Carla Daniela Suguimoto Leite

Análise de associação global do genoma para características produtiva e reprodutiva em suínos

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Zootecnia, para obtenção do título de *Doctor Scientiae*.

VIÇOSA
MINAS GERAIS - BRASIL
2013

Ficha catalográfica preparada pela Seção de Catalogação e Classificação da Biblioteca Central da UFV

T
L533a
2013

Leite, Carla Daniela Suguimoto, 1983-
Análise de associação global do genoma para características produtiva e reprodutiva em suínos / Carla Daniela Suguimoto Leite. – Viçosa, MG, 2013.
xiii, 52 f. : il. (algumas color.) ; 29 cm.

Orientador: Ricardo Frederico Euclydes.
Tese (doutorado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Suíno - Melhoramento genético. 2. Marcadores genéticos. 3. Genoma. I. Universidade Federal de Viçosa. Departamento de Zootecnia. Programa de Pós-Graduação em Zootecnia. II. Título.

CDD 22. ed. 636.40821

Carla Daniela Suguimoto Leite

Análise de associação global do genoma para características produtiva e reprodutiva em suínos

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Zootecnia, para obtenção do título de *Doctor Scientiae.*

APROVADA: 21 de fevereiro de 2013.

Marcelo Silva de Freitas

Fabyano Fonseca e Silva

Mônica Corrêa Ledur
(Coorientadora)

Robledo de Almeida Torres
(Coorientador)

Prof. Ricardo Frederico Euclydes
(Orientador)

Dedico este trabalho aos meus amados pais, Otacilio e Elisabeth,
Ao meu querido irmão Flávio e minha cunhada Carina,
À minha adorada sobrinha Laura.

"A mente que se abre a uma nova ideia jamais voltará ao seu tamanho original"

Albert Einstein

AGRADECIMENTOS

A DEUS, sempre presente em nossas vidas, por nos iluminar e proteger.

A CAPES, pela concessão da bolsa de estudos;

Ao programa departamento de Zootecnia, pela oportunidade de retornar à Universidade Federal de Viçosa para a instituição, para realizar o doutorado;

Ao amigo e professor Bajá, Ricardo Frederico Euclydes, pela sua orientação, atenção, carinho, amizade, pelos momentos de descontração, churrascos no sítio, as pizzas no shopping, sempre com o violão na mão!

Ao professor e amigo, Robledo de Almeida Torres, conselheiro de tantos momentos, pela sua amizade, apoio, orientação, minha admiração e gratidão.

À pesquisadora Dra. Mônica Ledur, fundamental para realização deste trabalho. Agradeço por toda paciência, atenção, dedicação, horas de skype, milhares de e-mails, pela ajuda incondicional, minha eterna gratidão.

Ao Marcelo Freitas, pelo exemplo de conduta profissional e pessoal, pelo auxilio e conselhos durante a elaboração da tese, pela amizade. Obrigada por colaborar, sempre!

Ao Prof. Fabyano Fonseca, pela valiosa colaboração na qualificação e participação na banca de defesa de tese, por todos ensinamentos na minha formação, pela dedicação, pelas cervejinhas no Moreiras!

Ao Prof. Antonio Policarpo, pela participação, conselhos e sugestões na banca de qualificação.

Aos demais professores dos departamentos de Zootecnia, de Estatística, de Genética, pelos valiosos ensinamentos, pelo incentivo, pela amizade.

Aos funcionários do departamento de Zootecnia, sempre queridos e atenciosos, em especial a Fernanda, por responder 10 vezes a mesma pergunta....

À BRF S.A., pela concessão do banco de dados do programa de melhoramento genético de suínos e pela oportunidade profissional.

Ao Rodrigo Torres, gerente de genética e inovação da BRF, pelo convite para realização da tese com os dados do programa de melhoramento genético de suínos, pela amizade, pelo conselhos, pelas oportunidades;

Ao Otaviano, Jader, supervisores e equipe da Gerência de Genética, em Faxinal dos Guedes, pela coleta das informações;

À equipe da Gerencia de Genética, em Curitiba, Marcos Yamaki, Marcelo Piassi, em especial a Luciana Freitas e Marcos Lagrotta, que muito me ajudaram na edição dos dados.

Ao meu chefe, Osório Dal Bello, por me provar que 1 + 1 pode ser diferente de 2, e pelos conselhos, pela atenção, pela compreensão, incentivo;

À equipe da EMBRAPA CNPSA, Jane, Ricardo, Mauricio, pessoas essenciais para a realização deste trabalho, pela dedicação, pela colaboração, pela atenção. Minha gratidão e admiração;

À equipe da EMBRAPA CNPTIA, pelo auxilio nas rotinas para a realização das análises;

Aos meus colegas do melhoramento da UFV: Ana Lucia, André, Cris, Gilberto, Jefferson, Joahsy, Luanna, Luciano, Rodrigo Batata, Rodrigo Pacheco, Luiz Fernando, Matilde, Carol, pela amizade incondicional, carinho, respeito, pelos trabalhos que realizamos e pelos agradáveis churrascos.

Aos estagiários da Granja de Melhoramento de Codornas, Aline, Ariane, Giovani, Helmut, Karol, Marcela, Vitor Abreu, que muito colaboraram com o programa de melhoramento de codornas da UFV, pelo empenho e dedicação nas pesagens, análise de componentes, nascimento, pesagem de ração...

Aos meus colegas da pós graduação, desde a UNESP até a UFV pelo companheirismo, respeito, momentos de estudo e de descontração. Admiro muito a todos que contribuem de forma magnífica à pesquisa brasileira.

Ao grande amigo, Nicola, por mesmo longe, ser sempre um amigo presente, por ser meu brother de coração!

Às minhas irmãs que a vida me permitiu escolher, Ana Paula (Bis), Carol (mãe do Pedro e do Arthur), Daiana (Dadá), Mariele (Maxi), Priscila (Amiga), pela amizade eterna, pelo amor, pelo carinho, por sermos uma família.

Às famílias que formamos em cada república que morei, em Jabuka, Viçosa e Curitiba, pelos momentos de conversas, novelas, festinhas, churrascos, enfim cada pessoa que convivi, nestes anos, fez parte do meu amadurecimento e por isso sou eternamente grata!

Aos colegas do escritório da BRF em Curitiba, pela amizade, pelo incentivo... Muito obrigada!

Aos meus familiares, por me incentivarem, apoiarem e torcerem pela minha felicidade!

Aos meus pais, Beth e Otacilio, por serem minha fortaleza, meus maiores incentivadores, pelo amor incondicional,

Ao meu irmão Flavio e minha cunhada Carina, por terem trazido ao mundo o maior motivo de alegria da nossa família, a doce e linda Laurinha.

Ao Wilton, por ser paciente, por ter me ajudado sempre, por tudo o que representa em minha vida!

À todos que direta ou indiretamente contribuíram para que pudesse alcançar mais esta etapa em minha vida!

BIOGRAFIA

CARLA DANIELA SUGUIMOTO LEITE - filha de Otacílio Batista Leite e Elisabeth Suguimoto Leite nasceu em Ribeirão Preto, Estado de São Paulo, em 24 de novembro de 1983. Em maio de 2002, iniciou a graduação em Zootecnia na Universidade Federal de Viçosa, MG, obtendo o título em fevereiro de 2007.

Em março de 2007 ingressou no Programa de Pós-Graduação em Genética e Melhoramento Animal pela Faculdade de Ciências Agrárias e Veterinárias, UNESP, campus de Jaboticabal-SP, como bolsista da CAPES (Coordenação de Aperfeiçoamento de Pessoal de nível Superior), sob orientação do Prof. Dr. Jeffrey Frederico Lui. Submeteu-se à defesa do trabalho de dissertação para obtenção do título de Mestre em Genética e Melhoramento Animal em 16 de fevereiro de 2009.

Em março de 2009 iniciou o curso de doutorado no Programa de Zootecnia, pela Universidade Federal de Viçosa, como bolsista da CAPES (Coordenação de Aperfeiçoamento de Pessoal de nível Superior), sob orientação do prof. Ricardo Frederico Euclides.

Em 13 de julho de 2011 foi contratada pela empresa BRF para integrar a equipe da Diretoria de Inovação e Desenvolvimento Agropecuário.

Submeteu-se a defesa de tese para obtenção do título de *doctor scientiae* em Zootecnia, com ênfase em Genética e Melhoramento Animal, no dia 21 de fevereiro de 2013.

RESUMO

LEITE, Carla Daniela Suguimoto, D.Sc., Universidade Federal de Viçosa, fevereiro de 2013. **Análise de associação global do genoma para características produtiva e reprodutiva em suínos.** Orientador: Ricardo Frederico Euclydes. Coorientadores: Mônica Corrêa Ledur, Robledo de Almeida Torres e Antonio Policarpo Souza Carneiro.

A análise de associação global do genoma (GWAS) visa identificar regiões cromossômicas associadas a características fenotípicas de interesse econômico, com base nas diferenças entre as frequências alélicas dos polimorfismos de base única (SNPs). O objetivo nesse estudo foi identificar SNPs no genoma suíno que influenciam características produtiva e reprodutiva, utilizando informações provenientes de animais de um programa de melhoramento de suínos. Neste trabalho foram utilizados dados genotípicos e fenotípicos de animais da raça Landrace (LA) e Large White (LW). Primeiramente foi realizado o controle de qualidade de amostras e dos SNPs. Após esta etapa, as análises GWAS para as características idade ajustada aos 100 kg de peso corporal (ID100) e a idade ao primeiro parto (IPP) foram realizadas com dados da raça LA. No controle de qualidade foram removidas amostras com problemas de discordância de sexo, paternidade, animais duplicados, eficiência de genotipagem (*call rate*) < 90%, desvios da heterozigosidade de ±3 desvios-padrão, além de verificar a estratificação e subestrutura da população. Para o controle de qualidade dos SNPs, foram removidos aqueles que apresentaram eficiência de genotipagem < 0,98, frequência do menor alelo (MAF) <0,03, equilíbrio de Hardy-Weinberg (EHW) com X^2 $<10^{-6}$ e SNPs coincidentes. Estas análises foram realizadas por meio de rotinas elaboradas no programa R e no programa Plink. Após a remoção das amostras e dos SNPs que não passaram pelo controle de qualidade, restaram 604 amostras e 42.360 SNPs da raça LA e 345 amostras e 40.166 SNPs da raça LW. Para as análises de GWAS, por meio de marcadores únicos, foram utilizados dados fenotípicos previamente corrigidos para os efeitos de grupo contemporâneo para cada uma das características. O modelo incluiu os efeitos fixos da média e dos SNPs e os aleatórios infinitesimal e do resíduo. A matriz de parentesco continha

informações das 3 últimas gerações dos animais genotipados. As análises de GWAS foram realizadas no programa QXPak. Foram considerados três limiares de significância: um conservador (valor de $p = 5x10^{-7}$), um moderado (valor de $p = 5x10^{-5}$) e outro baseado no critério de Bonferroni (valor de $p = 1,18x10^{-6}$). Pela análise GWAS para a característica ID100 foram encontrados 22 SNPs significativos no cromossomo 1 e um SNP no cromossomo 4, considerando o limiar moderado. Para IPP foram encontrados dois SNPs significativos, sem posição definida no genoma. Para cada característica foi estimado o efeito aditivo do SNP mais significativo, estes foi de grande magnitude, de 2,22 dias para ID100 e de 4,69 dias para IPP, indicando grande potencial na inclusão destas informações na seleção. A análise de controle de qualidade foi efetiva na remoção de amostras e de SNPs com problemas de genotipagem. Futuras análises incluindo maior número de animais genotipados e a avaliação de metodologias adicionais permitirá detectar SNPs de menor efeito para as características estudadas.

ABSTRACT

LEITE, Carla Daniela Suguimoto, D.Sc., Universidade Federal de Viçosa, fevereiro de 2013. **Genome wide association studies for production and reproduction traits in swine.** Advisor: Ricardo Frederico Euclydes. Co-Advisor: Mônica Corrêa Ledur, Robledo de Almeida Torres and Antonio Policarpo Souza Carneiro.

The aim of the genome wide association studies (GWAS) is to identify chromosome regions associated with phenotypic traits, based in differences between allele frequencies of single nucleotide polymorphisms (SNPs). The aim of this study was to identify SNPs in the swine genomes that influence productive and reproductive traits, using information from animals from a breeding swine program. It was used genotypic and phenotypic data from Landrace (LA) and Large White (LW) animals. First, it was performed quality control of samples and SNPs. After this step, the GWAS analysis for age adjusted to 100 kg and age at first farrowing traits, were performed only with LA data. The quality control for samples removed in the cases of disagreement on sex or paternity, duplicate animals, call rate < 90%, deviations of heterozygosity of more than ± 3 standard deviations, and check if there is stratification and substructure in the population. For the quality control of the SNPs, were removed those that have call rate < 0.98, the minor allele frequency (MAF) < 0.03, Hardy-Weinberg equilibrium (HWE) with $\chi^2 < 10^{-6}$, and the SNPs that were coincident. The softwares R and Plink were used for this analysis. After removing samples and SNPs that were problematic, for the analysis remained 604 samples and 42,360 SNPs for LA and for LW there were 345 samples and 40,166 SNPs. For GWAS analysis, by single marker, phenotypic data were previously corrected for contemporary group effects for both traits. The model included as fixed effects the mean and the SNPs and as random infinitesimal effect and residue. The relationship matrix contains information of the last 3 generations of genotyped animals. It was considered 3 thresholds: a conservative (p-value = $5x10^{-7}$), a moderate (p-value = $5x10^{-5}$) and one based on Bonferroni test (p-value = $1.18x10^{-6}$). The GWAS analysis for age at 100 kg found 22 significant SNPs on

chromosome 1 and one SNP on chromosome 4, considering the moderate threshold. For age at first farrowing, two SNPs were significant, with no defined position in the genome. For both traits, the additive effect estimated was of large magnitude, 2.22 days for age at 100 kg and 4.69 days for age at first farrowing, indicating a great potential of the inclusion of this information in selection to reduce these traits. The quality control analysis was effective to remove samples and SNPs with problems in genotyping. Further analysis including more genotyped animals and evaluation of additional methodologies may detect SNPs with minor effect for traits of interest.

ÍNDICE

INTRODUÇÃO GERAL

Tradicionalmente, os programas de melhoramento genético têm realizado a seleção dos animais com base apenas em informações fenotípicas e de pedigree. No entanto, a incorporação de informações genotípicas na avaliação genética apresenta grande potencial para aumentar a acurácia de seleção, e assim, acelerar os ganhos genéticos no melhoramento animal.

A identificação de genes que controlam características de interesse econômico pode ser obtida em estudos de mapeamento genético, por meio da descoberta de polimorfismos que, recentemente, estão sendo utilizados como auxilio na seleção em programas de melhoramento (Wagner, 2011).

Dentre os estudos realizados para aumentar a eficiência desse procedimento, baseado em dados genotípicos, Lande e Thompson (1990) propuseram a seleção auxiliada por marcadores moleculares (MAS), que utiliza, simultaneamente, dados fenotípicos e dados de marcadores moleculares, que estão em ligação gênica com alguns *loci* controladores de características quantitativas (QTL).

Os avanços no sequenciamento do genoma e as tecnologias de genotipagem de alto rendimento, como a utilização dos SNPs (polimorfismos de base única) tornaram os estudos de associação global do genoma (GWAS) práticos para explorar a associação de genes com características complexas. As análises baseadas no GWAS apresentam como vantagem a detecção de QTLs mais próximos das mutações causais, quando comparado com o mapeamento de QTL por análise de ligação (Hirschhorn, 2005; Jiang et al., 2010).

Os marcadores SNPs estão amplamente distribuídos no genoma, representando grande parte da variação genética presente em todos os cromossomos. Outra vantagem desses marcadores é a facilidade em realizar a genotipagem de grande número de SNPs a baixo custo. Com o advento dos painéis de SNPs, estes têm sido utilizados na detecção e

localização de QTL para características complexas em muitas espécies, com grande utilidade na identificação de mutações casuais em características de importância econômica em animais, bem como em doenças humanas (Ledur et al., 2009).

Desta forma, dentre os objetivos da aplicação da genética molecular no melhoramento genético animal é identificar e mapear os genes que interferem na expressão de características quantitativas de importância econômica, visando melhorar a compreensão do controle genético de características complexas de interesse para a cadeia de produção animal (Rosa e Fragoso, 2010). Outro objetivo é o uso se marcadores tipo SNPs diretamente na seleção visando o aumento dos ganhos genéticos principalmente devido à melhoria da acurácia da seleção.

REVISÃO BIBLIOGRÁFICA

Melhoramento de suínos

A carne suína é a mais consumida no mundo, e estima-se que 103,78 milhões de toneladas foram consumidas e 104,357 milhões de toneladas foram produzidas em 2012 (USDA, 2012). Se os produtores de suínos pretendem suprir a demanda da população mundial, aproximadamente sete bilhões de habitantes, novas estratégias de produção deverão ser desenvolvidas (Wagner, 2011). Além disso, devido à similaridade fisiológica com o humano, o suíno tem sido utilizado em estudos de doenças em humanos.

O aumento da produção suinícola é suportado pelas melhorias no ambiente de produção e no melhoramento genético, que se baseia na seleção de indivíduos superiores, considerando o seu desempenho fenotípico, corrigido pelos efeitos ambientais e pelas informações de pedigree, sem considerar o número de genes que afetam cada característica ou o efeito de cada gene sobre as características de interesse (Fan et al. 2010; Rothschild, 2008).

Assim, apesar do sucesso obtido com a seleção baseada nos dados fenotípicos, é crescente o interesse em utilizar a informação molecular na seleção, especialmente para características de baixa herdabilidade, de difícil mensuração, que podem ser medidas em apenas um sexo, ou tardiamente na vida do animal, e também naquelas que necessitam do abate do animal para serem mensuradas.

O sequenciamento do genoma suíno e os diversos estudos baseados na informação molecular possibilitaram a identificação de marcadores do tipo SNP. Os estudos de associação global do genoma por meio da genotipagem de SNPs são ferramentas básicas para possibilitar a seleção genômica. Existem diversos Consórcios envolvidos no sequenciamento e estudos do genoma, como o Wellcome Trust Case Control Consortium (WTCCC, 2007) na área de humanos. Este consórcio, formado por 50 pesquisadores, tornou-se referência no estudo de características ligadas a doenças em humanos e nos avanços nas tecnologias de genotipagem de alto rendimento.

Na área de suínos, dentre os consórcios internacionais destacam-se o *Swine Genome Sequencing Consortium* (SGSC), um consórcio internacional com envolvimento de universidades, governo e indústria para obter o completo sequenciamento do genoma suíno. Esse consórcio desenvolveu o *Illumina Porcine SNP60K iSelect Beadchip* com 62.621 SNPs, o que tornou possível a realização dos estudos de GWAS e seleção genômica em suínos (Ramos et al., 2009).

Atualmente, as principais empresas de genética de suínos que possuem atuação mundial estão instaladas no Brasil, entretanto, funcionando em menor nível de eficiência que nos países de origem. Desta forma, os programas nacionais de melhoramento genético de suínos passam a ter um papel fundamental, no sentido de poderem continuar sendo competitivos frente às genéticas internacionais e assim, poder continuar o país independente em termos de material genético de suínos e principalmente, continuar concorrendo e avançando no mercado mundial de carnes.

No Brasil, recentemente foi estabelecida a parceria entre a rede pública de pesquisa, com a participação de diversas unidades da Embrapa, como Embrapa Suínos e Aves, Embrapa Gado de Leite, Embrapa Gado de Corte, Embrapa Pecuária Sul, Embrapa Informática Agropecuária e Embrapa Recursos Genéticos e Biotecnologia, universidades (UFMG, ESALQ/USP, Universidade Autônoma de Barcelona), o MAPA e a empresa privada nacional BRF. Este grupo vem desenvolvendo de forma conjunta um projeto que visa à identificação de genes ou marcadores (SNPs) associados a características de interesse econômico na produção comercial de suínos e sua utilização na seleção assistida em programa de melhoramento nacional. Iniciativas como esta têm como escopo maior os esforços para melhorar a competitividade dos programas de melhoramento de suínos nacionais, e a tentativa de evitar a dependência da suinocultura nacional de programas genéticos internacionais.

Análise de associação global do genoma

A utilização de análises moleculares envolvendo características quantitativas em programas de melhoramento genético pode ocorrer por meio de identificação de mutações causais ou de marcadores indiretos (Dekkers, 2002). Anderson (2001) ressaltou que as mutações causais de características quantitativas são difíceis de serem encontradas, difíceis de serem comprovadas e poucos exemplos são disponíveis. Com relação aos marcadores indiretos, ligados a QTL, são abundantes em todo o genoma e baseia-se na associação dos genótipos dos marcadores com o fenótipo da característica em uma determinada população (Dekkers, 2002).

O mapeamento de loci de característica quantitativa (QTL) por meio de microssatélites foi comumente utilizado para a detecção de variação genética para características de importância econômica. Zhang et al. (2012) realizaram revisão nos estudos de identificação de QTL para diferentes características em diversos animais, e no caso de suínos,

foram relatados 7.451 QTLs representando 600 características diferentes (http://www.animalgenome.org/QTLdb/). No entanto, os autores ressaltaram que a maioria destes QTLs foi detectada utilizando marcadores microssatélites com mapas de baixa resolução e intervalo de confiança (CI) contemplando mais de 20 cM. Desta forma, seria difícil detectar genes importantes para características de interesse baseada apenas nesta informação.

O desenvolvimento de arranjos de SNP (SNP; Single Nucleotide Polymorphism) de alta densidade possibilitou a realização de análises de associação global do genoma (GWAS, *genome wide association studies*) em diversas características de várias espécies de interesse, como suínos, bovinos e ovinos (Rosa e Fragoso 2010).

Por meio da análise GWAS pode-se correlacionar o genótipo de milhares de SNPs com fenótipos de interesse. Os primeiros estudos envolvendo a associação global do genoma, realizados em humanos, possibilitaram a identificação de *loci* que afetam várias doenças complexas, como câncer de próstata, transtorno bipolar, doença arterial coronariana, doença de Crohn, hipertensão, artrite reumatóide, diabetes tipo 1 e diabetes tipo 2 (Yeager et al., 2007; WTCCC, 2007) e características quantitativas, como a altura em humanos (Visscher, 2008).

Em contraste com os estudos em humanos, em que as análises têm o propósito de identificar marcadores para doenças, a aplicação dos marcadores SNP nos animais domésticos visa aumentar a acurácia na seleção para acelerar o melhoramento genético para características de interesse econômico (Fan et al., 2011). Assim, o desenvolvimento de painéis de SNPs de alta densidade para animais domésticos permite a detecção de QTLs para essas características através de estudos de associação global do genoma com maior acurácia que as análises tradicionais de ligação (Wagner, 2011).

De acordo com Pearson et al. (2008) a análise de GWAS pode ser realizada em quatro passos, dentre os quais estão: a coleta de informação fenotípica, garantindo a consistência da informação; a genotipagem e verificação da qualidade das amostras e dos SNPs

genotipados; a análise de associação entre os SNPs e a característica de interesse; e a replicação para identificar a associação em uma população independente.

O controle de qualidade da genotipagem é realizado anteriormente à análise de GWAS, para evitar que problemas na genotipagem ocasionem associações falso-positivas (Panoutsopoulou et al., 2009). Dentre os critérios utilizados no controle de qualidade das amostras pode-se verificar a eficiência de genotipagem (*call rate*), a heterozigosidade média, se há amostras idênticas que estão em duplicidade, a estratificação da população e a subestrutura da população. A checagem de discordâncias quanto ao sexo também é realizada no controle de qualidade das amostras. Com relação aos SNPs, o controle de qualidade leva em consideração a eficiência de genotipagem, os desvios com relação ao equilíbrio de Hardy-Weinberg, frequência do menor alelo e SNP idênticos. A remoção de amostras e SNPs problemáticos também reduzem o número de testes a serem realizados na análise de associação, reduzindo os esforços computacionais (Turner et al., 2011).

O método mais simples de realizar GWAS é a associação de um único marcador SNP, que considera o efeito de cada SNP individualmente. Essa metodologia utiliza testes de hipótese para detectar associação, estatisticamente significativa, entre os loci e o fenótipo, dentro da população (Resende et al., 2011). Outra metodologia é a associação por múltiplos marcadores, que se baseia na formação de haplótipos, que podem reduzir a detecção de marcadores falsos positivos e aumentar o poder da análise (Li et al., 2006).

Para identificar o efeito de um SNP ou marcador a metodologia mais simples é por meio de um modelo incluindo somente os efeitos fixos, no entanto este modelo não considera os dados de pedigree ou número de famílias, ignorando a covariância do parentesco entre os indivíduos. A aplicação desta metodologia pode resultar na ocorrência de associações falso positivas (Aulchenko et al., 2007). Uma alternativa seria o uso de modelos mistos, considerando o parentesco entre os indivíduos como um componente aleatório no modelo, podendo incluir também outro efeitos

aleatórios. No entanto esta análise quando comparada com o modelo linear simples é bem mais lenta.

Devido à ocorrência de resultados falsos positivos, outro ponto relevante na análise de GWAS é a seleção dos SNPs que de fato podem ser considerados como associações verdadeiras, sendo necessária a identificação de critérios para seleção destes SNPs (WTCCC, 2007).

Diversos métodos foram propostos para estabelecer um limiar de seleção de SNPs e minimizar os problemas de associações falsos positivas resultantes dos testes múltiplos. Entre estes métodos pode-se destacar a taxa de descoberta de falsos positivos (FDR, false discovery rate); o teste permutação; fatores Bayesianos (BF); e a correção de Bonferronni (Duggal et al., 2008). Em adição aos limiares estabelecidos com base no valor de p, é recomendado utilizar outros critérios como distância entre os SNPs e a correlação (r^2), baseada no desequilíbrio de ligação.

A taxa de descoberta de falso positivo (FDR) controla a proporção esperada de falsos positivos entre todas as rejeições, proporcionando um controle menos rigoroso do erro Tipo I.

A correção de Bonferroni pode usar o número real de testes realizados (como por exemplo os SNPs genotipados) ou um valor teórico baseado no número total de testes possíveis (todos os SNPs possíveis). No entanto, o método de correção de Bonferroni tem como pressuposição a independencia dos testes, o que biologicamente não ocorre no caso dos SNPs, pois SNPs próximos não estão independentes. Assim, quando se emprega método de correção de Bonferronni tradicional, estaria ocorrendo uma “hipercorreção” no limiar de significância da análise de associação global do genoma (Potkin et al., 2009).

Ledur et al. (2009) estabeleceram um limiar de seleção baseado na frequência de cada SNP no modelo, que corresponde ao fato de ser um QTL verdadeiro, chamado de bootstrap probabilities (BPP), considerando um limiar arbitrário de BPP> 0,25 para uma ocorrência de associação verdadeira.

O critério de limiar pode ser estabelecido com base no p-valor do Teste F, sendo os mais empregados os estabelecidos pelo WTCCC (2007), assumindo uma probabilidade de 0,05 e um limiar de $5x10^{-7}$, sendo considerado altamente conservador. Assim um limiar de p-valor é estabelecido com uma restrição de distância mínima entre os SNPs selecionados, ou seja, quando dois SNPs significativos estiverem numa região de menos de 5 cM, apenas o SNP mais significativo será considerado. No entanto, nem todos os SNPs selecionados são necessariamente QTL genuínos (WTCCC, 2007).

Estudos de Associação Global do Genoma em Suínos

Os estudos de GWAS em suínos começaram a se destacar em 2010, com o lançamento do chip desenvolvido pela llumina® PorcineSNP60 BeadChip, versão 1.

Na literatura, diversos estudos com o uso de análises GWAS em suínos podem ser citados, como por exemplo, os estudos em características qualitativas, como cor de pelagem (Ren et al., 2011; Wagner, 2011; Cho I-C et al., 2011) e em características quantitativas, envolvendo composição e conformação corporal (Fan et al., 2011), características reprodutivas em porcas, como tamanho de leitegada, duração da gestação, idade a puberdade (Onteru et al., 2012; Nonneman et al., 2011), longevidade da porca (Tart, 2012), infertilidade do macho (Sironen et al., 2010) e também em estudos de qualidade de carne em suínos (Luo et al., 2012).

Fan et al. (2011) foram os primeiros a publicarem a utilização de GWAS em características de interesse econômico em suínos. Neste estudo foram identificadas regiões cromossômicas associadas a características de crescimento, composição e conformação corporal e estrutura de perna e pés. Os autores encontram diversas regiões cromossômicas que correspondem a QTL que já haviam sido descritos anteriormente, como os genes candidatos MC4R (para espessura de toucinho) e IGF2 (para área do lombo), e também relataram novos genes

incluindo o CHCHD3 (para espessura de toucinho), BMP2 (para área de lombo, conformação corporal e estrutura de pernas e pés).

Em estudo para seleção de fêmeas suínas baseada em longevidade, Tart (2012) encontrou grande número de regiões identificadas por meio de análises GWAS para características reprodutivas em suínos, confirmando a natureza poligênica destas características.

Luo et al. (2012), analisando características de qualidade de carne em uma população cruzada F2 (Large White × Minzhu), utilizando análises GWAS, por meio de um modelo misto, que considerou a informação de parentesco entre os indivíduos, encontraram que 36 dos 45 SNPs que foram significativamente associados com características de qualidade de carne em suínos estavam localizados em uma região que anteriormente já havia sido relatada a presença de QTLs.

Nonneman et al. (2011) utilizaram GWAS para analisar a idade a puberdade em população Duroc-Landrace-Yorkshire e identificaram regiões genômicas que explicaram aproximadamente 5% da variação fenotípica para esta característica.

Onteru et al. (2012) utilizaram a análise de GWAS para avaliar o número de leitões nascidos totais, nascidos vivos, mumificados e natimortos e a duração da gestação, por três parições, em população de 683 fêmeas suínas de Large White x Landrace. Foram relatadas diferentes regiões cromossômicas entre as três primeiras parições para as características estudadas, evidenciando o efeito diferenciado, ao longo do tempo, no controle genético dessas características.

Os estudos envolvendo GWAS e chips de genotipagem de alta densidade têm demonstrado ser eficiente na identificação de SNPs associados a características de interesse, possibilitando a identificação de genes candidatos e, também melhorando o entendimento de mecanismos de regulação de características de interesse econômico (Caetano, 2009)

OBJETIVOS

Objetivo Geral

O objetivo nesse estudo foi identificar SNPs no genoma suíno que influenciam características produtiva e reprodutiva, utilizando informações provenientes de animais do programa de melhoramento de suínos da BRF S.A.

Objetivos específicos

Capítulo 1: Realizar a análise de controle de qualidade das amostras e dos SNPs, provenientes do banco de dados de genotipagem de amostras, para posterior análise de associação global do genoma.

Capítulo 2: Estimar efeito aditivo e selecionar SNPs, por meio de análise de associação global do genoma (GWAS), para as características idade para atingir os 100 kg e idade ao primeiro parto, considerando a análise de um único marcador por vez.

REFERÊNCIA BIBLIOGRÁFICA

ANDERSSON, L. Genetic dissection of phenotypic diversity in farm animals. **Nature Review Genetics***,* v. 2, p. 130–138. 2001.

AULCHENKO, Y.S.; KONING, D.; HALEY, C. Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. **Genetics,** v. 177, p. 577–585, 2007.

CAETANO, A.R. Marcadores SNP: conceitos básicos, aplicações no manejo e no melhoramento animal e perspectivas para o futuro. **Revista Brasileira de Zootecnia**, v. 38, p.64-71, 2009 (supl. especial)

CHO, I-C.; ZHONG, T.; SEO, B-Y; JUNG, E-J; YOO, C-K; KIM, J-H; LEE, J-B; LIM, H-T; KIM, B-W; LEE, J-H; KO, M.S.; JEON, J. Whole-genome association study for the roan coat color in an intercrossed pig population

between Landrace and Korean native pig. **Genes & Genomics**, v. 33, p.17-23. 2011.

DUGGAL, P.; GILLANDERS, E.M.; HOLMES, T.N.; BAILEY-WILSON, J.E. Establishing an adjusted p-value threshold to control the family-wide type 1 error in genome wide association studies. **BMC Genomics**, v. 9:516, 2008.

FAN, B.; ONTERU, S.K.; DU, Z.Q.; GARRICK, D.J.; STALDER, K.J. ROTHSCHILD, M.F. Genome-Wide association study identifies loci for body composition and structural soundness traits in pigs. **PloS One**, v.6, p. 1-11, 2011.

HIRSCHHORN, J.N., DALY, M.J. Genome-wide association studies for common diseases and complex traits. **Nature Reviews Genetics**, v. 6, p. 95–108. 2005.

JIANG, L.; LIU, J.; SUN, D.; MA, P.; DING, X.; YU, Y.; ZHANG, Q. Genome wide association studies for milk production traits in chinese Holstein population. **Plos One,** v. 5, e: 13661, 2010.

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

LEDUR, M.C.; NAVARRO, N.; PÉREZ-ENCISO, M. Data modeling as a main source of discrepancies in single and multiple marker association methods. **BMC Proceedings,** v.3, s.9, 2009.

LI, J.; ZHOU, Y.; ELSTON, R.C. Haplotype-based quantitative trait mapping using a clustering algorithm. **BMC Bioinformatics**, v.7, p.258. 2006.

LUO, W.; CHENG, D.; CHEN, S.; WANG, L.; LI, Y.; MA, X.; SONG, X.; LIU, X.; LI, W.; LIANG, J.; YAN, H.; ZHAO, K.; WANG, C.; WANG, L.; ZHANG, L. Genome-wide association analysis of meat quality traits in a porcine large white × minzhu intercross population. **International Journal of Biological Sciences**, v. 8, p. 580-595, 2012.

NONNEMAN, D. J., G. A. ROHRER, L. A. REMPEL, R. T. WIEDMANN AND J. L. VALLET. Genome-wide associations for age at puberty in a Duroc-Landrace-Yorkshire swine population. **Plant and Animal Genome Conference Proceedings.** 2011

ONTERU, S.K.; Fan, B.; Du, Z.Q.; GARRICK, D.J.; STALDER, K.J.; ROTHSCHILD, M.F. A whole-genome association study for pig reproductive traits. **Animal Genetics**, v. 43, p. 18-26, 2012.

PANOUTSOPOULOU, k.; ZEGGINI, E. Finding common susceptibility variants

for complex disease: past, present and future. **Briefings in Functional Genomics and Proteomics**, v. 8, p. 345-352, 2009.

PEARSON, T.A.; MANOLIO, T.A. How to interpret a genome-wide association study. **JAMA**, v. 299, p. 1335-1344, 2008.

POTKIN, S.G.; TURNER, J.A.; GUFFANTI, G.; LAKATOS, A.; TORRI, F.; KEATOR, D.B.; MACCIARDI, F. Genome-wide strategies for discovering genetic influences on cognition and cognitive disorders: Methodological considerations. **Cognitive Neuropsychiatry,** v. 14, p. 391-418, 2009.

RAMOS, A.M.;, CROOIJMANS, R.P.M.A.; AMARAL, A.J.; ARCHIBALD, A.L.; BEEVER, J.E.; BENDIXEN, C.; DEHAIS, P.; AFFARA, N.A.; HANSEN, M.S.; HEDEGAARD, J.; HU, Z-L.; KERSTENS, H.H.; LAW, A.S.; MEGENS, H.J.; MILAN, D.; NONNEMAN, D.J.; ROHRER, G.A.; ROTHSCHILD, M.F.; SMITH, T.P.L.; SCHNABEL, R.D.; VAN TASSELL, C.P.; CLARK, R.; CHURCHER, C.; TAYLOR, J.F.; WIEDMANN, R.T.; SCHOOK, L.B.; GROENEN, M.A.M: Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology. **PLoS ONE,**v. 4, p. e6524. 2009.

REN, J.; MAO, H.; ZHANG, Z.; XIAO, S.; DING, N.; HUANG, L. A 6-bp deletion in the TYRP1 gene causes the brown colouration phenotype in Chinese indigenous pigs. **Heredity**, v.106, p. 862-868, 2011.

RESENDE, M. D. V. de; SILVA, F. F. e; VIANA, J. M. S.; PETERNELLI, L. A.; RESENDE JÚNIOR, M. F. R.; DEL VALLE, P. M. Métodos estatísticos na seleção genômica ampla. **Documentos 219,** EMBRAPA - CNPF, 104 p., 2011.

ROSA, A. J. M.; FRAGOSO, R. R. **Análise Genômica em Bovinos**. Planaltina, DF: Embrapa Cerrados, 58f. 2010.

ROTHSCHILD, M. F. e PLASTOW, G. S. Impact of genomics in animal agriculture and opportunities for animal health. **Trends in Biotechnology**, v. 26, p. 21-25. 2008.

SIRONEN, A.; UIMARI, P.; NAGY, S.; PAKU, S.; ANDERSSON, M.; VILKKI, J. Knobbed acrosome defect is associated with a region containing the genes STK17b and HECW2 on porcine chromosome 15. **BMC Genomics**, v. 11, 699, 2010.

TART, J.K. **Genomic Analysis of Characteristics in Swine Contributing to Sow Longevity.** 2012. 73f. *Thesis (Master of Science in Animal Science) - University of Nebraska-Lincoln*, Lincoln, 2012.

The Wellcome Trust Case Control Consortium: Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. **Nature**, v. 447, p.661-678, 2007.

TURNER, S.; ARMSTRONG, L.L.; BRADFORD,Y.; CARLSON, C.S.; CRAWFORD, D.C.; CRENSHAW, A.T.; ANDRADE, M.; DOHENY, K.F.; HAINES, J.L.; HAYES, G.; JARVIK, G.; JIANG, L.; KULLO, I.J.; LI, R.; LING, H.; MANOLIO, T.A.; MATSUMOTO, M.; MCCARTY, C.A.; MCDAVID, A.N.; MIREL, D.B.; PASCHALL, J.E.; PUGH, E.W.; RASMUSSEN, L.V.; WILKE, R.A.; ZUVICH, R.L.; RITCHIE, M.D. Quality control procedures for genome wide association studies. **Current Protocols Human Genetics**, v. 1.19, 2011.

USDA/FAS, 2012. Production, Supply and Demand Online - Downloadable Data Sets. Disponível em: http://www.fas.usda.gov/psdonline/psdHome.aspx.

VISSCHER, P. M.; HILL, W. G.; WRAY, N. R. Heritability in the genomics era - concepts and misconceptions. **Nature Reviews Genetics**, v. 9, p. 255-265, 2008.

WAGNER, E.K. **Mapping phenotypic traits in swine**. 2011. 118f. Dissertation (Doctor of Philosophy in Animal Science) – University of Illinois, Ilinois, 2011.

YEAGER, M.; ORR, N.; HAYES, R.B. et al.: Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. **Nat. Genet.,** v.39, p.645-649. 2007.

ZHANG, H.; WANG, Z.; WANG, S. E LI, H. Progress of genome wide association study in domestic animals. **Journal of Animal Science and Biotechnology**, v. 3, p.26. 2012.

Capítulo 1. APLICAÇÃO DO CONTROLE DE QUALIDADE PARA AMOSTRAS E SNPS EM DADOS DE GENOTIPAGEM DE DUAS LINHAGENS DE SUÍNOS

Resumo: Para a realização de análises de associação global do genoma mais acuradas primeiramente é necessário realizar o controle de qualidade das amostras e dos SNPs. Foram utilizados dados genotípicos de animais da raça Landrace (LA) e Large White (LW) para a realização destes controles. No controle de qualidade das amostras foram removidas amostras com problemas de discordância de sexo, paternidade, animais duplicados, eficiência de genotipagem (*call rate*) < 90%, desvios da heterozigosidade de ±3 desvios-padrão, além se verificar a estratificação e subestrutura da população. Para o controle de qualidade dos SNPs, foram removidos aqueles que apresentaram eficiência de genotipagem < 0,98, frequência do menor alelo (MAF) < 0,03, equilíbrio de Hardy-Weinberg (EHW) com $X^2 < 10^{-6}$ e SNPs coincidentes. No início da análise o banco de dados da raça LA continha 645 amostras enquanto da LW 355 amostras. Para ambas as raças utilizou-se um SNP chip contendo 61.643 SNPs. As rotinas para avaliação dos critérios de qualidade foram elaboradas no programa R e no programa Plink. Na raça LA foram genotipadas 606 amostras, enquanto na LW foram genotipadas 347 amostras. Após a genotipagem, o primeiro passo foi verificar a ocorrência de SNPs que falharam em todos os animais, sendo removidos 2.422 SNPs, em ambas as raças, LA e LW. Após a exclusão destes, a eficiência de genotipagem (*call rate*) foi de 99%, restando 59.741 SNPs, que foram considerados para as demais análises de controle de qualidade. Pelos testes controle de qualidade que removeram as amostras e os SNPs, restaram 604 amostras raça Landrace e 42.360 SNPs, para Large White restaram 345 amostras e 40.116 SNPs. Estas remoções tanto de amostras, como de SNPs proporcionará a realização de futuras análises, como a de análise de associação global do genoma, de forma mais acurada.

APPLICATION OF QUALITY CONTROL FOR SNPs AND SAMPLE GENOTYPED DATA FOR TWO BREEDS OFSWINE

Abstract: For more accurate analysis of genome wide association studies is necessary first to perform the quality control of samples and SNPs. It was used genotype data from Landrace (LA) and Large White (LW) animals. The quality control for samples removed in the cases of disagreement on sex or paternity, duplicate animals, call rate <90%, deviations of heterozygosity of more than ± 3 standard deviations, and check if there is stratification and substructure in the population. For the quality control of the SNPs, were removed those that have call rate < 0.98, the minor allele frequency (MAF) < 0.03, Hardy-Weinberg equilibrium (HWE) with $\chi^2 < 10^{-6}$, and the SNPs that were coincident. The routines for evaluation of the quality criteria were developed in the R and Plink softwares. It was genotyped 606 samples for LA, and 347 for LW. The genotyped data was analyzed to verify the occurrence of SNPs that failed in all animals, being removed 2,422 SNPs in both breeds, LA and LW. After excluding these SNPs, the call rate was 99%, and remaining 59,741 SNPs for further quality control analysis. By quality control tests it was removed the samples and SNPs, resulting in 604 samples and 42,360 SNPs for Landrace, and 345 samples and 40,116 SNPs for Large White. These deletions of both samples and SNPs may provide conducting further analysis, such as genome wide association more, accurate.

INTRODUÇÃO

O objetivo da análise de associação global do genoma (GWAS, *genome wide association studies*) é identificar regiões cromossômicas associadas com características fenotípicas de interesse. Para isso, as diferenças entre os efeitos de substituição alélica dos polimorfismos de base única (SNP, *single nucleotide polymorphism*) podem ser usadas para identificar estas associações.

Nas análises de GWAS vários testes estatísticos são realizados para reduzir as chances de uma falsa associação, porém como estas análises envolvem um grande número de marcadores, qualquer problema relacionado à genotipagem pode levar a associações falso positivas ou negativas (Zanella, 2011).

Para a realização de análises de associações mais acuradas, Anderson et al. (2010) destacaram como grande desafio a qualidade dos dados genotípicos, ou seja, provenientes de amostras de boa qualidade e sem erros de genotipagem.

Ao considerar um teste de associação com 60.000 SNPs, tendo uma proporção de 0,001 % dos SNPs falhando na genotipagem, caso não seja aplicado um controle de qualidade, a genotipagem poderia indicar a falsa associação de 60 SNPs. A ocorrência de resultados de associações falsos positivas ocasionam perdas em custo de tempo e financeiro na replicação do estudo (Turner et al., 2011).

Desta forma, para reduzir a ocorrência de associações falsas, devido a problemas na genotipagem, é necessário a realização do controle de qualidade das amostras e dos SNPs, antes das análises de GWAS. Outro aspecto bastante importante é a necessidade de avaliar a consistência e acurácia dos dados fenotípicos, pois problemas neste tipo de informação podem levar a associações viesadas.

O controle de qualidade baseia-se na remoção de amostras e SNPs com resultados duvidosos por meio de critérios adotados para excluir aqueles que não satisfazem os limiares estabelecidos e que possivelmente irão prejudicar os resultados reais da análise. O limiar empregado pode variar com o tipo de estudo, a característica avaliada e o quão conservativo serão os critérios definidos para avaliação, na busca do equilíbrio entre minimizar o número de amostras retiradas e maximizar a eficiência de genotipagem (Turner et al., 2011).

Controle de qualidade das amostras (por indivíduo)

O controle de qualidade por amostra (ou indivíduo) consiste em diversos critérios para detectar amostras que apresentam problemas na

genotipagem, dentre os quais podem ser destacados: discordância do sexo, eficiência da genotipagem (*call rate*), desvios da hetorozigosidade, animais duplicados, indivíduos com divergência de ancestrais e a estratificação da população.

As amostras podem ser removidas se apresentarem uma identidade duvidosa (erros no apontamento do sexo do indivíduo) ou qualidade questionável (estar aparentemente contaminada) (Laurie et al., 2011). A taxa de eficiência de genotipagem e de heterozigosidade são indicativos de baixa qualidade da amostra de DNA (Anderson et al., 2010), podendo também a heterozigosidade indicar problemas de contaminação da amostra.

As falhas na genotipagem e a ausência de genotipagem podem ser decorrentes da coleta e manipulação de amostras, e esta baixa qualidade das amostras pode estar relacionada à degradação do DNA, baixa concentração de DNA e altos níveis de contaminação por RNA ou proteínas (Huentelman et al., 2005; Settles et al., 2009).

Discordância de sexo

Uma das formas mais simples para se verificar a ocorrência de problemas com a manipulação das amostras é na identificação dos sexos, comparando se o sexo relatado confere com o previsto pela genotipagem (Turner et al., 2011). Pelo fato dos machos apresentarem apenas um cromossomo X, os testes para avaliar possíveis casos de discordância no sexo são baseados na taxa de heterozigosidade do cromossomo X.

Eficiência da genotipagem (Call rate)

O critério de qualidade de eficiência da genotipagem ou *call rate* indica a qualidade de DNA das amostras. Se uma grande proporção de SNPs falharem na genotipagem, este pode ser um indicativo de problemas com a amostra de DNA, e estas devem ser eliminadas das análises (Turner et al., 2011; Anderson et al., 2010). A eficiência de genotipagem é um bom indicador do desempenho da hibridização durante

o processo de genotipagem. Problemas durante esta etapa podem diminuir essa eficiência (Panoutsopoulou e Zeggini, 2009).

O limiar de eficiência de genotipagem considera que amostras com uma taxa de falha de genotipagem superior a 10% devem ser excluídas das análises por causa do aumento na possibilidade de gerar erros de associações (Anderson et al., 2010; WTCCC, 2007).

Heterozigosidade média

A distribuição da heterozigosidade média (excluindo os cromossomos sexuais) em todas as amostras é investigada para identificar indivíduos com proporções excessivas ou reduzidas de genótipos heterozigotos, podendo ser indicativo, respectivamente, de contaminação da amostra ou de endogamia (Anderson et al., 2010).

O cálculo da heterozigosidade é dado pela fórmula *(N-O) /N*, em que *N* é o número de genótipos esperados; e *O* é o número observado de genótipos homozigotos para dado individuo (Anderson et al., 2010). Pelo fato da heterozigosidade média ser relativa à população em estudo, considera-se que indivíduos com heterozigosidade da ordem de 3 desvios-padrão acima ou abaixo da média devem ser excluídos da análise (Anderson et al., 2010).

Estratificação da População

A estratificação da população pode ocorrer quando amostras em estudos envolvem vários grupos de indivíduos que diferem sistematicamente na composição genética e no fenótipo sob estudo (Turner et al., 2011).

A avaliação da estratificação da população tem a finalidade de verificar se os animais dividem a mesma composição genética e pode ser observada pela informação da distância de pares idênticos por estado (IBS, *identity-by-state*) entre todos os indivíduos, baseada na proporção média entre os alelos compartilhados em comum entre todos os SNPs genotipados (excluindo os cromossomos sexuais) (Anderson et al., 2010).

Considerando tanto análise de associação envolvendo estudo caso – controle ou características quantitativas, se não considerar a estrutura da população, a ocorrência de associações falso positivas podem ocorrer devido a diferenças na frequência alélica entre as diferentes populações.

O gráfico de escalonamento multidimensional (*multidimensional scaling*, MDS) mensura a similaridade dos alelos entre os loci independentes dentre todos os animais (Zanella, 2011). Para realizar este teste é necessário utilizar um grupo de SNPs que não estejam em desequilíbrio de ligação, considerando $r^2 < 0,02$, ou seja, SNPs independentes, de modo a não gerar viés na análise.

Controle de qualidade dos SNPS (por marcador)

O controle de qualidade dos SNPs dos dados utilizados em análises GWAS baseia-se na remoção de SNPs com baixa eficiência de genotipagem, baixa frequência do menor alelo (MAF) e SNPs que desviam significativamente do Equilíbrio de Hardy-Weinberg. No entanto, os critérios utilizados para filtrar os SNPs de baixa qualidade podem variar entre estudos. Anderson et al. (2010) destacam que no caso de estudos envolvendo características ligadas a doenças, deve-se tomar muito cuidado para remover apenas os SNPs mal caracterizados, pois cada marcador removido pode ser potencialmente, uma variante da doença.

Eficiência de genotipagem (Call Rate)

A eficiência de genotipagem (*call rate*) é igual ao número de SNP que receberam genótipo heterozigoto ou homozigoto dividido pelo número total de loci observados, conforme descrito por Zanella (2011). Os SNPs que falharam em um grande número de amostras podem gerar resultados de associações falso-positivas (Turner et al., 2011) ou podem indicar polimorfismo em torno do SNP investigado (Huentelman et al. 2005).

O limiar recomendado para remover SNPs com baixas taxas de genotipagem é de aproximadamente 98-99% (Turner et al., 2011).

Equilíbrio Hardy-Weinberg

A maioria das análises de GWAS excluem marcadores com desvios do equilíbrio de Hardy-Weinberg, pois podem ser um indicativo de erro de genotipagem (Anderson et al., 2010). Pelo principio do equilíbrio de Hardy-Weinberg (EHW) é esperado, que em uma população grande, sob acasalamento ao acaso, as freqüências alélicas e genotípicas mantenham-se estabilizadas ao longo das gerações. Pelas suposições do Equilíbrio de Hardy-Weinberg espera-se que as freqüências genotípicas de uma geração para outra seja: $p^2 + 2pq + q^2$, sendo *p* a frequência dos alelos dominantes e *q* dos recessivos (Deng et al., 2000).

O desvio do EHW, por ser uma medida robusta que cobre todo o genoma, pode indicar, além dos erros de genotipagem, a estratificação da população e contaminação da amostra (Zanella, 2011). A ocorrência do desvio no equilíbrio também pode ser um indício de seleção imposta nos animais em estudo (Turner et al., 2011).

Os limiares adotados para assumir que o SNP está em EHW podem variar entre os estudos e, na literatura, estes valores de p variam entre 0,001 e $5,7X10^{-7}$ (Anderson et al., 2010).

Frequência do Alelo Menor

Outro passo no controle de qualidade é filtrar SNPs baseados na freqüência do alelo menor (MAF, *minor allele frequency*) e SNPs monomórficos. A remoção destes SNPs da análise diminui a correção pelos múltiplos testes realizados na análise de associação e reduz os erros de associações (Turner et al., 2011).

O limiar escolhido depende do tamanho da amostra em estudo e do tamanho do efeito esperado (Turner et al., 2011), bem como pode variar com a característica em estudo. Geralmente o valor considerado para a remoção dos SNPs pode variar entre 1-2%, entretanto, em estudos com poucas amostras pode ser necessário limiares maiores (Anderson et al., 2010).

OBJETIVO

O objetivo deste estudo foi realizar a análise de controle de qualidade das amostras e dos SNPs provenientes do banco de dados de genotipagem de amostras oriundas do Programa de Melhoramento Genético de Suínos da empresa BRF, para posterior análise de associação global do genoma para características de interesse econômico.

MATERIAL E MÉTODOS

Foram coletadas 1000 amostras de tecido de suínos (amostra de cauda ou orelha) de duas linhas fêmeas, sendo 645 amostras oriundas da raça Landrace (LA) e 355 amostras da Large White (LW), pertencentes ao programa de melhoramento genético de suínos da BRF.

O preparo e a extração de DNA dessas amostras foram realizados na Embrapa Suínos e Aves (CNPSA) com kit *PureLinkTM Genomic DNA* (Invitrogen) para extração de DNA de tecido. A qualidade e integridade do DNA foram avaliadas em espectrofotômetro (*NanoDrop*). Alíquotas de aproximadamente 800 ng foram desidratadas em estufa a 37º e enviadas à prestadora de serviço *Geneseek*, EUA, para a realização da genotipagem.

As amostras foram genotipadas pelo *PorcineSNP60kBeadchip*, que contem 64.232 marcadores SNPs. No entanto, deste total, 62.163 SNPs são marcadores informativos e utilizados na genotipagem. O chip utilizado foi a versão 1, com um espaçamento em média de 28 kb.

Os resultados da genotipagem foram enviados a Embrapa Informática Agropecuária, responsável pelo recebimento dos dados do projeto e também pelo desenvolvimento do sistema integrado de dados genotípicos, fenotípicos e de pedigree.

Os arquivos de genótipos foram então extraídos para a realização do controle de qualidade dos dados genotípicos. Os dados genotípicos

passaram pela análise de controle de qualidade das amostras e dos SNPs, considerando os critérios descritos abaixo:

Controle de qualidade das amostras: foram verificados erros com relação a: a) discordância do sexo; b) possibilidade de ocorrência de animais idênticos ou quase idênticos, por possuírem o genótipo com similaridade > 99.995% - neste caso, para descartar a possibilidade de gêmeos, foi verificado se os animais tinham os mesmos pais; c) eficiência de genotipagem (*Call rate*), em que foram removidas as amostras abaixo de 90% de *call rate*; d) desvios da Heterozigosidade, pelo qual foram removidas amostras com ±3 desvios padrão; e) estratificação da população, verificada por meio do gráfico do MDS.

Controle de qualidade dos SNPs: foram removidos os SNPs que apresentaram eficiência de genotipagem (*Call rate*) menor que 98%; frequência do menor alelo (MAF) menor que 3%; Equilíbrio de Hardy-Weinberg (EHW) com X^2 menor que 10^{-6}; e SNPs coincidentes, ou seja, dentre os SNPs que apresentaram a mesma posição foram retirados aqueles que apresentaram menor MAF.

Para as análises de critério de qualidade foram utilizadas uma rotina em linguagem R desenvolvida em conjunto com a Embrapa Informática Agropecuária e a Embrapa Suínos e Aves. Para a análise de estratificação da população foi utilizado o programa Plink, versão 1.07 (Purcell et al., 2007), sendo o gráfico gerado em R.

RESULTADOS E DISCUSSÃO

No total foram coletadas e enviadas para genotipagem 1000 amostras, da orelha ou cauda, que seriam pertencentes a 1000 animais. No entanto, 47 foram previamente descartadas, sendo 39 Landrace e 8 Large White, por apresentarem erro na identificação das amostras ou por terem sido coletadas duas vezes, e não foram genotipadas. Desta forma,

na raça Landrace (LA) foram genotipadas 606 amostras, enquanto na Large White (LW) foram genotipadas 347 amostras.

Após a genotipagem, o primeiro passo foi verificar a ocorrência de SNPs que falharam em todos os animais, sendo removidos 2.422 SNPs, em ambas as raças, LA e LW. Após a exclusão destes, a eficiência de genotipagem (*call rate*) foi de 99%, para cada raça, restando 59.741 SNPs, que foram considerados para as demais análises de controle de qualidade.

Para a análise da qualidade das amostras foram utilizados os critérios de qualidade como a eficiência de genotipagem (*call rate*), a hetorozigosidade média e os indivíduos idênticos, sendo removidas do banco de dados as amostras que apresentaram problemas em pelo menos um dos critérios. Os resultados destas análises estão descritos na Tabela 1

.

Tabela 1. Número de amostras removidas pelo controle de qualidade das amostras nas populações Landrace (LA) e Large White (LW).

	LA	LW
Total de amostras (início)	606	347
Eficiência de genotipagem	2	1
Desvio Heterozigosidade	2	1
Indivíduos idênticos	-	-
Amostras Removidas (baixa eficiência + desvio da heterozigosidade + indivíduos idênticos)	2	1
Total de amostras restante	604	346

Na análise da eficiência de genotipagem (*call rate*) das amostras foram observadas 2 amostras da raça LA e 1 na raça LW que apresentaram *call rate* menor que 90% e foram removidas do banco de dados. As mesmas amostras também foram removidas por estarem com mais de 3 desvios padrão da heterozigosidade média. Neste controle não foram encontrados indivíduos idênticos.

A remoção de amostras está diretamente ligada a qualidade e integridade do DNA e pode ser afetada por diversas causas como a coleta do tecido, condições de armazenamento, o tempo de coleta, o tipo de tecido coletado, o método de extração do DNA, incluindo também os fatores que afetam o processo de genotipagem, como a qualidade do lote do reagente e os instrumentos utilizados (Laurie et al., 2010). Neste trabalho, poucas amostras foram excluídas, o que indica a boa qualidade das amostras de DNA enviadas à análise, ou seja, a integridade e quantidade de DNA foram suficientes para uma boa genotipagem.

O próximo passo foi realizar o controle de qualidade dos marcadores SNPs, em que foram utilizados os critérios: eficiência de genotipagem, frequência do alelo menor (MAF), desvio do Equilíbrio de Hardy-Weinberg (EHW) e SNPs na mesma posição (coordenada genômica). Os SNPs foram removidos quando se considerou que a frequência do alelo menor (MAF) foi menor que 3%, eficiência de genotipagem de 98% ou se os SNP não foram significativos no teste de Equilibrio de Hardy-Weinberg ($\chi^2 > 10^{-6}$) e podem ser observados na Tabela 2.

Tabela 2. Número de SNPs excluídos pelo controle de qualidade dos dados de genotipagem realizados nas populações Landrace (LA) e Large White (LW).

	LA	LW
Total de SNPs (Início)	59.741	59.741
Eficiência de genotipagem	1.350	1.361
MAF	16.151	18.417
Desvio EHW	154	112
SNPs coincidentes	16	16
Total de SNPs Removidos (MAF + Desvio EWH + SNPs coincidentes)*	17.381*	19.625*
SNPs utilizados na análise	42.360	40.116

MAF: frequência do Alelo Menor; EWH: Equilibrio Hardy-Weinberg.
* Os mesmos SNPs podem ter sido removidos por diferentes critérios.

Pelo critério de eficiência de genotipagem, foram removidos SNPs que apresentaram *call rate* inferior a 98%, ou seja, por apresentarem perda de informação de genótipo em mais de 2% dos indivíduos analisados. A retirada destes SNPs tem como objetivo reduzir a ocorrência de resultados falsos positivos na análise de associação.

Os SNPs com MAF inferior a 3% foram removidos por serem monomórficos ou por estarem em uma frequência muito baixa, o que pode significar erro de genotipagem. Dentre os 16.151 SNPs removidos na raça LA por apresentar frequência do alelo menor abaixo de 3%, 12.591 SNPs eram monomórficos, ou seja, estavam fixados na população, e por isso não são informativos. Na raça LW, dos 18.417 SNPs removidos pelo mesmo critério, 15.730 eram monomórficos. O maior número de SNPs monomórficos na raça LW pode ser um indício de uma população mantida com tamanho efetivo da população menor e/ou submetida à maior intensidade de seleção, apresentando menor variabilidade em relação a LA.

Para o equilíbrio de Hardy-Weinberg, foram excluídos SNPs com X^2 inferior a $1x10^{-6}$, pois desvios extremos podem indicar problemas na precisão da genotipagem.

Em relação ao critério de remoção de SNPs por apresentarem a mesma posição, ou seja, a mesma coordenada genômica, 16 SNPs foram removidos em cada raça estuda. Dentre os SNPs que localizados na mesma coordenada genômica, foram removidos os que apresentaram menores MAF. No entanto, este critério não é aplicado aos SNPs do cromossomo “0”, pois este cromossomo contempla os SNPs que ainda não foram mapeados no genoma suíno e não possuem posição definida.

O número total de SNPs removidos considerando aqueles abaixo dos limiares indicados para MAF, desvio do EHW e SNPs coincidentes, foi de 17.381 SNPs (27%) para raça Landrace e 19.625 (32%) para a Large White. De acordo com Ziegler et al. (2009), SNPs genotipados de forma inconsistente ou que não contribuirão para a acurácia das avaliações

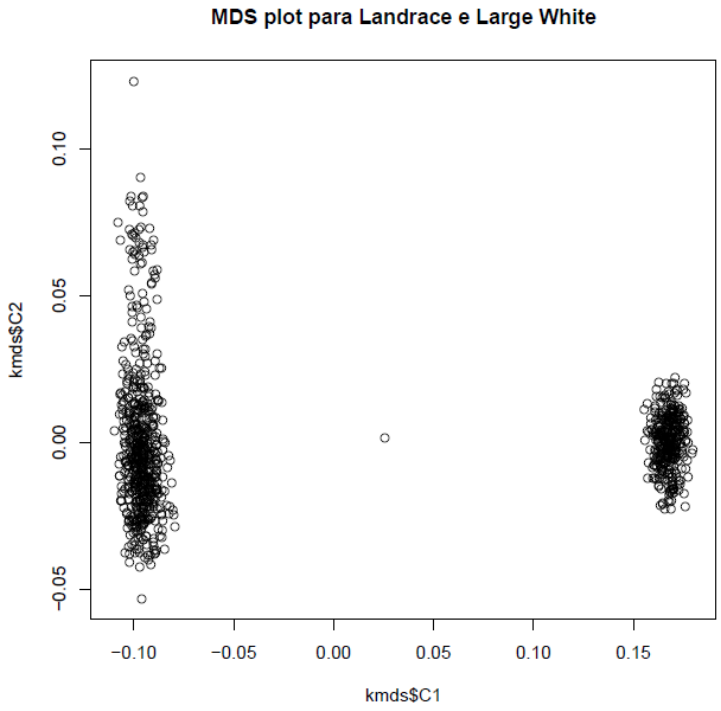
genéticas devem ser excluídos no sentido de reduzir esforços computacionais, diminuir a quantidade de resultados falsos e melhorar a precisão das análises realizadas com os polimorfismos restantes.

Onteru et al. (2011), analisando características reprodutivas em porcas, utilizaram um limiar de MAF > 0,001, eficiência de genotipagem >80% e X^2 < 0,0001 para verificar o EHW para exclusão de SNP com problemas de genotipagem. O Gentrain Score, utilizado nesta análise, baseia-se na análise de agrupamento dos genótipos, que são realizadas por meio do *Gen Call score* da Illumina. Após este controle de qualidade foram removidos 6.814 SNPs (11%) do banco de dados genotípico. Estes autores foram menos rigorosos com os limiares estabelecidos no critério de qualidade, considerando um maior número de SNPs nas posteriores análises de associação.

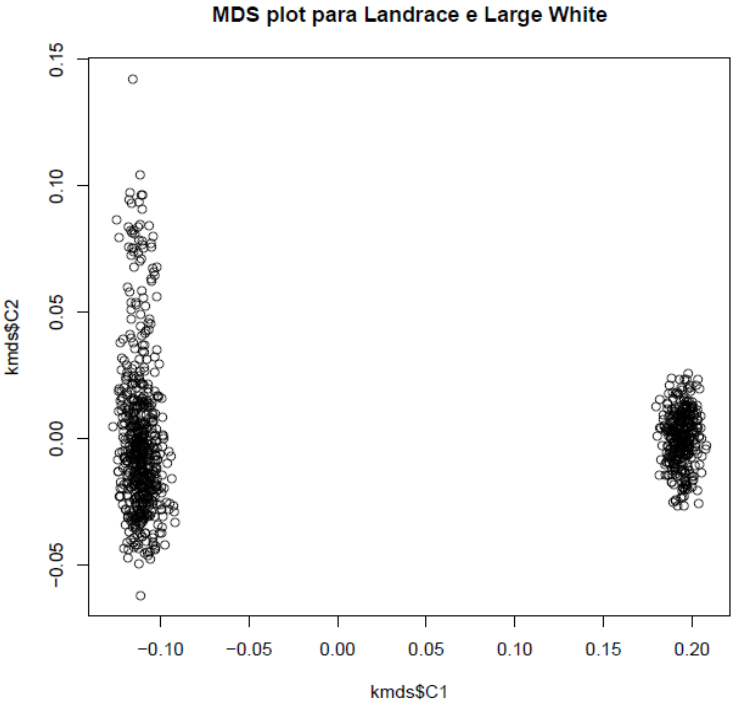
Fan et al. (2011), avaliando características de composição corporal em suínos, utilizaram um limiar de MAF >0,05, eficiência de genotipagem > 80% e p-valor < $1x10^{-6}$ no teste de X^2 para verificar o EHW para exclusão de SNP. Neste estudo, primeiramente, foram excluídos 5 amostras por apresentarem eficiência de genotipagem menor que 0,80 e no controle de qualidade por SNPs foram removidos 2.286 com eficiência de genotipagem inferior a 90% e por falharem na genotipagem (*no call*). Os autores consideraram 51.385 SNPs no banco de dados para as análises de associação.

Assim, observou-se que os valores de limiar adotados para os critérios de qualidade variaram entre os estudos, no intuito de reduzir a ocorrência de resultados falsos positivos nos posteriores estudos de associação. A adoção de determinado limiar para a remoção, especialmente de SNP, deve considerar a característica em estudo. Outro aspecto relevante a ser considerado em uma análise de controle de qualidade, cujo objetivo é melhorar a consistência dos dados genotípicos para posterior análise de associação, é avaliar a estratificação da população, que pode ser verificada por meio do gráfico da técnica multivariada de escalonamento multidimensional (MDS), considerando todos os SNPs (Gráfico 1A) e após a remoção dos SNPs pelo controle de

qualidade, em que foram removidos os SNPs que apresentaram baixa MAF e baixa eficiência de genotipagem (Gráfico 1B).



(A)



(B)

Gráfico 1. Gráfico MDS plot das populações Landrace e Large White, antes do critério de qualidade (A) e depois (B).

Antes do controle de qualidade, observou-se a presença de uma amostra que estaria entre as duas populações, podendo indicar o acasalamento entre as duas populações, ou contaminação entre amostras. No entanto, após a análise pelos critérios de qualidade, ocorreu que esta amostra foi eliminada, indicando possivelmente um erro de genotipagem.

As duas populações em estudo pertencem a duas raças distintas, desta forma há a formação de dois grupos distintos no gráfico 1B, evidenciando a diferença genética entre as populações.

Outro aspecto relevante é com relação à variabilidade apresentada dentro de cada população e este fato pode ser evidenciado pelo calculo do coeficiente de endogamia, que considerou o número de genótipos homozigotos observados versus esperados (Purcell et al., 2007). Neste caso a população da raça Landrace apresentou coeficiente de endogamia em média de -0,021, enquanto para a população da raça Large White foi de -0,025. Desta forma, na raça LW houve maior remoção de SNPs pela frequência do menor alelo (Tabela 2) em relação a Landrace, uma vez que a LW seria mais endogâmica, tendo mais alelos fixados, conforme resultados do Gráfico 1.

CONCLUSÃO

Neste capítulo foram apresentados os resultados obtidos no controle de qualidade elaborado pelos pesquisadores da EMBRAPA Informática Agropecuária e EMBRAPA Suínos e Aves. O controle de qualidade utilizado evidenciou que houve algumas amostras problemáticas e possíveis erros na genotipagem dos SNPs de duas raças, Landrace e Large White pertencentes ao programa de melhoramento da BRF. A remoção tanto de amostras, como de SNPs proporcionará a realização de futuras análises, como a de análise de associação global do genoma, de forma mai acurada.

REFERÊNCIAS BIBLIOGRÁFICA

ANDERSON, C.A.; PETTERSSON F.H.; CLARKE, G.M.; CARDON, L.R.; MORRIS, P.; ZONDERVAN, K.T. Data quality control in genetic case-control association studies. **Nature Protocols**, v. 5, p.1564–1573, 2010.

FAN, B.; ONTERU, S.K.; DU, Z.Q.; GARRICK, D.J.; STALDER, K.J.; ROTHSCHILD, M.F. Genome-wide association study identifies Loci for body composition and structural soundness traits in pigs. **PLoS One**, v. 6, e14726, 2011.

LAURIE, C.C.; DOHENY, K.F.; MIREL, D.B.; PUGH, E.W.; BIERUT, L.J.; BHANGALE, T.; BOEHM, F.; CAPORASO, N.E.; CORNELIS, M.C.; EDENBERG, H.J.; GABRIEL, S.B.; HARRIS, E.L.; HU, F.B.; JACOBS, K.B.; KRAFT, P.; LANDI, M.T.; LUMLEY, T.; MANOLIO, TA.; MCHUGH, C.; PAINTER, I.; PASCHALL, J.; RICE, J.P.; RICE, K.M.; ZHENG, X.; WEIR, B.S.; GENEVA INVESTIGATORS. Quality control and quality assurance in genotypic data for genome-wide association studies. **Genetic Epidemiology**, v. 34, p. 591-602, 2010.

ONTERU, S.K.; FAN, B.; NIKKILA, M.T.; GARRICK, D.J.; STALDER, K.J.; ROTHSCHILD, M.F. Whole-genome association analyses for lifetime reproductive traits in the pig. **Journal of Animal Science**, v. **89,** p. 988-995, 2011.

PANOUTSOPOULOU, k.; ZEGGINI, E. Finding common susceptibility variants for complex disease: past, present and future. **Briefings in Functional Genomics and Proteomics**, v. 8, p. 345-352, 2009.

PURCELL, S., NEALE, B.; TODD-BROWN,K.; THOMAS, L.; FERREIRA, M.A.; BENDER, D.; MALLER, J.; SKLAR, P.; DE BAKKER, P.I.; DALY, M.J.; SHAM, P.C. PLINK: a tool set for whole-genome association and population-based linkage analyses. **American Journal of Human Genetics**, v. 81, p. 559-575, 2007.

R Development Core Team. **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. Versão 2.15.2. Disponível em <http://www.R-project.org>, 2012. 60p.

SETTLES, M., ZANELLA, R.; MCKAY, S. D.; SCHNABEL, R. D.; TAYLOR, J. F.; WHITLOCK, R.; SCHUKKEN, Y.; VAN KESSEL, J. S.; SMITH, J. M.; NEIBERGS; H. A whole genome association analysis identifies loci associated with Mycobacterium avium subsp. paratuberculosis infection status in US Holstein cattle**. Animal Genetics**, v. 40, p. 655-662. 2009.

The Wellcome Trust Case Control Consortium: Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. **Nature**, v. 447, p.661-678, 2007.

TURNER, S.; ARMSTRONG, L.L.; BRADFORD, Y.; CARLSON, C.S.; CRAWFORD, D.C.; CRENSHAW, A.T.; ANDRADE, M.; DOHENY, K.F.; HAINES, J.L.; HAYES, G.; JARVIK, G.; JIANG, L.; KULLO, I.J.; LI, R.; LING, H.; MANOLIO, T.A.; MATSUMOTO, M.; MCCARTY, C.A.; MCDAVID, A.N.; MIREL, D.B.; PASCHALL, J.E.; PUGH, E.W.; RASMUSSEN, L.V.; WILKE, R.A.; ZUVICH, R.L.; RITCHIE, M.D. Quality Control Procedures for Genome Wide Association Studies. **Current Protocols Human Genetics**, v. 68, p.1.19.1–1.19.18, 2011.

ZANELLA, R. **Identification of chromosomal regions associated with infectious diseases in cattle**. 2011. 254f. Dissertation (Doctor of Philosophy in Animal Science) - Washington State University, Washington, 2011.

ZIEGLER, A. Genome-Wide Association Studies: Quality Control and Population-Based Measures, **Genetic Epidemiologic,** v.33, p. S45–S50, 2009.

Capítulo 2. ESTUDO DE ASSOCIAÇÃO GLOBAL DO GENOMA PARA CARACTERÍSTICAS PRODUTIVA E REPRODUTIVA EM SUÍNOS.

RESUMO: A análise de associação global do genoma (GWAS) visa identificar regiões cromossômicas associadas a características fenotípicas de interesse, com base nas diferenças entre as frequências alélicas dos polimorfismos de base única (SNPs). Esta metodologia consiste em correlacionar o genótipo de milhares de SNPs (SNP) com fenótipos de interesse. O objetivo deste capítulo foi realizar o estudo de associação global do genoma, em características de idade para atingir os 100 kg e idade ao primeiro parto da raça Landrace, considerando a análise de um único marcador por vez, estimando-se o efeito aditivo de cada SNP; e selecionar os SNPs significativos por meio de diferentes limiares de significância, observando-se também o desequilíbrio de ligação entre os *loci*. O banco de dados de informações genômicas utilizado, após o controle de qualidade para a remoção de amostras e SNPs, foi composto por 604 amostras e 42.360 SNPs, conforme descrito no capítulo 1. Os dados fenotípicos foram previamente corrigidos para o efeito de grupo contemporâneo em ambas as características. Para a GWAS foram utilizadas apenas as informações fenotípicas, corrigidas para o efeito de grupo contemporâneo, sendo consideradas 604 observações de ID100 e 507 observações para IPP, referentes aos animais genotipados. O modelo incluiu os efeitos fixos da média e dos SNPs e os aleatórios infinitesimal e de resíduo. A matriz de parentesco continha informações das 3 últimas gerações dos animais genotipados. Foram considerados 3 limiares de significância: um conservador (valor de $p = 5x10^{-7}$), um moderado (valor de $p=5x10^{-5}$) e outro baseado no critério de Bonferroni (valor de $p=1,18x10^{-6}$). Pela análise GWAS para a característica idade aos 100 kg foram encontrados 22 SNPs significativos no cromossomo 1 e um SNP no cromossomo 4, considerando o limiar moderado. Para a idade ao primeiro parto foram encontrados dois SNPs significativos, sem posição definida no genoma. Para ambas as características o efeito aditivo estimado foi de grande magnitude, sendo de 2,22 dias para idade

aos 100 kg e de 4,69 dias para idade ao primeiro parto, indicando grande potencial na inclusão destas informações na seleção para diminuir a idade aos 100 kg e idade ao primeiro parto. Pela análise de GWAS foi possível identificar marcadores potenciais localizados no cromossomo 1 e 4 para uso na seleção da população em estudo. Futuras análises incluindo maior número de animais genotipados e a avaliação de metodologias adicionais permitirão detectar SNPs de menor efeito para as características estudadas.

Genome wide association studies for production and reproduction traits in swine

Abstract: The aim of the genome wide association studies (GWAS) is to identify chromosome regions associated with phenotypic traits, based in differences between allele frequencies of single nucleotide polymorphisms (SNPs). The aim of this chapter was to study the genome wide association, for age to 100 kg of body weight (A100) and age at first farrowing (AFF) for Landrace, by single marker analyses, estimating the additive effect of each SNP, and select significant SNPs. The selection of the SNPS was done using different thresholds of significance, and also the linkage disequilibrium between *loci*. The genomic information data used, after the quality control to remove of samples and SNPs, was composed of 604 samples and 42,360 SNPs, as described in Chapter 1. The phenotypic data were previously corrected for the effects of contemporary group in both traits. For the GWAS were used only phenotypic information, corrected for the effect of contemporary group, considering 604 observations A100 and 507 observations for AFF, concerning animals genotyped. The model included as fixed effects the mean and the SNPs and as random infinitesimal effect and residue. The relationship matrix contains information of the last 3 generations of the genotyped animals. It was considered 3 thresholds: a conservative (p-value = $5x10^{-7}$), a moderate (p- value = $5x10^{-5}$) and one based on Bonferroni test (p-value = $1.18x10^{-6}$). The GWAS analysis for A100 found 22 significant SNPs on

chromosome 1 and one SNP on chromosome 4, considering the moderate threshold. For AFF, two SNPs were significant, with no defined position in the genome. For both traits, the additive effect estimated was of large magnitude, 2.22 days for age at 100 kg and 4.69 days for age at first farrowing, indicating a great potential of the inclusion of this information in selection to reduce these traits. The quality control analysis was effective to remove samples and SNPs with problems in genotyping. Further analysis including more animals genotyped and evaluation of additional methodologies may detect SNPs with minor effect for all traits.

INTRODUÇÃO

Do ponto de vista produtivo, a eficiência reprodutiva é considerado um dos fatores mais importantes que afetam a produção de suínos. Características reprodutivas, especialmente aquelas relacionadas à fertilidade, tamanho de leitegada e viabilidade antes do desmame, são importantes componentes para melhorar o índice de leitões produzidos/porca/ano (Irgang, 1985; Tart, 2012). Este aumento depende da redução da idade das leitoas à primeira concepção ou ao primeiro parto (precocidade sexual), do intervalo desmame-concepção ou intervalo entre partos, da duração do período de aleitamento e do intervalo desmame-descarte das porcas, bem como do aumento da taxa de concepção, do tamanho da leitegada ao nascer e da porcentagem de sobrevivência dos leitões (Irgang, 1985). A estimativa de herdabilidade para idade ao primeiro parto apresenta valores controversos na literatura. Serenius et al. (2008) estimaram herdabilidade de 0,16 (± 0,09) para idade ao primeiro parto, por meio de análise Bayesiana multivariada gaussiana. Valores próximos para mesma característica foram relatados por Ziedina et al. (2011), que estimaram a herdabilidade de 0,10 em uma população Landrace, ao avaliarem características reprodutivas por meio de análise multicaracterística. No entanto, Torres Filho et al. (2005)

encontraram herdabilidade aditiva direta superior, de 0,34 (± 0,09) ao incluir no modelo efeitos materno e comum de leitegada.

A idade para atingir determinado peso vivo é uma característica de grande importância, em virtude da necessidade em diminuir os custos de produção, principalmente ligados a ração. Quanto menor o intervalo necessário para o animal atingir o peso de abate, maior a oportunidade de utilização da infraestrutura relacionada ao sistema de produção. A característica dias para atingir 100 kg de peso corporal é uma variável previamente ajustada, ou seja, é a idade ajustada para 100 Kg de peso corporal. A herdabilidade para a característica idade aos 100 kg em suínos foi relatada entre 0,13 e 0,20 para machos e fêmeas da raça Large White, utilizando a metodologia da máxima verossimilhança restrita (Torres Filho et al., 2005) a 0,33 para uma população da mesma raça, utilizando a metodologia de Amostrador de Gibbs (Barbosa et al., 2008).

Em suínos, a avaliação de características de importância econômica é feita usando o valor genético predito, com base nas informações fenotípicas e de pedigree do indivíduo e de seus parentes. A partir do início da década de 90, este panorama começou a ser alterado, pelo início de grandes projetos de mapeamento genômico em animais domésticos, cujo objetivo principal era aplicar técnicas moleculares para identificação e clonagem de genes que controlam características quantitativas.

O avanço das técnicas moleculares permitiu um grande aumento na rapidez, quantidade e complexidade dos dados gerados, o que mudou o paradigma da Genética para uma ciência extremamente rica em dados. Dessa forma, o fator limitante tornou-se a análise e a interpretação desses dados, ao invés da geração dos mesmos, sendo necessário o avanço no desenvolvimento de metodologias de análise, ferramentas de bioinformática e estratégias de seleção (Ledur et al., 2009).

O estudo de associação global do genoma (GWAS) é um procedimento validado para identificar *loci* responsáveis pela variação de características poligênicas. Esta metodologia consiste em correlacionar o genótipo de milhares de SNP com fenótipos de interesse. Os primeiros

estudos envolvendo a associação global do genoma, realizados em humanos, possibilitaram a identificação de *loci* que afetam várias doenças complexas, como câncer de próstata, transtorno bipolar, doença arterial coronariana, doença de Crohn, hipertensão, artrite reumatóide, diabetes tipo 1 e diabetes tipo 2 (Yeager et al., 2007; WTCCC, 2007) e características quantitativas, como a altura em humanos (Vischer, 2008).

Em contraste com os estudos em humanos, em que as análises têm o propósito de identificar principalmente marcadores para doenças, a aplicação dos marcadores SNPs nos animais domésticos visa aumentar a acurácia na seleção para acelerar o melhoramento genético para características de interesse econômico (Fan et al., 2011). O emprego de ferramentas moleculares avançadas, como o painel de SNPs, tem maximizado os benefícios da seleção assistida por marcadores, permitindo a varredura uniforme do genoma para milhares de marcadores simultaneamente, e propiciando a obtenção de mapas genômicos densos extremamente informativos.

Esta tecnologia pode complementar a seleção tradicional e aumentar a acurácia, enquanto reduz o intervalo de geração (Meuwissen et al., 2001). Na literatura, há estudos recentes utilizando a metodologia GWAS para características reprodutivas como características de leitegada, duração da gestação e idade a puberdade e características estruturais, como a longevidade (Tart, 2012; Onteru et al., 2012; Nonneman et al., 2011; Fan et al., 2011). No entanto, nenhum estudo considerando a idade ao primeiro parto foi encontrado.

O método mais simples de analisar GWAS é a associação com um único marcador (SMA; *Single Marker Association*), ou seja, testa-se um marcador de cada vez. Essa metodologia utiliza testes de hipótese para detectar associação entre locos e caráter fenotípico em nível populacional com significância estatística (Resende et al., 2011). Este tipo de análise baseia-se em um modelo linear que considera o efeito individual de cada SNP e o efeito de fatores adicionais como leitegada e reprodutor (Bouwman et al., 2011). A análise de marcador único auxilia na determinação do número e da natureza do gene/QTL que controla a

característica. Para detectar a associação entre marcadores moleculares e características de interesse, uma variedade de análises estatísticas pode ser utilizada, incluindo o teste T, a ANOVA, a regressão, estimativas de máxima verossimilhança e log *likelihood ratios* (Resende, 2011).

Nonneman et al. (2011) utilizaram a metodologia de análise de associação global do genoma, considerando um marcador de cada vez, para avaliar a idade a puberdade em uma população Duroc-Landrace-Yorkshire.

No entanto, a aplicação desta metodologia leva ao grande número de associações falso-positivas (Aulchenko et al., 2007). A utilização dos modelos mistos, considerando o parentesco poligênico entre os indivíduos como um componente aleatório no modelo, podendo incluir também outros efeitos aleatórios, pode ser uma alternativa para reduzir o número de falsos positivos. A desvantagem desta análise quando comparada com o modelo linear simples é a velocidade de processamento, sendo bem mais lenta.

A alternativa proposta neste trabalho é utilizar os dados previamente corrigidos para os demais efeitos fixos, empregando todas as observações fenotípicas contempladas no banco de dados, e considerando na análise de associação somente o efeito infinitesimal e o efeito de cada SNP, além da média e o resíduo.

Na realização de análise de associação global do genoma, o uso da matriz de parentesco no modelo reduz a ocorrência de resultados viesados quando há animais relacionados, pois considera que as diferenças são decorrentes da composição genética e não somente do fenótipo em estudo (Deng e Chen 2001; Purcell et al., 2007).

Para verificar o ajuste do modelo e para verificar a existência de subestrutura da população pode-se utilizar o gráfico quantil-quantil (Q-Q plot) (Settles et al., 2009; Zanella et al., 2011). Este gráfico plota os valores de p esperado contra os valores de p observado. Na ausência de subestrutura da população ou associação, os valores devem perfazer uma linha de 45 graus. O desvio da linha observada (λ) no final do gráfico é representação dos SNPs associados com fenótipo a ser testado.

A interpretação dos resultados do estudo de associação global do genoma depende da escolha apropriada do limiar estatístico (valor de p). Na análise de associação por um único ou por múltiplos marcadores, o resultado pode conter um número excessivo de falsos positivos, exigindo uma estratégia que envolva além da escolha correta do limiar de significância, outro parâmetro que relacione a distância mínima entre SNPs, na seleção de SNPs significativos (Potkin et al., 2009), ou a própria correlação entre os SNPs (r^2).

OBJETIVO

O objetivo deste capítulo foi realizar o estudo de associação global do genoma, em características de idade para atingir os 100 kg e idade ao primeiro parto, com a análise de um único marcador por vez, considerando-se dados fenotípicos previamente corrigidos para os efeitos fixos de grupo contemporâneo, estimando-se o efeito aditivo de cada SNP; e consequentemente selecionar os SNPs significativos por meio de diferentes limiares de significância, observando-se também o desequilíbrio de ligação entre os loci.

MATERIAL E MÉTODOS

População, genótipos e fenótipos: Os dados utilizados neste estudo foram provenientes de animais da linha fêmea da raça Landrace, pertencentes ao Programa de Melhoramento Genético de Suínos da BRF SA. Amostras de DNA de 606 animais foram enviadas para serem genotipadas pela empresa *GeneSeek,* nos Estados Unidos. As amostras foram genotipadas pelo *PorcineSNP60kBeadchip*, que contem 64.232 marcadores SNPs. No entanto, deste total, 62.163 SNPs são marcadores informativos e utilizados na genotipagem. O chip utilizado foi a versão 1, com um espaçamento em média de 28 Kb. Os dados de genotipagem

foram verificados por meio de controle de qualidade desenvolvido em conjunto com a Embrapa Suínos e Aves e no servidor da Embrapa Informática Agropecuária, apresentado no capitulo anterior, para verificar amostras e SNPs com possíveis problemas na genotipagem. As rotinas para avaliação dos critérios de qualidade foram elaboradas no programa R.

No controle de qualidade das amostras foram verificados: problemas com discordância do sexo; teste de paternidade; animais duplicados; eficiência de genotipagem (*call rate*) menores que 90%; desvio da heterozigosidade fora de 3 desvios-padrão; estratificação da população; e subestrutura da população. Para o controle de qualidade dos SNPs, foram removidos aqueles que apresentaram eficiência de genotipagem menores que 0,98; frequência do menor alelo (MAF) menor que 0,03; equilíbrio de Hardy-Weinberg (EHW) com X^2 menor que 10^{-6} e SNPs coincidentes, ou seja, os que apresentaram a mesma posição, sendo retirados os de menor MAF. Foi plotado o gráfico Quantil – Quantil (*Q-Q Plot*) que avalia a subestrutura da população, podendo indicar contaminação de amostras, e também a ocorrência de associações verdadeiras entre os SNPs significativos e a característica.

Após as análises de critério de qualidade foram eliminadas amostras e SNPs, sendo considerado para análise de GWAS o total de 604 amostras e 42.360 SNPs.

Inicialmente, o banco de dados referente às observações fenotípicas continha 132.409 animais com observações de características produtivas e reprodutivas. Neste estudo foram avaliadas as características idade aos 100 kg, considerando 77.765 observações e idade ao primeiro parto, contendo 2.422 observações. Estas características foram escolhidas por apresentarem média e baixa herdabilidades, respectivamente, além de sua importância econômica.

Considerando as informações fenotípicas de toda população para ambas as características, os dados foram previamente corrigidos, considerando os seguintes modelos:

Idade aos 100 kg (1.1)

$Y_i = \mu + GC_i + e_i$;

em que Y_i é o valor observado da característica idade para atingir 100 kg de peso vivo; μ é a média das observações; GC_i é o efeito fixo do i-ésimo grupo contemporâneo, formado por sexo, granja e ano-semana de desmame; e_i é o efeito aleatório do resíduo, com μ= 0 e variância=σ_e^2

Idade ao primeiro parto (1.2)

$Yi = \mu + GC_i + e_i$,

em que Y_i é o valor observado da característica idade ao primeiro parto; μ é a média das observações; GC_i é o efeito fixo i-ésimo grupo contemporâneo, formado por granja, ano-mês de cobertura; e e_i o efeito aleatório do resíduo, com μ= 0 e variância=σ_e^2.

Foram eliminadas observações de grupos contemporâneos com menos de 5 observações para idade ao primeiro parto e 40 observações para idade aos 100 kg, considerando toda base de dados. As herdabilidades encontradas para a população foram de 0,44 para idade aos 100 kg e 0,08 para idade ao primeiro parto (dados não publicados), e foram obtidas por meio da máxima verossimilhança restrita, utilizando o programa REMLF90 (Mizstal, 2001).

Para a realização da edição do banco de dados foi utilizado o programa estatístico SAS, versão 9.2 (SAS Institute, 2011).

Análises de associação global do genoma: Para os estudos de associação global do genoma foi utilizada a análise de marcadores únicos para identificação dos SNPs de maior efeito sobre as características avaliadas.

Análises de associação global do genoma: Para os estudos de associação global do genoma foi utilizada a análise de marcadores únicos para identificação dos SNPs de maior efeito sobre as características avaliadas.

Foram utilizadas as observações fenotípicas previamente corrigidas pelos modelos 1.1 e 1.2, considerando somente as observações dos animais genotipados. O modelo para análise de associação global do genoma contemplou os efeitos de acordo com o modelo 1.3:

$$Y_{ij}^{*} = a_i + SNP_j + e_{ij} \quad (1.3);$$

Em que Y_{ij}^{*} é o fenótipo corrigido pelo modelo 1.1 e pelo modelo 1.2 para idade para atingir os 100 kg de peso vivo e idade ao primeiro parto, respectivamente; a_i é o efeito aleatório infinitesimal (dos poligenes); SNP_j é o efeito fixo de cada SNP, testando-se um SNP de cada vez, ou seja, a análise envolveu 42.360 SNPs; e e_{ij} é o efeito aleatório do resíduo.

O banco de dados com informações de parentesco para esta análise considerou apenas as informações de no máximo 3 gerações anteriores a dos animais genotipados, contendo 23.232 animais.

Foi utilizado o programa QxPak (Pérez-Enciso e Misztal, 2011), desenvolvido para análise de dados genômicos, empregando modelos mistos e o procedimento de máxima verossimilhança. O programa permite a inclusão de pedigrees complexos nas análises e atualmente realiza também a análise de SNPs em grande escala, sendo possível ajustar modelos aditivos e de dominância, considerando as informações genotípicas e fenotípicas conjuntamente. O programa permite ainda estimar efeitos aditivos e de dominância dos SNPsDefinição do limiar de significância e seleção de SNPs: Devido à análise de associação com um único marcador gerar um número grande de falsos positivos, além do limiar de significância também foi utilizado um critério envolvendo o desequilíbrio de ligação (LD) entre os loci para a seleção dos SNPs relevantes. Para o teste de significância da associação entre os SNPs e as características avaliadas, foram utilizados diferentes limiares de probabilidade, considerando um limiar mais conservativo, com valor de p de 5×10^{-7}, e outro moderado, de 5×10^{-5}, baseados na literatura (WTCCC, 2007); e o limiar baseado no critério de Bonferroni, também considerado bastante conservador, que consiste no valor de *p* desejado dividido pelo

número de testes. Para este estudo, o ponto de corte foi calculado por: 0,05/N, sendo N o número de SNPs utilizados na análise.

A análise do desequilíbrio de ligação foi realizada com o programa Haploview (Barret et al., 2005) e a medida utilizada foi o r^2, que mensura a correlação entre alelos de dois marcadores, sendo que r^2 igual a 1 indica uma correlação perfeita entre os dois loci.

RESULTADOS E DISCUSSÃO

Após as análises preliminares dos dados, para a análise de associação global do genoma foram utilizadas apenas as observações dos animais genotipados. Na tabela 1 estão as estatísticas descritivas das características idade para atingir os 100 kg e idade ao primeiro parto, considerando apenas os animais genotipados.

Tabela 1. Estatísticas descritivas para as características Idade aos 100 kg, em dias, e idade ao primeiro parto, em dias.

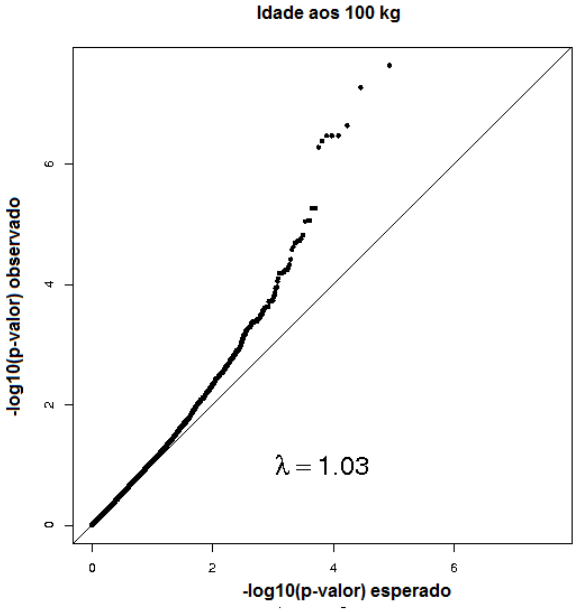
Característica	Número de Observações	Média	Desvio Padrão	Mínimo	Máximo
Idade aos 100 kg (dias)	604	132,33	6,75	112,2	152,1
Idade ao Primeiro Parto (dias)	507	329,27	13,07	304,43	383,36

Na população em estudo, ou seja, dos animais genotipados, a média para idade aos 100 kg foi de 132,33 dias na raça Landrace. Este valor é bem inferior ao encontrado por Torres Filho et al. (2004), que em uma população de 7.009 indivíduos da raça Large White, considerando somente os machos, encontraram média de 140,37 dias, enquanto no caso de fêmeas, considerando 10.541 fêmeas encontraram média de 154,16 dias. As diferenças encontradas na média para idade aos 100 kg entre as populações demonstram a precocidade dos animais da raça Landrace e também pela magnitude da diferença entre as raças e o ano

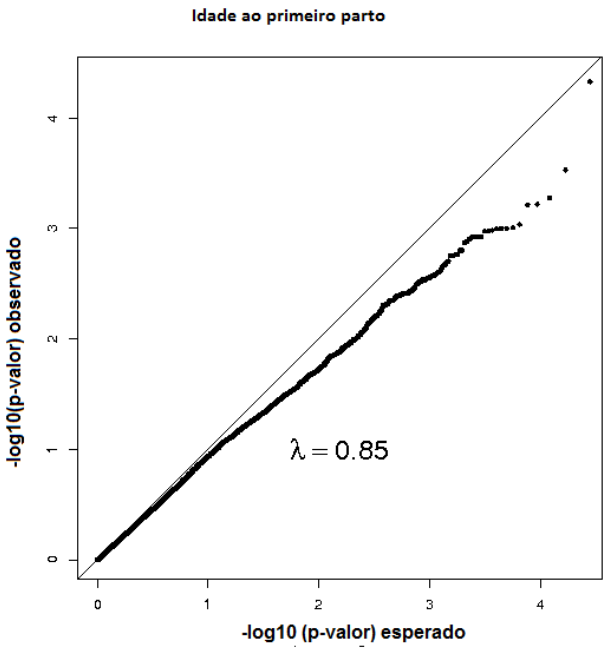
dos estudos pode ser um indicativo do progresso genético, pois em ambos os estudos os dados pertenciam ao mesmo programa de melhoramento genético

Para idade ao primeiro parto, o valor médio encontrado foi de 329,93 dias, no entanto, Torres Filho et al. (2004) encontraram média de 318,5 dias numa população com 1.875 fêmeas da raça Large White. Os valores obtidos neste estudo foram superiores ao relatado na literatura, contudo a população utilizada neste estudo, além de ser de outra raça, apresenta número reduzido de indivíduos.

O gráfico Quantil-Quantil (Q-Q plot) pode ser um indicativo de subestrutura da população e da existência de associação dos SNPs significativos com a característica estudada. Dessa forma, foram plotados os gráficos considerando os valores do –log10(p) observado contra o esperado, para idade aos 100 kg e idade ao primeiro parto, considerando os 42,360 SNPs (Gráfico 1).



(a)



(b)

Gráfico 1. Gráfico Quantil-Quantil dos valores de –log10(p) esperado contra observado para as características idade aos 100 kg (a) e idade ao primeiro parto (b).

De acordo com os gráficos, observa-se que não há subestrutura da população entre as amostras, uma vez que os valores observados encontram-se distribuídos de modo uniforme, próximos a diagonal, sendo o valor do desvio λ igual a 1,03 para idade aos 100 kg e de 0,85 para idade ao primeiro parto. O desvio dos valores observados, em relação à linha diagonal, demonstra possível associação dos SNPs com o fenótipo, no caso da característica idade aos 100 kg (gráfico 1a). No entanto, pelo gráfico 1 (b), verifica-se que os valores observados perfazem uma linha abaixo dos esperados, evidenciando o baixo poder de associação devido ao pequeno número de indivíduos na população genotipada para uma característica de baixa herdabilidade.

As análises de associação global do genoma foram realizadas utilizando-se um marcador por vez, considerando um modelo aditivo, obtendo-se este efeito para cada SNP. Para a seleção dos SNPs significativos, foram adotados 3 limiares, 2 baseados na literatura

(WTCCC, 2007), sendo um mais conservador ($5x10^{-5}$) e outro menos conservador ($5x10^{-7}$). O outro limiar adotado, também conservador, foi seguindo o critério de Bonferroni, em que o valor de p foi de $1{,}18x10^{-6}$, calculado por 0,05/42.360, em que se consideraram somente os SNPs remanescentes na análise após o critério de qualidade.

O Gráfico 2 representa o *Manhattan plot* da análise de associação global do genoma, em que é verificada a associação dos SNPs com a característica de interesse, representada pelo $-\log_{10}$ (valor de p) no eixo do Y e o SNP no eixo do X, por cromossomo, para a característica idade aos 100 kg.

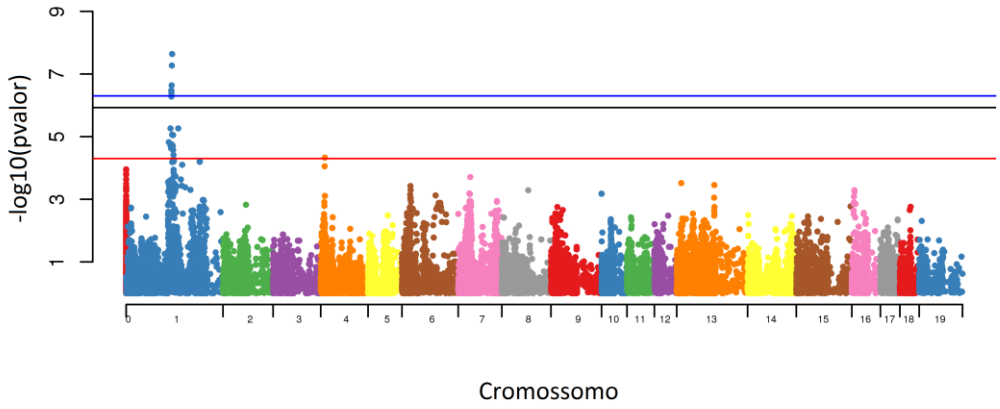


Gráfico 2. Manhattan plot da análise de associação global do genoma entre o –log10(p-valor) e o loci do SNP, em ordem sequencial, por cromossomo, para a característica idade aos 100 kg. Em vermelho o limiar –log10 (5x10-5), em azul o –log10 (5x10-7) e em preto o Bonferroni, -log10(1,18x10-6).

Para idade aos 100 kg, 23 SNPs foram significativos considerando-se o limiar moderado (5×10^{-5}), sendo que 22 SNPs estão localizados no cromossomo 1, indicando que esta região pode conter genes que controlam a característica em questão. No cromossomo 4, apenas 1 SNP foi significativo com o limiar moderado. Ao considerar os limiares mais conservadores, o número de SNPs selecionados se reduz para 7 com o limiar de $5X10^{-7}$ e para 8 com o limiar de Bonferroni, sendo todos estes localizados no cromossomo 1. Mesmo assim, muitos desses SNPs podem

ser falso positivos devido a sua proximidade no genoma e consequentemente podem estar correlacionados.

Na Figura 1 pode-se observar o desequilíbrio de ligação entre os 22 SNPs significativos pelo limiar moderadamente significativo, no cromossomo 1, em que o valor do r^2 encontrado entre os SNPs foi de alta magnitude.

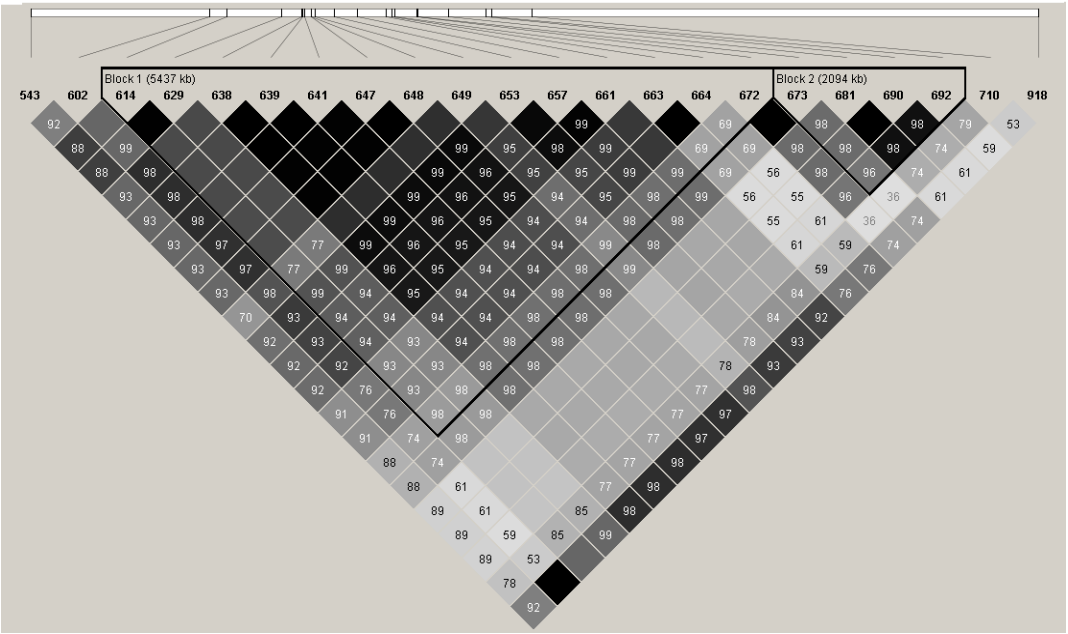


Figura 1. Desequilíbrio de ligação (r^2) entre os SNPs moderadamente significativos no cromossomo 1 para a característica idade aos 100 kg. Os quadrados pretos, sem número, representam desequilibrio de ligação completo entre os SNPs (r^2=1).

Pela Figura 1, observa-se que os SNPs significativos localizados na região do cromossomo 1 estão, e em sua maioria, em desequilíbrio de ligação completo, indicando a presença de associações falso positivas. Dessa forma, apenas um SNP pode estar realmente associado com a idade aos 100 Kg na região do pico bem definido. Entretanto, como se evidenciou a presença de dois blocos na Figura 1, é possível a existência de outra associação verdadeira com o SNP que quase atinge o limiar moderado.

O efeito aditivo do SNP mais significativo no cromossomo 1 foi de -2,24 (0,57) dias. Este efeito indica que este marcador deve ser investigado, pois poderia ser utilizando na seleção visando à redução da idade para atingir os 100 kg de peso vivo. A seleção de animais mais precoce para o abate impacta diretamente na produção, pois reduz os custos fixos, tornando mais lucrativa a produção (Torres Filho et al., 2004).

No banco de dados QTLdb (http://www.animalgenome.org/cgi-bin/QTLdb/SS/index) foi encontrado apenas um estudo envolvendo a característica idade aos 100 kg, no qual Wei et al. (2011) estudaram a região de complexo principal de histocompatibilidade no cromossomo 7. Estes autores identificaram dois SNPs associados a idade aos 100 Kg em animais cruzados Meishan x Large White, que poderiam ser utilizados em programas de melhoramento futuros. Soma et al. (2011) avaliaram características de crescimento em animais da raça Duroc, utilizando marcadores microssatélites e encontraram QTLs nos cromossomos 1, 13 e 15 para idade aos 105 kg.

O Gráfico 3 representa o *Manhattan plot* da análise de associação global do genoma entre o –log 10 (p-valor) e o SNP, por cromossomo, para a característica idade ao primeiro parto.

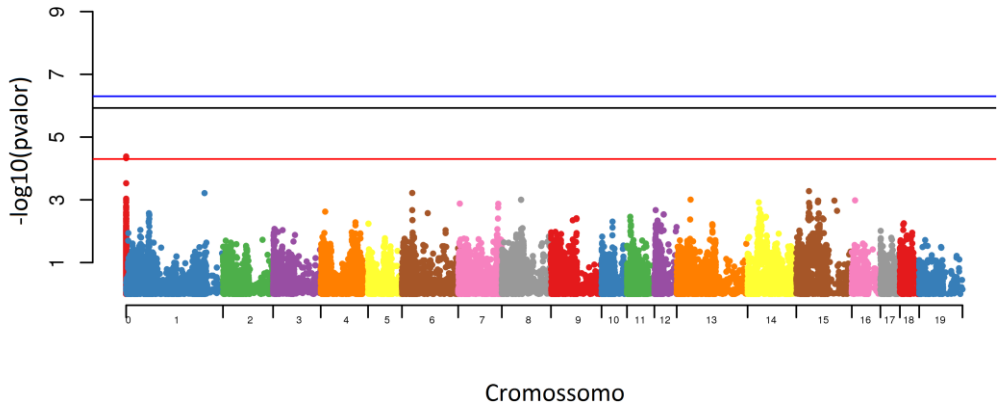


Gráfico 3. Manhattan plot da análise de associação global do genoma entre o $-\log_{10}$ (p-valor) e o loci do SNP, em ordem sequencial, para a característica idade para o primeiro parto. Em vermelho limiar $-\log_{10}$ ($5x10^{-5}$), em azul $-\log_{10}$ ($5x10^{-7}$) e em preto Bonferroni, $-\log_{10}(1{,}18x10^{-6})$.

Para a idade ao primeiro parto, somente dois SNPs significativos foram encontrados, considerando o limiar de significância moderado. Estes foram incluídos no cromossomo zero, pois ainda não têm localização definida no genoma. Dessa forma, os dois SNPs significativos podem ou não estar localizados no mesmo cromossomo. Versões atualizadas da anotação do genoma serão necessárias para que se possa estabelecer a correta posição desses SNPs.

O efeito aditivo do SNP mais significativo foi estimado em 4,69 (1,12) dias, sendo este efeito de grande magnitude sobre a característica em questão. Caso este SNP fosse utilizado na seleção da população em estudo, impactaria potencialmente na redução de aproximadamente 4,5 dias na idade ao primeiro parto.

O baixo poder de detecção da análise pode ter sido em decorrência do pequeno tamanho da amostra para uma característica de baixa herdabilidade. A utilidade dos marcadores SNPs encontrados para características complexas, como as características reprodutivas, são limitadas devido à natureza poligênica destes fenótipos (Tart, 2012).

Os resultados obtidos neste trabalho devem ser melhor explorados para posterior uso na seleção em programas de melhoramento. Uma das formas seria o aumento da amostra genotipada, o que daria maior poder de detecção de SNPs de menor efeito, além de outras metodologias como mencionado anteriormente. A utilização da metodologia de GWAS por múltiplos marcadores, baseado em haplótipos, deveria ser avaliada visando detectar SNPs de menor efeito, mantendo o número de associações falsos positivas em um nível aceitável.

CONCLUSÃO

A análise de associação global para a característica idade aos 100 kg e idade ao primeiro parto, considerando apenas um marcador por vez, estimou efeitos de grande magnitude para os SNPs associados a ambas as características. Os SNPs significativos para idade aos 100 kg foram localizados nos cromossomos 1 e 4, considerando um limiar de significância moderado. Ao considerar o limiar mais conservativo, observaram-se SNPs significativos somente no cromossomo 1, diminuindo a possibilidade de associações falso positivas. A análise do desequilíbrio de ligação permitiu identificar 2 possíveis regiões de QTL no cromossomo 1. Já para idade ao primeiro parto, os SNPs significativos com limiar de significância moderado estavam posicionados no cromossomo 0, ou seja, não apresentam posição definida no genoma atualmente.

REFERÊNCIAS BIBLIOGRÁFICA

AULCHENKO, Y.S.; KONING, D.; HALEY, C. Genome wide rapid association using mixed model and regression: a fast and simple method for genome wide pedigree-based quantitative trait loci association analysis. **Genetics,** v. 177, p. 577–585, 2007.

BARBOSA, L.; LOPES, P. S.; REGAZZI, A. J.; TORRES, R. A.; SANTANA JUNIOR, M. L.; VERONEZE, R. Estimação de parâmetros genéticos em

suínos usando Amostrador de Gibbs. **Revista Brasileira de Zootecnia**. v. 37, p. 1200 – 1206. 2008.

BARRETT, J.C.; FRY, B.; MALLER, J.; DALY, M. J. Haploview: analysis and visualization of LD and haplotype maps. **Bioinformatics**. [PubMed ID: 15297300]. 2005.

BOUWMAN, A. C.; BOVENHUIS, H.; VISKER, M. H.; ARENDONK, J. A. V. Genome-wide association of milk fatty acids in Dutch dairy cattle. **BMC Genetics**, v. 12, p. 1-12, 2011.

DEKKERS, J. C. M.; HOSPITAL, F. The use of molecular genetics in the improvement of agricultural populations. **Nature Reviews Genetics**, v.3, p. 22–32. 2002.

DENG, H. W.; CHEN, W. M. The power of the transmission disequilibrium test (TDT) with both case-parent and control-parent trios. **Genetic Research**, v. 78, p. 289-302. 2001.

FAN, B.; ONTERU, S. K.; DU, Z. Q.; GARRICK, D. J.; STALDER, K. J. ROTHSCHILD, M. F. Genome-Wide association study identifies loci for body composition and structural soundness traits in pigs. **PLoS One**, v.6, p. 1-11, 2011.

HUENTELMAN,M.J.; CRAIG, D. W.; SHIEH, A.D.; CORNEVEAUX, J.J.; HU-LINCE, D.; PEARSON, J.V.; STEPHAN, D.A. SNiPer: improved SNP genotype calling for Affymetrix 10K GeneChip microarray data. **BMC Genomics**, v. 6, p. 149. 2005.

IRGANG, R.. **Estimativas de herdabilidade para características que compões a produtividade anual de leitões por porca**. Concórdia: EMBRAPA-CNPSA, 1985. 4p.

LEDUR, M. C.; NAVARRO, N.; PÉREZ-ENCISO, M. Data modeling as a main source of discrepancies in single and multiple marker association methods. **BMC Proceedings,** v.3, s.9, 2009.

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

MISZTAL, I. **REMLF90 manual**. [2001]. Disponível em: ftp://nce.ads.uga.edu/pub/ignacy/blupf90/. Acesso em: 10/11/12.

NONNEMAN, D. J., G. A. ROHRER, L. A. REMPEL, R. T. WIEDMANN AND J. L. VALLET. Genome-wide associations for age at puberty in a Duroc-Landrace-Yorkshire swine population. **Plant and Animal Genome Conference Proceedings.** 2011.

ONTERU, S. K.; Fan, B.; Du, Z. Q.; GARRICK, D. J.; STALDER, K. J.; ROTHSCHILD, M. F. A whole-genome association study for pig reproductive traits. **Animal Genetics**, v. 43, p. 18-26, 2012.

PEREZ-ENCISO, M. e MISZTAL, I. Qxpak.5: old mixed model solutions for new genomics problems. **BMC Bioinformatics***,* v. 12, p. 202, 2011.

POTKIN, S. G.; TURNER, J. A.; GUFFANTI, G.; LAKATOS, A.; TORRI, F.; KEATOR, D. B.; MACCIARDI, F. Genome-wide strategies for discovering genetic influences on cognition and cognitive disorders: Methodological considerations. **Cognitive Neuropsychiatry,** v. 14, p. 391-418, 2009.

PURCELL, S., NEALE, B.; TODD-BROWN,K.; THOMAS, L.; FERREIRA, M.A.; BENDER, D.; MALLER, J.; SKLAR, P.; DE BAKKER, P.I.; DALY, M.J.; SHAM, P.C. PLINK: a tool set for whole-genome association and population-based linkage analyses. **American Journal of Human Genetics**, v. 81, p. 559-575, 2007.

R Development Core Team. **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. Versão 2.15.2. Disponível em <http://www.R-project.org>, 2012. 60p.

RESENDE, M. D. V. de; SILVA, F. F. e; VIANA, J. M. S.; PETERNELLI, L. A.; RESENDE JÚNIOR, M. F. R.; DEL VALLE, P. M. Métodos estatísticos na seleção genômica ampla. **Documentos 219,** EMBRAPA - CNPF, 104 p., 2011.

SAS/STAT. **User´s guide.** Versão 9.3. Cary: SAS Institut Inc., 2011.

SERENIUS, T..;STALDER, K.J.; FERNANDO, R.L. Genetic associations of sow longevity with age at first farrowing, number of piglets weaned, and wean to insemination interval in the Finnish Landrace swine population. **Journal of Animal Science**, v. 86, p. 3324-3329. 2008.

SOMA, Y.; UEMOTO, Y.; SATO, S.; SHIBATA, T. ; KADOWAKI, H. ; KOBAYASHI E.; SUZUKI, K. Genome-wide mapping and identification of new quantitative trait loci affecting meat production, meat quality, and carcass traits within a Duroc. **Journal of Animal Science**, v. 89, p. 601-608. 2011.

TART, J.K. **Genomic Analysis of Characteristics in Swine Contributing to Sow Longevity.** 2012. 73f. *Thesis (Master of Science in Animal Science) - University of Nebraska-Lincoln*, Lincoln, 2012.

The Wellcome Trust Case Control Consortium: Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. **Nature**, v. 447, p.661-678, 2007.

TORRES FILHO, R. A.; TORRES, R. A.; LOPES, P. S.; EUCLYDES, R. F.; ARAÚJO, C. V.; C. S.; ALMEIDA E SILVA, M. Avaliação de modelos para estimação de componentes de (co)variância em características de desempenho e reprodutivas em suínos. **Revista Brasileira de Zootecnia**, v. 33, p. 350-357, 2004.

TORRES, R.A.; TORRES, R.A.; LOPES, P.S.; PEREIRA, C.S.; EUCLYDES, R.F.; ARAUJO, C.V.; SILVA, M.A.; BREDA, F.C. Genetic parameters estimates for reproductive traits in swine. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, v. 57, n.5, p.684–689. 2005.

WTCCC - The Wellcome Trust Case Control Consortium: Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. **Nature**, v. 447, p.661-678, 2007.

WEI, W. H.; SKINNER, T.M; ANDERSON, J. A.; SOUTHWOOD, O. I.; PLASTOW, G; ARCHIBALD, A. L.; HALEY, C. S. Mapping QTL in the porcine MHC region affecting fatness and growth traits in a Meishan/Large White composite population. Animal Genetic, v. 42, p. 83 – 85. 2011.

YEAGER, M.; ORR, N.; HAYES, R.B.; JACOBS, K.B.; KRAFT, P.; WACHOLDER, S.; MINICHIELLO, M.J.; FEARNHEAD, P.; YU, K.; CHATTERJEE, N.; WANG, Z.; WELCH, R.; STAATS, B.J.; CALLE, E.E.; FEIGELSON, H.S.; THUN, M.J.; RODRIGUEZ, C.; ALBANES, D.; VIRTAMO, J.; WEINSTEIN, S.; SCHUMACHER, F.R.; GIOVANNUCCI, E.; WILLETT, W.C.; CANCEL-TASSIN, G.; CUSSENOT, O.; VALERI, A.; ANDRIOLE G.L.; GELMANN, E.P.; TUCKER, M.; GERHARD, D.S.; FRAUMENI, J.F.; HOOVER, R.; HUNTER, D.J.; CHANOCK, S.J.; THOMAS, G. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. **Nature Genetics**, v. 39, p. 645-649, 2007.

ZANELLA, R. **Identification of chromosomal regions associated with infectious diseases in cattle**. 2011. 254f. Dissertation (Doctor of Philosopy in Animal Science) - Washington State University, Washington, 2011.

ZIEDINA, I.; JONKUS, D.; PAURA, L. Genetic and phenotypic parameters for reproduction traits of landrace sows in Latvia. **Agriculturae Conspectus Scientificus,** v. 76, p. 219-222. 2011.

CONCLUSÕES GERAIS

Com base nos resultados obtidos nesta tese, conclui-se:

- A aplicação do controle de qualidade de amostras e SNPs, realizados com os critérios adotados no capítulo 1, foram efetivos na remoção de amostras e SNPs que apresentaram problemas de genotipagem;
- A análise de associação global para a característica idade aos 100 kg de peso corporal e idade ao primeiro parto, considerando apenas um marcador por vez, estimou efeitos de grande magnitude para os SNPs associados à ambas as características, sendo -2,24 dias para idade aos 100 kg de peso corporal e 4,5 dias para idade ao primeiro parto;
- Os SNPs significativos para idade aos 100 kg foram localizados nos cromossomos 1 e 4, considerando um limiar de significância moderado. Ao considerar o limiar mais conservativo, observou-se SNPs significativos somente no cromossomo 1, diminuindo a possibilidade de associações falso positivas;
- A análise do desequilíbrio de ligação entre os 22 SNPs significativos para a característica idade aos 100 kg de peso corporal demonstrou uma forte correlação entre os loci, indicando possíveis associações falso positivas, mas demonstrando a possibilidade de 2 regiões de QTL neste cromossomo;
- Para idade ao primeiro parto, foram encontrados 2 SNPs significativos, considerando o limiar de significância moderado, e estes não apresentaram posição definida no genoma atualmente.