UNIVERSIDADE FEDERAL DA GRANDE DOURADOS

Lara Endres da Silva

CONTROLE DE QUALIDADE DE DADOS DE SNP EM BOVINOS DA RAÇA NELORE

DOURADOS

ABRIL 2013

LARA ENDRES DA SILVA

CONTROLE DE QUALIDADE DE DADOS DE SNP EM BOVINOS DA RAÇA NELORE

Trabalho de Conclusão de curso de graduação apresentado para o título de Bacharel em Biotecnologia. Faculdade de Ciências Biológicas e Ambientais. Universidade Federal da Grande Dourados. Orientadora: Prof. Dra. Alexéia Barufatti Grisolia. Co-orientadora: Dra. Marcia Cristina Matos

DOURADOS

ABRIL 2013

LARA ENDRES DA SILVA

CONTROLE DE QUALIDADE DE DADOS DE SNP EM BOVINOS DA RAÇA NELORE

Trabalho de Conclusão de Curso aprovado como requisito parcial para obtenção de título de Bacharel em Biotecnologia na Universidade Federal da Grande Dourados, pela comissão formada por:

_____

Orientadora: Prof. Dra. Alexéia Barufatti Grisolia
FCBA – UFGD

_____

Dra. Marcia Cristina Matos

_____

Prof. Dr. Rodrigo Matheus Pereira
FCBA – UFGD

Dourados, abril de 2013

Agradecimentos

Agradeço primeiramente à Deus, pela minha vida e por todos os caminhos que me trouxeram até aqui.

À minha família, especialmente minha mãe Marli, por ouvir meus "desabafos", me incentivar e principalmente entender minhas ausências.

Aos amigos (mais que amigos, irmãos) do grupo de jovens EMAÙS,especialmente à Valeska, Daiane de Deus, Daya, Lucas Ricardo, Sthefany, Rodrigão, pelos momentos de risos e descontração durante o dia a dia.

Aos amigos da I turma de Biotecnologia, especialmente Danielly, Nicholas, Carla, Suellen e Luiz Augusto, por tudo o que vivemos e crescemos juntos. A amizade de vocês é algo que sempre guardarei.

Aos amigos do Laboratório de Biotecnologia Aplicada à Produção Animal (Jussara, Joyce, Dani, Bruno, Alexandre e André), pelos momentos de seriedade e crescimento profissional, mas também (e talvez principalmente) pela amizade, união e carinho. Vocês são exemplos para mim.

À profa.orientadora Alexéia Barufatti Grisolia, pelos ensinamentos durante os anos de iniciação científica, pela paciência e por todo apoio concedido até hoje.

Ao professor Dr. Leonardo de Oliveira Seno, por ser um profissional que admiro e levo como exemplo, e também pela oportunidade de "aproveitar" um pouco do seu conhecimento durante os anos de iniciação científica.

A todos os professores que tive a oportunidade de conhecer na graduação, a quem devo todo aprendizado, profissionalismo e ética adquiridos ao longo destes anos, e que me inspiram a seguir a carreira acadêmica para, quem sabe um dia, poder atuar como professora ao lado deles.

Aos técnicos administrativos e auxiliares que atuam na Faculdade de Ciências Biológicas e Ambientais, especialmente à Tatiane Zaratini, pela dedicação e trabalho exemplar.

À Universidade Federal da Grande Dourados, pelo apoio logístico e pela oportunidade de realizar o curso de graduação em Biotecnologia.

Aos órgãos de fomento CNPq e Fundect, pelo apoio financeiro à pesquisa realizada, bem como à bolsa de iniciação científica concedida durante o período de desenvolvimento da pesquisa.

À empresa Deoxi Biotecnologia Ltda., pela oportunidade de realizar os procedimentos de genotipagens em suas instalações.

A todos aqueles que não citei neste documento mas que, de uma forma ou de outra, me incentivaram e motivaram a seguir em frente.

SUMÁRIO

Lista de abreviaturas

AFLP*:* *Amplified Fragment Lenght Polymorphism* (Polimorfismo de Comprimento de Fragmento Amplificado)

DNA*:* *Desoxyribonucleic Acid* (Ácido Desoxirribonucleico)

EDTA*: Ethylenediamine tetraacetic acid* (Ácido Etilenodiamino Tetraacético)

HWE*:* *Hardy-Weinberg Equilibrium* (Equilíbrio de Hardy-Weinberg)

IBS*:* *Identity By State* (Identidade Por Estado)

GWAS*:* *Genome-Wide Association Studies* (Estudos de Associação Genômica Ampla)

GWS*:* *Genome-Wide Selection* (Seleção Genômica Ampla)

MAF*: Minor Allelle Frequency* (Frequência do Alelo Mínimo)

PCR*:* *Polymerase Chain Reaction* (Reação de Polimerase em Cadeia)

RFLP*:* *Restriction Fragment Length Polymorphism* (Polimorfismo do Comprimento de Fragmentos de Restrição)

RAPD*:* *Random Amplified Polymorphic DNA* (DNA Polimórfico Amplificado Aleatoriamente)

SNP*:* *Single Nucleotide Polymorphism* (Polimorfismo de Nucleotídeo Único)

SSR*: Simple Sequence Repeats* (Sequências Simples Repetidas)

Lista de Figuras

Lista de Tabelas

RESUMO

Considerando-se a quantidade de dados gerados em paineis de alta densidade de SNPs, até mesmo uma baixa taxa de erro pode ser prejudicial para as análises. Deste modo, o objetivo deste trabalho foi salientar a importância do controle de qualidade para análise dos dados e gerar conhecimentos referentes às genotipagens de SNPs por meio de SNP-chip. O grupo amostral utilizado para gerar o banco de dados foi constituído por 96 fêmeas Nelore. O DNA foi extraído a partir de sangue periférico, em seguida submetido aos procedimentos de preparo para hibridização. As amostras foram aplicadas no chip e seguiu-se com os procedimentos, até que o mesmo fosse submetido ao equipamento de genotipagem. Para controle de qualidade excluiu-se marcadores com freqüência do alelo mínimo abaixo de 0,02%, call rate abaixo de 98% e marcadores que não se encontravam em equilíbrio de *Hardy-Weinberg* de p<1e-5 para o teste de *Fisher*. No controle de qualidade por amostra foram excluídas aquelas que tinham IBS acima de 95% e call rate menor que 90%. Por não atenderem aos critérios de *call rate*, 97.334 marcadores foram removidos, 188.762 por não atenderem aos critérios da freqüência do menor alelo e 2.509 por não se encontrarem em equilíbrio de *Hardy-Weinberg*. Nenhuma amostra foi removida do conjunto de dados. De 742.909 SNPs que faziam parte do conjunto de dados iniciais, 333.574 foram removidos por meio do controle de qualidade, resultando no conjunto de dados de 409.335. Os parâmetros utilizados foram capazes de impedir a presença de marcadores e amostras não informativas no conjunto de dados, demonstrando a importância do controle de qualidade como etapa prévia nas análises de associação genômica.

Palavras-chave: Frequência do menor alelo; Polimorfismo de base única; Equilíbrio de *Hardy-Weinberg*

1. INTRODUÇÃO

Existem diversas linhas de pesquisa que visam melhoria da produção animal, e dentre elas as técnicas de biologia molecular destacam-se. Diferentes abordagens tecnológicas podem ser usadas na busca por genes responsáveis pela expressão de características zootécnicas, ou de regiões do genoma que estejam relacionadas com a manifestação destas características (HERR et al., 2004).

Avanços no sequenciamento de genomas de mamíferos e no desenvolvimento de ferramentas de bioinformática identificaram um marcador molecular denominado SNP (do inglês, *Single Nucleotide Polymorphism)*. Muitos SNPs existentes em genes de interesse econômico causam alteração no funcionamento dos mesmos, pois podem alterar os aminoácidos produzidos por eles, fazendo com que sua atuação seja diferenciada dos genes que não possuem tal polimorfismo. Deste modo, os SNPs são capazes de gerar modificações fenotípicas (HEATON et al., 2002).

Tradicionalmente a associação de SNPs com características fenotípicas de interesse econômico é feita analisando um gene por vez, ou, ocasionalmente, dois ou três genes, demandando tempo e quantidade razoavelmente grande de mão de obra especializada. Até o advento do sequenciamento do genoma bovino, e a consequente identificação de milhares de SNPs, esta metodologia tradicional era a possibilidade mais adequada aos pesquisadores para tais análises, sendo limitadas pela necessidade de se escolher apenas regiões específicas do genoma (HAWKEN et al., 2004).

Recentemente os SNPs têm sido estudados por meio dos painéis de genotipagem de alta densidade, que consistem em uma poderosa abordagem para identificação de variações genéticas ligadas a características fenotípicas produtivas de grande interesse comercial (CORNELIS et al., 2010; WEIGEL e MOTT, 2009).

A utilidade prática destas informações genéticas oriundas da técnica de SNP-chip dependerá da qualidade dos dados gerados, tornando necessária a adoção de medidas para o controle de qualidade das análises subseqüentes. Testes estatísticos de associação simples ficam

comprometidos quando realizados em conjuntos de dados que não foram apropriadamente filtrados, podendo levar a associações falso-negativas ou falso-positivas (TURNER et al., 2011).

Considerando-se a quantidade de dados gerados em GWAS (do inglês, *Genome Wide Association Study* – Estudos de Associação Genômica Ampla), até mesmo uma baixa taxa de erro pode ser prejudicial para as análises. Se, por exemplo, um milhão de marcadores moleculares são testados para associação e a proporção de genótipos falhos for 0.001, então mais de 1000 marcadores podem ser desnecessariamente usados para as análises posteriores devido a associações falso-positivas. Dado o grande número de marcadores analisados, a remoção de uma pequena porcentagem destes não diminui significativamente o potencial de abrangência do estudo em relação ao genoma analisado (ANDERSON et al., 2010).

Os procedimentos de controle de qualidade representam um desafio operacional. A cada análise, em diferentes conjuntos de dados, novas realidades são descobertas a respeito do GWAS, e melhores procedimentos são desenvolvidos. Os algoritmos utilizados para determinação de *genotype calling*, por exemplo, são continuamente aprimorados a fim de permitir que apenas marcadores e amostras confiáveis sejam utilizadas nas análises (TURNER et al., 2011).

A capacidade de gerar grande quantidade de informações em um pequeno intervalo de tempo faz com que a tecnologia de SNP-Chip se destaque dentre as demais. Entretanto, este potencial é desperdiçado quando não se adota medidas adequadas na abordagem das informações geradas. Deste modo, torna-se clara a necessidade de se avaliar metodologias para controle de qualidade na abordagem de dados genômicos.

2. REVISÃO BIBLIOGRÁFICA

2.1. Marcadores Moleculares

Os marcadores moleculares são pequenas sequências de DNA capazes de revelar polimorfismos e distinguir indivíduos de uma mesma espécie ou de espécies diferentes, apresentando segregação pelas gerações segundo padrão de herança mendeliana (FERREIRA e GRATTAPALIA, 1998).

Tais polimorfismos podem se encontrar em regiões codificadoras ou não codificadoras. Quando se encontram inseridos em regiões codificadoras, o polimorfismo pode resultar na alteração do aminoácido expressado naquela região, causando variações fenotípicas. Este tipo de polimorfismo é chamado de "não sinônimo". Quando o polimorfismo ocorre em uma região não codificadora isto também pode alterar a expressão de determinados genes, pois pode alterar sequências promotoras, bem como suprimir a expressão de códons de iniciação ou terminação. Já os polimorfismos "sinônimos" são aqueles que não causam variações fenotípicas, embora haja indícios de que os mesmos possam causar alterações na estabilidade da transcrição do DNA para o RNA (CAPON et al., 2004; CURI, 2004).

Devido a existência de várias técnicas de biologia molecular que são capazes de detectar tais polimorfismos, bem como as particularidades que alguns deles podem apresentar, deve-se determinar qual marcador deve ser utilizado de acordo com o propósito do estudo (VIGNAL et al., 2002).

Os primeiros marcadores utilizados foram os RFLPs (do inglês, *Restriction Fragment Length Polymorphism*). Este é um marcador que se baseia nas variações no tamanho de

sequências tratadas com enzimas de restrição, apresentando características de co-dominância, uma vez que os cromossomos homólogos podem revelar fragmentos de um mesmo tamanho (homozigotos) ou de tamanhos diferentes (heterozigotos). O desenvolvimento do RFLP se deu em 1974, e a partir dos anos 80 começaram a surgir as primeiras metodologias aplicadas em análises de bovinos com este marcador. Naquela época era necessário aproximadamente cinco dias de trabalho para que se conseguisse um genótipo por meio do uso de marcadores moleculares (GEORGES et al., 1987; GRODZICKER et al., 1974).

Há outro tipo de marcador molecular que também utiliza enzimas de restrições: os AFLPs (do inglês, *Amplified Fragment Lenght Polymorphism*), que associam os polimorfismos gerados pelas enzimas com a capacidade de detecção da PCR (do inglês, *Polimerase Chain Reaction* – Reação de Polimerase em Cadeia). Tal técnica possui vantagem de apresentar ampla cobertura do genoma e detectar grande número de *loci*. Um fator importante dos marcadores AFLP é que estes não exigem conhecimento prévio da sequência alvo, como é o caso de marcadores como microssatélites e SNPs. Os AFLPs são dominantes, pois os alelos de um mesmo *locus* são revelados por meio da presença ou ausência de determinada banda no gel e, ao contrário dos marcadores RFLPs, não é possível determinar se o *locus* observado está em homozigose ou heterozigose. Estes marcadores, bem como todos aqueles descritos a seguir, surgiram logo após a invenção da técnica de PCR, em 1987, pois dependem da amplificação exponencial de suas sequências (LOPES et al., 2002; OLIVEIRA et al., 2005).

Os RAPDs (do inglês, *Random Amplified Polymorphic DNA*) utilizam *primers* com sequências aleatórias na amplificação por PCR, tendo como base os *indels* (inserções ou deleções) e as mutações ocorridas nos locais de hibridização dos *primers*, de modo que os padrões de bandas gerados serão altamente distintos. Estes marcadores, desenvolvidos em 1990, são dominantes, e portanto são incapazes de distinguir genótipos homozigóticos ou heterozigóticos, como descrito anteriormente (LOPES et al., 2002).

Os marcadores microssatélites, ou SSR (do inglês, *Simple Sequence Repeats*), são pequenas sequências de DNA repetidas em tandem, e se encontram dispersas por todo o genoma, embora sejam raros nas regiões codificadores. Apresentam característica de co-dominância e alta taxa de variabilidade, o que os tornam marcadores muito informativos. A base das variações deste marcador se encontram no número de unidades repetidas, resultando em sequências com diferentes pares de bases (SCHLOTTERER e PERBENTON, 1998).

Cada marcador molecular possui características distintas e servem para os mais variados objetivos, porém, desde o surgimento dos paineis de genotipagem de alta densidade, o uso de marcadores SNP têm prevalecido nas pesquisas científicas.

2.1.1. Marcadores SNP

Descritos pela primeira vez em 1989, os marcadores SNP consistem na troca de uma única base nitrogenada da sequência de DNA. As transições (substituição de uma purina por outra purina – Adenina por Guanina – ou de pirimidina por pirimidina – Citosina por Timina) são as mutações mais comuns. Já as transversões (troca de uma purina por uma pirimidina ou vice-versa) são menos frequentes. Tais alterações podem acontecer em regiões codificadoras ou regulatórias do genoma, tendo assim maior probabilidade de afetar algum fenótipo, e também em regiões intergênicas sem função determinada (CAETANO 2009; ORITA et al., 1989; VIGNAL et al., 2002).

Os SNPs apresentam duas importantes vantagens em relação a outros marcadores: facilidade de identificação e baixa taxa de mutação. Representam o tipo de polimorfismo mais abundante nos genomas. Um único indivíduo pode ter, espalhado ao longo de seu genoma, milhares destes polimorfismos, presentes em maior número em espécies nas quais não existe grande taxa de endogamia. Estima-se que haja 1 SNP a cada 700 pb em *Bos taurus taurus* e 1 a cada 300 pb em *Bos taurus indicus,* indicando que existem mais variações no genoma do segundo. Porém, para ser considerado um polimorfismo e não apenas uma mutação, o polimorfismo em questão deve ter frequência alélica de pelo menos 1% da população (LI, LI e GUAN, 2008; RESENDE et al., 2008; The Bovine Hapmap Consortium, 2009).

Durante um tempo a utilização deste marcador molecular foi considerada limitada. Até relativamente pouco tempo atrás, a metodologia mais utilizada para identificação de novos SNPs era o seqüenciamento Sanger. Porém, para que isto fosse possível, eram necessários alguns pré-requisitos, tais como a existência de uma sequência consenso que serve de base para comparações com outras sequências (ou seja, o genoma da espécie estudada deve estar seqüenciado), a capacidade de geração de dados dos laboratórios e a capacidade de análise de tais dados. A sequência consenso servia como um “padrão” estabelecido para a espécie em

estudo, então alinhava-se diversas sequências de diferentes indivíduos com a sequência consenso e observava-se alterações em algumas bases nitrogenadas, detectando os SNPs (CAETANO, 2009).

2.2. Avanços nas técnicas de seqüenciamento e surgimento dos paineis de genotipagem de SNPs

Em 1977 foi desenvolvida uma técnica de seqüenciamento que revolucionou todas as áreas de pesquisa das ciências biológicas, uma vez que proporcionou ferramenta capaz de decifrar genes inteiros e, posteriormente, genomas completos. A técnica de sequenciamento Sanger, que recebeu o nome do pesquisador que a desenvolveu, foi o único método de sequenciamento de DNA utilizado durante 30 anos (SANGER, NICKLEN e COULSON, 1977; SCHUSTER, 2008).

Foi por meio da combinação da técnica de sequenciamento de Sanger e BACs (do inglês, *Bacterial Artificial Chromosome*), bem como da técnica de sequenciamento por capilaridade, que o genoma da espécie *Bos taurus taurus* foi desvendado a partir de dois animais da raça *Hereford*. Até o ano de 2009, grande parte dos 2 milhões de SNPs identificados no dbSNP (do inglês, *data bank* SNP – Banco de dados de SNP) era proveniente dos polimorfismos detectados nos dois animais (ECK et al., 2009).

Baseando-se no seqüenciamento Sanger, pesquisadores desenvolveram duas metodologias para a identificação de polimorfismos: a identificação de SNPs distribuídos aleatoriamente pelo genoma (por meio do alinhamento da sequência consenso com sequências aleatórias do genoma estudado) e a identificação de SNPs em regiões específicas do genoma (utilizando amplificação por PCR). Embora sejam eficazes, estas técnicas possuem custos relativamente altos, de US$3,00 e US$10,00 por polimorfismo detectado na primeira e na segunda metodologia, respectivamente (CAETANO, 2009).

Com o surgimento das chamadas tecnologias de sequenciamento de segunda geração, este quadro teve mudança significativa. Capazes de seqüenciar milhões de bases em aproximadamente 48 horas por meio de metodologia extremamente automatizada, conseguiu-se

reduzir o custo para até US$0,48 por polimorfismo detectado, utilizando o seqüenciador *Solexa Genome Analizer*, da empresa Illumina (VAN TASSEL et al., 2008).

Três empresas se destacaram no desenvolvimento de seqüenciadores de segunda geração: a empresa Life Technologies (com o sequenciador *Solid*), a Roche (com o sequenciador *Roche 454*) e a Illumina (com o sequenciador *Solexa*). Apesar de apresentarem performance muito mais eficiente que os métodos de seqüenciamento anteriores, os seqüenciadores de segunda geração ainda exigem que o genoma da espécie a ser estudada esteja seqüenciada, sendo necessário o preparo de uma biblioteca genômica fragmentada aleatoriamente (MARGULIES et al., 2005; SCHENDURE e JI, 2008; VALOUEV et al., 2008).

A partir disto, diversas pesquisas realizadas ao longo dos anos trouxeram informações fundamentais para que os painéis de genotipagem de alta densidade se tornassem realidade. Em pesquisa conduzida por Van-Tassel e colaboradores, por exemplo, utilizou-se 66 bovinos representando diferentes linhagens da raça *Holstein* e 7 raças de corte (*Angus, Red Angus, Charolais, Gelbvieh, Hereford Limousin* e *Simmental*), resultando na identificação de mais de 23.000 SNPs bovinos e sua consequente deposição em bancos de dados mundiais (VAN TASSEL et al., 2008).

O surgimento dos painéis de genotipagem de SNPs de alta densidade trouxe grandes vantagens no que diz respeito à identificação de polimorfismos espalhados no genoma. Capaz de identificar um número muito maior de polimorfismos em apenas algumas horas, o chip é manuseado apenas uma única vez, o que diminui a probabilidade de acontecer erros de manipulação das amostras. A grande vantagem destas plataformas é que não é mais necessário realizar um estudo de mapeamento, que utiliza pelo menos 150 marcadores microssatélites no caso dos bovinos, para que se identifique os polimorfismos. Em conseqüência, o custo deste processo foi reduzido: enquanto um estudo de mapeamento que identifica aproximadamente 18 milhões de pares de bases custa US$10,00 por genótipo, uma plataforma capaz de identificar 50.000 *SNPs* tem o custo unitário menor que US$400,00 por amostra (CAETANO, 2009).

Existem diversas aplicações práticas para os paineis de genotipaem de SNPs, sendo GWAS (do inglês, *Genome Wide Association Study*) e GWS (do inglês, *Genome Wide Selection*) as mais utilizadas na pecuária. Por meio da abordagem conhecida como GWS, produtores do mundo inteiro podem selecionar indivíduos com genótipos considerados superiores para a

produção, capturando todos os genes que afetam um caráter quantitativo. Já GWAS é realizado objetivando-se detectar polimorfismos que possam exercer alguma influência em características de interesse comercial (RESENDE et al., 2008).

Diversos chips estão disponíveis no mercado, variando desde o número de SNPs por plataforma até a espécie a partir do qual o chip foi construído. Duas empresas se destacam na produção destes, a Affymetrix e a Illumina. Ambas tiveram os primeiros chips voltados para estudos em humanos (LI, LI e GUAN, 2008).

Os paineis de genotipagem podem ser caracterizados pelo número de polimorfismos que são capazes de identificar quando comparados ao tamanho total do genoma analisado, de modo que são classificados por apresentarem alta ou baixa densidade de SNPs. Os principais paineis de alta densidade de bovinos atualmente disponíveis no mercado podem ser observados na tabela abaixo:

Tabela 1. Principais plataformas de genotipagem de *SNPs*

Identificação	Número de SNPs	Empresa
GeneChip® Bovine Mapping 10K SNP	~10.000	Affymetrix
Targeted Genotyping® Bovine 25k SNP	~25.000	Affymetrix
Bovine SNP50k® Beadchip	54.609	Illumina
Bovine HD Genotyping BeadChip® 777k	786.798	Illumina

Tais plataformas de genotipagem têm a capacidade de realizar uma revolução na área da genômica animal, uma vez que permitem a varredura de milhares de SNPs ao mesmo tempo, combinando baixo custo, rapidez e eficiência na técnica.

2.3. Parâmetros utilizados no controle de qualidade em dados de SNP

Os painéis de genotipagem de SNP de alta densidade possuem capacidade de gerar grande quantidade de dados, porém vieses na construção dos mesmos, bem como erros cometidos durante o procedimento de genotipagem, podem levar a associações falso-positivas e/ou falso-negativas. Para que tais interpretações errôneas não aconteçam, é necessário que se realize o controle de qualidade nos dados de SNP gerados por meio de tais painéis de genotipagem (ANDERSON et al., 2010).

Determinados parâmetros são estabelecidos a fim de filtrar dados não informativos, tais como a frequência do menor alelo, *call rate* por amostra e por marcador, IBS (do inglês, *Identity By State* – Identidade Por Estado), equilíbrio de *Hardy-Weinberg*, alta correlação entre marcadores e inferência de sexo.

A porcentagem de SNPs cujos genótipos foram adequadamente determinados é denominada "*call rate*". Sendo de fundamental importância, o *call rate* é um dos parâmetros que indica a qualidade da genotipagem.

A frequência do menor alelo (ou MAF) nada mais é do que a determinação da freqüência mínima que este alelo menos comum deve apresentar na população. A importância deste parâmetro consiste no fato de que alelos com baixa freqüência não possuem representatividade da população, tornando muito difícil associações estatísticas destes alelos raros com determinados fenótipos.

Utiliza-se o equilíbrio de *Hardy-Weinberg* a fim de se detectar contaminação das amostras genotipadas. De acordo com as pressuposições deste, as freqüências alélicas e genotípicas de uma população em equilíbrio podem ser estimadas a cada geração. Marcadores que não se encontram em equilíbrio podem ser resultado de erros de genotipagem, estratificação da população, endogamia, entre outros.

O cálculo de identidade por estado (ou IBS) consiste em um método simples de verificar a existência de possíveis falhas, por meio da identificação de indivíduos duplicados ou altamente relacionados. Tal identificação é baseada na proporção de alelos compartilhados em cada par indivíduos (ANDERSON et al., 2010).

A inferência do sexo também é utilizada para verificar existência de falhas de genotipagem, uma vez que compara o gênero genético (fornecido pela genotipagem) e o gênero relatado da amostra. Teoricamente, as duas informações devem estar de acordo uma com a outra. A determinação do sexo se dá por meio da avaliação da taxa de heterozigosidade do cromossomo X, de modo que as fêmeas (por possuírem 2 cromossomos X) terão maior taxa de heterozigosidade que os machos (TURNER et al., 2011).

O cálculo para se estimar a correlação existente entre dois marcadores é baseado no desequilíbrio de ligação, e é utilizado a fim de se eliminar marcadores não informativos do conjunto de dados. O desequilíbrio de ligação consiste na associação de dois ou mais alelos no genoma, que segregam juntos na população (GABRIEL, 2002).

3. OBJETIVO GERAL

Gerar conhecimento referente ao perfil de SNPs por meio de painel de genotipagem de SNP de alta densidade, bem como salientar a importância do controle de qualidade para análise dos dados.

3.1. OBJETIVOS ESPECÍFICOS

- Avaliar parâmetros de controle de qualidade por marcador: *call rate*, frequência do menor mlelo, equilíbrio de *Hardy-Weinberg*, alta correlação entre marcadores

- Avaliar parâmetros de controle de qualidade por amostra: *call rate*, identidade por estado, inferência de sexo

4. MATERIAL E MÉTODOS

4.1. Animais

O grupo amostral foi constituído por 96 fêmeas da raça Nelore, provenientes de um rebanho comercial localizado no município de Iguatemi, Mato Grosso do Sul.

4.2. Colheita das amostras e extração de DNA

As amostras de sangue periférico de cada animal foram coletadas à vácuo por meio de punção da veia jugular e armazenadas à -20°C até o processamento das análises laboratoriais.

A extração do DNA genômico a partir das amostras de sangue foi realizada por meio de kit comercial DNeasy Blood & Tissue (QIAGEN, Valencia, CA, Espanha). A determinação da concentração e avaliação da pureza dos ácidos nucléicos realizou-se por espectrofotometria de micro-volume (NanoDrop 2000, Thermo Scientific). O padrão determinado para as amostras era concentração de 50ng/µL e razão de 1,8. A razão nada mais é do que a divisão entre os valores da absorbância obtidos a 260nm e a 280nm, e considera-se puro o DNA cuja razão se encontra entre 1,8 e 2,0. Tais etapas foram realizadas no Laboratório de Biotecnologia Aplicada à Produção Animal da FCA/UFGD.

4.3. Protocolo de genotipagem

Os procedimentos de genotipagem foram realizados na empresa Deoxi Biotecnologia Ltda® (Araçatuba/SP). Os animais foram genotipados para a caracterização do perfil de SNPs em painel Bovine HD Genotyping BeadChip 777k (Illumina Inc., San Diego, CA), capaz de identificar até 786.798 SNPs. Cada painel é capaz de genotipar 8 amostras (Figura 1), de modo que utilizou-se o total de 12 paineis de genotipagem.

As etapas de genotipagem, baseadas no ensaio *Infinium HD* (Illumina Inc., San Diego, CA), são descritos a seguir:

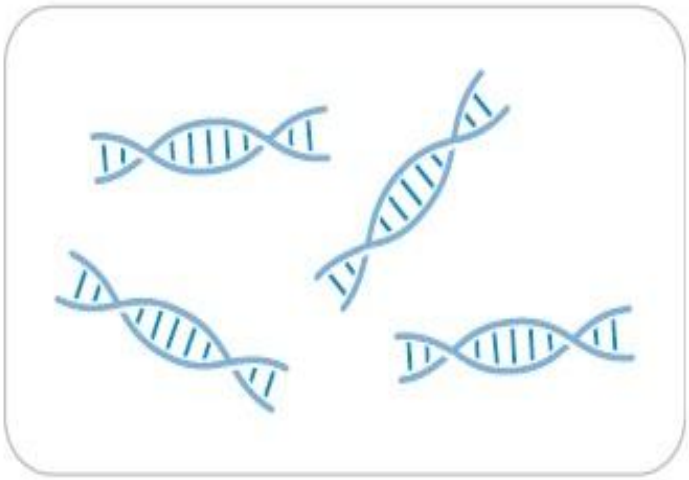
Figura 1. Bovine HD Genotyping BeadChip 777k (Illumina Inc., San Diego, CA)

4.3.1. Desnaturação e amplificação

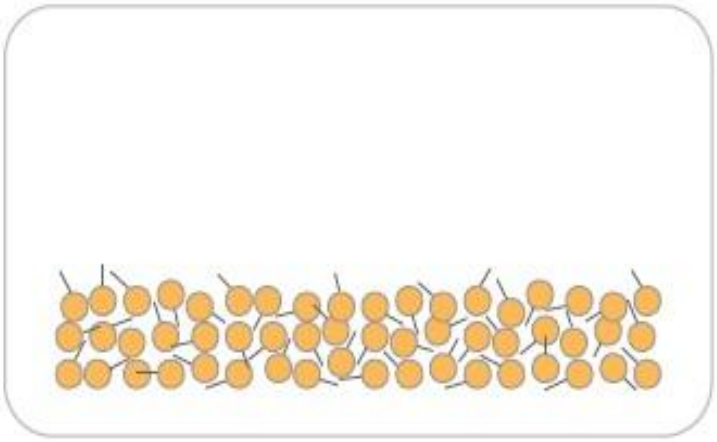
Inicialmente o DNA foi desnaturado e neutralizado com NaOH, que prepara os ácidos nucleicos para posterior amplificação. Em seguida, o DNA desnaturado foi isotermicamente amplificado a 36°C e incubado no equipamento *Hybex Microsample Incubator* (*Scigene* – Sunnyvale, California) por aproximadamente 16 horas. Este tipo de amplificação resulta em maiores quantidades de DNA, utilizando menos reagentes quando comparada à reação de PCR (do inglês, *Polymerase Chain Reaction*).

4.3.2. Fragmentação

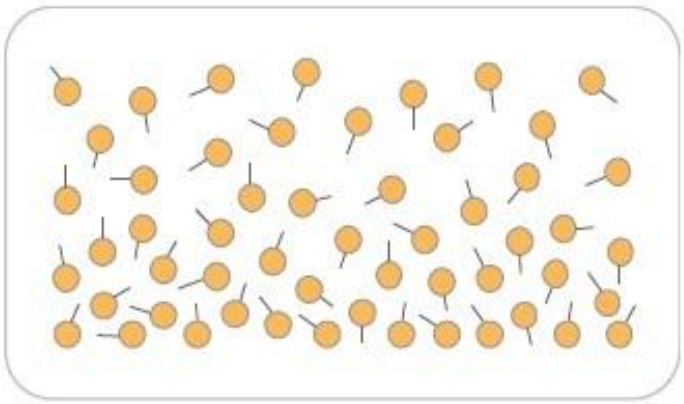
Após as 16h, as amostras amplificadas foram fragmentadas por um processo enzimático que não requer a realização da técnica de eletroforese em gel. Após fragmentação enzimática (Figura 2 – a), foi adicionado isopropanol no DNA (Figura 2 – b) e centrifugado a 4°C, em seguida descartou-se o isopropanol. O DNA foi ressuspendido (Figura 2 – c) em tampão de hibridização.



(a)



(b)



(c)

Figura 2. Ilustração esquemática dos diferentes estados do DNA: (a) fragmentação; (b) precipitação; (c) ressuspensão (Illumina, 2011)

4.3.3. Manuseio do painel de genotipagem

A plataforma de genotipagem foi preparada para hibridização por meio de reagentes do próprio Kit e as amostras foram aplicadas no mesmo (Figura 3 – a). Uma vez completado com as amostras, o chip foi incubado por 16 a 24 horas em forno com temperatura controlada de 47°C (Figura 3 – b). Durante este processo, os fragmentos de DNA das amostras se anelaram às sequências de DNA que se encontram ligadas por ligação covalente à nanopartícula *bead,* de modo que a hibridização de cada alelo com cada *bead* presente no chip representa um *lócus* diferente do DNA.

4.3.4. Lavagem e coloração da plataforma de genotipagem

Após a hibridização foi realizada a lavagem do chip, na qual o DNA não hibridizado foi removido (Figura 3 – c). O procedimento de lavagem durou aproximadamente de 20 minutos. O chip foi então submetido ao procedimento de coloração, que durou aproximadamente 2 horas. O processo de coloração do chip se baseia na extensão de uma única base nucleotídica da sequência hibridizada com cada *bead,* de modo que os corantes serão incorporados determinando os diferentes genótipos. Os reagentes utilizados para coloração eram provenientes de kits comerciais da empresa produtora do chip (Figura 3 – d). Tal procedimento ocorreu em um equipamento específico para tal tarefa, que possui umidade e temperatura controladas (Figura 3 – e). Terminado o processo de coloração, os chips permaneceram em local protegido para secagem durante 1 hora e 30 minutos (Figura 3 – f).

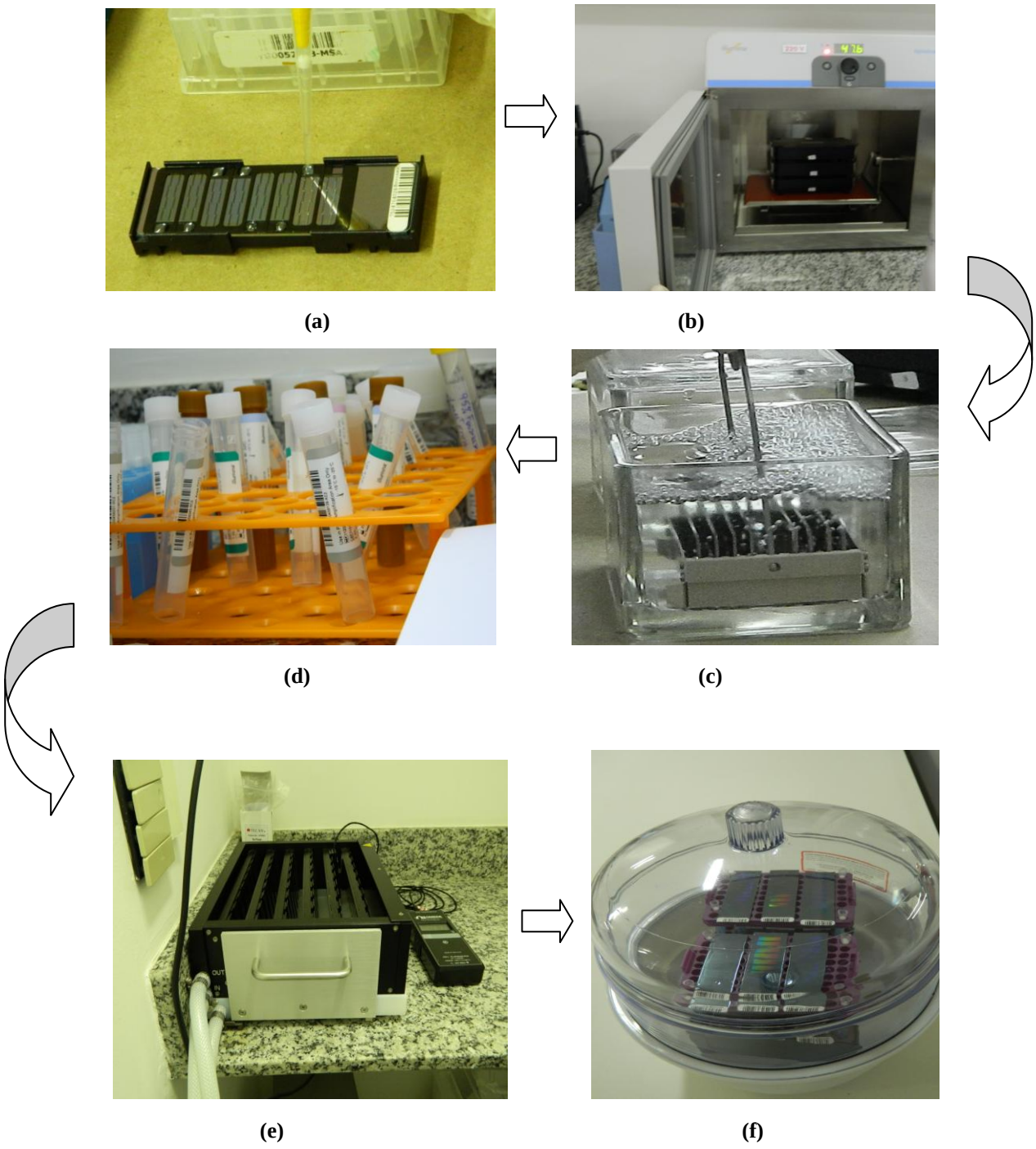


Figura 3. Preparo das amostras e do painel de genotipagem: (a) aplicação das amostras no chip; (b) hibridização do DNA em incubadora; (c) lavagem dos chips em cuba de vidro; (d) utilização de solução para coloração dos chips; (e) incubação de chips para coloração; (f) secagem dos chips em dessecador. **Fotos**: Lara Silva (2012)

4.3.5. Leitura a laser do painel de genotipagem

Os chips secos foram levados para o equipamento *Illumina iScan*® (Figura 3 - a), responsável por interpretar as intensidades de luz. Isto ocorre devido ao fato de as moléculas de fluoróforos que estavam ligados aos *beads* do chip serem excitadas por um raio laser emitido pelo aparelho. Diferentes colorações podem ser observadas e as reações podem ser acompanhadas pelo software *iScan Control*, conforme a Figura 3 – b.

De acordo com os alelos encontrados, haverá diferentes intensidades de coloração no chip, formando imagens de alta resolução da luz emitida. Obtém-se, desta forma, os resultados da hibridização entre o DNA e o chip.

(a) (b)

Figura 4. Visualização do procedimento de escaneamento: (a) Equipamento *Illumina iScan*®; (b) Processo de genotipagem visualizado pelo software *Illumina iScan Control*. **Fotos:** Lara Silva (2012)

4.4. Obtenção dos dados genotípicos

Os dados gerados passaram pelo software *GenomeStudio Data Analysis* (Illumina®), para que este leia e converta as intensidades de fluorescência em genótipos de SNPs. Deste modo, foi gerado um arquivo de intesidade com extensão ".idat", contendo informações a respeito dos genótipos.

Todos os SNPs não autossômicos (relacionados ao cromossomo X, Y e o mitocondrial) foram removidos.

4.5. Controle de qualidade

O controle de qualidade pretende, basicamente, determinar quais indivíduos e quais marcadores devem permanecer no conjunto de dados para as análises posteriores, bem como retirar todos os dados que são considerados não informativos. Tais etapas foram realizadas em ambiente estatístico R (R Development Core Team, 2011), em sistema operacional Linux Ubuntu.

4.5.1. Controle de qualidade por marcador

Alguns marcadores podem ter as mesmas coordenadas no chip, o que provavelmente se deve a erros durante a construção do mesmo. Para que tais marcadores não interfiram negativamente nas análises posteriores, decidiu-se descartar todos os 54 marcadores que apresentavm as mesmas coordenadas.

- MAF (do inglês, *Minor Allelle Frequency)*: marcadores com freqüência alélica menor que 2% na população foram removidos
- *Call Rate*: marcadores que não se encontravam em pelo menos 98% da população foram removidos
- HWE (do inglês, *Hardy Weinberg Equilibrium*): marcadores $p<10^{-5}$ para o teste de Fisher foram excluídos
- Alta correlação dos marcadores: foi analisado o desequilíbrio de ligação entre os marcadores, de modo que, quando $r^2 > 99,5\%$ entre dois marcadores, exclui-se um deles

4.5.2. Controle de qualidade por amostra

- IBS (do inglês, *Identity By State*): indivíduos com mais de 95% de semelhança são descartados do conjunto de dados a ser analisado
- Inferência de sexo: indivíduos que apresentam heterozigosidade do cromossomo X maior que 10% são considerados fêmeas
- *Call rate* por amostra: amostras que tinham menos de 90% de genótipos determinados pelo painel de genotipagem foram desconsideradas para as análises

5. RESULTADOS E DISCUSSÃO

O Bovine HD Genotyping BeadChip 777k possui espaçamento médio de 1 SNP a cada 3,43kb e mediana de 2,68kb. Utilizou-se, para as análises, o total de 742.909 SNPs. No cromossomo X há 40.235 SNPs, 1.423 no cromossomo Y, 346 em DNA mitocondrial e 1.876 no chamado cromossomo "0", ou seja, regiões que não foram mapeadas.

A eficiência dos testes de associação é diretamente influenciada pela etapa de identificação e remoção de marcadores e amostras com baixo *call rate*, uma vez que a presença dos mesmos pode reduzir a habilidade de encontrar associações verdadeiras. Foram removidos 97.334 marcadores que apresentaram *call rate* abaixo de 98%.

É aconselhável adotar *call rate* por marcador entre 93% e 98% em estudos GWA aplicados a humanos. A tecnologia de SNP-Chip foi inicialmente aplicada em humanos, de modo que atualmente as plataformas e os métodos de análise mais bem estabelecidos se aplicam a estes estudos. Há plataformas de genotipagem capazes de identificar mais de 5 milhões de marcadores por amostra, e geralmente o grupo amostral é constituído por milhares de pessoas, tornando os parâmetros de controle de qualidade extremamente rigorosos. Deste modo, pode-se observar que o *call rate* por marcador adotado em estudos com seres humanos é considerado rígido quando comparado aos padrões utilizados em estudos com animais (ANDERSON et al., 2010; Illumina®; TURNER et al., 2011).

O *call rate* estabelecido para as amostras do presente estudo foi de 90%, e a média apresentada pelas mesmas foi de 94%, de modo que nenhuma amostra foi excluída em função deste parâmetro.

Existem estudos que utilizam *call rate* das amostras maior que 95%, ou seja, são mais exigentes e consequentemente descartam maior número de amostras do conjunto de dados (JIANG et al., 2013; NISHIMURA et al., 2012;). Ao se escolher o *call rate* ideal para cada estudo, deve-se levar em consideração fatores importantes como o número amostral e a espécie estudada. Nos dois estudos citados anteriormente, o número amostral e/ou a raça bovina estudada eram propícios para tais valores adotados. Jiang e colaboradores (2013) trabalharam com mais de 600 animais da raça *Holstein* e adotaram *call rate* de 99.9%, nesse caso a raça em questão apresentou maior similaridade com as sequências de DNA utilizadas na plataforma quando

comparados aos animais da raça Nelore. Já Nishimura e colaboradores (2012) utilizaram animais da raça *Japanese Black* adotaram *call rate* de 95%, entretanto o grupo amostral era constituído 1156 animais.

Os SNPs geralmente são bialélicos, de modo que um dos alelos estará presente em menor freqüência na população. Por apresentarem frequência alélica menor que 2% no grupo amostral em estudo, 188.762 SNPs foram removidos do conjunto de dados.

Nishimura e colaboradores (2012) optaram por utilizar frequência do alelo mínimo de 1% em estudo realizado com mais de 1000 bovinos da raça *Japanese Black*, demonstrando que a rigorosidade da MAF adotada neste estudo é bem maior, uma vez que o grupo amostral é composto por 96 animais. De acordo com Anderson e colaboradores (2010), os valores ideais para MAF aplicados em humanos variam de 1% a 2%, o que indica que o valor adotado neste estudo está de acordo com outras pesquisas similares. Marcadores com freqüência alélica abaixo de 1% são extremamente raros e sem poder estatístico, de modo que se recomenda sua exclusão do banco de dados (TURNER et al., 2011).

Nas análises realizadas encontrou-se 2.509 SNPs que não se encontravam em equilíbrio de *Hardy-Weinberg,* que foram excluídos pelo controle de qualidade realizado e que devem portanto ser excluídos das análises posteriores. Na maioria de estudos GWA, marcadores que estão fora do equilíbrio são eliminados, entretanto alguns autores responsáveis por estudos em humanos afirmam que tais marcadores, se apresentarem desvios severos, não devem ser excluídos das análises, mas sim indicados para posteriores estudos de associação. Apesar disso, ainda que tais marcadores não sejam eliminados, não serão analisados segundo os mesmos métodos estatísticos que os outros marcadores, ou seja, não farão parte do mesmo conjunto de dados (ANDERSON et al., 2010; TURNER et al., 2011).

No grupo amostral utilizado, nenhum indivíduo apresentou IBS > 95%, não ocorrendo a remoção de nenhuma amostra.

Todas as amostras analisadas tiveram heterozigosidade maior que 10%, não sendo descartada nenhuma amostra nesta etapa.

Para o cálculo de correlação entre os pares adotou-se valor de r^2 > 99,5%, e 120.686 SNPs foram removidos do conjunto de dados a ser analisado posteriormente.

As plataformas de genotipagens são suscetíveis a erros na sua construção, gerando alguns dados que não são confiáveis. Deste modo, excluiu-se 54 marcadores que continham a mesma coordenada genômica contida na plataforma.

Na Figura 5 foi demonstrada a distribuição dos SNPs nos cromossomos antes e depois do controle de qualidade.

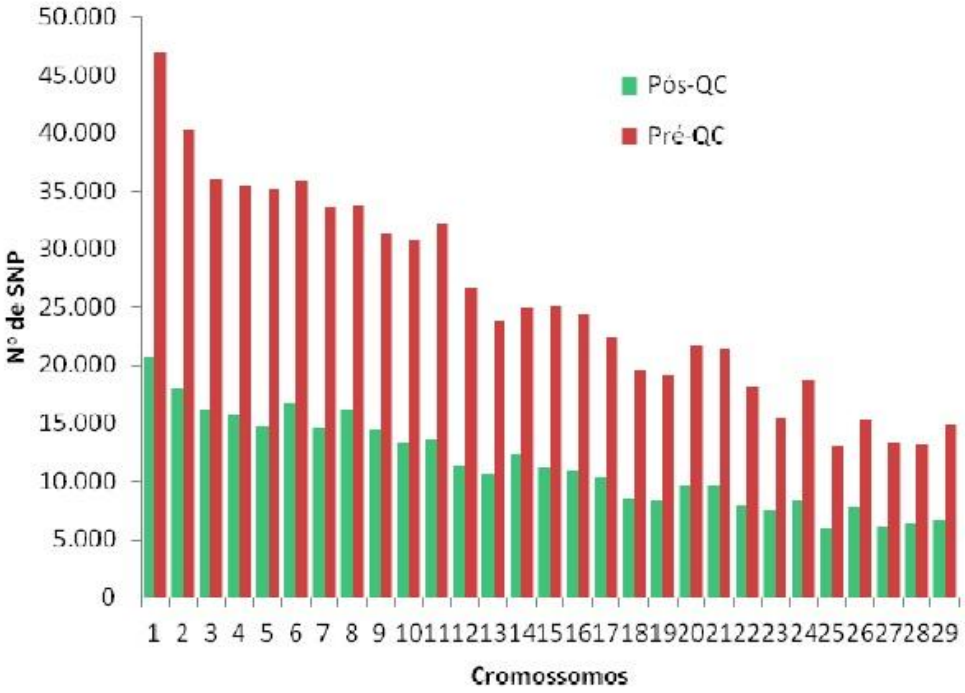


Figura 5. Número de marcadores distribuídos pelos cromossomos antes e depois do controle de qualidade

O controle de qualidade realizado teve como finalidade principal a exclusão de indivíduos e marcadores considerados inadequados para os posteriores estudos de associação. Dos 742.909 *SNPs* que correspondiam ao conjunto de dados inicial submetido ao controle de qualidade, apenas 333.574 permaneceram para os posteriores testes de associação, ou seja, 44, 9% dos marcadores foram considerados pouco informativos ou redundantes segundo os parâmetros estabelecidos, e nesse caso foram exluídos..

De acordo com a Tabela 2, observa-se que o critério responsável pelo maior número de remoção de SNPs foi a MAF. A remoção de marcadores que não apresentaram frequência alélica

acima de 2% se deve, geralmente, a erros na detecção destes marcadores ou à baixa qualidade da determinação do perfil de SNPs (TABANGIN et al., 2009).

Tabela 2. Marcadores removidos por meio do controle de qualidade

Critério Controle de Qualidade	Número de SNPs excluídos
Coordenada Genômica Idêntica	54
HWE (p< 1e-5)	2.509
Call Rate (< 98%)	97.334
Alta correlação aos pares (r^2 > 99,5%,)	120.686
MAF (< 2%)	188.762
TOTAL	**409.335**

Para o estabelecimento dos valores de cada critério analisado nas etapas de controle de qualidade, é fundamental que se leve em consideração o número de indivíduos que será analisado, pois, como mencionado anteriormente, a frequência na qual determinados alelos se encontram na população é um parâmetro de suma importância para as análises estatísticas.

6. CONCLUSÃO

Com base nos resultados obtidos por meio deste estudo pode-se concluir que o controle de qualidade é fundamental como etapa prévia nas análises de associação genômica. Tais critérios se mostraram capazes de impedir a presença de dados não informativos do conjunto de dados quando aplicado em estudo com a raça Nelore, o que facilitará as análises estatísticas posteriores, atribuindo maior confiabilidade às associações que podem ser encontradas.

7. REFERÊNCIAS BIBLIOGRÁFICAS

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

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.

CAPON, F.; ALLEN, M. H.; AMEEN, M.; BURDEN, A. D.; TILLMAN, D.; BARKER, J. N.; TREMBATH, R. C.; A synonymous SNP of the corneodesmosin gene leads to increased mRNA stability and demonstrates association with psoriasis across diverse ethnic groups, **Human Molecular Genetics**, v. 13, p. 2361-2368, 2004

CORNELIS, C.C.; AGRAWAL, A.; COLE, J.W.; et al..; The gene, Environment association studies consortium (GENEVA): Maximizing the knowledge obtained from GWAS by Collaboration Across Studies of multiples conditions, **Genetic epidemiology**, v.34, p.364-372, 2010.

CURI, R. A.; Relação entre os polimorfismos de genes envolvidos no controle do crescimento e na composição da carcaça e características de produção de bovinos de corte no modelo biológico superprecoce, 2004. Tese (Doutor em Genética) – Universidade Estadual Paulista, 2004

ECK, S. H.; BENET-PAGÈS, A.; FLISIKOWSKI, K.; MEITINGER, T.; FRIES, R.; STROM, T. M.; Whole genome sequencing of a single *Bos taurus* animal for single nucleotide polymorphism discovery, **Genome Biology**, v. 10, n. 8, 2009

FERREIRA, M.E.; GRATTAPALIA, D.; Introdução ao uso de marcadores moleculares em análises genéticas, EMBRAPA-CENARGEM: Brasília, 220p, 1998

GABRIEL, S. B.; The structure of haplotypes blocks in the human genome, **Science**, v. 296, p. 2225, 2002

GEORGES, M.; LEQUARRÉ,A. S.; HANSET, R.; VASSART, G.; Genetic variation of the bovine thyroglobulin gene studied at the DNA level. **Animal Genetics**, v.18, p. 41-50, 1987

GRODZICKER, T.; WILLIANS, J.; SHARP, P.; SAMBROOK, J.; Physical mapping of temperature-sensitive mutations of adenovirus, **Cold Spring Harbor Symposium Quantitative Biology**, p.439-446, 1974

HAWKEN, R.J; BARRIS, W.C.; McWILLIAM, S.M.; DALRYMPLE, D.P.; An interactive bovine *in silico* SNP database, **Mammalian Genome**, vol.15, 2004

HEATON, M. P.; HARHAY G. P.; BENETT, G. L.; STONE, R. T.; GROSSE, W. M.; CASAS, E.; KEELE, J. W.; SMITH, T. P. L.; CHITKO-MCKOWN, C. G.; LAEGREID, W. M. Selection and use of SNP markers for animal identification and paternity analysis in U. S. beef cattle. **Mammalian Genome**, v.13, p. 272- 281, 2002.

HERR, A.; GRUTZMANN, R.; MATTHAEI, A.; ARTELT, J.; SCHROCK, E.; RUMP, A.; PILARSKY, C.; High-resolution analysis of chromosomal imbalances using Affymetrix 10K SNP genotyping chip, **Genomics**, v. 85, p. 392-400, 2004

JIANG, L.; JIANG, J.; YANG, J.; LIU, X.; WANG, J.; WANG, H.; DING, X.; LIU, J.; ZHANG, Q.; Genome-wide detection of copy number variations using high-density SNP genotyping platforms in Holsteins, **BMC Genomics**, 2013

LI, M.; LI, C.; GUAN, W.; Evaluation of coverage variation of SNP chips for genome-wide association studies, **European Journal of Human Genetics**, v. 16, p. 635-643, 2008

LOPES, R.; LOPES, M.T.G.; FIGUEIRA, A.V.O.; CAMARGO, L.E.A.; FUNGARO, M.H.P.; CARNEIRO, M.S.; VIEIRA, M.L.C.; Marcadores moleculares dominantes (RAPD e AFLP), **Biotecnologia, Ciência & Desenvolvimento**, n.29, p.56-60, 2002

MARGULIES, M.; EGHOLM, M.; ALTMAN, W.E. et al.; Genome sequencing in microfabricated high-density picolitre reactors, **Nature**, v.437, n.15, p.376-380, 2005

NISHIMURA, S.; WATANABE, T.; MIZOSHITA, K.; TATSUDA, K.; FUJITA, T.; WATANABE, N.; SUGIMOTO, Y.; TAKASUGA, A.; Genome-Wide association study identified three major QTL for carcass weight including the *PLAG 1- CHCHD7* QTN for strature in Japanese Black cattle, **BMC Genetics**, 2012

OLIVEIRA, P.R.D.; SCOTTON, D.M.; NISHIMURA, D.S.; FIGUEIRA, A.; Análise da diversidade genética por AFLP e identificação de marcadores associados à resistência de doenças em videira, **Revista Brasileira de Fruticultura**, v.27, n.3, p.454-457, 2005

ORITA, M.; IWAHANA, H.; KANAZAWA, H.; Detection of polymorphisms of human DNA by gel electrophoresis as single-strand conformation polymorphisms, **Proceedings of the National Academy of Sciences USA,** v.86, n.8, p.2766-2770, 1989

R Development Core Team.; R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, 2011. Disponível em http://www.R-project.org/

RESENDE, M. D. V.; LOPES, P. S.; SILVA, R. L.; PIRES, I. E.; Seleção Genômica Ampla (GWS) e maximização da eficiência do melhoramento genético, **Pesquisa Florestal Brasileira**, n. 56, p. 63-77, 2008

SANGER, F.; NICKLEN, S.; COULSON, A.R.; DNA sequencing with chain-terminating inhibitors, **Proceedings of the National Academy of USA**, v.74, n.12, p.5463-5467, 1977

SCHENDURE, J.; JI, H.; Next-generation DNA sequencing, **Nature Biotechnology**, v.26, n.10, p.1135-1145, 2008

SCHUSTER, S.C.; Next-generation sequencing transforms today's biology, **Nature Methods**, v.5, n.1, p.16-18, 2008

SCHLOTTERER, C.; PERBENTON, J.; The use of microsatellite for genetic analysis for natural populations – a critical review, **Molecular Approaches to Ecology and Evolution**, p.71-86, 1998

TABANGIN, M. E.; WOO, J. G.; MARTIN, L. J.; The effect of minor allele frequency on the likelihood of obtaining false positives, **BMC Proceedings**, v. 3, 2009

THE BOVINE HAPMAP CONSORTIUM; Genome survey of SNP variation uncovers the genetic structure of cattle breeds, **Science**, p.528-532, 2009

TURNER, S.; ARMSTRONG, L. L.; BRADFORD, Y. et al; Quality control procedures for Genome Wide Association Studies, **Curr. Protoc. Hum. Genetics**, 2011

VALOUEV, A.; ICHIKAWA, J.; TONTHAT, T. et al.; A high-resolution nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning, **Genome Research**, v.18, p.1051-1063, 2008

VAN-TASSEL, C. P.; SMITH, T. P. L.; MATUKUMALLI, L. K.; TAYLOR, J. F.; SCHNABEL, R. D.; LAWLEY, C. T.; HAUDENSCHILD, C. D.; MOORE, S. S.; WARREN, W. C.; SONSTEGARD, T. S.; SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries, **Nature Methods**, v. 5, p. 247-252, 2008

VIGNAL, A.; MILAN, D.; SANCRISTOBAL, M.; EGGEN, A.; A review on SNP and other types of molecular markers and their use in animal genetics, **Genetics Selection Evolution**, v. 34, p. 275-305, 2002

WANG, W.Y.S.; BARRATT, B.; CLAYTON, D.G.; TODD, J.A.; Genome wide association studies: theoretical and practical concerns, **Nature reviews**, v.6, p.109-118, 2005.

WEIGEL, D.; MOTT, R.;The 1001 genome project for Arabidopsis thaliana, **Genome biology**, v.10, 2009.