Universidade Federal de Minas Gerais
Escola de Veterinária

Busca por estruturas causais fenotípicas com a utilização de informações genômicas

Raphael Rocha Wenceslau

2014

Raphael Rocha Wenceslau

Busca por estruturas causais fenotípicas com a utilização de informações genômicas

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

Área de Concentração: Genética e Melhoramento Animal

Prof. Orientador: Martinho de Almeida e Silva

Belo Horizonte
2014

_____

Prof. Martinho de Almeida e Silva

(Orientador)

_____

Prof. Aldrin Vieira Pires

_____

Prof. Idalmo Garcia Pereira

_____

Prof. José Aurélio Garcia Bergmann

_____

Prof. Robledo de Almeida Torres

À minha família,

Dedico.

There are more things in heaven and earth, Horatio,

Than are dreamt of in your philosophy.

William Shakespeare. *Hamlet.* Act 1, scene 5, Hamlet to Horatio.

Obrigado,

a Deus,
à minha família,
à Vivian Paula Silva Felipe,
ao Professor Martinho de Almeida e Silva,
ao Professor Guilherme Jordão Magalhães Rosa,
ao Professor Fabyano Fonseca e Silva,
ao Bruno Dourado Valente,
aos Professores Robledo de Almeida Torres, José Aurélio Garcia Bergman, Idalmo Garcia Pereira e Aldrin Vieira Pires,
aos demais Professores da Escola de Veterinária da Universidade Federal de Minas Gerais,
aos funcionários da Escola de Veterinária da Universidade Federal de Minas Gerais,
aos colegas da Escola de Veterinária da Universidade Federal de Minas Gerais,
aos meus amigos,
à Universidade Federal de Minas Gerais,
à University of Wisconsin – Madison,
à CAPES, CNPq e FAPEMIG.

Muito obrigado a todos pelos ensinamentos e apoio.

SUMÁRIO

LISTA DE TABELAS

LISTA DE FIGURAS

INTRODUÇÃO

A tentativa de conhecer a relação causal entre eventos subsequentes é desejo básico nas diversas áreas de conhecimento, assim como, nas tarefas cotidianas do ser humano. Frequentemente utiliza-se a expressão "por quê" ou "e se" tentando se referir à relação de causa de eventos ocorridos ou que aconteceriam caso houvesse intervenção externa no sistema de variáveis observadas. Na área da produção animal, o conhecimento da estrutura causal existente entre os fenótipos de características economicamente importantes possui implicação direta na decisão da melhor maneira de agir e fazer intervenções diante de determinado sistema de produção. Ainda nessa linha de raciocínio, pode-se imaginar um sistema de produção de bovinocultura, de corte ou leite, em que características de reprodução, como a idade ao primeiro parto ou o intervalo de partos apresentem grande valor econômico e relação causal com outras características de produção, como, por exemplo, quilogramas de leite produzidos ao longo da vida do animal ou a quantidade de quilogramas de carne produzida. Práticas comuns no manejo reprodutivo, como a inseminação artificial ou a transferência de embriões são intervenções no sistema que podem fazer os valores fenotípicos das características reprodutivas se modificarem e, assim, causar mudanças indiretas, também, nas características de produção. Caso a estrutura e a magnitude das relações de causa entre os fenótipos sejam conhecidas, as mudanças indiretas geradas nas características em virtude de práticas, como as descritas anteriormente, podem ser previstas. Também, na área de melhoramento animal, o valor genético de um animal poderia ser predito diante dos dois sistemas distintos, com ou sem intervenção, mesmo antes da mesma ser praticada. Fato importante, pois, pode haver diferenças significativas no ordenamento dos animais a serem selecionados para reprodução de acordo com o sistema de produção.

Para obtenção da rede de relações causais entre características algumas metodologias foram desenvolvidas, dentre elas, o algoritmo *Inductive Causation* –IC (Verma e Pearl, 1990) cuja utilização na área de produção animal foi proposta por Valente et al. (2010). Para a aplicação do algoritmo é necessário obter a matriz de variância residual que é acessada a partir do ajuste de modelos mistos conhecidos no melhoramento animal. Nesse, a matriz de parentesco dos coeficientes de Wright (matriz A) é utilizada nas equações de resolução do modelo. Essa matriz expressa a semelhança genética entre os animais incluídos na avaliação, ou seja, mede a covariância entre as observações dos animais a partir do cálculo de uma semelhança genética esperada. Porém, uma vez que informações moleculares do tipo SNP (*Single*

Nucleotide Polymorphism) estejam disponíveis, poderia ser formada uma matriz de parentesco realizada (matriz G), que demonstraria o verdadeiro compartilhamento genético entre os animais. A utilização dessa nova matriz poderia ser feita com o intuito de gerar informação mais segura e completa ao modelo utilizado, uma vez que já contaria os efeitos da segregação independente dos alelos, além de evitar desvios atribuídos a erros de anotação de *pedigree*. Diante das informações expostas até então, o presente trabalho propôs a extensão do estudo de Valente et al. (2010), por meio da adição das informações genômicas afim de se obter melhores resultados na busca da estrutura causal existente entre fenótipos.

No primeiro capítulo serão discutidos, em forma de revisão bibliográfica, assuntos importantes relativos às metodologias utilizadas ao longo do trabalho, dentre eles: o Modelo de Equações Estruturais, o Algoritmo *Inductive Causation*, a formação e utilização da matriz de relacionamento genético genômico. Já, no segundo capítulo da tese será relatada a pesquisa científica propriamente dita, com material e métodos, discussão dos resultados observados e conclusões.

CAPÍTULO 1

REVISÃO DE LITERATURA

1. Inferência causal: modelos de equações estruturais, algoritmo IC e aplicação na ciência animal

Uma tarefa comum do homem é tentar encontrar uma explicação coerente e satisfatória para determinado conjunto de observações, assim como propor relação entre elas. Essa relação pode ser descrita basicamente com duas diferentes visões, como medida de associação entre as variáveis ou com a existência de relação de causa-efeito entre elas.

Hume, citado por Gillies (2002), define causa como a ação existente de um objeto seguido por outro, em que caso seja encontrado um objeto similar ao primeiro, ele deve estar seguido de um objeto similar ao segundo. O autor, portanto, sugere que causa é uma relação de determinação. Outros conceitos de causa surgem, como, a causa não determinante, em que é possível haver a associação de causa e efeito e que o primeiro objeto seja causa de um segundo, mas não é necessária a presença do segundo objeto caso o primeiro ocorra, invocando, portanto, a presença de uma probabilidade de ocorrência, que pode ser vista como uma influência ou propensão. Spirtes et al. (2000) define causa como sendo uma relação entre eventos particulares: alguma coisa acontece e causa alguma outra coisa a acontecer. Cada causa é um evento particular e cada efeito é um evento particular. Um evento A pode ter mais de uma causa, sendo nenhuma delas suficientes para causar A e pode ser sobre determinado, ou seja, ter mais de uma causa suficiente para que ele possa ocorrer. Abandonada a teoria conceitual, tem-se na ordem prática, de modo geral, o conhecimento de causa com duas importantes funções: 1) permitir prever resultados com base em observações; 2) fornecer a capacidade de controle de eventos (Blaisdell et al., 2006).

Em estudos empíricos, a intenção fundamental é predizer efeitos de mudanças, sendo essas mudanças ocorridas de forma natural ou impostas deliberadamente. Na ciência animal, assim como em outras áreas de conhecimento e disciplinas, o objetivo central dos estudos geralmente se refere à inferência dos efeitos causais. Por exemplo, pesquisadores ligados à nutrição, reprodução e imunologia animal estudam fatores que têm efeito sobre a produção

dos indivíduos, como, o nível nutricional da dieta sobre o peso de suínos, o período luminoso sobre a maturidade sexual de aves e a qualidade de água sobre a incidência de doenças em peixes (Rosa e Valente, 2012).

Uma maneira de inferir efeitos de causa entre variáveis são os experimentos aleatorizados. Esses experimentos tentam controlar, por meio da casualização, todas as fontes de causa que não sejam o objeto de estudo por meio de casualização das unidades experimentais (animais, no caso da ciência animal) dentro dos grupos experimentais e distribuição também ao acaso dos tratamentos para esses grupos. Tais experimentos controlados são ferramenta poderosa para estudo de causalidade, permite saber a existência de efeitos de tratamentos, assim como a magnitude desses efeitos (Rosa e Valente, 2012). Porém, algumas críticas ainda são feitas na capacidade de se buscarem relações causais por meio de experimentos. A primeira seria com relação à capacidade de controle total de variáveis causais externas ao sistema de estudo. Por exemplo, imagine que o pesquisador estaria interessado no fato de que inalar fumaça de cigarro causa câncer nos pulmões. Para demonstrar tal fato, foi realizado um experimento inteiramente ao acaso, em que algumas pessoas foram amostradas ao acaso para fumar e outras para não entrar em contato com o cigarro. Imagine agora que o pesquisador não conhece que existe a presença de uma substância cancerígena no papel que embala o cigarro. Nesse caso, o ato de fumar e o câncer seriam estatisticamente dependentes, mesmo que inalar fumaça de cigarro sabidamente não cause câncer. Eles seriam dependentes porque existe um fator em comum entre eles. Dificilmente não existirá variável que foge do controle experimental (Spirtes et al., 2000). Em outras ocasiões, os experimentos inteiramente ao acaso não são viáveis, por exemplo, em razão de impedimentos legais, éticos ou logísticos. Às vezes, também, os experimentos devem ser realizados em condições extremamente específicas, o que dificulta a extrapolação dos resultados para condições reais (Rosa e Valente, 2012).

Uma alternativa aos dados experimentais seria a utilização de dados observacionais. Na área de estudo da veterinária e zootecnia, esses dados são coletados rotineiramente a campo e representam importante fonte de informação que podem ser usadas para inferência causal entre características. Os resultados obtidos poderiam servir para conhecimento dos efeitos causados por alteração no ambiente de criação, como por exemplo, mudança no manejo nutricional dos animais.

Apesar de não muito conhecidas, já existem ferramentas disponíveis para estudo das relações causais entre variáveis com base em dados observacionais. Será descrito, a seguir, um algoritmo de busca de estruturas causais entre fenótipos, a aplicação de uma metodologia que torna possível a medida de força da relação de causa entre variáveis, além de exemplos de utilização desses modelos na ciência animal. Antes de revisar algumas metodologias para descrever relações de causas entre variáveis, haverá a tentativa de elucidar o significado de parâmetros usados frequentemente na área de estatística em contraste com a noção de causalidade.

1.1 Medidas de associações e causalidade

Correlações são utilizadas para descrever associação estatística entre duas variáveis. Podem ser obtidas, por exemplo, na área de melhoramento animal, por meio de modelos multicaracterísticas tradicionais na avaliação genética dos animais. O conhecimento considerando a associação entre variáveis é certamente importante, e pode ser usado para inferir quão prováveis são determinados eventos. No entanto, essa informação não é suficiente para predizer como as probabilidades irão mudar como resultado de intervenções externas. Portanto, para elucidar relações causais as correlações são de utilidade limitada. Isso é atribuído ao fato que correlações não somente confundem associações diretas e indiretas, mas também não são capazes de distinguir variáveis e covariáveis, e então, causa e efeito (Opgen-Rhein e Strimmer, 2006).

Segundo Rosa e Valente (2012), um provérbio conhecido na comunidade estatística é: "*correlation does not imply causation*", traduzindo, correlação não implica causalidade. A frase enfatiza o argumento de que o conhecimento da correlação entre duas variáveis não é suficiente para descrever uma relação de causas entre elas. Diferentes tipos de associações causais podem ser fontes de correlação entre duas variáveis, x e y. Por exemplo, uma correlação poderia surgir a partir do efeito causal de x em y, representado como $x \rightarrow y$, ou por um efeito causal de y em x, representado por $x \leftarrow y$, ou até como o efeito de outra variável, z, afetando x e y ($x \leftarrow z \rightarrow y$). Uma correlação observada entre x e y pode ser explicada por esses três potenciais efeitos causais, ou seja, nenhuma das três hipóteses pode ser confirmada simplesmente por meio do conhecimento da correlação entre as variáveis. Como visto o conhecimento de associação estatística isoladamente oferece pouca ajuda no planejamento de práticas de manejo e na predição de efeitos de intervenções externas no sistema de produção

animal. Considerando-se esta questão, pode-se dizer que existe muito mais a aprender de observações do que simplesmente correlações e covariâncias entre variáveis. Portanto, obter valores que traduzem as relações causais é necessário (Pearl, 2009).

1.2 Inferindo efeitos causais de dados observacionais

A missão de se inferirem efeitos causais a partir de dados observacionais é sempre muito desafiadora e polêmica. Uma das razões para isso é que exige premissas adicionais comparada às inferências estatísticas tradicionais que não incorporam significado causal. Entretanto, as informações oriundas desses dois tipos de inferência são bem diferentes e não podem ser comparadas. Estatísticas tradicionais essencialmente descrevem a plausibilidade de determinado evento ocorrer, esse evento pode ser multivariado. Por outro lado, a informação causal permite a descrição de como o valor de uma variável é influenciado pelo valor de outras variáveis. A aplicação prática disso é que também é possível a predição de intervenções externas físicas na estrutura de relacionamento causal, o que pode ser muito útil no manejo de animais de produção. Tal predição não é possível de ser obtida de uma distribuição conjunta desse mesmo conjunto de variáveis (Spirtes et al., 2000; Pearl et al., 2009; Rosa e Valente, 2012). Diferentes modelos como Redes Bayesianas (*Bayesian Networks,* BN) e o modelo de equações estruturais (*Structural Equation Model*, SEM) podem ser usados para expressar relacionamentos causais entre diversas variáveis. O ajuste desses modelos permite que seja feita a inferência de efeitos de causa a partir de dados observacionais e condicionais a uma estrutura conhecida *a priori*. A interpretação causal derivada do ajuste de um modelo como, SEM ou BN, depende não apenas de premissas estatísticas, mas também de premissas causais. Por exemplo, inferências obtidas com o ajuste de BN assumem que a estrutura causal entre as variáveis é acíclica e que todas as variáveis que possuem efeito causal em duas ou mais variáveis contidas em BN estão representadas em BN. Premissas idênticas podem ser aplicadas em SEM e podem ser traduzidas como o uso de estruturas causais acíclicas e construção de uma matriz de covariância residual diagonal, o que é suficiente para garantir que qualquer relação de recursividade presente seja identificável. No entanto, essa premissa não é obrigatória para ajuste do modelo de equações estruturais, que pode representar relações simultâneas (de *feedback*) e considerar associações causadas por variáveis escondidas (*hidden*). No entanto, a identificabilidade do modelo não é garantida e deve ser verificada se essas características forem permitidas (Gianola e Sorensen, 2004). Em geral, a utilização de um desses modelos para fazer inferências a respeito de relações de causa exige importantes

premissas, como já citadas, especialmente considerar que associações entre variáveis mensuradas não são confundidas pela existência de variáveis desconhecidas que apresentam efeitos causais em duas ou mais outras variáveis contidas no sistema. Alternativamente, assume-se pelo menos que as trilhas (*paths*) que confundem a inferência causal e que estão escondidas são parcialmente conhecidas e que seja possível a mensuração das variáveis dessas trilhas. Uma crítica à inferência causal baseada em dados observacionais é que não é possível estar totalmente confiante de que não há fontes de confudimento ou que todas essas fontes estão controladas. Apesar de a crítica ser razoável, não significa que a busca por conhecimento mais aprofundado da relação entre variáveis deve ser abandonada. Existem vários graus de conhecimento ou grau de confiança entre a absoluta certeza sobre a relação ou a absoluta ignorância. Além disso, métodos puramente estatísticos, como a realização de experimentos inteiramente ao acaso, às vezes, não são alternativas, e, no mínimo, estudos com dados observacionais poderiam elucidar alguns aspectos da relação entre variáveis (Rosa e Valente, 2012).

1.3 Modelos de equações estruturais

Os modelos multivariados possuem grande importância na genética aplicada, evolucionária e quantitativa. No melhoramento animal, o relacionamento entre característica normalmente é estudado por meio do ajuste de modelos multicaracterísticas, estimando, por exemplo, parâmetros de covariância e correlação. Convencionalmente, a correlação genética é definida como a proporção da variância que duas características dividem em razão de causas genéticas, e isso indica o quanto que a influência genética em duas características é comum às duas. O conhecimento da correlação genética pode ser usado para decisões de seleção e maximizar o ganho genético de programas de melhoramento (Rosa et al., 2011).

Uma alternativa ao tradicional modelo multicaracterísticas é o Modelo de Equações Estruturais (SEM), que pode ser aplicado para o estudo de relacionamentos entre fenótipos dos tipos recursivos ou simultâneos em sistema multivariados. A modelagem por estruturas causais foi desenvolvida primeiramente por Wright (1921) e Haavelmo (1943) de forma que informações qualitativas de causa-efeito pudessem ser combinadas com procedimentos estatísticos para prover medida quantitativa das relações causa-efeito entre variáveis de interesse. De acordo com os autores, a condição que faz a equação de regressão $y = \beta x + e$ ser

estrutural, é que precisamente a conexão causal entre x e y é dada por β. Este último define a sensibilidade de $E(y)$ às manipulações experimentais de x (Pearl, 2009). O modelo de equações estruturais proporciona técnica estatística para estimar e testar relacionamentos funcionais entre características que não são reveladas pelos modelos lineares padrões. Quando se ajusta o SEM para um conjunto de variáveis é necessário que se defina *a priori* para cada variável o subconjunto de variáveis remanescentes que possuem efeito causal nele. Essa informação é chamada "estrutura causal" e pode ser representada por um gráfico direcionado em que as variáveis constituem vértices e as relações causais são representadas como linhas direcionadas (setas) entre os vértices. Por exemplo, considere o gráfico na Figura 1 adaptado de Rosa et al. (2011).

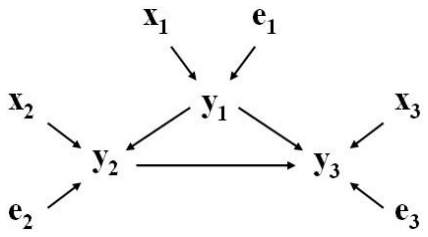


Figura 1. Exemplo de estrutura causal, em que *y*'s representam medidas de três características fenotípicas, *x*'s e *e*'s representam variáveis explanatórias conhecidas e fatores residuais que influem *y*'s, respectivamente.

O gráfico da Figura 2 pode ser representado por um conjunto de equações estruturais, dado por:

$$\begin{cases} y_1 = \beta_1 x_1 + e_1 , \\ y_2 = \lambda_{21} y_1 + \beta_2 x_2 + e_2 , \\ y_3 = \lambda_{31} y_1 + \lambda_{32} y_2 + \beta_3 x_3 + e_3 , \end{cases}$$

em que os β's são os parâmetros do modelos que representam os efeitos fixos das covariáveis x's em *y*'s, e λ's são os coeficientes estruturais representando a magnitude dos efeitos causais entre y's. Em notação matricial o modelo de equação estrutural pode ser representado como, $y = \Lambda y + x\beta + e$, em que Λ é uma matriz quadrada com zeros na diagonal e com coeficientes estruturais λ ou zeros nos elementos fora da diagonal. y, x, β e e são vetores ou matrizes de observações, variáveis exógenas, parâmetros do modelo e resíduo, respectivamente. As equações demonstradas podem ser interpretadas de forma que o lado esquerdo é determinado como uma função, geralmente linear, das variáveis do lado

direito. Além disso, o sinal de igualdade representa uma relação assimétrica definida como "é determinado por", que difere do conceito padrão de igualdade (Gianola e Sorensen, 2004; Pearl, 2009; Rosa et al., 2011; Rosa e Valente, 2012).

1.4 SEM utilizado dentro do modelo misto em genética quantitativa

O modelo de equações estruturais recebeu pouca atenção na genética quantitativa até recentemente. No entanto, vários sistemas biológicos poderiam ser estudados mais apropriadamente com o SEM do que os modelos lineares tradicionais. Por exemplo, animais com alta produção de leite apresentam maior predisposição à incidência de doenças do úbere e, em contrapartida, a presença de doenças tem efeito sobre a produção de leite (Wu et al., 2010). Gianola e Sorensen (2004) apresentaram um modelo misto linear que permite relações recursivas e de simultaneidade entre fenótipos envolvidos em um sistema multivariado assumindo herança genética aditiva. Nesse modelo, uma ou mais variáveis aparecem como variáveis explanatórias nas equações para outras variáveis respostas. Eles demonstraram que os parâmetros estruturais que definem os sistemas recursivos e simultâneos têm consequência na interpretação dos parâmetros genéticos. Desde que o trabalho foi publicado, o interesse no modelo de equações estruturais têm aumentado, sendo aplicado em estudos de diferentes espécies e características.

Um modelo com uma específica estrutura causal e efeitos genéticos aditivos aleatórios pode ser escrito como:

$y = \Lambda y_i + X_i\beta + u_i + e_i$,

em que y_i é o vetor (t x 1) de dados fenotípicos para o objeto i; Λ é a matriz de coeficientes estruturais descrevendo a estrutura causal escolhida; $X_i\beta$ representa os efeitos das variáveis exógenas como regressões lineares, em que a matriz X_i contém as covariáveis e β é um vetor dos coeficientes de regressão fixos. u_i e e_i são vetores (t x 1) de efeitos genéticos aditivos e resíduos do modelo, respectivamente, que são ambos associados ao objeto i e são distribuídos, como, $\begin{bmatrix} u_i \\ e_i \end{bmatrix} \sim N\left\{\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} G_0 & 0 \\ 0 & \Psi_0 \end{bmatrix}\right\}$, em que G_0 e Ψ_0 são as matrizes de

covariâncias genéticas aditivas e de resíduo, respectivamente. O modelo para *n* animais pode ser, então, descrito como $y = (\Lambda \quad I_n)y + X\beta + Zu + e$, com $\begin{bmatrix} u \\ e \end{bmatrix} \sim N\left\{\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} G_0 \otimes A & 0 \\ 0 & \Psi_0 \otimes I_n \end{bmatrix}\right\}$, em que *y*, u e e são vetores de observações, efeitos genéticos aditivos e resíduos do modelo ordenados por característica e indivíduo dentro de característica, enquanto *X* e *Z* são matrizes de incidência dos efeitos β e u no vetor *y*. Esse modelo pode ser reescrito como $[I_{tn} - (\Lambda \quad I_n)]y = X\beta + Zu + e$, então um modelo equivalente reduzido pode ser obtido, ainda de acordo com Gianola e Sorensen (2004), como,

$$y = [I_{tn} - \Lambda \quad I_n)]^{-1} X\beta + [I_{tn} - \Lambda \quad I_n)]^{-1} Zu + [I_{tn} - \Lambda \quad I_n)]^{-1} e .$$

A distribuição amostral resultada para y dado os parâmetros de local e a matriz de covariância residual é:

$p(y \mid \Lambda, \beta, u, \Psi_o) \sim N\left\{[I_{tn} - \Lambda \quad I_n)]^{-1}(X\beta + Zu), [I_{tn} - \Lambda \quad I_n)]^{-1} \Psi_o [I_{tn} - \Lambda \quad I_n)]'^{-1}\right\}$, em que $\Psi = \Psi_o \quad I_n$. Observa-se que no modelo reduzido, o sistema é resolvido para as variáveis resposta. Assim, os parâmetros de local e dispersão são transformados em parâmetros de um modelo multicaracterísticas padrão:

$$y = (I_t - \Lambda)^{-1} X_i\beta + (I_t - \Lambda)^{-1} u_i + (I_t - \Lambda)^{-1} e_i = \mu_i + u_i + e_i ,$$

em que, $\mu_i = (I_t - \Lambda)^{-1} X_i\beta$, $u_i = (I_t - \Lambda)^{-1} u_i$ e $e_i = (I_t - \Lambda)^{-1} e_i$. μ_i, u_i e e_i são respectivamente vetores de efeitos fixos, efeitos genéticos aditivos e resíduos de um modelo que não considera elos funcionais entre as variáveis respostas. Adicionalmente, pode-se representar a distribuição conjunta dos efeitos aleatórios do modelo multicaracterística padrão como:

$\begin{bmatrix} u_i^* \\ e_i^* \end{bmatrix} \sim N\left\{\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} G_0^* & 0 \\ 0 & R_0^* \end{bmatrix}\right\}$, com $G_0 = (I_t - \Lambda)^{-1} G_0 (I_t - \Lambda)'^{-1}$ e $R_0 = (I_t - \Lambda)^{-1} \Psi_0 (I_t - \Lambda)'^{-1}$. Desta forma, SEM's podem ser descritos como reparametrizações do modelo multicaracterísticas simples. As duas formas apresentadas são equivalentes, uma vez que geram a mesma distribuição para as variáveis respostas.

Apesar de o modelo multicaracterísticas ser justamente identificável, de forma que mudanças nos valores paramétricos resultem em mudança na distribuição conjunta de *y*, o SEM carrega parâmetros extras em Λ, resultando em não identificabilidade da função de verossimilhança (Rosa et al., 2011). É possível, no entanto, fazer restrições no modelo para ativar a identificabilidade dos parâmetros. Duas principais estratégias podem ser utilizadas: Formar

combinações lineares de parâmetros em uma mesma equação ou entre equações, ou fazer restrição à matriz de covariância residual, assumindo, por exemplo, que a matriz de covariância residual Ψ_0 seja zero (Wu et al., 2010).

Parâmetros de locação como efeitos genéticos aditivos e parâmetros de dispersão como componentes de variância e covariância genética são considerados tanto no modelo de equações estruturais quanto no modelo multicaracterísticas. Apesar da equivalência apresentada entre o modelo multicaracterísticas e o SEM totalmente recursivo (em que todas as entradas abaixo da diagonal para Λ são definidas como parâmetros livres) deve-se tomar cuidado com as interpretações desses parâmetros que se modificam de acordo com o modelo. Para exemplificar, considere um modelo recursivo em que A tem efeito causal em B. Sob modelo recursivo, a correlação genética representa a associação linear entre dois efeitos genéticos aditivos não observáveis, cada um influindo diretamente uma característica específica. Porém, esta não seria a única fonte de correlação genética caso o modelo multicaracterísticas fosse utilizado, uma vez que existe no modelo amostral utilizado uma associação indireta entre o efeito genético de A e o fenótipo de B, mediada pelo fenótipo de A (o efeito genético de A possui efeito sobre o fenótipo de A, que por sua vez apresenta efeito causal sobre o fenótipo de B). A correlação genética sob o modelo recursivo não considera esse efeito. Ao contrário, o modelo multicaracterísticas considera toda a associação de origem genética, seja ela indireta ou direta. Em razão deste fato, torna-se possível que a correlação genética seja diferente de zero mesmo que os valores genéticos aditivos sejam independentes no contexto recursivo. Adicionalmente, observa-se que o efeito genético aditivo de A, poderia ser considerado fonte de influência apenas da característica A, se o coeficiente estrutural λ_{BA} fosse igual a zero (Gianola e Sorensen, 2004).

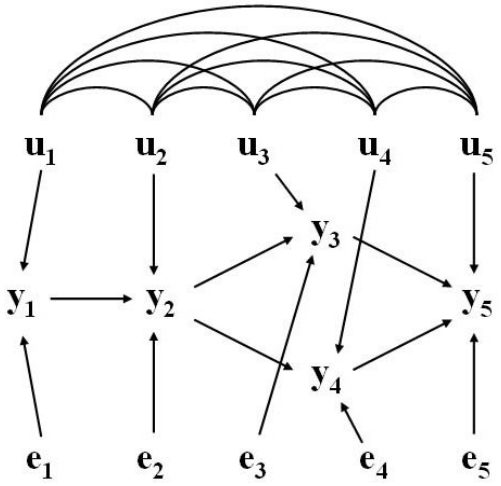


Figura 2. Exemplo de uma estrutura causal entre cinco características fenotípicas e seus efeitos aditivos genéticos (*u*'s) e residuais (*e*'s) correspondentes.

Os modelos de equações estruturais têm sido utilizados para estudar relações simultâneas e recursivas entre fenótipos em várias espécies e raças. Nos primeiros trabalhos utilizando dados reais e SEM, de los Campos et al. (2006a) e de los Campos et al. (2006b) estudaram efeitos simultâneos e recursivos entre a produção de leite e escores de células somáticas em vacas e cabras usando o modelo reprodutor. Os parâmetros foram estimados por meio de máxima verossimilhança e os modelos comparados por meio do critério BIC. Os resultados indicaram um efeito recursivo do escore sobre a produção de leite, com evidência de que a associação negativa entre produção de leite e saúde do úbere é mais provável de ser atribuida ao efeito da infecção sobre a produção de leite do que na direção contrária. Além desse resultado, os autores observaram relações simultâneas do efeito de escore de células somáticas entre as duas metades do úbere, esquerda e direita. Aspecto que indica que a infecção não é restrita a somente um lado do úbere. Wu et al. (2007) estenderam o modelo de Gianola e Sorensen (2004) para contabilizar relações heterogêneas entre produção de leite e escore de células somáticas para vacas de alta e baixa produção. Diferentes efeitos recursivos e simultâneos foram ajustados para as duas sub-populações. Os resultados observados confirmaram efeito negativo direto do escore de células somáticas sobre a produção de leite e pequenos efeitos recíprocos na direção oposta. Ainda, foi relatada diferença entre essas relações para os dois grupos estudados. Os efeitos da saúde do úbere sobre a produção foi maior para as vacas com alta produção de leite. Outro estudo utilizando SEM foi o de Varona et al. (2007) que usaram um modelo recursivo para analisar relações de causa entre o tamanho da leitegada e o peso ao nascer de leitões das raças Landrace e Yorkshire, assim como,

também, um modelo bi-característica tradicional. Eles reportaram diferenças marcantes entre as estimativas dos coeficientes estruturais entre as raças. Wu et al. (2008) propuseram modelo hierárquico limiar-normal com recursividade e simultaneidade para análise de relações entre características binárias e contínuas e utilizaram esse modelo para estudar as características presença de mastite clínica e produção de leite em vacas da raça Norwegian Red. Os primeiros 180 dias de lactação foram divididos em três períodos menores de 60 dias cada um, para investigação do relacionamento dessas características ao longo da lactação. O modelo recursivo mostrou efeitos negativos dentro de cada período da presença de mastite sobre a produção de leite em todos os períodos de lactação, e efeito positivo da produção de leite sobre a incidência de mastite clínica entre períodos. Os resultados sugerem efeitos desfavoráveis da produção na susceptibilidade à mastite e relações dinâmicas entre mastite e a produção de leite ao longo da lactação.

1.5 Busca por estruturas causais recursivas

A inferência de efeitos causais entre variáveis se baseia no fato de que a estrutura causal entre elas é conhecida, ou seja, para ajuste do SEM às observações que pertencem a um conjunto de características, é necessário definir *a priori* a estrutura causal entre as variáveis estudadas. Como visto, essa estrutura pode ser vista como um gráfico direcionado ou pela simples definição de quais entradas vão ser consideradas livres na matriz Λ. A seleção da estrutura causal a ser utilizada se torna um desafio ao ajuste do SEM na medida em que o número de estruturas possíveis aumenta consideravelmente quando se aumenta o número de variáveis no modelo. Desta forma a comparação exaustiva de modelos por critérios de comparação de modelos como o AIC ou o BIC é inviável. Os pesquisadores até então se utilizavam de conhecimento *a priori* para reduzir o número de hipóteses causais a serem testadas. Essa informação *a priori* pode ser proveniente de situações experimentais, conhecimento biológico das características estudadas ou até mesmo de uma sequência temporal em que os eventos ocorrem. Nos casos em que se admite que a relação entre as variáveis não seja conhecida, uma alternativa é utilizar algoritmos que permitem explorar espaços de estruturas causais, como *SGS*, *PC* e o *IC* (Verma e Pearl, 1990).

Metodologias como o algoritmo IC têm sido desenvolvidas para explorar a conexão entre estruturas causais recursivas e distribuições conjuntas a fim de recuperar gráficos acíclicos direcionados (DAG) subjacentes. Baseado em uma dada matriz de correlação, o algoritmo

questiona uma série de independências condicionais entre variáveis. Assumindo que essas independências refletem em d-separações (ver Anexo 1.1) nos DAGs, o algoritmo retorna um gráfico parcialmente orientado como resultado que representa uma classe de estruturas causais compatíveis com as independências condicionais obtidas. Essa classe normalmente já é uma restrição muito grande do espaço de hipóteses causais em relação ao espaço inicial (Spirtes et al., 2000; Rosa et al., 2011).

Considerando um conjunto *V* de variáveis aleatórias, o algoritmo IC pode ser descrito pelos seguintes passos:

1. Para cada par de variáveis *a* e *b* em *V*, procurar por um conjunto de variáveis S_{ab} tal que *a* é independente de *b* dado S_{ab}. Se *a* e *b* forem dependentes para todos os possíveis conjuntos condicionais, conecte *a* e *b* com uma linha não direcionada. Esse passo resulta em um gráfico não-direcionado *U*. O objetivo aqui é obter um gráfico que especifica variáveis adjacentes que na estrutura causal subjacente não são *d-separadas* (Anexo 1.1), portanto não são probabilisticamente independentes.
2. Para cada par de variáveis não adjacentes *a* e *b* com uma variável adjacente *c* em comum em *U* (por exemplo, *a – c – b*), procurar por um conjunto de variáveis S_{ab} que contém *c* de tal forma que *a* é independente de *b* dado S_{ab} . Se não existir, então adicionar cabeça de setas em direção a *c*. Se o conjunto existir então prossiga para o próximo passo. O passo tem como objetivo orientar linhas com base na procura de c*olliders*, que são vértices nos gráficos no qual setas convergem de ambos os sentidos em uma trilha. Estruturas internas em um gráfico compostas por um *collider* e que sofrem influências causais de dois vértices não conectados são chamados de *unshielded colliders* (Spirtes, 2000). Em tal estrutura, os pais são *d-separado* condicionalmente a pelo menos um conjunto de variáveis restantes no gráfico completo, mas não se o vértice *c* pertence a este conjunto. Condicionalmente a *c*, a trilha entre *a* e *b* por intermédio de *c* permite o fluxo de dependência, não ocorrendo *d-separação*. A consequência observacional é a dependência probabilística entre pais não adjacentes condicionalmente a todos os possíveis conjuntos de variáveis que contém o filho em comum.
3. No gráfico parcialmente orientado resultado, orientar o máximo de linhas possíveis de forma que não resulte em formação de novos *colliders* ou em ciclos. Em adição, restrições nesse passo devem ser feitas com base em informação *a priori*.

As decisões sobre declarar pares de variáveis como condicionalmente dependentes ou não são baseadas em correlações parciais obtidas da amostra, que envolve algum grau de incerteza. Para considerar isso, decisões estatísticas devem ser tomadas testando hipóteses de nulidade ou, na visão Bayesiana, usando intervalos de maior probabilidade para essas correlações.

O algoritmo IC é baseado na conexão entre a estrutura causal e a distribuição conjunta e requer algumas premissas. Talvez a mais forte seja a de suficiência causal, ou seja, assume-se que toda variável que influencia duas ou mais variáveis dentro do conjunto de variáveis estudadas já está contida no sistema. O resultado dessa premissa é a ausência de fontes de covariância residual entre as características. Como mencionado anteriormente, tal medida já é adotada nas aplicações de SEM para permitir que o modelo seja identificável. Portanto, as premissas do algoritmo IC não são mais fortes do que a premissa do modelo de equações estruturais já utilizado (Rosa et al., 2011).

1.6 Busca da estrutura causal no contexto de genética quantitativa

Para o ajuste de modelos de equações estruturais vistos sem a presença de variáveis escondidas, os resíduos dos modelos são considerados independentes e os efeitos recursivos são utilizados para descrever padrões de covariabilidade entre fenótipos observados. No entanto, em modelos mistos, a covariância entre características podem ser atribuída à causa direta entre os fenótipos ou ter origem genética, ou seja, os efeitos genéticos correlacionados podem ser fontes de confundimento na seleção da correta estrutura causal (Valente et al., 2010; Rosa e Valente, 2012). Na Figura 3 (adaptada de Valente et al., 2010) podem ser vistos cenários em que há relação de recursividade entre os fenótipos, mas a conexão entre a distribuição conjunta e a estrutura causal entre fenótipos não ocorre em razão do confundimento causado pela associação entre os efeitos genéticos aditivos. Por exemplo, na mesma Figura, em 3a, os fenótipos y_1 e y_2 não são marginalmente independentes. Nas outras figuras os mesmos fenótipos não são independentes condicionalmente a y_3. No entanto, informações de marcadores moleculares e de parentesco podem ser utilizadas para controlar esse confundimento.

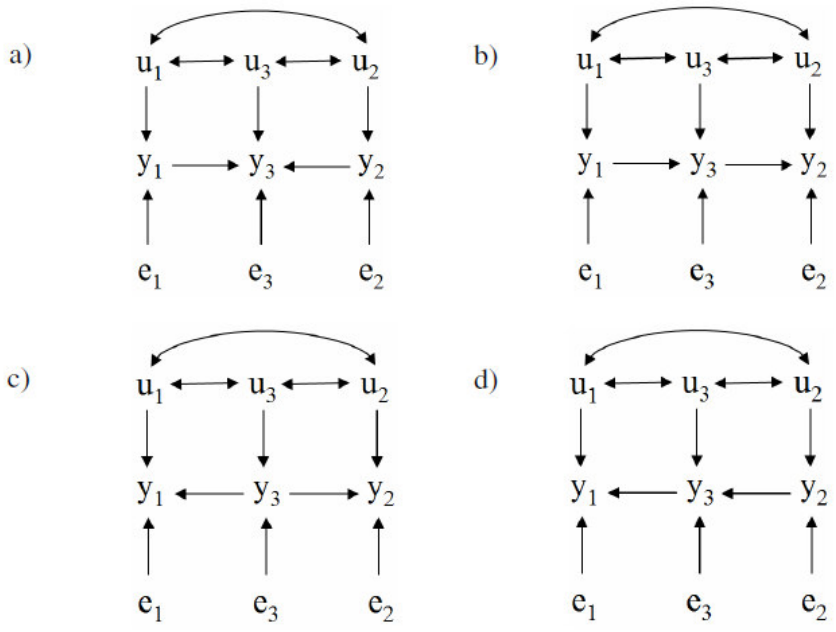


Figura 3. Estruturas causais para três variáveis observadas (y_1, y_2, y_3), com resíduos independentes (e_1, e_2, e_3) e efeitos genéticos aditivos correlacionados (u_1, u_2, u_3).

Valente et al. (2010) propuseram um método para busca de estruturas causais acíclicas em que d-separações refletem independências condicionais na distribuição dos fenótipos depois de considerar os efeitos genéticos aditivos. O modelo que considera os efeitos genéticos aditivos pode ser representado como, $y = (I_t - \Lambda)^{-1} X_i \beta + (I_t - \Lambda)^{-1} u_i + (I_t - \Lambda)^{-1} e_i$, que implica que a matriz de covariância do vetor de observações pode ser descrita, como:

$Var(y_i) = (I_t - \Lambda)^{-1} G_0 (I_t - \Lambda)'^{-1} + (I_t - \Lambda)^{-1} \Psi_0 (I_t - \Lambda)'^{-1}$. Note que o primeiro e o segundo termos do lado direito da equação correspondem, respectivamente, à matriz de covariância genética aditiva e residual obtidas de um modelo multicaracterísticas padrão que considera covariâncias genéticas e residuais entre características, mas não considera os efeitos causais entre os fenótipos (Gianola e Sorensen, 2004). A matriz de covariância de y_i pode ser descrita como, $Var(y_i \mid u_i) = (I_t - \Lambda)^{-1} \Psi_0 (I_t - \Lambda)'^{-1} = R_0$. Desta forma, estimativas de R_0 podem ser utilizadas para selecionar a estrutura causal entre fenótipos (Rosa et al., 2010). No trabalho de Valente et al. (2010) a matriz de covariância R_0 é obtida usando métodos MCMC (*Markov Chain Monte Carlo*), nos quais amostras são conhecidas a partir da distribuição *a posteriori* de R_0. As amostras são utilizadas para obter a incerteza a respeito dessa matriz, enquanto consideram todos os outros parâmetros de um modelo multicaracterísticas reduzido. O método consiste em três estádios:

1) Ajustar o modelo multicaracterísticas e obter amostras da distribuição *a posteriori* de R_0 .
2) Aplicar o algoritmo IC sobre as amostras *a posteriori* de R_0 para tomar as decisões estatísticas exigidas. Especificamente, para cada pergunta acerca da independência entre as variáveis *a* e *b* dado um conjunto de variáveis *S* e, implicitamente, os efeitos genéticos:
    a) Obtenha a distribuição a posteriori da correlação parcial residual $\rho_{a,b|S}$. Tais correlações parciais são funções de R_0 . Desta forma, suas distribuições *a posteriori* podem ser obtidas pelo cômputo da correlação correspondente para cada amostra utilizada para representar a distribuição *a posteriori* de R_0 .
    b) Obtenha o intervalo HPD 95% para a distribuição *a posteriori* de $\rho_{a,b|S}$.
    c) Se o intervalo HPD contém 0, declare $\rho_{a,b|S}$ como nulo. Caso contrário, declare *a* e *b* como condicionalmente dependentes.
3) Ajustar o SEM utilizando a estrutura causal obtida (ou uma das estruturas observacionalmente equivalentes recuperadas pelo algoritmo IC).

No mesmo trabalho, Valente et al. (2010) usaram dados simulados para validar a metodologia proposta mostrando que a estrutura causal verdadeira subjacente, pode de fato, ser recuperada. Uma primeira aplicação do método completo IC+SEM foi feita utilizando dados de codornas de corte por Valente et al. (2011), que estudaram a relação de causa-efeito entre cinco características, peso da ave ao nascimento, peso aos 35 dias de idade, idade à maturidade sexual, peso médio do ovo e taxa de postura. Foram feitas decisões estatísticas adotando diferentes intervalos HPD para as correlações parciais. A estrutura causal encontrada foi completada com a ajuda da informação temporal de expressão das características. Por exemplo, o fenótipo de peso ao nascer é expresso anteriormente ao fenótipo de peso aos 35 dias de idade e idade à maturidade sexual.

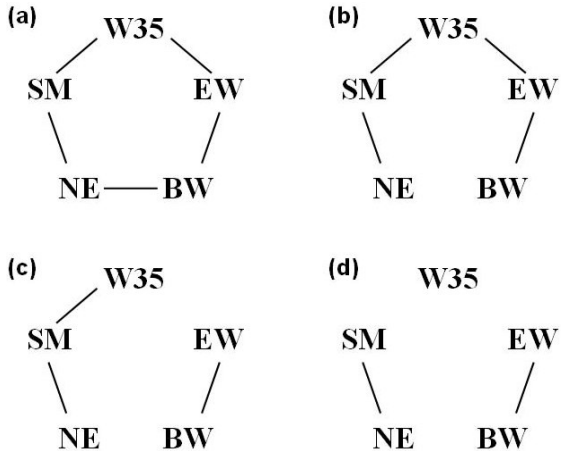


Figura 4. Estruturas recuperadas pelo algoritmo IC por Valente et al. (2010).

2. Matriz dos coeficientes de parentesco genômica e seu uso na avaliação genética animal

Com a disponibilidade de dados referentes a marcadores moleculares do tipo SNP (*single neucleotide polymorphism*) metodologias de avaliação genética começaram a ser derivadas para incluir informações de marcadores de DNA à avaliação genética animal. Talvez a forma mais natural de lidar com o novo problema seria a substituição, nas equações de modelo misto de Henderson (1984), da matriz dos coeficientes de parentesco estimados pela matriz de relacionamento genético realizado. Essa matriz refletiria a verdadeira proporção dos genes compartilhados entre os indivíduos, fator, que dentre outros, aumentaria a acurácia na predição dos valores genéticos. Reconhecendo a atual importância da utilização da matriz de parentesco genômico, ou, como simplesmente é chamada, matriz *G*, serão abordados temas relativos à sua construção, às metodologias de incorporação dessa matriz nas equações para solução dos valores genômicos, assim como à sua utilização na prática dos programas de avaliação genética animal.

2.1. Avaliação genética e matriz dos coeficientes de parentesco tradicional

Em um modelo clássico de genética quantitativa para características complexas o valor fenotípico é controlado por um número infinito de genes, cada gene contendo um efeito infinitesimal, assim como, também, por efeitos não genéticos ou ambientais. Sob esse modelo não é possível estabelecer o genótipo de um indivíduo para um loco específico e selecionar plantas ou animais com os alelos mais desejáveis. Consequentemente, a seleção é baseada na predição total dos efeitos dos genes que o indivíduo carrega, ou seja, seu valor genético. Os

modelos mistos exercem papel importante na predição de valores genéticos de plantas e animais. Sob a premissa que os valores genéticos seguem distribuição normal multivariada com covariância genética igual a *G**, melhores preditores lineares não viesados (BLUPs) podem ser usados para calcular valores genéticos a partir de dados fenotípicos (Endelman e Jannink, 2012). Esse modelo clássico é robusto e a seleção é efetiva mesmo quando o número de genes que controlam a característica é pequeno (Goddard, 2009).

Admitindo que o modelo que explica o fenótipo de determinada característica em um indivíduo seja $y = Xb + Za + e$, as equações de modelo misto, descritas matricialmente como:

$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + G^{-1} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} X'R^{-1}y \\ Z'R^{-1}y \end{bmatrix}$, podem ser utilizadas para a obtenção simultânea dos efeitos fixos e aleatórios, sendo a covariância genética estimada como, $\hat{G} = A\sigma_a^2$, em que *A* é a matriz dos coeficientes de parentesco dos animais para os quais se pretende obter valores genéticos. Incorporando a matriz *A* ao sistema de equações pode-se fazer a ligação entre fenótipos de diferentes indivíduos contidos em diferentes classes de efeitos fixos, assim, mesmo animais que não possuem dados coletados podem ter preditos seus valores genéticos. A matriz *A* também proporciona um ajuste dessa predição ao permitir que informações de parentes tenham pesos diferentes de acordo com a magnitude do relacionamento genético entre os animais.

Coeficientes de parentesco correspondem à covariância genética existente entre animais relacionados, expressados relativamente e independente da característica considerada (Bomcke et al., 2011). O coeficiente de parentesco entre dois indivíduos pode ser visto como a probabilidade de um gene amostrado aleatoriamente de um indivíduo ser idêntico por descendência a um gene homólogo também amostrado de forma aleatória do segundo indivíduo (Wright, 1922; Malécot, 1948). A probabilidade de que dois alelos são IBD (idênticos por descendência) deve ser definida com respeito a uma população base que possui variância genética igual a σ_a^2, ou seja, os dois alelos são originados de um mesmo ancestral na população base (Hayes et al., 2009; Powell et al., 2010; Endelman e Jannink, 2012). Tradicionalmente a probabilidade que dois alelos são IBD é calculada a partir de um pedigree conhecido, de tal forma que os animais no topo do pedigree (fundadores) formam naturalmente uma base populacional (Powell et al., 2010).

Apesar da seleção baseada nos melhores preditores lineares não viesados obtidos com a utilização da matriz *A* ter sido de grande valia ao melhoramento genético dos animais de produção, algumas situações fazem com que a esperança predita da proporção do genoma que é idêntico por descendência entre dois indivíduos, dado o pedigree, seja diferente da verdadeira proporção compartilhada. Isso ocorre, pois, o pedigree normalmente se apresenta incompleto, primeiro, porque existe uma edição dos dados de acordo com a data, em que os animais nascidos em épocas longínquas são cortados da avaliação, e, segundo, alguns animais que não possuem pedigree podem entrar na avaliação genética (Bomcke et al., 2011). O afastamento entre o coeficiente de parentesco predito com dados de pedigree e a proporção de alelos compartilhados realizada acontece, também, em razão da segregação mendeliana durante a formação dos gametas, o que resulta em variação da proporção do genoma que é IBD entre pares de animais que possuem mesmo coeficiente de parentesco. Por exemplo, o coeficiente de parentesco predito entre irmãos-completos é 0,5, sendo que essa predição apresenta desvio padrão de 0,04 em caso de espécies com 30 cromossomos com um Morgan de comprimento (Hayes et al., 2009). Segundo VanRaden (2008), a proporção de alelos que são idênticos por descendência, ou melhor, quando as predições para os coeficientes podem variar, é dependente do número de locos que influenciam a característica. O autor ainda atesta que o desvio-padrão da percentagem de alelos compartilhada predita para irmãos-completos é de 5%. Considerar essas pequenas diferenças pode aumentar significativamente a confiança na predição dos coeficientes de parentesco, e consequentemente aumentar também a acurácia dos valores genéticos preditos. Além disso, maior diferenciação entre irmãos é possível, e redução na seleção de indivíduos aparentados é esperada, já que os efeitos da segregação mendeliana são mais bem estimados. Como resultado, a endogamia média ao longo das gerações deve elevar mais lentamente do que com as avaliações tradicionais (Daetwyler et al., 2007).

2.2. Matriz dos coeficientes de parentesco genômicos

Com a disponibilidade de dados moleculares, informação de marcadores de DNA para vários locos espalhados ao longo do genoma podem ser usados para medir a similaridade de indivíduos e calcular a matriz de parentesco realizada com elementos que demonstram a verdadeira proporção do genoma em comum, sendo, portanto, mais precisa que as informações de pedigree. O parentesco genômico pode melhor estimar a proporção de segmentos cromossômicos compartilhados entre indivíduos uma vez que a obtenção de

genótipos de alta densidade permite a identificação de genes idênticos por estado (Forni et al., 2011).

Nejati-Javaremi et al. (1997) demonstraram que se os locos responsáveis por uma característica forem conhecidos, os mesmos poderiam ser usados para derivar a matriz de parentesco realizada e obter acurácia de predição dos valores genéticos maior comparada à utilização da matriz de parentesco tradicional. Porém, na prática a identificação de todos os QTLs referentes à variação de determinada característica é improvável. Villanueva et al. (2005) também demonstraram que a utilização da matriz de parentesco realizada para obtenção de valores genéticos poderia incrementar a acurácia, mesmo quando os locos de características quantitativas não são conhecidos e o modelo subjacente à característica é infinitesimal. Para tal, utilizaram de informações de pedigree conhecido, explorando informações de ligação, e obtiveram a matriz de parentesco para indivíduos não relacionados dentro de uma mesma população. Outra maneira em que informações de marcadores de DNA podem ser utilizadas para obtenção dos valores genéticos é a seleção genômica (Meuwissen et al., 2001). Nesse método, informações de marcadores em alta densidade são utilizadas para rastrear QTLs em desequilíbrio de ligação com eles. Os efeitos dos marcadores são então estimados e somados para predizer o valor genético de cada animal (Goddard e Hayes, 2007). Se houver grande quantidade de QTLs com efeitos normalmente distribuídos com variância constante entre eles, a seleção genômica se torna equivalente ao BLUP com incorporação da matriz de parentesco realizada (*G*) no lugar da matriz *A* (Meuwissen et al., 2001; Habier et al., 2007; VanRaden, 2008; Hayes et al., 2009).

A matriz de parentesco genômico pode ser calculada por diversos métodos utilizando dados de marcadores do tipo SNP. VanRaden (2008) apresentou três métodos utilizando a matriz M, que especifica quais marcadores de alelos cada animal herdou. A dimensão de M é o número de indivíduos (n) pelo número de locos (m). As equações podem incluir as informações de marcadores usando a matriz $n \; x \; n$, MM', ou, a matriz $m \; x \; m$, $M'M$. Se os elementos de M forem -1, 0 e 1 para os locos, homozigoto, heterozigoto e o outro homozigoto, respectivamente, a diagonal MM' conta o número de locos homozigotos para cada indivíduo, e os elementos fora da diagonal medem o número de alelos compartilhados entre os indivíduos aparentados. Já a matriz $M'M$ conta o número de indivíduos homozigotos para cada alelo na diagonal e o número de vezes que os alelos de diferentes

locos foram herdados pelo mesmo indivíduo. Se P é matriz que contém as frequências dos alelos subtraídas por 0,5 e então multiplicada por dois, como, $2(p_i - 0{,}5)$, a subtração $M - P$ é igual a Z, que ajusta os efeitos dos alelos para zero. Caso outros códigos para definir o genótipo dos diferentes locos forem utilizados, como, por exemplo, 0,1 e 2, a variância de W (genótipos) não se altera, porém a esperança é modificada. Assim, a matriz P deve ser diferente de modo que os efeitos dos alelos continuem centrados em zero (Aguilar, 2010; Gianola e de los Campos, 2012). Mais informações são registradas no Anexo 1.3. As frequências alélicas utilizadas em P devem ser provenientes da população base não selecionada ao invés daquela observada após a seleção ou endogamia. Uma população base mais antiga ou mais recente pode gerar menores ou maiores coeficientes de parentesco ou modificar o grau de endogamia (VanRaden, 2008). Como normalmente as frequências da população base em que não houve seleção são raramente disponíveis, Forni et al. (2011) propuseram outros métodos de obter as frequências alélicas na montagem da matriz P, sendo esses, atribuir o valor 0,5 para todos os marcadores, usar a média da frequência mínima dos alelos para todos os locos ou a frequência observada dos alelos em cada SNP. A subtração de P em M dá mais crédito aos alelos raros do que aos alelos comuns quando se faz o cálculo dos relacionamentos genômicos, ou seja, quando a frequência dos alelos na população base é diferente de 0,5, os alelos menos frequentes contribuem mais para a semelhança genética entre indivíduos do que aqueles alelos mais frequentes. Também, a endogamia genômica é maior se o indivíduo é homozigoto para alelos raros do que homozigoto para alelos comuns (VanRaden, 2008; Forni et al., 2011).

Para se obter a matriz de relacionamento genômico, o primeiro método usa a fórmula: $G = \frac{ZZ'}{2\Sigma p_i(1-p_i)}$. A divisão por $2\Sigma p_i(1-p_i)$ faz a matriz *G* ser análoga à matriz *A* (mais detalhes no Anexo 1.2). O coeficiente de endogamia genômico calculado para o indivíduo *j* é simplesmente $G_{jj} - 1$, e o parentesco genômico entre os indivíduos *j* e *K*, que são análogos ao coeficiente de parentesco de Wright (1922), são obtidos dividindo os elementos G_{jk} pelo desvio padrão dos elementos da diagonal G_{jj} e G_{kk}.

O segundo método para obtenção de *G* pesa os marcadores com as suas variâncias esperadas ao invés de somar as variâncias como na metodologia anterior, e então faz a divisão. *G* =

ZDZ', em que *D* é matriz diagonal com $D_{ii} = \frac{1}{m\,[2p_i(1-p_i)]}$. Esse modelo foi proposto anteriormente para utilização em estudos de genética humana por Leutenegger et al. (2003) e Amin et al. (2007).

O terceiro método demonstrado por VanRaden (2008) não requer o conhecimento das frequências alélicas e ao invés disso ajusta os genótipos por meio de uma regressão de MM' em *A*. O modelo é o seguinte: $MM' = g_0 11' + g_1 A + E$, em que g_0 é o intercepto e g_1 é a inclinação. A matriz E inclui a diferença entre a fração de DNA em comum esperada e a realizada mais a mensuração do erro, pois sequências de DNA completas não estavam disponíveis e apenas uma amostra de marcadores foram genotipados. A regressão foi ajustada considerando MM' como variável dependente e *A* como independente porque *A* é o valor esperado de *G* e não o contrário. Entretanto, por causa dessa característica é necessário que os cálculos sejam revertidos após ajuste do modelo, usando a fórmula: $G = \frac{MM' - g_0(11')}{g_1}$. As estimativas de g_0 e g_1 podem não ser óbvias, pois a variável dependente e independente são matrizes e não um vetor, mas as equações podem ser escritas com notação usando somas da seguinte maneira: $\begin{bmatrix} n^2 & \sum_j \sum_k A_{jk} \\ \sum_j \sum_k A_{jk} & \sum_j \sum_k A_{jk}^2 \end{bmatrix} \begin{bmatrix} g_0 \\ g_0 \end{bmatrix} = \begin{bmatrix} \sum_j \sum_k (MM')_{jk} \\ \sum_j \sum_k (MM')_{jk} A_{jk} \end{bmatrix}$.

Forni et al. (2011) ainda recordaram de dois outros métodos para obtenção da matriz G. O primeiro, descrito por Gianola et al. (2009), escala a matriz *ZZ'* de forma diferente levando em consideração a distribuição aleatória dos SNPs e suas frequências por meio de: $\frac{G = (M-P)(M-P)'}{\left[(p_0 - q_0)^2 + 2\left(\frac{\sum_{j=1}^{m} p_j(1-p_j)}{m}\right)\left(\frac{\alpha+\beta+2}{\alpha+\beta}\right)\right] m}$, em que $p_0 = \frac{\alpha}{\alpha+\beta}$ e $q_0 = 1 - p_0$ são a esperança das frequências dos alelos; α e β são os parâmetros da distribuição beta ajustando a frequência alélica base; m é o número de marcadores SNP.

O último método baseia-se na idéia de se obter uma matriz normalizada com a média dos elementos da diagonal igual a 1, o que pode ser conseguido com:

$G = \frac{(M-P)(M-P)'}{\{traço[(M-P)(M-P)'\}/n}$. O n da equação pode ser substituído pela soma de $(1+F)$ de todos os animais genotipados, sendo F o coeficiente de endogamia derivado do pedigree. Diferente de uma matriz de parentesco tradicional, os valores dos elementos da diagonal da matriz podem ser menor que 1. Uma média dos elementos da diagonal igual a 1 também pode ser obtida multiplicando a matriz acima por uma constante.

A matriz de relacionamento genômico é positiva semi-definida, mas pode ser singular se o número de marcadores é limitado ou dois indivíduos apresentam mesmo genótipo para todos os marcadores. A matriz também não irá apresentar inversa caso o número de marcadores SNP for menor que o número de animais com genótipo conhecido. Para evitar problemas potenciais com a inversão de *G*, essa pode ser calculada como: $G = \lambda G_{orig} + (1-\lambda)A$, em que, G_{orig} é a matriz *G* pura. Segundo VanRaden (2008) o peso λ é relacionado à variância do erro em se estimar as verdadeiras porções do genoma compartilhadas em comum entre os indivíduos. Christensen e Lund (2010) sugeriram que λ pode ser interpretado como o peso relativo do efeito poligênico necessário para explicar a variância genética aditiva total, assim, $\lambda = \frac{\sigma_a^2}{(\sigma_a^2 + \sigma_g^2)}$, em que σ_g^2 é a variância explicada pelos marcadores. Aguilar et al. (2010) relatam que a principal influência do peso λ deve ser relacionada à proporção da variância genética aditiva explicada pela informação genômica, o que por sua vez está relacionada à quantidade de marcadores e às quais animais têm genótipo observado, dentre outros fatores. Porém, os mesmos autores apontaram diferenças pequenas nos valores genéticos preditos dos animais quando o λ variou de 0,95 a 0,98. Christensen e Lund (2010) sugeriram calcular o valor de λ por meio de função de verossimilhança considerando um conjunto de diferentes valores para este parâmetro. VanRaden (2008) testou diferentes valores para λ, que variaram de 0,90 a 1. O autor observou melhoria na acurácia das predições de apenas 0,0002 por cento ao usar o peso igual a 0,95, que foi o maior ganho obtido em relação à utilização da matriz *G* pura. Ainda, segundo VanRaden (2008), os elementos de *G* possuem variância do erro de $\frac{0,125}{m}$. Quando a menor frequência dos alelos for menor do que 0,5, para que a informação proveniente dos marcadores seja refletida de forma precisa, a fórmula $\lambda = \frac{0,05}{\left(0,05^2 + \frac{0,125}{m}\right)}$

pode ser usada e *G* deve receber maior peso que *A* para $m > 50$ e quase todo o peso (>0,99) se $m > 5000$.

Forni et al. (2011) construíram diversas matrizes de parentesco, entre elas a *A* (tradicional), *G05* (matriz segundo VanRaden (2008) com frequência alélica utilizada igual a 0,5 para todo marcador), *GMF* (como em VanRaden (2008) utilizando a frequência mínima média para todos os SNPs), *GOF* (de acordo com VanRaden (2008) utilizando as frequências observadas), *GOF** (como proposto por Gianola et al., 2009) e *GN* (normalizada para obter média dos elementos da diagonal igual a 1). A comparação dos valores dos coeficientes de relacionamento genético contidos nessas matrizes pode ser observada na Tabela 1 adaptada do mesmo trabalho.

Tabela 1. Estatística dos coeficientes de parentesco estimados usando informações de pedigree e de marcadores moleculares do tipo SNP

Elementos da Diagonal				
	Média	Mínimo	Máximo	Variância
A	**1,000**	**1,000**	**1,075**	**0,00003**
G05	**1,253**	**1,178**	**1,462**	**0,00083**
GMF	**1,697**	**1,632**	**1,894**	**0,00073**
GOF	0,936	0,837	1,228	0,00176
GOF*	0,505	0,436	0,663	0,00051
GN	1.002	0,895	1,314	0,00201
Elementos de fora da diagonal				
	Média	Mínimo	Máximo	Variância
A	0,032	0,000	0,600	0,00172
G05	0,595	0,387	1,231	0,00160
GMF	1,022	0,822	1,654	0,00155
GOF	0,000	-0,198	1,000	0,00241
GOF*	0,000	-0,105	0,540	0,00070
GN	0,000	-0,212	1,070	0,00275

Segundo os autores, esperava-se que a variância dos coeficientes de parentesco fossem maiores para as matrizes genômicas, já que o parentesco genômico reflete a exata proporção de genes compartilhados enquanto o coeficiente de parentesco de Wright é uma predição. Porém, isso foi observado apenas nas matrizes *GOF* e *GN*. Ainda, pode ser visto que os elementos fora da diagonal das matrizes genômicas podem assumir valores negativos. Hayes et al. (2009) explicam que os coeficientes de parentesco são sempre relativos à uma população base, assim subtraindo os valores dos elementos contidos em *M* pela frequência média dos alelos, faz com que os relacionamentos contidos em *G* passem a ser relativos à atual população. Consequentemente, o parentesco médio se aproxima de 0 e alguns elementos podem assumir valores negativos. A normalização que ocorre para formação da matriz *GN* realmente gerou média dos elementos da diagonal igual a 1, assim como ocorre para a matriz de parentesco tradicional. A média dos coeficientes de fora da diagonal nas matrizes *GOF*, *GOF** e *GN* foi 0, similar à matriz *A*.

2.3. Metodologias de utilização da matriz de parentesco genômica

Um trabalho de grande impacto na área de melhoramento animal nos últimos anos foi escrito por Meuwissen et al. (2001) que propuseram pela primeira vez a seleção genômica e modelos que se adequam a ela. Entre os métodos descritos está o que os autores nomeiam de BLUP e que considera os efeitos de QTLs amostrados de uma distribuição normal com variância constante pelos segmentos de cromossomas. Esta variância, bem como a variância residual, é considerada conhecida. Os efeitos dos marcadores são obtidos pela resolução das equações de modelos mistos (Henderson, 1984). A demonstração do G-BLUP e RR-BLUP (*Ridge Regression* BLUP, *Random regression BLUP*) apresentada a seguir foi retirada de Gianola e De Los Campos (2012).

Considere o modelo:

$y_i = \mu + \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i$, em que y_i é o fenótipo do indivíduo *i*, μ é efeito comum a todos os indivíduos, x_{ij} são covariáveis (por exemplo, genótipos de marcadores), β_j é o efeito da covariável *j* e ε_i é o resíduo do modelo. Em notação matricial o modelo é expresso como:

$y = X\beta + \varepsilon$, em que $y = \{y_i\}$ é um vetor de fenótipos, $X = \{1, x_1, ..., x_p\}$ é uma matriz de incidência para o vetor de coeficientes de regressão $\beta = (\mu\mu\beta_1, ...\beta_p)'$ e $\varepsilon = \{\varepsilon_i\}$ é um vetor de resíduos do modelo.

A densidade conjunta de y e β é:

$\begin{bmatrix} y \\ \beta \end{bmatrix} \sim MVN\left[0, \begin{bmatrix} XX'\sigma_\beta^2 + I\sigma_\varepsilon^2 & X\sigma_\beta^2 \\ X'\sigma_\beta^2 & I\sigma_\beta^2 \end{bmatrix}\right]$. Logo, $X\beta$ também segue distribuição normal já que é uma soma de distribuições normais atribuídas aos efeitos dos marcadores. O BLUP para esses efeitos é:

$E[\beta | y, \sigma_\varepsilon^2] = X'\sigma_\beta^2[XX'\sigma_\beta^2 + I\sigma_\varepsilon^2]^{-1} y = X'[XX' + \lambda I]^{-1} y$ [Equação 1], que é a média a *posteriori* de β. Aqui, $\lambda = \frac{\sigma_\varepsilon^2}{\sigma_\beta^2}$. A expressão acima é linear com respeito aos dados e é não viesada com respeito à média à *prior*i, $E(\beta) = 0$. Para confirmar isso se tira a esperança com respeito à y e têm-se $E\{E[\beta | y, \sigma_\varepsilon^2]\} = X'[XX' + \lambda I]^{-1} E(y)$ e $E[y] = 0$, assim $E\{E[\beta \mid y, \sigma_\varepsilon^2]\} = 0$.

Agora é derivada a esperança dos valores genéticos genômicos condicionais aos dados:

$E[X\beta | y, \sigma_\varepsilon^2] = XE[\beta | y, \sigma_\varepsilon^2] = XX'[XX' + \lambda I]^{-1} y = [I + G^{-1}]^{-1} y$, em que G, a matriz dos genótipos para os marcadores nos diferentes indivíduos é XX'. Esse é o G-BLUP (VanRaden, 2008) para predição dos valores genéticos a partir de informações moleculares. A expressão é também o melhor preditor para os valores genômicos e é linear com respeito aos dados. Obtendo-se a esperança com relação aos fenótipos, $E\{[I + \lambda G^{-1}]y\} = [I + G^{-1}]^{-1} E\{y\} = 0$. Então, está demonstrado o BLUP dos valores genéticos genômicos.

De forma generalizada a expressão de BLUP para os efeitos dos marcadores mostrada acima é reescrita como, $[X'X + \lambda D]^{-1} \hat{\beta}_{RR} = X'Y$ [Equação 2], em que λ é constante adicionada à diagonal da matriz dos coeficientes e D é uma matriz diagonal com 0 na primeira entrada da diagonal (para evitar o encurtamento da estimativa do intercepto) 1 nas demais entradas da diagonal, é o sistema de equações para solução dos efeitos dos marcadores RR (*Ridge Regression*). Relembrando que ao considerar no sistema de equações acima $\lambda = \frac{\sigma_\varepsilon^2}{\sigma_\beta^2}$, obtêm-se o BLUP (RR-BLUP) para os efeitos de SNPs como proposto por Meuwissen et al. (2001).

Adicionar uma constante à diagonal da matriz dos coeficientes de parentesco genômica faz com que essa deixe de ser singular e encurta as estimativas dos coeficientes de regressão que não são o intercepto em direção a zero. Isso induz viés, mas reduz a variância dos estimadores. Em um problema com grande número de parâmetros a serem estimados e com pequeno número amostral, o quadrado médio do erro pode ser reduzido e predições mais acuradas são produzidas (Gianola e de los Campos, 2012). Apesar de λ ser constante para todos os SNPs, diferenças nos encurtamentos surgem em razão da variação na frequência alélica dos marcadores (Moser et al., 2009).

Como demonstrado, as equações 1 e 2 produzem estimativas equivalentes para os efeitos dos marcadores, a primeira necessita de inversão de uma matriz $n \times n$, já a segunda da inversão de uma matriz $p \times p$. Por levarem às mesmas estimativas, na literatura, o método G-BLUP normalmente é citado quando se deseja predizer o valor genético genômico dos animais sem a necessidade de saber os efeitos dos marcadores, em contrapartida o RR-BLUP é citado em trabalhos que desejam obter os efeitos dos SNPs.

Apesar dos métodos descritos acima serem promissores, obter genótipos de todos os indivíduos de uma população é impraticável por questões de custo e de logística, como por exemplo, descarte de animais (Legarra et al., 2009). Predições genômicas dos valores genéticos podem ser obtidas estimando os efeitos dos marcadores ou usando o modelo de equações mistas com a adição da matriz de relacionamentos genômicos. Porém o sistema de equações proposto por VanRaden (2008) prediz valores genômicos apenas para animais com genótipo conhecido ou necessita de metodologia com múltiplas etapas para obtenção de valores genéticos compostos pelos dois tipos de informação de relacionamento genético. Uma forma simples para utilizar a metodologia tradicional de modelos mistos de Henderson e aproveitar todos os dados de fenótipo, genótipo e de pedigree ao mesmo tempo é fazer a modificação da matriz A, tal que essa inclua também informações de marcadores. Esse método é chamado de *single-step genomic BLUP* - ssGBLUP (Legarra et al., 2009; Misztal et al., 2009; Aguilar et al., 2010). O resultado é uma distribuição conjunta dos valores genéticos dos animais genotipados e não genotipados com uma matriz pedigree-genômica de relacionamentos genéticos, *H*, que transmite a informação genômica também para a covariância entre todos os indivíduos sem genótipo (Legarra et al., 2009). A matriz *H* é representada como:

$H = \begin{bmatrix} A_{11} + A_{12}A_{22}^{-1}(G - A_{22})A_{22}^{-1}A_{21} & A_{12}A_{22}^{-1}G \\ GA_{22}^{-1}A_{21} & G \end{bmatrix}$, em que A_{11} é a partição da matriz de parentesco tradicional para animais não genotipados, A_{22} é relativa aos animais genotipados, A_{21} e A_{12} corresponde a parte da matriz A que indica os coeficientes de parentesco entre os animais com genótipo e sem genótipo, e *G* é a matriz de parentesco genômico. Christensen e Lund descreveram uma maneira fácil de obter a matriz H^{-1}, que será utilizada no sistema de equações:

$H^{-1} = A^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & G^{-1} - A_{22}^{-1} \end{bmatrix}$.

2.4. Resultados da utilização da matriz de relacionamento genômico

VanRaden (2008) descreveu diversas metodologias de cálculo da matriz *G* e comparou métodos para predição dos valores genéticos genômicos. Por meio de simulação de estrutura populacional e 50.000 marcadores do tipo SNP, o autor relatou acurácias, ou seja, correlação entre o valor genético predito e o valor genético simulado, maior para as metodologias que utilizaram a matriz de relacionamento genômico. Em média houve um ganho na acurácia de 30 por cento com relação à utilização da matriz de parentesco tradicional.

Utilizando os mesmos modelos do trabalho de 2008, mas com dados reais de bovinos da raça holandesa, VanRaden et al. (2009) relataram ganho na acurácia de predição do valor genético para diversas características ao se utilizar a matriz *G* no sistema de equações de modelos mistos.

Muitos métodos foram desenvolvidos com objetivo de acomodar as informações genômicas no modelo para obtenção de efeitos de marcadores, fenótipos futuros ou o valor genômico dos indivíduos. Talvez, outros métodos não citados até aqui, como, por exemplo, o Bayes B, tenham maior apelo teórico no sentido de que não possuem a premissa do modelo infinitesimal para o efeito dos marcadores, portanto tem a capacidade de reconhecer que alguns SNPs tem efeito 0, já outros apresentam efeitos maiores. Porém, na prática, o que se vê até aqui é que não há grande diferença na acurácia de predição para as diferentes metodologias genômicas. A incorporação da informação de marcadores de DNA por meio da matriz de parentesco genômica é bastante viável, pois é de fácil entendimento e fácil execução

computacional. Luan et al. (2009) relataram semelhança na acurácia de predição de valores genômicos para diversas características ao estudarem uma população de gado *Norwegian Red* utilizando chips de 25K.

Wolc et al. (2011) para quantificar o potencial benefício da utilização da seleção genômica, compararam a acurácia dos valores genéticos, estimados para diversas características em galinhas comerciais de postura, obtida por diferentes métodos usando matriz de parentesco derivada de pedigree ou a partir de informações de marcadores SNP. Nesse trabalho foi demonstrado que de modo geral as metodologias que são baseadas em informações genômicas apresentaram maior habilidade de predição do que a baseada em informações de pedigree, principalmente ao se utilizarem de dados de animais ainda jovens. Os autores ainda salientaram que tal diferença se deveu ao fato de haver grandes desvios com relação aos coeficientes de parentesco obtidos pelos diferentes métodos (Figura 5). Nesse estudo, o G-BLUP superou ligeiramente o método Bayes-C-Pi quanto à acurácia de predição dos valores genéticos.

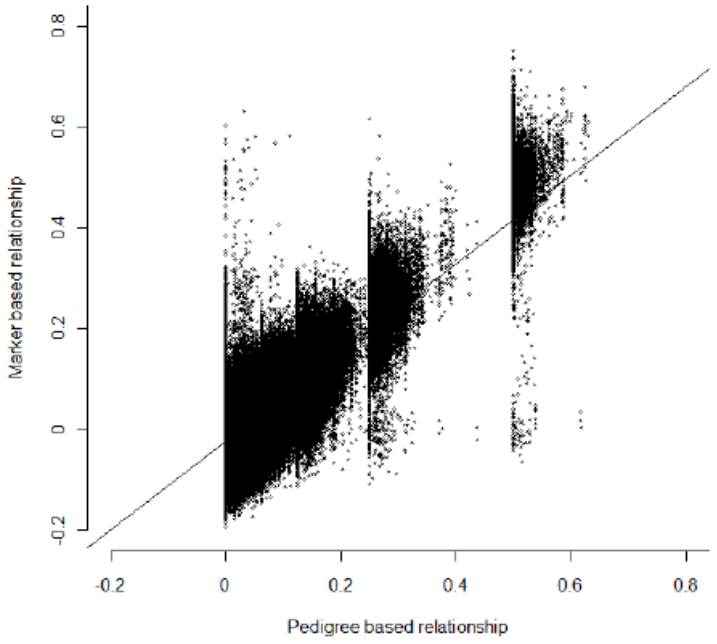


Figura 5. Relação dos coeficientes de parentesco tradicionais e genômicos (Wolc et al., 2011).

Hayes et al. (2009) fizeram revisão de diversos experimentos realizados nos Estados Unidos da América, Nova Zelândia, Austrália e Holanda, e relataram as acurácias dos valores genéticos genômicos obtidas. Esses experimentos utilizaram populações referência de 650 a 4.500 touros de raças para produção de leite, que foram genotipados para aproximadamente

50.000 marcadores do tipo SNP. As acurácias observadas dependeram de diversos fatores, entre eles, a herdabilidade da característica, o número de touros na população referência e o método estatístico utilizado para estimar os efeitos dos marcadores. Um achado comum, nos Estados Unidos, Nova Zelândia e Austrália, foi que o método GBLUP garantiu acurácias quase tão boas quanto às dos métodos mais complexos. Os autores ainda concluíram que o método GBLUP é atrativo, pois, a única informação *a priori* exigida é a variância genética aditiva da característica.

3. Avaliação genômica pelo método S*ingle-Step*

O desenvolvimento da genotipagem de alta densidade para marcadores do tipo SNP (*single nucleotide polymorphism*) favoreceu a utilização de informações moleculares para melhoria genética dos animais de produção. Meuwissen et al. (2001) apresentaram métodos para predição dos valores genéticos genômicos de animais, que acontecem em duas etapas:
a) os efeitos dos marcadores são estimados usando fenótipos e genótipos de indivíduos que fazem parte da população de treinamento do modelo; b) os valores genéticos genômicos são preditos para qualquer animal que apresenta genótipo conhecido e que estão na população de validação. Essa metodologia de predição e seleção com base em informações genômicas de alta densidade é chamada seleção genômica.

A seleção genômica pode ser realizada estimando efeitos individuais dos marcadores ou utilizando a metodologia BLUP com a adição da matriz de relacionamento genético genômico, G (VanRaden et al., 2008). Normalmente, apenas fração dos animais da população é genotipada. Consequentemente, animais com genótipo desconhecido não possuem valores genéticos genômicos preditos, e suas informações fenotípicas e de pedigree não são aproveitadas para enriquecimento das estimativas dos efeitos dos SNPs. Já existe metodologia que une as informações de parentesco derivadas do pedigree e dos marcadores. A seleção genômica por S*ingle-Step*, que usa uma matriz de parentesco combinada permite que em uma única avaliação sejam consideradas todas informações de fenótipos, genótipos e pedigree disponíveis sem limitações no modelo de análise (Legarra et al., 2009; Misztal et al., 2009; Aguilar et al., 2010).

A seguir será apresentado o método *single-step genomic BLUP* (ssGBLUP) e suas especificidades. Adicionalmente, serão relatados possíveis problemas e resultados de avaliações genéticas que utilizaram a metodologia.

3.1. Método Single-Step

Tradicionalmente, informações de pedigree têm sido utilizadas para estimar a herdabilidade e os efeitos genéticos para características complexas. Agora, já é viável gerar dados individuais genotípicos para grande número de *single nucleotide polymorphisms* (SNP) distribuídos ao longo de todo o genoma. A disponibilidade de marcadores de alta densidade levou à introdução recente de métodos para a seleção genômica. Esses modelos são normalmente baseados na estimativa simultânea de efeitos de marcadores, e a diferença entre eles é, em sua maioria, atribuída à distribuição *a priori* para os efeitos dos SNPs (Mizstal et al., 2009). Procedimentos eficientes computacionalmente já existem mesmo para grandes bancos de dados (Legarra e Misztal., 2008). A predição dos valores genéticos genômicos pode ser feita com a simples substituição, no modelo de equações de Henderson (1984), da matriz dos coeficientes de parentesco de Wright (A) pela matriz de relacionamento genômico (G) baseada em informações de marcadores. Vários autores citam ganhos na acurácia de predição dos valores genéticos ao adotar essa metodologia (VanRaden, 2008; VanRaden et al., 2009; Hayes et al., 2009). Esse modelo assume distribuição normal para os efeitos de marcadores com variância em comum entre eles, e, apesar de a premissa ser argumentável, na prática, creditar uma distribuição *a priori* mais complexa resulta em pouco ganho (VanRaden et al., 2009; Hayes et al., 2009; Luan et al., 2009; Wolc et al., 2011). Ainda, a matriz de relacionamento genômico é fácil de interpretar e manipular. Apesar do método G-BLUP proposto até então ser muito promissor para o melhoramento genético animal, assume-se que todos os animais têm genótipo conhecido. Porém, a genotipagem de uma população inteira não é viável, pois apresenta alto custo e difícil logística. Por exemplo, em suínos, é provável que apenas os machos reprodutores ou candidatos à seleção sejam genotipados (Christensen e Lund, 2010).

Como nem todos os animais podem ter genótipos conhecidos, um procedimento de dois ou três passos, *multi-step ou multiple-stage*, comumente é realizado. a) Primeiro, uma avaliação genética regular é feita, normalmente utilizando o modelo animal; b) Fenótipos corrigidos são utilizados no segundo passo, em que a seleção assistida por marcadores é efetivamente

aplicada e os efeitos dos marcadores são obtidos. Esses fenótipos são *DYD* (*daughter yield deviations*) ou *YD* (*yield deviation*); c) São calculados os valores genéticos genômicos diretos para os candidatos a seleção a partir dos efeitos dos marcadores estimados e dos genótipos desses indivíduos; d) Por último, combinam-se as informações dos valores genéticos tradicionais e genômicos usualmente por meio da teoria do índice de seleção. Os passos *b* e *c* podem ser combinados em que as estimativas dos efeitos de marcadores são obtidas pelo ajuste do modelo de equações mistas com utilização da matriz *G* de parentesco genômico (Meuwissen et al., 2011).

As vantagens do método *multiple-stage* incluem não haver necessidade de mudanças às metodologias já conhecidas de avaliação genética e conter passos simples para obtenção dos valores genéticos genômicos de animais jovens com genótipo conhecido. As desvantagens incluem requerer parâmetros nos passos *b* e *d*, por exemplo, variâncias *a priori* e coeficientes para o índice de seleção, perda de informação que reflete na diminuição de acurácia e viés atribuído à seleção. Além do mais, enquanto a utilização do modelo multicaracterísticas no passo *a* é possível para todos os animais, a avaliação no passo *b* utilizando o mesmo modelo não é óbvia e deve ser unicaracterística incorporando apenas os animais genotipados. A utilização de parâmetros incorretos para os passos *b* e *d* podem causar mudanças na predição de valores genéticos até mesmo de animais com alta acurácia (Misztal et al., 2009). Quanto à perda de informação, vários problemas existem na utilização de dados de *DYD*. Esses problemas são pesos (causados pelo diferente número de informações por animal no arquivo original), viés (causado por seleção, por exemplo), acurácia (para animais em rebanhos pequenos) e colinearidade (a estimativa de *YD* para duas vacas em um mesmo rebanho, por exemplo). Esses problemas podem ofuscar os benefícios da seleção assistida por marcadores (Legarra et al., 2009). Os atuais resultados para avaliações genômicas com *multiple-step* não são conclusivos, enquanto a acurácia para obtenção dos valores genéticos é maior comparada à utilização de apenas informações de pedigree, elas também se mostram inflacionadas. A utilização de animais jovens com avaliação genômica torna ainda mais importante obter métodos que alcancem alta acurácia e mínimo viés. A inflação nas avaliações genômicas faz com que jovens animais avaliados tenham uma vantagem sobre os animais velhos provados por meio de teste de progênie. Como visto anteriormente, os problemas com a avaliação em múltiplos estádios podem ser atribuídas à utilização de parâmetros incorretos e fortes premissas do método. Outra explicação para os resultados observados, até então, é a

genotipagem seletiva dos animais. Um problema ainda mais sério é quando as pseudo-observações são pobremente definidas e apresentam baixa qualidade (Aguilar et al., 2010).

Uma possibilidade para lidar com os problemas apontados e simplificar a atual estratégia seria fazer uma avaliação conjunta utilizando todos os dados fenotípicos, genotípicos e de pedigree. Uma medida poderia ser imputar marcadores para animais sem genótipos identificados via informações de marcadores e pedigree dos outros animais, assim a análise poderia ser realizada após imputação. No entanto essa alternativa seria difícil para um banco de dados moderado e com grande número de marcadores desconhecidos.

Misztal et al. (2009), Legarra et al. (2009) e Christensen e Lund (2010) propuseram um método que incorpora informação genômica dentro do passo *a* da avaliação *multiple-stage*, resultando em um procedimento de etapa única (*single-step*). Isso pode ser alcançado pela modificação da matriz de parentesco A de modo que informação a respeito dos marcadores é acrescentada. Nos trabalhos citados, é formada a matriz *H*, que combina a matriz *G* de relacionamento dos animais genotipados, com a matriz dos coeficientes de parentesco de Wright, A, construída a partir de informações dos animais com genótipo e sem genótipo conhecido (Meuwissen et al., 2011). Para o método G-BLUP, demonstra-se que os valores genéticos genômicos podem ser obtidos sem o cálculo dos efeitos dos marcadores, apenas incluindo a matriz G^{-1} nas equações do BLUP (Nejati-Javaremi et al., 1997; Meuwissen et al., 2001; Habier et al., 2007; VanRaden et al., 2009). Similarmente, a matriz H^{-1} pode ser utilizada para predição direta dos valores genéticos genômicos dos indivíduos. A metodologia *single-step* provê avaliação unificada, elimina várias premissas e oferece a oportunidade de cálculo mais acurado dos efeitos genômicos do que o procedimento *multiple-step* (Aguilar et al., 2010).

Assuma um sistema de equações de modelo misto regular como utilizado em uma avaliação genética tradicional, por simplicidade considere apenas um efeito aleatório:
$y = Xb + Zu + e$, em que y, é um vetor de dados, b é um vetor de efeitos fixos e u é um vetor de efeitos dos animais. Sob o modelo poligênico infinitesimal de herança $\text{var}(u) = A\sigma_a^2$, em que A é a matriz dos numeradores de parentesco baseada nas informações de pedigree. Além disso, $\text{var}(e) = I\sigma_e^2$ e X e Z são matrizes de incidência.

Misztal et al. (2009) sugeriram que a matriz dos coeficientes de parentesco (*A*) pode ser modificada para uma matriz (*H*) que acomode tanto as informações de pedigree quanto as diferenças entre o parentesco calculado com base no pedigree e com base nas informações dos marcadores (A_Δ), de tal forma que,

$H = A + A_\Delta$:

$H = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & G \end{bmatrix} = A + \begin{bmatrix} 0 & 0 \\ 0 & G - A_{22} \end{bmatrix}$, em que os índices subscritos 1 e 2 representam os animais não genotipados e genotipados, respectivamente. *G* é a matriz de parentesco genômico.

A matriz descrita é simples, mas não é construída corretamente (Legarra et al., 2009). Assumindo, por exemplo, que nenhum animal em *G* tem fenótipo, então, de acordo com *H,* o valor genético predito para os animais com genótipo conhecido (u_2) seria $u_2 \mid u_1 = A_{21}A_{11}^{-1}u_1$, em que u_1 é o valor genético predito para os animais não genotipados e G, portanto, não tem influencia sobre as predições (Aguilar et al., 2010). O uso da matriz G deve potencialmente modificar a covariância dos ancestrais e descendentes dos animais genotipados. Por exemplo, se dois irmãos completos apresentam relacionamento genômico de 0,6, utilizando apenas a matriz *A*, assume-se que o parentesco médio entre os seus filhos é de 0,25, quando na verdade seria 0,3 (Legarra et al., 2009).

Legarra et al. (2009) sugeriram a derivação da distribuição conjunta de u_1 e u_2 como $p(u_1,u_2) = p(u_1 \mid u_2)p(u_2)$. A distribuição $p(u_1 \mid u_2)$ é condicional ao pedigree por meio da teoria do índice de seleção e propriedades da distribuição normal e $p(u_2)$ é baseada apenas nas informações genômicas. A covariância da distribuição conjunta de u_1 e u_2 é chamada de *H*, que representa a matriz da covariância entre os valores genéticos incluindo a informação genômica, sendo, portanto:

$$H = \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} = \begin{bmatrix} A_{11} + A_{12}A_{22}^{-1}(G - A_{22})A_{22}^{-1}A_{21} & A_{12}A_{22}^{-1}G \\ GA_{22}^{-1}A_{21} & G \end{bmatrix}.$$

Considerando $A^{-1} = \begin{bmatrix} A^{11} & A^{12} \\ A^{21} & A^{22} \end{bmatrix}$, duas alternativas computacionalmente eficientes são:

$$H = \begin{bmatrix} (A^{11})^{-1} + (A^{11})^{-1} A^{12} G A^{21} (A^{11})^{-1} & -(A^{11})^{-1} A^{12} G \\ -G A^{21} (A^{11})^{-1} & G \end{bmatrix} \text{ e}$$

$$H = A + \begin{bmatrix} A_{12} A_{22}^{-1} & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} I \\ I \end{bmatrix} [G - A_{22}] \begin{bmatrix} I & I \end{bmatrix} \begin{bmatrix} A_{22}^{-1} A_{21} & 0 \\ 0 & I \end{bmatrix}.$$

A inversa da matriz H apresentada por Aguilar et al. (2010) e Christensen e Lund (2010) é:

$H^{-1} = A^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & G^{-1} - A_{22}^{-1} \end{bmatrix}$. Em que A_{22}^{-1} é a inversa da matriz de parentesco tradicional apenas para os indivíduos com genótipo conhecido. Entretanto, essa nova fórmula apresenta um problema, *G* é normalmente singular e não pode ser invertida sem passos adicionais.

3.2 A matriz G escalada

Segundo Chen et al. (2011), uma das preocupações que surgem ao aplicar a seleção genômica utilizando a matriz de parentesco realizada é o viés. Esse viés tem sido relatado para avaliação genômica por diversos autores (VanRaden, 2009; Aguilar et al., 2010; Forni et al., 2011). Com viés, a comparação entre animais genotipados e sem genótipo conhecido seria difícil. Esse viés pode ser causado parcialmente pela construção não adequada da matriz de relacionamento genômico. Tal matriz pode ser construída assumindo diferentes frequências alélicas, escalada ou ignorando alelos com frequência mínima. Mais detalhes sobre a formação da matriz *G* pode ser visto no item 2.2 dessa mesma revisão.

Como visto em uma das seções anteriores a fórmula de *H* inclui a expressão *G* – *A*, que é a diferença entre o coeficiente de parentesco genômico e de pedigree. Caso *G* esteja inflada ou de outro modo incompatível com *A*, a relação entre a informação proveniente dos marcadores e do pedigree vai estar errada. Diferentes matrizes de relacionamento genômico podem levar a diferentes estimativas de valores genéticos, fato que pode ser atribuído à incorreta escala de G.

Estudando gado holandês Aguilar et al. (2010) observaram que a utilização de frequência igual para os alelos na construção da matriz *P* resultou em maior acurácia e menor viés na predição dos valores genéticos genômicos. Já, Forni et al. (2011) relataram melhores resultados com utilização da frequência alélica observada atual da população. No estudo de

Aguilar et al. (2010), a acurácia do método *single-step* foi dependente da escolha de *G* e do peso atribuído à diferença entre *G* e *A*. Com a escolha apropriada a metodologia foi superior à da metodologia de estádios múltiplos. Os autores ainda citam que uma razão para a escolha de *G* ser fundamental é que as matrizes de relacionamento genético genômico e tradicional devem ser compatíveis em escala e estrutura. O peso da informação proveniente do pedigree com relação à informação genômica depende de λ e ainda mais das diagonais de G^{-1} e A_{22}^{-1}.

3.3. Críticas e novas metodologias

A matriz H^{-1} pode ser usada para obtenção direta dos valores genéticos genômicos. Nos trabalhos de Mizstal et al. (2009), Legarra et al. (2009), Aguilar et al. (2010) e Christensen e Lund (2010) é sugerida análise unificada com inclusão das informações de pedigree e dos marcadores. Se adotarmos novamente os índices subscritos 1 e 2 para, respectivamente, os animais com genótipo conhecido e não genotipados, então, A_{22}, denota o bloco de A relativo aos animais genotipados e A_{12} relaciona os animais com genótipo e sem genótipo conhecido.

Segundo Meuwissen et al. (2011). Da equação,

$$H = \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} = \begin{bmatrix} A_{11} + A_{12}A_{22}^{-1}(G - A_{22})A_{22}^{-1}A_{21} & A_{12}A_{22}^{-1}G \\ GA_{22}^{-1}A_{21} & G \end{bmatrix},$$ pode ser visto que os efeitos do desvio de *G* em relação à A_{22} nos outros blocos da matriz H é calculado pela regressão dos animais não genotipados nos animais genotipados, por exemplo, $A_{12}A_{22}^{-1}$. Isso é totalmente baseado nas informações de pedigree, ainda que os marcadores possam contribuir com informação para essas regressões. Por exemplo, dois filhos de pais meio-irmãos paternos são genotipados e contém por chance muitos genes idênticos por descendência, então esses genes foram provavelmente herdados do avô em comum, consequentemente a regressão deles no avô deveria ser aumentada. Outro problema, de acordo com Meuwissen et al. (2011), com a metodologia de *single-step* é que o termo $G^{-1} - A_{22}^{-1}$ resulta em viés que pode ser diminuído introduzindo fatores de escala, como, $G^{-1} = \lambda G_{orig}^{-1} + (1-\lambda)A_{22^{-1}}$. O objetivo do trabalho citado foi solucionar os dois problemas apontados, e, para tal, cinco métodos de utilização conjunta de informações de marcadores e de parentesco foram comparados com base na acurácia da obtenção dos valores genéticos genômicos.

O primeiro, já descrito, foi intitulado *MLAC*, pois foi sugerido por Mizstal et al. (2009), Legarra et al. (2009), Aguilar et al. (2010) e Christensen e Lund (2010). O potencial problema do método MLAC é que A_{22} e *G* devem estar na mesma escala, senão haverá diferença entre *G* e A_{22} mesmo se os coeficientes de parentesco baseados nos marcadores e pedigree forem os mesmos. Ambas matrizes são expressas relativas a uma população base (por exemplo, uma população original em que todos os animais são considerados não aparentados) mesmo se uma base real para *G* possa ser definida.

A idéia geral do segundo método é recalcular a matriz *G* (*G**) usando o coeficiente de endogamia da população calculado a partir de A_{22} e os coeficientes de parentesco e de endogamia dos animais obtidos de *G*. Isso vai fazer com que a endogamia geral da população de *G** se torna a mesma de A_{22} enquanto as estimativas de parentesco e endogamia dos indivíduos sejam relativas ao nível de *G*. A endogamia total do indivíduo *i* é:
$G_{ii} = A_{st} + (1 - A_{st})F_{is} + 1$, em que A_{st} é a média dos elementos da diagonal de A_{22} menos 1, F_{is} é o coeficiente de endogamia do animal relativo à endogamia média da população base (F_{st}) e G_{ii} é o elemento da diagonal de *G* escalado com a endogamia base mudada para a endogamia de A_{22}. Similarmente, os elementos de fora da diagonal também são escalados usando os mesmos valores de A_{st} e F_{st}, $G^*_{ij} = 2[A_{st} + (1 - A_{st})\varphi_{jis}]$, em que φ_{jis} é o parentesco entre os indivíduos *j* e *i* relativo à endogamia base F_{st}, $\varphi_{jis} = (G_{ij}/2 - F_{st})/(1 - F_{st})$. Esse método com correção para a base populacional foi chamado $MLAC_b$ (Meuwissen et al., 2011). Um problema que foi citado e atinge os métodos *MLAC* e $MLAC_b$ é que a propagação da mudança de A_{22} para *G* feita a partir do bloco 22 para os outros blocos da matriz *H* é feita por coeficientes de regressão baseados apenas em informações de pedigree.

O terceiro método, *FG*, foi proposto por Fernando e Grossman (1989), e combina informações de pedigree e de marcadores aproveitando a análise de ligação (*linkage analysis*). A aplicação do método *FG* resulta em uma matriz de relacionamento entre os alelos paternos e maternos para todos os animais e, portanto, possui quatro elementos para cada animal *i*. Esses quatro elementos são somados e divididos por 2 para se obter a matriz de parentesco na escala correta G_{FGK} para todos os *loci K*. A média dos elementos para cada loco foi obtida, que por sua vez resultou na matriz geral G_{FG}.

Outro método apresentado por Meuwissen et al. (2011) foi o *LDLA* (*Linkage Desiquilibrium Linkage Analysis*), que combina as duas últimas metodologias em uma só, nos seguintes passos: 1) Calcular a matriz G_{FG} de Fernando e Grossman. 2) Calcular a matriz *G* baseada nos marcadores tal como em MLAC, mas obter *G** ajustando *G* à mesma base populacional de G_{FG}. Notar que esse ajuste é o mesmo do método $MLAC_b$, porém com a importante diferença de utilizar os elementos da diagonal de G_{FG22} ao invés dos coeficientes de A_{22} para obter a nova endogamia base. 3) Utilizar o método $MLAC_b$ para combinar a matriz dos marcadores G* com a matriz G_{FG} afim de construir H_{LDLA} e $H_{LDLA^{-1}}$. No método *LDLA*, a matriz G_{FG} substitui A na formação de H e os fatores de propagação dos efeitos de substituição de G* por G_{FG} são, $G_{FG12}G_{FG22}^{-1}$. Essas regressões contabilizam a informação dos marcadores uma vez que ela está contida em G_{FG}.

Os coeficientes de parentescos obtidos com a utilização de *G* ou *G** contêm erros de estimativas porque um número finito de SNPs é usado. No método $LDLA_{b,}$ melhor estimativa de *G** é obtida regredindo *G** de volta à matriz *A*, de tal forma que, $\hat{G} = A + b(G^* - A) = bG^* + (1+b)A$. Então, $LDLA_b$ é o mesmo de *LDLA*, mas $\hat{G}$ é usado ao invés de *G*, e $b = \frac{Cov(G_t^*, G^*)}{Var(G^*)}$, em que *Cov()* e *Var()* denotam a covariância e variância entre os coeficientes de parentesco fora da diagonal e G_t^* é o verdadeiro parentesco.

O método $LDLA_b$ atingiu melhores acurácias e com apenas um pequeno viés quando a herdabilidade da característica estudada foi 0,1. Sob a situação em que os parentes foram genotipados e a progênie continha apenas informações de fenótipo e pedigree, todos os modelos apresentaram resultados similares, o que era previsto em razão do coeficiente de parentesco entre os pais e os filhos não ser afetado pela informação de marcadores dos pais. Na situação em que a progênie foi genotipada, os genótipos dos filhos devem ser usados para estimar os relacionamentos genéticos entre os animais acima no pedigree. Nessa situação, ficou claro que o método $MLAC_b$ é mais viesado e sua acurácia é reduzida de modo que seja menor que a acurácia dos métodos *LDLA* e $LDLA_b$. Para essa situação o método $LDLA_b$ claramente apresenta o melhor resultado. Os autores apesar de indicar a utilização desse método, já que ele faz melhor uso das informações dos marcadores, ressaltam que ele é

computacionalmente mais pesado, pois é necessário o cálculo das probabilidades de segregação dos alelos, a inversão da matriz G_{FG} e a obtenção do coeficiente de regressão *b*.

Gao et al. (2012), ao estudarem uma população de touros da raça holandesa, compararam quatro métodos de seleção genômica: o G-BLUP simples sem ajuste da matriz *G* para o efeito poligênico; o G-BLUP como proposto por VanRaden (2008) incluindo uma porção da variância explicada por *A* na formação da matriz *G*; o método de *single-step* como descrito por Legarra et al. (2009) e Aguilar et al. (2010); e o método *single-step*, em que a matriz *G* é escalada de forma que a média dos elementos da diagonal e fora da diagonal da matriz de parentesco genômico seja igual à média dos elementos correspondentes da matriz de parentesco tradicional para os animais que contém genótipo conhecido. Os autores observaram que os métodos de *single-step* podem aumentar a acurácia e reduzir o viés das predições genômicas. O método de *single-step* com o ajuste adicional obteve resultados levemente superiores à metodologia original relativos à acurácia e ao viés.

O método *single-step* permite que diversos modelos aplicados na avaliação genética tradicional possam ser adaptados à metodologia com inclusão de informações moleculares. Aguilar et al. (2011) apresentaram a avaliação multicaracterística no contexto *single-step*. Para isso utilizaram a taxa de concepção em três parições como características diferentes correlacionadas. A pouca disponibilidade de dados e a baixa herdabilidade fizeram com que os coeficientes de determinação e de regressão observados fossem também baixos. A adição de informações genômicas ao modelo aproximadamente dobrou o coeficiente de regressão e reduziu o viés das estimativas de valor genético. A utilização do modelo multicaracterísticas em relação ao unicaracterística triplicou a acurácia das estimativas dos valores genéticos genômicos para a taxa de concepção na primeira parição. O estudo mostrou a importância de se utilizarem todas as informações disponíveis no aumento da capacidade de predição de valores genéticos, principalmente, quando se tratam de características de baixa herdabilidade.

REFERÊNCIAS BIBLIOGRÁFICAS

AGUILAR, I.; MISZTAL, I.; JOHNSON D. L. et al. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *J. Dairy Sci*. 93:743–752, 2010.

AGUILAR, I.; MISZTAL, I.; TSURUTA, S.; et al. Multiple trait genomic evaluation of conception rate in Holsteins. *J. Dairy Sci*. 94: 2621–2624, 2011.

AMIM, N.; van DUIJN, C.M.; AULCHENKO, YS. A genomic background based method for associantion analysis in related individuals. *Plos One*, 12, 2012.

CHEN, C.Y.; MISZTAL, I.; AGUILAR, I.; et al. Effect of different genomic relationship matrices on accuracy and scale. *J. Anim. Sci*., 89: 2673-2679, 2011.

DAETWYLER, H. D.; VILLANUEVA, B.; BIJMA, P.; WOOLLIAMS, J. A. Inbreeding in genome-wide selection. *J. Anim. Breed. Genet*., 124: 369-376, 2007.

DE LOS CAMPOS, G.; GIANOLA, D.; BOETTCHER, P.; MORONI, P. A structural equation model for describing relationships between somatic cell score and milk yield in dairy goats. *J. Anim. Sci.* 84:2934-2941, 2006.

DE LOS CAMPOS, G.; GIANOLA, D.; HERINGSTAD, B. A structural equation model for describing relationships between somatic cell score and milk yield in first-lactation dairy cows. *J. Dairy Sci*. 89:4445-4455, 2006.

ENDELMAN, J. B.; JANNINK, J. Shrinkage estimation of the realized relationship matrix. *G3*, vol. 2, 11, 1405-1413, 2012.

FERNANDO, R. L.; GROSSMAN, M. Marker-assisted selection using best linear unbiased Prediction. *Genet. Sel. Evol*., 21, 467–477, 1989.

FORNI, S.; AGUILAR, I.; MISZTAL, I. Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. *Genet. Sel. Evol.*, 43, 1, 2011.

GAO, H.; CHRISTENSEN, O.F.; MADSEN, P.; et al. Comparison on genomic predictions using three GBLUP methods and two single-step blending methods in the Nordic Holstein population. *Genetics Selection Evolution*, 44: 8, 2012.

GIANOLA D.; DE LOS CAMPOS G. A. *Course: Statistical methods for genome-enabled selection*. May 6-10, 2012.

GIANOLA, D.; DE LOS CAMPOS G. A.; HILL, W. G. et al. Additive genetic variability and the Bayesian alphabet. *Genetics,* 183:347–363, 2009.

GIANOLA, D.; SORENSEN, D. Quantitative genetic models for describing simultaneous and recursive relationships between phenotypes. *Genetics,* 167:1407-1424, 2004.

GILLIES, D. Causality, Propensity, and bayesin networks. *Synthese*, 132 (1-2): 63-88, 2002.

GODDARD M. Genomic selection: prediction of accuracy and maximization of long term response. *Genetica*, 136:245-257, 2009.

GODDARD, M. E.; HAYES, B. J. *J. Anim. Breed. Genet*., 124, 323–330, 2007.

GODDARD, M. E.; HAYES, B. J.; MEUWISSEN, T. H. E. (2011) Using the genomic relationship matrix to predict the accuracy of genomic selection. *J. Anim. Breed. Genet.*, 128, 409–421.

HAAVELMO, T.: The statistical implications of a system of simultaneous equations. *Econometrica*, 11:1-12, 1943.

HABIER, D.; FERNANDO, R. L.; DEKKERS, J. C. M. The impact of genetic relationship information on genome-assisted breeding values. *Genetics*, 177: 2389-2397, 2007.

HAYES, B. J.; BOWMAN, P. J.; CHAMBERLAIN, A. J.; GODDARD, M. E. Invited review: Genomic selection in dairy cattle: Progress and challenges. *J. Dairy Sci*. 92:433–443, 2009.

HAYES, B. J.; VISSCHER, P. M.; GODDARD, M. E. Increased accuracy of artificial selection by using the realized relationship matrix**.** *Genet Res*, 91:47-60, 2009.

HENDERSON C. R. *Applications of linear models in animal breeding***.** University of Guelph, Guelph, 1984.

HENDERSON, C. R. A simple method for computing the inverse of a numerator relationship matrix use in prediction of breeding values. *Biometrics*, 32: 69–83, 1976.

LEGARRA, A.; AGUILAR, I.; MISZTAL, I. A relationship matrix including full pedigree and genomic information. *J. Dairy Sci*., 92: 4656–4663, 2009.

LEGARRA, A.; MIZSTAL, I. Technical note: Computing strategies in genome-wide selection. *J. Dairy Sci.*, 91: 360-366, 2008

LEUTENEGGER, A. L.; PRUM, B.; GENIN, E. et al. Estimation of the inbreeding coefficient through use of genomic data. *Am. J. Hum. Genet*., 73: 516–523, 2003.

LUAN T.; WOOLLIAMS, J. A.; LIEN, S. et al. The accuracy of genomic selection in Norwegian Red Cattle assessed by cross-validation. *Genetics*, 183: 1119-1126, 2009.

MALECOT, G. *Les mathematiques de l'heredite*. Masson et Cie, Paris, France, 1948.

MEUWISSEN, T. H.; HAYES, B. J.; GODDARD. M. E. Prediction of total genetic value using genome-wide dense marker maps. *Genetics,* 157:1819–1829, 2001.

MEUWISSEN, T.H.E.; LUAN, T.; WOOLLIAMS, J.A. The unified approach to the use of genomic and pedigree information in genomic evaluations revisited. *J. Anim. Breed. Genet.*, 128: 429-439, 2011.

MOSER, G.; TIER, B. ; CRUMP, R. E. et al. A comparision of five methods to predict breeding values of dairy bulls from genome-wide SNP markers. *Genetics Selection Evolution*, 41:56, 2009.

NEJATI-JAVAREMI, A.; SMITH, C.; GIBSON, J. P. Effect of total allelic relationship on accuracy of evaluation and response to selection. *J. Anim. Sci*., 75:1738–1745, 1997.

OPGEN-RHEIN, R.; STRIMMER, K. From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. *BMC System Biology*, 1: 37, 2007.

PEARL, J.: *Causality: Models, Reasoning and Inference*. 2 edition. Cambridge, UK: Cambridge University Press; 2009.

POWELL, J. E.; VISSCHER, P. M.; GODDARD, M. E. Reconciling the analysis of IBD and IBS in complex trait studies. *Nat. Genet.*, 11: 800–805, 2010.

ROSA, G.J.M.; VALENTE, B.D. Breeding and Genetics Symposium: Inferring causal effects from observational data in livestock. *Journal of Animal Science*, 91: 553-564, 2012.

ROSA, G.J.M.; VALENTE, B.D.; DE LOS CAMPOS, G.; et al. Inferring causal phenotype Networks using structural equation models. *Genetics Selection Evolution*, 43:6, 2011.

SPIRTES, P.; GLYMOUR, C.; SCHEINES, R. *Causation, Prediction and Search*. 2 edition. Cambridge, MA: MIT Press; 2000.

VALENTE, B.D.; ROSA, G.J.M.; DE LOS CAMPOS, G.; ET AL. Searching for recursive causal structures in multivariate genetic mixed models. *Genetics*, 185: 633-644, 2010.

VALENTE, B.D.; ROSA, G.J.M.; SILVA, M.A.; et al. Searching for phenotypic causal networks involving complex traits: na application to European quail. *Genetics Selection Evolution*, 43: 37, 2011.

VANRADEN, P. M. Efficient methods to compute genomic predictions. *J. Dairy Sci.*, 91: 4414–4423, 2008.

VANRADEN, P. M.; VAN TASSELL, C. P.; WIGGANS, G. R. Invited review: Reliability of genomic predictions for North American Holstein bulls. *J. Dairy Sci*. 92: 16–24, 2009.

VARONA, L.; SORENSEN, D.; THOMPSON, R. Analysis of litter size and average litter weight in pigs using recursive model. *Genetics*, 177:1791-1799 , 2007.

VERMA, T., PEARL, P. *Equivalence and synthesis of causal models.* In Proceedings of the 6th Conference on Uncertainty in Artificial Intelligence. Volume 6. Cambridge, MA; 1990:220-227, Reprinted in Uncertainty in Artificial Intelligence, 6: 255:268, Elsevier, Amsterdam.

VILLANUEVA B.; PONG-WONG R.; FERNANDEZ J.; TORO M. A. Benefits from marker-assisted selection under an additive polygenic genetic model. *J. Anim. Sci*., 83, 1747–1752, 2005.

WOLC, A.; STRICKER, C.; ARANGO, J. et al. Breeding value prediction for production traits in layer chickens using pedigree or genomic relationships in a reduced animal model. *Genetics Selection Evolution*, 43:5, 2011.

WRIGHT, S. Coefficients of inbreeding and relationship. *Am. Naturalist*, 56:330–338, 1922.

WRIGHT, S. Correlation and causation. *J. Agric. Res*., 201:557-585, 1921.

WU, X-L.; HERINGSTAD, B.; GIANOLA, D. Bayesian structural equation models for inferring relationships between phenotypes: a review of methodology, identifiability, and applications. *J. Anim. Breed. Genet*., 127:3-15, 2010.

WU, X-L.;,HERINGSTAD, B.; CHANG, Y.M.; et al. Inferring relationships between somatic cell score and milk yield using simultaneous and recursive models. *J. Dairy Sci*., 90:3508-3521, 2007.

WU, X-L.; HERINGSTAD, B.; GIANOLA, D. Exploration of lagged relationships between mastitis and milk yield in dairy cows using a Bayesian structural equation Gaussian-threshold model. *Genetics Selection Evololution*, 40:333-357, 2008.

Anexo 1.1. D-Separação

Uma estrutura causal recursiva pode ser representada por um gráfico acíclico direcionado (DAG), que é composto por um conjunto de variáveis (vértices) conectadas por linhas direcionadas (setas). Os pares de vértices conectados representam uma relação de efeito causal. Uma trilha na estrutura causal é uma sequência de vértices conectados. Marginalmente, existe um fluxo de dependência entre variáveis nos extremos da trilha, a menos que exista uma variável *collider* (variável com setas convergindo para ela). Essas variáveis bloqueiam o fluxo de dependência em uma trilha, o que faz com que as variáveis *a* e *b* na estrutura $a \longrightarrow c \longleftarrow b$ sejam independentes, por exemplo. Condicionar a uma variável que não está em um dos extremos da trilha troca o estado de dependência considerando o fluxo através dela. Neste caso, uma variável *collider* permitiria o fluxo de dependência. Duas variáveis em um DAG são ditas ser d-separadas condicionalmente a um sub-conjunto *S* de variáveis remanescentes, se não existem caminhos entre *a* e *b* tal que todos os vértices contidos neles permitem fluxo de dependência. Sob algumas premissas, as d-separações em uma estrutura causal resultam em independência na distribuição conjunta de *y*. Isso pode ser usado para guiar a seleção de estruturas causais que é compatível com a distribuição conjunta dos dados (Spirtes et al., 2000; Pearl, 2009; Rosa et al., 2011) .

Anexo 1.2. Codificação, esperança e variância dos genótipos

A seguinte explicação pode ser encontrada em Gianola e de los Campos (2012). A variável aleatória W denota o genótipo em um loco bialélico, para identificação dos alelos presentes pode-se utilizar códigos para os possíveis genótipos, como, por exemplo:
$W(aa) = -1$, W(Aa) = 0, $W(AA) = 1$ (codificação 1)
ou $W(aa) = 0$, $W(Aa) = 1$ e $W(AA) = 2$ (codificação 2)

As esperanças e variâncias de W sob equilíbrio de Hardy-Weinberg são:
1) Para a primeira codificação:

$$E_{HW}(W) = p^2 - q^2 = (p - q) = \mu$$

$$Var_{HW}(W) = E(x^2) - E^2(x) = p^2 + q^2 - (p - q)^2 = 2pq$$

2) Para a segunda codificação:

$E_{HW}(W) = 2p^2 + 2pq = 2p(p+q) = 2p$

$Var_{HW}(W) = 4p^2 + 2pq - 4p^2 = 2pq$

Pode-se notar que a codificação não altera a variância dos genótipos, mas a média se modifica. Os desvios em relação à média e os desvios padronizados também são invariantes, como pode ser visto a seguir:

$W - E(W) \mid codificação1$:	$W - E(W) \mid codificação2$
$-1-(p-q) = -1-p+q = -2p$	$0-2p$
$0-(p-q) = q-p = 1-2p$	$1-2p$
$1-(p-q) = 1-p+q = 2(1-p)$	$2-2p$

$$W * \frac{W - E(W)}{\sqrt{2pq}}$$

Anexo 1.3. Relação entre a escala da matriz G e A

Segundo Gianola e de los Campos (2012), sob as premissas de distribuição normal dos efeitos genéticos aditivos, variância homogênea atribuída aos diferentes pedaços de cromossomo, mesmo mecanismo de amostragem dos efeitos genéticos ao longo das gerações e de linearidade, o BLUP de um sinal (f) que simboliza $X\beta$ pode ser descrito, como:

$y = f + e = X\beta + e$, em que X é fixo,

$f \sim N(0, Var(f)$ e $Var(f) = XX'Var(\beta)$.

$$BLUP(f) = \left[I + (XX')^{-1} \frac{\sigma_e^2}{Var(\beta)} \right]^{-1} y.$$

Esse preditor se transforma no BLUP descrito por VanRaden (2008):

$$BLUP(\hat{g}) = \left[I + G^{-1} \frac{\sigma_e^2}{Var(\beta) \mid V_{marcadores,HW}} \right]^{-1} y$$, em que

$G = \frac{(X - E(X))(X - E(X))'}{2\sum_{j=1}^{p} p_j(1-p_j)} = \frac{X * X*'}{V_{marcadores,HW}}$ e a matriz de incidência dos genótipos é centrada usando informações de frequência alélica. Para se obterem fenótipos futuros basta inverter a fórmula:

$y = \left[I + G^{-1} \frac{\sigma_e^2}{Var(\beta)\left|V_{marcadores,HW}\right.} \right] \hat{g}$. Pode-se também estimar os efeitos dos marcadores utilizando a teoria clássica do BLUP sob normalidade:

$\hat{\beta} = X'(XX')^{-1} \left[I + (XX')^{-1} \frac{\sigma_e^2}{\sigma_\beta^2} \right]^{-1} y$. Portanto, $\hat{\beta} = X'(XX')^{-1} \hat{g}$ e $\hat{g} = X\hat{\beta}$.

X é a matriz que identifica os genótipos dos animais para os diferentes locos (no Anexo 1.1 acima são apresentados diferentes códigos para os elementos de *X*). Assim,

$$XX' = \begin{bmatrix} \sum_{j=1}^{p} x_{1j}^2 & \sum_{j=1}^{p} x_{1j}x_{2j} & . & \sum_{j=1}^{p} x_{1j}x_{nj} \\ . & \sum_{j=1}^{p} x_{2j}^2 & . & . \\ . & . & . & . \\ . & . & . & \sum_{j=1}^{p} x_{nj}^2 \end{bmatrix}$$

. Sob equilíbrio Hardy-Weinberg,

$E\left(\sum_{j=1}^{p} x_{ij}^2\right) = \sum_{j=1}^{p} Var(x_{ij}) + \sum_{j=1}^{p} E^2(x_{ij}) = \sum_{j=1}^{p} 2p_j q_j + \sum_{j=1}^{p} (p_j - q_j)^2$ e

$E\left(\sum_{j=1}^{p} x_{1j}x_{2j}\right) = \sum_{j=1}^{p} Cov(x_{1j}x_{2j}) + \sum_{j=1}^{p} E(x_{1j})E(x_{2j}) = \sum_{j=1}^{p} 2\phi_{ij} p_j q_j + \sum_{j=1}^{p} (p_j - q_j)^2$, em que ϕ é o relacionamento genético aditivo entre dois indivíduos. Seguindo,

$$E(XX') = \begin{bmatrix} \sum_{j=1}^{p} 2p_jq_j + \sum_{j=1}^{p} (p_j - q_j)^2 & a_{12}\sum_{j=1}^{p} 2p_jq_j + \sum_{j=1}^{p} (p_j - q_j)^2 & . & a_{n3}\sum_{j=1}^{p} 2p_jq_j + \sum_{j=1}^{p} (p_j - q_j)^2 \\ . & \sum_{j=1}^{p} 2p_jq_j + \sum_{j=1}^{p} (p_j - q_j)^2 & . & . \\ . & \textit{simétrica} & . & a_{n,n-1}\sum_{j=1}^{p} 2p_jq_j + \sum_{j=1}^{p} (p_j - q_j)^2 \\ . & . & . & \sum_{j=1}^{p} 2p_jq_j + \sum_{j=1}^{p} (p_j - q_j)^2 \end{bmatrix}$$

Igualmente, se os elementos x são centrados:

$$E\left\{\left[X - E(X_{nxp})\right]\left[X - E(X_{nxp})'\right]\right\} = \left(\sum_{j=1}^{p} 2p_j q_j\right)\begin{bmatrix} 1 & a_{12} & . & a_{1n} \\ . & . & . & . \\ . & \textit{simétrica} & . & a_{n,n-1} \\ . & . & . & 1 \end{bmatrix} = A\left(\sum_{j=1}^{p} 2p_j q_j\right) \text{e}$$

$$\frac{E\left[X - E(X_{nxp})\right]\left[X - E(X_{nxp})'\right]}{\left(\sum_{j=1}^{p} 2\mathrm{p}_j q_j\right)} = A.$$

Então, a matriz de parentesco genômica é:

$G = \dfrac{E\left[X - E(X_{nxp})\right]\left[X - E(X_{nxp})'\right]}{2\sum_{j=1}^{p} p_j(1-p_j)} = \dfrac{X * X*'}{V_{marcadores,HW}}$, que é a realização do processo. Se o processo está em equilíbrio Hardy-Weinberg, então a sua esperança é:

$\dfrac{E\left[X - E(X_{nxp})\right]\left[X - E(X_{nxp})'\right]}{\left(\sum_{j=1}^{p} 2\mathrm{p}_j q_j\right)} = A$. Por exemplo, espera-se que pai e filho tenham parentesco de 0,5, mas esse poderia ser maior ou menor.

CAPÍTULO 2

Busca por estruturas causais fenotípicas com a utilização de informações genômicas

Resumo: Objetivou-se com o estudo utilizar informações de marcadores moleculares para se obter estrutura causal entre fenótipos por meio do algoritmo *Inductive Causation* (IC), e comparar a inserção da matriz dos coeficientes de parentesco de Wright e da matriz de relacionamento genético genômico em equações multivariadas de modelos mistos a fim de se recuperar a rede causal fenotípica. Para tal, foram feitas simulações de três estruturas populacionais que se diferiram pelo número de indivíduos selecionados por geração e pelo tamanho da progênie originada a partir de cada par de indivíduos formado para reprodução, o que influenciou os níveis de parentesco e endogamia. Foram obtidos na simulação os fenótipos, genótipos e informações de *pedigree*, necessários para as análises. O algoritmo IC foi utilizado para a busca da estrutura causal simulada. As inserções das matrizes de parentesco clássica e genômica no modelo multicaracterísticas para obtenção das amostras *a posteriori* das (co)variâncias residuais, que são utilizadas como entrada do algoritmo, foram comparadas com base no gráficos acíclicos direcionados obtidos. Para cada situação, variou-se o número de observações de *pedigree*, genótipos e fenótipos utilizadas nas análises. As diferentes formas de matrizes e a quantidade variável de fenótipos foram utilizadas a fim de saber a necessidade quantitativa de informações para se obter a correta estrutura causal entre as características, e também para avaliar a sensibilidade dos métodos frente à quantidade de informação disponível. A fim de se buscar redes causais fenotípicas por meio do algoritmo IC, a inserção da matriz de relacionamento genético genômico no método, de acordo com os gráficos acíclicos direcionados observados e obtidos para as inúmeras repetições de simulação, apresenta vantagens quando comparada à utilização da matriz de parentesco tradicional, principalmente, em situações específicas com estrutura populacional formada por indivíduos com alto grau de parentesco e endogâmicos, ou ainda, quando há escassez de informações fenotípicas e de parentesco.

Palavras-chave: algoritmo *Inductive Causation*, causalidade, matriz dos coeficientes de parentesco de Wright, matriz de relacionamento genético genômico, redes causais fenotípicas, simulação

Abstract: *It was aimed to use molecular markers information to obtain causal structure between phenotypes by means of the Inductive Causation (IC) algorithm, and to compare the insertion of the Wright's coefficients matrix and the genomic relationship matrix in multivariate mixed model equations in order to recover the phenotype causal network. For such goal, it was made a simulation of three different population structures that distinguish each other by the number of selected individuals per generation and the number of progeny of each pair of individuals formed for reproduction, what influenced the levels of parenthood and endogamy. It was obtained phenotypes, genotypes and pedigree information necessary for analyzes. The IC algorithm was implemented for the search of the simulated causal structure. The incorporation of the classic and genomic relationship matrices in the multiple trait models for the achievement of the residual (co)variances posterior distributions that are used as the algorithm input was compared by means of the obtained acyclic semi-directed graphs. For each scenario, it was used different number of observations from pedigree, genotypes and phenotypes in the analyzes. The different matrices and number of phenotypes were used in order to know the quantitative necessity of information to obtain the correct causal structure between traits, as well as, to evaluate the sensibility of such methods under the available information. In order to search causal structures between phenotypes using the IC algorithm, the insertion of the genomic relationship matrix in the method, according to the observed directed acyclic graphs obtained for several simulation repetitions, shows advantages when compared to the traditional relationship matrix, especially in specific scenarios where the population structure is composed by highly related and endogamic individuals, or when there is a lack of phenotype and pedigree information.*

Keywords: causality, genomic relationship matrix, Inductive Causation algorithm, phenotype causal networks, simulation, Wright's coefficients relationship matrix

INTRODUÇÃO

A causalidade pode ser definida como uma relação de eventos particulares, em que um primeiro evento ocorre e causa um segundo evento a acontecer (Spirtes, 2000). Em outras palavras, existe causalidade quando um evento secundário é, pelo menos em parte, determinado pela existência de um evento primário. A percepção do relacionamento causal em um sistema de variáveis poderia gerar informação mais poderosa do que a obtenção de parâmetros de associação linear simples entre características, como, por exemplo, a correlação. O conhecimento da causalidade permite quantificar resultados que são consequências de intervenções externas, provendo a habilidade de controle de eventos futuros (Blaisdell et al., 2006; Pearl, 2009; Valente et al., 2013b). Especificamente na área da produção animal, na qual é muito importante o planejamento de práticas de manejo, é de grande interesse a predição dos efeitos de mudanças aplicadas ao sistema, como, por exemplo, o efeito do nível nutricional de diferentes dietas sobre características de produção ou o efeito da qualidade da água sobre a incidência de doenças (Rosa & Valente, 2012). Ainda, em melhoramento animal, o ajuste de modelos que consideram relações causais, como, o Modelo de Equações Estruturais (SEM), culmina em informações adicionais aos modelos clássicos. Por exemplo, pode ser feita a predição de valores genéticos específicos para diferentes cenários frente à modificação da rede causal fenotípica. Um cenário em que isso se aplica seria quando há a fixação do valor de uma característica por meio de intervenções externas ao sistema, como em casos de práticas de manejos reprodutivos ou ao se abater o animal a certa idade (Valente et al., 2013b).

Existem basicamente dois meios pelos quais a magnitude do relacionamento causal entre características fenotípicas pode ser estimada. Um, é o uso de experimentos delineados, em que todas as fontes de variação fenotípica são controladas por casualização e o objeto de estudo é isolado como a causa de uma característica particular. O segundo se dá a partir da utilização de dados observacionais coletados a campo, que é o caso da maioria dos dados disponíveis em estudos de produção e melhoramento genético animal. Embora, a primeira ferramenta possa ser vista como mais controlada e simples que a segunda, ela não é sempre uma representação fidedigna de um cenário real ou não pode ser realizada por causa de preocupações éticas e logísticas.

A modelagem causal com base em dados observacionais foi apresentada por Wright (1921) e Haavelmo (1943) de uma maneira em que a qualidade da informação causa-efeito poderia ser combinada com procedimentos estatísticos a fim de proporcionar uma medida quantitativa do relacionamento entre as características de interesse. No modelo, uma ou mais variáveis podem aparecer como variáveis explanatórias nas equações da variável resposta, a qual pode ser causa de outras variáveis, representando relacionamentos simultâneos ou recursivos. Por esse motivo, o Modelo de Equações Estruturais poderia ser uma alternativa ao tradicional modelo multicaracterísticas (MTM). Em genética e melhoramento animal, essa metodologia não recebeu muita atenção até que Gianola e Sorensen (2004) adaptaram o SEM ao contexto das Equações de Modelos Mistos considerando herança genética. Desde então, o interesse no SEM têm aumentado e vários estudos foram publicados usando dados de diferentes espécies e características (de Los campos et al., 2006a; de Los campos et al., 2006b; Varona et al., 2007; Wu et al., 2008; Valente et al., 2011).

A inferência de efeitos causais por meio do SEM assume que a estrutura causal real entre características é conhecida, ou seja, é necessário que se defina a *priori* a estrutura causal entre as características de interesse ao se ajustar o modelo. Aparentemente essa premissa pode ser cumprida pela comparação exaustiva de todos os modelos de todas as possíveis estruturas por algum critério, como, o AIC (Akaike, 1974) ou o BIC (Schwartz, 1978). Entretanto, o número de possíveis estruturas causais é grande e seria muito laborioso compará-las. Ainda, o número de estruturas possíveis aumenta consideravelmente com o aumento da quantidade de variáveis no sistema (Verma & Pearl, 1990; Pearl, 2000; Valente et al., 2010). Uma alternativa, como primeiramente foi utilizada por muitos autores (de Los campos et al., 2006a; de Los campos et al., 2006b; Varona et al., 2007; Wu et al., 2008), é a pré-seleção de algumas estruturas a serem comparadas usando o conhecimento a *priori* sobre a biologia das características, informações obtidas de experimentos ou, até mesmo, a sequência temporal de expressão dos fenótipos. Em casos em que não se admite o conhecimento do relacionamento entre os fenótipos, poderiam ser usados alguns algoritmos que exploram o espaço de estruturas causais, como, o algoritmo *Inductive Causation* - IC (Verma & Pearl, 1990). Tal metodologia foi desenvolvida para recuperar Gráficos Acíclicos Direcionados (DAGs), que representam estruturas causais recursivas entre variáveis, por meio da distribuição conjunta dos dados. Um gráfico causal pode ser interpretado como uma família de modelos causais dos quais relacionamentos causais qualitativos podem ser deduzidos. Apesar de não apresentar informação quantitativa sobre o efeito de uma variável sobre outra, os gráficos podem ser

muito eficientes em representar independências condicionais entre variáveis que necessariamente seguem uma estrutura causal que as originam (Valente et al., 2013a).

A utilização do algoritmo IC na área da genética e melhoramento animal foi proposta por Valente et al. (2010), depois disso, alguns estudos já têm usado a metodologia para obtenção de redes causais entre fenótipos e posterior utilização do Modelo de Equações Estruturais (Valente et al., 2011, Bouwman et al., 2014). Esse algoritmo utiliza a matriz de (co)variância residual do clássico modelo multicaracterísticas, previamente ajustado, visando fazer decisões estatísticas. Mais especificamente, são utilizadas distribuições a *posteriori* das correlações residuais parciais entre pares de variáveis para determinar dependência condicional entre elas. Considerando isso, é plausível afirmar que a adição de informações que gerem predições sobre as (co)variâncias residuais mais precisas poderia enriquecer o desempenho para a busca da verdadeira estrutura causal existente. Tal ganho pode ser alcançado, por exemplo, pela adição de informações moleculares no modelo de predição quando o fato de possuir somente os dados de pedigree e de fenótipos não é suficiente. Com a atual disponibilidade de informações de marcadores de DNA para muitos locos espalhados ao longo do genoma, pode ser construída a matriz de relacionamento genético genômico (matriz *G*), na qual os elementos mostram a verdadeira proporção do genoma compartilhado para medir a similaridade entre indivíduos. Essa matriz é considerada mais precisa do que a matriz de coeficientes de parentesco tradicional construída a partir de dados de *pedigree* (matriz *A*), uma vez que a última reproduz apenas a expectativa de compartilhamento de alelos (VanRaden et al., 2008; Forni et al., 2011). Uma consideração adicional é que a falta e erros de dados de *pedigree*, que comumente ocorrem em anotações de observações obtidas a campo, podem ser superados pela genotipagem de toda a população de interesse, assim a rede fenotípica causal poderia ser de toda forma recuperada.

Objetivou-se com o estudo utilizar informações de marcadores moleculares para se obter estrutura causal entre fenótipos por meio do algoritmo *Inductive Causation*. Assim como, comparar a inserção da matriz de parentesco clássica e da matriz de relacionamento genético genômico em equações multivariadas de modelos mistos a fim de se recuperar a rede causal fenotípica.

MATERIAL E MÉTODOS

Simulação

Para se obterem amostras dos fenótipos, assim como, dos genótipos e do pedigree dos indivíduos, informações necessárias para a construção da matriz dos coeficientes de parentesco tradicional (matriz *A*) e a matriz de relacionamento genético genômico (matriz *G*), algumas simulações foram realizadas usando o programa R (códigos pessoais). Foram feitas 10 repetições de simulação para cada uma das três diferentes estruturas populacionais estudadas, as quais se diferiam pelo número de indivíduos selecionados por geração e pelo tamanho da progênie originada a partir de cada par de indivíduos formado para reprodução, o que também influencia o parentesco entre os animais da população e o grau de endogamia.

Para cada simulação foi gerado um genoma com 30 cromossomos de 1 Morgan cada, os quais continham 10 locos bi-alélicos equidistantes que foram designados como QTLs putativos e foram flanqueados por 70 marcadores do tipo polimorfismos de nucleotídeos simples (SNPs) em cada lado. Assim, somou-se um total de 300 QTLs e aproximadamente 20.000 SNPs igualmente distribuídos ao longo do genoma. A composição de cada um dos cromossomos pode ser representada, como:
$M_1, ..., M_{70}, Q_1, M_1, ..., M_{70}, Q_2, ..., Q_{10}, M_1,..., M_{70}$. Em que, M_i é um marcador e Q_i é um QTL.

No início da simulação foram computadas 1000 gerações de evolução da população (t_1 a t_{1000}) sob seleção e acasalamento ao acaso com tamanho de população de 100 indivíduos, sendo considerados 50 indivíduos de cada sexo. O balanço da frequência gênica foi definido por mutação, em uma taxa de 2.5×10^{-3} por loco e por geração, e pela recombinação gênica (*crossing-over*), com probabilidade de ocorrência dependente da distância entre locos, de forma a se permitir criar desequilíbrio de ligação entre os alelos. Todos os alelos na geração base foram fixados como "0" e depois mutados para "1" de acordo com a probabilidade de mutação. Como resultado, cerca de 99 por cento dos locos estavam segregando ao final das 1000 gerações. Como esperado, a distribuição da frequência alélica teve a forma de U (Figura 1). Após 1000 gerações foi feita uma expansão da população por meio de acasalamento fatorial entre cada um dos 50 machos da última geração e 10 fêmeas escolhidas ao acaso. Para então, formação de 11 gerações (g_1 a g_{11}) com 500 indivíduos em cada, os quais tiveram a

informação de parentesco (*pedigree*) armazenada**.** O genótipo e fenótipo dos indivíduos das últimas quatro gerações (g_8 a g_{11}) também foram salvos. Nas simulações não houve sobreposição das gerações. A partir da geração g_1 cessou-se a mutação, portanto, passaram a atuar sobre a frequência alélica apenas as forças de recombinação gênica e seleção. No Anexo 2.1 está exposta parte do código que é comum para todos os cenários populacionais desenvolvidos e que representa a evolução das primeiras 1001 gerações, como descrito acima.

A maneira com que as gerações g_1 a g_{11} evoluíram dependeu do número de animais amostrados e acasalados aleatoriamente para originar a próxima geração e do número de progênie para cada casal formado.

Situação1: Para a primeira estrutura populacional simulada, machos e fêmeas de cada uma das gerações g_1 a g_{10} foram aleatoriamente escolhidos e acasalados ao acaso para produzir 500 descendentes por geração. Assim, os números de pais, irmãos completos e meio-irmãos por indivíduo poderiam variar entre as repetições feitas, sendo esperado que o relacionamento genético entre os indivíduos seja fraco e existente principalmente por conta da relação pai-filho em seguidas gerações.

Situação 2: Na segunda estrutura populacional, para cada geração, cinco machos e 100 fêmeas foram escolhidos aleatoriamente e acasalados em esquema fatorial a fim de gerar os indivíduos da próxima geração, assim a maior fonte de informação de relacionamento genético entre os indivíduos é proveniente de famílias de meio-irmãos e o parentesco médio esperado entre os indivíduos é maior que o do cenário anterior.

Situação 3: Para a terceira estrutura populacional, dois machos e 25 fêmeas foram amostrados e acasalados ao acaso, produzindo 10 indivíduos por casal formado. Nessa situação, cada um dos indivíduos das diferentes gerações teria parentesco com vários meio-irmãos e nove irmãos-completos, o que causa magnitudes de coeficiente de parentesco e de endogamia média ainda maior que no segundo cenário.

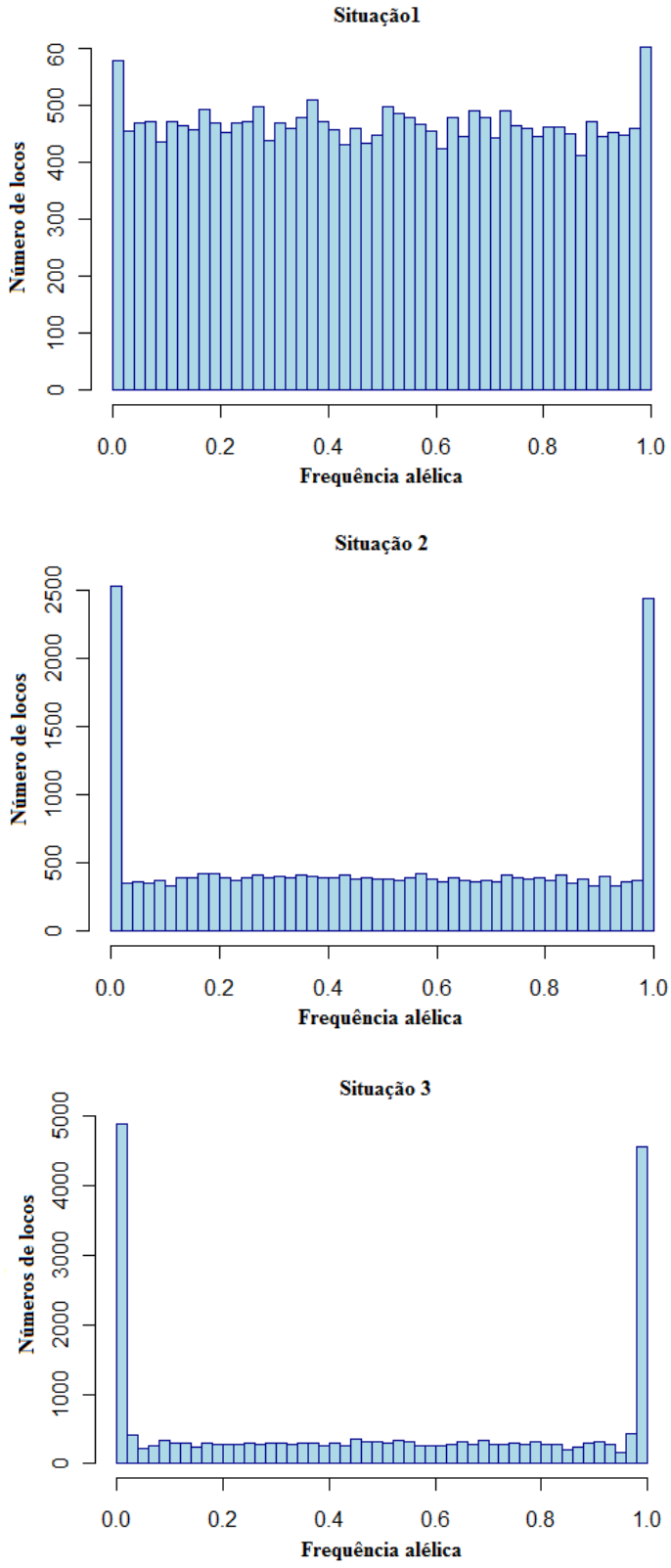


Figura 1. Freqüência alélica nas quatro últimas gerações de indivíduos de uma repetição de simulação tomada aleatoriamente para cada uma das três estruturas populacionais estudadas.

A fim de descrição e controle das diferentes endogamias médias resultantes das simulações, para as três estruturas populacionais, os coeficientes de parentesco de Wright (Wright, 1922) e os coeficientes de relacionamento genético genômico (VanRaden, 2008) foram calculados. Os resultados para uma repetição tomada aleatoriamente em cada situação são demonstrados na Tabela 1.

Tabela 1. Coeficientes médios de parentesco de Wright e genômicos, endogamia média, e correlação entre elementos da matriz *A* e *G* obtidos de uma repetição de simulação escolhida ao acaso para cada uma das três situações de simulação realizadas

	Situação 1		Situação 2		Situação3	
	Matriz *A*	Matriz *G*	Matriz *A*	Matriz *G*	Matriz *A*	Matriz *G*
MED	1,008	0,9992	1,163	0,9868	1,396	0,9690
MEF	0,0174	-0,0004	0,3482	-0,0166	0,8166	-0,0004
CED		0,4978		0,4034		-0,3049
CEF		0,8539		0,8211		0,6410

MED = Média dos elementos da diagonal; MEF = Média dos elementos fora da diagonal; CED = Correlação entre os elementos da diagonal das matrizes *A* e *G*; CEF = Correlação entre os de fora da diagonal das matrizes *A* e *G*

Os efeitos de QTL foram amostrados de uma distribuição normal multivariada parametrizada de modo a refletir aproximadamente variâncias e covariâncias genéticas pré-estabelecidas entre cinco características. O valor fenotípico de cada indivíduo para cada uma das cinco características foi composto por uma soma da média populacional da característica, do valor genético aditivo verdadeiro (soma dos efeitos de todos os QTLs presentes em cada cromossomo do indivíduo) e do efeito aleatório residual, que foi amostrado a partir de uma distribuição normal com média zero e variância igual a uma razão que permitisse que a herdabilidade da característica fosse a desejada. Para todas as cinco características estudadas a herdabilidade considerada foi de 0,3333.

O modelo de equações estruturais que determina a estrutura causal utilizada para obtenção dos fenótipos (Figura 2) foi:

$$y_{i1} = \mu_1 + u_{i1} + e_{i1}$$

$$y_{i2} = \mu_2 + \lambda_{21} y_{i1} + u_{i2} + e_{i2}$$

$$y_{i3} = \mu_3 + \lambda_{32} y_{i2} + u_{i3} + e_{i3}$$

$$y_{i4} = \mu_4 + \lambda_{42} y_{i2} + u_{i4} + e_{i4}$$

$$y_{i5} = \mu_5 + \lambda_{53} y_{i3} + \lambda_{54} y_{i4} + u_{i5} + e_{i5} ,$$

em que, y_{ij} e e_{ij}, são os efeitos fenotípicos e residuais para a característica *j* (j = 1,...,5) observados no animal i, μ_j é a média da característica *j*, u_{ij} é o efeito genético aditivo do animal *i* para a característica *j*, e $\lambda_{jj'}$ é a taxa de mudança da característica *j* com respeito à característica *j'*. As médias correspondentes a cada característica, μ_1, μ_2, μ_3, μ_4 e μ_5 consideradas nas equações foram 80, 60, 70, 140 e 60, respectivamente. Os coeficientes estruturais λ_{21}, λ_{32}, λ_{42}, λ_{53} e λ_{54} usados para a simulação foram, respectivamente, 0,5; 0,35; -0,5; 0,8 e -0,4. O código em R desenvolvido para a obtenção dos fenótipos dos indivíduos vinculados à estrutura causal e herdabilidade desejadas pode ser observado no Anexo 2.2.

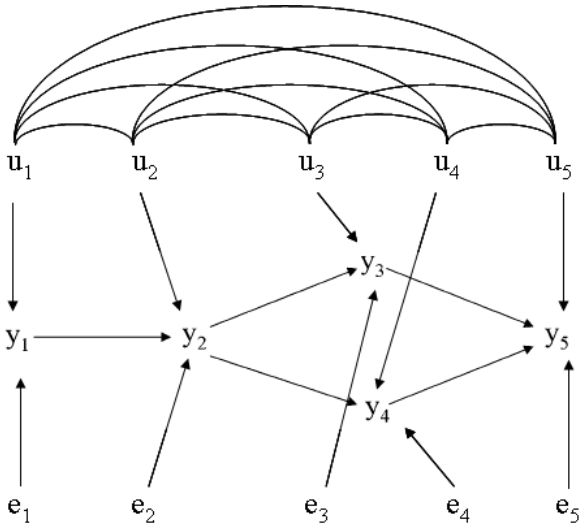


Figura 2. Diagrama do modelo do qual os dados simulados foram obtidos. y_j é uma medida observada da característica j, u_j é o efeito genético aditivo que contribui para a característica j, e e_j é o resíduo associado à característica j. Os arcos conectando $u's$ representam as correlações genéticas aditivas.

O gráfico acíclico direcionado (DAG) que expressa a estrutura causal entre características fenotípicas no modelo causal acima pode ser observado abaixo na Figura 3. Esta estrutura foi o alvo de busca nas avaliações realizadas nesse estudo.

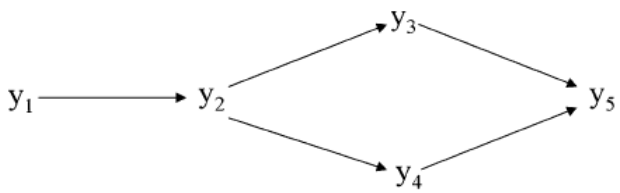


Figura 3. Gráfico acíclico direcionado que representa a estrutura causal simulada entre as características e que se espera obter após utilização do algoritmo IC.

Busca pela estrutura causal

Obtidos os dados para todas as estruturas populacionais simuladas (*pedigree* para as últimas 11 gerações, fenótipos e genótipos para as últimas quatro gerações), foram realizadas análises para teste da hipótese do trabalho.

A busca pela estrutura causal foi feita por meio do algoritmo *Inductive Causation* – IC (Pearl, 2000). Como demonstrado no diagrama do modelo de simulação de fenótipos (Figura 2), os resíduos foram considerados independentes e os efeitos recursivos foram as fontes de covariância entre as variáveis. No contexto padrão de modelos de equações estruturais, resíduos independentes refletindo a premissa de suficiência causal (Pearl 2000, Spirtes et al, 2000) atribuem às relações causais toda associação entre características. Este cenário permite a aplicação do algoritmo IC. No entanto, no contexto de modelos mistos, associações também são originadas por fatores genéticos. Neste caso, os efeitos genéticos aditivos correlacionados podem atuar como elementos de confusão na tentativa de buscar a rede causal fenotípica por meio da distribuição conjunta dos fenótipos. Para contornar esse problema na busca pela estrutura causal fenotípica, Valente et al. (2010) adaptaram a abordagem de uso do algoritmo IC para situações de genética quantitativa. A proposta é aplicar o algoritmo de busca em informações da distribuição conjunta das características condicionais aos efeitos genéticos, o que bloqueia os fatores de confusão. Esta informação é fornecida pela matriz de variância e covariância residual após ajuste do modelo multicaracterísticas tradicional. Esse método foi o escolhido no atual estudo e pode ser descrito da seguinte maneira:

1. Ajuste do modelo multicaracterísticas e obtenção de amostras da distribuição *a posteriori* da matriz de variâncias e covariâncias residuais.

2. Aplicação do algoritmo *Inductive Causation* (Pearl, 2000) às amostras *a posteriori* do resíduo para tomar uma série de decisões estatísticas que são baseadas na correlação parcial entre características. Especificamente, para cada questionamento sobre a independência estatística entre variáveis, *a* e *b*, dado um determinado conjunto de variáveis *S* no algoritmo foi primeiramente obtida a distribuição da correlação parcial do resíduo, $\rho_{a,b|S}$. Posteriormente, o intervalo de maior probabilidade de 95% das amostras, HPD95, foi computado e se zero estava contido nele, foi declarado que $\rho_{a,b|S}$ era nula. De outra forma, *a* e *b* foram considerados condicionalmente dependentes.

O algoritmo IC aplicado pode ser separado em três passos:

1º. As correlações parciais são utilizadas para buscar traços que conectam variáveis adjacentes e são obtidos gráficos não direcionados. Para tal, se todas as correlações parciais entre duas variáveis condicionais a todos os possíveis conjuntos de variáveis remanescentes forem diferentes de zero, então um traço é colocado entre essas variáveis.

2º. As correlações parciais são utilizadas para buscar *unshielded colliders* - vértices nos gráficos no qual setas convergem de dois outros vértices não conectados (Spirtes, 2000), como pode ser visto entre as variáveis y_3, y_4 e y_5 (Figura 3) - em vista de orientar alguns traços do gráfico não direcionado obtido no passo anterior. Para isso, se todas as correlações parciais entre duas variáveis não adjacentes, *a* e *b*, que possuem uma variável adjacente em comum, *c*, tomadas condicionalmente a qualquer possível conjunto de variáveis que possuem essa mesma variável *c,* não conterem zero no HPD95, duas setas são orientadas em direção a *c*.

3º. Todos os traços não direcionados devem ser orientados de uma forma que não se formem novos *unshielded colliders* e nem ciclos, caso possível.

Com o objetivo de comparar a utilização da matriz de parentesco tradicional (*A*), e da matriz de parentesco genômico (*G*) para a busca da estrutura causal verdadeira (Figura 3), o modelo multicaracterísticas foi ajustado com a inclusão de cada uma, para cada cenário simulado. Para o estudo das três estruturas populacionais, foram consideradas diferentes quantidades de informações de genótipos e de parentesco utilizadas para construção das matrizes de relacionamento genético, que constituíram as diferentes situações de análise. A diferenciação em número de informações foi dada pelo aproveitamento de dados das diversas gerações de indivíduos existentes. Como foram feitas 10 repetições de simulação para cada uma das três estruturas populacionais estudadas, tiveram que ser construídas as diferentes formas de matrizes de relacionamento genético para todas essas repetições.

Para a formação da matriz *A*, foram utilizadas as informações de *pedigree* dos indivíduos de todas as últimas 11 gerações, assim também, como a partir apenas das informações de parentesco de indivíduos das quatro últimas gerações, das três últimas gerações, das duas últimas gerações e da última geração.

A matriz *G* foi construída a partir dos genótipos dos indivíduos das quatro últimas gerações, das três últimas gerações, das duas últimas gerações, e, ainda, com apenas os genótipos dos indivíduos da última geração.

Após todas as gerações de simulação, no final da geração g_{11}, a porcentagem de alelos segregando variou de aproximadamente 75 a 98 por cento, sendo que apenas os SNPs segregantes foram utilizados nas análises. Portanto, o número de marcadores utilizados na construção da matriz de relacionamento genético genômico também variou.

Visto isso, as análises comparativas foram feitas com base em cinco tipos de matriz *A* e quatro formas de matriz *G*, que eram diferentes na quantidade de informações utilizadas para sua construção. O número de informações fenotípicas utilizadas nas análises também variou. Foram realizadas avaliações que comportaram todas as informações de fenótipo disponíveis, ou seja, das quatro últimas gerações de indivíduos, assim também, como, informações referentes aos fenótipos dos indivíduos das três últimas gerações, das duas últimas ou apenas da última geração. As diferentes formas de matrizes e a quantidade variável de fenótipos foram utilizadas a fim de saber a necessidade quantitativa de informações para se obter a correta estrutura causal entre as características, e também para avaliar a sensibilidade dos

métodos frente à quantidade de informação disponível. Na Tabela 2 são especificadas todas as avaliações realizadas e o número de gerações consideradas para obtenção das informações de genealogia, de genótipos e de fenótipos em cada situação.

Tabela 2. Avaliações realizadas e número de informações consideradas

Análise	**Genótipos (NG / NI)**	**Pedigree (NG / NI)**	**Fenótipos (NG / NI)**
Situação 1/1[a]	-	11 / 5500	4 / 2000
Situação 1/1[b]	-	4 / 2000*	4 / 2000
Situação 1/1[c]	4 / 2000	-	4 / 2000
Situação 1/2[a]	-	11 / 5500	3/ 1500
Situação 1/2[b]	-	3 / 1500*	3 / 1500
Situação 1/2[c]	3/ 1500	-	3 / 1500
Situação 1/3[a]	-	11 / 5500	2 / 1000
Situação 1/3[b]	-	2 / 1000*	2 / 1000
Situação 1/3[c]	2 / 1000	-	2 / 1000
Situação 1/4[a]	-	11 / 5500	1 / 500
Situação 1/4[b]	-	1 / 500*	1 / 500
Situação 1/4[c]	1 / 500	-	1 / 500
Situação 2/1[a]	-	11 / 5500	4 / 2000
Situação 2/1[b]	-	4 / 2105	4 / 2000
Situação 2/1[c]	4 / 2000	-	4 / 2000
Situação 2/2[a]	-	11 / 5500	3/ 1500
Situação 2/2[b]	-	3 / 1605	3 / 1500
Situação 2/2[c]	3/ 1500	-	3 / 1500
Situação 2/3[a]	-	11 / 5500	2 / 1000
Situação 2/3[b]	-	2 / 1105	2 / 1000
Situação 2/3[c]	2 / 1000	-	2 / 1000
Situação 2/4[a]	-	11 / 5500	1 / 500
Situação 2/4[b]	-	1 / 605	1 / 500
Situação 2/4[c]	1 / 500	-	1 / 500

Situação 3/1[a]	-	11 / 5500	4 / 2000
Situação 3/1[b]	-	4 / 2027	4 / 2000
Situação 3/1[c]	4 / 2000	-	4 / 2000
Situação 3/2[a]	-	11 / 5500	3/ 1500
Situação 3/2[b]	-	3 / 1527	3 / 1500
Situação 3/2[c]	3/ 1500	-	3 / 1500
Situação 3/3[a]	-	11 / 5500	2 / 1000
Situação 3/3[b]	-	2 / 1027	2 / 1000
Situação 3/3[c]	2 / 1000	-	2 / 1000
Situação 3/4[a]	-	11 / 5500	1 / 500
Situação 3/4[b]	-	1 / 527	1 / 500
Situação 3/4[c]	1 / 500	-	1 / 500

NG / NI = Número de Gerações e Número de Informações utilizadas. O número de informações fenotípicas é igual ao número de indivíduos em cada geração (500) vezes o número de gerações. Todos os indivíduos apresentaram fenótipos para todas as características; [a] = Análise realizada com a inclusão da matriz *A* no modelo misto considerando as onze gerações de *pedigree*, por isso a presença de um traço no número de genótipos considerados na mesma linha; [b] = Análise realizada com a inclusão da matriz *A* no modelo misto considerando número de gerações de *pedigree* igual ao número de gerações de fenótipo; [c] = Análise realizada com a inclusão da matriz *G* no modelo misto, por isso a presença de um traço no número de informações de *pedigree* utilizadas na mesma linha; * = Número aproximado de observações de *pedigree* utilizadas. O número de informações genealógicas é igual ao número de gerações vezes o número de indivíduos em cada geração (500), mais o número de pais que originaram a primeira geração considerada. Na Situação 1 esse número varia de acordo com a repetição de simulação para essas análises, pois o número de progenitores é variável

Em todas as análises as matrizes de relacionamento genético construídas foram incorporadas ao sistema de equações de modelo misto e obtidas as soluções por meio de inferência bayesiana. O software utilizado para as avaliações foi o Gibbs1F90 (MISZTAL et al., 2012). O modelo multicaracterísticas foi ajustado para se obter as amostras *a posteriori* da matriz de (co)variância residual. O modelo ajustado incluiu apenas o efeito sistemático da média populacional, o efeito aleatório do valor genético do animal e o efeito residual. O modelo geral pode ser descrito em notação matricial da seguinte maneira:

$$y = X\beta + Za + e,$$

em que y é o vetor coluna de fenótipos; X é a matriz de incidência dos efeitos sistemáticos; β é o vetor do valor médio de cada característica; Z é a matriz de incidência dos efeitos

aleatórios; *a* é o vetor dos valores genéticos aditivos dos animais para cada característica; e *e* é o vetor de efeitos residuais.

Com distribuição conjunta dos vetores *a* e *e* igual a:

$$\begin{bmatrix} a \\ e \end{bmatrix} \sim N\left\{\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} G^* \otimes P & 0 \\ 0 & R \otimes I \end{bmatrix}\right\},$$

em que, *G** é a matriz de (co)variância genética aditiva; *P* é a matriz de relacionamento genético dos indivíduos, e nesse trabalho será *A* (matriz dos coeficientes de parentesco de Wright) ou *G* (matriz de relacionamento genético genômico); *R* é referente à matriz de variâncias e covariâncias residuais do modelo; *I* é uma matriz identidade.

A seguinte distribuição *a priori* conjunta foi considerada para os parâmetros de local e dispersão do modelo:

$p(\beta, a, G^*, R) = p(\beta)\,p(a|G^*)\,p(G^*)\,p(R) \propto N(a|0, G^* \otimes P)\,IW(G^*|\nu_S, S)\,IW(R|\nu_R, R^*)$, em que, $N(a|0, G^* \otimes P)$ é uma distribuição Normal Multivariada centrada em 0 e com matriz de covariância $G^* \otimes P$; $IW(G|\upsilon_S, S)$ é uma distribuição Wishart Invertida com parâmetros de escala com ν_S graus de liberdade e matriz de parâmetros de escala *S*; e $IW(R|\nu_R, R^*)$ é uma distribuição Wishart Invertida com parâmetros de escala com ν_R graus de liberdade e matriz de parâmetros de escala R^*. Foi considerada para o efeito sistemático uma distribuição constante imprópria.

A probabilidade *a posteriori* conjunta de todas as variáveis do modelo é então:

$p(\beta, a, G^*, R|y) = p(y|\beta, a, R)\,p(a|G^*)\,p(G^*)\,p(R)$. O amostrador de Gibbs foi utilizado para se obter as amostras dessa distribuição usando as distribuições condicionais completas *a posteriori*.

O tamanho da cadeia amostral considerada em cada análise para cada situação de simulação variou de 50.000 a 500.000 iterações dependendo do diagnóstico de convergência que foi realizado por meio de inspeção visual da cadeia amostral de cada parâmetro do modelo após construção do gráfico de traço e da determinação do tamanho efetivo de cadeia e erro de Monte Carlo. O número de primeiras amostras a serem descartadas também dependeu dessa inspeção realizada.

O método para a construção da matriz de relacionamento genômico (G) utilizada foi proposto por VanRaden (2008), em que:

$$G = \frac{(M - P)(M - P)'}{2\sum_{j=1}^{m} p_j(1 - p_j)},$$

em que, M é uma matriz que especifica quais marcadores cada animal herdou. M_{ij} é -1 se o animal *i* para o SNP j é homozigoto 00; 0 se é heterozigoto 01, e 1 se é homozigoto 11 . Seja p_j a frequência para o alelo 1 em determinado loco na população, P contém a subtração dessa frequência por 0,5 e multiplicada por 2, tal como, $2(p_j - 0{,}5)$. Para evitar problemas potenciais com a inversão de G na resolução das equações do modelo misto, foi utilizada a seguinte matriz mesclada nas avaliações, $G = \varphi G_{orig} + (1 - \varphi)A$, em que G_{orig} é a matriz *G* antes de ser modificada, e o φ escolhido foi igual a 0,05 de acordo com os resultados apresentados no mesmo trabalho de VanRaden (2008).

As utilizações da matriz de parentesco tradicional e genômica no modelo foram comparadas quanto à capacidade de recuperar a saída esperada de acordo com a estrutura causal verdadeira. Para determinar a escolha do melhor método, foi contado o número de estruturas corretas entre as dez repetições, assim também, como, utilizada uma escala de proximidade da estrutura encontrada após o segundo passo do algoritmo IC e a estrutura correta que deveria ser observada. A seguir, é apresentada essa escala com ordem de descrição igual à importância relativa das características observadas nos gráficos acíclicos direcionados obtidos:

1) Existência de todas as relações de dependência entre as características com direcionamento correto do fluxo causal; 2) Presença correta apenas da existência da dependência causal entre características, mas sem o direcionamento de setas; 3) Presença de traços representando fluxo causal entre características onde no gráfico verdadeiro não existe; 4) Direcionamento incorreto dos fluxos causais entre características; 5) Ausência de relação causal entre as características onde deveria haver o traço representando dependência entre elas.

Como as análises demandaram grande tempo e memória computacional, foi necessária a utilização de programação em paralelo em um sistema de integração de computadores de diversas universidades americanas, Condor (Litzkow et al., 1988), o qual permitiu que fossem

feitos diversos trabalhos simultâneos. As análises foram realizadas por meio de recursos computacionais e de servidores situados na Universidade de Wisconsin – Madison, EUA.

RESULTADOS E DISCUSSÃO

Após aplicação do segundo passo do algoritmo IC às amostras *a posteriori* das (co)variâncias residuais obtidas por ajuste do modelo multicaracterísticas deve-se recuperar as estruturas causais existentes entre as características representadas pelo seguinte gráfico acíclico semi direcionado:

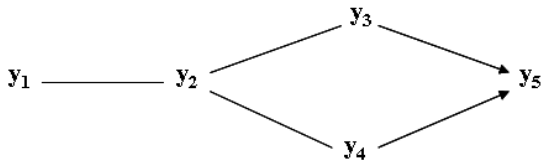


Figura 4. Gráfico acíclico semi direcionado obtido após o segundo passo do algoritmo IC que deve ser observado caso a estrutura causal correta seja recuperada.

O gráfico acima representa um conjunto de quatro estruturas causais estatisticamente equivalentes em que está contida a verdadeira estrutura simulada (Figura 5). Por meio das probabilidades condicionais não é possível distinguir o gráfico que representa a estrutura verdadeira, alvo de estudo nesse trabalho, das outras três remanescentes. Há, portanto, limite de aprendizado quando se usa o algoritmo IC para certas redes causais de fenótipos. Contudo, a informação obtida com a utilização de tal algoritmo é totalmente aproveitável no sentido de que o número enorme de estruturas possíveis de se obter quando se tem cinco características no modelo é reduzido para apenas quatro estruturas que se adéquam à distribuição de probabilidades dos dados. Após obtenção desse número limitado de estruturas e caso haja interesse no ajuste de modelos que consideram associações causais entre características, a fim de se obter a magnitude da relação causal entre essas variáveis, como o Modelo de Equações Estruturais, pode-se fazer a escolha objetiva da estrutura a ser utilizada por uso do conhecimento biológico a respeito das características, por exemplo, se a expressão fenotípica de uma característica sabidamente ocorre antes da outra. A utilização de métodos de comparação de modelos baseados em verossimilhança não seria alternativa uma vez que as observações fenotípicas refletiriam mesma probabilidade de ocorrência condicionalmente às estruturas de mesma classe (Valente et al., 2010; Valente et al., 2011).

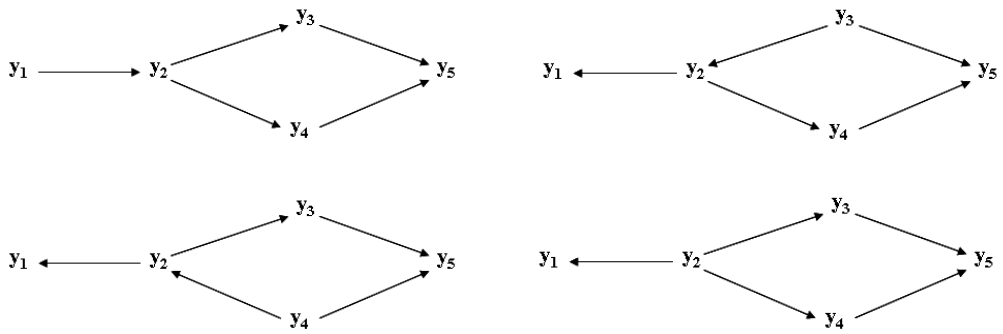


Figura 5. Estruturas causais estatisticamente equivalentes resultantes do terceiro passo do algoritmo IC.

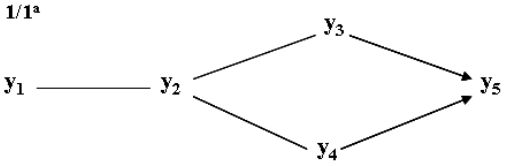
Para testar a consistência das estruturas observadas com a aplicação do algoritmo IC em uma situação real onde não se conhece a estrutura causal alvo, as decisões estatísticas sobre a existência ou não de correlação parcial entre as características poderiam ser feitas considerando o intervalo de maior probabilidade das amostras da distribuição *a posteriori* das correlações com diferentes amplitudes, por exemplo, assumindo HPD99%, HPD95%, HPD90%, HPD80%. Isso pode gerar estruturas de diferentes classes, e essas, por sua vez, serem comparadas com bases em critérios como o *Akaike Information Criterion* (AIC) ou o *Deviance Information Criterion* (DIC). Em situações que levam a distribuições das correlações parciais serem menos agudas, se as decisões estatísticas forem tomadas a partir de intervalos de maior probabilidade altos, o algoritmo IC perde a eficiência em detectar correlações parciais mais fracas. Como não há preferência entre recuperar associações não existentes e perder conexões verdadeiras entre características, seria razoável reduzir o intervalo de HPD em situações pouco informativas que geram distribuições de probabilidade das correlações parciais com forma mais plana (Valente et al., 2010). No atual estudo foi considerado para as decisões estatísticas o intervalo de HPD de 95% para todas as situações, já que o número de repetições foi grande para se analisar individualmente a sensibilidade da busca da estrutura causal frente à mudança de rigor do algoritmo e tentou-se fazer comparação justa em todas as análises.

De modo geral, como relatado também por Valente et al. (2010), o algoritmo IC foi capaz de recuperar o conjunto de estruturas no qual está inserida a estrutura causal verdadeira. Porém, a confiança que pode ser depositada na veracidade do gráfico, observado após o segundo passo do algoritmo, depende da quantidade e da qualidade das informações de parentesco e dos

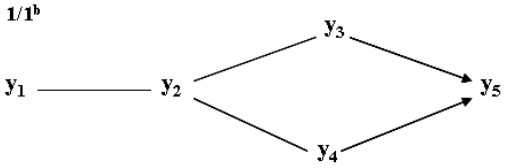
fenótipos disponíveis. Como é possível observar na Figura 6 e na Tabela 3, em situações em que o número de fenótipos e de informações de parentesco disponíveis é maior, o algoritmo IC conseguiu resgatar o gráfico acíclico semi-direcionado esperado. Apesar da dificuldade em separar o efeito da conexão genética entre os animais e o número de observações sob a busca da estrutura causal, principalmente, no caso da primeira estrutura populacional (situação 1) em que a maior fonte de informação de parentesco ocorre entre gerações com relações como pai-filho, os resultados demonstram que cerca de 1500 observações fenotípicas são suficientes para que a utilização das duas diferentes forma de matriz de relacionamento genético fossem eficientes. Valente et al. (2010) simularam uma população com estrutura próxima à situação 2 aqui apresentada em que 1800 fenótipos foram utilizados para a busca da estrutura causal idêntica à do presente estudo e isso foi suficiente para o algoritmo IC realizar as decisões corretas a partir da obtenção das amostras *a posteriori* das (co)variâncias residuais por meio do ajuste do modelo de equações estruturais totalmente recursivo incorporando a matriz *A* nas equações para retirar o efeito da correlação genética na associação entre as características. Ainda, pode-se observar na Figura 6 e Tabela 3, que para as situações 2 e 3, em que endogamia e parentesco médio são mais altos, a utilização da metodologia que incorpora a matriz *G* se mostrou menos sensível à diminuição do número de informações fenotípicas.

Ao se compararem os três grandes conjuntos de situações (1, 2 e 3), que se diferenciam, principalmente, pela magnitude da relação de parentesco entre os animais, pode-se inferir que a substituição da matriz de parentesco tradicional pela matriz de relacionamento genômico gera maior beneficio quando o nível de endogamia e coeficiente de parentesco médio da população é maior. Isso pode ser explicado pela grande diferença das matrizes nessas situações, proporcionada, provavelmente, por conta da segregação mendeliana dos alelos. Em estruturas populacionais com alto índice de endogamia e parentesco entre os animais, a esperança de compartilhamento de alelos computada pela matriz *A* tende a se afastar do real compartilhamento de alelos computado pela matriz *G*.

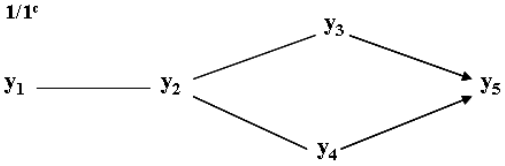
1/1[a]



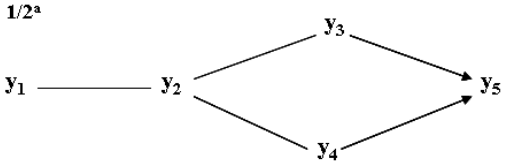
1/1[b]



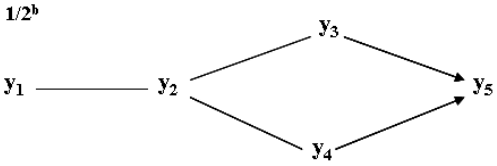
1/1[c]



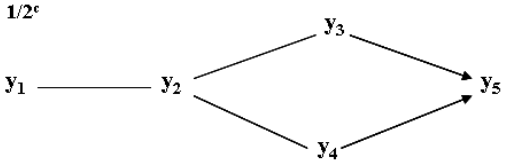
1/2[a]



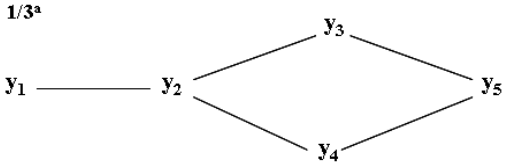
1/2[b]



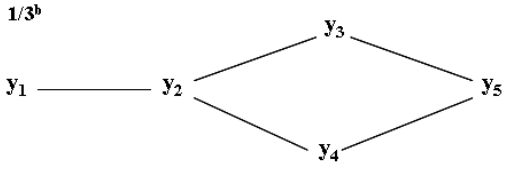
1/2[c]



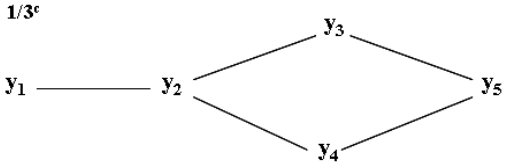
1/3[a]



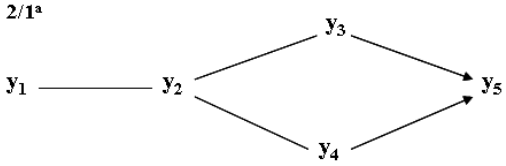
1/3[b]



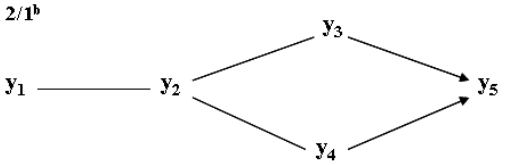
1/3[c]



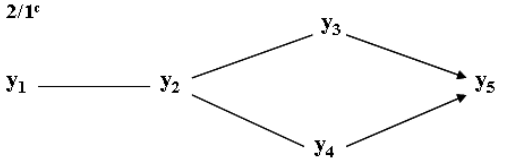
2/1[a]



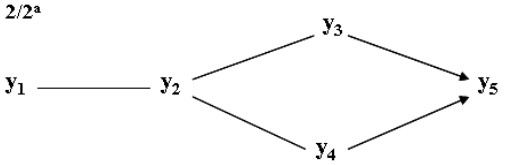
2/1[b]



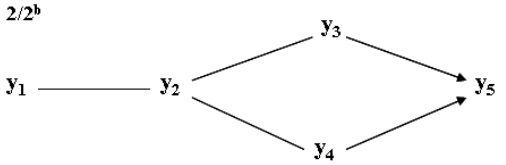
2/1[c]



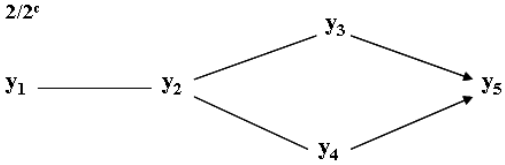
2/2[a]



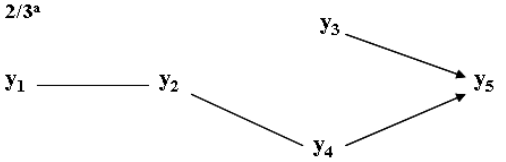
2/2[b]



2/2[c]



2/3[a]



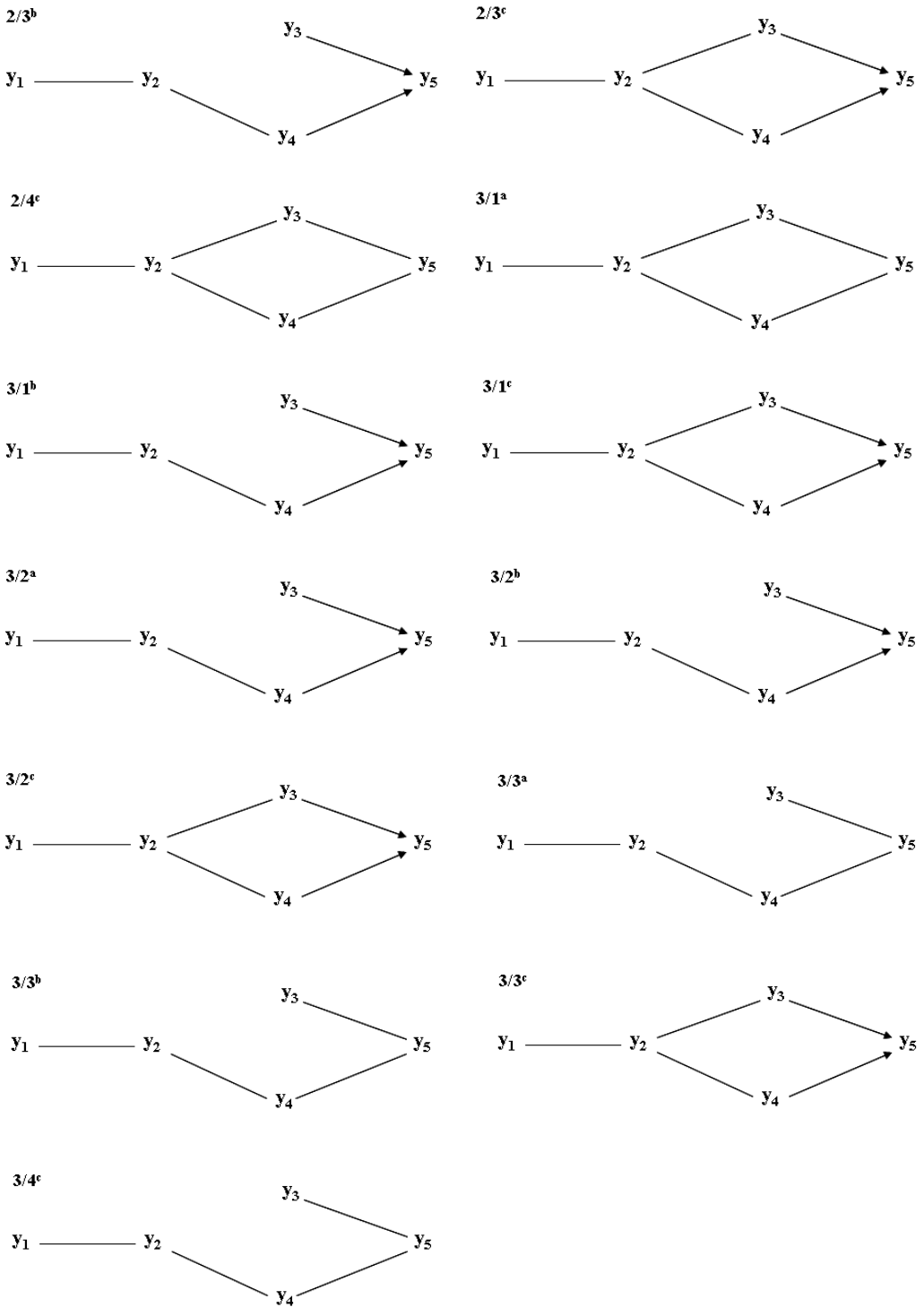


Figura 6. Estruturas causais obtidas mais observadas após o segundo passo do algoritmo IC entre as dez repetições de cada situação. Para as situações $1/4^{a}$, $1/4^{b}$, $1/4^{c}$, $2/4^{a}$, $2/4^{b}$, $3/4^{a}$ e $3/4^{b}$ não foram observadas duas ou mais estruturas causais similares dentre as dez repetições.

O número de estruturas corretas recuperadas entre as dez repetições para cada situação é maior ou no mínimo similar quando a matriz G é utilizada nas avaliações em relação à inclusão da matriz *A* nos modelos (Tabela 3). VanRaden (2008), Wolc et al. (2011) e Forni et al. (2011) relataram aumento da acurácia de predição dos valores genéticos quando a matriz de relacionamento genético genômico é empregada no modelo misto quando comparada à metodologia tradicional de avaliação genética. Uma vez que os efeitos genéticos são melhor identificados, a matriz de (co)variância residual se torna também melhor predita, o que favorece a recuperação da estrutura causal entre fenótipos a partir dos dados observacionais.

Quando a comparação é feita em termos mais específicos, como pela proximidade das estruturas obtidas e a verdadeira, usando os critérios aqui adotados, notou-se que sempre a utilização da matriz *G* trouxe resultados melhores, mesmo em situações em que não foi possível recuperar a estrutura simulada. A identificação do *unshielded collider* (3→5←4) existente pelas metodologias comparadas foi satisfatória nas situações 1/1 e 1/2. Porém, para essas situações, quando o número de informações de parentesco e fenotípicas diminuiu, a distinção não foi realizada. Para as situações 2 e 3, ou seja, em cenários em que o nível de endogamia e parentesco é maior na população, os *unshilded colliders* foram identificados com maior precisão quando utilizada a matriz de relacionamento genético genômico no modelo. A distinção do *unshielded collider* é fundamental para o correto direcionamento do fluxo causal entre características durante os passos 2 e 3 do algoritmo IC.

As ligações causais mais fracas tendem a ser mais dificilmente recuperadas pelo algoritmo IC, como demonstrado por Valente et al. (2010). No atual estudo, a primeira ligação a não ser identificada de forma correta, quando houve diminuição do número de informações, foi a presença da associação causal entre as características 2 e 3 (2→3), que apresenta coeficiente estrutural igual a 0,35. A relação causal entre essas variáveis foi mais difícil de ser observada com a diminuição do número de informações disponíveis, em especial, nas situações 2 e 3 quando utilizada a matriz *A* de parentesco. A sensibilidade à perda de informações foi menor quando a matriz G foi considerada. A associação causal mais forte, no modelo representada pela ligação entre as características 3 e 5 (2→3), com magnitude igual a 0,8, foi corretamente identificada em quase todas as situações por ambas metodologias.

Tabela 3. Resultados do segundo passo do algoritmo IC na busca pela estrutura causal simulada utilizando a matriz de parentesco tradicional e genômica para as 10 repetições realizadas

Situação	**NGR**	**UC**	**LC53**	**LC32**
1/1[a]	6	10	10	10
1/1[b]	7	10	10	10
1/1[c]	8	10	10	10
1/2[a]	9	9	10	10
1/2[b]	9	9	10	10
1/2[c]	8	8	10	10
1/3[a]	1	4	10	9
1/3[b]	1	4	10	8
1/3[c]	3	4	10	10
1/4[a]	0	1	7	5
1/4[b]	0	2	7	4
1/4[c]	0	4	9	3
2/1[a]	7	9	10	9
2/1[b]	6	9	10	9
2/1[c]	9	10	10	10
2/2[a]	4	8	10	8
2/2[b]	4	8	10	8
2/2[c]	6	9	10	9
2/3[a]	1	6	10	5
2/3[b]	1	6	10	4
2/3[c]	7	9	10	8
2/4[a]	0	3	8	1
2/4[b]	0	1	8	3
2/4[c]	1	2	10	6

3/1[a]	2	6	10	6
3/1[b]	2	6	10	5
3/1[c]	9	9	9	9
3/2[a]	0	2	10	6
3/2[b]	1	6	10	5
3/2[c]	8	8	10	10
3/3[a]	0	2	10	3
3/3[b]	0	1	9	3
3/3[c]	5	7	9	9
3/4[a]	0	1	9	4
3/4[b]	0	0	8	2
3/4[c]	2	6	10	6

NGR = Número de gráficos recuperados corretamente após o segundo passo do algoritmo IC para cada análise após o segundo passo do algoritmo IC; UCE = *Unshielded Collider* da estrutura causal identificado de forma correta; LC53 = Ligação causal entre as características 3 e 5 recuperada; LC32 = Ligação causal entre as características 2 e 3 recuperada.

Considerando os resultados observados após todas as análises pode-se dizer que a busca pela estrutura causal por meio do algoritmo IC quando se utiliza a matriz de relacionamento genético genômico para ajuste do modelo multicaracterística é mais consistente e menos sensível à variação das estruturas populacionais e mudança no número de informações disponíveis. A substituição da matriz de coeficientes de parentesco de Wright pela matriz *G* na metodologia proposta por Valente et al. (2010) causa maior confiabilidade na estrutura causal obtida como sendo a estrutura que representa a verdadeira relação causal entre características, principalmente, quando a endogamia média da população é alta e a informação de fenótipos é escassa e proveniente de animais com grande relacionamento genético.

A indicação de genotipagem de animais e utilização da matriz *G* para fins de se obter a rede de fenótipos em um determinado sistema, irá depender da necessidade de conhecimento da estrutura causal entre os fenótipos, da importância econômica da predição de mudanças nesse sistema frente à intervenções, do custo de genotipagem e da quantidade de informações já disponíveis de fenótipos e de *pedigree*. Em situações em que é importante conhecer a estrutura causal fenotípica entre as características de interesse, e a população apresenta alta taxa de endogamia ou as observações disponíveis são provenientes em sua maioria de irmãos

completos, ou ainda, quando não se tem informações de fenótipo e parentesco satisfatórias, a aplicação da metodologia seria aconselhável para maior confiabilidade do resultado obtido.

CONCLUSÕES

O algoritmo IC pode ser utilizado com objetivo de recuperar a estrutura causal entre fenótipos a partir da inclusão de matrizes de relacionamento genético entre indivíduos, construídas com base em informações genealógicas ou de marcadores moleculares, aos modelos mistos multicaracterísticas. A inserção da matriz de relacionamento genético genômico no ajuste do modelo multicaracterísticas para obtenção das amostras das (co)variâncias residuais, que serão aproveitadas na aplicação do algoritmo *Inductive Causation*, apresenta vantagens quando comparada à utilização da matriz de parentesco tradicional, principalmente, em situações específicas com estrutura populacional formada por indivíduos com alto grau de parentesco e endogâmicos, ou ainda, quando há escassez de informações fenotípicas e de parentesco. A indicação da aplicação da metodologia depende da importância do conhecimento da rede causal fenotípica entre as características, do custo de obtenção de genótipos dos indivíduos, da estrutura da população, e da quantidade e qualidade de informações de fenótipo e *pedigree* disponíveis.

REFERÊNCIAS BIBLIOGRÁFICAS

AKAIKE, H. Information theory and an extension of the maximum likelihood principle. 1978: 267–291, In: *2nd International Symposium on Information Theory*, edited by B. N. Petrov and F. Csaki. Publishing House of the Hungarian Academy of Sciences, Budapest.

BLAISDELL A. P.; AWA, K.; LEISING, K.J.; WALDMANN, M.R. Causal Reasoning in rats. *Science*. 311:1020-1022, 2006.

BOUWMAN, A.C.; VALENTE, B.D.; JANSS, L.L.G.; BOVENHUIS, H.; ROSA, G.J.M. Exploring causal networks of bovine milk fatty acids in a multivariate mixed model context. *Genetics Selction Evolution*, 46: 2, 2014.

DE LOS CAMPOS, G.; GIANOLA, D.; BOETTCHER, P.; MORONI, P. A structural equation model for describing relationships between somatic cell score and milk yield in dairy goats. *J. Anim. Sci.* 84:2934-2941, 2006a.

DE LOS CAMPOS, G.; GIANOLA, D.; HERINGSTAD, B. A structural equation model for describing relationships between somatic cell score and milk yield in first-lactation dairy cows. *J. Dairy Sci*. 89:4445-4455, 2006b.

FORNI, S.; AGUILAR, I.; MISZTAL, I. Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. *Genet. Sel. Evol.*, 43, 1, 2011.

GIANOLA D.; DE LOS CAMPOS G. A. *Course: Statistical methods for genome-enabled selection*. May 6-10, 2012.

GIANOLA, D.; DE LOS CAMPOS G. A.; HILL, W. G. et al. Additive genetic variability and the Bayesian alphabet. *Genetics,* 183:347–363, 2009.

GIANOLA, D.; SORENSEN, D. Quantitative genetic models for describing simultaneous and recursive relationships between phenotypes. *Genetics,* 167:1407-1424, 2004.

HAAVELMO, T.: The statistical implications of a system of simultaneous equations. *Econometrica*, 11:1-12, 1943.

LITZKOW, M.; LIVNY, M.; MUTKA, M. Condor – A Hunter of Idle Workstations. *Proceedings of the 8th International Conference of Distributed Computing Systems*. Madison, WI. 104-111, 1988.

PEARL, J.: *Causality: Models, Reasoning and Inference*. 2 edition. Cambridge, UK: Cambridge University Press; 2009.

R. DEVELOPMENT CORE TEAM. R: *A Language and Environment for Statistical Computing*. R Foudation for Statistical Computing, Viena; 2009. http://www.R-project.org.

ROSA, G.J.M.; VALENTE, B.D. Breeding and Genetics Symposium: Inferring causal effects from observational data in livestock. *Journal of Animal Science*, 91: 553-564, 2012.

ROSA, G.J.M.; VALENTE, B.D.; DE LOS CAMPOS, G.; et al. Inferring causal phenotype Networks using structural equation models. *Genetics Selection Evolution*, 43:6, 2011.

SPIRTES, P.; GLYMOUR, C.; SCHEINES, R. *Causation, Prediction and Search*. 2 edition. Cambridge, MA: MIT Press; 2000.

SCHWARZ, G. Estimating the dimension of a model. *Ann. Stat.* 6: 461–464, 1978.

VALENTE, B.D.; MOROTA, G.; ROSA, G.J.M.; GIANOLA, G.; WEIGEL, K. *The causal meaning of genomic predictors and how it affects the construction and comparision of genome-enabled selection models*. arXiv, 1401.1165, 2013a.

VALENTE, B.D.; ROSA, G.J.M.; DE LOS CAMPOS, G.; ET AL. Searching for recursive causal structures in multivariate genetic mixed models. *Genetics*, 185: 633-644, 2010.

VALENTE, B.D.; ROSA, G.J.; GIANOLA, D.; WU, X.L.; WEIGEL, K. Is structural equation modeling advantageous for the genetic improvement of multiple traits? *Genetics*. 194: 561-72, 2013b.

VALENTE, B.D.; ROSA, G.J.M.; SILVA, M.A.; et al. Searching for phenotypic causal networks involving complex traits: na application to European quail. *Genetics Selection Evolution*, 43: 37, 2011.

VANRADEN, P. M. Efficient methods to compute genomic predictions. *J. Dairy Sci.*, 91: 4414–4423, 2008.

VARONA, L.; SORENSEN, D.; THOMPSON, R. Analysis of litter size and average litter weight in pigs using recursive model. *Genetics*, 177:1791-1799 , 2007.

VERMA, T., PEARL, P. *Equivalence and synthesis of causal models.* In Proceedings of the 6th Conference on Uncertainty in Artificial Intelligence. Volume 6. Cambridge, MA; 1990:220-227, Reprinted in Uncertainty in Artificial Intelligence, 6: 255:268, Elsevier, Amsterdam.

WOLC, A.; STRICKER, C.; ARANGO, J. et al. Breeding value prediction for production traits in layer chickens using pedigree or genomic relationships in a reduced animal model. *Genetics Selection Evolution*, 43:5, 2011.

WRIGHT, S. Coefficients of inbreeding and relationship. *Am. Naturalist*, 56:330–338, 1922.

WRIGHT, S. Correlation and causation. *J. Agric. Res*., 201:557-585, 1921.

WU, X-L.; HERINGSTAD, B.; GIANOLA, D. Bayesian structural equation models for inferring relationships between phenotypes: a review of methodology, identifiability, and applications. *J. Anim. Breed. Genet*., 127:3-15, 2010.

WU, X-L.;,HERINGSTAD, B.; CHANG, Y.M.; et al. Inferring relationships between somatic cell score and milk yield using simultaneous and recursive models. *J. Dairy Sci*., 90:3508-3521, 2007.

WU, X-L.; HERINGSTAD, B.; GIANOLA, D. Exploration of lagged relationships between mastitis and milk yield in dairy cows using a Bayesian structural equation Gaussian-threshold model. *Genetics Selection Evololution*, 40:333-357, 2008.

Anexo 2.1. Código de simulação desenvolvido no programa R referente às 1001 primeiras gerações que não se diferiram para as três situações de estruturas populacionais propostas

```
##############################################################################
#### SIMULAÇÃO DA POPULAÇÃO
##############################################################################
rm(list=ls())

#### Função de meiose com recombinação
meiosis <- function
nloci=dim(ksome)[2]
gamete=matrix(0,1,nloci)
nxover=rpois(1,sum(dis))
origin1=rep(sample(c(TRUE,FALSE),1),nloci)
 if(nxover>0) {
  places=runif(nxover,min=0,max=sum(dis))
  for (i in 1:nxover) {
     aux=(cumsum(dis)>places[i])
     origin1[2:nloci][aux]=!origin1[2:nloci][aux]
 }
}
gamete[origin1]=ksome[1,][origin1]
gamete[!(origin1)]=ksome[2,][!(origin1)]
gamete
}

####Função de mutação
 mutation <- function(hap, rate) {
 nloci <- length(hap)
 hap.mut <- hap
 mut <- runif(nloci) < rate
 hap.mut[mut] <- abs(hap.mut[mut]-1)
 return (hap.mut)
```

```
}

#### Parâmetros do genoma
size.chrm <- 1
nChrm <- 30
nQTL <- 10
markerDens <- 70
nMarker <- (nQTL+1)*markerDens
nLoci <- nQTL + nMarker
nInterval <- nQTL + nMarker - 1
distance <- rep(size.chrm/nInterval, nInterval)

QTLpos <-  NULL
for (i in 1:nChrm) {
      QTLpos<-c(QTLpos,seq((nQTL+nMarker)*(i-1)+(markerDens+1),    by=markerDens+1,
length=nQTL))
}
Markerpos <- (1:((nQTL+nMarker)*nChrm))[-QTLpos]

mut.QTL <- 2.5e-3 # mutation rate at QTLs
mut.marker <- 2.5e-3 # mutation rate at markers

#### Gerações 1 a 1000
Ne <- 100
nMale <- Ne/2
nFemale <- Ne/2

Male <- matrix(0, nrow=2*nMale, ncol=(nQTL+nMarker)*nChrm)
Female <- matrix(0, nrow=2*nFemale, ncol=(nQTL+nMarker)*nChrm)
chrm <- rep(0, (nMarker+nQTL)*nChrm)
for (i in 1:(2*nMale)) {
        Male[i,] <- chrm
        Female[i,] <- chrm
```

```
}

#### Começando a evolução da população
nGen <- 1000

individual.new<- matrix(0, nrow=2*(nMale+nFemale), ncol=(nQTL+nMarker)*nChrm)

for (gen in 1:nGen) {
 for (i in 1:Ne) {
  sire <- sample(1:nMale,1)
  dam <- sample(1:nFemale,1)
  stmp <- (sire-1)*2 + 1
   for (chr in 1:nChrm) {individual.new[(i-1)*2+1,((chr-1)*nLoci+1):(chr*nLoci)] <-
    meiosis(Male[stmp:(stmp+1),((chr-1)*nLoci+1):(chr*nLoci)], distance)}
   individual.new[(i-1)*2+1,QTLpos]<-mutation(individual.new[(i-1)*2+1,QTLpos],
   mut.QTL)
   individual.new[(i-1)*2+1,Markerpos]<-mutation(individual.new[(i-
   1)*2+1,Markerpos], mut.marker)
   dtmp <- (dam-1)*2 + 1 # get row number in matrix "Female"
   for (chr in 1:nChrm) {individual.new[(i-1)*2+2,((chr-1)*nLoci+1):(chr*nLoci)] <-
   meiosis(Female[dtmp:(dtmp+1),((chr-1)*nLoci+1):(chr*nLoci)], distance)}
   individual.new[(i-1)*2+2,QTLpos]<-mutation(individual.new[(i-1)*2+2,QTLpos],
   mut.QTL)
   individual.new[(i-1)*2+2,Markerpos]<-mutation(individual.new[(i-
   1)*2+2,Markerpos], mut.marker)
 }

 Male <- individual.new[1:(2*nMale),]
 Female <- individual.new[(2*nMale+1):(2*Ne),]
}

#### Checando a porcentagem de alelos segregantes
locus_af <- apply(rbind(Male, Female), 2,mean)
```

```
seg_locus <- which(locus_af > 0 &  locus_af < 1)
cat("% seg. loci in the last historical gen:", length(seg_locus)/length(locus_af), "\n")

#### Gerando g1 com expansão populacional
Animal <- NULL
Sire <- NULL
Dam <- NULL

n_G1 <- 500
G <- matrix(0, nrow=2*n_G1, ncol=(nQTL+nMarker)*nChrm) #matrix containing individuals of G1
rand <- sample(1:n_G1)
i <- 1
nDams <- 10

for (sire in 1:nMale) {
  dam <- sample(1:nFemale, nDams, replace=F)
  stmp <- (sire-1)*2 + 1

  for (mating in 1:nDams) {
    r <- rand[i]

    for (chr in 1:nChrm) {G[(r-1)*2+1,((chr-1)*nLoci+1):(chr*nLoci)]<-
    meiosis(Male[stmp:(stmp+1),((chr-1)*nLoci+1):(chr*nLoci)], distance)}
    dtmp <- (dam[mating]-1)*2 + 1 # find dam's position in matrix "Female"
    for (chr in 1:nChrm) {G[(r-1)*2+2,((chr-1)*nLoci+1):(chr*nLoci)] <-
      meiosis(Female[dtmp:(dtmp+1),((chr-1)*nLoci+1):(chr*nLoci)], distance)}
    i <- i+1
  }
}

Male_G <-G[1:n_G1,]
Female_G <- G[(n_G1+1):(2*n_G1),]
```

```
Xhap <- rbind(Male_G,Female_G)
Xhap1 <- Xhap
currentAnim <- 1:n_G1
currentSire <- rep(-9, n_G1)
currentDam <- rep(-9, n_G1)
Animal  <- c(Animal, currentAnim)
Sire <- c(Sire, currentSire)
Dam <- c(Dam, currentDam)
```

Anexo 2.2. Código de simulação desenvolvido no programa R para obtenção dos valores genéticos e fenótipos dos indivíduos com base na estrutura causal e nas herdabilidades propostas

```
#############################################################################
#### SIMULAÇÃO DOS FENÓTIPOS
#############################################################################
#### Conversão dos dados de haplótipos para genótipos para cálculo dos valores genéticos e fenotípicos
Xg1 <- Xhap1
n_X1 <- dim(Xg1)[1]/2
 for (i in seq(1,(2*n_X1-1), by=2)) {
   Xg1[i,] <- Xg1[i,] + Xg1[i+1,]
}

Xg1 <- Xg1[seq(1,(2*n_X1-1), by=2), ]
ImQTL1 <- Xg1[, QTLpos]
MQTL1 <- ImQTL1 - 1
Xg <- Xhap3
n_X <- dim(Xg)[1]/2
  for (i in seq(1,(2*n_X-1), by=2)) {
    Xg[i,] <- Xg[i,] + Xg[i+1,]
}
```

```
Xg <- Xg[seq(1,(2*n_X-1), by=2), ]

ImQTL <- Xg[, QTLpos[segQTL]]
MQTL <- ImQTL-1

write.table(file="Xg.dat", Xg, col.names=F, row.names=F, quote=F, sep= " ")
write.table(file="ImQTL.dat", ImQTL, col.names=F, row.names=F, quote=F, sep= " ")
write.table(file="MQTL.dat", MQTL, col.names=F, row.names=F, quote=F, sep= " ")

#### Parâmetros das características
traits<- 5
h2 <- c(0.333,0.333,0.333,0.333,0.333)
h2 <- as.vector(h2)
h2 <- t(h2)
MCov <- cbind(c(100.00, 47.373, 20.283, -38.839, 9.773),
            c(47.373, 100.000, 31.993, -46.357, -49.791),
            c(20.283, 31.993, 100.000, 60.625, -14.557),
            c(-38.839, -46.357, 60.625, 100.000, 6.490),
            c(9.773, -49.791, -14.557, 6.490, 100.000))

QTL.eff <- matrix(rnorm(traits*length(QTLpos)), ncol=traits, nrow=length(QTLpos))

PQ1<-NULL
 for (i in 1:(length(QTLpos)))
 {
  pq1<-(sum(ImQTL1[,i])/(nrow(ImQTL1)*2))*(1-
  (sum(ImQTL1[,i])/(nrow(ImQTL1)*2)))
  PQ1 <- cbind(PQ1,pq1)
}

PQ1mean<- sum(PQ1)/length(PQ1)
 for (i in 1:(length(QTLpos)))
 {
```

```
  QTL.eff[i,]<- QTL.eff[i,]%*%chol(MCov/(2*PQ1mean*length(QTLpos)))
}

pQTL1 <- apply(Xhap1[,QTLpos], 2, mean)
pQTL1 <- as.vector(pQTL1)
pQTL1 <- t(pQTL1)
qQTL1 <- 1-pQTL1

write.table(file="ZQTL.dat", ZQTL, col.names=F, row.names=F, quote=F, sep= " ")
add.eff <- MQTL%*%QTL.eff[segQTL,]
cov.gen <- var(add.eff)
var.gen <- diag(cov.gen)
var.gen <- as.vector(var.gen)
var.gen <- t(var.gen)

R <- matrix (nrow=nrow(add.eff), ncol=length(var.gen))
 for (j in 1:length(var.gen))
 {
  R[,j] <- rnorm(nrow(add.eff), 0, sqrt(var.gen[,j]/h2[,j]-var.gen[,j]))
}

Ynf = add.eff + R
Y1 =   80 + Ynf[,1]
Y2 =   60 + 0.5*Y1 + Ynf[,2]
Y3 =   70 + 0.35*Y2 + Ynf[,3]
Y4 =  140 - 0.5*Y2 + Ynf[,4]
Y5 =   60 + 0.8*Y3 -0.4*Y4 + Ynf[,5]
Y = cbind(Y1,Y2,Y3,Y4,Y5)

write.table(file="add.eff.dat", add.eff, col.names=F, row.names=F, quote=F, sep= " ")
write.table(file="covgen.dat", cov.gen, col.names=F, row.names=F, quote=F, sep= " ")
write.table(file="R.dat", R, col.names=F, row.names=F, quote=F, sep= " ")
write.table(file="Y.dat", Y, col.names=F, row.names=F, quote=F, sep= " ")
```