UNIVERSIDADE ESTADUAL PAULISTA

FACULDADE DE CIÊNCIAS AGRÁRIAS E VETERINÁRIAS

CÂMPUS DE JABOTICABAL

IDENTIFICAÇÃO DE POLIMORFISMOS DE NUCLEOTÍDEO ÚNICO A PARTIR DO SEQUENCIAMENTO DO GENOMA COMPLETO DE TOUROS DA RAÇA GIR

Gustavo Henrique Russo

Jaboticabal - SP

UNIVERSIDADE ESTADUAL PAULISTA

FACULDADE DE CIÊNCIAS AGRÁRIAS E VETERINÁRIAS

CÂMPUS DE JABOTICABAL

IDENTIFICAÇÃO DE POLIMORFISMOS DE NUCLEOTÍDEO ÚNICO A PARTIR DO SEQUENCIAMENTO DO GENOMA COMPLETO DE TOUROS DA RAÇA GIR

Gustavo Henrique Russo

Orientador: Prof. Dr. Danísio Prado Munari

Co-orientadora: Ma. Larissa Graciano Braga

Co-orientador: Thomaz Marquez Sena

Trabalho apresentado à Faculdade de Ciências Agrárias e Veterinárias - UNESP, Câmpus de Jaboticabal, para obtenção do título de Bacharel em Ciências Biológicas.

Jaboticabal - SP

12/07/2022

R969i

Russo, Gustavo Henrique

Identificação de polimorfismos de nucleotídeo único a partir do sequenciamento do genoma completo de touros da raça Gir. / Gustavo Henrique Russo. -- Jaboticabal, 2022

20 p.

Trabalho de conclusão de curso (Bacharelado - Ciências Biológicas) - Universidade Estadual Paulista (Unesp), Faculdade de Ciências Agrárias e Veterinárias, Jaboticabal

Orientador: Danisio Prado Munari

Coorientadora: Larissa Graciano Braga

1. Polimorfismo de nucleotídeo único. 2. Genética animal. 3. Sequenciamento de nucleotídeos de alto rendimento. 4. Controle de qualidade. 5. Gado Leiteiro. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Faculdade de Ciências Agrárias e Veterinárias, Jaboticabal. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

unesp

UNIVERSIDADE ESTADUAL PAULISTA

CÂMPUS DE JABOTICABAL

DEPARTAMENTO: Engenharia e Ciência Exatas

CERTIFICADO DE APROVAÇÃO

TRABALHO DE CONCLUSÃO DE CURSO

TÍTULO: Identificação de polimorfismos de nucleotídeo único a partir do sequenciamento do genoma completo de touros da raça Gir.

ACADÊMICO: Gustavo Henrique Russo

CURSO: Ciências Biológicas

ORIENTADOR (ES): Prof. Dr. Danísio Prado Munari
MSc. Larissa Graciano Braga
MSc. Thomaz Marques Sena

Aprovado e corrigido de acordo com as sugestões da Banca Examinadora

BANCA EXAMINADORA:

	(Nomes)	(Assinaturas)
Presidente	MSc. Larissa Graciano Braga	
Membro	MSc. Ana Carolina Almeida Rollo de Paz	
Membro	Dr. Rafael Nakamura Watanabe	

Jaboticabal 05 / 08 / 2022

Aprovado "ad referendum" do Conselho do Departamento em: 5 / 8 /2022.

Prof. Dr. Danísio Prado Munari – chefe de departamento de ensino

Agradecimentos

Começo escrevendo essas dedicatórias a 10.000 mil pés de altura, no meu primeiro voo de avião, indo para meu primeiro congresso científico, pois já estava no mestrado quando conclui esse trabalho.

Primeiramente dedico esses agradecimentos a minha mãe e meu pai, que antes de tudo decidiram me dar a dádiva da vida, decisão está que torna tudo possível até o momento que escrevo essas dedicatórias. Pai e mãe, obrigado por todo esforço, empenho e responsabilidade, coragem e amor acima de tudo, que empregaram aos meus cuidados e educação, não sou nada além, portanto, nem mesmo esse trabalho, do que fruto do empenho e esforço de vocês. Vocês são as pessoas mais honestas e humildes que eu conheço. E também a minha irmã Milena que sempre esteve ao meu lado para superar os problemas que vivemos e sempre me enche de orgulho ainda mais agora por ter escolhido ser mais uma bióloga na família e ser aprovada no curso de Ciências Biológicas da UFSCar.

Em segundo quero agradecer a mim mesmo, pois apesar de todo a dificuldade e a luta comigo mesmo, permaneci persistente, teimoso e curioso até o final desse trabalho e graduação, ainda que tudo tenha sido "feito ao meu tempo", mas creio que o importante de qualquer tarefa que a vida nos entrega, é ao menos tentar e com sorte concluir.

Agradeço a minha querida Tia Sônia, que é uma forte influência na minha vida, quem me deu os primeiros livros que tive na infância e quem me inspirou a ser interessado nos estudos, no conhecimento, na cultura, na arte, na política e na solidariedade com aqueles mais necessitados que estão à margem da nossa sociedade, não tenho palavras para você Tia, tu és um ser humano incrível.

Agradeço também a minha namorada Vitória (ou Angel), por todo carinho, amor, dedicação, por todo apoio e ajuda que foram fundamentais para a conclusão desse trabalho, e também pela finalização incrível no designer que você deu nos slides da apresentação desse TCC. Você é uma das pessoas mais maravilhosas que a vida me apresentou, ter te conhecido foi uma das grandes sortes que eu tive.

Obrigado por ser essa pessoa tão amável, educada e carinhosa que você é. Além de tudo muito inteligente, dedicada e também engraçada.

Agradeço aos meus amigos queridos sem o apoio de vocês eu não teria chegado até aqui, principalmente aos amigos da República Tua Ksa, por esses cinco anos de convivência, vocês são hoje a minha segunda família, obrigado por terem sempre me influenciado a ser dedicado à faculdade, pelo apoio emocional que sempre foi crucial para mim, pelo companheirismo e por todos as vezes que vocês me fizeram rir, como eu sempre digo nesses cinco anos eu nunca passei nenhum dia sem dar risada naquela casa.

Principalmente ao Cèléra, quem me fez gostar de genética e sempre foi um professor para mim, Mela-Mina meu professor particular de bioquímica e pessoa única, Biro-Juice, Aleluia, Sandy, Graza-Deus, Indyana, Ô-zama, Sid e Ai-Fudi meus companheiros de quarto com quem sempre tive ótimas conversas e desabafos, Pepita por todas as risadas, conversas, aventuras culinárias, músicas que ouvimos, cafés e cervejas que bebemos juntos, kpí-tche pelas longas conversas, aventuras de coletas em campo que encaramos e viagens que fizemos você sempre somou muito conhecimento para mim e admiro sua inteligência, Miusa, Falcatrua, Só-Caminha, Mussum, Cookies, Magia, Livinho, Raxadura, Deslizo, Xapeleiro, Calisto e a todos os da nova geração também que serão o futuro da nossa República.

Aos meus amigos de São Carlos, Augusto, Lucas, Bruno, Gustavo, Edgar, Felipe, Et, Brabo, Yuri, Cadu, Gustavo G., com os quais cresci ao meu lado desde a infância e na minha adolescência. Vocês são os mais doidos que eu já conheci e as mentes mais brilhantes também, o senso de realidade que vocês têm não encontrei em mais ninguém, sem palavras para vocês.

Agradeço imensamente também ao Professor Danísio, Thomaz e Larissa por terem me aberto essa porta que foi um grande passo para mim, obrigado por toda orientação e aprendizado que obtive com vocês, mesmo com a maior parte da nossa comunicação sendo distância infelizmente, devido a pandemia, obrigado por todo apoio dado. E a todos os membros EAGMA, desejo sorte e sucesso na trajetória de vocês.

Agradeço a Embrapa Gado de Leite, CNPq (processo 431629/2016-1), Laboratório Multiusuário de Bioinformática da Embrapa Agricultura Digital pela infraestrutura computacional e recurso de TI, especialmente a Adhemar Neto e Leandro Cintra. Agradeço aos órgãos de fomento CNPq e a Pró-Reitoria de pesquisa (PROPe) pela bolsa concedida para a iniciação científica e também à CAPES pelas bolsas concedidas ao meus coorientadores.

A Unesp-Jaboticabal e a todos seus ótimos professores que foram essenciais nesses anos de formação, obrigado por todos os ensinamentos e conhecimento passado.

Sumário

IDENTIFICAÇÃO DE POLIMORFISMOS DE NUCLEOTÍDEO ÚNICO A PARTIR DO SEQUENCIAMENTO DO GENOMA COMPLETO DE TOUROS DA RAÇA GIR

RESUMO – A chamada de variantes é o processo pelo qual novas variantes são identificadas quando sequências são alinhadas a um genoma de referência. No entanto, após descobertas, essas variantes necessitam passar por um controle de qualidade a fim de retirar possíveis chamadas falso-positivas. Neste trabalho, o objetivo foi relatar o efeito de diferentes valores da profundidade de leitura mínima na filtragem de Variantes de Nucleotídeo Único (SNV) em sequências de genoma completo de touros da raça Gir. Neste estudo foi sequenciado o genoma completo de 30 touros da raça Gir, em que foram produzidas leituras de 2x150 pb, totalizando média de 16,7x de cobertura entre as amostras. As leituras foram alinhadas ao genoma de referência ARS-UCD 1.2 e consecutivamente foi realizada a chamada de variantes pela opção HaplotypeCaller do GATK. Após a chamada de variantes, essas foram armazenadas em um arquivo no formato VCF contendo as variantes dos 30 animais amostrados. Esse arquivo passou por processo de filtragem de variantes, que foi dividido em cinco diferentes controles de qualidade denominados “QC_A, QC_B, QC_C, QC_D e QC_E”. Esses controles diferiram entre si quanto à profundidade mínima de leitura dos sítios onde foram detectadas as variantes. A média de SNVs para os 30 animais amostrados foi de 17.757.945. A amostra com menor número de SNVs foi de 17.123.018 e a com maior número obteve 20.617.141 SNVs. O arquivo VCF *raw* (cru) apresentou um total de 38.597.271 SNVs. Este número foi reduzido consecutivamente após a aplicação dos controles de qualidade, do menos restritivo ao mais restritivo, demonstrando que há variação da cobertura na montagem do genoma. Esse estudo traz uma perspectiva do efeito da profundidade de leitura na exclusão de variantes que podem representar possíveis chamadas falso-positivas.

IDENTIFICATION OF SINGLE NUCLEOTIDE POLYMORPHISM FROM THE WHOLE GENOME SEQUENCING OF GIR BREED BULLS

ABSTRACT – Variant calling is the process by which new variants are identified when sequences are aligned to a reference genome. However, after these variants are discovered, they need to undergo a quality control in order to remove possible false-positive calls from the variant file. In this study, the aim was to report the effect of different minimum read depth values on the filtering of Single Nucleotide Variants (SNV) in whole genome sequences of Gir bulls. In this study, the whole genome of 30 Gir bulls was sequenced, obtained from the Illumina NovaSeq 6000 platform, in which 2x150 bp readings were produced, totaling an average of 16.7x coverage between samples. The readings were submitted to quality control following the recommended parameters from the protocol of the 1000 Bull Genomes Project, aligned to the ARS-UCD 1.2 reference genome and consecutively, variants were called by the HaplotypeCaller option of the GATK software. After the variant calling, these were saved in a VCF file containing the variants of the 30 animals considered in this study. This file went through the process of variant filtering, which was divided into five different quality controls named “QC_A, QC_B, QC_C, QC_D and QC_E”. These, in turn, had practically the same combination of parameters as those most used in filtering variants, but, however, they differed from each other in terms of the minimum reading depth of the sites where the variants were detected. The mean SNV for the 30 animals sampled was 17,757,945. The sample with the lowest SNV count was 17,123,018 and the one with the highest count had 20,617,141 SNVs. The raw VCF file, in other words, the file containing all the variants but which had not yet gone through any filtering process, had a total of 38,597,271 SNVs. This number was consecutively reduced after the quality controls application, from the least restrictive to the most restrictive, demonstrating that there is variation in genome assembly. This study provides a perspective on the effect of read depth in excluding variants that may represent possible false-positive calls.

1.Introdução

O projeto de sequenciamento do genoma bovino foi concluído em 2009 (BOVINE GENOME SEQUENCING AND ANALYSIS CONSORTIUM et al, 2009) e foi um dos primeiros genomas de mamíferos sequenciados, uma vez que os bovinos são um dos principais animais de produção e fornecedores de proteína animal, também por conta do seu posicionamento filogenético representando a ordem *Ruminantia* (TELLAM et al, 2009). Desde então, outros grupos de pesquisa apresentaram novas versões de genomas de referência.

O consórcio de sequenciamento do genoma bovino apresentou os genomas de referência bovinos, UMD e Btau (LIU et al, 2009; ZIMIN et al, 2009). A versão ARS-UCD1.2 resultou em uma montagem mais completa quando comparada ao UMD 3.1.1. Esta versão apresentou contigs maiores (N50= 26,3 versus 0,097 megabases) e menor número de gaps (393 versus 72.051) (ZIMIN et al, 2009; ROSEN et al, 2020).

As tecnologias de sequenciamento ou sequenciamento de alto rendimento tem possibilitado que o genoma completo de diversas raças bovinas fosse resequenciadas para identificar números consideráveis de variantes como: polimorfismo de nucleotídeo único (SNP) e inserções e deleções (InDels) (ECK et al, 2009; KAWAHARA-MIKI et al, 2011; STOTHARD et al, 2011; CHOI et al, 2013; STAFUZZA et al, 2017; IQBAL et al, 2019).

O surgimento de versões mais completas do genoma de referência permite, por si, descobertas de novas variantes, uma vez que são preenchidas as lacunas do genoma. Segundo Liu et al. (2020), a incorporação de novas variantes observadas em dados de sequenciamento também pode ser utilizada para ampliar a cobertura presente em painéis de genotipagem, tendo como principal benefício o aumento em acurácia na predição dos valores genéticos.

Um SNP é uma mutação pontual de um único nucleotídeo de DNA genômico. Entretanto para ser considerado um polimorfismo a frequência dentro de uma população do alelo menor deve ser maior que um limite (geralmente 1-5%), o que a distingue das variantes raras. Contudo, há um movimento na área da genética e da oncologia para generalizar um termo mais amplo: SNV (*single*

nucleotide variant) em oposição a SNP, que por sua vez engloba mutações comuns, mas também raras (CLINE; KARCHIN, 2011).

A chamada de variantes genômicas tem como objetivo determinar com acurácia as posições em que pelo menos uma das bases de uma amostra de DNA genômico difere em relação a uma sequência de referência, ou também conhecida como chamada de SNP ou *SNP calling* (NIELSEN et al, 2011). O arquivo de entrada normalmente é um conjunto de leituras alinhadas salvas em um arquivo do formato BAM ou semelhante. A chamada de variantes é um conjunto de estratégias algorítmicas baseadas nos tipos de variantes em que se busca, como variantes de nucleotídeo único (SNV), pequenas inserções ou deleções (InDels), variação no número de cópias (CNVs) e grandes alterações estruturais (inserções, inversões e translocações). A acurácia da chamada de variantes é altamente dependente da qualidade das bases chamadas e dos alinhamentos das leituras, para isso a recalibração de escore de qualidade de base, como processamento anterior à chamada de variantes, é frequentemente usada para garantir chamadas de variantes acuradas e eficientes (ROY et al, 2018).

Variantes falso-positivas podem ocorrer, e para isso é necessário um procedimento de controle de qualidade pós chamada de variantes. Esse processo exclui variantes potencialmente falso-positivas do arquivo VCF original (*raw*), com base em vários alinhamentos de sequência e chamadas de variantes associadas a metadados, tais como qualidade de mapeamento, qualidade de chamada de base, profundidade de leitura, entre outros (ROY et al, 2018). Porém, os parâmetros a serem utilizados em uma filtragem de variantes ainda não são bem definidos, e o assunto ainda precisa de mais estudos e detalhamentos, bem como uma melhor descrição do efeito individual de cada parâmetro, como por exemplo a profundidade de leitura, na exclusão das variantes falso-positivas.

2. Objetivo

O objetivo deste trabalho foi relatar o efeito de diferentes valores da profundidade de leitura mínima na filtragem de Variantes de Nucleotídeo Único (SNV) em sequências de genoma completo de touros da raça Gir

3.Revisão de literatura.

3.1 A raça Gir.

O Gir é uma raça zebuína nativa da Índia, especificamente da península de Kathiawar (oeste) que foi trazida ao Brasil no século XIX, entre os anos de 1870 e 1962, nesse período 6262 animais zebuínos foram importados. Desses, aproximadamente 700 animais eram oriundos da raça Gir (SANTANA et al., 2014).

O gado zebuíno demonstrou capacidade de adaptação ao clima brasileiro, e logo sua população cresceu rapidamente. Esse processo rápido se deu inicialmente pelo uso de fêmeas disponíveis no Brasil, como as do gado Crioulo, que derivam do gado ibérico (O'BRIEN et al., 2015). A raça Gir representa cerca de 10% do rebanho zebuíno nacional e é reconhecida como a raça zebuína com maior capacidade leiteira, o que tem favorecido seu uso na produção de leite (SOUZA SOBRINHO et al., 2003).

Em 1938, a Associação Brasileira de Criadores Zebuínos (ABCZ) situada em Uberaba-MG, criou o livro genealógico das raças zebuínas, contribuindo para a regularização e disseminação da raça Gir pelo país. Nos anos 60, parte dos criadores selecionavam animais para dupla aptidão (carne e leite) enquanto outro grupo menor selecionava apenas para a produção de leite, o que deu origem ao Gir Leiteiro. Após décadas de seleção, foi criado em 1985 o Programa Nacional de Melhoramento do Gir Leiteiro (PNMGL) (SANTANA et al., 2014).

A partir de 2018, foi iniciado o uso da informação genômica na avaliação genética do Gir Leiteiro. O PNMGL tem avaliado características associadas à produção de leite e seus componentes (proteína, gordura e sólidos), temperamento, conformação corporal, sanidade e longevidade. O PNMGL utiliza também marcadores moleculares para a predição de valores genéticos genômicos de machos e de fêmeas (PANETTO et al., 2020), o que justifica a importância de estudos genômicos aplicados a essa raça.

3.2 Variantes de nucleotídeo único.

Os marcadores SNPs se dão pela alteração de uma única base em uma sequência de DNA genômico, no qual existem diferentes sequências alternativas (alelos) em indivíduos de determinada população, com uma alternativa comum de

dois nucleotídeos possíveis em determinada posição. Entretanto, para que essa mudança de base seja considerada um polimorfismo, é necessário que o alelo menos frequente tenha uma frequência maior que 1% em uma população (BROOKES, 1999).

Usualmente os SNPs são considerados bi-alélicos e a probabilidade de duas mudanças de base independentes ocorrer numa mesma posição é muito baixa. Outra razão para isso é o fato de existir viés nas mutações visto que há o dobro de possibilidades de transversões em relação a transições. Em mutações aleatórias espera-se que a proporção das transições sobre as transversões seja de 0,5, entretanto os estudos têm demonstrado claro viés para transições. Possivelmente esse viés está relacionado a alta taxa espontânea de desaminação de 5-metil citosina (5mC) em timidina nos dinucleotídeos CpG, resultando na geração de níveis mais elevados de SNPs (C ↔T), vistos como SNPs (G↔A) na fita reversa (VIGNAL et al, 2002; BROOKES, 1999).

Os SNVs podem apresentar um impacto funcional de diversas formas, dependendo da região do DNA em que ocorrerem e pode afetar a maquinaria transcricional de uma célula, se estiverem em regiões que contenham sinais reconhecidos por fatores de transcrição e/ou potenciadores transativacionais. Se ocorrer em um sítio de *splicing*, em um sítio que se ligam intensificadores ou repressores de *splicing* exônicos ou intrônicos, pode resultar em isoformas de *splicing* alternativas ou aberrantes de um gene transcrito. Que pode interferir na maquinaria de tradução da célula. Os SNVs que ocorrem em códons que alteram aminoácidos podem interferir no dobramento, localização, estabilidade, ligação ou catálise da proteína (CLINE; KARCHIN, 2011).

3.3 O arquivo de variantes

O arquivo VCF ou formato de chamada de variante (*variant call format*) é um arquivo com formato padronizado para armazenar os tipos predominantes de variação de sequências, incluindo SNPs e InDels, juntamente com anotações de enriquecimento de genes e vias metabólicas. O formato foi desenvolvido com a intenção inicial de representar a variação genética humana, mas seu uso não se limita a genomas diplóides e pode ser aplicado a diferentes contextos. Sua flexibilidade e extensibilidade do usuário permitem a representação de uma ampla

diversidade de variação genômica em relação a uma única sequência de referência (DANECEK et al, 2011).

3.4 Descoberta de novas variantes genômicas e aplicação no melhoramento genético animal.

Dados de sequenciamento possibilitam o contínuo progresso genético e o aumento nas acurácias das predições genômicas (ZHANG et al., 2016). As características influenciadas por alelos raros na população ou causadas por mutações recentes podem ser as mais beneficiadas pelo uso de sequenciamentos de alto rendimento, porque essas variantes podem não estar presentes nos painéis de marcadores SNP. Para as demais características são esperados ganhos com melhor estimação da variabilidade genética e maior persistência nas estimativas do efeito dos marcadores genéticos, além de GWAS que identifiquem as mutações causais e possibilitem maior compreensão dos mecanismos de herança (HAYES; GODDARD, 2010).

Diferentes estudos para avaliação de sequências genômicas foram conduzidos com raças zebuínas, tais como N elore (CANAVEZ et al, 2012), Brahman (BARRIS et al, 2012), Gir (LIAO et al 2013), Guzerá e Girolando (STAFUZZA et al, 2017), e raças africanas (KIM et al, 2017). Liao et al. (2013), em um "*pool*" gamético com 14 animais sendo três da raça Gir, identificaram 9.990.733 variantes, 604 inserções e 308 deleções. Entre as variantes, 62,34% representaram novos polimorfismos no genoma e 83,62% das inserções e deleções ainda não haviam sido reportadas.

4. Material e Métodos

Amostras de DNA, obtidas a partir de sangue e sêmen, foram extraídas de 30 touros da raça Gir Leiteiro (obtidas por meio de financiamento, processo CNPq 431629/2016-1). A qualidade e a normalização do material extraído foram avaliadas por meio de fluorescência no Qubit fluorometer 2.0 (Life technologies, Grand Island, NY). Para a preparação da biblioteca foi utilizada a plataforma IlluminaTruSeq Nano kit (Illumina Inc., San Diego, CA, USA), sendo cada amostra marcada com um código de barras (*barcode*).

O sequenciamento do genoma completo dos animais avaliados foi realizado na plataforma Illumina NovaSeq 6000 (Illumina Inc., San Diego, CA, USA), que gerou bibliotecas do tipo *paired-end*, ou seja, houve sequenciamento de ambas as extremidades dos fragmentos de DNA, em que foram produzidas leituras com tamanho de 2 x 150 pb, totalizando uma média de 16,7 vezes de cobertura por amostra. Os procedimentos de construção das bibliotecas foram realizados de acordo com os protocolos sugeridos pelo fabricante.

A qualidade dos dados de sequenciamento dos 30 touros da raça Gir foi verificada pela ferramenta FastQC v.0.11.8 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Com base nos resultados do FastQC, foram aplicados aos dados, os critérios de controle de qualidade das leituras por meio do programa SeqyClean (ZHBANNIKOV et al., 2017). Conforme recomendado pelo protocolo do 1000 Bull Genomes Project, foram removidas:(1) leituras com três ou mais bases não identificadas (N) nas sequências; (2) média de qualidade para *phred score* inferior ou igual a 20 ou seja, a probabilidade média de que as bases estejam incorretas foi de no mínimo 0,01; (3) comprimento menor do que 50 bases nas sequências. Além da remoção das leituras com baixa qualidade, (4) foram removidas as sequências de adaptadores e possíveis contaminantes. Com a remoção dos adaptadores, buscou-se evitar que estas sequências fossem incorporadas no alinhamento do genoma, o que poderia gerar mapeamento errôneo das bases no genoma.

Com base nas sequências resultantes do controle de qualidade, foi iniciado o alinhamento das sequências com base no genoma referência bovino ARS-UCD 1.2 (https://sites.ualberta.ca/~stothard/1000_bull_genomes/) mediante recomendações de parâmetros do 1000 Bull Genomes Project (http://www.1000bullgenomes.com/) por meio do algoritmo BWA opção MEM (v. 0.7.15-r1144-dirty) (LI; DURBIN, 2009).

As estatísticas dos alinhamentos foram realizadas por meio da ferramenta Samtools na opção Flagstat e a conversão para o formato binário, ordenação e indexação foi realizada pelo Samtools (v. 1.8) (LI et al., 2009; LI, 2011), por meio das opções *view*, *sort* e *index*, respectivamente.

As duplicatas ópticas e de PCR foram removidas pela opção *MarkDuplicates* do Picard Tools (v. 2.18.2-SNAPSHOT) (“Picard toolkit”, 2019). A opção *flagstat* do

Samtools (v. 1.8) (LI et al., 2009; LI, 2011) e um “script” em linguagem *perl* foram utilizados para o cálculo de estatísticas do alinhamento e da cobertura de alinhamento no genoma, respectivamente.

A recalibração do escore de qualidade das bases foi realizada pelo BaseRecalibrator e PrintReads do Genome Analysis Toolkit (GATK, v. 3.8-1-0-gf15c1c3ef), resultando em arquivos com maior confiabilidade por base. Todas as etapas seguiram as recomendações de parâmetros do 1000 Bull Genomes Project (http://www.1000bullgenomes.com/). Consecutivamente foi realizada a chamada de variantes pela opção HaplotypeCaller do GATK (v. 3.8-1-0-gf15c1c3ef), contudo os arquivos de saída dessa etapa o “g.vcf” foram combinados e convertidos para o formato “vcf” pela opção “Combine GVCF” do GATK.

Após a geração do arquivo VCF (raw) com as variantes chamadas dos 30 animais, foram aplicadas as opções de filtragem no arquivo de variantes por meio dos *softwares* VCFtools (versão 0.1.15) (DANECEK et al., 2011), pela opção vcf annotate do VCFtools e bcftools.

Cinco cenários de filtragem foram aplicados após a chamada de variantes (QC_A, QC_B, QC_C, QC_D e QC_E). Para todas as opções foram aplicados filtros que removeram variantes seguindo os critérios: variantes com escore de qualidade média menor que 20, qualidade de mapeamento menor que 30, variantes com distância menor que 10 pb uma da outra, variantes que não estivessem presentes em ao menos quatro animais. Também foram removidas para todas os cinco cenários, variantes que apresentaram valor de profundidade máxima maior que a média das coberturas mais 3 vezes o desvio padrão. Os parâmetros escolhidos para as filtragens tiveram como base os trabalhos de Nascimento (2018) e Larmer (2016) e também foram utilizados em nosso grupo de pesquisa.

As filtragens diferiram quanto ao valor da profundidade mínima de cobertura das variantes. No QC_A foram removidas as variantes com valor de cobertura mínima menor que 10 vezes; QC_B: variantes com valor de cobertura mínima menor que 10% da cobertura média (2); QC_C: variantes com valor de cobertura mínima menor que 25% (4) da média das coberturas; QC_D: variantes com cobertura mínima menor que 50% (8) da média das coberturas; QC_E: variantes com cobertura mínima menor 75% (13) da média das coberturas (os números entre parênteses são os valores reais da cobertura correspondente a cada porcentagem).

5. Resultados e Discussão

Os resultados da chamada de variantes são apresentados na Tabela 1. A média de leituras obtidas entre as amostras foi de, com máximo de 486.209.902 e o mínimo de 273.119.499. Em média 99,527 % das leituras foram mapeadas e em média 92,28% das leituras mapearam de forma esperada. A cobertura média foi de 16,83x com mínimo de cobertura em 11,6x e máximo de 25x. A dispersão dos valores de cobertura entre as amostras está representada na Figura 1**.**

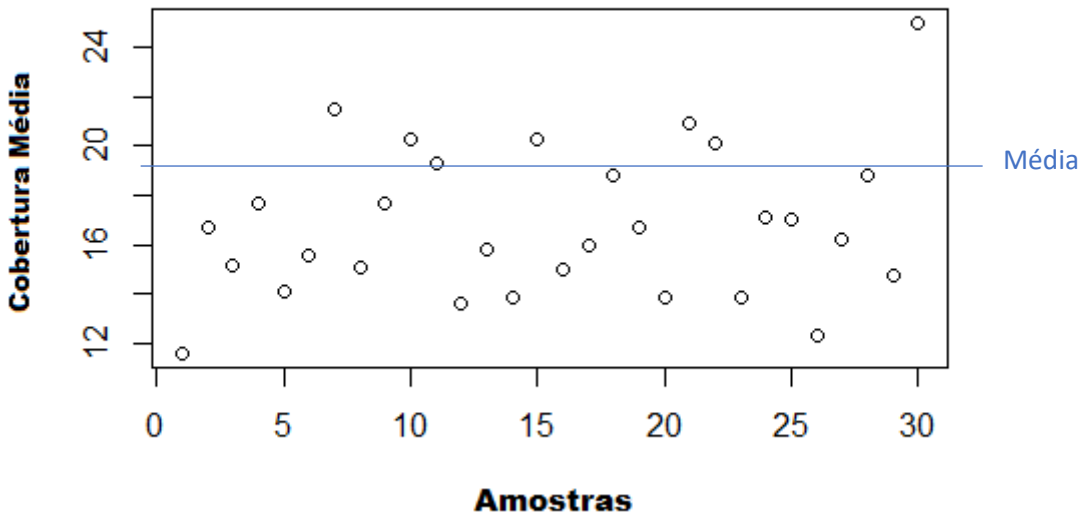

Figura 1. Dispersão dos valores de cobertura da montagem de genoma para cada um dos 30 animais amostrados.

A média de SNV para os 30 animais amostrados foi de 17.757.945 (DP=701.900). A amostra com menor número de SNV foi de 17.123.018 e a com maior número obteve 20.617.141 SNV. Para as InDels a média encontrada foi de 2.543.052(DP=90.288,97) com máximo em 2.780.002 e mínimo de 2.337.132. A taxa de Transições/Transversões (TS/TV) apresentou média de 2,13 (DP=0,032) e com mínimo em 2,02 e máximo em 2,17.

Tabela 1. Resultados da chamada de variantes por animal amostrado (arquivo bruto, ou seja, valores obtidos anteriormente à filtragem de variantes).

Amostra	Total de Leituras	LM*	LPP*	Cb*	Nº de SNV	Nº de InDels	Nº de SM*	Taxa TS/TV
1501	313.768.952	99,74	75,59	11,6	17.700.006	2.489.221	20.128.554	2,16
1447	359.056.254	99,7	90,39	16,7	17.337.445	2.499.907	19.773.822	2,14
1442	307.936.476	97,79	95,76	15,2	17.280.109	2.446.450	19.665.444	2,13
1440	366.598.393	99,87	95,76	17,7	17.652.742	2.500.769	20.087.671	2,12
1439	325.143.647	99,79	86,33	14,1	18.063.334	2.557.034	20.555.588	2,15
1438	367.083.050	99,72	84,38	15,6	17.908.330	2.570.758	20.412.602	2,15
1295	452.832.085	99,81	95,98	21,5	18.109.611	2.626.993	20.663.179	2,11
1258	310.223.610	99,79	92,94	15,1	17.414.037	2.533.233	19.881.178	2,16
1200	364.206.965	99,79	93,89	17,7	17.583.648	2.539.501	20.057.864	2,14
907	434.008.561	99,88	91,13	20,3	18.069.960	2.635.039	20.631.883	2,11
906	395.116.189	99,78	95,81	19,3	17.652.024	2.566.234	20.150.489	2,13
893	273.119.499	99,79	95,04	13,6	17.251.472	2.476.778	19.666.686	2,17
890	321.650.759	99,85	94,35	15,8	17.464.609	2.535.757	19.934.195	2,15
870	276.449.451	99,8	96,56	13,9	17.303.082	2.488.083	19.728.884	2,17
863	401.851.858	99,81	97,98	20,3	17.701.455	2.592.860	20.223.761	2,12
846	323.707.726	96,59	89,11	15	17.571.990	2.547.458	20.053.339	2,15
831	325.540.150	99,7	93,32	16	17.197.606	2.513.543	19.646.532	2,15
823	365.787.842	99,79	97,3	18,8	17.616.137	2.585.655	20.131.683	2,13
805	328.843.430	97,56	95,54	16,7	17.463.798	2.545.232	19.943.600	2,16
791	312.239.045	99,63	88,37	13,9	19.390.149	2.754.422	22.064.850	2,17
789	404.099.122	99,72	97,8	20,9	17.607.216	2.581.223	20.117.646	2,12
756	405.614.167	99,87	96,8	20,1	17.123.018	2.503.315	19.560.233	2,11
754	346.287.110	99,87	85,93	13,9	18.517.275	2.337.132	20.792.272	2,02
753	372.228.475	99,64	94,32	17,1	17.510.609	2.357.229	19.807.612	2,15
752	340.249.508	99,79	95,49	17	17.544.050	2.541.975	20.020.242	2,1
721	356.956.710	99,81	72,94	12,3	20.617.141	2.780.002	23.318.554	2,09
714	335.406.862	99,71	92,29	16,2	17.590.435	2.541.521	20.066.179	2,15
713	380.210.484	99,78	96,15	18,8	17.373.810	2.529.352	19.836.825	2,12
707	305.000.681	99,66	94,08	14,8	17.277.102	2.489.040	19.704.102	2,16
702	486.209.902	99,78	97,17	25	17.846.146	2.625.855	20.395.866	2,08

*LM= Leituras mapeadas, *LPP= Leituras propriamente pareadas, *Cb= Cobertura, *SM=Sítios Multialélicos.

O arquivo VCF *raw* (Tabela 1), contém as variantes, que ainda não passaram por nenhum processo de filtragem, e apresentou total o de 38.597.271 SNVs. Os resultados após a aplicação dos parâmetros de filtragem sobre o arquivo VCF são apresentados na Tabela 2. O arquivo QC_B resultou no total de 10.620.122 SNVs a menos do que no arquivo raw quando foram aplicadas as filtragens. A diferença entre o QC_B e o QC_C foi de 8.850 SNVs a menos no QC_C, em que foram

filtrados todos os SNVs com cobertura menor do que 4x e assim consecutivamente, o QC_D apresentou 682.759 SNVs a menos que o QC_C. O QC_A resultou em 513.547 SNVs a menos que o QC_D e a diferença entre o QC_A e o QC_E foi de 1.475.315 SNVs a menos no QC_E. Se comparado o QC_B (que exige menor profundidade de leitura mínima) com o QC_E (que exige a maior profundidade de leitura mínima), a diferença é de 2.685.462 SNVs.

Tabela 2.Número de SNV e taxa de transição/transversão (Ti/Tv) em arquivo original (raw) e após filtragens

	Número de SNV	Ti/Tv	Diferença
RAW	38.597.271	2.10	
QC_B	27.977.149	2.26	-10.620.122
QC_C	27.968.299	2.26	-10.628.972
QC_D	27.280.549	2.27	-11.316.722
QC_A	26.767.002	2.28	-11.830.269
QC_E	25.291.687	2.30	-13.305.584

RAW= arquivo cru sem nenhum processo de filtragem, QC_B = apenas variantes com cobertura média maior que 10% da média, QC_C * 25% da média, QC_D * 50% da média, QC_A= apenas valores de cobertura mínima maior que 10, QC_E * 75 % da média.*

O número de SNVs encontrados neste estudo mesmo após a filtragem com critério de exclusão de variantes mais rígido (QC_E) foi superior se comparado com estudo de Liao et al. (2013) com três indivíduos e mais um “pool” gamético de 11 animais da raça Gir, em que foram descobertos 9.990.733 SNPs e 604.308 InDels, nossos resultados foram superiores, possivelmente por incluir um número maior de animais. Iqbal et al. (2019), em que sequenciaram 20 animais (número menor do que no presente estudo), com 11 das principais raças paquistanesas mais importantes de *Bos indicus* (gado zebuíno) foram encontrados 67.303.469 SNPs e

1.083.842 InDels, nossos resultados obtiveram um número inferior se comparado a estes, isso possivelmente se explica devido ao fato do mesmo ter incluído 11 espécies em sua análise, contra apenas uma deste presente trabalho. Nesse mesmo estudo, obteve-se uma média de cobertura entre as 20 amostras de 16X (variando entre 9X a 27X). Esses valores são muito próximos dos encontrados por nós, porém os autores utilizaram uma plataforma de sequenciamento diferente deste trabalho, o BGISEQ-500 system. Já o estudo de Liao et al. (2013) obteve cobertura de aproximadamente 3 a 4X por amostra, utilizando a plataforma SOLiD 4.

A diferença encontrada no número de SNV resultantes dos diferentes controles de qualidade indica que há variação na profundidade de leitura ao longo da montagem do genoma, e que um número considerável de variantes possui uma profundidade de leitura inferior a 10% da média de cobertura geral da montagem (10.620.122), contudo os controles de qualidade menos restritivos apresentaram número maior de variantes.

6. Conclusão

Portanto conclui-se que a filtragem de variantes baseada na profundidade de cobertura mínima resulta em diferentes conjuntos de variantes totais após a filtragem, quanto maior o valor de profundidade cobertura mínima requerido menor é o número total de SNVs resultantes.

7. Referências

BARRIS, W. et al. Next generation sequencing of African and Indicine cattle to identify single nucleotide polymorphisms. Animal Production Science, v. 52, n. 3, p. 133-142, 2012.

BOVINE GENOME SEQUENCING AND ANALYSIS CONSORTIUM et al. The genome sequence of taurine cattle: a window to ruminant biology and evolution. **Science**, v. 324, n. 5926, p. 522-528, 2009

BROOKES, Anthony J. The essence of SNPs. **Gene**, v. 234, n. 2, p. 177-186, 1999.

CANAVEZ, F. C. et al. Genome sequence and assembly of Bos indicus. **Journal of Heredity**, v. 103, n. 3, p. 342-348, 2012.

CHEN, N. et al. Whole-genome resequencing reveals world-wide ancestry and adaptive introgression events of domesticated cattle in East Asia. **Nature communications**, v. 9, n. 1, p. 1-13, 2018.

CHOI, J.W. et al. Massively parallel sequencing of Chikso (Korean brindle cattle) to discover genome-wide SNPs and InDels. **Molecules and cells**, v. 36, n. 3, p. 203-211, 2013.

CLINE, Melissa S.; KARCHIN, Rachel. Using bioinformatics to predict the functional impact of SNVs. **Bioinformatics**, v. 27, n. 4, p. 441-448, 2011.

DANECEK, Petr et al. The variant call format and VCFtools. **Bioinformatics**, v. 27, n. 15, p. 2156-2158, 2011.

ECK, S. H. et al. Whole genome sequencing of a single Bos taurus animal for single nucleotide polymorphism discovery. **Genome biology**, v. 10, n. 8, p. R82, 2009.

HAYES, B.; GODDARD, M. Genome-wide association and genomic selection in animal breeding. **Genome**, v. 53, n. 11, p. 876-883, 2010.
IQBAL, Naveed et al. Genomic variants identified from whole-genome resequencing of indicine cattle breeds from Pakistan. **PloS one**, v. 14, n. 4, p. e0215065, 2019.

KAWAHARA-MIKI, R. et al. Whole-genome resequencing shows numerous genes with nonsynonymous SNPs in the Japanese native cattle Kuchinoshima-Ushi. **BMC genomics**, v. 12, n. 1, p. 103, 2011.

KIM, Jaemin et al. The genome landscape of indigenous African cattle. **Genome biology**, v. 18, n. 1, p. 1-14, 2017.

LARMER, Steven G. **Next generation sequencing data in bovine: quality control, imputation, and application**. 2016. Tese de Doutorado. University of Guelph.

LI, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. **Bioinformatics**, v. 27, n. 21, p. 2987–2993, 1 nov. 2011. Disponível em: <https://academic.oup.com/bioinformatics/articlelookup/doi/10.1093/bioinformatics/btr509>. Acesso em: 9 fev. 2021.

LI, H. et al. The Sequence Alignment/Map format and SAMtools. **Bioinformatics**, v. 25, n. 16, p. 2078–2079, ago. 2009.

LI, H.; DURBIN, R. Fast and accurate short read alignment with Burrows-Wheeler transform. **Bioinformatics**, v. 25, n. 14, p. 1754–1760, 15 jul. 2009. Disponível em:<https://academic.oup.com/bioinformatics/articlelookup/doi/10.1093/bioinformatics/btp324>. Acesso em: 25 set. 2019.7

LIAO, X. et al. Whole genome sequencing of Gir cattle for identifying polymorphisms and loci under selection. **Genome**, v. 56, n. 10, p. 592-598, 2013.

LIU, A. et al. Improvement of genomic prediction by integrating additional single nucleotide polymorphisms selected from imputed whole genome sequencing data. **Heredity**, v. 124, n. 1, p. 37-49, 2020.

LIU, Y. et al. Bos taurus genome assembly. **BMC genomics**, v. 10, n. 1, p. 180, 2009.

NASCIMENTO, Guilherme Batista do. Estratégias de imputação e associação genômica com dados de sequenciamento para características de produção de leite na raça Gir. 2018.

NIELSEN, Rasmus et al. Genotype and SNP calling from next-generation sequencing data. Nature Reviews Genetics, v. 12, n. 6, p. 443-451, 2011.

O'BRIEN, Ana M. Perez et al. Low levels of taurine introgression in the current Brazilian Nelore and Gir indicine cattle populations. **Genetics Selection Evolution**, v. 47, n. 1, p. 1-7, 2015.

PANETTO JC do C et al. (2020) **Programa Nacional de Melhoramento do Gir Leiteiro Sumário Brasileiro de Touros 3a Avaliação Genômica de Touros**

PEVSNER, Jonathan. Bioinformatics and functional genomics. John Wiley & Sons, 2015.

Picard toolkit. **Broad Institute, GitHub repository** Broad Institute, 2019.

ROSEN, B.D. et al. De novo assembly of the cattle reference genome with single-molecule sequencing. **GigaScience**, v. 9, n. 3, giaa021, 2020.

ROY, Somak et al. Standards and guidelines for validating next-generation sequencing bioinformatics pipelines: a joint recommendation of the Association for Molecular Pathology and the College of American Pathologists. **The Journal of Molecular Diagnostics**, v. 20, n. 1, p. 4-27, 2018.

SANTANA JR, M. L. et al. History, structure, and genetic diversity of Brazilian Gir cattle. **Livestock Science**, v. 163, p. 26-33, 2014.

SOUZA SOBRINHO, F. de et al. Relatório técnico da Embrapa Gado de Leite 2001-2003. **Embrapa Gado de Leite-Documentos (INFOTECA-E)**, 2003.

STAFUZZA, N.B. et al. Single nucleotide variants and InDels identified from whole-genome re-sequencing of Guzerat, Gyr, Girolando and Holstein cattle breeds. **PLoS One**, v. 12, n. 3, 2017.

STOTHARD, P. et al. Whole genome resequencing of black Angus and Holstein cattle for SNP and CNV discovery. **BMC genomics**, v. 12, n. 1, p. 559, 2011.

TELLAM, Ross L. et al. Unlocking the bovine genome. **BMC genomics**, v. 10, n. 1, p. 1-4, 2009.

VERLI, Hugo. Bioinformática: da biologia à flexibilidade molecular. 2014.

VIGNAL, Alain et al. A review on SNP and other types of molecular markers and their use in animal genetics. **Genetics selection evolution**, v. 34, n. 3, p. 275-305, 2002.

ZHANG, Q. et al. Genome-wide association study for longevity with whole-genome sequencing in 3 cattle breeds. Journal of dairy science, v. 99, n. 9, p. 7289-7298, 2016.

ZHBANNIKOV, I. Y. et al. SeqyClean: a pipeline for high-throughput sequence data preprocessing. In: **Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics**. 2017. p. 407-416.

ZIMIN, A. V. et al. A whole-genome assembly of the domestic cow, Bos taurus. Genome biology, v. 10, n. 4, p. R42, 2009.