Seleção genômica ampla (GWS) e maximização da eficiência do melhoramento genético

Marcos Deon Vilela de Resende[1], Paulo Sávio Lopes[2], Rogério Luíz da Silva[3] e Ismael Eleotério Pires[3]

[1]Embrapa Florestas, Estrada da Ribeira Km 111, Caixa Postal 319, CEP 83411-000, Colombo-PR. Email: deon@cnpf.embrapa.br; [2]Universidade Federal de Viçosa, Centro de Ciências Agrárias, Departamento de Zootecnia, Campus Universitário CEP 36571-000, Viçosa- MG. Email: plopes@pq.cnpq.br; [3]Universidade Federal de Viçosa, Departamento de Engenharia Florestal, Campos Universitário, Avenida Peter Henry Rolfs, s/n, Campus Universitário, CEP 36570-000, Viçosa-MG. Email: rogerio_ufv@yahoo.com.br, iepires@ufv.br

Resumo - A seleção genética tem sido praticada pelo procedimento BLUP, usando dados fenotípicos avaliados a campo. Uma primeira proposição realizada para aumentar a eficiência desse procedimento, baseado em dados fenotípicos, foi a seleção auxiliada por marcadores (MAS) moleculares, a qual usa simultaneamente dados fenotípicos e moleculares. Posteriormente, foi proposto um novo método de seleção denominado seleção genômica ampla (*genome wide selection* – GWS), o qual apresenta alta acurácia seletiva para a seleção, baseada exclusivamente em marcadores, após terem seus efeitos genéticos estimados a partir de dados fenotípicos em uma amostra da população de seleção. A GWS é excelente para caracteres de baixa herdabilidade, ao contrário da MAS, que não é útil para caracteres de baixa herdabilidade. O presente trabalho tem como objetivos apresentar a metodologia GWS e simular um caso de aplicação da mesma, visando enfatizar as suas vantagens sobre a MAS. Objetiva também demonstrar a relação entre o BLUP tradicional e o BLUP genômico associado à GWS. Adicionalmente, discute aspectos referentes ao tamanho amostral adequado para estimação dos efeitos genotípicos dos marcadores. Os resultados revelam que a GWS poderá ter grande utilidade ao melhoramento genético. No entanto, é preciso adquirir experiência prática com a GWS, visando inferir sobre sua efetividade.

Termos para indexação: Seleção genômica, análise de desequilíbrio de ligação, mapeamento fino, marcadores moleculares.

Genome wide selection (GWS) and maximization of the genetic improvement efficiency

Abstract - Genetic selection has been practiced by the best linear unbiased prediction (BLUP) method using phenotypic records. A first proposal for enhancement of the efficiency of this procedure was the marker assisted selection (MAS). Later, another method called genome wide selection – GWS was reported, which presents high accuracy for the selection based exclusively on markers, after predicting their genetic effects from phenotypic data in a sample of the population of selection. GWS is excellent for low heritable traits, while MAS is not. This paper presents the GWS methodology and simulates a case of its application, aiming at emphasizing its advantages over MAS. The relation between traditional BLUP and genomic BLUP is also detailed as well as the sample size required for precise estimation of the genetic values of the markers. Results revealed that the GWS can be worthy for genetic improvement. Practical experience is much needed to infer about its effectiveness.

Index terms: Genomic selection, linkage disequilibrium analysis, fine mapping, genetic markers.

Introdução

A eficiência do melhoramento genético depende basicamente de duas ações do geneticista: a criação e a identificação de genótipos superiores. Em ambas as ações, a seleção desempenha papel fundamental, na definição dos cruzamentos a serem realizados, visando à criação de novos genótipos e na indicação dos indivíduos superiores a serem usados comercialmente. A seleção genética tem sido praticada com base em dados fenotípicos avaliados a campo. Uma primeira proposição realizada para aumentar a eficiência desse procedimento baseado em dados fenotípicos foi descrita por Lande e Thompson (1990), por meio da seleção auxiliada por marcadores (MAS) moleculares. A MAS utiliza simultaneamente dados fenotípicos e dados de

marcadores moleculares em ligação gênica próxima com alguns locos controladores de características quantitativas (QTL). Em geral, os dados de marcadores são utilizados como covariáveis na explicação dos valores fenotípicos dos indivíduos em avaliação ou como efeitos aleatórios incorporados no modelo para o fenótipo. Esses marcadores são eleitos ou não como determinantes dos efeitos de QTLs após modelagem estatística associada a erros do tipo II (probabilidade de aceitar uma hipótese falsa, ou seja, tomar como verdadeira uma hipótese falsa de ausência de efeitos).

A seleção baseada na MAS apresenta as seguintes características: (i) requer o estabelecimento (análise de ligação) de associações marcadores-QTLs para cada família em avaliação, ou seja, essas associações apresentam utilidade para seleção apenas dentro de cada família mapeada em espécies alógamas; (ii) para ser útil, precisa explicar grande parte da variação genética de uma característica quantitativa, que é governada por muitos locos de pequenos efeitos. Isto não tem sido observado na prática, exatamente em função da natureza poligênica e alta influência ambiental nos caracteres quantitativos, fato que conduz à detecção apenas de um pequeno número de QTLs de grandes efeitos, os quais não explicam suficientemente toda a variação genética; (iii) só apresenta superioridade considerável em relação à seleção baseada em dados fenotípicos, quando o tamanho de família avaliado e genotipado é muito grande (da ordem de 500 ou mais). Em função desses aspectos, a implementação da MAS tem sido limitada e os ganhos em eficiência muito reduzidos (DEKKERS, 2004).

O grande atrativo da genética molecular em benefício do melhoramento genético aplicado é a utilização direta das informações de DNA na seleção, de forma a permitir alta eficiência seletiva, grande rapidez na obtenção de ganhos genéticos com a seleção e baixo custo, em comparação com a tradicional seleção baseada em dados fenotípicos. Visando a esses objetivos, Meuwissen et al. (2001) propuseram um novo método de seleção denominado seleção genômica (GS) ou seleção genômica ampla (*genome wide selection* – GWS), a qual pode ser aplicada em todas as famílias em avaliação nos programas de melhoramento genético de espécies alógamas, apresenta alta acurácia seletiva para a seleção baseada exclusivamente em marcadores (após terem seus efeitos genéticos estimados a partir de dados fenotípicos em uma amostra da população de seleção) e não exige prévio conhecimento das posições (mapa) dos QTLs, não estando sujeita aos erros tipo II associados à seleção de marcadores ligados a QTLs.

Esse método permaneceu discreto por cerca de cinco anos, devido ao fato dos marcadores moleculares disponíveis à época serem caros e restritos. Recentemente, com o desenvolvimento e baixo custo dos marcadores tipo SNP (*single nucleotide polymorphism*), o método tornou-se muito atrativo e geneticistas e melhoristas renomados e adeptos de métodos tradicionais têm demonstrado e confirmado a superioridade e exeqüibilidade prática do método em benefício do melhoramento animal (SCHAEFFER, 2006; KOLBEHDARI et al. 2007; MEUWISSEN, 2007; GODDARD; HAYES, 2007; LONG et al. 2007; LEGARRA; MISZTAL, 2008) e vegetal (BERNARDO; YU, 2007). Esses trabalhos mostraram, definitivamente, que a seleção genômica terá grande utilidade no melhoramento genético, via métodos do tipo BLUP/GWS, que equivalem ao procedimento BLUP (melhor predição linear não viciada) aplicado sobre dados moleculares e permitem a predição de valores genéticos genômicos. A GWS é excelente para caracteres de baixa herdabilidade, ao contrário da MAS, que não é útil para caracteres de baixa herdabilidade (MUIR, 2007).

A MAS baseia-se na detecção, mapeamento e uso de QTLs de grande efeito na seleção. Ou seja, enfatiza a determinação do número, posição e efeitos dos QTLs marcados. A GWS é definida como a seleção simultânea para centenas ou milhares de marcadores, os quais cobrem o genoma de uma maneira densa, de forma que todos os genes de um caráter quantitativo estejam em desequilíbrio de ligação com pelo menos uma parte dos marcadores. Esses marcadores em desequilíbrio de ligação com os QTL´s, tanto de grandes quanto de pequenos efeitos, explicarão quase a totalidade da variação genética de um caráter quantitativo. O número de SNP´s é de tal magnitude que a probabilidade de se encontrar um QTL em desequilíbrio de ligação com pelo menos um marcador é muito alta. Este aspecto é importante uma vez que somente os marcadores em desequilíbrio de ligação com os QTL´s serão úteis na determinação dos fenótipos e na explicação da variação genética. Os efeitos dos marcadores são estimados em uma amostra de indivíduos pertencentes a várias famílias. Assim, o impacto de determinadas famílias específicas (com específicos padrões de desequilíbrio de ligação) nas estimativas dos efeitos dos marcadores será

minimizado. É importante enfatizar que os marcadores terão seus efeitos genéticos estimados a partir de uma amostra de pelo menos 1.000 indivíduos genotipados e fenotipados, ou seja, com base em pelo menos 1.000 repetições experimentais. Assim, embora a herdabilidade de cada marcador efetivo (aquele que identifica um dos poligenes com precisão) seja muito baixa, com 1.000 repetições essa herdabilidade se torna alta. Em outras palavras, o efeito de ambiente será minimizado por meio do uso de um número de repetições muito alto. Essa é a mesma filosofia da avaliação e seleção de características quantitativas com base em fenótipos em experimentos de campo.

A GWS é ampla porque atua em todo o genoma, capturando todos os genes que afetam um caráter quantitativo. E isso sem a necessidade prévia de identificar os marcadores com efeitos significativos e de mapear QTLs, como no caso da MAS. Valores genéticos genômicos associados a cada marcador ou alelo são usados para fornecer o valor genético genômico global de cada indivíduo. Há uma diferença básica na predição de valores genéticos tradicionais e na predição de valores genéticos genômicos. Nos primeiros, informações fenotípicas são utilizadas visando inferências sobre os efeitos dos genótipos dos indivíduos e, nos últimos, informações genotípicas (genótipos para os alelos marcadores) são usadas visando às inferências sobre os valores fenotípicos futuros (ou valores genéticos genômicos preditos) dos indivíduos. Em outras palavras, os métodos tradicionais usam o fenótipo para inferir sobre o efeito do genótipo e a GWS usa o genótipo, com efeito genético pré-estimado em uma amostra da população, para inferir sobre o fenótipo a ser expresso nos candidatos à seleção.

Os efeitos dos marcadores não serão necessariamente os mesmos em diferentes estudos e ambientes. Na GWS, os efeitos genéticos dos marcadores são estimados e usados na seleção para cada população de melhoramento e em um determinado ambiente. Modelos de estimação, incluindo a interação genótipos x ambientes, podem também ser usados, visando verificar a possibilidade de se obter estimativas válidas para um conjunto de ambientes. Mas isso dependerá da magnitude da interação envolvendo os vários ambientes.

A GWS pode basear-se no uso de: (i) apenas dos marcadores; (ii) de haplótipos ou intervalos definidos por dois marcadores; (iii) haplótipos definidos por mais de dois marcadores, incluindo a covariância entre haplótipos devida à ligação. Segundo Calus et al. (2008), para caracteres de baixa herdabilidade (10 %) não existem diferenças significativas entre essas três abordagens. Solberg et al. (2006) mostraram que é possível praticar a GWS eficientemente com o uso apenas dos marcadores, ou seja, com a predição direta dos efeitos dos marcadores. Relatam também que isso é vantajoso porque não há necessidade de estimar as fases de ligação entre os marcadores, as quais são estimadas com algum erro. Não apenas marcadores SNPs podem ser usados na GWS. Marcadores microssatélites também se prestam a esse fim. Solberg et al. (2006) relatam que o uso de SNPs requer quatro a cinco vezes maior densidade de marcadores do que o uso de microssatélites. Isto se deve à natureza bi-alélica (bi-nucleotídica) dos SNPs e multi-alélica dos microssatélites.

O presente trabalho tem como objetivos apresentar a metodologia GWS e simular um caso de aplicação da mesma, visando enfatizar as suas vantagens sobre a MAS. Objetiva também demonstrar a relação entre o BLUP tradicional e o BLUP genômico associado à GWS. Adicionalmente discute aspectos referentes ao tamanho amostral adequado para estimação dos efeitos genotípicos dos marcadores, usando simulação.

Material e Métodos

Fundamentos da GWS

A GWS fundamenta-se nos marcadores genéticos moleculares do tipo SNP (polimorfismo de um único nucleotídeo), o qual se baseia na detecção de polimorfismo resultante da alteração de um único par de base no genoma. E para que uma variação seja considerada SNP, essa deve ocorrer em pelo menos 1 % da população. Os SNPs são a forma mais abundante de variação do DNA em genomas e são preferidos em relação a outros marcadores genéticos devido à sua baixa taxa de mutação e facilidade de genotipagem. Milhares de SNPs podem ser usados para cobrir o genoma de um organismo com marcadores que não estão a mais de 1 cM um do outro no genoma inteiro.

A GWS atua mais proximamente aos QTNs (nucleotídeos de características quantitativas) ou sobre marcadores fortemente ligados a esses. Os QTNs são polimorfismos funcionais, causadores diretos da variação quantitativa observada. A análise de SNP´s permite a detecção de polimorfismos funcionais ou polimorfismos

em forte desequilíbrio de ligação com os QTNs. Tecnologias para genotipagem de milhares de SNPs em microarranjos estão disponíveis atualmente. Microarranjos são sistemas de arranjos de DNA que utilizam lâminas de vidro e sondas fluorescentes e permitem depositar milhares de seqüências de DNA. Nessa técnica são utilizados nucleotídeos marcados capazes de emitir fluorescência ao invés de radioatividade.

O desenvolvimento teórico da GWS coincide com a tecnologia SNP, a qual é acurada e relativamente barata. A GWS usa associações de um grande número de marcadores SNPs em todo o genoma com os fenótipos, capitalizando no desequilíbrio de ligação entre os marcadores e QTLs proximamente ligados, sem uma prévia escolha de marcadores com base nas significâncias de suas associações com o fenótipo. Predições são então obtidas para os efeitos dos haplótipos marcadores ou dos alelos em cada marcador. Essas predições derivadas de dados fenotípicos e de genótipos SNPs em alta densidade em uma geração são então usadas para obtenção dos valores genéticos genômicos (VGG) dos indivíduos de qualquer geração subseqüente, tendo por base os seus próprios genótipos marcadores. Os hapótipos são definidos como intervalos resultantes de combinações de dois alelos marcadores vizinhos. A seleção genômica baseada simultaneamente em um grande número de marcadores contrasta com a MAS, que é baseada em um número limitado de marcadores ou genes. Os marcadores moleculares do tipo microssatélites podem também ser usados na GWS. Tais marcadores são eficientes por serem co-dominantes, multi-alélicos, abundantes e apresentarem alta transferibilidade entre indivíduos e espécies.

Se os marcadores estão ligados aos QTLs e não aos nucleotídeos causadores da variação alélica responsável pela variação no caráter quantitativo (QTNs), variantes das fases de ligação fazem com que os marcadores sejam incorretos em algumas famílias ou populações. Com o uso dos SNPs, existe a vantagem de que os mesmos tendem a estar intimamente ligados aos próprios QTNs.

Quando o desequilíbrio de ligação entre marcadores não é completo, as freqüências alélicas conjuntas envolvendo dois locos podem mudar substancialmente através das gerações, conduzindo a mudanças nos haplótipos. Também, se houver dominância e epistasia em magnitudes consideráveis, os efeitos dos marcadores necessitarão ser re-estimados para manter a acurácia da GWS em várias gerações (DEKKERS, 2007). O desequilíbrio de ligação ou desequilíbrio de fase gamética é uma medida da dependência ou não entre alelos de dois ou mais locos. Em um grupo de indivíduos, se dois alelos são encontrados juntos com freqüência maior do que aquela esperada com base no produto de suas freqüências, infere-se que tais alelos estão em desequilíbrio de ligação. Valores de desequilíbrio de ligação próximos de zero indicam equilíbrio ou independência entre os alelos de diferentes genes e valores próximos de um indicam desequilíbrio ou dependência (ligação) entre alelos de diferentes genes.

Com desequilíbrio de ligação completo e ausência de dominância e epistasia, os VGG são caracteres genéticos, com herdabilidade 1 e cujos efeitos permanecem constantes através das gerações em um mesmo ambiente. Embora sejam estimativas, esses caracteres genéticos podem ser vistos como herdados de maneira poligênica, porém sem efeitos ambientais. Os VGG equivalem à soma dos valores genéticos de cada alelo e de cada gene (loco) do caráter quantitativo.

Procedimento REML/BLUP/GWS

A estimação dos VGGs usa um conjunto de dados de referência que inclui indivíduos com ambos os conhecidos, os genótipos (marcadores) e os fenótipos. Os valores genéticos estimados dos haplótipos ou marcadores em um grande número de supostos caracteres quantitativos são usados para a predição dos valores genéticos genômicos de indivíduos jovens canditatos à seleção e que foram genotipados para os marcadores, mas não possuem informação fenotípica. Se toda variação genética puder ser explicada pelos haplótipos ou marcadores, não há necessidade de inclusão no modelo de predição, do efeito poligênico para levar em consideração a variação genética não explicada (variação genética residual). Na prática, se não há uma cobertura completa (mapa denso de marcadores) do genoma com SNPs, a inclusão do efeito poligênico pode tornar-se necessária.

A estimação dos valores genéticos genômicos para haplótipos marcadores individuais ou alelos individuais do QTL baseia-se em um número relativamente grande de haplótipos e outro relativamente pequeno de indivíduos. Marcadores SNPs circundando cada região genômica de 1 cM são combinados em um haplótipo marcador. Com um mapa de marcadores denso, alguns

marcadores estarão muito próximos dos QTLs e provavelmente em desequilíbrio de ligação com eles. Assim, alguns alelos marcadores estarão correlacionados com efeitos positivos no caráter quantitativo através de todas as famílias e poderão ser usados na seleção sem a necessidade de estabelecer a fase de ligação em cada família. Segmentos cromossômicos que contém os mesmos haplótipos marcadores raros apresentam alta probabilidade de identidade por descendência e então carregam os mesmos alelos do QTL. A precisão do mapeamento de QTL pelos métodos tradicionais de análise de ligação é pouco melhorada pelo uso de mapas de marcadores densos. Mas, pela abordagem da GWS, os efeitos nos QTLs, de pequenos segmentos de cromossomo definidos pelos haplótipos dos alelos marcadores que eles carregam, são estimados com alta precisão, e os referidos mapas densos são muito úteis (MEUWISSEN et al., 2001).

Para um genoma de 3 mil cM, apenas 3.001 marcadores a intervalos de 1 cM seriam necessários. No entanto, tais marcadores necessitam ser informativos e um painel com 10 mil marcadores aumentaria as chances de sucesso. Cada par contíguo de marcadores define um haplótipo ou intervalo. Existem apenas dois alelos para cada marcador, pois os SNPs têm diferenças em um único par de bases. Dessa forma, para cada par de marcadores, existem quatro haplótipos possíveis. A freqüência de cada haplótipo depende da freqüência dos alelos em cada marcador e a distância entre marcadores depende dos eventos de recombinação. Assim, um número suficiente de indivíduos deve ser genotipado de forma que todos os haplótipos estejam representados nos indivíduos com avaliações fenotípicas (SCHAEFFER, 2006).

A estimação dos efeitos do elevado número de haplótipos a partir de um número limitado de dados conduz ao problema da estimação por quadrados mínimos com insuficiente número de graus de liberdade para ajustar todos esses efeitos simultaneamente. O método BLUP, por outro lado, permite ajustar todos os efeitos alélicos simultaneamente, mesmo quando existem mais efeitos a serem preditos do que o número total de observações fenotípicas.

Os efeitos estimados dos haplótipos são assumidos como estimativas válidas para toda a população e não para apenas um grupo de indivíduos. Dessa forma, VGG podem ser estimados para quaisquer indivíduos da população, desde que os mesmos sejam genotipados e os haplótipos marcadores sejam determinados. Assim, cada indivíduo pode ter uma estimativa de VGG desde o momento em que é gerado.

Os efeitos de cada intervalo em um caráter em um dado ambiente podem ser estimados para todos os intervalos simultaneamente em um modelo linear misto em que os efeitos de intervalo são tratados como aleatórios. Os genótipos marcadores dos indivíduos podem ser usados para predição de qualquer caráter, mas as estimativas dos efeitos dos intervalos ou haplótipos serão diferentes para cada caráter. Os intervalos com maiores efeitos em cada caráter conterão um ou mais QTLs. A maioria dos intervalos, no entanto, apresentará efeitos relativamente menores, refletindo o que acontece no modelo infinitesimal (muitos genes de pequenos efeitos associados ao caráter quantitativo).

Para uso dos SNPs na GWS, inicialmente, devem ser identificados aqueles informativos e, posteriormente, um *software* deve ser usado para construir haplótipos a partir dos genótipos SNP. De posse dos haplótipos, as predições de seus efeitos podem ser feitas por meio de *softwares* específicos de genética quantitativa e estatística.

O seguinte modelo linear misto geral é usado para estimar os efeitos de haplótipos:

$y = Xb + Zh + e$,

em que: y é o vetor de observações fenotípicas, b é o vetor de efeitos fixos, h é o vetor dos efeitos aleatórios de haplótipos (intervalos) e *e* refere-se ao vetor de resíduos aleatórios. X e Z são as matrizes de incidência para *b* e *h*.

A dimensão de h é igual ao número de intervalos multiplicado por 4 (número de haplótipos possíveis para cada intervalo). A matriz de incidência Z contém os valores 0, 1 e 2 para o número de alelos (do suposto QTL) ou haplótipos do tipo h_i no indivíduo diplóide j.

A estrutura de médias e variâncias é definida como:

$$h \sim N(0,\ G) \qquad E(y) = Xb$$

$$e \sim N(0,\ R = I\sigma_e^2) \qquad Var(y) = V = ZGZ' + R$$

$$G = \sum_i^n I_h \sigma_i^2$$

em que: I_h é de ordem 4 e σ_i^2 é a variância dos efeitos dos haplótipos no i-ésimo intervalo e n é o número total de intervalos.

As equações de modelo misto para a predição de h, via o método BLUP/GWS, equivalem a:

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + I\dfrac{\sigma_e^2}{(\sigma_g^2/n)} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{h} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

em que σ_g^2 refere-se à variância genética total do caráter e σ_e^2 é a variância residual. O valor genético genômico global do indivíduo j é dado por:

$VGG = \hat{y}_j = \sum_i Z_i \hat{h}_i$, em que Z_i equivale a 0, 1 ou 2.

As equações de predição apresentadas anteriormente assumem, a priori, que todos os locos explicam iguais quantidades da variação genética. Assim, a variação genética explicada por cada loco é dada por σ_g^2 / n, em que σ_g^2 é a variação genética total e n é o número de intervalos ou haplótipos. Essa estratégia foi adotada por Meuwissen et al. (2001), Muir (2007), Bernardo e Yu (2007) e Kolbehdari et al. (2007). Bernardo e Yu (2007) relatam que essa suposição de iguais variâncias, por loco, não conduz a perdas significativas na acurácia da GWS. A variação genotípica σ_g^2 pode ser estimada por REML sobre os dados fenotípicos da maneira tradicional ou pela própria variação entre os haplótipos ou variância dos segmentos cromossômicos de QTL.

Os efeitos do vetor h são ajustados como covariáveis aleatórias associadas às observações fenotípicas. Muir (2007) denomina o método BLUP/GWS como regressão de cumeeira ou "*ridge regression*" (RR). No caso, o parâmetro de regressão é função de $\sigma_e^2 / (\sigma_g^2 / n)$.

O procedimento BLUP/GWS é similar ao BLUP tradicional. Porém, na predição dos referidos efeitos aleatórios, não há necessidade de uso da matriz de parentesco (SCHAEFFER, 2006). A matriz de parentesco baseada em *pedigree*, usada no BLUP tradicional, é substituída por uma matriz de parentesco estimada pelos marcadores. Essa matriz de parentesco é a própria matriz Z'Z presente nas equações de modelo misto, em que Z é a matriz de incidência para os efeitos de marcadores. Esse procedimento é superior ao uso do *pedigree*, pois efetivamente captura a matriz de parentesco realizada e não uma matriz de parentesco médio associada ao *pedigree*.

Na situação em que os marcadores não explicam toda a variação genética, o modelo pode ser estendido para englobar o efeito poligênico residual (variação genética não explicada pelos marcadores). Esse modelo estendido é da forma:

y = Xb + Zh + Wg* + e,

em que g* é o vetor dos efeitos poligênicos residuais (aleatórios) e W é a matriz de incidência para g*.

Com o uso de mapa denso de marcadores, a inclusão dos efeitos poligênicos g* não aumenta a acurácia da GWS (CALUS; VEERKAMP, 2007). No entanto, para capitalizar o ganho genético no longo prazo, a inclusão desses efeitos é recomendada (MUIR, 2007). No longo prazo, o BLUP tradicional obtém informação no genoma inteiro em cada geração. A GWS sem o efeito poligênico seleciona de forma muito acurada para a mesma parte do genoma em cada geração. Uma forma de aliviar esse problema é por meio da re-estimação dos efeitos de marcadores, visando à exploração de novas associações marcadores-QTL.

Os modelos apresentados devem ser estendidos para a incorporação de outros fatores de efeitos aleatórios, visando contemplar ajustes para efeitos ambientais tais quais efeitos de blocos incompletos e, também, para a incorporação de covariáveis ambientais de efeitos fixos. Podem ser estendidos também para incorporar efeitos de dominância, tais quais efeitos de capacidade específica de combinação. Os modelos apresentados assumem genes de efeitos aditivos e, portanto, estimam os efeitos médios dos genes.

Intervalos que contém QTL podem ser localizados por meio da soma dos efeitos absolutos dos haplótipos dentro de cada intervalo. Se não existe um QTL em um intervalo, todas as estimativas dos efeitos dos haplótipos dentro dele serão de pequena magnitude em módulo. Intervalos com as maiores somas desses efeitos absolutos provavelmente contém um QTL ou estará adjascente a um intervalo contendo um QTL. Dessa forma, a posição do QTL pode ser encontrada e a descoberta de QTL com grande efeito é facilitada. No entanto, a MAS não será de fato necessária, pois todos os QTLs poderão ser selecionados simultaneamente usando GWS. A alta acurácia da GWS e a possibilidade de seleção na fase de embriões conduzirão a grandes alterações nas estratégias de melhoramento em várias espécies.

Procedimentos Bayesianos para a Estimação dos Efeitos dos Haplótipos ou Marcadores

Vários métodos de predição de valores genéticos genômicos foram propostos: quadrados mínimos (LS), BLUP/GWS, BayesA e BayesB (MEUWISSEN et al., 2001) e aprendizado de máquina (AM de LONG et al., 2007). Essas abordagens diferem na suposição sobre o modelo genético associado ao caráter quantitativo. O BLUP assume o modelo infinitesimal com muitos locos de pequenos efeitos; o AM assume que existe um número limitado de genes e de SNPs a serem ajustados; o método BayesB é intermediário entre esses dois, assumindo poucos genes de grandes efeitos e muitos genes com pequenos efeitos. No método BayesB, muitos efeitos de marcadores são assumidos como zero, *a priori*. Isso reduz o tamanho do genoma, por meio da concentração nas partes do mesmo onde existem QTLs. O melhor método é aquele que reflete melhor a natureza biológica do caráter poligênico em questão, em termos de efeitos gênicos.

O método LS é ineficiente devido a: impossibilidade de estimar todos os efeitos simultaneamente, pois o número de efeitos é maior do que o número de dados; estimando um efeito de cada vez e verificando a sua significância, conduz a superestimativas dos efeitos significativos; a acurácia do método é baixa; somente QTLs de grande efeito serão detectados e usados e, conseqüentemente, nem toda a variação genética será capturada pelos marcadores.

O método LS assume distribuição *a priori* para os QTLs, com variância infinitamente grande, fato que é incompatível com a conhecida variância genética total. O BLUP/GWS assume os efeitos de QTL com distribuição normal, com variância constante através dos segmentos cromossômicos. A distribuição dos efeitos de QTL é conhecida em poucos caracteres e espécies. Em gado bovino leiteiro, Goddard e Hayes (2007) relatam a presença de 150 QTLs para o caráter produção de leite e estimaram a distribuição de seus efeitos como aproximadamente exponencial. Com distribuição exponencial e não muitos efeitos com valor zero, o melhor estimador dos efeitos alélicos é denominado LASSO (TIBSHIRANI, 1996). Entretanto, com muitos efeitos com valor zero, o LASSO não é adequado.

O método ideal de predição de valores genéticos genômicos equivale ao cálculo da média condicional do valor genético dado o genótipo do indivíduo em cada QTL. Essa média somente pode ser calculada usando uma distribuição *a priori* dos efeitos dos QTLs. Considerando cada QTL em separado, essa esperança condicional é dada por $\hat{h} = E(h|dados)$. O estimador apropriado segue o teorema de Bayes e é dado por

$$\hat{h} = \frac{\int h * f(m|h) \quad f(h) \; d \; h}{\int \quad f \; (m|h) \quad f(h) \; d \; h}$$

em que que $f\ (m|h)$ é a verossimilhança dos dados (m), e $f(h)$ é a distribuição *a priori* dos efeitos dos QTLs. Esse estimador mostra que o método ideal depende da distribuição *a priori* $f(h)$ dos efeitos de QTL. A presença de QTLs é testada em muitas posições (10 mil SNPs) e, portanto, não existe QTLs em muitas posições. Dessa forma, a distribuição *a priori* deve ter uma alta probabilidade para $f(0)$. Para especificar essa alta probabilidade, deve-se ter uma noção de quantos QTLs controlam o caráter (GODDARD; HAYES, 2007).

Nessa situação, com muitos efeitos h iguais a zero, o método BLUP/GWS resulta em muitas estimativas de h próximas de zero, porém não iguais a zero. Na soma dessas estimativas, esse efeito acumulado pode introduzir algum erro na predição. Os métodos bayesianos BayesA e BayesB relatados por Meuwissen et al. (2001) consideram mais adequadamente a distribuição *a priori* dos efeitos dos QTLs.

O método BayesA equivale ao método BLUP, porém as variâncias dos segmentos cromossômicos diferem para cada segmento e são estimadas sob esse modelo, considerando a informação combinada dos dados e da distribuição *a priori* para essas variâncias. Essa distribuição é tomada como uma qui-quadrado invertida e escalada. Para obtenção dessa informação combinada ou da distribuição *a posteriori* das variâncias, adota-se o procedimento da amostragem de Gibbs. Detalhes da estimação bayesiana são apresentados por Resende (2002) e Sorensen e Gianola (2002).

O método BayesB usa uma distribuição *a priori* dos efeitos dos QTLs com alta densidade em $\sigma_g^2 = 0$ e distribuição qui-quadrado invertida para $\sigma_g^2 > 0$. Assim, considera que em muitos locos não existe variação genética, ou seja, não estão segregando. A distribuição *a priori* do método BayesA não tem um pico de densidade em $\sigma_g^2 = 0$. Uma vez que não é possível uma amostragem de $\sigma_g^2 = 0$, o método da amostragem de Gibbs não pode ser usado no método BayesB. Assim, o algoritmo de Metropolis-Hastings deve ser usado.

Implementação da Seleção Genômica Ampla

Na prática da seleção genômica ampla, três populações ou conjuntos de dados são necessários, conforme descrito na seqüência, com base em Goddard e Hayes (2007).

População de Descoberta. Esse conjunto de dados contempla um grande número de marcadores SNPs avaliados em um número moderado de indivíduos, os quais devem ter seus fenótipos avaliados para os vários caracteres de interesse. Equações de predição de valores genéticos genômicos são obtidas para cada caráter de interesse. Essas equações associam a cada intervalo marcador o seu efeito (predito por BLUP) no caráter de interesse.

População de Validação. Esse conjunto de dados é menor do que aquele da população de descoberta e contempla indivíduos avaliados para os marcadores SNPs e para os vários caracteres de interesse. As equações de predição de valores genéticos genômicos são testadas para verificar suas acurácias nessa amostra independente. Para computar essa acurácia, os valores genéticos genômicos são preditos (usando os efeitos estimados na população de descoberta) e submetidos à análise de correlação com os valores fenotípicos observados. Como a amostra de validação não foi envolvida na predição dos efeitos dos haplótipos marcadores, os erros dos valores genéticos genômicos preditos e dos valores fenotípicos são independentes e toda correlação entre esses valores é de natureza genética e equivale à própria acurácia. Para cômputo dessa correlação, podem ser usados valores genotípicos preditos com base nos fenótipos, em vez dos valores fenotípicos brutos.

População de Seleção. Esse conjunto de dados contempla apenas os marcadores SNPs avaliados nos candidatos à seleção. Essa população não necessita ter os seus fenótipos avaliados. As equações de predição derivadas na população de descoberta são então usadas na predição dos VGG ou fenótipos futuros dos candidatos à seleção. Mas, a acurácia seletiva associada refere-se àquela calculada na população de validação.

Segundo Meuwissen (2007), quando dezenas a centenas de milhares de haplótipos são estimados, existe o risco de superparametrização, ou seja, erros nos dados serem explicados pelos efeitos de marcadores. A validação cruzada é então de grande importância para contornar esse problema.

Sob seleção genômica, todos os candidatos à seleção (indivíduos sem observação fenotípica) poderão ser avaliados para quaisquer ambientes, desde que tais ambientes possuam equações de predição derivadas para os próprios e com alta acurácia. Acurácia da ordem de 85 % para a GWS foi relatada por Meuwissen et al. (2001) para uma população com 2.200 indivíduos com avaliações fenotípicas. Tais autores relataram também que equações de predição acuradas (71 %) foram obtidas mesmo para populações de descoberta de tamanho modesto, tal qual com 500 indivíduos com avaliações fenotípicas. Assim, indivíduos poderão ser comercializados com base em seus valores fenotípicos preditos (valores de cultivo e uso – VCU), derivados de um catálogo de marcadores associados aos candidatos à seleção. Também os produtos animais e vegetais (leite, carne, alimentos, fibras) poderão ser remunerados com base em seus marcadores genéticos (GODDARD; HAYES, 2007).

Simulação e Aspectos Computacionais da Seleção Genômica Ampla

Um caráter quantitativo controlado por 20 locos com dois alelos e efeitos aditivos foi simulado. Conforme Lande e Thompson (1990) e Bernardo e Yu (2007), em cada loco, o efeito do alelo favorável foi computado por $a = (L - 1) / (L + 1)$, em que L refere-se ao L-ésimo QTL. O efeito do alelo menos favorável foi tomado como $-a$. Esse modelo segue a parametrização clássica em que a refere-se à metade da diferença entre os dois genótipos homozigotos.

Foram simulados 1.000 indivíduos com fenótipos gerados segundo o modelo $f = g + e$, em que g são os efeitos genéticos totais dados pelo somatório dos efeitos genéticos em cada loco e e são os efeitos ambientais, gerados segundo uma distribuição normal com média zero e variância compatível com uma herdabilidade individual de 20 %. Em cada indivíduo, genótipos do tipo 11, 12 ou 22 foram sorteados aleatoriamente em cada loco, em que 1 denota alelo favorável e 2 denota alelo desfavorável.

Foram considerados 30 locos marcadores, 20 explicando exatamente cada um dos 20 locos e 10 com efeitos nulos, ou seja, sem qualquer associação com os 20 QTLs. A partir dos dados fenotípicos gerados, foram estimados os efeitos de cada um dos 30 locos marcadores por meio do procedimento BLUP/GWS e esses efeitos foram somados para fornecer os valores genéticos genômicos preditos. Considerou-se uma variância genética comum a todos os locos, dada por σ_g^2 / n, em que σ_g^2 é a variação genética total associada aos efeitos genéticos simulados e n é o número de locos marcadores.

O principal problema computacional da GWS é que a matriz de informação Z'Z implica densa matriz dos coeficientes das equações de modelo misto e grandes esforços e recursos computacionais. Legarra e Misztal (2008) indicam o método de Gauss-Seidel com atualização de resíduos como o mais apropriado para a predição BLUP e estimação de componentes de variância nesse contexto. Para predição dos efeitos dos locos marcadores e dos valores genéticos genômicos dos indivíduos, empregou-se o *software* Selegen-Genômica-REML/BLUP/GWS que implementa a GWS para algumas situações.

A acurácia da GWS foi computada, por meio da correlação entre os efeitos genéticos genômicos preditos e os efeitos genéticos paramétricos simulados.

Resultados e Discussão

Relação entre BLUP Tradicional e BLUP Genômico

O efeito genético aditivo (a) de um indivíduo i, predito pelo BLUP tradicional é dado por:

$$\hat{a}_i = 0{,}5\ (\hat{a}_p + \hat{a}_m) + \hat{a}_d$$

$$= 0{,}5\ (\hat{a}_p + \hat{a}_m) + h_d^2\ (y - X\hat{b} - 0{,}5\ \hat{a}_p - 0{,}5\ \hat{a}_m)$$

em que $h_d^2 = (1/2\ \sigma_a^2)/(1/2\ \sigma_a^2 + \sigma_e^2)$ é a herdabilidade dentro de famílias de irmãos germanos, σ_a^2 é a variância genética aditiva e σ_e^2 é a variância residual. As demais quantidades são:

$\hat{a}_p$: efeito genético aditivo predito do genitor paterno.

$\hat{a}_m$: efeito genético aditivo predito do genitor materno.

$\hat{a}_d$: efeito genético aditivo predito do indivíduo dentro de família, ou seja, desvio em relação à média dos efeitos aditivos paterno e materno, explicado pela segregação de amostragem mendeliana que ocorre durante a formação de gametas.

Por esse procedimento, a fração $0{,}5\ (\hat{a}_p + \hat{a}_m)$ é predita com alta acurácia, pois, os efeitos aditivos dos genitores são preditos com base em várias repetições experimentais. Por outro lado, a quantidade $\hat{a}_d$ é predita com baixa acurácia devido ao fato de que cada indivíduo é único e, portanto, não propicia repetições experimentais, exceto se for clonado. Adicionalmente, a variação dentro de famílias $(1/2\ \sigma_a^2)$ é considerada comum a todas, o que não é verdadeiro, exatamente devido às diferentes segregações associadas a cada família.

Para contornar esse problema, as seguintes medidas podem ser adotadas: (i) realizar teste clonal dos indivíduos e estimar diretamente o valor genético total do indivíduo; (ii) adotar o procedimento BLUP-VEG, conforme descrito por Resende (2007), o qual considera variação dentro de família específica para cada família; (iii) adotar a MAS; (iv) adotar a GWS. Dentre essas alternativas, as melhores são o teste clonal e a GWS. O BLUP-VEG não permite repetições experimentais e não avalia diretamente (via DNA) as segregações realizadas. A MAS não abrange adequadamente todo o caráter poligênico.

Em resumo, todas essas técnicas visam melhorar a acurácia da estimativa $\hat{a}_d$, referente aos efeitos da segregação mendeliana. A GWS é o método que explora adequadamente a segregação de amostragem mendeliana que ocorre por ocasião da formação de gametas, pois captura a matriz de parentesco realizada e não uma matriz de parentesco médio associada ao *pedigree*. Uma vez que a GWS avalia diretamente o DNA associado (via marcadores) a cada loco de todo o caráter poligênico, avalia diretamente cada segregação em nível individual e não em nível médio. Assim, a GWS avalia diretamente o genótipo dos filhos, permite conhecer cada segregação e produz estimativas mais acuradas de valores genéticos por meio da melhor predição do termo referente à segregação mendeliana $\hat{a}_d$. Conforme Goddard e Hayes (2007), sob o modelo infinitesimal com grande número de locos de pequeno efeito, a GWS prediz os valores genéticos de maneira mais acurada do que o BLUP tradicional baseado em pedigree e dados fenotípicos. Bernardo e Yu (2007) relatam a superioridade da GWS sobre a MAS.

A GWS enfatiza mais o termo referente à segregação mendeliana $\hat{a}_d$, dando mais peso a esse componente do que o faz o BLUP tradicional. Isso leva à seleção de menos indivíduos aparentados do que o faz o BLUP, reduzindo assim o incremento da endogamia na população. Daetwyler et al. (2007) relatam que a GWS aumenta em torno de 67 % a acurácia da predição de $\hat{a}_d$ em comparação com o BLUP tradicional e, conseqüentemente, eleva, em determinada situação, a acurácia da seleção individual de 71 % para 85 %.

No contexto da GWS, o modelo para o valor fenotípico y de um indivíduo é dado por:

$y = g + e = h + g^* + e$, em que g representa os efeitos genéticos, e refere-se aos efeitos ambientais, h refere-se aos efeitos genéticos explicados pelos

marcadores e g* representa os efeitos genéticos residuais não explicados pelos marcadores. O efeito genético h de um indivíduo é estimado por meio de $\hat{h} = \sum_j (\hat{h}_j^p + \hat{h}_j^m)$, em que j refere-se a uma região genômica e $\hat{h}_j^p$ e $\hat{h}_j^m$ são as estimativas BLUP dos haplótipos marcadores paternais e maternais na região genômica j. O estimador $\hat{h}$ pode ser expresso alternativamente por:

$$\hat{h} = 0{,}5\ (\hat{h}_p + \hat{h}_m) + [(\sum_j \hat{h}_j^p - 0{,}5\ \hat{h}_p) + (\sum_j \hat{h}_j^m - 0{,}5\ \hat{h}_m)] = \sum_j (\hat{h}_j^p + \hat{h}_j^m)$$

Essa expressão é similar à expressão apresentada anteriormente para $\hat{a}_i$, sendo que o termo entre colchetes representa os efeitos da segregação mendeliana. No entanto, tal termo tem herdabilidade igual a 1 e não igual a h_d^2, pois o efeito genético genômico é condicional aos genótipos marcadores e previamente derivado de estimativas dos efeitos dos marcadores e, portanto, não contém efeito residual. Logicamente, isso assume que os haplótipos podem ser determinados sem erro (DEKKERS, 2007). É importante relatar que a acurácia global da GWS não é 1, pois g paramétrico é dado por $g = h + g^*$. Ou seja, existe uma parte (g*) de g que não é explicada pelos marcadores e não é contemplada pela GWS.

Simulação e Acurácia da GWS

A acurácia ($r_{q\hat{m}}$) da seleção GWS depende da proporção (r_{mq}^2) da variação genética explicada pelos marcadores e da acurácia ($r_{m\hat{m}}$) da predição dos efeitos dos marcadores que estão em desequilíbrio de ligação com os QTL's, segundo a expressão $r_{q\hat{m}} = (r_{m\hat{m}}^2 r_{mq}^2)^{1/2}$.

O parâmetro r_{mq}^2 depende da densidade de marcadores e da extensão e padrão do desequilíbrio de ligação que existe na população. Por sua vez, o parâmetro $r_{m\hat{m}}$ depende da quantidade e precisão dos dados disponíveis para estimar os efeitos dos marcadores, além da eficiência da estratégia e dos métodos estatísticos usados na predição. Nas Tabelas 1, 2 e 3 são apresentados resultados referentes à eficiência da GWS, obtidos a partir da simulação de dados.

Tabela 1. Eficiência da GWS para diferentes valores da proporção da variância genética explicada pelos marcadores usados na seleção. Análise de dados simulados.

Número de Marcadores Usados na Seleção	Número de QTLs Usados na Seleção	Proporção da Variação Genética Total Explicada pelos Marcadores (r_{mq}^2 em %)	Acurácia Seletiva da GWS (%)	Coeficiente de Determinação Genética na Seleção pela GWS (%)
20	20	100,00	94,59	89,47
30	20	100,00	91,78	84,24
10	10 maiores	62,50	77,72	60,40
7	7 maiores	44,84	65,76	43,25
5	5 maiores	32,45	55,00	30,25
Seleção Massal	-	-	45,00	20,00
3	3 maiores	19,70	43,88	19,25
5	5 menores	12,95	30,61	9,37

Com base na Tabela 1, verifica-se que quando a densidade de marcadores e o padrão de desequilíbrio de ligação são suficientes para explicar 100 % da variação genética, a acurácia da GWS por meio da utilização de 1.000 indivíduos na população de descoberta é da ordem de 92 % ou mais, nas condições da presente simulação. Verifica-se, também, comparando-se os resultados das duas primeiras linhas da Tabela 1, que o uso de marcadores não informativos, além daqueles que explicam os QTLs, praticamente não reduz a acurácia do processo seletivo. A redução na acurácia foi de 94,5 % para 92 %. Isto confirma que não é necessário o conhecimento *a priori* da significância dos efeitos dos marcadores. Essa inferência é corroborada também pelos resultados da Tabela 2.

Na situação em que apenas os 10 QTLs de maior efeito são explicados pelos marcadores, a proporção explicada da variação genética equivaleu a 62,5 % e a acurácia seletiva foi de 77,7 %. Quando a proporção explicada da variação genética foi de 44,8 %, a acurácia equivaleu a 65,8 %, a qual é baixa no contexto de um programa de avaliação genética (Tabela 1). Essa situação é análoga ao caso em que se detecta um QTL de grande efeito que explica 45 % da variação genética de um caráter. Nesse caso, o uso de apenas esse QTL de grande efeito na seleção assistida ou introgressão via transgenia é insuficiente para melhorar adequadamente o caráter em questão.

Quando apenas os 3 QTLs de maior efeito foram usados, a proporção da variação genética explicada foi de apenas 19,7% e a acurácia equivaleu a apenas 43,9% (Tabela 1). Essa situação pode ser análoga ao caso em que a MAS baseia-se apenas em um limitado número de QTLs com efeitos individuais estatisticamente significativos. Esses resultados revelam que pequenos efeitos de locos individuais, não significativos isoladamente, quando somados têm grande influência no caráter quantitativo. A filosofia da GWS é capitalizar todos os efeitos, pequenos ou grandes, de forma acumulada.

A última linha da Tabela 1 mostra a situação em que apenas marcadores de QTLs de pequenos efeitos estão disponíveis ou são informativos. Nesse caso, a eficiência da GWS é baixa e isso revela a importância de se ter uma alta densidade de marcadores, de forma que a maioria dos QTLs, de grandes e pequenos efeitos, esteja marcada.

Na Tabela 2 é apresentada a eficiência do procedimento BLUP/GWS em estimar os efeitos dos marcadores em cada QTL.

Tabela 2. Eficiência do procedimento BLUP/GWS em estimar os efeitos dos marcadores em cada QTL.

Número de Marcadores Usados na Seleção	Número de QTLs Usados na Seleção	Proporção da Variação Genética Total Explicada pelos Marcadores (r^2_{mq} em %)	Acurácia ($r_{q\hat{m}}$) Observada na Estimação do Efeito Genético de Cada QTL (%)	Coeficiente de Determinação Observado na Estimação do Efeito Genético de Cada QTL (%)
20	20	100,00	94,97	90,19
30	20	100,00	94,88	90,01

Verifica-se, pela Tabela 2, que o procedimento BLUP/GWS baseado em 1.000 indivíduos conduziu a uma acurácia de 95 % na estimação dos efeitos dos marcadores em cada QTL. Nesse caso, o uso apenas dos marcadores informativos ou de todos os marcadores conduziu à mesma eficiência.

Na Tabela 3 são apresentados os números de indivíduos necessários na população de descoberta.

Tabela 3. Aumento da acurácia da estimação dos efeitos de cada QTL em função do aumento do tamanho da população de descoberta. Herdabilidade de 20 % e número de QTLs igual a 20.

Número de Indivíduos	Herdabilidade de Um Indivíduo	Herdabilidade de Médias de Vários Indivíduos	Acurácia Esperada na Estimação do Efeito do QTL
100	0,01	0,50	0,71
200	0,01	0,67	0,82
500	0,01	0,83	0,91
1000	0,01	0,91	0,95
2000	0,01	0,95	0,98
4000	0,01	0,98	0,99
8000	0,01	0,99	0,99

A partir da herdabilidade ou contribuição de um indivíduo para inferir sobre os efeitos de cada QTL marcado, pode-se determinar o aumento da acurácia na estimação desses efeitos quando aumenta-se o tamanho da amostra. A equação para obtenção da acurácia esperada em função do tamanho (N) da amostra de indivíduos é dada por:

$r_{q\hat{q}} = \sqrt{(Nh_q^2)/[1+(N-1)h_q^2]}$. Essa equação depende de N e também da herdabilidade de um loco em um indivíduo, dada por:

$h_q^2 = (\sigma_g^2/n)/(\sigma_g^2+\sigma_e^2) = (\sigma_g^2/n)/\sigma_f^2$.

A quantidade $(Nh_q^2)/[1+(N-1)h_q^2]$ está associada ao fator de regressão (shrinkage) dos efeitos de marcadores nas equações de modelo misto para obtenção do BLUP. Os componentes σ_g^2, σ_e^2 e σ_f^2 referem-se às variâncias genética, ambiental e fenotípica, respectivamente, e n equivale ao número de locos.

Verifica-se que, nas condições estudadas (herdabilidade de 20 % e 20 locos), 500 ou mais indivíduos são necessários para se ter uma acurácia adequada (igual ou maior que 91 %). Com 1.000 indivíduos a acurácia esperada é de 95 % e para obter acurácia superior a 98 %, 2 mil indivíduos são necessários (Tabela 3). Esses números diferem daqueles relatados por Meuwissen et al. (2001): acurácia de 71 % para uma população de 500 indivíduos e acurácia de 85 % para uma população de 2.200 indivíduos. Isto é devido aos diferentes números de locos, herdabilidades e desequilíbrio de ligação considerados nos dois trabalhos. Novas simulações envolvendo outros valores de herdabilidade e número de locos são necessárias visando à determinação do tamanho amostral adequado à cada situação. Tais simulações serão relatadas em um outro trabalho. Para o presente caso, com 1.000 indivíduos, a acurácia esperada foi de 95 % (Tabela 3) e a acurácia observada ou calculada foi de 95 % (Tabela 2). Isso confirma a adequação da equação para obtenção da acurácia esperada.

Para validação cruzada adotou-se 100 indivíduos na população de validação e 900 indivíduos na população de descoberta. A acurácia média obtida foi de 0,4561, dada pela correlação entre os valores fenotípicos observados dos 100 indivíduos e seus valores genéticos genômicos preditos. Essa acurácia ou correlação equivale à raiz quadrada da herdabilidade individual. No presente caso, a raiz quadrada da herdabilidade assumida (0,20) equivale a 0,4472. Os valores 0,456 e 0,447 são próximos e indicam que as acurácias nas populações de descoberta e de validação são similares, ou seja, estão explicando os mesmos QTLs. As acurácias são, então, cerca de 45 % para explicar os fenótipos individuais e

em torno de 92 % para explicar os valores genéticos verdadeiros por meio da seleção baseada em 30 marcadores, informando sobre os 20 QTLs.

De maneira mais genérica, a acurácia da predição do valor genético de um loco de um caráter quantitativo é dada por:

$$r^2_{q\hat{q}} = \sqrt{r^2_{m\hat{m}} r^2_{mq}} = \sqrt{(Nh^2_m r^2_{mq}) / [1 + (N-1)h^2_m]}$$

em que r^2_{mq} é a proporção da variação genética explicada pelos marcadores, $r_{m\hat{m}}$ é a acurácia da predição dos efeitos dos marcadores que estão em desequilíbrio de ligação com os QTL´s e $h^2_m = (\sigma^2_g r^2_{mq} / n) / (\sigma^2_g r^2_{mq} + \sigma^2_e)$. Os resultados das Tabelas 2 e 3 assumem r^2_{mq} igual a 1, ou seja, assumem uma alta densidade de marcadores e grande desequilíbrio de ligação na população. Essa é uma premissa da GWS. No entanto, situações diferentes da ideal podem ser avaliadas, por meio da consideração de diferentes valores de r^2_{mq}. A proporção r^2_{mq} pode ser expressa por $r^2_{mq} = d^2 l^2$

O componente d^2 é a densidade efetiva de marcadores ligados a QTLs, em proporção do número de QTLs que devem ser capturados. Por sua vez, o componente l^2 é a proporção da variação em um loco explicada pelo marcador, a qual é dependente da extensão do desequilíbrio de ligação entre marcador e QTL e, conseqüentemente, da freqüência de recombinação entre eles. Os valores de r^2_{mq} na Tabela 1 assumem l^2 igual a 1, ou seja, tem-se um perfeito conhecimento dos alelos do QTL por meio dos alelos do marcador. Assim, na Tabela 1, r^2_{mq} é dado por $r^2_{mq} = d^2$, ou seja, tal proporção varia apenas em função da densidade efetiva dos marcadores e sua capacidade de capturar determinado número de QTLs com dada variação genética acumulada. E os resultados das Tabelas 2 e 3 assumem tanto d^2 quanto l^2 iguais a 1.

Valores de l^2 diferentes de 1 podem ser considerados visando ajustar os resultados das Tabelas 1, 2 e 3, em função de diferentes níveis de desequilíbrio de ligação nas populações. Na Tabela 4, considerou-se $r^2_{mq} = l^2$ e foram considerados os seguintes valores de l^2: 0,1; 0,3; 0,5; 0.7 e 0,9. O parâmetro l^2 aqui não equivale à estatística r^2 de desequilíbrio de ligação, mas denota a proporção da variação do QTL explicada pelo marcador, devida ao desequilíbrio de ligação e distância entre os locos do marcador e do QTL. Engloba também situações em que são usados haplótipos marcadores compostos por vários marcadores adjacentes (2 a 6, geralmente).

Tabela 4. Aumento da acurácia da estimação dos efeitos de cada QTL em função do aumento do tamanho da população de descoberta e de diferentes valores de r^2_{mq}. Herdabilidade de 20 % e número de QTLs igual a 20.

Número de Indivíduos	Acurácia Esperada na Estimação do Efeito do QTL: r^2_{mq} = 0,1	Acurácia Esperada na Estimação do Efeito do QTL: r^2_{mq} =0,3	Acurácia Esperada na Estimação do Efeito do QTL: r^2_{mq} = 0,5	Acurácia Esperada na Estimação do Efeito do QTL: r^2_{mq} = 0,7	Acurácia Esperada na Estimação do Efeito do QTL: r^2_{mq} = 0,9
100	0,10	0,28	0,42	0,55	0,66
200	0,14	0,35	0,51	0,65	0,76
500	0,19	0,44	0,61	0,74	0,86
1000	**0,23**	**0,48**	**0,65**	**0,79**	**0,90**
2000	0,27	0,51	0,68	0,81	0,92
4000	0,29	0,53	0,69	0,82	0,94
8000	0,30	0,54	0,70	0,83	0,94

Verifica-se que, para valores de r^2_{mq} iguais a 0,70, acima de 1.000 indivíduos são necessários para se ter acurácias superiores a 80 %. Para valores de r^2_{mq} menores que 0,50, as acurácias são baixas mesmo com tamanhos de população tão altos quanto 8 mil indivíduos. Assim, valores de r^2_{mq} iguais ou superiores a 0,70 são desejáveis. Com a descoberta de SNPs, muitos valores de l^2 tendendo a 1 têm sido reportados em gado de leite e gado de corte (HAYES et al., 2006). Outra forma, de aumentar o valor de l^2 é por meio do uso de haplótipos marcadores em lugar do uso de marcadores simples para identificar os alelos do QTL carregados por cada indivíduo. Hayes et al. (2007) relatam um aumento de 0,20 para 0,55 na proporção da variância do QTL explicada pelos marcadores, com o uso de haplótipos com seis marcadores ao invés de um só marcador, em gado bovino.

Outros números de locos foram também avaliados. Para a herdabilidade de 20 %, um r^2_{mq} de 0,7 e 50 locos controlando o caráter, os tamanhos amostrais de 1.000 e 2 mil na população de descoberta conduzem a acurácias de 72 % e 77 %, respectivamente.

Quanto aos valores adequados de LD (medidos por r^2), Hayes et al. (2001) relatam que com $r^2 = 0,3$, o uso de onze marcadores por cM resulta em haplótipos explicando 98 % da variância do QTL, ou seja, $r^2_{mq} = 0,98$.

Conclusões

Os resultados de simulação revelam um grande potencial da GWS em aumentar a eficiência do melhoramento. Esse benefício só se concretizará se houver um alto grau de desequilíbrio de ligação envolvendo marcadores SNPs proximamente espaçados e se os efeitos genéticos dos marcadores nas características sob melhoramento forem estimados (a partir dos fenótipos) com alta acurácia e usados na própria população e ambientes em que forem estimados. É preciso adquirir experiência prática com a GWS, visando inferir sobre sua efetividade.

O estudo da relação entre o BLUP tradicional e o BLUP genômico associado à GWS revelou a superioridade desse último por meio da possibilidade de avaliar diretamente cada segregação em nível individual e não em nível médio. A GWS avalia diretamente o genótipo dos filhos, permite conhecer cada segregação e produz estimativas mais acuradas de valores genéticos por meio da melhor predição do termo referente à segregação mendeliana, o qual compõe o valor genético estimado de um indivíduo.

Tamanhos amostrais da ordem de 1.000 ou mais são necessários para uma estimação acurada dos valores genéticos genômicos.

Referências

BERNARDO, R; YU, J. Prospects for genome wide selection for quantitative traits in maize. **Crop Science**, v. 47, p.1082-1090, 2007.

CALUS, M. P. L.; MEUWISSEN, T. H. E.; ROOS, A. P. W.; VEERKAMP, R. F. Accuracy of genomic selection using different methods to define haplotypes. **Genetics**, v. 178, p. 553-561, 2008.

CALUS, M. P. L.; VEERKAMP, R. F. Accuracy of breeding value when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM. **Journal of Animal Breeding and Genetics**, v. 124, p. 362-368, 2007.

DAETWYLER, H. D.; VILLANUEVA, B.; BIJMA, P.; WOOLIAMS, J. A. Inbreeding in genome-wide selection. **Journal of Animal Breeding and Genetics**, v. 124, p. 369-376, 2007.

DEKKERS, J. C. M. Commercial application of marker and gene assisted selection in livestock: strategies and lessons. **Journal of Animal Science**, v. 82, p.313-328, 2004.

DEKKERS, J. C. M. Prediction of response to marker assisted and genomic selection using selection index theory. **Journal of Animal Breeding and Genetics**, v. 124, p. 331-341, 2007.

GODDARD, M. E.; HAYES, B. J. Genomic selection. **Journal of Animal Breeding and Genetics**, v. 124, p. 323-330, 2007.

HAYES, B. J.; CHAMBERLAIN, A. J.; GODDARD, M. E. Use of markers in linkage disequilibrium with QTL in breeding programs. In: WORLD CONGRESS ON GENETICS APPLIED TO LIVESTOCK PRODUCTION, 8., 2006, Belo Horizonte, MG; **Proceedings**... Belo Horizonte: Instituto Prociência, 2006. 1 CD-ROM.

HAYES, B. J.; CHAMBERLAIN, A. J.; McPARTLAN, H.; MACLEOD, I.; SETHURAMAN, L.; GODDARD, M. E. Accuracy of marker assisted selection with single markers and marker haplotypes in cattle. **Genetical Research**, v. 89, p. 215-220, 2007.

HAYES, B. J.; BOWMAN, P. J.; GODDARD, M. E. Linkage disequilibrium and accuracy of predicting breeding values from marker haplotypes. In: **Association for the Advancement of Animal Breeding and Genetics**, Jopson N (ed), Proceedings of the Association for the Advancement of Animal Breeding and Genetics, v. 14, p.269-272, 2001.

KOLBEHDARI, D.; SHAEFFER, L. R.; ROBINSON, J. A. B. Estimation of genome-wide haplotype effects in half-sib designs. **Journal of Animal Breeding and Genetics**, v. 124, p. 356-361, 2007.

LANDE, R.; THOMPSON, R. Efficiency of marker assisted selection in the improvement of quantitative traits. **Genetics**, v. 124, p.743-756, 1990.

LEGARRA, A.; MISZTAL, I. Computing strategies in genome-wide selection. **Journal of Dairy Science**, v. 91, n.1, p. 360-366, 2008.

LONG, N.; GIANOLA, D.; ROSA, G. J. M.; WEIGEL, K. A; AVENDAÑO, S. Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality in broilers. **Journal of Animal Breeding and Genetics**, v. 124, p. 377–389, 2007.

MEUWISSEN, T. H. E. Genomic selection: marker assisted selection on genome-wide scale. **Journal of Animal Breeding and Genetics**, v. 124, p. 321-322, 2007.

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

MUIR, W. M. Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. **Journal of Animal Breeding and Genetics**, v. 124, p. 342-355, 2007.

RESENDE, M. D. V. **Matemática e estatística na análise de experimentos e no melhoramento genético**. Colombo: Embrapa Florestas, 2007. 561 p.

RESENDE, M. D. V. **Genética biométrica e estatística no melhoramento de plantas perenes**. Brasília: Embrapa Informação Tecnológica, 2002. 975 p.

SCHAEFFER, L. R. Strategy for applying genome-wide selection in dairy cattle. **Journal of Animal Breeding and Genetics**, v. 123, p. 218-223, 2006.

SOLBERG, T. R.; SONESSON, A.; WOOLIAMS, J.; MEUWISSEN, T. H. E. Genomic selection using different marker types and density. In: WORLD CONGRESS OF GENETICS APPLIED TO LIVESTOCK PRODUCTION, 8., 2006. **Proceedings**. Belo Horizonte: Ed. da UFMG, 2006. 1 CD-ROM.

SORENSEN, D; GIANOLA, D. **Likelihood, Bayesian and MCMC Methods in Quantitative Genetics**. New York: Springer Verlag, 2002.

TIBSHIRANI, R. Regression shrinkage and selection via the Lasso. **Journal of the Royal Statistics Society Series B**, v. 58, p. 267-288, 1996.

Recebido em 06 de agosto de 2007 e aprovado em 19 de agosto de 2008