de covariância mais utilizadas estão as baseadas nos Polinômios Ortogonais de Legendre (Kirkpatrick et al., 1990) e nos Polinômios Segmentados (Wegman and Wright, 1983), nestes últimos havendo destaque para os Polinômios Segmentados do tipo B (De Boor and Mathématicien, 1978, Meyer, 2005). Estas funções, denominadas funções não paramétricas, são decorrentes de uma transformação que se faz na variável independente que garante uma série de propriedades estatísticas interessantes, como a redução dos problemas de multicolinearidade.

O modelo animal unicaracterístico de regressão aleatória pode ser representado como segue:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Qa} + \mathbf{Vp} + \mathbf{e}$$

em que $\mathbf{y}$ é o vetor de observações, $\boldsymbol{\beta}$ é o vetor de efeitos fixos e dos coeficientes da regressão fixa do modelo, $\mathbf{a}$ e $\mathbf{p}$ são os vetores dos coeficientes de regressão aleatórios atribuídos aos efeitos genéticos aditivos e de ambiente permanente, respectivamente, e $\mathbf{e}$ é o vetor aleatório de resíduos. $\mathbf{X}$, $\mathbf{Q}$ e $\mathbf{V}$ são as matrizes de incidência dos polinômios relativas aos referidos efeitos. Além disso, assume-se que $\mathbf{a} \sim N(0, \mathbf{A} \otimes \mathbf{G})$, $\mathbf{p} \sim N(0, \mathbf{I} \otimes \mathbf{P})$ e, considerando homogeneidade de variância residual, $\mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$. As matrizes $\mathbf{A}$ e $\mathbf{I}$ foram definidas anteriormente; e as matrizes $\mathbf{G}$ e $\mathbf{P}$ referem-se, respectivamente, às matrizes de variância e covariância dos efeitos genéticos e de ambiente permanente para os polinômios ajustados.

Muitos estudos tem demostrado que particionar a variância residual em classes (ou seja, considerar heterogeneidade de variância residual no modelo), permite um melhor ajuste do modelo aos dados, visto que a variância residual muda ao longo da vida do animal (Flores and Werf, 2015, Pereira et al., 2013).

O sistema de equações de modelos mistos de Henderson, para uma característica, é dado por

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Q} & \mathbf{X'V} \\ \mathbf{Q'X} & \mathbf{Q'Q} + \mathbf{A}^{-1} \otimes \mathbf{G}^{-1} & \mathbf{Q'V} \\ \mathbf{V'X} & \mathbf{V'Q} & \mathbf{V'V} + \mathbf{P}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{a}} \\ \hat{\mathbf{p}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Q'y} \\ \mathbf{V'y} \end{bmatrix}$$

cujas simbologias foram definidas anteriormente.

Contudo, vale ressaltar que os valores genéticos preditos ($\hat{\mathbf{a}}$) são para os coeficientes da regressão que representam o comportamento dos efeitos aleatórios genéticos aditivos. Assim, para obter o valor genético predito de cada indivíduo i para cada tempo j ($\hat{\mathbf{u}}_{ij}$) é necessário multiplicar a matriz de polinômios gerada pela função de covariância pelo vetor de valores genéticos preditos de cada indivíduo (Mrode, 2014, Schaeffer e Dekkers, 1994, Schaeffer, 2004), ou seja

$$\hat{\mathbf{u}}_{ij} = \mathbf{T}_i \, \hat{\mathbf{a}}_i$$

em que $\hat{\mathbf{a}}_i$ é o vetor de valores genéticos para os coeficientes do animal i; e $\mathbf{T}_i$ é a matriz de polinômios gerada pela função de covariância usada.

Em análises multicaracterísticas é possível combinar diferentes funções para descrever diferentes características em um mesmo modelo, conforme demonstrado por Oliveira et al. (2016).

Modelos Multicaracterísticos

Os modelos multicaracterísticos são utilizados quando se considera mais de uma característica simultaneamente na avaliação genética. Dessa maneira, nos modelos animais multicaracterísticos há a possibilidade de inclusão de indivíduos com observações em apenas algumas características. A grande vantagem dos modelos multicaracterísticos é a exploração da correlação genética entre as características, o que aumenta consideravelmente a acurácia na predição dos valores genéticos dos indivíduos.

O modelo animal utilizado na avaliação de características múltiplas, ou seja, várias características simultaneamente, é representado por (Henderson, 1976b, Henderson, 1984, Martins et al., 1997)

$$\begin{bmatrix} \mathbf{y_1} \\ \mathbf{y_2} \\ \vdots \\ \mathbf{y_t} \end{bmatrix} = \begin{bmatrix} \mathbf{X_1} & \phi & \phi & \phi \\ \phi & \mathbf{X_2} & \phi & \phi \\ \phi & \phi & \ddots & \phi \\ \phi & \phi & \phi & \mathbf{X_t} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_t \end{bmatrix} + \begin{bmatrix} \mathbf{Z_1} & \phi & \phi & \phi \\ \phi & \mathbf{Z_2} & \phi & \phi \\ \phi & \phi & \ddots & \phi \\ \phi & \phi & \phi & \mathbf{Z_t} \end{bmatrix} \begin{bmatrix} \mathbf{a_1} \\ \mathbf{a_2} \\ \vdots \\ \mathbf{a_t} \end{bmatrix} + \begin{bmatrix} \mathbf{e_1} \\ \mathbf{e_2} \\ \vdots \\ \mathbf{e_t} \end{bmatrix}$$

em que $\mathbf{y_i}$ é o vetor de observações das $\mathbf{i}\ (\mathbf{i = 1,2,\cdots t})$ características dos $\mathbf{n}$ animais; $\mathbf{X_i}$, a matriz de incidência de efeitos fixos; $\boldsymbol{\beta}_\mathbf{i}$, o vetor de efeitos fixos; $\mathbf{Z_i}$, a matriz de incidência dos valores genéticos dos animais; $\mathbf{a_i}$, o vetor de dos valores genéticos dos animais; e $\mathbf{e_i}$, o vetor de erros aleatórios.

O sistema de equações de modelos mistos para características múltiplas é dado por

$$\begin{bmatrix} \mathbf{X_1'R^{11}X_1} & \cdots & \mathbf{X_1'R^{1t}X_t} & \mathbf{X_1'R^{11}Z_1} & \cdots & \mathbf{X_1'R^{1t}Z_t} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{X_t'R^{1t}X_1} & \cdots & \mathbf{X_t'R^{tt}X_t} & \mathbf{X_t'R^{1t}Z_1} & \cdots & \mathbf{X_t'R^{tt}Z_t} \\ \mathbf{Z_1'R^{11}X_1} & \cdots & \mathbf{Z_1'R^{1t}X_t} & \mathbf{Z_1'R^{11}Z_1+G^{11}} & \cdots & \mathbf{Z_1'R^{1t}Z_t+G^{1t}} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{Z_t'R^{1t}X_1} & \cdots & \mathbf{Z_t'R^{tt}X_t} & \mathbf{Z_t'R^{1t}Z_1+G^{1t}} & \cdots & \mathbf{Z_t'R^{tt}Z_t+G^{tt}} \end{bmatrix} \begin{bmatrix} \beta_1^0 \\ \vdots \\ \beta_t^0 \\ \hat{\mathbf{a}}_1 \\ \vdots \\ \hat{\mathbf{a}}_t \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{X_1'R^{11}y_1} & \cdots & \mathbf{X_1'R^{1t}y_t} \\ \vdots & \ddots & \vdots \\ \mathbf{X_t'R^{1t}y_1} & \cdots & \mathbf{X_t'R^{tt}y_t} \\ \mathbf{Z_1'R^{11}y_1} & \cdots & \mathbf{Z_1'R^{1t}y_t} \\ \vdots & \ddots & \vdots \\ \mathbf{Z_t'R^{1t}y_1} & \cdots & \mathbf{Z_t'R^{tt}y_t} \end{bmatrix}$$

em que

$$\mathbf{R} = \mathbf{R_0} \otimes \mathbf{I} = \begin{bmatrix} \sigma_{e1}^2 & \cdots & \sigma_{e1et} \\ \vdots & \ddots & \vdots \\ \sigma_{ete1} & \cdots & \sigma_{et}^2 \end{bmatrix} \otimes \mathbf{I}, \quad \mathbf{G} = \mathbf{G_0} \otimes \mathbf{A} = \begin{bmatrix} \sigma_{a1}^2 & \cdots & \sigma_{a1at} \\ \vdots & \ddots & \vdots \\ \sigma_{ata1} & \cdots & \sigma_{at}^2 \end{bmatrix} \otimes \mathbf{A},$$

$$\mathbf{R^{-1}} = \mathbf{R_0^{-1}} \otimes \mathbf{I} = \begin{bmatrix} \mathbf{R^{11}} & \cdots & \mathbf{R^{1t}} \\ \vdots & \ddots & \vdots \\ \mathbf{R^{t1}} & \cdots & \mathbf{R^{tt}} \end{bmatrix}, \quad \mathbf{G^{-1}} = \mathbf{G_0^{-1}} \otimes \mathbf{A_0^{-1}} = \begin{bmatrix} \mathbf{G^{11}} & \cdots & \mathbf{G^{1t}} \\ \vdots & \ddots & \vdots \\ \mathbf{G^{t1}} & \cdots & \mathbf{G^{tt}} \end{bmatrix}$$

sendo, σ_{ai}^2 e σ_{ei}^2, as variâncias genética aditiva e residual da i-ésima característica; σ_{aij} e σ_{eij}, as covariâncias genética aditiva e residual entre a i-ésima e a j-ésima característica.

Avaliação genética em populações multirraciais

Ponto fundamental no planejamento de um sistema produtivo animal é a definição da raça ou raças que serão utilizadas, mas também a definição do sistema de cruzamento ao qual os animais serão submetidos. Populações compostas por animais de raças puras e seus cruzamentos são conhecidas como populações multirraciais (ex., Braford, Brangus, Canchim, Girolando, dentre outras). A adoção de um sistema de cruzamento traz diversos benefícios e são predominantes nos sistemas de produção de aves e suínos (Christensen et al. 2014), mas também é presente na pecuária de corte e leite.

Animais cruzados se destacam, principalmente, por apresentarem, de forma geral, desempenho produtivo superior em relação aos animais de raça pura, sendo tal

superioridade atribuída a heterose e aos efeitos de complementariedade (Gregory et al. 1999). A heterose pode ser definida como o evento contrário ao da endogamia (Dickerson 1973). A endogamia ocorre quando alelos idênticos se unem durante o processo de fecundação, sendo responsável por causar mudança na frequência alélica da população ao fixar alelos. Consequentemente, é observada redução do poder adaptativo (vigor híbrido) dos animais e afeta mais fortemente características de baixa herdabilidade, tais como características reprodutivas (ex., idade ao primeiro parto e intervalo de partos). Esse fenômeno é conhecido como depressão endogâmica. A heterose, principalmente quando é resultante do acasalamento entre animais de raças divergentes, proporciona a recuperação do potencial gênico encoberto pela presença de genes em homozigose.

A partir do uso de sistemas de cruzamento também é possível explorar os efeitos de complementariedade ao selecionar raças com características fenotípicas distintas e economicamente interessantes. Por exemplo, a resistência a carrapatos de bovinos da raça Nelore e o maior marmoreio (gordura intramuscular) da carne de bovinos da raça Hereford. Além disso, os efeitos de dominância (resultado da interação alélica intra-*locus*) se tornam mais evidentes nos animais cruzados. Hill et al. (2008) apresentam argumentos de que os efeitos genéticos não aditivos (dominância e epistasia) são pouco relevantes quando comparados com o efeito genético aditivo. Porém, em populações multirraciais o efeito de dominância ocupa lugar de destaque porque um dos objetivos nessas populações é selecionar animais com base no potencial de combinação de cruzamento considerando simultaneamente os efeitos genéticos aditivos e não aditivos (Ibánẽz-Escriche et al. 2009). De forma geral, espera-se que o vigor híbrido seja reestabelecido no cruzamento. No entanto, em algumas situações é observado desvio do desempenho dos animais em função de diferentes níveis de heterozigose. Para isso, Dickerson (1973) introduziu o conceito de perda por recombinação ao explorar os efeitos de heterose e de complementariedade. As perdas por recombinação ocorrem em decorrência da quebra das ligações epistáticas favoráveis estabelecidas pela seleção a longo prazo dentro da raça. Segundo Kinghorn (1980), esse fenômeno é possivelmente um dos mais importantes efeitos negativos que surge pelo uso de sistemas de acasalamento.

Grande parte do sucesso em um sistema produtivo animal depende da adequada definição dos cruzamentos, contudo, novas raças e cruzamentos têm sido introduzidos sem prévias avaliações de desempenho em diferentes ambientes e/ou avaliados somente em função de modelos aditivos (Elzo e Famula 1985). A seleção de animais puros com o

objetivo de predizer o desempenho dos animais cruzados não reflete em sucesso devido a existência de diferentes fatores (genéticos e ambientais) entre as populações puras e cruzadas (Ibánẽz-Escriche et al. 2009). Nesse contexto, diversas metodologias têm sido propostas para a avaliação genética de populações multirraciais com a inclusão de efeitos (fixos) extras e de componentes de (co)variância específicos para raças (ou combinações raciais) (Elzo 1990, Elzo e Famula 1985, Lo et al. 1995, Lo et al. 1993, Lo et al. 1997). Dentre os efeitos fixos comumente incluídos na avaliação genética de populações multirraciais podem ser citados os de grupos raciais (do pai, da mãe, do avô-materno, do avô-paterno) e/ou covariáveis de heterose, composição racial, perda por recombinação (epistasia) e endogamia.

Um modelo animal aditivo multirracial pode ser representado como

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Za} + \mathbf{e}$$

em que $\mathbf{y}$ é o vetor de observações, $\boldsymbol{\beta}$ é o vetor de efeitos fixos incluindo heterose, composição racial, perda por recombinação e grupos de contemporâneos, $\mathbf{a} \sim N(\mathbf{0}, \mathbf{G}^*)$ é o vetor de efeitos genéticos aditivos, $\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma_e^2)$ é o vetor de resíduos. As matrizes $\mathbf{X}$ e $\mathbf{Z}$ relacionam as observações aos respectivos níveis dos efeitos dos indivíduos.

As equações de modelos mistos do modelo acima podem ser representadas como se segue

$$\begin{bmatrix} \mathbf{X'R^{-1}X} & \mathbf{X'R^{-1}Z} \\ \mathbf{Z'R^{-1}X} & \mathbf{Z'R^{-1}Z + G^{-1}} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta^0} \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'R^{-1}y} \\ \mathbf{Z'R^{-1}y} \end{bmatrix}$$

em que $\mathbf{R^{-1}}$ é a matriz de (co)variâncias residuais e $\mathbf{G^{-1}} = (\mathbf{G}^*)^{-1}$ é inversa da matriz de (co)variância de efeitos genéticos aditivos multirraciais, sendo que os componentes de variância intra- e inter-raciais são calculados como combinações lineares. Os termos da matriz $\mathbf{G}^*$ podem ser calculados de acordo com Lo et al. (1993), os quais consideram a variância de segregação alélica devido às diferenças raciais na derivação dos termos de (co)variância para todas as composições raciais de uma população multirracial. A covariância genética aditiva entre os indivíduos *i* e *i'* são calculadas como a média das covariâncias entre o indivíduo *i'* e os pais do indivíduo *i* (*j* e *k*).

$$\mathrm{Cov}(g_i, g_{i'}) = \frac{1}{2}\left[g_{ji'} + g_{ki'}\right]$$

Os termos de variância (diagonal) da matriz são calculados como combinações lineares entre os componentes de variância de raças e seus cruzamentos da seguinte maneira

$$\mathrm{Var}(g_{ii}) = \sum_{p=1}^{P} f_p^i \sigma_p^2 + \frac{1}{2} g_{SD} + 2 \sum_{p=1}^{P} \sum_{p'>p}^{P} (f_p^S f_{p'}^S + f_p^D f_{p'}^D) \sigma_{pp'}^2$$

em que f_p^i é a proporção racial do individuo *i* na raça *p*, $f_p^S (f_p^D)$ é a proporção racial do pai (mãe) na raça *p*, $f_{p'}^S$ $(f_{p'}^D)$ é a proporção do pai (mãe) na raça *p'*, σ_p^2 é a variância da raça *p*, $\sigma_{pp'}^2$ é a variância genética aditiva dos indivíduos do cruzamento *pp'* e g_{SD} é a covariância genética aditiva entre os pais do indivíduo *i*.

Apesar de disponíveis metodologias para a avaliação de populações multirraciais, grande parte dos programas de melhoramento genético adotam metodologias tradicionais na avaliação genética dessas populações. Isso é resultado de fatores complicadores na avaliação genética dessas populações, tais como a dificuldade de definição do modelo multirracial, a determinação da conexidade genética entre os grupos de contemporâneos, as metodologias de estimação de componentes de (co)variância (calculados como combinações lineares entre componentes intra- e inter-raciais) e a complexidade computacional. Note que a matriz $\mathbf{G}^{-1}$ é construída a partir dos componentes de variância das raças puras e da segregação alélica dos cruzamentos. Isso faz com que a maximização da função de verossimilhança nos modelos multirraciais seja complexa (Christensen et al. 2014, García-Cortés e Toro 2006). Uma alternativa seria a aplicação de métodos Bayesianos para a estimação de tais componentes de variância. Entretanto, as distribuições condicionais completas *a posteriori* para os componentes de variância são complexas, o que demanda a implementação do algoritmo Metropolis-Hastings (Cardoso e Tempelman 2004).

Não é raro surgirem problemas de multicolinearidade e grande desbalanceamento nas subclasses em modelos multirraciais, visto que é improvável que todas as raças estejam representadas em número igual e em todos os grupos de contemporâneos, especialmente em situações práticas. Além disso, em uma população multirracial é importante que conexidade genética seja garantida dentro e entre os grupos de contemporâneos, pois isso evitará problemas de colinearidade. Em uma etapa inicial, as conexões genéticas são avaliadas em função dos reprodutores (presentes em múltiplos grupos raciais e grupos de contemporâneos) devido a facilidade de disseminação de material genético (sêmen).

Segundo Elzo (1995), a possibilidade de comparação de animais de diferentes raças (e composições raciais) é uma das vantagens da avaliação multirracial. Isso pode ser realizado avaliando a habilidade de combinação dos machos com fêmeas de diferentes

composições raciais utilizando as informações de valores genéticos totais. Por sua vez, os valores genéticos aditivos nesses modelos podem ser utilizados em comparações diretas entre os animais das diferentes raças (ou composições raciais) de forma análoga a uma avaliação tradicional. Porém, a correta definição dos pares de acasalamento depende de as etapas anteriores terem sido realizadas adequadamente.

O sucesso no estabelecimento de um sistema de avaliação genética multirracial depende do desenvolvimento de estratégias de divulgação dos valores genéticos multirraciais (aditivos, não aditivos e totais) concomitantemente com adequada orientação do produtor sobre o uso dessas informações. Uma das principais discussões sobre a adoção de modelos multirraciais é a estratégia de uso dos valores genéticos obtidos por essas metodologias. Segundo Elzo e Borjas (2004), independente do modelo de avaliação multirracial adotado é importante que o valor genético aditivo seja o primeiro critério de seleção, porque determina mudanças genéticas permanentes e de longo prazo. Somente em uma próxima etapa as informações não aditivas deveriam ser utilizadas. Dessa forma, seria dado maior peso de importância aos animais que proporcionam mudanças permanentes (herdadas geneticamente), diferentemente das informações não aditivas (dominância e epistasia).

O Brasil é um país de dimensões continentais e com grande diversidade ambiental (ex., clima, nutrição e práticas reprodutivas), o que de certa forma implica em uma quantidade considerável de diferentes fatores que devem ser previstos na modelagem multirracial. Todas essas peculiaridades impõem um grande desafio para as avaliações genéticas, principalmente aos programas de âmbito nacional. Então, a viabilidade de um sistema de avaliação genética multirracial nacional depende do custo-benefício de implementação e da estratégia de uso das informações geradas nesses modelos.

Componentes de Variância

Na predição dos valores genéticos dos animais pressupõe-se o conhecimento prévio dos componentes de variâncias e covariâncias das características. Entretanto, na maioria das situações, esses componentes não são conhecidos e geralmente o que se faz é estimá-los a partir dos próprios dados.

Diversas metodologias têm sido sugeridas para estimação desses componentes: métodos de Henderson (Henderson, 1953, citado por Lopes et al., 1998), máxima verossimilhança (*Maximum likelihood* – *ML*, Hartley e Rao, 1967), máxima verossimilhança restrita (*Restricted maximum likelihood* – REML, Patterson e

Thompson, 1971) e os métodos Bayesianos (Gianola e Fernando, 1986; Sorensen e Gianola, 2002).

Os métodos de Henderson, principalmente o método III, foram amplamente utilizados até a década de 1990. A partir daí o método da máxima verossimilhança restrita (REML) passou a ser o recomendado na estimação dos componentes de variância, especialmente em dados desbalanceados, o que é comum nos programas de melhoramento genético animal.

As estimativas obtidas pelo REML sempre caem dentro do espaço paramétrico, ou seja, as estimativas de componentes de variância são sempre não-negativas. Esse método elimina o viés atribuído às mudanças nas frequências gênicas (oscilação genética ou seleção); e o viés resultante da perda de graus de liberdade na estimação dos efeitos fixos do modelo (Anderson, 1984, Henderson, 1984 e 1986, Sorensen e Kennedy, 1984).

O REML baseia-se na maximização do logaritmo da função densidade de probabilidade das observações, tomando essa função como a soma de duas funções independentes, uma referente aos efeitos fixos e outra aos aleatórios (Patterson e Thompson, 1971, citados por Lopes et al., 1998).

Ao considerar o modelo animal utilizado na avaliação genética incluindo apenas o efeito genético aditivo que é representado por

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{a} + \mathbf{e}$$

na derivação do REML, o logaritmo da função densidade de probabilidade de y, dado por

$$\lambda = -\frac{1}{2}\mathbf{n}\mathbf{q}\ln(2\pi) - \frac{1}{2}\ln\left|\mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}\right| - \frac{1}{2}\left[\mathbf{y}'\left(\mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}\right)^{-1}\mathbf{y} - 2\mathbf{y}'\left(\mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}\right)^{-1}\mathbf{X}\boldsymbol{\beta}\right] + \boldsymbol{\beta}'\mathbf{X}'\left(\mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}\right)^{-1}\mathbf{X}\boldsymbol{\beta},$$

é dividido em duas partes, uma referente aos contrastes entre os efeitos fixos e outra referente aos contrastes linearmente independentes entre as partes aleatórias das observações. Na estimação pelo REML, a função densidade de probabilidade referente aos contrastes linearmente independentes entre as partes aleatórias das observações é derivada em relação aos componentes de variância e covariância genética aditiva e residual das características e tais derivadas são igualadas a zero. Tal procedimento é apresentado em Lopes et al. (1998). O sistema de equações do REML é dado por

$$\mathbf{T}\mathbf{s} = \mathbf{v},$$

que equivale a

$$\begin{bmatrix} t'_{1e} \\ t'_{12e} \\ \vdots \\ t'_{1qe} \\ \vdots \\ t'_{(q-1)qe} \\ t'_{qe} \\ t'_{1a} \\ t'_{12a} \\ \vdots \\ t'_{1qa} \\ \vdots \\ t'_{(q-1)qa} \\ t'_{qa} \end{bmatrix} [s] = \begin{bmatrix} y'PR_1Py \\ y'PR_{12}Py \\ \vdots \\ y'PR_{1q}Py \\ \vdots \\ y'PR_{(q-1)q}Py \\ y'PR_qPy \\ y'PG_1Py \\ y'PG_{12}Py \\ \vdots \\ y'PG_{1q}Py \\ \vdots \\ y'PG_{(q-1)q}Py \\ y'PG_qPy \end{bmatrix}$$

em que

$$s' = \left[\sigma^2_{e1} \quad \sigma_{e12} \quad \cdots \quad \sigma_{e1q} \quad \cdots \quad \sigma_{e(q-1)q} \quad \sigma^2_{eq} \quad \sigma^2_{a1} \quad \sigma_{a12} \quad \cdots \quad \sigma_{a1q} \quad \cdots \quad \sigma_{a(q-1)q} \quad \sigma^2_{aq}\right],$$

$$t_{ie} = \begin{bmatrix} tr(PR_iPR_1 \\ tr(PR_iPR_{12} \\ \vdots \\ tr(PR_iPR_{1q} \\ \vdots \\ tr(PR_iPR_{(q-1)q} \\ tr(PR_iPR_q \\ tr(PR_iPG_1 \\ tr(PR_iPG_{12} \\ \vdots \\ tr(PR_iPG_{1q} \\ \vdots \\ tr(PR_iPG_{(q-1)q} \\ tr(PR_iPGq \end{bmatrix}, \; t_{ie} = \begin{bmatrix} tr(PR_{ij}PR_1 \\ tr(PR_{ij}PR_{12} \\ \vdots \\ tr(PR_{ij}PR_{1q} \\ \vdots \\ tr(PR_{ij}PR_{(q-1)q} \\ tr(PR_{ij}PR_q \\ tr(PR_{ij}PG_1 \\ tr(PR_{ij}PG_{12} \\ \vdots \\ tr(PR_{ij}PG_{1q} \\ \vdots \\ tr(PR_{ij}PG_{(q-1)q} \\ tr(PR_{ij}PGq \end{bmatrix},$$

$$\mathbf{t_{ie}} = \begin{bmatrix} \mathbf{tr(PG_iPR_1} \\ \mathbf{tr(PG_iPR_{12}} \\ \vdots \\ \mathbf{tr(PG_iPR_{1q}} \\ \vdots \\ \mathbf{tr(PG_iPR_{(q-1)q}} \\ \mathbf{tr(PG_iPR_q} \\ \mathbf{tr(PG_iPG_1} \\ \mathbf{tr(PG_iPG_{12}} \\ \vdots \\ \mathbf{tr(PG_iPG_{1q}} \\ \vdots \\ \mathbf{tr(PG_iPG_{(q-1)q}} \\ \mathbf{tr(PG_iPGq} \end{bmatrix}, \mathbf{t_{ie}} = \begin{bmatrix} \mathbf{tr(PG_{ij}PR_1} \\ \mathbf{tr(PG_{ij}PR_{12}} \\ \vdots \\ \mathbf{tr(PG_{ij}PR_{1q}} \\ \vdots \\ \mathbf{tr(PG_{ij}PR_{(q-1)q}} \\ \mathbf{tr(PG_{ij}PR_q} \\ \mathbf{tr(PG_{ij}PG_1} \\ \mathbf{tr(PG_{ij}PG_{12}} \\ \vdots \\ \mathbf{tr(PG_{ij}PG_{1q}} \\ \vdots \\ \mathbf{tr(PG_{ij}PG_{(q-1)q}} \\ \mathbf{tr(PG_{ij}PGq} \end{bmatrix}$$

sendo

$\mathbf{P = V^{-1} - V^{-1}X\left(X'V^{-1}X\right)^{-}X'V^{-1}}$, o projetor ortogonal da parte aleatória das observações no espaço coluna da matriz X (RAO, 1965), $\mathbf{R_i = Z_i \otimes Q_i}$, $\mathbf{R_{ij} = Z_{ij} \otimes Q_{ij}}$, $\mathbf{G_i = Z_iAZ_i' \otimes Q_i}$ e $\mathbf{G_{ij} = Z_iAZ_j' \otimes Q_{ij}}$, $\mathbf{Z_i}$ e $\mathbf{Z_{ij}}$ são matrizes de incidência, $\mathbf{Q_i}$ e $\mathbf{Q_{ij}}$ indicam a posição dos componentes nas matrizes de variância e covariância a serem estimados, conforme apresentado por Lopes et al. (1998).

As soluções para os componentes de variância e covariância são obtidas de forma iterativa e geralmente implica em grande esforço computacional, principalmente em se tratando de análise de características múltiplas. Para contornar esse problema das equações REML, Graser et al. (1987) propuseram um algoritmo denominado livre de derivadas (*Dervative-Free Restricted Maximum Likelihood* – DFREML).

Esse algoritmo foi implementado em programas computacionais por Meyer (199x) (DFREML) e por Boldman et al. (1993) (MTDFREML – *Multiple Trait Dervative-Free Restricted Maximum Likelihood*). Entretanto, o DFREML não tem sido recomendado em análise com várias características simultaneamente e modelos contendo muitos efeitos. Com isso, diversos programas computacionais que utilizam os algoritmos EM ou AI foram implementados e têm sido amplamente utilizados na estimação de componentes de variância: REMLF90 (Misztal, 2016), WOMBAT (Meyer, 2007), ASREML (Gilmour et al., 2009), dentre outros.

Inferência Bayesiana

Um dos principais objetivos da estatística é realizar inferência sobre os parâmetros de um modelo. Na abordagem frequentista os parâmetros desconhecidos são considerados fixos e toda a análise é baseada nas informações contidas na amostra dos dados, ou seja, distribuições inferenciais são assumidas para os estimadores dos parâmetros, e não propriamente para os parâmetros.

A inferência Bayesiana trata o vetor de parâmetros desconhecidos (θ) como quantidades aleatórias, e qualquer informação inicial sobre eles pode ser representada por meio de distribuições de probabilidade, as quais são denominadas de distribuições *a priori*. Assim, tal abordagem permite incorporar algum conhecimento sobre esses parâmetros antes que os dados tenham sido coletados. Essas distribuições podem ser obtidas através de análises anteriores, experiência do pesquisador na área em questão ou em revisões de literatura sobre o assunto que se deseja tratar.

Em resumo, para se realizar a inferência Bayesiana é necessário assumir uma distribuição de probabilidade *a priori* ($P(\theta)$) para os parâmetros a serem estimados; e uma distribuição para os dados amostrais, denominada de função de verossimilhança $(P(\theta|Y))$. O teorema de Bayes combina estas duas fontes informações em por meio da distribuição *a posteriori* $(P(\theta|Y))$ como segue

$$P(\theta|Y)=\frac{P(Y|\theta)P(\theta)}{\int P(Y|\theta)P(\theta)d\theta}=\frac{P(Y|\theta)P(\theta)}{P(Y)}$$

Uma vez que o denominador não depende de θ, este pode ser omitido, sendo possível reescrever o Teorema de Bayes da seguinte maneira:

$$P(\theta|Y)\propto P(Y|\theta)P(\theta),$$

representando, *Dist. Posteriori* $\propto$ *Função de Verossimilhança x Dist. Priori,* onde $\propto$ indica proporcionalidade.

Sendo θ um conjunto de parâmetros (ou seja, inferência multiparamétrica), a distribuição $(P(\theta|Y))$ é denominada de distribuição conjunta *a posteriori*. Desta forma, para se inferir em relação a qualquer elemento de θ, deve-se então integrar tal distribuição em relação a todos os outros parâmetros (θ_2, θ_3, ..., θ_p) simultaneamente. Assim, se o interesse do pesquisador se concentra sobre determinado conjunto de θ, por exemplo θ_1, é necessário obter a distribuição $P(\theta_1|Y)$, denominada de distribuição marginal *a posteriori* de θ_1, a qual é dada por

$$P(\theta_1|Y)=\int_{\theta\neq\theta_1} P(\theta|Y)d\theta_{\theta\neq\theta_1}.$$

A integração da distribuição conjunta *a posteriori* para a obtenção das distribuições marginais geralmente não é analítica, sendo necessário o uso de algoritmos iterativos especializados como o Gibbs Sampler e o Metropolis-Hastings. Estes algoritmos são denominados de algoritmos MCMC (*Markov Chain - Monte Carlo*). Para a utilização dos mesmos é necessário que se obtenha as distribuições condicionais completas para cada parâmetro a partir da distribuição conjunta *a posteriori*.

Na área de melhoramento animal a abordagem Bayesiana é utilizada para ajustar uma gama variada de modelos mistos, de forma que quanto mais complexos estes modelos, mais evidentes são as vantagens desta abordagem. Para o modelo misto tradicional de Henderson, $\mathbf{y} = \mathbf{X\beta} + \mathbf{Za} + \mathbf{e}$, a inferência Bayesiana é realizada por meio do algoritmo Gibbs Sampler, cujas vantagens são descritas a seguir, segundo Van Tassel et al. (1995): (i) não requer soluções para as equações de modelo misto; (ii) permite a análise de conjunto de dados maiores do que quando se usa REML com técnicas de matrizes esparsas; (iii) propicia estimativas diretas e acuradas dos componentes de variância, valores genéticos e intervalos de credibilidade para tais estimativas e, (iv) possibilita tratar modelos complexos, tais como aqueles envolvendo distribuições não-normais para os dados (tais como Binomial e Poisson) e modelos multi-característicos combinando diferentes distribuições para cada característica.

O ponto de partida para a inferência Bayesiana do modelo misto de Henderson amplamente utilizado no Melhoramento Animal é assumir a distribuição para os dados amostrais, a qual sob normalidade é dada por

$$\mathbf{y}|\boldsymbol{\beta},\mathbf{a},\sigma_a^2,\sigma_e^2 \sim N\left(\mathbf{X\beta}+\mathbf{Za},\ \mathbf{I}\sigma_e^2\right).$$

A partir desta distribuição Normal, tem-se a seguinte função de verossimilhança (distribuição conjunta dos dados amostrais) a ser empregada no teorema de Bayes

$$P(\mathbf{y}|\boldsymbol{\beta},\mathbf{a},\sigma_a^2,\sigma_e^2) \propto \left(\frac{1}{\sigma_e^2}\right)^{\left(\frac{n}{2}\right)} \exp\left\{-\frac{1}{2\sigma_e^2}(\mathbf{y}-(\mathbf{X\beta}+\mathbf{Za}))'(\mathbf{y}-(\mathbf{X\beta}+\mathbf{Za}))\right\}$$

Para o modelo em questão, geralmente as seguintes distribuições a priori para os parâmetros são consideradas (Sorensen e Gianola, 2002):

P(**β**) ~ constante (ou seja, para os efeitos sistemáticos geralmente são utilizadas distribuições a pirori denominadas de não-informativas, ou constantes);

$\mathbf{a} \sim N\left(\mathbf{0},\ \mathbf{A}\sigma_a^2\right)$, sendo **A** a matriz dos numeradores dos coeficientes de parentesco de Wright;

$$\sigma_a^2 \sim \chi^{-2}\left(\nu_a, S_a^2\right) \text{e} \qquad \sigma_e^2 \sim \chi^{-2}\left(\nu_e, S_e^2\right),$$

em que χ^{-2} representa a distribuição de qui-quadrado invertida com parâmetros S^2 (valor esperado para o componente de variância) e ν (grau de confiança depositado no valor esperado S^2).

As distribuições qui-quadrado invertidas assumidas como *priori* para os componentes de variância são dadas, respectivamente, por

$$P\left(\sigma_a^2 \mid S_a^2, \nu_a\right) \propto \left(\frac{1}{\sigma_a^2}\right)^{\left(\frac{\nu_a}{2}+1\right)} \exp\left\{-\frac{\nu_a s_a^2}{2\sigma_a^2}\right\} \text{ e } \quad P\left(\sigma_e^2 \mid S_e^2, \nu_e\right) \propto \left(\frac{1}{\sigma_e^2}\right)^{\left(\frac{\nu_e}{2}+1\right)} \exp\left\{-\frac{\nu_e s_e^2}{2\sigma_e^2}\right\}.$$

É importante realçar que ao assumir $\nu_a = \nu_e = -2$, e $S^2_a = S^2_e = 0$, obtêm-se funções constantes, caracterizando distribuições *a priori* não-informativas.

De acordo com a distribuição *a priori* normal assumida para o vetor de efeitos genéticos aditivos (**a**), tem-se

$$P\left(\mathbf{a} \mid \sigma_a^2\right) \propto \left(\frac{1}{\sigma_a^2}\right)^{\frac{q}{2}} \exp\left\{-\frac{1}{\sigma_a^2}\left(\mathbf{a}'\mathbf{A}^{-1}\mathbf{a}\right)\right\}.$$

Combinando a função de verossimilhança com as distribuições *a priori* por meio do teorema de Bayes, obtém-se a seguinte distribuição conjunta *a posteriori*

$$\mathrm{P}\left(\boldsymbol{\beta}, \mathbf{a}, \sigma_a^2, \sigma_e^2 \mid \mathbf{y}\right) \propto \mathrm{P}\left(\mathrm{y} \mid \boldsymbol{\beta}, \mathbf{a}, \sigma_a^2, \sigma_e^2\right) \mathrm{P}(\boldsymbol{\beta}) \mathrm{P}\left(\mathbf{a} \mid \sigma_a^2\right) P\left(\sigma_a^2 \mid S_a^2, \nu_a\right) P\left(\sigma_e^2 \mid S_e^2, \nu_e\right),$$

ou seja,

$$\mathrm{P}\left(\boldsymbol{\beta}, \mathbf{a}, \sigma_a^2, \sigma_e^2 \mid \mathbf{y}\right) \propto \left(\frac{1}{\sigma_e^2}\right) \exp\left\{-\frac{1}{2\sigma_e^2}^{\left(\frac{\nu_{a+q+2}}{2}\right)} \left[\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{a}\right)'\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{a}\right)+\nu_e s_e^2\right]\right\}$$

$$\times\left(\frac{1}{\sigma_a^2}\right) \exp\left\{-\frac{1}{2\sigma_a^2}\left(\frac{\nu_{a+q+2}}{2}\right)\left[\left(\mathbf{a}'\mathbf{A}^{-1}\mathbf{a}\right)+\nu_a s_a^2\right]\right\}$$

A inferência para cada parâmetro é feita por meio das distribuições marginais *a posteriori*, que devido ao fato de não apresentarem soluções analíticas via integração múltipla, precisam ser obtidas indiretamente por meio de algoritmos MCMC. Como mencionado anteriormente, para utilização destes algoritmos é necessário obter as distribuições condicionais completas *a posteriori* para cada parâmetro. Esta distribuição

para um dado parâmetro é obtida simplesmente ao assumir que os demais parâmetros são constantes, permitindo assim obter uma distribuição condicional. Se esta distribuição for conhecida (ou seja, representa uma densidade de probabilidade já estudada), utiliza-se o algoritmo Gibbs Samper; caso contrário, se obtida uma distribuição para qual ainda não se tem uma densidade de probabilidade conhecida, utiliza-se o algoritmo Metropolis-Hastings. De acordo com a distribuição conjunta *a posteriori* obtida para os parâmetros do modelo misto de Henderson, tem-se, respectivamente, as seguintes distribuições condicionais completas *a posteriori* para σ^2_a, σ^2_e, $\boldsymbol{\beta}$ e $\mathbf{a}$ (Sorensen e Gianola, 2002)

$$P\left(\sigma_a^2 \mid \sigma_e^2, \mathbf{a}, \boldsymbol{\beta}, \mathbf{y}\right) \propto \left(\frac{1}{\sigma_a^2}\right)^{\left(\frac{v_a+q+2}{2}\right)} \exp\left\{-\frac{1}{2\sigma_a^2}\left[\left(\mathbf{a}'\mathbf{A}^{-1}\mathbf{a}\right)+v_a s_a^2\right]\right\}$$

ou seja,

$$\sigma_a^2 \mid \sigma_e^2, \mathbf{a}, \boldsymbol{\beta}, \mathbf{y} \sim \chi^{-2}\left\{\frac{v_a+q}{2}, \frac{\mathbf{a}'\mathbf{A}^{-1}\mathbf{a}+v_a s_a^2}{2}\right\};$$

$$P\left(\sigma_e^2 \mid \sigma_a^2, \mathbf{a}, \boldsymbol{\beta}, \mathbf{y}\right) \propto \left(\frac{1}{\sigma_e^2}\right)^{\left(\frac{v_a+q+2}{2}\right)} \exp\left\{-\frac{1}{2\sigma_e^2}\left[\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u}\right)'\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u}\right)+v_e s_e^2\right]\right\}$$

ou seja,

$$\sigma_e^2 \mid \sigma_a^2, \mathbf{a}, \boldsymbol{\beta}, \mathbf{y} \sim \chi^{-2}\left(\frac{v_e+n}{2}, \frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u})+v_e s_e^2}{2}\right);$$

$$P\left(\boldsymbol{\beta} \mid \sigma_a^2, \mathbf{a}, \sigma_e^2, \mathbf{y}\right) \propto \left(\frac{1}{\sigma_e^2}\right)^{\left(\frac{p}{\sigma_e^2}\right)} \exp\left\{-\frac{1}{2\sigma_e^2}\left[\boldsymbol{\beta}-\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\left(\mathbf{y}-\mathbf{Z}\mathbf{a}\right)\right]\left(\frac{\mathbf{X}'\mathbf{X}}{\sigma_e^2}\right)\left[\boldsymbol{\beta}-\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\left(\mathbf{y}-\mathbf{Z}\mathbf{a}\right)\right]\right\},$$

ou seja,

$$\boldsymbol{\beta} \mid \sigma_a^2, \mathbf{a}, \sigma_e^2, \mathbf{y} \sim N\left(\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\left(\mathbf{y}-\mathbf{Z}\mathbf{u}\right), \left(\mathbf{X}'\mathbf{X}\right)^{-1}\sigma_e^2\right);$$

$$P\left(\mathbf{a} \mid \sigma_e^2, \sigma_a^2, \boldsymbol{\beta}, \mathbf{y}\right) \propto \left(\frac{1}{\sigma_e^2}\right)^{\left(\frac{q}{2}\right)} \exp\left\{-\frac{1}{2}\left[\mathbf{a}-\left(\mathbf{Z}'\mathbf{Z}+\mathbf{A}^{-1}\mathrm{K}\right)^{-1}\mathbf{Z}'\left(\mathbf{y}-\boldsymbol{\beta}\mathbf{X}\right)\right]'\left(\frac{\mathbf{Z}'\mathbf{Z}+\mathbf{A}^{-1}\mathrm{K}}{\sigma_e^2}\right)\left[\mathbf{a}-\left(\mathbf{Z}'\mathbf{Z}+\mathbf{A}^{-1}\mathrm{K}\right)^{-1}\mathbf{Z}'\left(\mathbf{y}-\boldsymbol{\beta}\mathbf{X}\right)\right]\right\}$$

,

ou seja,

$$\mathbf{a} \mid \sigma_e^2, \sigma_a^2, \boldsymbol{\beta}, \mathbf{y} \sim N\left(\left(\mathbf{Z}'\mathbf{Z}+\mathbf{A}^{-1}\mathrm{K}\right)^{-1}\mathbf{Z}'\left(\mathbf{y}-\boldsymbol{\beta}\mathbf{X}\right), \left(\mathbf{Z}'\mathbf{Z}+\mathbf{A}^{-1}\mathrm{K}\right)^{-1}\sigma_e^2\right), \text{ sendo } \mathrm{K}= \sigma^2_e /\sigma^2_a.$$

Uma vez que as distribuições condicionais completas *a posteriori* constituem distribuições de probabilidade conhecidas (qui-quadrado invertida para σ^2_a e σ^2_e; e normal

multivariada para **β** e **a**), utiliza-se então o algoritmo Gibbs Sampler para a obtenção das distribuições marginais *a posteriori*. Este algoritmo é um processo iterativo que constitui em gerar valores aleatórios das distribuições condicionais completas *a posteriori* deduzidas anteriormente. Em resumo, este algoritmo pode ser apresentado da seguinte forma (sendo os parâmetros denotados por θ os componentes de variância σ^2_a e σ^2_e, e os vetores de efeitos sistemáticos e genéticos aditivos, **β** e **a**):

I – Inicialize o contador de iterações (t) da cadeia MCMC, ou seja, faça t=0;

II – Especifique os valores iniciais para cada parâmetro para t=0, $\boldsymbol{\theta} = \left(\theta_1^0, \theta_2^0, ..., \theta_p^0\right)$;

III – Obtenha um novo valor de $\boldsymbol{\theta}^t = \left(\theta_1^t, \theta_2^t, ..., \theta_p^t\right)$ a partir de $\boldsymbol{\theta}^{t+1}$ através da geração sucessiva de valores aleatórios a partir das distribuições condicionais completas *a posteriori*;

$$\theta_1^t \sim P\left(\theta_1 \middle| \theta_2^{t-1}, \theta_3^{t-1}, ..., \theta_p^{t-1}, \mathbf{y}\right)$$
$$\theta_2^t \sim P\left(\theta_2 \middle| \theta_2^{t}, \theta_3^{t-1}, ..., \theta_p^{t-1}, \mathbf{y}\right)$$
$$\vdots$$
$$\theta_p^t \sim P\left(\theta_p \middle| \theta_1^{t}, \theta_2^{t}, ..., \theta_{p-1}^{t}, \mathbf{y}\right)$$

IV - Incremente o contador de t para $t+1$ e retorne ao passo II até obter convergência.

Como os algoritmos MCMC são processos iterativos, a avaliação da convergência se faz necessária. Dentre os métodos mais utilizados para avaliação da convergências das cadeias MCMC destacam-se o Heidelberger e Welch (1983), Geweke (1992) e Raftery e Lewis (1992). Todos estes métodos estão implementados no pacote *boa* do software R.

Uma vez constatada a convergência, os resultados da análise Bayesiana são apresentados como uma cadeia de valores gerados em cada iteração para cada um dos parâmetros. Estes valores representam então amostras das distribuições marginais *a posteriori* (as quais são impossíveis de serem obtidas analiticamente via integrais). As estimativas pontuais para cada parâmetro podem ser representadas pela, média, moda ou mediana destas distribuições marginais. A inferência intervalar pode ser obtida diretamente por meio dos quantis destas distribuições, por exemplo ao se trabalhar com os quantis 2,5 e 97,5% tem-se, respectivamente, os limites inferiores e superiores dos intervalos de credibilidade.

Sendo a herdabilidade (h^2) uma função direta dos componentes de variância, valores de h^2 podem ser obtidos a cada iteração por meio dos valores gerados para os componentes de variância. Assim, o valor de h^2 na iteração t é dado por: $h^{2\,(t)} = \sigma^2_a{}^{(t)} / (\sigma^2_a{}^{(t)} + \sigma^2_e{}^{(t)})$. Este procedimento permite obter uma distribuição marginal *a posteriori* para h^2, de forma que estimações intervalares podem ser realizadas diretamente por meio do cálculo dos intervalos de credibilidade. Tais estimações são impraticáveis, ou

assumem aproximações pouco precisas quando se trabalha via abordagem frequentista (Máxima Verossimilhança ou Máxima Verossimilhança Restrita).

A inferência Bayesiana para o modelo misto de Henderson apresentada de forma detalhada anteriormente, bem como para outros modelos mais complexos (multi-característicos e regressão aleatória) está implementada em diferentes softwares livres tais como MTGSAM, GIBBSF90, DMU, e pacote MCMCglmm do R, dentre outros.

Análise de Dados Genômicos

O melhoramento genético animal tradicional possui uma história de sucesso contribuindo fortemente para o aumento da produtividade das espécies de interesse zootécnico. Apesar do sucesso, a muitos anos os pesquisadores possuem o desejo de incorporar informações moleculares as avaliações genéticas e compreender melhor os mecanismos genéticos que contribuem para a variação das características quantitativas. No ano de 2001, utilizando dados simulados, Meuwissen et al. (2001) apresentaram métodos para a utilização de marcadores moleculares para a predição de valores genéticos, a qual posteriormente passou a ser chamada de seleção genômica (SG).

O rápido desenvolvimento das técnicas de biologia molecular possibilitou a genotipagem em larga escala para grande número de animais e marcadores do tipo polimorfismo de base única (SNP) a custos relativamente baixos. Essas evoluções das técnicas moleculares permitiram que os métodos propostos por Meuwissen et al. (2001) e métodos desenvolvidos por vários outros grupos de pesquisa passassem a ser testados utilizando dados reais.

Atualmente, a seleção genômica tem sido incorporada a maioria dos programas de melhoramento por se tratar de uma metodologia que permite predizer valores genéticos mais acurados em comparação com os métodos tradicionais para indivíduos jovens ou que não possuem fenótipo. Sendo que, a SG passa a ser ainda mais atraente para espécies que possuem longo intervalo de geração, uma vez que ela permite reduzir esse intervalo e aumentar a acurácia, o que resulta em aumento do ganho genético. Além disso, a SG é vantajosa na seleção de características limitadas ao sexo, características difíceis de mensurar ou que exigem o abate do animal, e características que são mensuradas em idades avançadas (Hayes e Goddard, 2010).

Os métodos baseados nas equações de modelos mistos são os mais utilizados na prática para a implementação da seleção genômica, por utilizarem a mesma estrutura do método tradicional, facilitando a análise e interpretação dos resultados e pela facilidade em implementar modelos mais complicados como multi-característicos e modelos de limiar, uma vez que utiliza toda a estrutura das equações de modelos mistos com uma matriz de parentesco genômica.

BLUP Genômico (GBLUP)

O método GBLUP foi introduzido por VanRaden (2008) utiliza toda teoria do método tradicional de obtenção do BLUP por meio das equações de modelo misto, porém utilizando uma matriz de parentesco (G*) construída com dados de marcadores em substituição a matriz de parentesco baseada em pedigree. A utilização da matriz de parentesco genômica é vantajosa por possibilitar a correção de erros de pedigree e solucionar os problemas de pedigrees incompletos. Além disso, a matriz de parentesco baseada em pedigree é calculada como o valor médio esperado, ou seja, não considera a segregação mendeliana o que leva a super ou subestimação do parentesco.

Ao considerar o seguinte modelo linear misto

$$\mathbf{y = X\beta + Za + e}$$

em que $\mathbf{y}$ é vetor de observações fenotípicas dos animais; $\mathbf{X}$, matriz de incidência de efeitos fixos; β, vetor de efeitos fixos; $\mathbf{Z}$, matriz de incidência dos valores genéticos dos animais; $\mathbf{a}$, vetor de valores genéticos dos animais; $\mathbf{e}$, vetor de erros aleatórios. Para predição de $\mathbf{a}$ utilizando o método GBLUP em modelo unicaracterístico a notação matricial é dada por

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z + G^{*-1}\dfrac{\sigma_e^2}{\sigma_a^2}} \end{bmatrix} \begin{bmatrix} \beta^0 \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \end{bmatrix}$$

em que, $\mathbf{G^*}$ é a matriz de parentesco genômica; σ_e^2, a variância residual e σ_a^2 a variância genéticas aditiva.

Seguindo o método proposto por VanRaden (2008) para o cálculo da matriz genômica, constrói-se uma matriz $\mathbf{M}$ que específica os alelos dos indivíduos para cada marcador, sendo, portanto, uma matriz nxm, em que n é o número de indivíduos e m o número de marcadores. A matriz $\mathbf{M}$ é composta pelos genótipos codificados como -1, 0

e 1 para o homozigoto, heterozigoto e o outro homozigoto. Nesse método é realizada uma correção para a frequência do alelo menor, a qual faz com que a soma dos coeficientes de parentesco dos animais para um marcador seja igual a zero.

Seja a frequência do alelo menor no locus *i* igual a p_i, constrói-se uma matriz **P** a qual possui as frequências alélicas expressas como diferença de 0,5 e multiplicada por dois, ou seja, a coluna *i* da matriz **P** é dada por $2(p_i - 0,5)$. É então calculada uma matriz $\mathbf{W = M - P}$, essa correção faz com que maior peso seja dado aos alelos raros no cálculo do parentesco e também faz com que o coeficiente de endogamia seja maior se o indivíduo é homozigoto para alelos raros (VanRanden, 2008). A matriz de parentesco genômica é então obtida por $\mathbf{G^* = W'W} \Big/ \left[2\sum p_i \left(1 - p_i\right) \right]$. A divisão por $2\sum p_i \left(1 - p_i\right)$ faz com que a matriz **G*** seja colocada na mesma escala da matriz de parentesco tradicional (**A**).

BLUP Genômico em Passo Único (ssGBLUP)

Os métodos de seleção genômica não consideram o fato de que na maioria dos casos apenas uma subpopulação não aleatória será genotipada e que existirão um número considerável de animais não genotipados mas que possuem informações fenotípicas. Negligenciar as informações dos animais não genotipados pode levar a predições menos acuradas e possivelmente viesadas devido a seleção (VanRaden et al., 2009a; VanRaden et al., 2009b).

O ssGBLUP proposto por Legarra et al. (2009) contorna esses problemas uma vez que permite avaliar simultaneamente animais com e sem genótipo. O método consiste basicamente na substituição da matriz genômica por uma matriz **H**, a qual combina as informações da matriz de parentesco baseada em pedigree (**A**) com a matriz de parentesco genômica (**G***) dos animais genotipados.

O considerar o modelo linear misto apresentado anteriormente para predição de $\mathbf{a}$ utilizando o método ssGBLUP em modelo unicaracterístico, a notação matricial é dada por

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z + H^{-1}} \frac{\sigma_e^2}{\sigma_a^2} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}^0 \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \end{bmatrix}$$

em que, $\mathbf{H}$ é uma matriz de parentesco que combina informações genômicas (**G***) e de pedigree (**A**). A matriz **H** é dada por $\mathbf{H = A + A_{\Delta}}$,

em que, $\mathbf{A} = \begin{bmatrix} \mathbf{A_{11}} & \mathbf{A_{12}} \\ \mathbf{A_{21}} & \mathbf{A_{22}} \end{bmatrix}$ e $\mathbf{A_{\Delta}} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G^{*}\text{-}A_{22}} \end{bmatrix}$, sendo que os subscritos 1 e 2 representam indivíduos não genotipados e genotipados, respectivamente. Sendo assim, a matriz **H** é dada por $\mathbf{H} = \begin{bmatrix} \mathbf{A_{11}} & \mathbf{A_{12}} \\ \mathbf{A_{21}} & \mathbf{G^{**}} \end{bmatrix}$.

Aguilar et al. (2010) mostraram que a utilização de diferentes frequências alélicas para a construção **G**** resulta em diferentes acurácias de predição, o que de acordo com Forni et al. (2011) pode ter ocorrido devido a diferenças na escala de **G**** e **A**. Segundo esses mesmos autores a compatibilidade das matrizes pode ser obtida utilizando as frequências alélicas observadas e mudando a escala da matriz de parentesco genômica de modo que a média dos elementos da diagonal seja igual a um. Forni et al. (2011) mostraram que a utilização de uma matriz normalizada (**GN**) produz estimativas realísticas da variância aditiva e não inflaciona a acurácia dos valores genéticos. A matriz **GN** é dada por

$$\mathbf{GN} = \frac{\mathbf{WW'}}{\left\{\mathbf{trace}\left[\mathbf{WW'}\right]\right\}/\mathbf{n}}$$

em que, **W** é a matriz descrita anteriormente e *n* o número total de indivíduos genotipados.

Estrutura de Dados na Avaliação Genética

A predição de valores genéticos com maior acurácia depende, dentre outros fatores, da estrutura dos dados. A importância da estrutura de dados remete-se a disponibilidade de dados consistentes com rigoroso controle de qualidade no processo de coleta e edição; a correta identificação dos fatores não genéticos que influenciam diretamente as características estudadas; a formação correta dos grupos de contemporâneos (GC) e o grau de conexidade entre os GC (Oliveira, 1995; Mathur et al., 2002; Roso et al., 2004; Fouilloux et al., 2006).

Nas avaliações genéticas há uma série de fatores não genéticos inerentes a estrutura dos dados como importantes fontes de variação que devem ser identificados corretamente e considerados. Entre os principais fatores pode ser destacado o rebanho, ano de nascimento, mês de nascimento, tipo de manejo a que estão submetidos os animais,

o sexo e outros fatores específicos de cada característica avaliada (Oliveira, 1995). Portanto, por definição, os grupos de contemporâneos (GC) permitem a remoção dos efeitos dos fatores não genéticos para que os animais expressem seus fenótipos nas mesmas condições ambientais (Oliveira, 1995; Cobuci et al., 2006).

Os GC são formados na edição dos dados a partir das informações (fatores não genéticos) contidas no conjunto de dados. Embora seja de responsabilidade do melhorista, a pessoa responsável pelas decisões do manejo na propriedade pode facilitar e aumentar a eficiência da formação dos grupos de contemporâneos, identificando corretamente os grupos de animais que são submetidos ao mesmo manejo ou que estão colocados em um mesmo lote desde o nascimento até o momento que foram coletadas as medidas (Oliveira, 1995).

A formação de grupos de contemporâneos (GC) apresenta um paradoxo conceitual quanto ao tamanho do grupo de animais e a definição das condições de ambientes homogêneos comuns a esse mesmo grupo. O agrupamento dos animais que atende a maximização da homogeneidade dentro do GC remete ao número reduzido de amostra que pode interferir diretamente na acurácia dos valores genéticos e na maior perda de registro na consistência dos dados, portanto, não existe uma fórmula exata para definir a divisão dos grupos que fornecerá o melhor resultado, mas deve-se buscar um ponto de equilíbrio (Cobuci et al., 2006). Os erros na identificação dos GC induzem a decisões incorretas em função de valores genéticos superestimados dos animais que de alguma maneira foram favorecidos diante das condições de ambientes ou subestimados para animais que foram privados dessas condições para expressar seu potencial genético (Ferraz e Eler, 1998).

Na expectativa de identificar esses dados discrepantes, como desempenhos diferenciados devido a fatores não genéticos dentro dos GC são aplicados critérios para remoção de *outliers* dentro de cada GC. O desvio-padrão é uma alternativa para remover os *outliers*.

Ao estudar critérios para formação de GC em bovinos da raça Nelore, Silva (2015) definiu nove estruturas de dados com base em três níveis de desvios-padrão em relação à média para remoção de *outliers* (±2, ±2,5 e ±3,5) e número mínimo de animais por GC (3, 7 e 15). A estrutura de dados com 3,5 desvios-padrão e mínimo de 15 animais por GC apresentou maior variância genética aditiva (82,65±2,93) e a menor (60,23±1,96) foi obtida com 2 desvios-padrão e mínimo de 3 animais por GC. A remoção de *outliers* com 2 desvios-padrão e GC com mínimo de 3 animais apresentou menores estimativas de

variância genética aditiva e herdabilidade. Maiores ganhos genéticos foram obtidos ao adotar 2,5 e 3,5 desvios-padrão para remoção de *outliers* e mínimo de 15 animais por GC. Portanto, estruturas com 2 desvios-padrão para remoção de *outliers* e/ou mínimo de 3 animais por GC não seriam recomendadas para avaliação genética de bovinos da raça Nelore.

A eliminação de dados ocorre devido a formação incorreta dos GC relacionados aos animais que não são contemporâneos e caracterizados como animais isolados. Esses dados não devem constar nas análises e quando constam não fornecem informações úteis para o progresso genético. Isto reforça a necessidade dos devidos cuidados para a perfeita formação dos GC (Cobuci et al., 2006).

Outro problema relacionado a estruturação dos dados é a baixa conexidade de dados entre os GC. Isso ocorre quando as populações são isoladas ou semi-isoladas geneticamente em razão de pequena troca de material genético entre populações. O uso de rebanhos fechados, a baixa intensidade do uso de inseminação artificial e o uso de matrizes de parentesco esparsas são algumas das principais causas da baixa conexidade de dados (Carneiro et al., 2001).

A conexidade de dados pode ser definida como uma medida estatística por influenciar diretamente a acurácia de predição. Neste caso, a conexidade entre os dados é definida em termos da correlação entre os efeitos aleatórios relativos às observações efetuadas em diferentes níveis de efeitos fixos. Se a correlação existe as comparações são passíveis de predição não-viesada, mas a variância do erro de predição (PEV) destas comparações é afetada pelo grau de conexidade dos dados. Nos modelos fixos, a conexidade assume caráter descontínuo, ou seja, os dados são conectados (1) ou não (0), enquanto nos modelos mistos a conexidade assume caráter contínuo, variando de 0 a 100%, (Martins, 1994).

A relação genética existente entre os GC de animais avaliados é determinante no aumento do grau de conexidade, mas se dois GC de animais não têm relação genética, ambos podem ser conectados através do envio de material genético para um terceiro grupo, o que aumenta o grau de conexidade entre os rebanhos (Mathur et al., 1999).

Dentre as metodologias para quantificar os níveis de conexidade entre os GC descritas na literatura destacam-se: Médias das relações genéticas (MRGA) (Banos e Cady, 1988); Índice de conexidade (IC) e Índice geral de conexidade (IGC) (Foulley et al., 1992); Coeficiente de determinação (r^2) (Laloë, 1993); PEV das diferenças em valores genéticos entre os animais (PEV(x)), Fluxo de genes (FG), Variância da deriva genética

(VDG) e Variância das diferenças entre os efeitos de GC (Kennedy e Trus, 1993); Número total de laços genéticos diretos entre GC (NTLGD) (Fries e Roso, 1997); Grau de conexidade (GrC) (Mathur et al., 2002); e Correlação entre as PEV dos valores genéticos (rij) (Lewis e Simm, 2000).

Kennedy e Trus (1993) afirmaram que a conexidade de dados é motivo de preocupação para os criadores na tomada de decisões na seleção entre os animais em diferentes rebanhos. Os valores genéticos dos animais provenientes de rebanhos conectados podem ser considerados para todos GC conectados avaliados, enquanto que os valores genéticos de animais provenientes de grupos desconectados devem ser usados somente dentro de rebanhos ou dotados de um aviso que as comparações entre rebanhos mal conectados podem ser viesadas (Tarrés et al., 2010).

O grau de conexidade dos dados influencia a acurácia das avaliações genéticas que envolvem diferentes rebanhos, regiões ou países. Enfim, o problema básico da baixa conexidade dos dados é que as funções lineares de efeitos fixos não são estimáveis, ou são viesadas, e a predição de efeitos aleatórios é de baixa acurácia (Carneiro et al., 2001).

Considerações Finais

O Brasil é um dos maiores produtores mundiais de carne bovina, de frango e de suínos, além de ser também um dos maiores exportadores. A produção de leite e de ovos no país é suficiente para abastecer toda a demanda interna da população. Os índices de produção e de produtividade dos rebanhos nacionais se equiparam aos dos países da Europa e da América do Norte, o que torna o país altamente competitivo na agropecuária mundial. Esse aumento na eficiência de produção dos rebanhos brasileiros deve-se à melhoria de uma série de fatores, como manejo, nutrição, pastagens, reprodução, sanidade e, principalmente, melhoramento genético.

Há enormes desafios na pecuária nacional, dados os diversos sistemas de produção adotados pelos criadores, especialmente na área de ruminantes (bovinos de corte e leite, caprinos, ovinos, bubalinos, etc.). No entanto, o melhoramento genético pode contribuir, substancialmente, para o equacionamento das dificuldades de produzir, em razão dessa grande variação, nos sistemas de produção, utilizando, por exemplo, sistemas de cruzamentos que atendam a ambientes de criação específicos. Outra alternativa seria a inclusão da interação genótipo x ambiente ou heterogeneidade de variância nas avaliações genéticas, com vistas a identificar genótipos ou grupos genéticos que sejam mais produtivos em determinados sistemas de produção.

A avaliação genética de populações multirraciais pode ser também uma alternativa viável, visto que existem metodologias adequadas para avaliar essas populações compostas por animais de raça pura e seus cruzamentos.

Algumas características, tais como ganho de peso em aves, em bovinos de corte e em suínos; espessura de toucinho e tamanho de leitegada em suínos; produção de leite em algumas raças de bovinos de leite; dentre outras, já atingiram os limites de seleção ou de produção. No entanto, as altas intensidades de seleção, adotadas ao longo dos anos para essas características, resultaram em respostas correlacionadas desfavoráveis a uma série de outras características, como, por exemplo, qualidade de carne em suínos e aves; leitegadas desuniformes em suínos; longevidade em bovinos de leite; e, principalmente, a características de conformação relacionadas com vigor físico dos animais. Nesse contexto, ações que visem à obtenção de novos fenótipos têm sido adotadas, visto que há carência de observações fenotípicas altamente especializadas.

Um aspecto relevante nas avaliações genéticas é a estrutura de dados. Alguns fatores são essenciais para que estes sejam consistentes, como rigoroso controle de qualidade no processo de coleta e edição; correta identificação dos fatores não genéticos que influenciam, diretamente, as características estudadas; formação correta dos grupos de contemporâneos; e alto grau de conexidade entre estes. Isso tem sido um grande gargalo na avaliação genética de bovinos de leite e de corte no Brasil.

Quando se trata de grande volume de dados (por exemplo, registros de peso corporal de bovinos da raça Nelore, no Brasil), há necessidade de hardwares, softwares e metodologias e procedimentos estatísticos adequados. Dessa forma, são necessários *workstations* com grande capacidade de memória, softwares robustos e modelos parcimoniosos com uso de inferência Bayesiana.

A seleção genômica deverá contribuir, significativamente, para obtenção de maiores ganhos genéticos em algumas características, não só pelo aumento da acurácia na predição dos valores genéticos, mas também pelo aumento da intensidade de seleção e pela redução do intervalo de gerações. Assim, novos impulsos poderiam ser dados aos programas de melhoramento genético animal, uma vez que características difíceis de mensurar, como rendimento de carne magra, qualidade de carne e resistência a doenças; limitadas ao sexo, como produção de leite, produção de ovos e taxa de ovulação; obtidas somente de animais adultos, como idade ao primeiro parto, intervalo de partos e tamanho de leitegada; e de elevado custo de mensuração, como consumo de alimentos e eficiência alimentar, serão devidamente avaliadas mediante seleção genômica.

Referências

AGUILAR, I.; MISZTAL, I.; JOHNSON, D.L.; LEGARRA, A.; TSURUTA, S.; LAWLOR, T.J. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. **Journal of Dairy Science**, v. 93, p. 743–752, 2010.

ANDERSON, R.D. Variance components, In: QUAAS, R.L.; ANDERSON, R.D., GILMOUR, A.R. **Use of mixed models for prediction and for estimation of (co)variance components**. Armidale: University of New England-AGBU, p.77-145, 1984.

ANDONOV, S.; ØDEGÅRD, J.; SVENDSEN, M.; ÅDNØY, T.; VEGARA, M.; KLEMETSDAL, G. Comparison of random regression and repeatability models to predict breeding values from test-day records of Norwegian goats. **Journal of Dairy Science**, v. 96, p. 1834-1843, 2013.

BANOS, G.; CADY, R. A. Genetic relationship between the United States and Canadian Holstein bull populations. **Journal of Dairy Science,** v. 71, n. 5, p. 1346-1354, 1998.

BOLDMAN, K.G.; KRIESE, L.A.; VAN VLECK, L.D.; VAN TASSEL, C. P.; KACHMAN, S.D. **A Manual for Use of MTDFREML. A Set of Programs to Obtain Estimates of Variances and Covariances [DRAFT]**. Lincoln: USDA/ARS, 1995. 120p.

CARDOSO, F.; GOMES, C.; SOLLERO, B.; OLIVEIRA, M.; ROSO, V.; PICCOLI, M.; HIGA, R.; YOKOO, M.; CAETANO, A.; AGUILAR, I. Genomic prediction for tick resistance in Braford and Hereford cattle. **Journal of Animal Science**, v. 93, p. 2693-2705, 2015.

CARDOSO, F. F.; TEMPELMAN, R. J. Hierarchical Bayes multiple-breed inference with an application to genetic evaluation of a Nelore-Hereford population. **Journal of Animal Science**, v. 82, n. 6, p. 1589-1601, 2004.

CARNEIRO, A. P. S.; TORRES, R. A.; EUCLYDES, R. F.; SILVA, M. A.; LOPES, P. S.; CARNEIRO, P. L. S.; TORRES FILHO, R. A. Efeito da conexidade de dados sobre a acurácia dos testes de progênie e performance. **Revista Brasileira de Zootecnia**, v.30, n. 2, p. 342-347, 2001.

CHRISTENSEN, O. F.; MADSEN, P.; NIELSEN, B.; SU, G. Genomic evaluation of both purebred and crossbred performances. **Genetics Selection Evolution,** v. 46, n. 1, p. 1, 2014.

COBUCI, J. A.; ABREU, U. G. P.; TORRES, R. A. **Formação de Grupos de contemporâneos em Bovinos de Corte.** Corumbá, MS: Embrapa Pantanal, 2006. 27p.

DE BOOR, C.; MATHÉMATICIEN, E.-U. **A practical guide to splines**. Springer-Verlag New York, 1978.

DICKERSON, G. E. Inbreeding and heterosis in animals. **Journal of Animal Science**, v. 1973, n. Symposium, p. 54-77, 1973.

ELZO, M. A.; BORJAS, A. Perspectivas da avaliação genética multirracial em bovinos no Brasil. **Ciência Animal Brasileira**, v. 5, n. 4, p. 171-185, 2004.

ELZO, M. Covariances among sire by breed group of dam interaction effects in multibreed sire evaluation procedures. **Journal of Animal Science**, v. 68, n. 12, p. 4079-4099, 1990.

ELZO, M.; FAMULA, T. Multibreed sire evaluation procedures within a country. **Journal of Animal Science**, v. 60, n. 4, p. 942-952, 1985.

ELZO, M. Considerations for the genetic evaluation of straightbred and crossbred bulls in large multibreed populations. In: PROC. SYMP. WRCC-100 MEETING, Brainerd, MN, 1995. **Proceedings...** Brainerd, MN, 1995. p. 1-21.

FERRAZ, J. B. S.; ELER, J. P. Qualidade dos dados coletados. In: SIMPÓSIO NACIONAL DA SOCIEDADE BRASILEIRA DE MELHORAMENTO ANIMAL, 1998. Uberaba, MG, **Anais...** Uberaba: Sociedade Brasileira de Melhoramento Animal, 1998., p. 265-269.

FLORES, E.; WERF, J. Random regression test day models to estimate genetic parameters for milk yield and milk components in Philippine dairy buffaloes. **Journal of Animal Breeding and Genetics**, v. 132, p. 289-300, 2015.

FORNI, S.; AGUILAR, I.; MISZTAL, I. Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. **Genetics Selection Evolution**, v. 43, n. 1, 2011.

FOUILLOUX, M. N.; MINERY S.; MATTALIA S.; LALOE D. Assessment of connectedness in the international genetic evaluation of Simmental and Montbeliard breeds. **Interbull Bulletin**, n. 35, p. 129-135, 2006.

FOULLEY, J. L., HANOCQ, E., BOICHARD, D. A criterion for measuring the degree of connectedness in linear models of genetic evaluation. **Genetics Selection Evolution**, v. 24, n. 4, p. 315-330, 1992.

FRIES, L. A.; ROSO, V. M. Conectabilidade em avaliações genéticas de gado de corte: uma proposta heurística. In: REUNIÃO DA SOCIEDADE BRASILEIRA ZOOTECNIA, 34., 1997, Juiz de Fora. **Anais...** Juiz de Fora: SBZ, 1997. p 159-161.

GARCÍA-CORTÉS, L. A.; TORO, M. Á. Multibreed analysis by splitting the breeding

values. **Genetics Selection Evolution**, v. 38, n. 6, p. 1, 2006.

GEWEKE, J. **Evaluating the accuracy of sampling-based approaches to calculating posterior moments**. New York: Oxford University, 1992, 631 p.

GIANOLA, D.; FERNANDO, L. Bayesian methods in animal breeding theory. **Journal of Animal Science**, v.63, p.217- 244, 1986.

GILMOUR, A.R., GOGEL, B.J., CULLIS, B.R., THOMPSON, R. **ASReml user guide release 3.0**. VSN International Ltd, Hemel Hempstead, UK. 2009.

GRASER, H.V.; SMITH, S.P; TIER, B. A derivative free approach for estimating variance components in animal models by restricted maximum likelihood. **Journal of Animal Science**, v.64, n.5, p.1362-1370, 1987.

GREGORY, K. E.; CUNDIFF, L. V.; KOCH, R. M. **Composite breeds to use heterosis and breed differences to improve efficiency of beef production**. Springfield: USDA. p. 1-81, 1999. Disponível em:
<http://naldc.nal.usda.gov/naldc/download.xhtml?id=CAT10879520&content=PDF>.

HARTLEY, H.O.; RAO, J.N.K. Maximum likelihood estimation for the mixed analysis of variance model. Biometrika, London, v.54, p.93-108, 1967.

HAYES, B.; GODDARD, M. Genome-wide association and genomic selection in animal breeding. **Genome**, v. 53, p. 876–883, 2010.

HAZEL, L.N. The genetic basis for constructing selection indexes. **Genetics**, v. 28, n. 6, p. 476-490, 1943.

HEIDELBERGER, P.; WELCH, P. Simulation run length control in the presence of an initial transient. **Operation Research**, v. 31, n. 6, p. 1109-44, 1983.

HENDERSON, C.R. Estimation of Variance and Covariance Components. **Biometrics**, v. 9, n. 2, p. 226-252, 1953.

HENDERSON, C.R. Selection index and expected genetic advance. In: NATIONAL ACADEMY OF SCIENCE / NATIONAL RESEARCH COUNCIL – NAS / NRC. **Statistical genetics and plant breeding**. Washington DC. p. 141-163, 1963.

HENDERSON, C.R. Sire evaluation and genetic trends. In: ANIMAL BREEDING GENETIC SYMPOSIUM IN HONOR OF DR. J.L. LUSH, ASAS/ADSA, 1973, Champaign. **Proceedings...** Champaign. p. 10-41, 1973.

HENDERSON, C.R., A simple method for computing the inverse of a relationship matrix. **Biometrics,** v. 32, n. 1, p. 69-83, 1976a.

HENDERSON, C.R.; QUAAS, R.L. Multiple Trait Evaluation Using Relatives' Records. **Journal of Animal Science**, v. 43, p. 1188-1197, 1976b.

HENDERSON, C.R. **Applications of linear models in animal breeding**. University of Guelph. Guelph, CA, 1984. 462p.

HENDERSON, C.R. Recent Developments in Variance and Covariance Estimations. **Journal of Animal Science**, v. 63, n. 1, p. 208-21. 1986.

HENDERSON JR, C. R. Analysis of covariance in the mixed model: higher-level, nonhomogeneous, and random regressions. Biometrics, p. 623-640, 1982.

HILL, W. G.; GODDARD, M. E.; VISSCHER, P. M. Data and theory point to mainly additive genetic variance for complex traits. **PLoS Genetics**, v. 4, n. 2, 2008.

IBÁNÊZ-ESCRICHE, N.; FERNANDO, R. L.; TOOSI, A.; DEKKERS, J. C. Genomic selection of purebreds for crossbred performance. **Genetics Selection Evolution**, v. 41, n. 1, p. 1-10, 2009.

KENNEDY, B. W., TRUS, D. Considerations on genetic connectedness between management units under an animal model. **Journal of Animal Science,** v. 71, n. 9, p. 2341-2352, 1993.

KINGHORN, B. The expression of "recombination loss" in quantitative traits. **Zeitschrift für Tierzüchtung und Züchtungsbiologie**, v. 97, n. 1-4, p. 138-143, 1980.

KIRKPATRICK, M.; LOFSVOLD, D.; BULMER, M. Analysis of the inheritance, selection and evolution of growth trajectories. **Genetics**, v. 124, p. 979-993, 1990.

LAIRD, N. M. & WARE, J. H. Random-effects models for longitudinal data. **Biometrics**, p. 963-974, 1982.

LALOE, D. Precision and information in linear models of genetic evaluation. **Genetics Selection Evolution,** v. 25, p. 557-576, 1993.

LEGARRA, A; AGUILAR, I.; MISZTAL, I. A relationship matrix including full pedigree and genomic information. **Journal of Dairy Science**, v. 92, p. 4656–4663, 2009.

LEWIS, R. M.; SIMM, G. Selection strategies in sire referencing schemes in sheep. **Livestock Science**, v. 67, p. 129-141, 2000.

LO, L. L.; FERNANDO, R. L.; CANTET, R. J. C.; GROSSMAN, M. Theory for modelling means and covariances in a two-breed population with dominance inheritance. **Theoretical and Applied Genetics**, v. 90, 1995.

LO, L. L.; FERNANDO, R. L.; GROSSMAN, M. Covariance between relatives in multibreed populations: Additive model. **Theoretical and Applied Genetics**, v. 87, 1993.

LO, L. L.; FERNANDO, R. L.; GROSSMAN, M. Genetic evaluation by BLUP in two-breed terminal crossbreeding systems under dominance inheritance. **Journal of Animal**

Science, v. 75, n. 11, p. 2877-2884, 1997.

LOPES, P.S.; MARTINS, E.N.; SILVA, M.A.; REGAZZI, A.J. **Estimação de componentes de variância**. Viçosa: UFV, 1998. 61p.

MAPHOLI, N.; MAIWASHE, A.; MATIKA, O.; RIGGIO, V.; BISHOP, S.; MACNEIL, M.; BANGA, C.; TAYLOR, J.; DZAMA, K. Genome-wide association study of tick resistance in South African Nguni cattle. **Ticks and Tick-Borne Diseases**, v. 7, p. 487-497, 2016.

MARTINS, E. N. Uso de modelos mistos no melhoramento animal. In: SIMPÓSIO INTERNACIONAL DE PRODUÇÃO de NÃO-RUMINANTES, 31., 1994, Maringá. **Anais...** Maringá: SBZ, 1994. p.41-46.

MARTINS, E.N.; LOPES, P.S.; SILVA, M.A.; TORRES, R.A.A. **Uso de modelos mistos na avaliação genética animal.** Universidade Federal de Viçosa, 1997. 121p (Cadernos didáticos, 18).

MATHUR, P. K., SULLIVAN, B., CHESNAIS, J. **Estimation of the degree connectedness between herds or management groups in the canadian swine population**. Ottawa: Canadian Center for Swine Improvement, 1999. 19 p. Disponível em: <http://www.ccsi.ca/include/docs/connectedness/connectedness-_asds_1998.pdf>. Acesso em 25 de julho de 2014.

MATHUR, P. K.; SULLIVAN, B.; CHESNAIS, J. Measuring connectedness: concept and application to a large industry breeding program. In: World Congress on Genetics Apllied Livestock Production, 7. 2002, Montpellier. **Anais eletrônicos…** Montpellier: INRA, 2002. Disponível em: <http://www.ccsi.ca/include/docs/con-nectedness/connectedness_wcgalp_2002.pdf>. Acesso em 03 de Agosto de 2014.

MENÉNDEZ-BUXADERA, A.; MOLINA, A.; ARREBOLA, F.; GIL, M.; SERRADILLA, J. Random regression analysis of milk yield and milk composition in the first and second lactations of Murciano-Granadina goats. **Journal of Dairy Science**, 93, 2718-2726, 2010.

MEUWISSEN, T.H.; HAYES, B.J.; GODDARD, M.E. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, v. 157, p. 1819–29, 2001.

MEYER, K. **DFREML - Programs to estimate variance components by REML using a dirivative-free algorithm, user notes**. Armidale: University of New England, 1991. 97p.

MEYER, K. Random regression analyses using B-splines to model growth of Australian

Angus cattle. **Genetics Selection Evolution**, v. 37, n. 1, 2005.

MEYER, K. WOMBAT: a tool for mixed model analyses in quantitative genetics by restricted maximum likelihood (REML). **Journal of Zhejiang University Science B,** v. 8, p. 815–821, 2007.

MEYER, K.; CARRICK, M.; DONNELLY, B. Genetic parameters for milk production of Australian beef cows and weaning weight of their calves. **Journal of Animal Science**, v. 72, p. 1155-1165, 1994.

MEYER, K.; HILL, W. G. Estimation of genetic and phenotypic covariance functions for longitudinal or 'repeated'records by restricted maximum likelihood. **Livestock Production Science**, v. 47, p. 185-200, 1997.

MISZTAL, I. 2001. Software disponível em: nce.ads.uga.edu/software. Acessado em 08 de Setembro de 2016.

MRODE, R. A. **Linear models for the prediction of animal breeding values**. Cabi, 2014.

OLIVEIRA, H. N. Grupos de contemporâneos e conectabilidade. In: CURSO SOBRE A AVALIAÇÃO GENÉTICA EM BOVINOS DE CORTE, 1995, Ribeirão Preto. **Anais...** Ribeirão Preto: FPCRC, 1995. p. 1-13.

OLIVEIRA, H. R.; SILVA, F.; SIQUEIRA, O.; SOUZA, N.; JUNQUEIRA, V.; RESENDE, M.; BORQUIS, R.; RODRIGUES, M. Combining different functions to describe milk, fat, and protein yield in goats using Bayesian multiple-trait random regression models. **Journal of Animal Science**, v. 94, p. 1865-1874, 2016.

PATTERSON, H.D.; THOMPSON, R. Recovery of inter-block information when block sizes are unequal. **Biometrik**a, v.58, p.545-554, 1971.

PEREIRA, R.; BIGNARDI, A.; EL FARO, L.; VERNEQUE, R.; VERCESI FILHO, A.; ALBUQUERQUE, L. Random regression models using Legendre polynomials or linear splines for test-day milk yield of dairy Gyr (Bos indicus) cattle. **Journal of Dairy Science**, v. 96, p. 565-574, 2013.

PTAK, E.; SCHAEFFER, L. Use of test day yields for genetic evaluation of dairy sires and cows. **Livestock Production Science**, v. 34, p. 23-34, 1993.

QUAAS, R. L. Computing the diagonal elements and inverse of a large numerator relationship matrix. **Biometrics,** v. 32, n. 4, p. 949-953. 1976.

QUAAS, R. L.; POLLAK, E. J. Mixed Model Methodology for Farm and Ranch Beef Cattle Testing Programs. **Journal of Animal Science**, v. 51, n. 1277-1287, 1980.

RAFTERY, A.E.; LEWIS, S. How many iterations in the Gibbs sampler? In:

BERNARDO, J.M. et al. (Eds.). **Bayesian statistics**. Oxford, USA: University Press, 1992, p. 763-773.

RAO, C.R. **Linear statistical inference and its applications**. New York: John Wiley & Sons, 1965. 522p.

ROSO, V. M.; SCHENKEL F. S.; MILLER S. P. Degree of connectedness among groups of centrally tested beef bulls. **Canadian Journal Animal Science,** v. 84, p. 37-47, 2004.

SCHAEFFER, L. Application of random regression models in animal breeding. **Livestock Production Science**, v. 86, p. 35-45, 2004.

SCHAEFFER, L.; DEKKERS, J. 1994. Random regressions in animal models for test-day production in dairy cattle. **Proceedings...** 5th World Congress of Genetics Applied to Livestock Production.

SILVA, D. A. **Influência da conexidade e da estrutura dos grupos de contemporâneos na avaliação genética de bovinos Nelore via inferência Bayesiana**. 2015. 51 p. Dissertação (Mestrado) – Universidade Federal de Viçosa, Viçosa, MG.

SORENSEN, D.; GIANOLA, D. **Likelihood, Bayesian and MCMC methods in quantitative genetics**. New York, Springer-Verlag, 758 p., 2002.

SORENSEN, D.A.; KENNEDY, B. W. Estimation of Genetic Variances from Unselected and Selected Populations. **Journal of Animal Science**, v. 59, n. 5, p. 1213-1223, 1984

SUZUKI, M.; VAN VLECK, L. D. Heritability and repeatability for milk production traits of Japanese Holsteins from an animal model. **Journal of Dairy Science**, v. 77, p. 583-588, 1994.

TARRÉS, J.; FINA, M.; PIEDRAFITA, J. Connectedness among herds of beef cattle bred under natural servisse. **Genetics Selection Evolution**, v. 42, n. 6, p. 1-9, 2010.

VANRADEN, P.M. Efficient methods to compute genomic predictions. **Journal of Dairy Science**, v. 91, p. 4414–23, 2008.

VANRADEN, P.M.; TOOKER, M.E.; COLE, J.B. Can you believe those genomic evaluations for young bulls? **Journal of Dairy Science**, v. 92, 2009.

VANRADEN, P.M.; VAN TASSELL, C.P.; WIGGANS, G.R.; SONSTEGARD, T.S.; SCHNABEL, R.D.; TAYLOR, J.F.; SCHENKEL, F.S. Invited review: reliability of genomic predictions for North American Holstein bulls. **Journal of Dairy Science**, v. 92, p. 16–24, 2009.

VAN TASSEL, C. P.; CASELLA, G.; POLLAK, E. J. Effects of selection on estimates of variance components using Gibbs sampling and restricted maximum likelihood.

Journal of Dairy Science, v. 78, p. 678 - 692, 1995.

WEGMAN, E. J.; WRIGHT, I. W. Splines in statistics. **Journal of the American Statistical Association**, v. 78, p. 351-365, 1983.

Desafios estatísticos na seleção genômica

Fabyano Fonseca e Silva[1], Marcos Deon Vilela de Resende[2], Camila Ferreira Azevedo[3], José Marcelo Soriano Viana[4] Moysés Nascimento[5] Paulo Sávio Lopes[6]

Introdução

A utilização de painéis de marcadores moleculares de alta densidade (principalmente da classe SNP) simultaneamente aos fenótipos para fins de predição de valores genéticos é o fundamento da seleção genômica (SG). Esta apresenta como principal vantagem o aumento da acurácia na avaliação genética (Meuwissen et al., 2001) de características cuja expressão fenotípica é muito influenciada por fontes de variação não controladas (baixa herdabilidade) e/ou aquelas de alto custo de medição e/ou avaliadas tardiamente nos indivíduos sob seleção.

As características mencionadas podem não seguir necessariamente a distribuição normal (dados categóricos); serem avaliadas ao longo do tempo (dados longitudinais) e sofrerem a influência da interação entre os genótipos e os ambientes (GxE). Embora um rico referencial teórico já esteja consolidado em relação à SG, estudos envolvendo dados categóricos, longitudinais e GxE ainda são escassos na literatura, demandando o desenvolvimento e a aplicação de novas metodologias estatísticas. Estas podem ser vistas como desafios estatísticos na SG, uma vez que modelos exclusivos para SG devem ser adaptados para contemplarem tais dados. Os dados mencionados são comuns nas áreas de melhoramento animal e vegetal, e uma vez superados estes desafios estatísticos, espera-se que a GS seja utilizada de forma ainda mais abrangente nestas áreas.

Exemplos de características discretas na área de melhoramento animal podem ser o tamanho de leitegada (número de leitões) e susceptibilidade ao carrapato (número de carrapatos/animal); enquanto que na área de melhoramento de plantas pode-se pensar no número de frutos e no número de folhas infectadas por um dado agente fitopatológico.

[1]Zootecnista, M.S., D.S. e Professor da Universidade Federal de Viçosa. E-mail: fabyanofonseca@ufv.br
[2]Estatístico e Engenheiro Agrônomo, M.S., D.S. e Pesquisador da Embrapa Floresta. E-mail: marcos.deon@gmail.com
[3]Matemática, M.S., D.S. e Professora da Universidade Federal de Viçosa. E-mail: camila.azevedo@ufv.br
[4]Engenheiro Agrônomo, M.S., D.S. e Professor da Universidade Federal de Viçosa. E-mail: jmsviana@ufv.br
[5]Estatístico, M.S., D.S. e Professor da Universidade Federal de Viçosa. E-mail: moysesnascim@ufv.br
[6]Zootecnista, M.S., D.S. e Professor da Universidade Federal de Viçosa. E-mail: plopesa@ufv.br

Todas estas se caracterizam como contagens discretas, e podem ser descritas, por exemplo, por meio da distribuição Poisson.

Outras características que muitas vezes são consideradas não normais, mas não necessariamente discretas, são aquelas representadas por intervalos de tempo, tal como idade ao primeiro parto e idade ao abate na área animal; e idade ao florescimento na área vegetal. Geralmente este tipo de característica apresenta certa assimetria (motivo pelo qual estas se afastam da condição de normalidade), porém o fato mais importante relacionado a elas é a condição de censura. Se esse intervalo de tempo não pode ser contabilizado, seja pela falta de informações disponíveis (o evento, parto ou florescimento, não ocorreu durante um certo período pré-estabelecido) ou por erros de anotações (trocas de datas de avaliação), esta característica pode ser denominada de censurada. O tratamento estatístico de características censuradas se distingue daqueles amplamente utilizados no melhoramento, dificultando a utilização das mesmas em avaliações genéticas. A opção mais simples é desconsiderar as observações censuradas na análise, o que leva a perda de um grande número de informações e diminui a acurácia de predição, além de mascarar a variabilidade real da característica e as diferenças genéticas entre os indivíduos. Outra opção é usar modelos estatísticos por meio dos quais os valores para as observações censuradas são aproveitados de forma mais efetiva na avaliação genética.

Em relação aos dados longitudinais, os mesmos também são muito utilizados nas áreas animal e vegetal, uma vez que a coleta de dados no decorrer do tempo é uma prática comum no meio agropecuário. Na área animal os principais exemplos são as observações de peso e de produção de leite medidas ao longo do tempo, as quais são geralmente denominadas de curvas de crescimento e de lactação, respectivamente. Já na área vegetal, têm-se principalmente as produções de culturas perenes ao longo dos anos.

A GxE também é de grande interesse para ambas as áreas, visto que a designação dos melhores genótipos para determinados ambientes é especialmente útil em países com grande extensão territorial e variada caracterização edafoclimática, tal como o Brasil. Neste contexto, modelos que contemplam o ranqueamento dos indivíduos por meio de seus méritos genéticos separadamente para cada ambiente, e também suas estabilidades e adaptabilidades, representam uma importante ferramenta estatística para a agropecuária. Dentre estes modelos destacam-se aqueles denominados de normas de reação, os quais são especialmente indicados quando os diferentes ambientes se caracterizam como um gradiente contínuo, tal como temperatura, umidade, e nível nutricional, dentre outros.

Diante do exposto, pretende-se apresentar aspectos estatísticos relacionados com a modelagem de dados não normais, dados censurados, dados longitudinais e de GxE considerando a inclusão de informações genômicas (genótipos de marcadores SNPs). Serão apresentadas e discutidas estratégias para predizer o mérito genômico dos indivíduos via valor genético genômico (*genomic estimated breeding values* - GEBV), bem como efeitos de marcadores SNPs ao longo do genoma para cada um destes tipos de dados. Também serão abordadas particularidades computacionais de cada metodologia via utilização de softwares livres amplamente utilizados na área de genética e melhoramento.

Vale ressaltar que todos os modelos apresentados no presente material consideram apenas o efeito aditivo de marcadores, portanto demais efeitos como dominância e epistasia não serão abordados aqui. Porém, todos estes modelos podem contemplar tais efeitos via matriz de parentesco genômica, ou mesmo diretamente como variáveis independentes aleatórias, principalmente nos modelos Bayesianos de SG. Para o efeito de dominância, estas abordagens já se encontram disponível na literatura de forma detalhada sob o ponto de vista teórico e computacional conforme apresentado por Azevedo et al. (2015).

Dados não normais

O enfoque geral para utilização de dados não normais em modelos de SG é o de modelos lineares generalizados mistos. Tais modelos estão fundamentados na utilização de distribuições específicas para cada tipo de dados considerados, e também na especificação de um efeito aleatório (valor genético individual) e efeitos fixos (efeitos ambientais). Por exemplo, para dados de contagens, geralmente se utiliza a distribuição Poisson, enquanto que para dados binários (presença ou ausência) utiliza-se a distribuição Binomial. Dada a complexidade destes modelos, geralmente os mesmos têm-se sido utilizados por meio da abordagem bayesiana via algoritmos MCMC (*Markov chain Monte Carlo*).

Seguindo a metodologia proposta por Hadfield e Nakagawa (2010), a implementação bayesiana do modelo linear generalizado misto assumindo distribuição de Poisson para os dados é baseada na especificação de uma variável latente (l), de forma que o referido modelo é dado por:

$$\boldsymbol{l}=\mathbf{X\beta}+\mathbf{Za}+\mathbf{e},$$

em que $\boldsymbol{l}$ é o vetor da variável latente para a característica (**y**) estudada, $y_i \sim Poi(\lambda_i = exp(l_i))$, onde i=1,2,...,N (número de observações); λ é o parâmetro canônico desta distribuição (média de Poisson) e *exp* é a função de ligação inversa da distribuição de Poisson. Assim, é postulado que o vetor da variável latente assume uma distribuição normal multivariada dada por: $\boldsymbol{l}|\boldsymbol{\beta},\mathbf{a},\sigma^2_a, \sigma^2_e \sim N(\mathbf{X\beta}+\mathbf{Za}, \sigma^2_e \mathbf{I})$.

As matrizes **X** e **Z** se referem a incidência dos efeitos fixos e aleatório genético aditivo, respectivamente; **β** é o vetor de efeitos fixos, $\boldsymbol{\beta} \sim N(\mathbf{0}, \sigma^2_\beta I)$, sendo σ^2_β um valor assumido como conhecido (geralmente 10^{10}) que representa um conhecimento vago *a priori*; **a** é o vetor do efeito aleatório genético aditivo, para o qual assume-se a seguinte distribuição *a priori*: $\mathbf{a}|\ \sigma^2_a \sim N(\mathbf{0}, \sigma^2_a \mathbf{G})$, sendo **G** a matriz de parentesco genômica aditiva (Van Raden, 2008) entre os indivíduos e σ^2_a a variância genética aditiva. Além disso, assume-se que os componentes de variância $\sigma^2_a \sim \chi^{-2}(\nu_a, S^2_a)$ e $\sigma^2_e \sim \chi^{-2}(\nu_e, S^2_e)$, indicando que estes seguem distribuições de qui-quadrado invertida com parâmetros S^2 (valor esperado para o componente de variância) e ν (grau de confiança depositado no valor esperado S^2).

No modelo em questão é assumido que vetor da variável latente ($\boldsymbol{l}$) é também considerado um parâmetro desconhecido, para o qual não há uma distribuição condicional *a posteriori* reconhecível, implicando na adoção do algoritmo *Metropolis-Hastings* para gerar amostras da distribuição *posterior* para $\boldsymbol{l}$. Para os outros parâmetros (**β**, **a**, σ^2_a, σ^2_e), dado que as distribuições condicionais completas *a posteriori* são conhecidas, o algoritmo *Gibbs sampler* é utilizado.

Para implementação do modelo apresentado, pode-se utilizar a função *MCMCglmm* (Hadfield e Nakagawa, 2010) do pacote homônimo no *software* R por meio da opção pedigree=c("poisson"). Esta função admite a utilização da matriz de parentesco genômica inversa ao usar a opção "ginverse".

As amostras da distribuição marginal *a posteriori* para a herdabilidade são obtidas a partir das estimativas dos componentes de variância gerados em cada iteração (k) dos algoritmos MCMC como segue: $h^2 = \sigma^{2(k)}_a / (\sigma^{2(k)}_a + \sigma^{2(k)}_e)$. Como os algoritmos MCMC são processos iterativos, a avaliação da convergência se faz necessária. Dentre os métodos mais utilizados para avaliação da convergências das cadeias MCMC destacam-se o Heidelberger e Welch (1983), Geweke (1992) e Raftery e Lewis (1992). Todos estes métodos estão implementados no pacote *boa* do software R.

Geralmente ao se trabalhar com dados não normais, é interessante comparar a qualidade do ajuste do modelo considerando distribuições específica, tal como a Poisson

no modelo apresentado, com o modelo assumindo distribuição Normal. Sob o enfoque Bayesiano os modelos podem ser comparados por meio do DIC (*Deviance Information Criterion)* tal como proposto por Spiegelhalter et al. (2002): $\text{DIC} = \text{D}(\bar{\theta}) + 2\text{p}_\text{D}$, em que $\text{D}(\bar{\theta})$ é valor do logaritmo da função de verossimilhança obtido por meio da substituição dos parâmetros pelas suas médias *a posteriori*, $\text{D}(\bar{\theta})$ =-2 log ($\boldsymbol{l}|\mathbf{y},\hat{\boldsymbol{\beta}},\hat{\mathbf{a}},\hat{\sigma}_a^2,\hat{\sigma}_e^2$), e p_D é o número efetivo de parâmetros no modelo. A ideia é que modelos com menores DIC sejam preferidos a modelos com maiores DIC. Para facilitar a interpretação dos valores de DIC em termos de superioridade de um modelo sobre outro, a probabilidade *a posteriori* do modelo pode ser calculada de acordo com a aproximação apresentada por Silva et al. (2011), a qual é dada por: $\text{p}(\text{M}_\text{t} \mid \boldsymbol{l}) = \exp\left(-\Delta_\text{t}/2\right) \Big/ \sum_{\text{t}=1}^{\text{T}} \exp\left(-\Delta_\text{t}/2\right)$, em que $\text{p}(\text{M}_\text{t} \mid \boldsymbol{l})$ probabilidade *a posteriori* do modelo t (t=1,2,...,T, sendo T o número de modelos comprados), Δ_t é diferença entre os valores do DIC do modelo t em relação ao melhor modelo dentre aqueles comprados (neste caso, aquele com o menor DIC), portanto o valor Δ_t para o melhor modelo é sempre igual a zero. Assim, quanto maior o valor desta probabilidade, mais indicado é modelo para descrever os dados em questão.

No modelo considerado, uma vez que se utilizou a matriz de parentesco genômica (**G**), o vetor **â** contém os GEBV para cada indivíduo sob seleção. Com o intuito de fazer predições genômicas para indivíduos não fenotipados, ou seja, para aqueles cujas observações fenotípicas não foram coletadas (população de validação), é necessário apenas assumir estes valores fenotípicos como observações perdidas (*missing values*). Desta forma, via parentesco baseado na identidade por estado entre os marcadores, é possível predizer os GEBVs para tais indivíduos. Outra possibilidade de se inferir sobre GEBVs de indivíduos não fenotipados é estimar o vetor dos efeitos dos marcadores SNPs (**b̂**) indiretamente por meio do vetor de predições (médias *a posteriori*) dos GEBVs (**â**). Uma vez que a utilização da matriz **G** implica no método denominado GBLUP, é possível obter este vetor por meio da seguinte transformação: $\hat{\mathbf{b}} = (\mathbf{M}'\mathbf{M})^{-}\mathbf{M}'\hat{\mathbf{a}}$, em que **M** é matriz de genótipos (contendo 0, 1 e 2, respectivamente para aa, aA e AA) utilizada na construção de **G** (Van Raden, 2008) na qual o número de linhas corresponde ao número de indivíduos fenotipados e genotipados, e o número de colunas corresponde ao número de marcadores utilizados.

O modelo apresentado já foi utilizado com sucesso na área de genômica para descrever as características número de tetas (NT) e tamanho de leitegada (TL) em suínos (Verardo et al., 2016a). No estudo em questão, o modelo Poisson foi mais adequado para descrever TL, enquanto o modelo Gaussiano (Normal) mostrou-se mais adequado para descrever a característica NT. Adicionalmente, dada a natureza iterativa dos algoritmos MCMC utilizados na inferência bayesiana, também foi possível obter distribuições *a posteriori* para o efeito de cada marcador via equação $\hat{\mathbf{b}}^{(k)} = (\mathbf{M}'\mathbf{M})^{-}\mathbf{M}'\hat{\mathbf{a}}^{(k)}$, em que k representa cada iteração (tal como previamente mostrado para obtenção de distribuições marginais *a posteriori* para herdabilidade). Isto implicou na possibilidade de se fazer inferências sobre o efeito de cada marcador, ou seja, de se verificar a significância de cada um via intervalos de credibilidade de 95%. Assim, se o zero está incluído no intervalo o efeito do marcador é não significativo. Além disso, visto que todos os marcadores foram considerados simultaneamente no modelo, não houve a necessidade de se corrigir para múltiplos testes (via critério de Bonferroni ou FDR - *False discover rate*). Em resumo, esta aplicação possibilitou explorar resultados de associação genômica ampla (GWAS) por meio de um modelo utilizado em SG, neste caso o GBLUP, com o diferencial de ser um GBLUP generalizado para distribuição Poisson.

O modelo descrito até o momento pode ser generalizado para descrever simultaneamente duas ou mais características não-normais. Neste caso, tem-se um modelo linear generalizado misto multi-característico. Este já foi utilizado para descrever tamanho de leitegada em diferentes partos (Ventura et al., 2015), de forma que TL em cada parto foi considerada uma característica diferente. Porém nesta aplicação, não se utilizou a matriz **G**, e sim a matriz tradicional de parentesco (**A**).

Dados censurados

Diferentes métodos para o tratamento de dados censurados têm sido propostos. O método de dados incompletos ou faltantes (*missing*) se baseia na restrição e/ou exclusão de informações censuradas. Pode citar como exemplo desta censura as exclusões das observações de dias para o parto, ou seja, fêmeas que não tiveram sucesso em emprenhar durante a estação de monta. Este é o procedimento mais simples, porém como já mencionado anteriormente, geralmente implica em perda acurácia e de precisão na estimação dos parâmetros genéticos.

O Método de simulação (ou imputação) considera que as observações para os dados censurados podem ser geradas (simuladas) a partir de distribuições preditivas, isto,

logicamente, quando se trabalha sob a abordagem bayesiana. Assim, pressupõe-se que os dados censurados seguem uma distribuição normal truncada (neste caso, a distribuição preditiva), e aleatoriamente são simulados dados da característica a serem imputados à cada observação censurada (Korsgaard et al., 2003; Donoghue et al., 2004). Os valores gerados para cada indivíduo leva em consideração os efeitos fixos (por exemplo, grupos de contemporâneos) e genéticos, este último inclusive afetando a matriz de covariância por meio da matriz de parentesco. Sob o enfoque de SG, esta última pode ser a matriz **G** (parentesco genômico) anteriormente mencionada. Este método considera encontra-se disponível nos softwares TM (Legarra et al., 2008) e GIBBSF90 (Misztal et al., 2014), nos quais as informações genômicas podem ser inseridas diretamente como arquivo de dados, ou já disponibilizadas na forma da matriz **G** inversa ($\mathbf{G}^{-1}$), a qual pode ser calculada externamente por outros softwares, tais como R (pacotes *cpgen, synbreed* e *pedigree*, dentre outros) e PreGSF90 (Misztal et al., 2014).

O método de limiar-linear está fundamentado na avaliação conjunta (multi-característica) entre diferentes formas de representação das características com dados censurados, de forma que indivíduos sem informações fenotípicas sejam incluídos na análise considerando seus registros das características como dados faltantes. O principal objetivo desta metodologia consiste em obter estimativas mais acuradas para os GEBVs em decorrência do uso das correlações genéticas existentes entre as diferentes formas de representação das características, diminuindo assim o problema de censura dos dados. Em resumo, para cada indivíduo têm-se duas observações, uma categórica (informando se é censurado ou não, portanto binário, zero ou um) e uma contínua (valor da característica, neste caso apenas para indivíduos não censurados). Assim, tem-se um modelo multi-característico, porém uma das características, a binária, exige a utilização de um modelo de limiar (denominado *threshold*, ou *probit*), enquanto que a outra, a característica contínua observada, é assumida como normal (Urioste et al., 2007; Hou et al., 2009). Tal como o método da simulação, o limiar-linear também pode ser implementado nos softwares TM e GIBBSF90, além do pacote MCMCglmm do R via opção pedigree=c(“threshold”, “poisson”).

O método de análise de sobrevivência (AS) é um dos mais sofisticados paraavaliação genética de características censuradas. Para dados censurados que são caracterizados como contínuos, é definida uma função de risco que representa um modo de quantificar o risco instantâneo de um evento ocorrer em um dado tempo t, ou seja,

estima a chance de que o evento (parto ou florescimento, por exemplo) aconteça em um período específico.

Existem diferentes modelos de AS que podem ser utilizados na área de genética e melhoramento. Um dos mais utilizados é o modelo de riscos proporcionais de Weibull (Ducrocq e Casella, 1996), o qual assume que o tempo de ocorrência do evento de interesse segue a distribuição de Weibull. Este é dado por:

$h(t) = \lambda_0(t)\exp(\mathbf{X\beta}+\mathbf{Za})$,

em que: h(t) é o vetor de funções de risco de um dado indivíduo no tempo t; $\lambda_0(t)$ é a função de risco base de Weibull no tempo t com um parâmetro de escala positivo λ e um parâmetro de forma ρ, isto é $\lambda_0(t) = \lambda\rho(\lambda t)^{\rho-1}$; $\boldsymbol{\beta}$ é o efeito fixo (com matriz de incidência **X**) e **a** o efeito genético aditivo (com matriz de incidência **Z**), assumido como $\mathbf{a}|\ \sigma^2_a \sim N(\mathbf{0}, \sigma^2_a\ \mathbf{G})$, sendo **G** a matriz de parentesco genômica aditiva (Van Raden, 2008) entre os indivíduos e σ^2_a a variância genética aditiva. Este modelo pode ser implementado no software *SurvivalKit* (Mészáros et al., 2013), e a informação genômica incluída por meio da matriz $\mathbf{G}^{-1}$ previamente calculada em outros softwares (R e PreGSF90).

Outro modelo de AS de ampla utilização na área de melhoramento é o semi-paramétrico de Cox, caracterizado como um método semi-paramétrico, pois não requer a escolha de uma distribuição de probabilidade para representar o tempo de sobrevivência e, como consequência disto, é muitas vezes considerado como mais robusto (Hou et al., 2009). Em resumo, este é uma aproximação do renomado modelo de riscos proporcionais de Cox (Cox, 1972), com a diferencial de permitir a inclusão de um efeito aleatório, que neste caso é o genético aditivo. Este modelo é dado por:

$\log[h(t)]=\log[h_0(t)]+(\mathbf{X\beta+Za})$

em que: $h_0(t)$ é o vetor da função segmentada constante no tempo t; e os demais parâmetros são os mesmos do modelo Weibull previamente apresentado. O modelo em questão também pode ser implementado no software *SurvivalKit* (Mészáros et al., 2013), e no software R, por meio da função *coxme* do pacote homônimo (Therneau, 2013).

O modelo semi-paramétrico de Cox já foi utilizado com sucesso na área de seleção genômica por (Santos et al., 2015). Nesta aplicação, a variável estudada foi idade ao abate de suínos, e os resultados demonstraram que ao considerar a censura, o modelo de Cox com efeito aleatório normal foi o mais apropriado por apresentar maiores acurácias de predição em relação ao modelo normal com dados imputados. Inclusive, as estimativas dos efeitos de marcadores obtidas como $\hat{\mathbf{b}} = (\mathbf{M}'\mathbf{M})^{-}\mathbf{M}'\hat{\mathbf{a}}$, possibilitou a seleção daqueles

mais relevantes, os quais foram utilizados para maximizar a capacidade preditiva do modelo em questão.

Dados longitudinais

Embora recentemente vários métodos estatísticos tenham sido propostos para SG, em se tratando de características longitudinais, como crescimento animal e produção de leite, tais propostas ainda são escassas na literatura. A maioria dos métodos propostos até o presente momento vêm sendo aplicados apenas para características pontuais, como pesos em idades específicas e/ou taxas de crescimento em certos períodos de tempo; ou ainda produção total de leite em uma lactação. Assim, surge o interesse em aplicar estes métodos a trajetórias completas do fenótipo de interesse ao longo do tempo. Na área de melhoramento vegetal dados desta natureza também são de grande interesse, principalmente para culturas perenes, cujas produções ao longo dos diferentes anos também podem ser caracterizadas como dados longitudinais.

Devido à extrema importância das características de desempenho (peso) na área de melhoramento animal, pesquisas direcionadas para a avaliação genética de curvas de crescimento contemplando informações de marcadores SNPs são relevantes para o sucesso da pecuária moderna. Da mesma forma, produções totais de culturas perenes ao longo dos anos, bem como altura da planta e comprimento de seus componentes (principalmente raiz, folhas e frutos) medidos em diferentes tempos também são de interesse para a fitotecnia.

Existem diferentes propostas para avaliação de dados longitudinais na área de genética e melhoramento, porém iremos nos ater a apresentar uma proposta geral composta por dois estágios distintos, a qual visa adaptar para a SG a metodologia inicialmente descrita por Varona et al. (1999). No primeiro estágio, são ajustados modelos de curvas de crescimento e de lactação aos dados longitudinais de cada indivíduo, de forma que são obtidas estimativas individuais dos parâmetros dos modelos em questão. Vale lembrar que pela natureza dos modelos ajustados, tais parâmetros apresentam interpretações biológicas. No caso de curvas de crescimento, tais parâmetros se caracterizam como velocidades de crescimento e pesos máximos. No segundo estágio, são considerados modelos específicos para seleção genômica ampla considerando como fenótipos as estimativas dos parâmetros obtidas no estágio anterior.

É interessante ressaltar que esta metodologia em dois estágios tem sido aplicada com sucesso em estudos de curvas de crescimento na área de SG. Pong-Wong e

Hadjipavlou (2010) utilizaram uma metodologia composta pelo ajuste de um modelo de crescimento e técnicas de seleção genômica ampla em conjuntos de dados simulados. Para o problema em questão, os autores primeiramente ajustaram o modelo de crescimento Gompertz aos dados simulados de peso-idade e obtiveram os parâmetros estimados para cada indivíduo avaliado. Posteriormente a SG foi aplicada para obter GEBVs para os fenótipos (parâmetros estimados) provenientes do primeiro passo. Silva et al. (2013) ajustaram o modelo Logístico para descrever curvas de crescimento de suínos, e as estimativas dos parâmetros deste modelo foram considerados como observações fenotípicas no modelo LASSO bayesiano de SG. Desta forma foi possível obter GEBVs para os parâmetros do modelo, e assim confeccionar, o que os autores denominaram, de curvas de crescimento genômicas para todos os indivíduos.

Na Tabela 1 são apresentados os modelos de regressão não-linear mais utilizados na literatura para descrever curvas de crescimento de animais de produção (peso); e também altura de plantas e comprimento de seus componentes (tais como raízes, folhas e frutos).

Tabela 1. Modelos não-lineares mais utilizados para descrever crescimento de uma dada característica (y) em diferentes idades (t), considerando k=1,2,...,K observações longitudinais.

Modelos	Equações	Modelos	Equações
Richards	$y_k = \frac{\beta_1}{\left(1+e^{(\beta_2-\beta_3 t_k)}\right)^{\frac{1}{\beta_4}}} + e_k$	von Bertalanffy	$y_k = \beta_1\left(1-\beta_2 e^{-\beta_3 t_k}\right)^3 + e_k$
Gompertz	$y_k = \beta_1 e^{\left(-e^{\left(\beta_2-\beta_3 t_k\right)}\right)} + e_k$	Brody	$y_k = \beta_1\left(1-\beta_2 e^{-\beta_3 t_k}\right) + e_k$
Logístico	$y_k = \frac{\beta_1}{\left(1+e^{(\beta_2-\beta_3 t_k)}\right)} + e_k$	Michaelis-Menten Modificado	$y_k = \frac{\beta_2\beta_3{}^{\beta_4} + \beta_1 t_k{}^{\beta_4}}{\beta_3{}^{\beta_4} + t_k{}^{\beta_4}} + e_k$

O ajuste dos modelos mostrados na Tabela 1 pode ser efetuado por meio da função *nls* do *software* R e *PROC NLIN* do *software* SAS. Em ambos os *softwares* utiliza-se o método dos quadrados mínimos ordinários com soluções obtidas via processo iterativo de Gauss-Newton. Para todos os modelos apresentados na Tabela 1, o parâmetro β_1 representa o peso adulto, ou peso assintótico, do animal (e altura máxima, ou comprimento máximo dos componentes, no caso de plantas); e o parâmetro β_3 a taxa de

maturidade, ou velocidade de crescimento (velocidade com que se atinge o máximo). Os modelos que apresentam o parâmetro $\beta 4$ possuem ponto de inflexão variável, cuja localização é determinada pelo parâmetro em questão. Os demais modelos, ou apresentam o ponto de inflexão fixo (Gompertz, Logístico e von Bertalanffy), ou não o possuem, como é o caso do modelo Brody. De forma geral, não há uma interpretação prática para o parâmetro β_2, sendo este uma constante de integração.

Em síntese, o modelo apresentado na Tabela 1 que melhor se ajusta aos dados de crescimento deve ser escolhido para que as estimativas de seus parâmetros sejam utilizadas como fenótipos no segundo estágio (modelos de SG). Para comparar os modelos em questão, geralmente são usados os seguintes avaliadores de qualidade de ajuste: coeficiente de determinação (R^2), quadrado médio do erro (QME), critério de informação de Akaike (AIC), critério de informação Bayesiano (BIC), erro quadrático médio de predição (MEP), coeficiente de determinação de predição (R^2p) e porcentagem de convergência (%C).

Conforme já relatado, após a indicação do melhor modelo de crescimento, as estimativas dos parâmetros dos referidos modelos são consideradas como fenótipos nos modelos de SG. Para tanto, tais estimativas são previamente corrigidas para efeitos fixos e efeitos ambientais aleatórios a fim de remover possíveis efeitos não genéticos que podem influenciar as características em questão.

Uma vez obtidos os fenótipos corrigidos (denominados a partir de agora por y) para efeitos fixos, os mesmos são submetidos às análises estatísticas por meio de modelos específicos de SG. Neste contexto, serão apresentados aqui apenas alguns modelos bayesianos (denominado de "alfabeto bayesiano da seleção genômica"), porém outros modelos tais como GBLUP e RR-BLUP também podem ser facilmente implementados. Os modelos bayesianos em questão são os seguintes: Bayes A (BA), B (BB), C (BC), LASSO Bayesiano (BL) e Regressão "*Ridge*" Bayesiana (BRR), os quais podem ser apresentados por meio do seguinte modelo geral de SG apresentado por Meuwissen et al., (2001):

$$\mathbf{y} = \mathbf{1}\mu + \sum_i \mathbf{x}_i g_i + \mathbf{e}, \qquad [1]$$

em que: **y** é o vetor de fenótipos corrigidos (estimativas dos parâmetros dos modelos de crescimento), **1** é o vetor de mesma dimensão de **y** com todas as entradas iguais a 1, μ é a média da característica estudada, g_i é o efeito do marcador (i=1,2,...,p), $\mathbf{x}_i$ é matriz de incidência de cada marcador i, e **e** é o vetor de resíduos do modelo, $\mathbf{e} \sim N(0, \mathbf{I}\sigma^2_e)$.

As distribuições *a priori* para os efeitos dos marcadores assumidas para os modelos BRR, BA, BB, BC e BL são dadas, respectivamente, por: $g_i \sim N(0,\sigma^2)\ \forall\ i=1,2,...,I$, sendo I o número total de marcadores; $g_i \sim N(0,\sigma_i^2)$; $g_i \sim (1-\gamma_i)N(0,\sigma_{i0}^2=0) + \gamma_i N(0,\sigma_{i1}^2)$; $g_i \sim (1-\gamma_i)N(0,\sigma_0^2=0) + \gamma_i N(0,\sigma_1^2)$ e $g_i \sim N(0,\tau_i^2\sigma_e^2)$. Especificamente para os modelos BB e BC, a probabilidade de gerar a variável indicadora binária γ_i (relacionada com a seleção de variáveis) é proveniente de uma distribuição Beta(α_1, α_2). Especificamente para o modelo BL, a distribuição *a priori* para o parâmetro τ_i^2 é assumida como sendo uma Exponencial, $\tau_i^2 \sim Exp(\lambda^2)$, na qual o parâmetro λ denominado de "penalização" ou "regularização", é por sua vez assumido por seguir uma distribuição Gama, tal que $\lambda^2 \sim G(\phi_1,\phi_2)$. Em todos estes modelos, o GEBV é calculado por: $\widehat{GEBV}_j=\hat{y}_j=\hat{\mu}+\sum_i x_{ij}\hat{g}_i = \hat{\mu} + x_{1j}\hat{g}_1 + x_{2j}\hat{g}_2 + \cdots + x_{pj}\hat{g}_p$.

Em termos práticos, vale ressaltar que a diferença entre os modelos BRR, BA, BB, BC e BL está relacionada com a arquitetura genética que os mesmos representam. O modelo BRR assume, *a priori*, que todos os marcadores apresentam a mesma variância; diferentemente o modelo BA assume uma variância para cada marcador. Já o modelo BB, tal como o BA, também assume uma variância para cada marcador, porém exerce adicionalmente uma seleção de covariáveis (i.e. assume que alguns marcadores não têm efeitos sobre a característica em questão). Esta seleção de variáveis é realizada por meio da mistura de distribuições Normais baseadas na proporção de valores 1 e 0 gerados para γ_i. O modelo BC, tal como BB, também assume esta seleção de variáveis, porém diferentemente do BA e similarmente ao BRR, assume apenas uma única variância para todos os marcadores. O modelo BL, tal como o BB, assume uma variância para cada marcador e também a seleção de variáveis, porém esta seleção não está fundamenta em misturas de distribuições, e sim na utilização de um parâmetro de regularização (λ) que direciona para zero marcadores com efeitos irrelevantes.

As distribuições *a priori* para todos os componentes de variância referentes ao modelo em [1] são assumidas como sendo qui-quadrada invertida escalada, de forma que exceto para os modelos BB e BC, o algoritmo amostrador de *Gibbs* pode ser utilizado para gerar amostras das distribuições marginais *a posteriori* para todos os parâmetros. Para estes modelos mencionados (BB e BC), a geração da variável latente indicadora (γ_i) é realizada por meio do algoritmo *Metropolis-Hastings*. Todos estes modelos bayesianos

podem ser facilmente implementados no software R por meio pacote BGRL (*Bayesian Generalized Linear Regression*).

Os modelos bayesianos apresentados podem ser comparados via análise de validação cruzada (VC). Esta pode ser implementada de diferentes formas, dependendo do número de indivíduos nos grupos de treinamento e validação. Uma proposta interessante, e teoricamente considerada a mais indicada (apesar de demandar alto custo computacional), é realizar a VC por meio do procedimento *Jacknife* (ou *leave-one-out*). Neste, em cada repetição, um indivíduo é removido do conjunto de dados para compor a população de validação, e os outros N-1 indivíduos são então utilizados na estimação dos GEBVs. Assim, uma vez estimados todos os efeitos de marcadores ($\hat{g}_1, \hat{g}_2, ..., \hat{g}_p$) na população de treinamento, estes são aplicados na população de validação para predizer o valor genômico ($\widehat{GBV}_j=\hat{y}_j$) para cada indivíduo da população de validação. Dessa forma, ao final da análise é obtido o fenótipo estimado ($\widehat{GBV}_j$) para todos os N indivíduos, de forma que a correlação entre $\widehat{GBV}_j$ e os fenótipos observados corrigidos constituem a medida de qualidade dos modelos de SG (quanto maior o valor desta correlação, maior capacidade de predição).

Uma vez obtidos os valores genômicos de cada animal j ($\widehat{GBV}_j$) para as estimativas dos parâmetros do modelo selecionado (melhor modelo da Tabela 1) é possível obter as "curvas de crescimento genômicas" ao substituir estes valores genômicos na equação do modelo de crescimento em questão. Por exemplo, assumindo que o modelo Brody (Tabela 1) tenha apresentado o melhor ajuste no primeiro estágio, pode-se obter as curvas mencionadas para cada animal j da seguinte maneira:

$$\hat{y}_{jk} = (\hat{\mu}_{\beta_1} + \widehat{GBV}_{\beta 1j})\left(1-(\hat{\mu}_{\beta_2} + \widehat{GBV}_{\beta 2j})e^{-(\hat{\mu}_{\beta_3}+\widehat{GBV}_{\beta 3j})t_k}\right), \quad [2]$$

em que: $\hat{y}_{jk}$ é o valor genômico do animal j para cada um dos k tempos nos quais os indivíduos foram mensurados, k=1,2,...,K; $\hat{\mu}_{\beta_1}$, $\hat{\mu}_{\beta_2}$ e $\hat{\mu}_{\beta_3}$ são, respectivamente, as médias estimadas para cada fenótipo $\hat{\beta}_1$, $\hat{\beta}_2$ e $\hat{\beta}_3$ (estimativas dos parâmetros do modelo Brody obtidas no primeiro estágio); $\widehat{GBV}_{\beta 1j}$, $\widehat{GBV}_{\beta 2j}$ e $\widehat{GBV}_{\beta 3j}$ são, respectivamente, os valores genômicos estimados para cada fenótipo $\hat{\beta}_1$, $\hat{\beta}_2$ e $\hat{\beta}_3$ (estimativas dos parâmetros do modelo obtidas no primeiro estágio); t_k é cada um dos k tempos nos quais os indivíduos foram avaliados, k=1,2,...,K.

A grande vantagem deste enfoque é que além de obter os valores genômicos para

cada tempo k observado, também é possível obtê-los para qualquer outro tempo de interesse dento da amplitude observada, necessitando apenas substituir o tempo este interesse no termo t_k do modelo em [2]. A metodologia de avaliação genética proposta para de curvas de crescimento em dois estágios permite ainda fazer o ranqueamento dos indivíduos por meio de seus valores genômicos diretamente para as estimativas dos parâmetros ($\hat{\beta}_1$, $\hat{\beta}_2$ e $\hat{\beta}_3$, no caso de curva de crescimento), uma vez que estas apresentam interpretações biológicas (valor máximo e velocidade de crescimento); e também ranquear os animais por meio de seus valores genômicos para a característica avaliada (peso dos animais, ou altura das plantas) em cada tempo, seja este observado ou não. Assim, cabe ao melhorista delinear a melhor estratégia de seleção para alterar, via seleção genômica, a forma das curvas de crescimento para obtenção de indivíduos mais eficientes (muitas denominados de precoces) em relação ao crescimento.

Para escolher tal estratégia, é necessário levar em consideração a herdabilidade (h^2) dos diferentes possíveis fenótipos a serem usados: pode-se considerar estimativas dos parâmetros dos modelos de crescimento ($\hat{\beta}_1$, $\hat{\beta}_2$ e $\hat{\beta}_3$) o próprio valor da característica ($\hat{y}_k$) em cada tempo observado (k) ou não (tempos de interesse dentro da amplitude de tempo observada). Para calcular h^2, independentemente do fenótipo e/ou do método de seleção genômica usado, a seguinte expressão pode ser utilizada: fenotípica $h^2 = \sigma_g^2/(\sigma_g^2 + \sigma_e^2)$, em que: σ_g^2 é a variância genética aditiva e σ_e^2 é a variância residual. Ao considerar como fenótipos as estimativas dos parâmetros dos modelos de crescimento ($\hat{\beta}_1$, $\hat{\beta}_2$ e $\hat{\beta}_3$), a variância genética aditiva é dada por (Gianola et al., 2009): $\sigma_g^2 = \sum_i 2p_i(1\text{-}p_i)\sigma_{gi}^2$, em que p_i é a freqüência alélica do marcador i.

Ao se considerar como fenótipos os pesos individuais (ou alturas de planas) $\hat{y}_k$ em cada tempo de interesse, uma vez que os valores genômicos para os mesmos são obtidos indiretamente via equação [2], deve-se adotar uma estratégia diferente. Nesta, a variância genética aditiva é dada por: $\sigma_g^2 = \text{var}(\hat{y}_k)$ e a variância fenotípica (σ_p^2) é dada por $\sigma_p^2 = \sigma_g^2 + \sigma_e^2 = \text{var}(\hat{y}_k^*)$, em que $\hat{y}_k^*$ são os valores de pesos (ou altura de plantas) em cada tempo k obtidos do ajuste do modelo de crescimento no primeiro estágio. Por exemplo, para o modelo Brody, a equação de predição fenotípica de peso (ou altura de planta) é dada por $\hat{y}_{jk}^* = \hat{\beta}_{1j}\left(1 - \hat{\beta}_{2j}e^{-\hat{\beta}_{3j}t_k}\right)$.

Diante das informações de herdabilidade para cada fenótipo, é possível estabelecer, conforme já comentado, estratégias de seleção a fim observar se maiores herabilidades são verificadas para as estimativas dos parâmetros dos modelos de crescimento ou para pesos em diferentes idades, além disso, estas duas informações podem ser combinadas com intuito de estabelecer índices de seleção para alterar geneticamente (Fitzhugh Jr., 1976) a forma das curvas de crescimento dos indivíduos sob seleção.

A metodologia proposta possibilita ainda estimar os efeitos dos diferentes marcadores para cada um dos possíveis fenótipos previamente mencionados. Para os fenótipos caracterizados pelas estimativas dos parâmetros dos modelos de crescimento, a estimação dos referidos efeitos é realizada diretamente por meio dos valores $\hat{g}_1, \hat{g}_2, ..., \hat{g}_p$, os quais podem ser agrupados no vetor $\hat{\mathbf{g}} = [\hat{g}_1, \hat{g}_2, ..., \hat{g}_p]$. Assim, é possível identificar os marcadores mais importantes para cada fenótipo ($\hat{\beta}_1$, $\hat{\beta}_2$ e $\hat{\beta}_3$).

Porém, até o momento, um aspecto ainda não explorado na literatura é a estimação dos efeitos de marcadores para peso (ou altura de plantas) $\hat{y}_k$ em qualquer tempo k de interesse. A importância de tal estimação está relacionada com a possibilidade de identificar marcadores mais importantes, ou de maiores efeitos, ao longo da trajetória temporal de crescimento, permitindo assim avaliar aqueles que atuam de forma mais significativa nas diferentes etapas de crescimento. Estes efeitos de marcadores em diferentes tempos (k) podem ser obtidos indiretamente por meio da expressão: $\hat{\mathbf{g}}_k = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\hat{\mathbf{y}}_k$, sendo $\mathbf{X}$ a matriz de incidência dos marcadores (contendo 0, 1 e 2, respectivamente para aa, aA e AA) e $\hat{\mathbf{y}}_k$ o vetor de valores genômicos estimados para peso (ou altura de planta) em cada tempo k de interesse. Uma vez identificados tais marcadores de maior relevância, os mesmos podem ser submetidos a análises posteriores para um maior entendimento de seu processo de atuação nas diferentes etapas de crescimento. Estas análises posteriores correspondem, por exemplo, a estudos de mapeamento fino, identificação de genes, ou expressão gênica que irão produzir informações sobre a atuação conjunta de diferentes genes que atuam sobre o processo de crescimento como um todo. Informações desta natureza foram exploradas, por exemplo, no estudo de Crispim et al. (2015), que utilizando a metodologia mencionada reportaram genes que atuam sobre toda a trajetória de crescimento de bovinos da raça Brahman.

Interação genótipo x ambiente

O entendimento dos fatores que levam os genótipos a apresentarem méritos genéticos diferenciados em vários ambientes sempre foi um interesse de programas de melhoramento animal e vegetal. Para tanto, é necessário o envolvimento com dados a partir dos quais se possa explorar a denominada interação genótipos x ambientes (GxE). Porém, a situação atual globalizada na área agropecuária, caracterizada pela rápida disseminação de material genético desenvolvidos em diferentes locais; e também o eminente apelo por soluções para as destacadas "mudanças climáticas", que implica em constante alterações de ambientes previamente estabelecidos como estáveis; aumentaram ainda mais o interesse pela GxE nos programas de melhoramento. Associado a este crescente interesse, tem-se ainda a inserção cada vez maior de informações genômicas nas avaliações genéticas, o que leva ao desafio de desenvolver estratégias para explorar a GxE no contexto de SG.

Na área de melhoramento animal muitos estudos sobre GxE têm sido realizados com análises multi-características, sendo que a mesma característica passa a ser considerada como característica diferente em função do nível do fator ambiental em estudo, o qual deve ser considerado um fator classificatório (como exemplo, sistema de criação – confinado ou pasto, ou regiões – norte e sul). Tal abordagem mostra-se interessante quando é possível definir um número pequeno de ambientes, de forma que análises multi-características sejam exequíveis.

Devido ao elevado número de ambientes (diversidade climática e/ou edafoclimática) existentes no Brasil, a modelagem da GxE por meio de análises multi-características fica limitada em função do elevado número de parâmetros (em geral (co)variâncias) a serem estimados. Uma alternativa para esta modelagem é a utilização de normas de reação (NR) com modelos de regressão aleatória, especialmente se associados às matrizes de parentesco genômicas, pois estas permitem maior conexidade dos rebanhos, independente das relações de pedigree entre indivíduos pertencentes a estes diferentes rebanhos.

Os modelos NR são especialmente interessantes porque usam funções lineares (nos parâmetros) para relacionar o mérito genético de um indivíduo a uma possível mudança ambiental. As estimativas dos parâmetros destes modelos permitem inferir sobre diferenças genéticas entre os indivíduos, bem como a herdabilidade e correlações genéticas em diferentes índices (gradientes) ambientais (Kolmodin et al, 2002, Calus e Veerkamp, 2003;. Cardoso e Tempelman, 2012). Em síntese, a NR de um genótipo é a

mudança sistemática no valor genético em resposta a uma mudança sistemática na variável ambiental, assim o modelo NR permite descrever como o mérito genético se altera gradualmente e continuadamente sobre um gradiente ambiental, possibilitando a identificação de diferenças de sensibilidade ambiental entre os indivíduos sob seleção, o que pode vir a indicar uma reclassificação do mérito genético dos mesmos no decorrer dos níveis ambientais. Geralmente os índices mais utilizados são aqueles baseados nos efeitos pré-estimados dos efeitos fixos considerados no modelo misto, tais como grupos de contemporâneos (rebanho-ano-estação de nascimento). Porém, índices mais específicos tais temperatura, umidade, fertilidade de solo, dentre outros, também podem ser combinados para representar este gradiente ambiental contínuo.

A utilização de informações genômicas (marcadores SNPs) para investigar a presença de GxE é uma importante ferramenta para a área de melhoramento, pois possibilita estimar os valores genômicos (GEBVs) dos indivíduos em diferentes ambientes e também identificar regiões cromossômicas relacionadas diretamente com a adaptação dos indivíduos a estes ambientes (Silva et al., 2014).

No presente material será apresentado um modelo NR para SG em dois passos por meio do método REML-BLUP (Calus et al., 2002; Kolmodin et al., 2002). De acordo com Cardoso e Tempelman (2012), esta é a abordagem mais comumente utilizada, e sob um ponto de vista prático, este modelo é preferível por permitir sua implementação em software livres como o R, WOMBAT e BLUPF90, visto que todos estes permitem a inclusão de uma matriz de parentesco externa (neste caso a matriz de parentesco genômica **G**).

Nesta análise em dois passos, o primeiro fornece estimativas dos efeitos fixos sobre o fenótipo usando um modelo geral que ignora a GxE. Assim, é possível obter o fenótipo corrigido para estes efeitos por meio da subtração dos mesmos em relação aos valores originais da característica de interesse. Neste primeiro passo, um modelo misto tradicional será utilizado:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \mathbf{e},$$

[3]

em que: **y** é o de observações fenotípicas; $\boldsymbol{\beta}$ é o vetor de efeitos fixos com respectiva matriz de incidência **X** e **u** é o vetor de efeito genético aditivo, $\mathbf{u} \sim N(\mathbf{0}, \sigma_u^2\mathbf{G})$, sendo **G** a matriz de parentesco genômico (Van Raden, 2008) e **e** o vetor de resíduos, $\mathbf{e} \sim N(\mathbf{0}, \sigma_\varepsilon^2\mathbf{I})$.

O objetivo deste primeiro passo é fornecer um vetor de fenótipos pré-corrigidos ($\mathbf{y}^*$) para os efeitos fixos, dado por: $\mathbf{y}^* = \mathbf{y} - \left(\mathbf{X}\widehat{\boldsymbol{\beta}}\right) = \mathbf{Z}\hat{\mathbf{u}} + \hat{\mathbf{e}}$.

No segundo passo, o modelo NR [4] é ajustado considerando como variável dependente os valores y^* obtidos no primeiro passo, e como variável independente o índice ambiental denotado por W (que na verdade representa as soluções para os efeitos fixos provenientes do primeiro passo, ou seja, $\widehat{\boldsymbol{\beta}}$ = W). Desta forma, tem-se:

$y_{ij}^* = \mu + a_i + b_i W_{ij} + \xi_{ij}$,

[4]

em que: y_{ij} são os valores fenotípicos (pré-corrigidos para efeitos fixos) do indivíduo i no nível j do índice ambiental (W), μ é a média geral, a_i e b_i serão, respectivamente, o intercepto aleatório e a inclinação da regressão aleatória para o valor genético aditivo (a_i) sobre os níveis de W, e ξ_{ij} é o termo residual, $\xi_{ij} \sim N\left(0, \sigma_\xi^2\right)$. A distribuição conjunta de $\mathbf{a}=[a_1, a_2, \ldots, a_{113}]'$ e $\mathbf{b}=[b_1, b_2, \ldots, b_{113}]'$, sendo $\boldsymbol{\theta} = [\mathbf{a}, \mathbf{b}]'$, é assumida como $\boldsymbol{\theta} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{ab} \otimes \mathbf{G})$, em que $\boldsymbol{\Sigma}_{ab} = \begin{bmatrix} \sigma_a^2 & \sigma_{ab}^2 \\ \sigma_{ab}^2 & \sigma_b^2 \end{bmatrix}$ e $\mathbf{G}$ é a matriz de parentesco genômica aditiva (Van Raden, 2008) entre os indivíduos.

A fim de avaliar se o modelo [4] é realmente necessário (ou seja, se realmente existe a GxE), geralmente compara-se o mesmo sua versão reduzida, $y_{ij}^* = \mu + a_i + \xi_{ij}$ (ausência de índice ambiental W afetando valores genéticos). Esta comparação é realizada por meio do teste de razão de verossimilhança (LRT), cuja estatística é D = - 2*(log L para o modelo reduzido) + 2*(logL para modelo alternativo). Neste caso, o modelo reduzido é o modelo de intercepto aleatório e o alternativo é o modelo com intercepto aleatório e inclinação aleatória [4].

No modelo considerando GxE [4], o valor genômico para o indivíduo i no nível ambiental j ($\hat{u}_{ij}$), bem como as estimativas de variância genética ($\hat{\sigma}_{u_j}^2$) e a herdabilidade ($\hat{h}_j^2$) para a característica em cada um destes níveis ambientais j podem ser calculados respectivamente por: $\hat{\mathbf{u}}_j = \mathbf{K}_j'\hat{\boldsymbol{\theta}}_i = \hat{\mathbf{a}}_i + \hat{\mathbf{b}}_i\mathbf{W}_j$, $\hat{\sigma}_{u_j}^2 = \mathbf{K}_j'\hat{\boldsymbol{\Sigma}}_{ab}K_j = \hat{\sigma}_a^2 + 2\hat{\sigma}_{ab}W_{ij} + \hat{\sigma}_b^2 W_{ij}^2$, e $\hat{h}_j^2 = \hat{\sigma}_{u_j}^2 / (\hat{\sigma}_{u_j}^2 + \sigma_\xi^2)$, em que $K_j' = \left[\mathbf{1}\ \mathbf{W}_j\right]$. Da mesma maneira, as covariâncias ($\hat{\sigma}_{u_{jj'}}$) e correlações genéticas ($\hat{r}_{jj'}$) estimadas entre a característica no nível ambiental j e no

nível j', são obtidas por: $\hat{\sigma}_{u_{jj}}=\mathbf{K}_j'\hat{\boldsymbol{\Sigma}}_{ab}\mathbf{K}_j = \hat{\sigma}_a^2 + 2\hat{\sigma}_{ab}W_{jj'} + \hat{\sigma}_b^2 W_j W_{j'}$ e $\hat{r}_{jj'} = \frac{\hat{\sigma}_{u_{jj}}}{\sqrt{\hat{\sigma}_{u_j}^2 \hat{\sigma}_{u_{j'}}^2}}$.

Para calcular o vetor de efeitos de marcadores SNPs ($\widehat{\boldsymbol{\phi}}_j$) para cada nível do índice ambiental W será utilizada a equação desenvolvida por Silva et al. (2014), a qual é a seguinte:

$$\widehat{\boldsymbol{\phi}}_j = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\hat{\mathbf{u}}_j) = (\mathbf{X}'\mathbf{X})^{-1}\left(\mathbf{X}'(\hat{\mathbf{a}} + \hat{\mathbf{b}}\mathbf{W}_j)\right) = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\hat{\mathbf{a}} + \mathbf{X}'\hat{\mathbf{b}}\mathbf{W}_j)$$

$$= \underbrace{(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\hat{\mathbf{a}})}_{\text{Efeito de SNP para a: } \hat{\mathbf{a}}_{SNP}} + \underbrace{(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\hat{\mathbf{b}})\mathbf{W}_j}_{\text{Efeito de SNP para b: } \hat{\mathbf{b}}_{SNP}} .$$

Tal desenvolvimento possibilita a obtenção de uma equação de predição linear geral do efeito de um dado marcador para cada nível j do gradiente ambiental:

$$\widehat{\boldsymbol{\phi}}_j = \hat{\mathbf{a}}_{SNP} + \hat{\mathbf{b}}_{SNP}\mathbf{W}_j.$$

Esta equação permite então estimar um vector de efeitos SNPs para cada nível j do índice ambiental W. Para estudar se os SNPs têm o mesmo efeito ao longo dos níveis ambientais pode-se, por exemplo identificar o top 1% de SNP que apresentaram os maiores efeitos absolutos em cada um desses níveis. Esta informação é de grande relevância para a área de genética e melhoramento, pois permite identificar regiões cromossômicas relacionadas com a adaptação dos indivíduos às condições ambientais em questão. Assim, uma vez identificadas estas regiões, é possível propor testes genômicos específicos para selecionar indivíduos detentores de genótipos favoráveis para certos genes (e/ou QTLs) que controlam fatores fisiológicos ligados ao sucesso da adaptação dos indivíduos geneticamente superiores em diferentes ambientes de interesse.

A modelagem apesentada já foi aplicada com sucesso por Verardo et al. (2016b) para explorar a GxE para características reprodutivas de suínos considerando como gradiente ambiental as soluções para os efeitos fixos de país-rebanho-ano-mês de nascimento. Neste contexto, também foram identificados genes localizados nas regiões dos SNPs mais relevantes em cada nível ambiente, de forma que redes gênicas foram propostas com o objetivo de elucidar a adaptação dos animais aos diferentes ambientes. A abordagem bayesiana dos modelos apresentados também foi aplicado com sucesso ao estudo de GxE para resistência ao carrapato em bovinos de corte por Mota et al. (2016).

Refinamento da Seleção Genômica pelo uso de QTNs em lugar de SNPs

A evolução da tecnologia genômica é previsível e a mutação causal de uma variação genética em nível de nucleotídeo (os QTNs) poderá ser acessada em um futuro próximo. Assim, a seleção genômica poderá ser aperfeiçoada pelo uso direto dos QTNs em lugar dos SNPs. O uso dos QTNs trará as seguintes vantagens: (i) a SG não dependerá do desequilíbrio de ligação, pois o QTN será acessado diretamente e não via marcadores; isto aumentará a durabilidade da predição genômica, a qual será útil também a longo prazo; (ii) a predição genômica poderá ter validade (transferibilidade) por meio de diferentes populações e espécies em um mesmo gênero; (iii) a predição genômica usará QTNs específicos para cada caráter, ao contrário do G-BLUP via SNPs, o qual usa a mesma matriz de parentesco G para todas as características; (iv) os índices de seleção multicaracterísticos ponderarão diretamente os QTNs e não as características fenotípicas; (v) a seleção genômica poderá usar um menor número de gerações (apenas as últimas) para a composição da matriz G, isto trará maior ganho genético e menor massa de dados a serem processados; (vi) as frequências alélicas dos QTNs serão acessadas diretamente e não via desequilíbrio de ligação com SNPs.

Referências

AZEVEDO, C.F., RESENDE, M.D.V, SILVA, F.F., Viana, J.M.S., VALENTE, M.S.F., RESENDE, M.F.R., MUÑOZ, P. Ridge, Lasso and Bayesian additive-dominance genomic models. **BMC Genetics**, v. 16, 2015.

CALUS, M. P. L., AND R. VEERKAMP. Estimation of environmental sensitivity of genetic merit for milk production traits using a random regression model. **Journal Dairy Science**. 86: 3756-3764. 2003.

CARDOSO, F.F., AND R.J. TEMPELMAN. Linear reaction norm models for genetic merit prediction of Angus cattle under genotype by environment interaction. **Journal Animal Science**. 90(7):2130-2141. 2012.

COX, D. R. Regression models and life tables (with discussion). **Journal Royal Statistical Society**, B, v. 34, n. 2, p. 187 – 220, 1972.

CRISPIM A.C., KELLY M.J., GUIMARÃES S.E.F., SILVA F.F., FORTES M.R.S. Multi-Trait GWAS and New Candidate Genes Annotation for Growth Curve Parameters in Brahman Cattle. **PLoS ONE** 10(10), 2015.

DONOGHUE, K.A.; REKAYA, R.; BERTRAND, J.K. Comparison of methods for handling censored records in beef fertility data: simulation study. **Journal of Animal**

Science, v.82, n.2, p.351–356, 2004.

DUCROCQ, V.; CASELLA, G. A Bayesian analysis of mixed survival models. **Genetics Selection Evolution**, v.28, p.505–529, 1996.

FITZHUGH Jr., H.A. Analysis of growth curves and strategies for altering their shape. **Journal of Animal Science**, v.42, n.4, p.1036-1051, 1976.

GEWEKE, J. **Evaluating the accuracy of sampling-based approaches to calculating posterior moments**. New York: Oxford University, 1992, 631 p.

GIANOLA D, DE LOS CAMPOS G, HILL WG, MANFREDI E, FERNANDO R. Additive genetic variability and the Bayesian alphabet. **Genetics**, 183:347-363, 2009.

HADFIELD, J. D. and NAKAGAWA, S. General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters. **Journal of Evolutionary Biology**, 23: 494–508. 2010.

HEIDELBERGER, P.; WELCH, P. Simulation run length control in the presence of an initial transient. **Operation Research**, v. 31, n. 6, p. 1109-44, 1983.

HOU, Y. et al. Genetic analysis of days from calving to first insemination and days open in Danish Holstein using different models and censoring scenarios. **Journal of Dairy Science**, v.92, p.1229–1239, 2009.

KOLMODIN, R., E. STRAMBERG, P. MADSEN, AND H. JORJANI. Genotype by environment interaction in Nordic dairy cattle studied using reaction norms. **Acta Agriculture Scandinavia**, 52:11-24.2002.

KORSGAARD, I.R. et al. Multivariate Bayesian analysis of Gaussian, right censored Gaussian, ordered categorical and binary traits using Gibbs sampling. **Genetics Selection Evolution**, v.35, p.159–183, 2003.

LEGARRA A. (2008) **TM Threshold Model** (available at: http://acteon.webs.upv.es/; last accessed 17 March 2014).

MÉSZÁROS, G., SÖLKNER, J., DUCROCQ, V. (2013): The Survival Kit: Software to analyze survival data including possibly correlated random effects. **Computer Methods and Programs in Biomedicine**, 110(3): 503-510. 2013.

MEUWISSEN, T. H. E.; HAYES, B. J.; GODDARD, M. E. Prediction of total genetic value using genome wide dense marker maps. **Genetics**, v. 157, n. 4, p. 1819 – 1829, 2001.

MISZTAL, I., TSURUTA, S., LOURENÇO, D., AGUILAR, I., LEGARA, A., VITEZICA, Z.(2014). **Manual for BLUPF90 family of programs**. Georgia: Athens: University of Georgia.

MOTA, R. R.; LOPES, P. S.; TEMPELMAN, R. J.; SILVA, F. F.; AGUILAR, I.; GOMES, C. C. G.; CARDOSO, F. F.. Genome-enabled prediction for tick resistance in Hereford and Braford beef cattle via reaction norm models. **Journal of Animal Science**, v. 94, p. 1834, 2016.

PONG-WONG R, HADJIPAVLOU GA. A two-step approach combining the Gompertz growth with genomic selection for longitudinal data. **BMC Proc**. 2010

RAFTERY, A.E.; LEWIS, S. How many iterations in the Gibbs sampler? In: BERNARDO, J.M. et al. (Eds.). **Bayesian statistics**. Oxford, USA: University Press, 1992, p. 763-773.

SANTOS, V.S. ; MARTINS FILHO, S. ; RESENDE, M.D.V. ; AZEVEDO, C.F. ; LOPES, P.S. ; GUIMARÃES, S.E.F. ; GLÓRIA, L.S. ; SILVA, F.F. . Genomic selection for slaughter age in pigs using the Cox frailty model. **Genetics and Molecular Research**, v. 14, p. 12616-12627, 2015.

SILVA, F. F.; ROSA, G. M. J.; GUIMARÃES, S. E. F.; LOPES, P. S.; CAMPOS, G. S. Three-step Bayesian factor analysis applied to QTL detection in crosses between outbred pig populations. **Livestock Science**, v. 142, n. 1, p. 210-215, 2011.

SILVA, FF; RESENDE, MDV ; ROCHA, GS ; DUARTE, DAS ; LOPES, PS ; BRUSTOLINI, OJB ; THUS, S ; VIANA, JMS. ; GUIMARÃES, S E.F. . Genomic growth curves of an outbred pig population. **Genetics and Molecular Biology** (Impresso), v. 36, p. 1, 2013.

SILVA, F.F., MULDER, H.A., KNOL, E.F., LOPES, M.S., GUIMARÃES, S.E.F., LOPES, P.S., MATHUR, P.K.,, VIANA, J.M.S., BASTIAANSEN, J.W.M. Sire evaluation for total number born in pigs using genomic reaction norms approach. **Journal Animal Science.** p. 382, 2014.

SPIEGELHALTER, D. J., BEST, N. G., CARLIN, B. P. AND VAN DER LINDE, A. Bayesian measures of model complexity and fit. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, 64: 583–639. 2002.

THERNEAU, T. M. Coxme: **Mixed Effects Cox Models**. R package version 2.2-3, 2012. Disponível em <http://CRAN.R-project.org/package=coxme> Acesso em: 10 de fev. 2013.

URIOSTE, J.I.; MISZTAL, I.; BERTRAND, J.K. Fertility traits in springcalving Aberdeen Angus cattle. 2. Model comparison. **Journal of Animal Science**, v. 85, p.2861–2865, 2007.

VAN RADEN, P.M. Efficient methods to compute genomic predictions. **Journal of Dairy Science**, v. 91, n. 11, p. 4414 – 4423, 2008.

VARONA, L,; MORENO, C.; GARCIA CORTÉS, L,A; YAGÜE, G,; ALTARRIBA,J. 1999. Two-step versus joint analysis of Von Bertalanffy function, **J. Anim, Breed, Genet**, 116, p,331-338.

VENTURA, H.T.; SILVA, F.F.; VARONA, L. ; FIGUEIREDO, E.A.P. ; COSTA, EV ; SILVA, LP ; VENTURA, R ; LOPES, PS. Comparing multi-trait Poisson and Gaussian Bayesian models for genetic evaluation of litter traits in pigs. **Livestock Science** (Print), v. 1, p. 1, 2015.

VERARDO, L. L. ; LOPES, M. S. ; WIJGA, S. ; MADSEN, O. ; SILVA, F. F. ; GROENEN, M. A. M. ; KNOL, E. F. ; LOPES, P. S. ; GUIMARÃES, S. E. F. . After genome-wide association studies: Gene networks elucidating candidate genes divergences for number of teats across two pig populations. **Journal of Animal Science**, v. 94, p. 1, 2016a.

VERARDO, L., SILVA F F. ; LOPES, M S. ; MADSEN, O ; BASTIAANSEN, J W. M., KNOL, E F. ; KELLY, M ; VARONA, L; LOPES, PS ; GUIMARÃES, S E. F. . Revealing new candidate genes for reproductive traits in pigs: combining Bayesian GWAS and functional pathways. **Genetics Selection Evolution**, v. 48, p. 1, 2016b.

Bioinformática e os avanços computacionais nas análises biométricas aplicadas ao melhoramento

Cosme Damião Cruz[1], Isabela de Castro Sant'Anna[2]

Introdução

A computação estuda e implementa os métodos de obtenção de soluções para problemas matemáticos e estocásticos. Seus fundamentos teóricos apresentaram grande evolução a partir do surgimento e da popularização do computador, um instrumento eletrônico com alta capacidade para efetuar operações aritméticas e lógicas com grande velocidade e sem a intervenção humana. O computador é hoje equipamento indispensável para realizar as mais diversas tarefas, dentre elas, auxiliar o pesquisador em trabalhos científicos, pois reúne qualidades básicas que o torna fundamental para gerar informações por meio da análise e processamento de dados. Além da alta velocidade no processamento das informações, é possível o armazenamento de grande volume de dados e, por meio do desenvolvimento de programas específicos, é possível executar um conjunto enorme de procedimentos estabelecidos pelo programador.

A necessidade de se realizar cálculos é tão antiga quanto a própria história da humanidade, mas somente a partir da invenção de instrumentos básicos é que se teve início ao processamento moderno. Constata-se que o uso de recursos computacionais tem sido cada vez mais intenso em todas as atividades do ser humano. Grande parte das dificuldades proporcionadas pela pequena disponibilidade de equipamentos tem sido superada, em particular no meio científico. O computador tem sido equipamento presente em laboratórios de pesquisa, permitindo o pronto processamento, gerenciamento de bancos de dados e edição de textos, gráficos e imagens. Tem sido um instrumento que permite, de forma dinâmica, levar conhecimentos a público, em especial na área de genética, permitindo tanto o ensino como o aprendizado de forma interativa e agradável.

Os geneticistas devem estar sempre atentos às mudanças e aos avanços técnicos que se apresentam na atualidade, pois atuam numa área da ciência que desperta grande interesse por tratar de temas que estão diretamente relacionados com a vida do homem. Na genética são enfatizados temas relativos a evolução, a hereditariedade e ao

[1]Engenheiro Agrônomo, M.S., D.S. e Professor da Universidade Federal de Viçosa. E-mail: cdcruz@ufv.br
[2]Bióloga, M.S. e Doutoranda da Universidade Federal de Viçosa. E-mail: isabela.santanna@ufv.br

melhoramento e, é possível que muitos dos avanços da humanidade para o próximo milênio, certamente virão de descobertas e aplicações de princípios genéticos. Apesar da grande motivação e interesse de sempre se incorporar estes novos conhecimentos e estas facilidades operacionais no trabalho do melhorista, os educadores e formadores de especialistas e profissionais, nos diversos níveis, tem razão de terem preocupação adicional sobre estes temas pois grande volume de informação tem sido gerado em diversos níveis, destacando dados fenotípicos e, ou, genotípicos originados de populações, indivíduos, cromossomos e DNA. A integração destas informações é um desafio para a ciência na busca de soluções para os mais diversos problemas da sociedade.

Processamento de dados

Geralmente o geneticista realiza suas pesquisas, em diferentes áreas, gerando informações ou dados a respeito de seus estudos biológicos. Essas informações quando quantificáveis são utilizadas para análise e interpretação e, por fim, para tomada de decisão que resulte na otimização de recursos e maximização de ganhos.

O processamento permite converter um conjunto complexo de dados em outros ordenados, simplificados e fáceis de serem interpretados. Ele pode ser feito de forma manual, quando exclusivamente realizado pelo pesquisador, semi-automático, quando se lança mão de qualquer instrumento de cálculo para auxiliar as tarefas de processamento, e automático, quando a transformação dos dados de entrada em solução, ou saída, é feita exclusivamente pela máquina. Neste caso, a máquina que tem realizado tarefas prolongadas, rápidas e eficientes, é o computador.

O computador é capaz de realizar o processamento, com base em um modelo biológico definido, porque é instruído de programas, em que há uma sequência de instruções claras e lógicas. Vários pesquisadores, ou programadores, têm se dedicado ao desenvolvimento destes programas ou aplicativos computacionais, com rotinas específicas a serem utilizados em cada área da ciência. Na maioria dos casos o computador tem pouca inteligência, no contexto de aprendizagem, pois não toma decisões por si só, apenas executa uma série de instruções fornecidas por um programador. Uma área da Inteligência computacional é, atualmente, bem promissora mas com resultados limitados em certas áreas.

Se as instruções forem bem estabelecidas, lógicas e corretas, o computador executará o processamento com sucesso, caso contrário poderá ter seu processamento interrompido por algum tipo de erro, ou proporcionar uma resposta incorreta, algumas

vezes absurda e outras imperceptíveis, provocando consequências danosas. Atribuir erros de análise ao computador não é coerente, pois o erro, em essência, é do programador que estabeleceu o programa e, outras vezes, é do usuário que não soube operar convenientemente o aplicativo ou fornecer os dados atendendo às suas particularidades uma vez que não estava atento em ultrapassar os limites ou deixar de atender as restrições do aplicativo.

Quando se realiza um estudo e há necessidade do processamento dos dados deve se ter algumas preocupações básicas, como considerar a qualidade dos dados disponíveis e a adequação dos mesmos ao modelo biométrico a ser empregado. A qualidade dos dados, depende do cuidado do pesquisador, de seu zelo em garantir que a informação obtida reflita exatamente o fenômeno estudado e que seu experimento seja implantado e conduzido adotando-se estratégias que permitam minimizar as influências de fatores externos, conduzindo a menores erros experimentais. Algumas vezes a análise é comprometida por erros grosseiros de coleta de dados, pela falta de controle ambiental, inerente ao delineamento utilizado ou pelo mal planejamento dos ensaios.

Muitos aplicativos computacionais estão disponíveis na comunidade científica, sendo utilizado em muitos estudos da área de genética e melhoramento. Muitos programas são de uso mais restrito ou abrange apenas certos procedimentos biométricos, mas são igualmente importantes e, por meio deles, tem sido viabilizados muitos estudos genéticos. A escolha do aplicativo depende de uma série de fatores, o principal é a qualidade do processamento, confiabilidade e capacidade de gerar resultados de interesse para o usuário. Existem programas amplos, de padrão internacional reconhecido, que devem ser preferidos, pois além de sua qualidade, apresentam grande consistência e confiabilidade pois estão sob responsabilidade de equipe técnica qualificada. O uso destes programas, entretanto, tem sido limitado pelo custo do software ou mesmo pela abrangência que, de tão ampla, acaba indo além do interesse do usuário, que passa a ter certa dificuldade em reconhecer, entre as várias opções disponíveis, aquelas que seriam mais apropriadas a sua análise. É por este motivo que programas mais específicos têm ocupado lugar de importância e o desenvolvimento de software constitui-se numa contribuição relevante para pesquisa, principalmente em genética e melhoramento.

A seguir destacaremos algumas importantes abordagens, que demanda recursos computacionais adequados, que vem proporcionando considerável avanço no melhoramento genético por meio do processamento de dados e geração de informações para entendimento de fenômenos biológicos, para otimização de recursos e para tomada

de decisão.

Bioinformática no melhoramento convencional

Uma das evidências da importância da bioinformática encontra-se nas atividades rotineiras de um programa de melhoramento genético. Sabe-se que, para a obtenção de materiais genéticos superiores, é necessário que os indivíduos selecionados reúnam, simultaneamente, uma série de atributos favoráveis que lhes confiram rendimento comparativamente mais elevado e que satisfaça às exigências do consumidor. É importante também a realização de experimentos fidedignos, dos quais é obtido grande volume de dados experimentais. Por meio do processamento adequado destes dados os parâmetros genéticos são estimados e os fenômenos biológicos são interpretados. Nesta etapa de análise e interpretação de resultados é fundamental a existência de recursos computacionais e aplicativos eficientes à disposição do pesquisador.

De maneira geral constata-se que na área científica há disponibilidade de vários softwares destinados ao processamento de dados segundo variados modelos estatísticos. Esses aplicativos, tanto os desenvolvidos em nosso país quanto no exterior, permitem uma gama de análises que auxiliam de forma considerável o pesquisador por permitir uma condensação de seus dados sem perda de informações indispensáveis à sua pesquisa. Entretanto, quando há necessidade de análises por meio de modelos bioestatísticos, como aqueles usualmente empregados na Genética, tais aplicativos são deficientes por não apresentarem as opções desejadas ou por apresentá-las com informações de certa forma limitadas.

O desenvolvimento de aplicativos na área de Genética e Melhoramento torna-se fundamental pela escassez dos mesmos, tanto no Brasil quanto no exterior. Sua disponibilidade visa atender a uma demanda crescente de usuários nas diversas instituições de pesquisa, que manipulam um grande volume de dados, os quais requerem um processamento adequado, para que parâmetros estatísticos e biológicos sejam convenientemente estimados. Apesar de serem de grande importância, a carência em aplicativos computacionais ocorre, apesar de que, atualmente, há menores limitações quanto a disponibilidade de equipamentos, ambientes e linguagens de programação. Isto ocorre em razão de se ter poucos profissionais que reúnam conhecimentos indispensáveis para identificar os problemas específicos da área de genética e formular soluções computacionais adequadas. O que se verifica é que profissionais da área de informática tem pouco conhecimento dos problemas da área de genética, que apresentam certa complexidade por tratar de fenômenos biológicos e envolver princípios e, de certa forma,

complexas distribuições probabilísticas na sua análise. Por outro lado, são raros os geneticistas com conhecimento e aptidão para atuarem na área de informática. Entretanto, com o melhoramento intensivo de muitas espécies, as tomadas de decisões a partir do processamento de dados tornam fundamentais e, cada vez mais, indispensáveis para nortearem o uso de estratégias que garantam o êxito do programa melhoramento.

Nas diversas etapas do melhoramento os melhoristas têm a necessidade de utilizar informações, expressas em parâmetros de modelos biométricos, que normalmente não estão disponíveis nas saídas da maioria dos softwares disponíveis para a área científica. Assim, por exemplo, metodologias de análises dialélicas, para escolha de progenitores para hibridações e formação de populações-base para seleção, de avaliação da estabilidade e adaptabilidade, para recomendação de cultivares, de estimação de parâmetros genéticos tais como herdabilidade, correlações etc., para avaliar e direcionar programas de melhoramento, não são geralmente encontradas nos aplicativos difundidos em nossa comunidade científica.

Programas destinados à área de Genética devem ser desenvolvidos atendendo a finalidade básica de análise e processamento de dados, com base em modelos biométricos. Estes programas, com os padrões sofisticados de interface entre usuário e máquina, têm de ser de fácil manuseio, objetivos e abrangentes, atendendo prontamente as exigências do usuário. Adicionalmente, sua compatibilidade com os sistemas operacionais disponíveis e a característica amigável com outros aplicativos e demais recursos computacionais são fundamentais na sua difusão e adoção na comunidade científica.

Alguns aplicativos eficientes e aplicáveis à área de Genética e Melhoramento estão disponibilizados. Entretanto, é fácil de perceber a necessidade de investimentos contínuos na área de programação, em que se têm ciclos bem definidos de avanços, demanda e uso. Cada vez que novos procedimentos são disponibilizados, novas alternativas surgem, em consequência de estudos diversificados e possibilidade de análise e interpretação de dados. A escassez que existia de recursos computacionais e de programas para análise, tornavam o uso de certas metodologias restrito, predominando somente algumas poucas por motivos diversos, tais como eficiência, simplicidade, divulgação ou mesmo modismo. Algumas destas metodologias ainda continuam sendo utilizadas, mesmo apresentando problemas ou que não satisfaziam completamente, por falhas de pressuposição, conceitos e adequação.

Exemplos de diversificações de metodologias que são empregadas em estudos ou pesquisas, em função da existência de programas computacionais, são encontrados em várias áreas. Nas metodologias para análise de estabilidade e adaptabilidade, por exemplo, verifica-se que na década de 80 a 90, o uso da metodologia baseada em regressão era rotineiro. Procedimentos como o de Eberhart e Russell (1966) eram utilizados tanto em trabalhos científicos como rotineiramente em trabalhos de melhoramento, apesar de existirem na literatura vários outros métodos, alguns complementares outros alternativos. Atualmente, os softwares já dispõem de várias metodologias, de modo que o melhorista tem maior flexibilidade e o fenômeno biológico estudado é mais eficientemente interpretado. O mesmo raciocínio se estende para índice de seleção, cuja eficiência já é comprovada na literatura, mas seu uso é limitado pela dificuldade de se ter dados prontamente disponíveis para o processamento.

A difusão de técnicas multivariadas é outro exemplo evidente do papel fundamental dos programas computacionais. Estas técnicas permitem interpretação simultânea de caracteres, sendo fundamentais em estudos de diversidade genética, seleção simultânea, relação entre caracteres, dentre outros. Alguns softwares podem ser considerados como instrumentos fundamentais por viabilizarem o uso destas técnicas, promovendo melhoria da qualidade da informação científica por permitirem o processamento adequado dos dados.

Ainda outro aspecto digno de destaque é a determinação da própria conduta no processamento dos dados. Sabe-se que a qualidade das informações depende da qualidade dos dados experimentais disponíveis e, em outros casos, mesmo com bons dados, o processamento pode ser inadequado por questões de pressuposições. Como exemplo cita-se os problemas danosos e resultados absurdos encontrados em análises de regressão, análise de trilha, correlações canônicas, índice de seleção, quando existem problemas de multicolinearidade. Pesquisadores sem conhecimento mais apropriado usam indiscriminadamente seus dados em certos procedimentos, gerando resultados equivocados. O diagnóstico de multicolinearidade, possível de ser realizado em alguns aplicativos computacionais, tem dado contribuição significativa na avaliação da qualidade e adequação dos dados a análises específicas.

Um exemplo de contribuição da bioinformática para esta área é o programa GENES é constituído por cerca de 420 programas executáveis. Os procedimentos são utilizados em diversos estudos genéticos aplicados a animais de grande e pequeno porte ou de culturas de plantas em geral. O programa apresenta procedimentos de análise

estatística que podem ser utilizados por grande número de pesquisadores de qualquer área de pesquisa. Assim, estatísticas descritivas, análise de variância, de regressão, de correlação e testes comparativos de médias que são amplamente utilizados por pesquisadores das diversas áreas. Além deste procedimento, o GENES conta com procedimentos específicos de biometria aplicadas a dados obtidos de estudos de genética e melhoramento. Dados binários ou multicategóricos, normalmente obtidos de estudos de Genética Molecular, também são processados, permitindo processamento fundamento em importantes modelos genéticos.

Bioinformática fundamentada em análises de marcadores moleculares

Com o desenvolvimento dos marcadores moleculares e o avanço em técnicas de biologia molecular, criou-se a expectativa de que as informações genotípicas dos marcadores moleculares, uma vez correlacionados com características fenotípicas de interesse, pudessem ser amplamente utilizadas na obtenção e seleção de indivíduos com maior valor genético (Resende Júnior, 2010). Assim, muitos estudos têm sido feitos com o objetivo de auxiliar o melhoramento genético agregando estas informações moleculares. Nestes estudos destacam-se abordagens sobre mapeamento genético, detecção de QTL (locos controladores de características quantitativas) e predições de ganhos por técnicas de seleção assistida por marcadores (SAM).

Um atrativo da genética molecular em benefício do melhoramento genético é a utilização direta das informações de DNA na seleção, de modo a permitir alta eficiência seletiva, rapidez na obtenção de ganhos genéticos com a seleção e baixo custo (Resende et al., 2008). Com a perspectiva de aumento nos ganhos de seleção e redução nos ciclos de melhoramento via seleção assistida por marcadores, muitas pesquisas foram feitas e QTL (locos controladores de características quantitativas) foram detectados e mapeados nas mais variadas culturas (Frary et al.,2000; Yano et al.,2000; Liu et al.,2002).

Com os avanços de tecnologias de genotipagem em larga escala, tornou-se possível uma cobertura completa do genoma, o que levou os pesquisadores a criar nova forma de utilização dessa informação genética gerada (Almeida, 2013). Meuwissen et al. (2001) propuseram método de seleção denominado seleção genômica (GS) ou seleção genômica ampla (GWS), a qual pode ser aplicada em todas as famílias em avaliação nos programas de melhoramento genético de espécies autógamas e alógamas.

Com todo o avanço pode-se visualizar uma nova ênfase no melhoramento genético, em que as informações pontuais sobre determinados genes, sua expressão e sua

interação com ambiente sejam ressaltadas. Ficou, então, explícita a necessidade de reconhecer os genes, manipula-los e regular sua expressão.

Para acompanhar todo avanço na área molecular houve a necessidade de desenvolvimento de algoritmos e aplicativos de fácil acesso e interpretação. Alguns exemplos de contribuição nesta área são os aplicativos MapeMaker, QTLcartograph e o GQMOL desenvolvidos para analisar dados obtidos de estudos de genética molecular e quantitativa. Geralmente estão disponíveis procedimentos para testes de segregação, mapeamento para populações exogâmicas (irmãos completos e meio-irmãos) e derivadas de populações controladas (retrocruzamentos, RILs, duplo-haplóides, F2 a Fn) e análise de QTLs, por marcas simples, intervalo simples e intervalo composto. Atualmente, bibliotecas no aplicativo R têm sido disponibilizadas para fins de processamento de grande volume de dados focados na predição de valores genético e estimação de efeitos de marcadores em análises de seleção genômica ampla usando modelos e abordagens biométricas variadas.

Bioinformática fundamentada em modelos mistos aplicados

Nos programas de melhoramento muitas vezes somos desafiados a trabalhar com populações já em alto nível de produtividade e que, por ventura, apresentam herança complexa para caracteres de interesse. Diante disso, faz se necessário relacionar a variação fenotípica a muitas variantes genéticas. Um dos problemas da modelagem de fatores poligênicos é captar a proporção da variância fenotípica que pode ser explicada por genótipos disponíveis, de modo a utilizar dessas informações para a predição de fenótipos.

A avaliação genética deve ser realizada sobre valores genéticos e não em médias fenotípicas, pois essas não se repetirão em plantios futuros (Resende e Duarte, 2007). Uma boa predição de fenótipos futuros ou mesmo a obtenção de boa eficiência seletiva só será atingida ao reconhecermos a complexidade dos caracteres quantitativos e proporcionar aos mesmos, dentro de programas de melhoramento, metodologias que garantam maior precisão experimental e o uso de métodos estatísticos apropriados para a seleção dos indivíduos. Sendo assim, um importante o parâmetro a ser utilizado na seleção genética é a acurácia (Resende, 2007).

Diante disto, reconhecemos que a predição de valores genéticos e a estimação de componentes de variância são atividades essenciais no melhoramento de plantas e, para tal, o procedimento de estimação/predição por meio do REML/BLUP (máxima

verossimilhança restrita/melhor predição linear não viciada) é o recomendado. Decorre deste fato, a necessidade de recursos computacionais para prover maneiras simples e amigáveis para que o processamento possa ser incorporado rotineiramente em um programa de melhoramento.

O uso da bioinformática para solucionar estes problemas é fundamental sobre vários contextos. Os teóricos ressaltam a necessidade de um processamento de dados mais robusto. Assim, a implementação da metodologia de modelos mistos, com a Melhor Estimação/Predição Linear Não-Viesada REML/BLUP de Henderson (1975), tem sido recomendada pelas vantagens práticas que esse procedimento oferece como a comparação de indivíduos ou variedades através do tempo e espaço, permitir a simultânea correção para os efeitos ambientais, a estimação de componentes de variância e a predição dos valores genéticos. Permite também lidar com estruturas complexas de dados (medidas repetidas, diferentes anos, locais e delineamentos), entre outros (Resende, 2007).

Já pesquisadores com foco nas questões mais práticas também argumentam sobre a utilização dos modelos mistos via REML/BLUP (máxima verossimilhança restrita e melhor predição linear não viesada) como o procedimento mais adequado para a seleção de indivíduos superiores por considerar características de espécies que apresentam longos ciclos de avaliação, em situações que ocorrerem perdas nas parcelas por diversos motivos e quando há interesse de capitalizar informações sobre o controle da estrutura da população e sobre o parentesco entre indivíduos mensurados.

Uma grande contribuição da área de bioinformática é a disponibilização do software SELEGEN REML/BLUP (RESENDE, 1997), que foi desenvolvido para servir de base ao melhoramento genético de perenes e permite trabalhar com vários sistemas reprodutivos e métodos de seleção. O seu objetivo é predizer valores genéticos de indivíduos ou de famílias, estimar parâmetros genéticos orientar a escolha de indivíduos para obtenção de máximos progressos genéticos imediatos, porém compatíveis com a manutenção de variabilidade genética suficiente para o melhoramento em longo prazo.

Na modelagem genética-estatística também destacam algumas bibliotecas disponíveis no softwares livre R (R Development Core Team, 2013). Dentre os pacotes criados, para a utilização de modelos lineares mistos voltados à estimação de componentes de variância e predição de valores genéticos, pode se citar o pedigreemm (Vazquez et al., 2009). Este tem a capacidade do pacote lme4, para ajustar modelos lineares mistos, com a diferença de que permite a correlação genética entre os níveis dos efeitos aleatórios do modelo, com a incorporação da matriz de parentesco. Entretanto, a

utilização desse pacote em dados de melhoramento vegetal, com delineamentos genéticos de famílias, apresenta certa dificuldade, devido o pedigreemm não considerar a informação dos genitores na matriz de incidência dos efeitos aditivos, gerando predições incorretas para os indivíduos. Resende et al. (2012) criaram equações que corrige a saída do pedigreemm para famílias de meio irmãos e irmãos germanos, considerando também algumas modificações, como a inclusão das informações dos genitores e a duplicação da última linha do arquivo de dados, para que o pacote pudesse ler e ajustar as referidas informações de famílias.

Bioinformática fundamentada em análises bayseanas

Como já foi relatado, a predição de valores genéticos e a estimação de componentes de variância são atividades essenciais no melhoramento de plantas e, pela sua eficiência, utiliza-se, na maioria das vezes, o procedimento REML/BLUP. Entretanto, uma deficiência do procedimento REML/BLUP para a estimação/predição de componentes de variância/valores genéticos refere-se ao fato de que o método REML propicia intervalos de confiança apenas aproximados para os parâmetros genéticos, através do uso de aproximações e suposições de normalidade assintótica. Isto porque a distribuição e variância dos estimadores não são conhecidos e, assim, questões referentes à efetividade da seleção a ser praticada não podem ser respondidas com rigor.

A análise Bayesiana baseia-se no conhecimento da distribuição a posteriori dos parâmetros genéticos e possibilita a construção de intervalos de confiança (melhor definido como intervalo de probabilidade ou intervalo de confiança Bayesiano) exatos para as estimativas dos parâmetros genéticos. Assim, esta análise propicia uma descrição mais completa sobre a confiabilidade dos parâmetros genéticos do que o método REML (Gianola & Fernando, 1986; Resende, 1997; 1999).

Em muitas situações há vantagens em se utilizar técnicas Bayesianas na estimação dos parâmetros genéticos, uma vez que estas podem incorporar informações prévias ao procedimento de estimação, as quais são especificadas por meio da distribuição a priori. Além desta possibilidade de incorporação de informações, outras vantagens relacionadas com a Inferência Bayesiana são caracterizadas pela ausência de pressuposições quanto aos modelos utilizados e pela facilidade da adoção da estimação por intervalo, neste caso denominado de intervalo de credibilidade, o qual é obtido diretamente pelos quantis da distribuição a posteriori (HOLSINGER, 2005).

Em outras áreas da Genética também tem sido apresentado trabalhos importantes considerando as técnicas Bayesianas. Assim, em Genética de Populações, AYRES & BALDING (1998) aplicaram a metodologia Bayesiana no estudo do coeficiente de endogamia populacional e KARHU (2001) utilizou, a partir de marcadores de microssatélites, para estimar o coeficiente de endogamia em populações de pinus. Mais recentemente, uma extensão da análise Bayesiana utilizada para estudar a estrutura genética de populações, sendo a distribuição Beta usada como aproximação da distribuição a posteriori do coeficiente de endogamia, foi realizada por HOLSINGER & WALLACE (2004). ARMBORST (2005) relata um caso de estimação multiparamétrica que pode ser tratado com o uso de técnicas Bayesianas e o método MCMC, que é a estimação das proporções alélicas e da medida de endocruzamento de forma simultânea.

Assim, considera-se que a utilização da metodologia Bayesiana pode suprir as carências relatadas pelo uso de procedimentos convencionais. Porém, para se realizar uma inferência Bayesiana, é necessário, além dos dados amostrais, o estabelecimento de uma informação a priori sobre o(s) parâmetro(s) e o cálculo da distribuição a posteriori do(s) parâmetro(s). A informação a priori é dada pela densidade de probabilidade $P(\theta)$, a qual expressa, de alguma forma, o conhecimento do pesquisador sobre o(s) parâmetro(s) a ser(em) estimado(s). Quando em determinado estudo o pesquisador tem pouca ou nenhuma informação para se incorporar, considera-se uma priori não-informativa, por exemplo, a priori de Jeffreys (JEFFREYS, 1961).

De maneira geral, considera-se os dados $Y=\{y1 , y2 , ..., yn \}$, representados por uma amostra aleatória de uma população com densidade f são utilizados na análise Bayesiana por meio da função de verossimilhança $L (y1 , ..., yn \mid \theta)$, que é o valor da densidade conjunta da amostra aleatória, obtido para a amostra dada. Portanto, a partir do momento que se opta por uma distribuição a priori, seja ela informativa ou não-informativa, e obtém-se a função de verossimilhança, é possível, por meio do Teorema de Bayes representado pela expressão sendo $Y=\{y1 , y2 , ..., yn \}$, obter a distribuição a posteriori de θ , de forma que qualquer conclusão a seu respeito é realizada a partir desta distribuição .

O denominador é uma constante, chamada de constante de integração, pois só depende da amostra dada e não depende de θ. Portanto temos: $P(\theta|Y) \alpha L(Y \mid \theta)P(\theta)$, ou seja, essa expressão pode ser entendida como: Posteriori α Verossimilhança x Priori, em que α é um símbolo que representa “proporcionalidade”. A distribuição a posteriori de

um parâmetro θ contém toda a informação probabilística de interesse a respeito do mesmo. Assim, a inferência sobre o parâmetro é realizada por meio desta distribuição.

Segundo ROSA (1998), para se inferir em relação a qualquer elemento (componente) de θ, a distribuição a posteriori conjunta dos parâmetros, P(θ |Y), deve ser integrada em relação a todos os outros elementos de θ. Assim, se θ =(θ1 , θ2) e, se o interesse do pesquisador se concentra sobre os componentes de θ1, tem-se a necessidade da obtenção da distribuição p(θ1|Y), denominada de marginal, a qual é dada por: A integração da distribuição conjunta a posteriori para a obtenção das distribuições a posteriori marginais pode não ser fácil de realizar, necessitando de algoritmos iterativos especializados como, por exemplo, o Gibbs Sampler, o qual faz parte de uma classe de algoritmos denominada de MCMC (Markov Chain-Monte Carlo).

Porém, para a utilização desse algoritmo, é necessário que se obtenha, a partir da distribuição a posteriori, um conjunto de distribuições chamadas de distribuições condicionais completas. O algoritmo Gibbs Sampler é uma ferramenta extremamente útil na resolução de problemas que envolvem a estimação de mais de um parâmetro, como é o caso do modelo de COCHERHAM (1969), porém, para utilização desse algoritmo, exige-se que as distribuições condicionais a posteriori dos parâmetros tenham formas conhecidas, ou seja, que sejam dadas por distribuições de probabilidade conhecidas.

Bioinformática fundamentada em análises de inteligência computacional

Outras metodologias disponíveis, que demandam uso intensivo de recursos e aplicativos computacionais fundamentados em inteligência artificial, tais como as redes neurais, a lógica fuzzy, a árvore aleatória, dentre outras. Estes procedimentos são úteis para a seleção de genótipos, na predição de valores fenotípicos e genéticos e em estudos classificatórios. A inteligência artificial tem permitido uma nova abordagem no processo de tomada de decisão em diversas áreas da ciência com grande potencial no melhoramento genético animal e vegetal (Ventura et al., 2013; Nascimento et al., 2013).

Em diversas situações o pesquisador dispõe de dados experimentais apropriados para realização das analises biométricas necessárias para avaliação dos experimentos com objetivo de discriminar indivíduos ou populações, porém as análises biométricas nem sempre são capazes de produzir resultados satisfatórios, pois o modelo adotado e as estatísticas requeridas (médias, variâncias e covariâncias) podem ser insuficientes para descrever e caracterizar convenientemente as particularidades de cada população. Neste contexto, a realização de análises por meio de métodos computacionais que sejam capazes

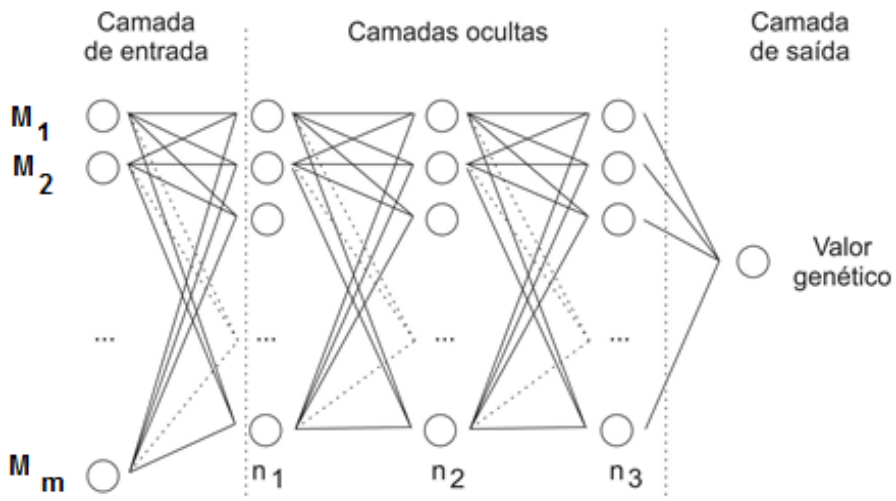
de aprendizagem e generalização a partir de toda a informação disponível, sendo tolerantes a ruídos, representa um grande avanço para os estudos envolvendo procedimentos estatísticos e para o melhoramento genético.

Esse novo paradigma trata da inteligência artificial que vem sendo cada vez mais utilizada e tem permitido, com o avanço da tecnologia e do entendimento da neurociência, a criação de modelos de neurônios artificiais muito próximos dos neurônios biológicos que conseguem se conectar formando as Redes Neurais Artificiais. Essas conexões as tornam capazes de análises de diversas situações, aprendizagem, reconhecimento de padrões e generalização (Braga et al., 2007).

Diferentemente das outras modelagens estocásticas utilizadas, os princípios de aprendizado e de inteligência computacional de um conjunto amplo de informação do desempenho do genótipo envolvendo médias, máximos, mínimos, variância e toda ordem de informação possível de ser direta ou indiretamente mensurada. Assim, as redes neurais funcionam com o um cérebro humano, captam toda informação disponível para gerar um critério de tomada de decisão (Silva, 2014).

O desempenho da rede é determinado pelas conexões entre os seus elementos. E, por isso, pode-se treinar uma rede neural para executar uma função particular ajustando-se os valores das conexões entre os elementos (Haykin, 2001). Esse processo permite uma adaptação da RNA às particularidades de um problema, que a torna capaz de generalizações, permitindo as mesmas respostas para estímulos similares.

As RNAs caracterizam-se pela sua arquitetura e pelo ajustamento de seus pesos às conexões durante o processo de aprendizado. A arquitetura de uma rede neural é definida pelo número de camadas (camada única ou múltiplas camadas), pelas conexões entre camadas, pelo número de neurônios em cada camada, pelo tipo de conexão entre eles (feedforward ou feedback) e pelo algoritmo de aprendizado (Haykin, 2001). Exemplo de uma arquitetura de uma rede neural evidenciando camadas, entradas e saída (representada pelo valor genético) é apresentado a seguir:



Dentre as vantagens da utilização da RNAs, ressaltam-se duas: primeiramente, a sua estrutura não linear, capaz de "captar complexas características entre o conjunto de dados de entrada" (Galvão et al., 1999); e em segundo lugar, a sua capacidade de não requerer informação detalhada sobre os processos físicos do sistema a ser modelado (Sudheer et al., 2003). Talvez por esses motivos, a utilização das redes neurais tem se mostrado mais promissora devido a possibilidade de um desempenho superior aos modelos convencionais utilizam.

Dentre os aplicativos computacionais destacam o R e o Matlab. Este último dispõe de uma biblioteca bastante abrangente de funções matemáticas, geração de gráficos e manipulação de dados que auxiliam muito o trabalho do programador. Também possui vasta coleção de bibliotecas -denominadas *toolboxes*- para áreas específicas como: equações diferenciais ordinárias, estatística, processamento de imagens, processamento de sinais e finanças e, em especial, redes neurais. Deve-se também mencionar o aplicativo Genes que permite a integração com o Matlab permitindo aproveitar seus recursos priorizando o seu em análises relacionadas a inteligência computacional envolvendo abordagens de redes neurais e lógica Fuzzy no melhoramento genético.

Bioinformática fundamentada em análises de simulação

Uma das grandes contribuições da informática é viabilizar o estudo de fenômenos, por meio da simulação de uma situação complexa, em que são estabelecidos parâmetros e restrições, de tal forma que o efeito de certos fatores controláveis possam ser convenientemente estudados. A simulação tem sido definida como a maneira de imitar, por meio de recursos computacionais, o comportamento de um sistema real, para estudar seu funcionamento em condições alternativas (Dachs, 1988), envolvendo certos tipos de

modelos lógicos, que permitam descrever, da melhor forma possível, o sistema natural (Naylor, 1971).

Quando se faz estudos baseados em simulação é necessário considerar que nenhuma parte do universo é tão simples que possa ser compreendida e controlada sem abstração. A abstração, segundo os autores, consiste em substituir a parte do universo por um modelo semelhante, porém com estrutura mais simples. Portanto, outro aspecto importante na simulação é a modelagem. O modelo deve ser suficientemente simples para ser operacionalizado, e interpretado adequadamente, mas seu desempenho deve ser comprável com o modelo real e, se a defasagem for grande, ele deve ser eliminado ou refinado. McNitt (1985) lembra que a simulação envolve modelos que representam a entidade a ser investigada, porém é mais que um simples modelo, é uma metodologia para avaliação destes.

A simulação inclui modelos físicos, matemáticos, biológicos e computacionais, dentre outros. Estes modelos podem também ser classificados como estáticos ou dinâmicos. Os dinâmicos diferem dos estáticos por envolver um conjunto de mudanças, inseridas de forma discreta ou em fluxo contínuo, representando o comportamento e a evolução do fenômeno estudado em resposta às variações no tempo ou no espaço. A adoção de modelos estáticos ou dinâmicos depende do próprio objetivo da simulação.

Apesar da modelagem anteceder o computador, seu uso e sua importância têm crescido consideravelmente com a disponibilização de computadores digitais, de alta velocidade. Atualmente a disponibilidade de recursos computacionais não está restrita a uma elite técnica ou instituição com grandes recursos financeiros. A viabilidade em termos de custo e qualidade de equipamento, tem permitido que pesquisadores tenham acesso contínuo e direto a recursos computacionais poderosos, permitindo maior prestação de serviços à área científica.

A simulação tem sido de grande utilidade em estudos genéticos sob vários contextos, incluindo estudos de populações, do indivíduo ou do próprio genoma. Ela demanda dos geneticistas o desenvolvimento de modelos biológicos adequados, que retratem da melhor forma possível os fenômenos de interesse, e dos programadores as rotinas para o processamento adequado segundo parâmetros e restrições, para que a influência de certos fatores possa ser avaliada.

Algumas vezes a simulação é realizada para encontrar uma solução ou um valor ótimo, a partir de um conjunto fixo de fatores, dentro de um modelo. Assim, pode-se estar interessado em prever o comportamento da variância genotípica, aditiva, e devido a

dominância considerando um loco gênico, em populações em equilíbrio de Hardy-Weinberg, com diferentes graus de dominância. Em outras, o objetivo pode ser o de monitorar o comportamento de um dado sistema sob condições variáveis. Os resultados destas observações permitem, em muitos casos, fazer comparações e previsões de técnicas alternativas, disponíveis em condições reais. Alguns estudos de simulação já foram realizados avaliando-se a eficiência de diferentes estratégias de melhoramento, em termos de ganhos de seleção. Cada ciclo de melhoramento pode ser avaliado ou monitorado sob diferentes contextos, como mudanças na composição genotípica, na freqüência alélica, na média, na variabilidade, no tamanho efetivo, dentre outros.

Sob contexto científico distinguem-se duas linhas de simulação. A primeira é a simulação para resolução de problemas matemáticos, completamente determinísticos, normalmente utilizados para obtenção de soluções de problemas de difícil trato matemático. A segunda é a simulação baseada em conceitos probabilísticos, que envolvem geração de números aleatórios com distribuição desejada. Os modelos determinísticos, que são não-randômicos, fornecem sempre o mesmo resultado para um conjunto de dados. Os estocásticos ou probabilísticos envolvem probabilidade, estimação de parâmetros e de tendência e seu resultado é incerto, mesmo que o esperado tenha sido previamente determinado.

A área de simulação em genética é tão vasta que deu origem a um conjunto de softwares, com características específicas. O programa DIALLEL, desenvolvido por Burow e Coors (1994), permite simular e analisar dados a partir de modelos aplicados a análise dialélica. O programa MENDEL, destina-se a simulação considerando as informações em nível de genes em indivíduos ou populações. Euclydes (1996) descreve o programa GENESYS, que permite a simulação de genomas de certa complexidade para estudos comparativos de métodos de seleção, testes de pressuposições e avaliação da eficácia de novas metodologias de seleção, inclusive daquelas assistidas por marcadores moleculares. O programa GENES permite, por simulação, estabelecer o número ótimo de famílias ou de plantas dentro de parcelas para representar uma população. O programa GQMOL permite, também por simulação, estimar o número ótimo de marcas moleculares para caracterizar a diversidade genética entre os indivíduos de uma população.

Referências

ALMEIDA, I. F. **Métodos de estimação e validação na seleção genômica na presença ou ausência de correção de fenótipos.** 2013. 68 p. Tese (Doutorado em Genética e Melhoramento de Planta) –Universidade Federal de Viçosa. 2013.

AYRES, K.L.; BALDING, D.J., 1998. Measuring departures from Hardy–Weinberg: a Markov chain Monte Carlo method for estimating the inbreeding coefficient. **Heredity**, v.*80*, p.769-777.

BRAGA, A.P.; CARVALHO, A. P. L. F.; LUDERMIR, T. B. **Redes Neurais Artificiais - Teoria e aplicações** – 2. ed. Rio de Janeiro: LTV, 2011. 226p.

BUROW, M.D.; COORS, J.G. Diallel : a microcomputer program for the simulation and analysis of diallel crosses. **Agronomy Journal**, v.86,p.154-158.1994.

COCKERHAM, C.C., 1969. Variance of gene frequencies. **Evolution**, p.72-84. Vancouver, 1969.

DACHS, N. **Estatística computacional**. ed. São Paulo: Livros Técnicos e Científicos Editora. S.A. 236p. 1998.

EBERHART,S.A., RUSSELL,W.A. Stability parameters for comparing varieties. **Crop Science.,** Madison, v.6, p.36-40, 1966.

EUCLYDES, R.F. **Uso de sistema Genesys na avaliação de métodos de seleção clássicos e associados a marcadores moleculares**. 1996. 135 p. Tese (Doutorado em Genética e Melhoramento de Planta) – Universidade Federal de Viçosa, Viçosa. 1996.

FRARY, A.; NESBITT, T.C.; GRANDILLO, S.; KNAAP, E.; CONG, B.; LIU, J.; MELLER, J.; ELBER,R.; ALPERT, K.B.; TANKSLEY, S.D. Fw2.2: A Locus Key Quantitative Trait para a Evolução da Tomato Fruit Tamanho. **Science**. 289:85-88, 2000.

GALVÃO, C. O.; VALENÇA, M. J. S.; VIEIRA, V. P. P. B.; DINIZ, L. S.; LACERDA, E. G. M.; CARVALHO, A. C. P. L. F.; LUDERMIR, T. B. (1999). **Sistemas inteligentes: Aplicações a recursos hídricos e ciências ambientais.** UFRGS/ABRH, Porto Alegre, 246p.

GIANOLA, D.; FERNANDO, R.L. Bayesian methods in animal breeding theory. **Journal of Animal Science,** v.63, p.217-244, 1986.

HAYKIN, S. S. **Redes Neurais: princípios e práticas** / Symon Haykin; trad Paulo Martins Engel. 2 ed. – Porto Alegre: Bookman, p.900. 2001.

HENDERSON CR. Best linear unbiased estimation and prediction under a selection model. **Biometrics**. 1975 Jun 1:423-47.

HOLSINGER, K.E.; WALLACE, L.E. Bayesian approaches for the analysis of population a genetic structure: an example from *Platanthera leucophaea* (Orchidaceae). **Molecular Ecology**, v.13, n.4, p.887-896, 2004.

HOLSINGER, B. (2005). **The premodern condition: Medievalism and the making of theory**. University of Chicago Press.

JEFFREYS, H. (1961). Theory of Probability. Clarendon Press.

KARHU, A. (2001). Estimation of inbreeding in radiata pine populations using microsatellites. **Finland: University of Oulu**.

LIU, J.; VAN ECK, J.; CONG, B.; TANKSLEY, S.D A new class of regulatory genes underlying the cause of pear-shaped tomato fruit. Proc Natl Acad Sci USA, v.99, p. 13302-13306, 2002.

MCNITT, L.L. **Simulação em Basic**. Livros Técnicos e Científicos Editora S.A. RJ. 348p. 1985.

MEUWISSEN, T. H. E.; HAYES, B. J.; GODDARD, M. E. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. **Genetics**. v.157, p. 1819-1829, 2001.

NASCIMENTO, M.; PETERNELLI, L. A.; CRUZ, C. D.; NASCIMENTO, ANA CAROLINA CAMPANA; FERREIRA, R. P.; BHERING, L. L.; SALGADO, C. C. Artificial neural networks for adaptability and stability evaluation in alfafa genotypes. **Crop Breeding and Applied Biotechnology**, v.12, p.152-156, 2013.

NAYLOR, T.H.; BALINTFY, J.L.; BURDICK, D.S.; CHU, K. **Técnicas de simulação em computadores**. Ed. Vozes LTDA, São Paulo, 401p. 1971.

RESENDE JÚNIOR, M.F.R. **Seleção genômica ampla no melhoramento vegetal**. Viçosa: Universidade Federal de Viçosa, 2010. 76 p. Dissertação (Mestrado em Genética e Melhoramento de Planta) –Universidade Federal de Viçosa, Viçosa. 2010.

RESENDE, M.D.V., SILVA, F.F., LOPES, P.S., AZEVEDO, C. F. 2012. *Seleção Genômica Ampla (GWS) via Modelos Mistos (REML/BLUP),* **Inferência Bayesiana (MCMC), Regressão Aleatória Multivariada (RRM) e Estatística Espacial**. Viçosa: Universidade Federal de Viçosa/ Departamento de Estatística. 291p.

RESENDE, M. D. V. **Genômica quantitativa e seleção no melhoramento de plantas perenes e animais**. Colombo: Embrapa Florestas, p. 330, 2008.

RESENDE, M.D.V. **Matemática e estatística na análise de experimentos e no melhoramento genético**. Colombo: Embrapa Florestas, p. 560, 2007.

RESENDE, M.D.V. de. **Selegen–Reml/Blup: Sistema Estatístico e Seleção Genética Computadorizada via Modelos Lineares Mistos**. Colombo: Embrapa Florestas, p. 361, 2007.

RESENDE, M.D.V. de; DUARTE, J.B. Precisão e controle de qualidade em experimentos de avaliação de cultivares. **Pesquisa Agropecuária Tropica***l*, v.37, p.182-194, 2007.

ROSA, G. J. M. (1998). Analise Bayesiana de Models Lineares Mistos Ro- bustos Via Amostrador de Gibbs. Ph. D. thesis, Escola Superior de Agricultura Luis de Queiroz, Piracicaba, Sao Paulo, Brazil. –Universidade Federal de Viçosa. 2013.

R Development Core Team (2013) **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

SILVA, G. N. **Redes Neurais Artificiais: novo paradigma para a predição de valores genéticos**. 2013. 106 p. Dissertação (Mestrado em Estatística Aplicada e Biometria). Universidade Federal de Viçosa, Viçosa 2013.

SUDHEER, K.; GOSAIN, A.; RAMASASTRI, K. Estimating actual evapotranspiration from limited climatic data using neural computing technique. **Journal of Irrigation and Drainage Engineering**,v. 129, n. 3, p. 214-218, 2003.

YANO, M.; KATAYOSE, Y.; ASHIKARIB, M.; YAMANOUCHI, U.; MONNA, L.; FUSE, T.; BABA, T.; YAMAMOTO, K.; UMEHARA, Y.; NAGAMURA, Y.; SASAKIA, T. Hd1, a major photoperiod sensitivity quantitative trait locus in rice, is closely related to the arabidopsis flowering time gene CONSTANS. **Plant Cell**., v.182: p.2473-2483, 2000.

VASQUEZ, JOHN A. **The war puzzle revisited.** Cambridge University Press, 2009.

VENTURA, R.; SILVA, M.; MEDEIROS, T.; DIONELLO, N.; MADALENA, F.; FRIDRICH, A.; VALENTE, B.; SANTOS, G.; FREITAS, L.; WENCESLAU, R. Use of artificial neural networks in breeding values prediction for weight at 205 days in Tabapuã beef cattle. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, v.64, n.2, p.411-418, 2012.