UNIVERSIDADE FEDERAL DE MINAS GERAIS
ESCOLA DE VETERINÁRIA
COLEGIADO DO PROGRAMA DE PÓS-GRADUAÇÃO EM ZOOTECNIA

SEQUENCIAMENTO DE GENOMA COMPLETO DO CAVALO NORDESTINO PARA IDENTIFICAÇÃO DE VARIANTES GENÔMICAS (SNVs E INDELs) E VALIDAÇÃO EM POPULAÇÕES REMANESCENTES

Danielle Cunha Cardoso

Belo Horizonte
2020

Danielle Cunha Cardoso

Sequenciamento de genoma completo do cavalo Nordestino para identificação de variantes genômicas (SNVs e INDELs) e validação em subpopulações remanescentes

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

Área de concentração: Genética e Melhoramento Animal.

Prof[a.] orientadora: Dra. Denise Aparecida Andrade de Oliveira.
Prof. coorientador: Dr. Idalmo Garcia Pereira.

Belo Horizonte
2020

Cardoso, Danielle Cunha, – 1983
C268s Sequenciamento de genoma completo do cavalo Nordestino para identificação de variantes genômicas (SNVs e INDELs) e validação em subpopulações remanescentes / Danielle Cunha Cardoso, -2020.

88f:il.

Orientadora: Denise Aparecida Andrade de Oliveira
Coorientador: Idalmo Garcia Pereira.
Tese (Doutorado) apresentado à Escola de Veterinária da Universidade Federal de Minas Gerais.
Área de Concentração: Genética e Melhoramento Animal.
Bibliografia f. 82 – 88.

1. Cavalo - Teses - 2. Genética – Estudo - Teses – 3. Zootecnia – Teses – I. Oliveira, Denise Aparecida Andrade de – II. Pereira, Idalmo Garcia – III. Universidade Federal de Minas Gerais, Escola de Veterinária – IV. Título.

CDD – 636.08
Bibliotecária responsável Cristiane Patrícia Gomes – CRB2569

Escola de Veterinária
UFMG

ESCOLA DE VETERINÁRIA DA UFMG
COLEGIADO DO PROGRAMA DE PÓS-GRADUAÇÃO EM ZOOTECNIA
Av. Antônio Carlos 6627 - CP 567 - CEP 30123-970 - Belo Horizonte- MG
TELEFONE (31)-3409-2173

www.vet.ufmg.br/academicos/pos-graduacao
E-mail: cpgzootec@vet.ufmg.br

ATA DE DEFESA DE TESE DE DANIELLLE CUNHA CARDOSO

Às 13:30 horas do dia 28 de fevereiro de 2020, reuniu-se, na Escola de Veterinária da UFMG a Comissão Examinadora de Tese, indicada pelo Colegiado em reunião no dia 11/11/2019, para julgar, em exame final, a defesa da tese intitulada: SEQUENCIAMENTO DE GENOMA COMPLETO DO CAVALO NORDESTINO PARA IDENTIFICAÇÃO DE VARIANTES GENÔMICAS (SNVS E INDELS) E VALIDAÇÃO EM POPULAÇÕES REMANESCENTES, como requisito final para a obtenção do Grau de **Doutor em Zootecnia área de concentração Genética e Melhoramento Animal.**

Abrindo a sessão, o Presidente da Comissão, Profa. Denise Aparecida Andrade de Oliveira, após dar a conhecer aos presentes o teor das Normas Regulamentares da Defesa de Tese, passou a palavra ao (a) candidato (a), para apresentação de seu trabalho. Seguiu-se a argüição pelos examinadores, com a respectiva defesa do candidato (a). Logo após, a Comissão se reuniu, sem a presença do candidato e do público, para julgamento da tese, tendo sido atribuídas as seguintes indicações:

	Aprovada	Reprovada
Prof.(a)/Dr.(a) Ângela maria Quintão Lana	X	
Prof.(a)/Dr.(a) Denise Aparecida Andrade de Oliveira	X	
Prof.(a)/Dr.(a) Eduardo Paulino da Costa	X	
Prof.(a)/Dr.(a) Adriana Mexia Teixeira	X	
Prof.(a)/Dr.(a) Lívia Paloda dos Santos Peres	X	

Pelas indicações, o (a) candidato (a) foi considerado (a): [X] Aprovado (a) [ ] Reprovado (a)

Para concluir o Doutorado, o(a) candidato(a) deverá entregar 10 volumes encadernados da versão final da tese acatando, se houver, as modificações sugeridas pela banca, e a comprovação de submissão de pelo menos um artigo científico em periódico recomendado pelo Colegiado dos Cursos. Para tanto terá o prazo máximo de 60 dias a contar da data defesa.

O resultado final, foi comunicado publicamente ao (a) candidato (a) pelo Presidente da Comissão. Nada mais havendo a tratar, o Presidente encerrou a reunião e lavrou a presente ata, que será assinada por todos os membros participantes da Comissão Examinadora e encaminhada juntamente com um exemplar da tese apresentada para defesa.

Belo Horizonte, 28 de fevereiro de 2020.

Assinatura dos membros da banca:

Profa. Ângela Maria Quintão Lana
Coordenadora do Colegiado de
Pós-Graduação em Zootecnia

(Vide Normas Regulamentares da defesa de Tese no verso)

(Este documento não terá validade sem assinatura e carimbo do Coordenador)

“Não importa quanto a vida possa ser ruim, sempre existe algo que você pode fazer, e triunfar. Enquanto há vida, há esperança”.

Stephen Hawking.

Dedico essa tese aos meus pais, David Cardoso e Efigênia, e ao meu marido Rodrigo, os quais torceram por mim com todo amor e estiveram presentes em cada etapa da construção desse caminho.

AGRADECIMENTOS

À Deus, pelo amor, dom da vida, as melhores oportunidades, saúde e proteção sem medidas.

Aos meus amados pais, David e Efigênia pelo amor incondicional, orações e torcida. Essa conquista pertence à vocês.

Ao meu marido Rodrigo, pelo amor, suporte, exemplo de persistência, garra, companheirismo, amizade e especiamente por alegrar meus dias com seu sorriso diário.

À minha amada filha Laura, que de dentro do meu ventre, me deu forças na reta final desse trabalho e me fez companhia nos experimentos finais, análise de dados e escrita da tese.

Aos meus irmãos Fabrício e Leo, por me amarem e apoiarem sempre.

Aos amigos e familiares pela torcida, especialmente à tia Cidinha e à Vera, pelo carinho e preocupação comigo.

Aos professores que contribuíram com a construção da minha formação.

À professoa Denise, pela oportunidade de desenvolver o doutorado, de participar de cursos de aperfeiçoamento e pela confiança depositada em mim, não apenas nesse projeto, mas tamém nos outros projetos de pesquisa, coordenados por ela, para os quais tive oportunidade de contribuir.

À CAPES pela bolsa concedida.

Ao INCT PECUÁRIA (projetos CNPq 573899/2008-8 e FAPEMIG APQ-0084/08) pela disponibilização de parte dos recursos utilizados.

Ao Eduardo por me apresentar esse projeto para que eu o desenvolvesse, pela amizade, confiança, apoio e gentileza sempre.

À TODOS os colegas do LGEV, pela agradável convivência e ajuda durante todos esses anos e às amizades fortes que se estabeleceram e ficarão pra sempre! Por nossos momentos de descontração, saídas, almoços, viagens, compras! Muito obrigada pela convivencia maravilhosa!

Às amigas de doutorado (Veterinária e ICB), Mayra, Virgínia, Nathália, Joelma e Gabi, pela troca de experiências, auxílios e momentos de descontração.

Ao professor Evanguedes (ICB) pelos ensinamentos e por disponibilizar os equipamentos do laboratório que coordena, para que eu fizesse as análises finais.

À equipe de bioinformática da Embrapa Recursos Genéticos e Biotecnologia, por me receber, pelos ensinamentos em bioinformática e suporte oferecido na análise dos resultados desse trabalho.

À Brenda (Embrapa) e Carol (ICB) pela importante ajuda na geração e interpretação dos dados do capítulo 1.

À professora Dênea pelo envio das amostras e apoio.

À professora e amiga Glacy e à Universidade Paranaense, pela ajuda financeira para a aquisição de parte dos insumos.

À todos os integrantes da banca de defesa, pela importantíssima contribuição.

RESUMO

A chamada de variantes genômicas, a fim de identificação de novas variantes de nucleotídeo único (SNVs) e inserções/deleções (InDELs), potencialmente associadas às características de rusticidade e resistência a ambientes áridos em equinos da raça Nordestina, a partir de dados de sequenciamento de genoma completo foi a estratégia utilizada no presente estudo para compreensão dos mecanismos adaptativos da raça para sobrevivência no semiárido do nordeste brasileiro. Essas variantes podem estar associadas à resistência exibida pelo cavalo Nordestino, por meio do padrão racial, que é altamente especializado para o desempenho de atividades que exigem força e resistência, sem quaisquer prejuízos à sua perpetuação. O genoma de um representante típico da raça foi sequenciado, sendo este o primeiro trabalho de sequenciamento de genoma completo de uma raça equina naturalizada brasileira, que utilizou a estratégia de genômica comparativa para chamada de variantes, usando-se como referência a mais recente atualização do genoma equino disponível (Ensemble, EquCab 3.0). A chamada de variantes genômicas por ferramentas baseadas em algoritmos identificou 1.598.210 SNVs e 138.139 InDels pelo software *Freebayes* e 88.838 SNVs e 25.232 InDELs pelo software *GATK*. As variantes foram classificadas quanto ao tipo, localização e efeito de impacto e foi feita a anotação genômica das regiões onde estão localizadas. Aquelas de impacto alto e moderado foram selecionadas para análise de enriquecimento funcional e para validação populacional em 60 animais provenientes dos estados da Bahia, Pernambuco e Piauí. SNVs de impacto alto ocorreram exclusivamente em genes de GTPases. Genes da família de Receptores Olfativos (OR), que continham SNVs de impacto moderado tiveram maior representatividade na análise funcional. Todos os dez *loci* testados, contendo SNVs características da raça apresentaram alelos polimórficos nas três populações, com frequências genotípicas elevadas de indivíduos heterozigotos e homozigotos contendo polimorfismo alélico. As frequências genotípicas de oito *loci* analisados foram significativamente diferentes entre as três populações, sugerindo a fixação populacional dos alelos polimórficos e diversidade genética. A presença de SNVs em OR genes, que estão diretamente relacionados aos mecanismos adaptativos, representa importantes informações genéticas para compreender a adaptação da raça ao semiárido do nordeste brasileiro e para associá-las ao fenótipo peculiar exibido por essa raça.

Palavras-chave: Cavalo Nordestino, SNVs, validação populacional, genes de GTPases, genes de Receptores Olfativos.

ABSTRACT

Genomic variant calling, in order to identify new single nucleotide variants (SNVs) and insertions/deletions (InDELs), potentially associated with the characteristics of rusticity and resistance to arid climate in Nordestino horse breed, based in whole genome sequencing was the strategy used in the present study to understand the adaptive mechanisms of the breed to semiarid climate of Brazilian northeastern. These variations are possibly associated with the resistance exhibited by the breed, through the highly specialized racial pattern for the performance of activities that require strength and endurance, without any damage to its perpetuation. Complete genome sequencing of a typical representative of the breed was carried out, this being the first whole genome sequencing of a Brazilian naturalized equine breed, as a comparative genomics strategy for variants calling with the most recent reference genome equine update (Ensemble EquCab 3.0). Tools based on algorithms for variants calling identified 1,598,210 SNVs and 138,139 InDels (Freebayes software) and 88,838 SNVs and 25,232 InDELs (GATK software). Variants were classified as to the type, location and impact effect and the genomic regions were anoted. High and moderate impact variants were selected to functional enrichment analysis and for population validation in 60 animals from the states of Bahia, Pernambuco and Piaui. High-impact SNVs occurred exclusively in GTPase genes. Olfactory Receptor (OR) genes family containing moderate-impact SNVs had high representativeness in the functional analysis. All 10 *loci* tested, containing breed-specific SNVs, presented polymorphic alleles in the three populations, with high genotypic frequencies of heterozygous and homozygous with allelic polymorphism. The genotypic frequencies for 8 *loci* were significantly different between the three populations, suggesting allelic fixation and genetic diversity. The presence of SNVs in OR genes, which are related to adaptive mechanisms, represents important genetic information to understand the adaptation of the breed to the semiarid region of northeastern Brazil and to associate them with the peculiar phenotype exhibited by that breed.

Keywords: Nordestino horse breed, SNVs, populational validation, GTPase gene, Olfactory Receptor gene.

LISTA DE FIGURAS E TABELAS

REVISÃO BIBLIOGRÁFICA

CAPÍTULO 1

CAPÍTULO 2

LISTA DE ABREVIATURAS

Sigla: Nome

ul: Microlitros

A: Adenina

ATP: Adenosina trifosfato

C: Citosina

ºC: grau Celcius

cm: Centímetro

DAVID: "Data Bank for Anotation, Visualization and Integrated Discovery"

DL: Desequilíbrio de Ligação

DNA: Ácido Desoxirribonucleico

dNTP: Desoxirribonucleotídeos fosfatados

EDTA: ácido etilenodiamino tetra-acético

FDR: "False Discovery Rate"

FREEBAYES: "Haplotype-based variant detection from short-read sequencing" (Software para chamada de variantes genômicas)

G: Guanina

Gb: Giga Bases

GATK: "Genome Analysis Toolkit" (Software para chamada de variantes genômicas)

GDP: Guanosina difosfato

GO: Gene Ontology (Ontologia de genes)

GTP: Guanosina trifosfato

GWAS: Genome Wide Association Study (Estudos de associação genômica ampla)

ID: Identidade gênica

InDel: Inserção/Deleção

Kb: Quilobases

KEEG: "Kyoto Encyclopedia of Genes and Genomes" (Enciclopédia de genes e genomas de Kyoto)

M: Molar

N: Normal

ng: nanogramas

NGS: "New/Next Generation Sequencing" (sequenciamento de nova geração)

PCR: Polymerase Chain Reaction (Reação em Cadeia da Polimerase)

pb: pares de bases

qPCR: Quantitative Polymerase Chain Reaction (Reação em Cadeia da Polimerase quantitativa)

pM: picomolar

SNV: Single Nucleotide Variants (Variantes de nucleotídeo único)

SNP: Single Nucleotide Polymorphisms (Polimorfismos de nucleotídeo único)

T: Timina

UTR: Untranslated region (região não traduzida)

SUMÁRIO

I INTRODUÇÃO

Dentre as raças nacionais de cavalo, a raça Nordestina destaca-se por apresentar patrimônio genético de valor inestimável, o qual precisa ser conservado em função das suas características peculiares que permitem a sobrevivência da mesma em condições inóspitas do semiárido nordestino, como insolação excessiva, solos pedrosos, baixa disponibilidade hídrica e escassa oferta de forragem. Tais características permitem a sobrevivência do cavalo Nordestino sem quaisquer prejuízos ao seu desempenho ou perpetuação.

Outras raças equinas não conseguiriam desenvolver, nesse ambiente, as atividades que exigem força e resistência. Entre as características do padrão racial, destacam-se os cascos rígidos, resistentes e pequenos, que dispensam uso de ferraduras, bem adaptados à dureza e sinuosidade do terreno, o porte médio, ossatura e musculaturas fortes e narinas dilatadas (Costa *et al*., 2001).

Em função disso, no nordeste brasileiro a raça é amplamente utilizada em atividades agropecuárias, como na lida com o gado, transporte de produção agrícola, de mercadorias, pessoas e em atividades relacionadas à equinocultura, destacando-se pela grande importância econômica e sociocultural. Apesar dessa significativa importância da raça para a região, os cruzamentos desordenados com a raça Quarto de Milha e Mangalarga Marchador e a desativação da Associação Brasileira de Criadores do Cavalo Nordestino tem contribuído para o quase desaparecimento da raça, exigindo ações urgentes para sua conservação (Costa *et al*., 2001). A realização de programas para conservação desse genoma é imprescindível, em virtude dos atuais sistemas de produção e das transformações climáticas do planeta (Pires *et al*., 2008).

Há poucos estudos sobre o cavalo Nordestino. Alguns são relacionados à caracterização fenotípica (Travassos 2004; Melo *et al*., 2006; Pires *et al*., 2008; Melo *et al*., 2013; Pires, 2012) e outros visando estimar quais raças nacionais e exóticas influenciaram na sua formação (Pires *et al*., 2012; Pires *et al*., 2014); porém, não há nenhum estudo relacionado ao uso de ferramentas de genética molecular para compreensão das bases genéticas relacionadas às características fenotípicas de resistência e rusticidade, que permitiram a perpetuação da raça no semiárido nordestino.

Nesse contexto, informações genômicas, especialmente em larga escala, por meio das tecnologias de sequenciamento de nova geração são essenciais para explicar a excelente adaptação da raça ao ambiente, bem como as bases moleculares do fenótipo exibido pela mesma. Nesse sentido, auxiliam na conservação e fornecem informações inéditas aos programas de melhoramento genético já existentes para raças equinas. O avanço dos estudos

moleculares possibilita aos programas de melhoramento genético a realização da seleção precoce de animais geneticamente desejáveis, por meio da análise de marcadores genéticos associados às características de interesse.

Desse modo, foi realizado o sequenciamento do genoma completo de um exemplar macho e típico da raça e foram identificadas todas as variantes de nucleotídeos únicos (SNVs) e inserções/deleções (InDels), em comparação com o genoma equino referência de montagem mais atualizada e disponível publicamente (Emsembl 3.0), da raça Puro Sangue Inglês.

Possíveis associações da presença dessas variantes à rusticidade e à resistência exibida pela raça equina Nordestina foram estabelecidas, por meio da análise de enriquecimento funcional dos genes que continham variações de impacto relevante (variantes de nucleotídeo único _SNVs_ que podem afetar sítios de *splicing*, códons de início e fim da transcrição, e variantes não sinônimas).

Em seguida, em continuidade a esse trabalho, como segundo capítulo dessa tese de doutorado, dez SNVs em quatro *loci* gênicos foram validados em três subpopulações remanescentes, em um total de 60 animais. As variantes foram escolhidas pelo nível de impacto (impacto alto e moderado) sobre a expressão dos genes nos quais estão presentes (família de GTPases e Receptores Olfativos) e com base na análise de enriquecimento funcional nos termos do Gene Onthology (GO) e de vias metabólicas da base de dados KEEG (Kyoto Encyclopedia of Genes and Genomes).

A presença de alelos polimórficos para todas as variantes testadas foi observada nas três populações de remanescentes da raça, provenientes dos estados da Bahia, Pernambuco e Piauí. As frequências alélicas e genotípicas indicam possível fixação dos alelos polimórficos nas populações e diversidade genética entre elas, uma vez que as frequências alélicas de oito dos dez *loci* validados diferiram significativamente entre as populações. Além disso, as maiores frequências de indivíduos homozigotos para os alelos contendo as SNVs ocorreu no estado da Bahia, o que sugere um menor controle sobre os cruzamentos entre indivíduos aparentados.

Esse estudo apresenta dados genéticos consistentes para que as SNVs identificadas e validadas, uma vez estudadas em outras populações da raça, sejam caracterizadas como possíveis polimorfismos de nucleotídeos únicos (SNPs), presentes em genes e vias metabólicas possivelmente relacionadas ao fenótipo exclusivo exibido pela raça. Os polimorfismos de impacto alto ou moderado, identificados em genes de GTPases e da ampla família de OR genes_ genes envolvidos na percepção sensorial do olfato e, consequentemente no reconhecimento do meio, de feromônios, da prole, do alimento e ameaças _ podem ter forte contribuição com a excelente adaptação da raça equina Nordestina ao semiárido do nordeste brasileiro.

A informação genômica obtida será disponibilizada em bancos de dados genômicos públicos tendo grande utilidade aos estudos de melhoramento genético animal, especialmente equinos. Esses dados podem ser ainda utilizados para pesquisas envolvendo estudos de associação genômica (GWAS) dentro da raça ou em outras raças equinas, além de outras aplicações, incluindo disponibilização de informações genômicas aos programas de conservação de recursos genéticos da raça Nordestina e outras.

II REVISÃO DE LITERATURA

1) Origem e caracterização da raça equina Nordestina

A raça Nordestina originou-se de raças europeias, especialmente ibéricas, trazidas pelos colonizadores. Nas primeiras décadas do século XVI, a capitania de Pernambuco (atual área dos estados de Pernambuco, Alagoas e Sergipe) destacava-se em prosperidade econômica, desenvolvimento impulsionado pelo trabalho animal equino e bovino. É nesse contexto que os cavalos foram difundidos pelo nordeste do Brasil (Braga, 2000), em especial na Bahia, por sediar a primeira capital do país. A região recebeu centenas de cavalos, vindos principalmente de Cabo Verde, a maioria da raça Barbo-Árabe. Costa e colaboradores (1974) destacam a descendência principal da raça Nordestina a partir do cavalo Barbo-Árabe, devido a semelhanças morfológicas como: perfil convexo, orelhas mal implantadas, garupa caída e cauda de inserção baixa.

Ao longo dos últimos cinco séculos, as raças equinas foram submetidas à seleção natural e artificial (praticada pelo homem) em determinados ambientes, de modo que as características específicas que permitiam a sobrevivência a tais condições foram conservadas e perpetuadas. Estas raças, aqui desenvolvidas, passaram a ser conhecidas como “crioulas”, “locais” ou “naturalizadas” e apresentam hoje características de adaptação de extrema importância para trabalhos de melhoramento animal, devido à sua maior rusticidade e longevidade, sucesso reprodutivo e à maior resistência à doenças e estresses ambientais (Mariante *et al*., 2011).

Em função da importância da raça para o agreste nordestino e dos traços genéticos adaptativos peculiares, somados à atual situação de redução das populações, alguns estudos relacionados à caracterização fenotípica vêm sendo realizados. Tais estudos revelaram predominância das pelagens castanha e alazã, perfil cefálico retilíneo, chanfro reto e garupa horizontal. São animais de média conformação corporal (não ultrapassando 145 cm para machos e 140 para fêmeas), peso corporal médio de 280 kg, crinas e caudas escassas, membros finos e delgados, inserção de cauda média, narinas dilatadas, corpo ligeiramente arqueado, lábios finos e orelhas médias (Travassos, 2004; Melo, 2011).

Além disso, Melo *et al*. (2006) e Melo (2011) destacam os cascos escuros, com ranilhas elásticas e profundas, conferindo-os a capacidade de caminhar demasiadamente por solo pedregoso e escarpado do Semiárido Nordestino, sem indícios de enfermidades nos cascos.

Esses autores também descrevem que esses animais apresentam rusticidade, resistência e vivacidade, mesmo após caminhadas longas, sob intensa insolação.

Outra característica particular da raça é a resistência à escassez de água e alimentos por dias, devido à incrível adaptação às condições de baixa e irregular disponibilidade de forragem e às secas periódicas. Assim, suportam a desidratação e elevação da temperatura corporal bem mais do que qualquer outra raça equina melhorada, obtendo-se ótimo desempenho em trabalhos agrícolas, especialmente na lida com o gado (Pires *et al.*, 2012). Segundo Melo (2011), a resistência ao calor é explicada em parte, pelo fato de a relação massa/superfície corpórea favorecer que o mesmo seja mais facilmente dissipado.

Infelizmente, o cavalo Nordestino encontra-se em cenário de ameaça bastante grave, devido principalmente à desativação dos núcleos de preservação e seleção da raça e da Associação Brasileira dos Criadores do Cavalo Nordestino (ABCCN), há vinte anos, reduzindo de forma expressiva o número de criadores e registros de animais. Os remanescentes da raça dizem respeito à populações distribuídas no semiárido Nordestino, principalmente em determinadas regiões dos estados da Bahia, Ceará e Piauí (Pires *et al.*, 2012).

Segundo a lista mundial de observação para a diversidade dos animais domésticos, publicada pela FAO/DAD-IS (FAO – DOMESTIC ANIMAL DIVERSITY INFORMATION SYSTEM, 2000), o grau de risco para raça equina Nordestina é considerado desconhecido, devido à ausência de dados atualizados da raça (Melo, 2011).

Ainda não há quaisquer estudos na raça acerca das variações estruturais genéticas, incluindo SNVs e inserções/deleções (InDels) e associação destes à resistência exibida pela raça às condições áridas, nem estudos funcionais relacionados à peculiaridade genética exibida pela mesma.

De acordo com o exposto, fica clara a importância do cavalo Nordestino como patrimônio genético e biológico nacional, bem como para a economia e cultura do Nordeste brasileiro. É incontestável a necessidade de estudo dessa raça, a fim de preservá-la para ampliar e diversificar as linhas de pesquisa sobre a mesma.

A compreensão da base genética envolvida na adaptação ao semiárido e caracterização de regiões genômicas polimórficas associadas terão papéis de extrema importância para programas de melhoramento animal, que visem, por exemplo, resistência à altas temperaturas, à escassez de água e forragem (decorrentes das mudanças climáticas), a determinadas doenças, força, entre outras.

2) Conservação de raças naturalizadas

Segundo Egito *et al.* (2002) as cinco razões principais que justificam esforços para manutenção da diversidade genética das raças de animais são: razões biológicas, econômicas, científicas, culturais e históricas.

Raças naturalizadas, como a raça Nordestina encontram-se ameaçadas de extinção, principalmente devido a cruzamentos indiscriminados com animais de raças exóticas, estratégias ineficientes de gestão desses recursos e programas de controle de doenças mal elaborados (Egito *et al*., 2002).

Com o número reduzido de animais em uma população ocorre aumento de consanguinidade, efeitos de deriva e cruzamentos indiscriminados resultando em perda de diversidade. Para evitar o desaparecimento destes importantes e insubstituíveis materiais genéticos, os programas de conservação de raças naturalizadas vêm crescendo em todo o mundo. No Brasil, a Embrapa decidiu incluir raças localmente adaptadas, inclusive equinas, no seu Programa de Pesquisa em Recursos Genéticos, em parceria com empresas estaduais de pesquisa e universidades (Egito *et al*., 2002).

A preservação vem sendo realizada também por meio de "núcleos de conservação", que ajudam a implementar bancos de germoplasma, sob coordenação do Centro Nacional de Pesquisa de Recursos Genéticos e Biotecnologia (CENARGEN). Os bancos de germoplasma podem ser compostos de rebanhos animais de uma raça, que ficam submetidos à seleção natural (*in situ*), ou de material genético congelado, como sêmen, embriões e ovócitos (*ex-situ*). Diversas raças localmente adaptadas estão presentes nestes bancos.

No caso específico da raça Nordestina, é necessário manter a máxima variabilidade existente na população remanescente, por meio de alternativas como troca de reprodutores entre as propriedades, inclusão da raça em núcleos de conservação animal, coleta de germoplasma e criopreservação (Almeida, 2009). Esforços nesse sentido estão sendo feitos por grupos de pesquisa que estudam a raça.

Segundo Egito *et al*., (2002) para a preservação de raças em risco de extinção e caracterização das mesmas é fundamental o conhecimento da estrutura genética e da dinâmica populacional. Assim, o estudo genético utilizando-se marcadores microssatélites possibilita definir a diversidade genética entre animais e raças, auxiliando a implementação dos programas de conservação, uma vez que a utilização de técnicas para identificação genética de indivíduos e análise de parentescos permite direcionar acasalamentos para manutenção da diversidade genética.

Para a raça equina Nordestina, esse primeiro passo foi realizado por Pires *et al*. (2014), que avaliaram a estrutura e diversidade de quatro subpopulações, oriundas das principais regiões de ocorrência. Os dados revelaram elevada diversidade genética, não evidenciando níveis relevantes de consanguinidade, o que indicou que há cruzamentos entre as subpopulações.

3) Variantes genômicas: InDELs e SNVs

O termo variantes genômicas é usado para definir mutações herdáveis no material genético. Inserções e deleções, referidas como InDELs, e substituições são mutações pontuais, uma vez que alteram poucas bases. Alterações de uma única base são conhecidas como variantes de nucleotídeo único (SNVs, *Single Nucleotide Variants*) e são resultantes de erros de incorporação de base durante o processo de replicação do DNA (Griffin e Smith, 2000).

Alterações que não resultam em modificações nos aminoácidos são chamadas silenciosas, devido à redundância do código genético, ou por estarem presentes fora de regiões de transcrição. Quando SNVs ocorrem em exons e geram variação em um aminoácido da proteína, estas são classificadas como não sinônimas, enquanto que SNVs sinônimos, embora não causem variação na sequencia de aminoácidos, podem gerar desestabilização da molécula de RNA mensageiro. Substituições que geram códons de terminação e interrompem a sequência da proteína, impossibilitando sua atividade, apresentam o maior grau possível de efeito de impacto (Modrek e Lee, 2002).

O efeito de impacto também é considerado alto quando variantes afetam sítios de *splicing* ou quando ocorre uma inserção ou deleção (InDEL) dentro da região codificadora de um gene, alterando o quadro de leitura (*frameshift*), acarretando mudança dos aminoácidos da proteína. Quando o InDEL não é uma sequência múltipla de três, se apenas um nucleotídeo for excluído da sequência ou inserido, todos os códons antes e após a mutação terão um quadro de leitura interrompido, podendo resultar na incorporação de aminoácidos incorretos. Entretanto, se três nucleotídeos forem inseridos ou excluídos, não haverá mudança na estrutura de leitura do códon; mas haverá um aminoácido extra ou um aminoácido ausente na proteína final, gerando de toda forma, efeito de impacto alto sobre a expressão do gene onde se localiza a variante (Chan *et al*, 2009).

Variantes de nucleotídeo único, para serem consideradas SNPs (Single Nucleotide Polymorphisms), assume-se que para um determinado *locus* que contém a variante, o alelo menos frequente na população deve ter uma abundância acima de 1%. Em geral são bialélicos,

podendo ocorrer em exons, introns, regiões transcritas e não traduzidas e regiões promotoras, sendo elementos associados à regulação da expressão gênica e ao metabolismo basal do DNA (Rafalski, 2002).

Wray (2007) evidencia que SNPs presentes em regiões de regulação em *Cis* tem um impacto relevante sobre o fenótipo, pois podem alterar a ligação de fatores de transcrição a sítios específicos, alterando a transcrição e efetivamente, a disponibilidade de determinada proteína. É comum a segregação de dois ou mais SNPs, chamados SNPs haplótipos, que caracterizam um desequilíbrio de ligação (DL), o qual pode ser compreendido como a falta de segregação independente entre os alelos em dois ou mais *loci*.

Desse modo, a análise de SNPs permite inferir análise de diversidade populacional, pois o DL pode ser aumentado ou diminuído, em função de endocruzamentos, tamanho populacional, isolamento genético, pressões seletivas, altas taxas de recombinação, entre outros fenômenos (Remington, 2001).

4) Estudo em larga escala de genomas: Identificação de regiões genômicas de interesse zootécnico

A genômica animal é uma realidade presente nos programas de melhoramento e conservação de recursos genéticos e os impactos das aplicações desses métodos podem ser notados em várias áreas da agropecuária.

A genômica gera dados em larga escala, principalmente por meio das novas metodologias de sequenciamento de DNA, promovendo identificação de regiões relacionadas às características de interesse, seja por análise comparativas de genomas para busca de polimorfismos ou sequenciamento de transcriptoma, identificando genes candidatos, diferencialmente expressos sob condições distintas. O uso dessas e outras técnicas de biologia molecular para identificar as bases moleculares do fenótipo, tem permitido avanços imensuráveis nos programas de melhoramento genético para o desenvolvimento de animais que carreguem características de interesse diversos.

Esses estudos genômicos são feitos a partir da geração e análise minuciosa de um grande volume de dados e permitem a identificação pontual da marca genética associada no DNA. Desse modo, essas variações de bases nitrogenadas, incluindo-se diversos tipos de polimorfismos ou mutações, podem ser identificadas precocemente, ainda nas primeiras fases do desenvolvimento embrionário. Isso promove, por exemplo, um ganho de tempo nos

programas de melhoramento, pois não é necessário aguardar a manifestação do fenótipo para dar continuidade aos trabalhos de melhoramento genético.

É importante considerar que as bases genéticas dessas características de interesse são complexas e chamadas quantitativas, pois são controladas por vários genes e sofrem influência do ambiente. Womack (2005) destaca que a associação do componente genético à característica de interesse, visando o melhoramento genético, há alguns anos vinha sendo feita, por mapeamento de loci de características quantitativas (*Quantitative Trait Loci*, QTLs), a qual baseia-se na definição de regiões cromossômicas associadas à variação genética e ao fenótipo em estudo.

Hoje, as técnicas de análises genômicas em larga escala, como sequenciamento de DNA de próxima geração (*Next Generation Sequencing*, NGS), vem sendo utilizadas em conjunto com metodologias clássicas e análises em *softwares* e técnicas computacionais de bioinformática. Assim, é possível explorar características genômicas complexas, como identificação da localização e mapeamento de polimorfismos e variantes genômicas, de modo que, aquelas que ocorrem em regiões gênicas transcritas ou envolvidas na transcrição podem ser caracterizadas quanto aos efeitos sobre o produto desses genes.

Não faltam exemplos de trabalhos que realizaram a associação de ambas as técnicas (QTLS e NGS) em zootecnia, a fim de gerar dados de sequenciamento genômico de regiões contendo QTLS previamente mapeados. Os resultados desses trabalhos de associação revelaram ação de genes relacionados e o efeito dos mesmos sobre características de interesse, como por exemplo, variação da porcentagem de gordura/proteína e do peso corporal em bovinos, qualidade de ovos em aves, resistência a parasitas em ovinos, entre outros (Dalrymple *et al*., 2007; Groenen *et al*., 2011; Khatkar *et al*., 2004).

Em equinos, aplicações da genômica são mais frequentes em estudos de pigmentação genética da pelagem, doenças hereditárias e desempenho atlético, em função, principalmente, do grande interesse mundial pelo mercado esportivo. Porém, com o genoma completo disponível, montado e reanalisado (Kalbfleisch *et al*., 2018) é possível explorar profundamente genes envolvidos em diversos outros aspectos, como na adaptação a determinado ambiente, resistência a condições adversas, características de rusticidade, doenças simples e complexas, herdáveis ou não, regulação genética, entre outros, obtendo-se eficiência seletiva e bom custo-benefício, em comparação com a tradicional seleção baseada em dados fenotípicos, além de ser excelente para características de baixa herdabilidade (Resende *et al*., 2008).

As análises de NGS e, consecutivamente, mapeamento e montagem de genomas, anotação de genes e análises de enriquecimento funcional, por permitirem obter a sequência,

posição dos genes e variantes no genoma, também aceleram o desenvolvimento de várias estratégias para estudos genômicos.

Entre essas, destacam-se as plataformas de microarranjos para detecção da presença de SNPs, os chamados *SNPs Arrays*, que podem ser aplicados para genotipagem, identificação de SNPs específicos e estudos de associação genômica (*GWAS*), além de análises de expressão gênica em larga escala (*RNAseq*) e/ou expressão diferencial (modulação do transcriptoma em resposta à condições distintas) (Anderson e Georges, 2004).

O uso comercial de marcadores moleculares na produção animal é aplicado na seleção precoce de animais, por exemplo, com melhor qualidade e maciez de carne (suínos e bovinos), maior produção de leite (bovinos) e maior eficiência reprodutiva (ovinos) (Dekkers, 2004).

Hoje, esses marcadores são mais amplamente identificados e caracterizados pelo uso de NGS. O grupo de pesquisa do Laboratório de Genética da Escola de Veterinária da UFMG, por exemplo, está concluindo um trabalho para identificação de polimorfismos em genes associados à precocidade sexual em bovinos da raça Guzerá. O sequenciamento em larga escala de genes alvo foi feito via plataforma Illumina ™, a partir de um conjunto de 450 *read*s representativos de genes candidatos, em 50 animais com precocidade sexual, contrastados com outros 50 animais com inicialização tardia do desenvolvimento sexual. Uma vez identificados e caracterizados, esses polimorfismos podem ser utilizados como marcadores genéticos para seleção dos animais com o fenótipo de interesse.

Alfonso (2005) ressalta que, uma vez identificado o polimorfismo associado à característica de interesse, faz-se necessária a condução de estudos funcionais para estabelecer uma relação de causa e efeito entre o mesmo e a característica fenotípica, pois a variação da sequência pode ou não ocorrer nos genes relacionados à característica fenotípica em estudo.

Em genética animal, a identificação de variantes, especialmente em genes, a partir de dados de sequenciamento completo de genomas e/ou transcriptomas, tem se revelado uma estratégia valiosa para amenizar limitações das técnicas anteriores, permitindo compreender como a variação genética influencia a característica de interesse, uma vez que esta é mapeada de modo preciso no genoma. Além disso, é possível mensurar o efeito das variantes, relacionando-o com o fenótipo em estudo.

Alterações fenotípicas associadas às mutações acompanham a domesticação como efeito do impacto entre a reprodução seletiva, controlada por ações humanas e pela atuação da seleção natural (Andersson *et al*., 2012). Como resultado, a maioria das raças equinas apresentam populações com alta uniformidade fenotípica e genotípica dos indivíduos dentro da raça, mas entre raças há muita variação.

Atualmente, a ciência utiliza apenas pequena parte do extraordinário potencial total das aplicações da genômica para a seleção e melhoramento animal e muitos avanços estão ainda previstos para um futuro próximo.

5) Genoma equino

O projeto para sequenciar o genoma equino (*Equus caballus*) teve início em meados da década de noventa, em Lexington, Kentucky, EUA, onde foi lançada a ideia para desenvolver o *Horse Genome Project*. Em 1997, a comunidade internacional participante do projeto reuniu-se como parte de uma pesquisa nacional patrocinada pelo *United States Departament of Agriculture* (USDA) e em 2005 foi apresentado requerimento junto ao *National Human Genome Research Institute* (NHGRI), mostrando as contribuições da pesquisa do genoma dos equídeos na compreensão do genoma humano. O sequenciamento e a organização do genoma equino foram realizados pelo *Broad Institute* do Instituto de Tecnologia de Massachussets (MIT) e pela Universidade de Harvard.

Foi produzido um rascunho de alta qualidade, com cobertura de 6,8 vezes, a partir do DNA genômico de uma égua Puro-Sangue Inglês, chamada Twilight, de propriedade da Universidade de Cornell em Nova Iorque, EUA. Aproximadamente trezentos mil BACs (cromossomo artificial de levedura, usados como vetores de transformação que comportam insertos grandes) com extremidades sequenciadas foram gerados na Alemanha pela Universidade de Medicina Veterinária em Hanover, e pelo centro Helmholtz para a Pesquisa de Infecções em Braunschweig.

Entre 2006 e 2007 as BACs foram sequenciadas e ordenadas gerando um mapa inicial, contendo um rascunho de alta qualidade, de tamanho aproximado de 2,7 Giga bases (Gb) a um custo aproximado de quinze bilhões de dólares, mostrando que o genoma equino é menor que o humano e pouco maior que o do cão doméstico. Sucessivos aprimoramentos da montagem resultaram na publicação completa e disponibilização no banco genômico Ensembl em 2009. (Wade *et al*., 2009).

A partir desse mapa inicial, o projeto de sequenciamento do genoma equino do *Broad Institute* vem continuando os projetos de montagem em diversas parcerias, disponibilizando dados dos genomas das raças Árabe, Andaluz, Akhal-teke, Islandesa, Standardbred, Puro Sangue Inglês e Quarto de Milha. Como resultado desses projetos em conjunto, há acessível o mapa contendo a localização de mais de 940 mil SNPs distribuído por todo o genoma equino, exceto no cromossomo Y.

Em 2012 foi publicada a sequência do genoma de uma égua da raça Quarto de Milha, via plataforma de sequenciamento Illumina ™. Foram gerados 59,6 Gb de DNA sequenciado, com cobertura do genoma de 24,7 vezes. As sequências foram mapeadas em 97% do genoma referência do cavalo Puro-Sangue Inglês e, complementando o projeto, identificou-se ainda 3,1 milhões de SNPs (Doan *et al*., 2012).

A partir desses dados, a Illumina ™ disponibilizou, em 2013, um chip de *microarray* de alta densidade, contendo mais de 74 mil SNPs uniformemente distribuídos ao longo do genoma (Neogen, 2013).

Como consequência, vários estudos têm realizado sequenciamento do genoma completo de diversas raças de equinos, incluindo raças domésticas, a fim de compreender os mecanismos genéticos associados ao padrão e estabelecimento racial, por meio de busca de variações estruturais, incluindo a primeira análise de re-sequenciamento de genoma completo em raças domésticas chinesas (Zhang *et al*., 2018).

No início de 2018, uma nova e aprimorada montagem do genoma equino (EquCab3.0) foi disponibilizada (Kalbfleisch *et al*., 2018), impulsionando sucessivos trabalhos para busca de variantes. Além disso, a recente disponibilidade de sequências de genoma completo de raças equinas possibilitou o desenvolvimento de uma nova geração de *SNP array* equino de alta densidade (670k), compreendendo informações genômicas de indivíduos representativos de 24 raças equinas distintas. O estudo catalogou 23 milhões de novas variantes genéticas (Schaefer *et al*., 2017). *SNP arrays* de alta densidade permitem aprimorar abordagens baseadas em população para identificar sinais de seleção e índices de diversidade. Há vários estudos de SNPs associados à características de interesse em animais domésticos a partir de *Chips* de alta densidade, como equinos (McCue *et al.,* 2012), ovelhas (Kijas *et al*., 2012) bovinos (Zhan *et al.,* 2011; Salomon-Torres *et al*., 2016; Valente *et al.,* 2016) e outras espécies de interesse zootécnico.

Quando pretende-se iniciar o estudo de características complexas e peculiares de determinada raça, para qual não há qualquer informação genômica, a partir de sequenciamento de genoma completo, a identificação de todos os genes variantes no genoma é um primeiro passo para a descoberta das variantes causais, eventualmente associadas à essas características (Das *et al*., 2015). Existe grande interesse em variantes genéticas em novas raças de equinos, especialmente SNVs para a criação de base de dados de SNPs e integração de mapeamento de características quantitativas e mapas de ligação a estas, como realizado para a raça Puro Sangue Inglês, a fim de contribuir com as estratégias de melhoramento genético da raça (Joon-Ho *et al*., 2014).

Embora o sequenciamento de genoma completo tenha se tornado uma técnica acessível e de fácil execução para a busca de variantes, a maioria desses estudos são direcionados para raças equinas voltadas para práticas esportivas, tornando-se necessário buscar variantes também em raças naturalizadas. Isso é fundamental para a elucidação genômica e conservação, uma vez que podem conter ricas informações genéticas associadas à características adaptativas peculiares, as quais podem ser inseridas nos programas de melhoramento genético equino, através do desenvolvimento de tecnologias genômicas que utilizem essas informações.

Diante do grande volume de estudos baseados em dados genômicos em larga escala, o número de empresas que desenvolvem tecnologias de NGS tem crescido, baseadas em diferentes princípios químicos para detecção de bases, com maior rapidez, cobertura (repetibilidade) e chances mínimas de erros. Atualmente, é possível gerar dados de sequenciamento de genomas ou transcriptomas por custos muito pequenos, comparado à complexidade dos dados. Em função disso, o uso tem aumentado, impulsionando estudos genéticos e genômicos.

Em genética animal, as novas tecnologias de NGS tem sido estratégias importantes para a conservação de recursos genéticos, estudos filogenéticos, melhoramento genético para diversos fins, bem como para elucidação de diversos processos biológicos, base para qualquer linha de pesquisa aplicada.

6) Genes candidatos aplicados a estudos genômicos

Genes candidatos é um termo que vem ganhando destaque em genômica para estudos que visam compreender as bases genéticas e moleculares associadas a determinadas características fenotípicas.

É comum a seleção de genes que possam estar relacionados a determinado traço quantitativo, ao qual múltiplos genes estão envolvidos, para estudos em indivíduos ou populações que manifestem características contrastantes. Em seguida, o perfil de expressão em larga escala é avaliado ou o sequenciamento desses genes é feito a fim de buscar polimorfismos possivelmente associados às características em estudo (Metzger *et al*.;2014).

É possível sequenciar diretamente os genes de interesse (candidatos), os produtos destes (transcritos) ou quaisquer regiões específicas do genoma, para finalidades diversas, desde estudos bioquímicos e moleculares, até funcionais. A estratégia de sequenciamento *TruSeq Custom Amplicon*, disponibilizada pela empresa Illumina ™ tem sido recentemente aplicada para sequenciar genes candidatos. A seleção destes, previamente ao sequenciamento, deve ser

feita após estudo minucioso dos genes envolvidos na característica fenotípica em investigação. No caso da adaptação dos animais ao calor excessivo, por exemplo, não apenas o fenótipo aparente deve ser considerado, mas também características hematológicas e fisiológicas, uma vez que trata-se de uma adaptação complexa e, portanto, que envolve múltiplas vias metabólicas e ação coordenada de genes (Marai *et al.*, 2007; Starling *et al.*, 2002; Silva, 2000).

Deve-se considerar que a adaptação de equinos da raça Nordestina ao semiárido envolve o conceito genético de adaptação, segundo o qual ocorre um conjunto de alterações herdáveis nas características que favorecem a sobrevivência de uma população de indivíduos em um determinado ambiente, podendo envolver modificações evolutivas em muitas gerações (Egito *et al*., 2002).

As características da pelagem, casco e pele que afetam as trocas de energia incluem cor, densidade, comprimento, diâmetro, profundidade, transmissividade e absorção do calor (Marai e Habeeb, 2010). Destaca-se também, ainda no contexto da adaptação, as características fisiológicas como: frequência respiratória, frequência cardíaca e a evaporação cutânea, além de transpiração, circulação sanguínea, transposte hematológico de oxigênio, entre outras características fisiológicas (Silva, 2000; Marai *et al.*, 2007).

Segundo Marai *et al.* (2007), o aumento dos níveis hormonais de T3, T4 e cortisol são fatores importantes para a adaptação de animais a determinado ambiente. Segundo os autores, para perder calor por evaporação respiratória, o animal aumenta sua frequência respiratória. Nessas situações de estresse calórico, há redução do consumo de oxigênio, a fim de diminuir o metabolismo e, consequentemente, a produção de calor metabólico.

Outro fator essencial para a adaptação animal é a pigmentação da epiderme, pois ela tem associação com a termorregulação. A pigmentação branca, por exemplo, é mais sujeita aos efeitos da radiação ultravioleta (Cunningham, 2004).

Essas informações são essenciais para seleção dos genes candidatos associados a esses processos metabólicos da respiração, estresse oxidativo, termorregulação, estrutura muscular e transporte de oxigênio, além de composição física, como cor de pelagem, diâmetro, espessura e constituição dos cascos e outros anexos da epiderme em geral.

Há bases de dados que permitem busca de genes e análise de enriquecimento pela base de dados do Gene Ontology (GO) a partir de palavras-chave, relacionadas às características fenotípicas para as quais pretende-se identificar genes candidatos. E a partir da simbologia desses genes, pode-se fazer a customização de *kits* de sequenciamento de genes específicos.

É importante destacar que, para identificar variantes, tanto de nucleotídeos únicos, quanto inserções e deleções, é recomendável que seja feito a partir de dados de sequenciamento

de genoma completo (e não apenas genes alvo), uma vez que variantes podem ser identificadas em famílias gênicas não esperadas, principalmente no caso de adaptações exclusivas e peculiares, como no presente trabalho. Sendo assim, NGS para busca de variantes é uma maneira de não perder informações que podem ser extremamente relevantes.

As análises de enriquecimento funcional dos genes que apresentaram as variantes, nos termos do Gene Onthology (GO), o qual é uma iniciativa de bioinformática para a representação dos genes e produtos gênicos, permite a classificação de genes, com base na função molecular, processo biológico e componente celular.

Desse modo, identifica-se as vias metabólicas nas quais esses genes atuam, a partir da simbologia gênica universal do banco de dados do NCBI (*National Center for Biotechnology Information*) ou da busca de bases de dados que disponibilizam publicamente os dados de montagem e anotação de genomas de diversas espécies. Um desses bancos de dados é o Ensembl, o qual contém o genoma referência equino que foi utilizado para o mapeamento das *reads*, obtidas a partir do sequenciamento completo do genoma do cavalo Nordestino.

7) Sequenciamento de próxima geração – Plataforma Illumina

A plataforma de sequenciamento da empresa Illumina™, começou a ser comercializada em 2006 com o nome Solexa. O princípio de funcionamento das duas plataformas comercializadas atualmente pela Illumina ™ (MiSeq e HiSeq) baseia-se na química de sequenciamento por síntese, com nucleotídeos terminadores reversíveis e marcados com fluoróforos, com sinal de fluorescência específico para cada uma das quatro bases do DNA (Illumina, 2017).

Neste método, é feita a construção das bibliotecas de DNA, onde são ligados adaptadores de sequência conhecida às duas terminações, 5’ e 3’ do DNA previamente fragmentado, randomicamente. São feitas milhões de cópias desse complexo DNA-adaptador em termociclador comum e então a biblioteca é inserida no sequenciador, onde há ciclos de incremento das quatro bases modificadas (A, C, T, G) e marcadas, cada uma com um corante fluorescente diferente. Somente quando a base certa é incorporada, o fluoróforo é liberado e o sinal para a geração da imagem é emitido, havendo dessa forma a detecção da cor referente à base. A base incorreta não liga, pois não haverá complementariedade e, portanto, não ocorrerá liberação do fluoróforo. Em seguida, realiza-se um processo de lavagem, no qual as bases não ligadas são removidas e um novo ciclo de incremento de bases é iniciado. Isso ocorre à medida que todo o sequenciamento prossegue.

Todo esse processo ocorre em uma lâmina, chamada *Flow Cell*, onde há fragmentos de DNA imobilizados sobre uma superfície. Esses fragmentos são oligonucleotídeos complementares aos adaptadores adicionados ao DNA que se deseja sequenciar, mencionados acima. Cada fragmento de fita simples imobilizado cria uma estrutura em forma de ponte, ou seja, a outra extremidade do fragmento recém sequenciado hibridiza com os oligonucleotídeos da superfície sólida, sendo clonados e formando agrupamento (cluster).

Os ciclos de adição das bases prosseguem na formação dos clusters, de forma que um mesmo fragmento é sequenciado várias vezes, reduzindo-se muito as chances de erros de posicionamento de bases. A quantidade de ciclos que se deseja obter é ajustável, havendo disponíveis no mercado reagentes de sequenciamento de até 600 ciclos. Com essa tecnologia é possível sequenciar centenas de milhões de clusters simultaneamente (Illumina, 2017).

O sistema HiSeq é adequado para projetos que tenham necessidade de produção de grandes volumes de dados de sequenciamento, em curto período. Ele apresenta dois modos de corrida (*Rapid Run* e *High Output Run*), além da capacidade de sequenciar uma ou duas lâminas *(flow cells)* simultaneamente. Esse equipamento gera até 1 terabase (Tb) de dados.

O sistema MiSeq poder gerar até 15 gigabases (Gb) de dados, sendo o mais indicado para sequenciamento de regiões alvo do DNA, ou conjunto de genes candidatos, de genomas microbianos e sequenciamento de transcriptoma, pela metodologia *RNAseq*. Para genomas maiores, o uso desse sistema deve ser feito em mais de uma corrida, a fim de obter cobertura de sequenciamento adequada, geralmente recomendável que seja de, pelo menos, dez vezes. Isso significa que cada fragmento de DNA da biblioteca, as chamadas *reads*, de tamanho médio entre duzentas e trezentas bases, seja sequenciada pelo menos dez vezes (Illumina, 2017). O princípio de funcionamento de ambos os sistemas (MiSeq e HiSeq), conforme explicado acima, baseia-se na química de sequenciamento por síntese, de forma simultânea para bilhões de fragmentos de DNA durante a corrida (figura 1).

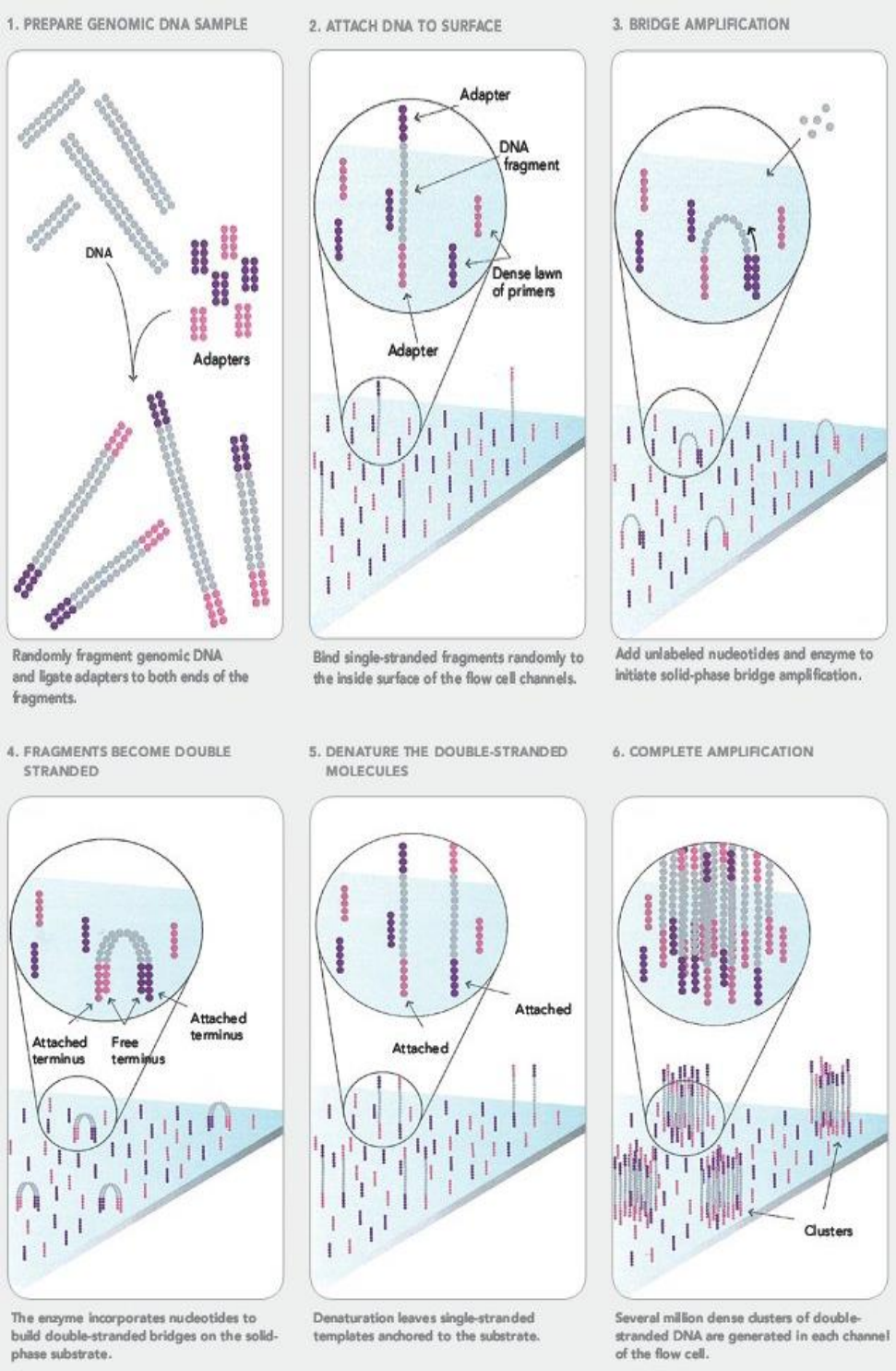

Figura 1: Esquema do preparo da amostra de DNA e da química de sequenciamento por síntese pela plataforma Illumina. 1. Preparo Genômico da amostra, 2. Fixação do fragmento de DNA à superfície sólida da lâmina, 3. Amplificação em ponte, 4. Formação da dupla fita de DNA no fragmento, 5. Denaturação das moléculas de DNA dupla fita, 6. Amplificação completa dos grupos de fragmentos de DNA (clones de fragmentos de DNA) (fonte catálogo online Illumina Inc.).

8) Busca de variantes genômicas

Os milhares de *reads* (fragmentos de DNA de tamanho médio de 300pb geradas como resultado de sequenciamento pela plataforma Illumina ™), resultantes do sequenciamento NGS, necessitam ser alinhadas com o genoma referência (no caso, Ensemble EquCab 3.0, mencionado anteriormente). Antes, porém, aplica-se o controle de qualidade pré-alinhamento para detectar possíveis problemas ou erros sistemáticos que potencialmente afetam a reação de sequenciamento. Estes podem introduzir vieses nas etapas de interrogação das bases pelo sequenciador, produção das bibliotecas, designação de valores de qualidade às bases interrogadas e, finalmente, repercutir em todas as etapas de processamento dos dados, culminando em erros na classificação das variantes e sua eventual interpretação errônea. Comumente utiliza-se o programa FastQC para essa etapa (Souza e Carvalho, 2014).

As divergências com o genoma referência são denominadas variantes, as quais podem ser consideradas importantes causas da variação fenotípica. O "software" BWA é utilizado amplamente pela comunidade científica para o mapeamento ou alinhamento das *reads* contra o genoma referência. Esse *software* implementa três algoritmos, sendo que o BWA-MEM é o indicado para mapeamento de *read*s do tamanho gerado em projetos de NSG pela plataforma Illumina ™ (Church *et al*., 2015).

Os *softwares* Freebayes (Garrison e Marth, 2012) e GATK (*Genome Analysis Toolkit*) (DePristo *et al*., 2011) são as ferramentas baseadas em algoritmos mais utilizadas para a chamada de variantes no processo de mapeamento das *read*s com o genoma referência.

O número de *read*s gira em torno de milhões e varia de acordo com o tamanho do genoma e método de NGS utilizado, enquanto que o número de variantes encontradas depende de uma série de fatores, como o software escolhido para a chamada de variantes, que pode ter filtros distintos para eliminação de variantes falso-positivos, de acordo com o algoritmo e linhas de corte próprios. Esse número varia também com o tamanho do genoma, com a taxonomia do *especimen* em estudo e com a qualidade dos dados de NGS (Paten *et al*., 2014).

Após aplicação de métricas de discriminação de variantes por erros de alinhamento ou sequenciamento e aplicação de filtros com parâmetros pré-definidos por cada software escolhido, deve-se fazer a anotação das chamadas "variantes reais". O VEP (*Variant Effect Predictor*) (McLaren *et al.*, 2010) é a ferramenta de anotação bastante utilizada e insere informações de genes e transcritos que contém a variante em questão, fornece a localização desta no contexto genômico em que ela se insere e faz predição de consequências na codificação de proteínas, do potencial deletério e associações entre esta variante e condições conhecidas,

entre outras análises. É em função disso que em um projeto de busca de variantes, o número de efeitos dessas variantes é muito maior que a contagem de variantes propriamente dita.

III OBJETIVOS

Objetivo Geral

Identificação de variantes genômicas (SNVs e InDels) na raça equina Nordestina, com possível associação às características de rusticidade e resistência às condições inóspitas do semiárido nordestino, a partir de chamada de variantes em dados de sequenciamento de genoma completo de um representante clássico da raça.

Objetivos Específicos

- Validar SNVs de impacto alto e moderado, os quais tem potencial de alteração do transcrito, em estudo populacional da raça (60 animais) em três regiões do sertão nordestino brasileiro (estados da Bahia, Pernambuco e Piauí), por meio do sequenciamento automático capilar das regiões gênicas contendo as SNVs.
- Comparar as frequências genotípicas e alélicas dos *loci* polimórficos.
- Selecionar grupos de genes que possam estar envolvidos na resistência fenotípica, característica do padrão racial do cavalo nordestino.
- Utilizar estratégias de genômica comparativa a partir de dados de NGS entre as raças equinas Nordestina e Puro Sangue Inglês.
- Classificar as variantes (SNVs e InDELs quanto à localização no genoma e quanto ao efeito de impacto (alto, moderado, baixo e modificador-*modifier*).
- Disponibilizar os dados genômicos em larga escala e a chamada de variantes, a fim de promover a inclusão da raça equina Nordestina entre raças naturalizadas envolvidas no Programa de Pesquisa em Recursos Genéticos, coordenado pelo Centro Nacional de Pesquisa de Recursos Genéticos e Biotecnologia (CENARGEN).

IV CAPÍTULO 1

Whole-genome sequencing of a Brazilian naturalized horse breed resistant to arid climate for identifying single nucleotide variants and insertions/deletions

Danielle Cunha Cardoso[1*¶], Eduardo Geraldo Alves Coelho[1#], Brenda Neves Porto[2&], Glacy Jaqueline da Silva[3€], Denea de Araújo Fernandes Pires[4$], Denise Aparecida Andrade de Oliveira[1≠£].

[1] Departamento de zootecnia, Universidade Federal de Minas Gerais, Escola de Veterinária, Belo Horizonte, Minas Gerais, Brazil

[2] Embrapa Recursos Genéticos e Biotecnologia, Brasília, Distrito Federal, Brazil, Brazil

[3] Departamento de Biotecnologia**,** Universidade Paranaense, Umuarama, Paraná, Brazil

[4] Instituto Federal de Estado de Pernambuco, Barreiros, Pernambuco, Brazil

These authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

*corresponding author

E-mail: danielleccard@gmail.com

¶[#] Conceived of and designed the experiments.

¶[&€] Analyzed the data.

[$] Contributed to sampling.

¶[€≠] Contributed to the writing of the manuscript.

[£] Contributed to the acquisition of reagents and materials

All the authors reviewed and approved the final manuscript.

Abstract

In this study, we perform the search for variants (SNVs and InDels) in the genome of a Brazilian Naturalized horse breed using FreeBayes and GATK variant calling tools. This breed presents exclusive adaptive traits of extreme importance to semi-arid condition, such as those that allow survival under excessive sunlight, rainfall, low forage availability and stony ground. Moreover, these traits are expressed without any detriment to the performance and perpetuation of the breed. A total of 305,588,364 *read*s were mapped to the horse reference genome (EquCab3.0 from the Ensembl database) and 1,598,210 single nucleotide variations (SNVs) and 138,139 insertions/deletions (InDels) were detected by FreeBayes and 88,838 (SNVs) and 25,232 (InDels) by GATK. Both have been used in order to increase the safety of variant calls, identify in which regions of the genome they are present and check for variants in genes possibly associated with the peculiar traits exhibited by the breed. The variants annotation identified numerous non-synonymous SNVs and frameshift InDels which could affect phenotypic variation. We found 28 and 392 Emsembl gene IDs containing high and moderate impact SNVs, respectively, in agreement between GATK and FreeBayes, including GTPase family members, olfactory receptors, mitochondrial complex and defense genes. Functional enrichment analysis was performed and revealed that variants in the olfactory transduction pathway was over represented. The variability identified in these genes, possibly has relevant importance in the gorgeous adaptation of the Nordestino horse breed to the semi-arid climate of the Brazilian Northeast.

Introduction

The Nordestino horse breed is a Brazilian naturalized breed, developed from the introduction of Iberian and Barb-Arabic breeds in the Northeast of Brazil, after Portuguese colonization. It presents adaptive traits of extreme importance to semi-arid conditions. The animals representing the breed exhibit greater resistance to disease and phenotypic rusticity in the racial pattern, such as rigid hooves, medium and arcuate body conformation, strong musculature and bones, rectilinear cephalic profile, and dilated nostrils [1,2]; which allow survival under excessive sunlight, rainfall, low forage availability and stony ground, without any detriment to the performance and perpetuation of the breed.

It is unlikely that other non-specialized horse breeds would develop activities in this environment that require strength and endurance. Understanding the molecular basis associated with these traits becomes essential to advance breeding programs and breed conservation, since it is in danger of disappearing and therefore requires urgent actions for its conservation [3]. To date, there are no studies in this breed on genetic structural variations, including single nucleotide variations (SNVs) and insertions/deletions (InDels) or the association to the resistance exhibited by the breed to the arid conditions, or even functional studies related to the genetic peculiarity exhibited by the breed.

Phenotypic variations associated with mutations co-occur with domestication as an effect of the impact between selective breeding, controlled by human actions and the performance of natural selection [4]. As a result, most equine breeds present populations with high phenotypic and genotypic uniformity within the breed, but between breeds there is a lot of variation.

The equine genome project (*Equus caballus*) has publicly made available a full and high quality genome data of a Thoroughbred female, representing a breakthrough in genomics and veterinary medicine [5]. As a consequence, several studies have performed complete genome sequencing of several equine breeds, including domestic breeds, in order to understand the genetic mechanisms associated with pattern and racial establishment, from pursuit of structural variations, including the first analysis of re-sequencing of a complete genome, identifying significant variations in the Quarter Horse breed [6] and in Chinese horses (Lichuan and Kazakh breeds) [7].

In early 2018, a new and improved equine genome assembly (EquCab3.0) was made available [8], boosting successive searches for variants. Furthermore, the recent availability of complete genome sequences of horse breeds allowed the development of a next generation,

high-density equine SNP array (670k), comprising genomic information from individuals representative of 24 different equine breeds. The study cataloged 23 million new genetic variants [9]. High-density SNP array enables the enhancement of population-based approaches to identify selection signals and diversity indexes.

There are several studies of SNPs associated with traits of interest in domestic animals, from high density Chips, such as equines [10], sheep [11] cattle [12, 13, 14] and other species of zootechnical interest.

When it is intended to start the study of complex and peculiar traits of a particular breed, for which there is no genomic information, starting from whole genome sequencing, the identification of all variant genes in the genome is a first crucial step for the discovery of causal variants, possibly associated with these traits [15]. There is a great interest in genetic variants in new equine breeds, especially SNVs, for the creation of SNPs database and integration of quantitative and linkage maps, as performed for the Thoroughbred breed, in order to contribute to breeding strategies [16].

Although the whole genome re-sequencing has become an accessible and easy-to-perform technique for variant search, most of these studies are targeted to equine breeds aimed to sports practices. Thus, it is necessary to search for variants in naturalized breeds, in order to elucidate genomic and conservation, since these may contain rich genetic information associated with peculiar adaptive traits. This informations can be inserted into equine breeding programs through the development of genomic technologies.

Here we present the first complete genomic sequence and characterization of the genetic variations of a Brazilian naturalized breed specimen, a male of the Nordestino horse breed, including SNVs and InDels, with genetic annotation analyzes. That annotation allow to identify, locate and associate variations related to the complex traits of resistance that are peculiar to the breed, as well as subsequent studies on origin, genomic characterization and population studies, especially about the segregation of variants in the remaining population.

Materials and methods

Ethics statement

The blood sample was collected from a male horse in a private property with written consent of the owner, without experimental planning on the property, or experimental interventions that cause damage or non-momentary pain and suffering to the animal. Therefore,

no specific ethical approval is needed (Brazil law number 11794, from October 8th, 2008, Chapter 1, Art. 3, paragraph III).

Sample collection

The DNA sample was extracted from a blood aliquot of a male specimen typical of the Nordestino horse breed, from Pernambuco state, belonging to the Caatinga biome (semi-arid climate). The specimen presents all the phenotypic traits of the breed, such as mean weight, height (138 cm ± 8), hull type, characteristic stiffness and head size [17]. The sample was kindly provided by the research group of Dr. Jânio Benevides Melo and Dr. Denea de Araújo Fernandes Pires, co-authors of the present study.

Genomic DNA extraction, library preparation and genome sequencing

Genomic DNA samples were extracted in replicates, using the DNeasy Blood & Tissue Kit (QIAGEN Pty. Ltd., Venlo, Netherlands), according to the protocol provided by the manufacturer. The DNA quality was determined by NanoDrop 1000 spectrophotometer (Thermo Scientific, Wilmington, DE, USA). Then the samples were quantified in Qubit 2.0 fluorometer (Thermo Scientific, Wilmington, DE, USA) using the Qubit ™ dsDNA BR (Broad Range) Assay Kit, following the manufacturer's instructions. DNA libraries were synthesized from 50 ng of genomic DNA, using the Nextera DNA Sample Preparation Kit and the Nextera Index Kit (Illumina, San Diego, CA, USA), according to the manufacturer´s protocol. Size estimation of the library was performed on a 4200 Tape Station (Agilent Technologies) and quantified using a KAPA library quantification kit (Kapa Biosystems, MA, USA) according to Illumina's library quantification protocol. Based on the qPCR quantification, the libraries were normalized to 12 pM, denatured using 0.1 N NaOH and sequenced using the MiSeq Reagent Kit v3-600cycle (2 × 301 bp paired-end *read*s) in Illumina MiSeq Sequencer (Illumina, San Diego, CA, USA). Sequencing Control Software (Illumina, San Diego, CA, USA) was used to process the raw fluorescent images and the called sequences. Upon completion of the sequencing, DNA libraries that remained frozen on double stranded were dissociated and normalized, repeating the sequencing, in order to obtain twice the coverage of the genome sequencing.

Filtering and mapping processes

Before the mapping process of the sequenced *reads*, raw *reads* were filtered using FastQC software, version 0.11.7 (cutoff read length for high quality, 70%; cutoff quality score, 20) [18]. For the *reads* mapping, it was used the horse reference genome (Ensembl EquCab3.0). Clean sequencing *reads* were mapped to the reference assembly using the Burrows-Wheeler Aligner tool (BWA, version 0.7.10-r789) with default parameters [19]. PCR duplicates were detected and removed using the Picard tools (version 1.54) (http://broadinstitute.github.io/picard/). Then, a re-alignment of the *reads*, using one of the Genome Analysis tools, Toolkit (GATK, version 3.8), was done to improve the mapping quality [20]. Downstream processing was carried out using typical GATK pipeline according to parameters applied by Cornish and Guda [21], for the base quality score recalibration (BQSR) step of GATK.

Variant detection and annotation

Variant calling was conducted with two tools: FreeBayes (https://github.com/ekg/freebayes) [22] and GATK (https://software.broadinstitute.org/gatk) [23] in order to ensure greater reliability of the search for variants. All SNVs and InDels were identified as differences from the reference genome sequences. In the Variant call conducted with FreeBayes, the variant list was filtered by vcffilter (https://github.com/vcflib/vcflib). We filter calls using GATK's recommended hard filters, instead of Variant Quality Score Recalibration (VQSR).

The SNVs and InDels were functionally annotated with the SnpEff software [24], with default settings. For each putative SNP, the useful annotation and position were identified based on the gene annotation of the horse reference genome, obtaining the effect of the variants and their impact; and according to the effect, the functional class of the variant, possible codon and/or amino acid change, gene name, biotype gene, gene coding, transcript identity and position of the exon or intron.

Functional enrichment

For functional enrichment, we selected all Emsembl IDs containing SNVs of high and moderate impact, present in the variant call analysis according to both GATK and FreeBayes.

The Gene Ontology (GO) terms were obtained using the Databank for Annotation, Visualization and Integrated Discovery (DAVID) [25]. This databank was used to evaluate enrichment in the GO terms using known annotations of horse genes with *Equus caballus* selected as background. For GO term analysis, we considering a 10% FDR (False Discovery Rate) threshold for significance.

Results and Discussion

Genomic variants in Nordestino horse breed

A total of 28 Gb of paired-end sequence data were produced from whole-genome sequence data of a male of the Nordestino horse breed, with 11.2 fold the genome coverage, considering the sum of the sequencing runs performed. A total of 305,588,364 *reads* were mapped to the horse reference genome (EquCab3.0 from the Ensembl database) with a mapping rate of 96.05%.

At an effective genome size of 2,462,676,227 bases, the total of 1,741,210 variants were identified using FreeBayes; therefore, a variation rate of 1 variant per 1,414 bases, relative to the reference genome. Among these, 1,598,210 were classified as SNVs; 57,580 as insertions and 80,529 as deletions. In particular InDels analysis, a total of 4,964 were classified as structural variants, being 54 insertions, 3 deletions and 4,907 mixed.

When we applied the Genome Analysis Toolkit (GATK), we identified 88,848 variants classified as SNVs (1 variant every 27,470 bases), for an effective genome size of 2,440,521,205 bases. In the search for InDels, 10,006 insertions and 15,226 deletions (1 InDel per 96,300 bases) were identified for an effective genome size of 2,429,851,222 bases (Table1).

Table 1. Variants number in the Nordestino horse genome by FreeBayes and GATK variant calling tools.

Variants	FreeBayes	GATK
SNV	1,598,210	88,838
INS	57,580	10,006
DEL	80,559	15,226
MIXED	4,861	_
SNV Rate	1/1,540	1/27,470
INDEL Rate	1/17,221	1/96,300

Using recommended quality metrics for each software, was found that the total number of variants detected by FreeBayes were higher than that of GATK, which is expected since GATK exhibits higher sensitivity while maintaining a lower number of false positive SNVs [21].

We do not intend to compare variant calling tools. However, both have been used in order to increase the safety of variant calling and to identify in which regions of the genome they are present and to check for variants in genes possibly associated with the peculiar traits exhibited by the breed. From these data, in a successive study, we intend to validate SNPs in a population of nordestinos horses. In addition, we do not intend, at this time, to compare new SNVs and already known and present in SNPs arrays, once this is the first study in the breed and we initially made the search for variants from the complete genome of a single specimen of the breed.

Characterization of SNVs and InDels

Calling variants were distributed through 3,113 supercontigs in the analysis of FreeBayes and through 1,520 supercontigs using the GATK, which constitute, respectively, 98.23% and 97.34% of the horse genome (Additional file1). We found 12.23% of SNVs effects in intergenic regions, 37.63% of introns, and 1.02% of exons in FreeBayes analysis. According to GATK, 19.19% of SNVs are located in intergenic regions, 33.05% in introns and 1.16% in exons. In both variants calls tools, SNVs and InDels had very low occurrence in splice-site donor and acceptor sequences (values close to 0.01%), splice-site regions (approximately 0.08%) and UTR 5’ (0.17%) and somewhat higher occurrence in UTR 3’ region (values next to 0.4%) (Fig 1). The actual effects number is greater than the number of variants because those that are found between genes positioned in close proximity can have their effects categorized as both downstream and upstream.

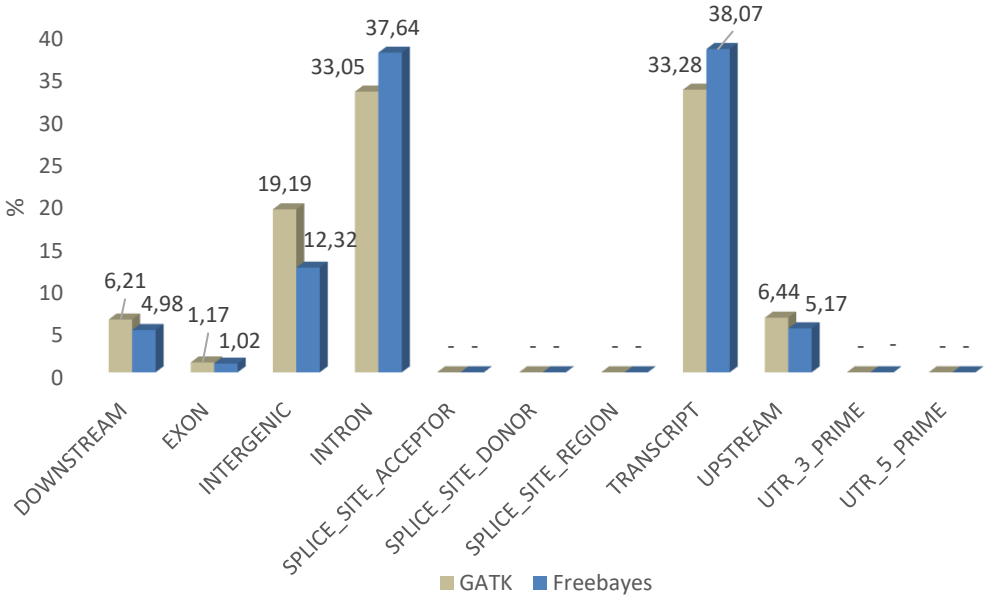

Fig 1. SNVs effects percentage by genomic region through the FreeBayes and GATK variant calling tools. SNPEff software was used to categorize the variants effects based on position in the genome. These include exons, introns, untranslated regions (5 'UTR and 3' UTR), splice site donor, splice site acceptor and region, transcripts and intergenic regions. “Downstream” and “Upstream” is defined as regions 5 kilobase (kb) downstream of the most distal polyA addition site and 5 kilobase (kb) upstream of the most distal transcription start site respectively [24]. Splice_region means that a variant is within 2 bp of a splice junction. Splice_acceptor means that the variant hits a splice acceptor site (defined as 2 bases before the exon start site, except for the first exon). Splice_donor means that the variant hits a splice donor site (defined as 2 bases after the end of the coding exon, except for the last exon [7].

The percentage of SNVs effects per genotypic region of the Nordestino horse presented here was very similar to the average percentage found in Chinese native breeds (Lichuan and Kazakh breeds, small and rugged horses), as demonstrated by Zhang *et al*. [7] by SNVs calling conducted with GATK and functional annotation based on SnpEff software, from whole-genome sequencing data. From the total of single nucleotide Variations, defined by the authors as single nucleotide polymorphisms (SNPs), the most intense effects was transcript, intron and intergenic (29.07%, 28.12% and 27.02%, respectively) and the smaller effects were also exactly the same as found here and with very similar percentages.

Using the SnpEff program [24], we also classified the effects of variants (SNVs and InDels) by impact as modifiers in large part, high, moderate and low impact of variants called by GATK and FreeBayes. We showed the additional effect of SNVs and InDels on FreeBayes data and the exclusive effect of SNVs on GATK data (Fig 2). The effects of SNVs and InDels in all categories were much higher in FreeBayes analysis, due to the combined effect of these

variations. GATK analysis revealed greater sensitivity for exclusion of "false-positive" (PF) variants, as previously mentioned.

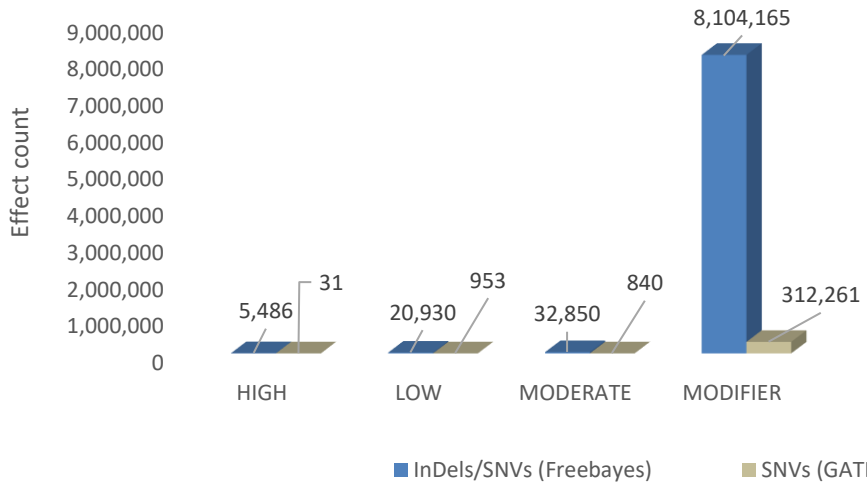

Fig 2. Number of variants effects by impact acoording to FreeBayes (InDels + SNVs) and GATK (SNVs) software's. SNV effects were categorized by impact as high (affecting splice-sites, stop and start codons), low (synonymous coding/start/stop, start gained), moderate (non-synonymous) and modifier (upstream, downstream, intergenic, UTR).

Selection of genes containing high and moderate effects SNVs

Based on the effect of the variants and their annotation by SnpEff, we have identified all genes or Emsembl IDs in which high, low, moderate and modifier impact effects (SNVs and InDels) occur, based in variants calling results both by GATK, and FreeBayes. In order to prioritize single nucleotide variants, which can be characterized as SNPs in subsequent population studies, we have screened the genes that have at least one such variation that has high impact (disruptive impact in the protein causing protein truncation, loss of function or triggering nonsense mediated decay) and then check which genes are present in the analyzes by both variants calling tools. We found 28 Emsembl IDs from multigenic families containing at least one high impact SNVs, in agreement between GATK and FreeBayes (Table 2). Among these, a pseudogene and *GTPase* Family members 7, 4, 2 and 1.

Table 2. Total of Nordestino horse genes containing High Impact SNVs in agreement between the GATK and FreeBayes variant calling tools.

				GATK	FreeBayes
GeneName	Transcript	Product	BioType	SNVs impact HIGH	SNVs impact HIGH
LOC102149342	*gene32392*		pseudogene	1	1
LOC100146699				1	1
(id196201)	*rna15620*	GTPase family member 7	protein_coding		
id196202	*rna15621*	GTPase family member 7	protein_coding	1	1
id196203	*rna15622*	GTPase family member 7	protein_coding	1	1
id196204	*rna15623*	GTPase family member 7	protein_coding	1	1
id196205	*rna15624*	GTPase family member 7	protein_coding	1	1
id196206	*rna15625*	GTPase family member 7	protein_coding	1	1
id196207	*rna15626*	GTPase family member 7	protein_coding	1	1
id196208	*rna21264*	GTPase family member 7	protein_coding	1	1
id196209	*rna21265*	GTPase family member 7	protein_coding	1	1
id196210	*rna21266*	GTPase family member 7	protein_coding	1	1
id196211	*rna25375*	GTPase family member 7		1	1
id196213	*rna38835*	GTPase family member 7	protein_coding	1	2
id196214	*rna38989*	GTPase family member 7	protein_coding	1	1
id196215	*rna40155*	GTPase family member 7	protein_coding	1	2
id196216	*rna43275*	GTPase family member 7	protein_coding	1	1
id196217	*rna51848*	GTPase family member 7	protein_coding	1	1
id196219	*rna63380*	GTPase family member 7		1	1
id196220	*rna63381*	GTPase family member 7		1	1
LOC111773116				1	1
(id196221)	*rna71433*	GTPase family member 4	protein_coding		
id196222	*rna71664*	GTPase family member 4		1	1
id196223	*rna73843*	GTPase family member 4	protein_coding	1	1
id196224	*rna73844*	GTPase family member 4	protein_coding	1	1
LOC100054458				1	1
(id196225)	*rna73845*	GTPase family member 2	protein_coding		
id196226	*rna76474*	GTPase family member 2	protein_coding	2	5
id196227	*rna76774*	GTPase family member 2	protein_coding	1	1
id196228	*rna76804*	GTPase family member 2	protein_coding	1	1
LOC100063777 (id196229)	*rna76872*	GTPase family member 1	protein_coding	1	2

Considering also the relevance of moderate impact effects, we identified 392 Ensembl IDs containing SNVs with this effect, in agreement between GATK and FreeBayes (Additional

file 2). Among these, we selected 70 genes for functional enrichment analysis and Gene Ontology (GO) terms using the Databank for Annotation, Visualization and Integrated Discovery (DAVID). The screening for 70 genes was done with the purpose of allowing the data presentation in a non-additional table. These data allowed the timely identification of regions where there are variations of high and moderate impacts, including variations in genes in which impacts on gene transcription can occur and verify the occurrence of these variations in candidate genes, possibly related to the resistance traits to the arid conditions exhibited by the Nordestino horse breed.

The 70 Emsembl IDs with SNVs of moderate impact selected from the 392, are representative from 33 human orthologous genes (*ABCD2, ARHGAP20, ARHGAP28, ATP6, CHFR, COX2, COX3, CYT, ERP27, EXOC6, FDFT1, GBP7, GIMAP7, HYAL4, KIF1, KLRK1, LCORL, LY49F, NBPF7, OR10D4, OR52L2, OR56A3, OR56A4, OR6B2, OR7G2, OR8S1, PDPR, RNASEL, RWDD3, SLC45A1, SWT1, WAPL, WD40*). We explored these genes functions (which contained high and moderate SNVs impact) associated with various biological processes. The P value of .05 was considered significant for GO annotations. Gene Ontology enrichment analysis for these genes revealed eight GO biological processes and nineteen GO molecular functions, acting in nine metabolic pathways (Table 3).

Table 3. Gene ontology (GO) terms and enriched KEGG pathways (False Discovery rate (FDR)<0.10) of the selected gene set containing high and moderate impact SNVs in agreement between GATK and FreeBayes.

Category	ID	Name	FDR	Genes with high and moderate impact SNVs
GO Molecular Function	GO:0004984	olfactory receptor activity	5,37E-01	*OR8S1,OR56A3,OR56A4,OR6B2,OR7G2,OR52L2P,OR10D4P*
	GO:0015078	proton transmembrane transporter activity	1,81E+00	*MT-ATP6,MT-CO2,MT-CO3,MT-CYB*
	GO:0007186	G protein-coupled receptor signaling	1,48E+01	*OR8S1,OR56A3,OR56A4,OR6B2,OR7G2,OR52L2P,OR10D4P*
	GO:0009055	electron transfer activity	1,48E+01	*MT-CO2,MT-CO3,MT-CYB*
	GO:0004888	transmembrane signaling receptor activity	1,48E+01	*KLRK1,OR8S1,OR56A3,OR56A4,OR6B2,OR7G2,OR52L2P,OR10D4P*
	GO:0016676	oxidoreductase activity	1,48E+01	*MT-CO2,MT-CO3*
	GO:0004129	cytochrome-c oxidase activity	1,48E+01	*MT-CO2,MT-CO3*
	GO:0015002	heme-copper terminal oxidase activity	1,48E+01	*MT-CO2,MT-CO3*
	GO:0016675	oxidoreductase activity	1,48E+01	*MT-CO2,MT-CO3*
	GO:0038023	signaling receptor activity	1,81E+01	*KLRK1,OR8S1,OR56A3,OR56A4,OR6B2,OR7G2,OR52L2P,OR10D4P*
	GO:0051996	squalene synthase activity	1,81E+01	*FDFT1*
	GO:0004310	farnesyl-diphosphate farnesyltransferase activity	1,81E+01	*FDFT1*

	GO:0015077	transmembrane transporter activity	3,02E+01	*MT-ATP6,MT-CO2,MT-CO3,MT-CYB*
	GO:0032394	MHC class Ib receptor activity	3,10E+01	*KLRK1*
	GO:0060089	molecular transducer activity	3,89E+01	*KLRK1,OR8S1,OR56A3,OR56A4,OR6B2,OR7G2,OR52L2P,OR10D4P*
	GO:0004741	[pyruvate dehydrogenase phosphatase activity	3,89E+01	*PDPR*
	GO:0022857	transmembrane transporter activity	3,89E+01	*SLC45A1,MT-ATP6,MT-CO2,ABCD2,MT-CO3,MT-CYB*
	GO:0016491	oxidoreductase activity	4,77E+01	*PDPR,MT-CO2,MT-CO3,MT-CYB,FDFT1*
GO				
Biological		chemical stimulus in sensory perception		
Process	GO:0050911	of smell	1,42E+00	*OR8S1,OR56A3,OR56A4,OR6B2,OR7G2,OR52L2P,OR10D4P*
	GO:0007608	sensory perception of smell	1,42E+00	*OR8S1,OR56A3,OR56A4,OR6B2,OR7G2,OR52L2P,OR10D4P*
	GO:0050907	sensory perception	1,42E+00	*OR8S1,OR56A3,OR56A4,OR6B2,OR7G2,OR52L2P,OR10D4P*
	GO:0009593	detection of chemical stimulus	1,67E+00	*OR8S1,OR56A3,OR56A4,OR6B2,OR7G2,OR52L2P,OR10D4P*
	GO:0007606	sensory perception of chemical stimulus	1,67E+00	*OR8S1,OR56A3,OR56A4,OR6B2,OR7G2,OR52L2P,OR10D4P*
	GO:0050906	detection of stimulus involved in sensory perception	1,67E+00	*OR8S1,OR56A3,OR56A4,OR6B2,OR7G2,OR52L2P,OR10D4P*
	GO:0051606	detection of stimulus	8,72E+00	*OR8S1,OR56A3,OR56A4,OR6B2,OR7G2,OR52L2P,OR10D4P*
	GO:0022900	electron transport chain	4,92E+01	*MT-CO2,MT-CO3,MT-CYB*
KEEG				
Pathway	1269583	Olfactory Signaling Pathway	9,52E-01	*OR8S1,OR56A3,OR56A4,OR6B2,OR7G2,OR52L2P,OR10D4P*
	1270121	The citric acid cycle and respiratory electron transport	9,52E-01	*PDPR,MT-ATP6,MT-CO2,MT-CO3,MT-CYB*
	1270127	Respiratory electron transport, ATP synthesis and heat production.	3,36E+00	*MT-ATP6,MT-CO2,MT-CO3,MT-CYB*
	82942	Oxidative phosphorylation	3,36E+00	*MT-ATP6,MT-CO2,MT-CO3,MT-CYB*
	93344	Cardiac muscle contraction	7,00E+00	*MT-CO2,MT-CO3,MT-CYB*
	1270128	Respiratory electron transport	1,27E+01	*MT-CO2,MT-CO3,MT-CYB*
	83087	Olfactory transduction	1,27E+01	*OR8S1,OR56A3,OR56A4,OR6B2,OR7G2*
	1269574	GPCR downstream signaling	2,66E+01	*OR8S1,OR56A3,OR56A4,OR6B2,OR7G2,OR52L2P,OR10D4P*
	142287	epoxysqualene biosynthesis	3,56E+01	*FDFT1*
Gene				
Family	167	Olfactory receptors, family 56	1,31E+00	*OR56A3,OR56A4*
	643	Mitochondrial complex: cytochrome c oxidase subunits	3,10E+00	*MT-CO2,MT-CO3*
	721	Rho GTPase activating proteins	1,44E+01	*ARHGAP28,ARHGAP20*
	808	ATP binding cassette subfamily D	3,03E+01	*ABCD2*
	580	GTPases, IMAP	4,31E+01	*GIMAP7*
	1055	Exocyst complex	4,31E+01	*EXOC6*
	642	Mitochondrial complex III	4,31E+01	*MT-CYB*
	621	CD molecules|Killer cell lectin like receptors	4,52E+01	*KLRK1*

The *GTPse* gene family was the only one that presented high impact SNVs in agreement with the variant calling between both softwares. Members of this family are key regulators of most processes in the cell, including proliferation, differentiation, vesicle and organelle dynamics, transport and regulation of the cytoskeleton [26]. These are evolutionarily conserved proteins

couple extracellular signals to various cellular responses through an ability to undergo conformational changes in response to the alternate binding of GDP and GTP. The GDP-bound "off" or "on" state recognize distinct effector proteins, thereby allowing these proteins to function as binary molecular switches [27]. Considering the functional importance of this gene family in basal cellular processes, the presence of SNVs with possible disruptive impact in the protein causing protein truncation or loss of function should be carefully evaluated. Thus, SNVs present in this restricted group of genes will be reevaluated and validated in a Nordestino breed population in our successive study. A study conducted by Zhang *et al*. [7] also identified high impact SNVs in the metabolic pathway associated with GTPases in enriched biological processes from GO analysis in horses of 14 breeds. The highest count of genes with this type of SNVs impact effect (55) are G protein–coupled receptor signaling pathway (GO:0007186), in which we also observed high gene representativity, but with SNVs of moderate effect (7 out of 33 selected for functional analysis). G protein–coupled receptors activate signal transduction pathways and coupling with G proteins, they pass through the cell membrane seven times, being called seven-transmembrane receptors [28].

When we analyzed moderate impact SNVs, the variety of gene families in which they occur is wide, drawing attention to olfactory receptors family, which had high representativeness. These genes involved in the Olfactory Signaling pathway act in the perception of odor through olfactory, interact with odorant molecules in the nasal epithelium, to initiate a neuronal response that triggers the perception of a smell [29]. The biochemical signaling events related to this (super pathway) act in food recognition and consequently food preference [30], identification of sexual partners [31], mother-infant bonding [32] and several other aspects of animal survival. Among them, we can also highlight the variable susceptibility to intranasal infections, as the study by Kupke *et al*. [33], which analyzed the proteins expression of this pathway in the equine nasal epithelium, in association with this susceptibility. The SNVs present in genes associated with this pathway, once validated in more individuals of the Nordestino horse breed, may be associated with resistance, including respiratory diseases and the phenotypic profile of rusticity exhibited by Nordestino horse, a profile characterized by Melo *et al*. [2]. The high representativeness of this genetic family in our moderate impact SNVs calling can be explained by the fact that Mammalian Olfactory Receptor (OR) Genes constitute a large family. In humans, for example, they are 390 OR genes and 465 pseudogenes [34], since these receptors recognize varied binders, from chemical compounds to peptides [30].

Stafuzza *et al*. [35] in a cattle variant calling study, the olfactory transduction pathway was over represented in all four important cattle breeds in Brazil: Guzerat, Gyr, Girolando and

Holstein. Metzger *et al*. [36] identified InDels with codon shift effect of OR genes on horses of the Hanoverian and Arabian breeds, including the *O56A3* gene, in which we also identify SNVs with moderate impact on the Nordestino horse breed (Table 3). They also investigated codon changes due to private InDels occurrence in breed horses compared to non-breed (Przewalski) horses and revealed higher occurrence of these variants in genes involved in immune system processes in breed horses.

Jun *et al*. [37] characterized the genome of the Marwari horse (from the complete genome sequencing of a male Marwari horse) an Indian rare breed with unique phenotypic traits. The variant calling results by SAMtools software and functional enrichment analysis also showed that the genes with Nonsynonymous SNVs and/or InDels in coding regions were highly enriched in olfactory functions.

Immune regulation and metabolic processes also contained variants of impact on gene transcription. As mentioned in the study by Metzger *et al*. [36], the high density of mutations in domestic equine breeds seems to be concentrated in metabolic pathways related to the signaling of basal cellular mechanisms, known as house-keeping, to the signaling of the immune system and mostly in olfactory genes, also associated with the perception of chemical stimuli. This variability, specifically in these last two gene classes, seems to have great importance in promoting the adaptation of these domestic breeds to specific environments, being exactly the one observed for the breed studied in the present work, which exhibits high adaptation to the inhospitable environment of the semi-arid region of northeastern Brazil.

It should be considered that due to the small sampling of this research, care is needed in the interpretation of the over-represented pathways and the terms and results of the GO. However, these results provide genomic information of extreme importance to investigate the genetic mechanisms associated with the exclusive phenotypic differences of the Nordestino horse breed.

Conclusion

This is the first genomic data for a Naturalized Brazilian horse breed and it is an invaluable resource for future studies of genetic variation associated with the exclusive phenotype of the Nordestino horse breed. Comparing its genome to the horse reference genome, approximately 89 thousand SNVs and 10 thousand InDels were identified. We prioritized variants of high (affecting splice-sites, stop and start codons) and moderate impacts (non-synonymous), especially SNVs, and identified 28 Ensembl IDs in which high impact SNVs are present and 392 Ensembl IDs contain moderate impact SNVs. The functional enrichment analysis indicated that the GTPase IMAP Family was the only one that presented high impact SNVs and the genes with non-synonymous SNVs in coding regions were highly enriched in olfactory functions, sensory perception of smell and metabolic processes. It is possible that the variability in these gene families has relevant importance in the gorgeous adaptation of the breed to the semi-arid climate of the Brazilian Northeast. Therefore, this study provides the basis for validation of variants in a population study of this breed to identify genomic markers, such as SNPs, associated with the exclusive phenotype and the molecular mechanisms involved. The genomic insights may aid in breed conservation and in development of resistance markers to arid climate conditions.

References

1. Mariante AS, Albuquerque MSM, Ramos AF. Criopreservação de recursos genéticos animais brasileiros. Rev. Bras. Reprod. Anim. 2011; 35(2): 64-68.

2. Melo JB, Pires DAF, Ribeiro MN. Perfil fenotípico do remanescente do cavalo nordestino no nordeste do Brasil. Archivos de Zootecnia. 2013; 62(328) 171-180. doi:104321/s0004-05933013000200002.

3. Pires DAF, Coelho EGA, Melo JB, Oliveira DAA, Ribeiro MN, Gus Cothran E, et al. Genetic diversity and population structure in remnant subpopulations of nordestino horse breed. Archivos de Zootecnia. 2014; 63(242): 349-358. doi:10.4321/S0004-05922014000200013.

4. Andersson L. How selective sweeps in domestic animals provide new insight into biological mechanisms. J Intern Med. 2012; 271(1):1-14. doi: 10.1111/j.1365-2796.2011.02450.x.

5. Wade CM, Giulotto E, Sigurdsson S, et al. Genome sequence, comparative analysis and population genetics of the domestic horse. Science. 2009; 326(5954):865–867. doi:10.1126/science.1178158.

6. Doan R, Cohen ND, Sawyer J, Ghaffari N, Johnson CD, Dindot SV. Whole-genome sequencing and genetic variant analysis of a Quarter Horse mare. BMC Genomics. 2012; 13:78. doi:10.1186/1471-2164-13-78.

7. Zhang C, Ni P, Ahmad HI, Gemingguli M, Baizilaitibei A, Gulibaheti D, et al. Detecting the Population Structure and Scanning for Signatures of Selection in Horses (*Equus caballus*) from Whole-Genome Sequencing Data. Evol Bioinform. 2018; 4(14): 1-9. doi: 10.1177/1176934318775106.

8. Kalbfleisch TS, Rice ES, DePriest Jr MS, Walenz BP, Hestand MS, Vermeesch JR, et al. (2018) EquCab3, an updated reference genome for the domestic horse. B*ioRxiv*. 2018. doi:10.1101/306928. Cited 18 November 2018.

9. Schaefer RJ, Schubert M, Bailey E, Bannasch DL, Barrey E, Bar-Gal GK, et al. Developing a 670k genotyping array to tag ~2M SNPs across 24 horse breeds. BMC Genomics. 2017; 18(1): 565. doi: 10.1186/s12864-017-3943-8.

10. McCue ME, Bannasch DL, Petersen JL, Gurr J, Bailey E, Binns MM, et al. A High Density SNP Array for the Domestic Horse and Extant Perissodactyla: Utility for Association Mapping, Genetic Diversity, and Phylogeny Studies. PLoS Genet. 2012; 8(1): e1002451. doi:10.1371/journal.pgen.1002451.

11. Kijas JW, Lenstra JA, Hayes B, Boitard S, Porto Neto LR, San Cristobal M, et al. Genome-wide analysis of the world's sheep breeds reveals high levels of historic mixture and strong recent selection. PLoS Biol. 2012; 10(2): e1001258. doi:10.1371/journal.pbio.1001258.

12. Zhan B, Fadista J, Thomsen B, Hedegaard J, Panitz F, Bendixen C. Global assessment of genomic variation in cattle by genome resequencing and high-throughput genotyping. BMC Genomics. 2011; 12(1): 557. doi: 10.1186/1471-2164-12-557.

13. Salomón-Torres R, González-Vizcarra VM, Medina-Basulto GE, Montaño-Gómez MF, Mahadevan P, Yaurima-Basaldúa VH, et al. Genome-wide identification of copy number variations in Holstein cattle from Baja California, Mexico, using high-density SNP genotyping arrays. Genet Mol Res. 2015; 14(4): 11848-11859. doi:10.4238/2015.

14. Valente TS, Baldi F, Sant'Anna AC, Albuquerque LG, Paranhos da Costa MJ. Genome-Wide Association Study between Single Nucleotide Polymorphisms and Flight Speed in Nellore Cattle. PLoS One. 2016; 14: 11(6): e0156956. doi:10.1371/journal.pone.0156956.

15. Das A, Panitz F, Gregersen VR, Bendixen C, Holm LE. Deep sequencing of Danish Holstein dairy cattle for variant detection and insight into potential loss-of-function variants in protein coding genes. BMC Genomics. 2015; 16: 1043. doi:10.1186/s12864-015-2249-y.

16. Joon-Ho L, Taeheon L, Hak-Kyo L, ByungWook C, Dong-Hyun S, Kyoung-Tag D, et al. Thoroughbred horse single nucleotide polymorphism and expression joodatabase: HSDB. Asian-Australs J. Anim. Sci. 2014; 27(9): 1236–1243. doi:10.5713/ajas.2013.13694

17. ABCCN. Associação Brasileira dos Criadores do Cavalo Nordestino. Regulamento do registro genealógico do cavalo Nordestino. ABCCN. Recife; 1987. pp.33-34.

18. Andrews S. FastQC: a quality control tool for high throughput sequence data, 2010. Available at http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/. (last accessed date December 14, 2018).

19. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25(14): 1754-1760. doi:10.1093/bioinformatics/btp324.

20. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010; 20(9): 1297-303. doi:10.1101/gr.107524.110

21. Cornish A, Guda C. A comparison of variant calling pipelines using genome in a bottle as a reference. Biomed. Res. Int. 2015. doi: 10.1155/2015/456479

22. Garrison E and Marth G. Haplotype-based variant detection from short-read sequencing. 2012. Preprint. Available from: arXiv:1207.3907. Cited 17 December 2018.

23. DePristo MA, Banks E, Poplin R, Garimella KV, Hartl C, Philippakis AA et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics. 2011; 43(5):491-501. doi:10.1038/ng.806.

24. Cingolani P, Platts A, le Wang L, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs

in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly. 2012; 6(2): 80-92. doi:10.1186/1471-2164-15-275.

25. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. Nat Protoc. 2009; 4(1): 44-57. doi:10.1038/nprot.2008.211.

26. Bos JL, Rehmann H, Wittinghofer A. GEFs and GAPs: critical elements in the control of small G proteins. Cell. 2007; 129(5): 865-77. doi:865-877. 10.1016/j.cell.2007.05.018.

27. Bos JL. Linking rap to cell adhesion. Current Opinion in Cell Biology. 2005; 17(2): 123-128. doi:10.1016/j.ceb.2005.02.009.

28. Trzaskowski B, Latek D, Yuan S, Ghoshdastider U, Debinski A, Filipek S. Action of molecular switches in GPCRs-theoretical and experimental. Current Medicinal Chemistry. 2012; 19 (8):109. doi:10.2174/092986712799320556.

29. Antunes G, Sebastião AM, Souza FM. Mechanisms of Regulation of Olfactory Transduction and Adaptation in the Olfactory Cilium. PLoS One. 2014; 9(8): e105531. doi:10.1371/journal.pone.0105531.

30. Ma M. Encoding olfactory signals via multiple chemosensory systems. Crit Rev Biochem Mol Biol. 2007; 42(6): 463-480. doi: 10.1080/10409230701693359

31. Kang N, Kim H, Jae Y, Lee N, Ku CR, Margolis F, et al. Olfactory Marker Protein Expression Is an Indicator of Olfactory Receptor-Associated Events in Non-Olfactory Tissues. PLoS One. 2015; 10(1): e0116097. doi: 10.1371/journal.pone.0116097.

32. Doucet S, Soussignan R, Sagot P, Schaal B. The secretion of areolar (Montgomery's) glands from lactating women elicits selective, unconditional responses in neonates. PLoS One. 2009; 4(10) e7579. doi:10.1371/journal.pone.0007579.

33. Kupke A, Wenisch S, Failing K and Herden C. Intranasal location and lmmunohistochemical characterization of the equine olfactory epithelium. Front. Neuroanat. 2016; 10:97. doi:10.3389/fnana.2016.00097.

34. Olender T, Lancet D, Nebert DW. Update on the olfactory receptor (OR) gene superfamily. Human Genomics. 2008; 3(1): 87-97. doi: 10.1186/1479-7364-3-1-87.

35. Stafuzza NB, Zerlotini A, Lobo FP, Yamagishi MEB, Chud TCS, Caetano AR, et al. Single nucleotide variants and InDels identified from whole-genome re-sequencing of Guzerat, Gyr, Girolando and Holstein cattle breeds. PLoS One. 2017; 12(3): e0173954. doi:10.1371/journal.pone.0173954.

36. Metzger J, Tonda R, Beltran S, Águeda L, Gut M, Distl O. Next generation sequencing gives an insight into the characteristics of highly selected breeds versus non-breed horses in the course of domestication. BMC Genomics. 2014; 15(1): 562-575. doi: 10.1186/1471-2164-15-562.

37. Jun J, Cho YS, Hu H, Kim H, Jho S, Gadhvi P, et al. Whole genome sequence and analysis of the Marwari horse breed and its genetic origin. BMC Genomics. 2014; 15(9): S4. doi:10.1186/1471-2164-15-S9-S4.

V CAPÍTULO 2

VALIDAÇÃO DE VARIANTES DE NUCLEOTÍDEO ÚNICO (SNVs) DE GENES DA FAMÍLIA GTPase (GIMAP) E OLFACTORY RECEPTOR (OR) EM POPULAÇÕES REMANESCENTES DA RAÇA EQUINA NORDESTINA

Cardoso, Danielle Cunha[1*]; Coelho, Eduardo Alves[1]; Silva, Glacy Jaqueline[2]; Pires, Denea de Araújo Fernandes[3]; Oliveira, Denise Aparecida Andrade[1].

[1] Laboratório de Genética da Escola de Veterinária. Universidade Federal de Minas Gerais. Belo Horizonte-MG. Brasil.
[2] Universidade Paranaense, Umuarama, Paraná, Brasil.
[3] Instituto Federal de Pernambuco, Barreiros, Pernambuco, Brasil.

*Corresponding author at: Danielle Cunha Cardoso, laboratório de genética da Escola de Veterinária. Universidade Federal de Minas Gerais. Belo Horizonte-MG. Brasil. 31270-901.
E-mail address: danielleccard@gmail.com

Resumo

Variantes de nucleotídeo único (SNVs) características da raça equina Nordestina, de impacto alto (afetando sítios de splicing e códons de início e término da transcrição) e impacto moderado (SNVs não sinônimas), presentes em genes da família de GTPases (GIMAP, membros 1 e 4 e 7) e genes de Receptores Olfativos (OR), identificados em trabalho prévio desse grupo de pesquisa, a partir de chamada de variantes em dados de sequenciamento de genoma completo, foram validadas em 60 indivíduos de três subpopulações remanescentes da raça, dos estados da Bahia, Pernambuco e Piauí. Esses genes atuam em vias de sinalização relacionadas à capacidade de percepção do meio e, consequentemente, a mecanismos de adaptação a novos ambientes e domesticação. O cavalo Nordestino apresenta características adaptativas exclusivas e de extrema importância para a adaptação nas condições inóspitas do semiárido do nordeste brasileiro, como baixa disponibilidade de forragem, altas temperaturas e insolação excessiva, pouca disponibilidade de chuvas e solos pedregosos. Tais condições não geram quaisquer prejuízos à perpetuação e desempenho da raça, que apresenta fenótipo específico de resistência física no padrão racial. Devido a essas particularidades, o cavalo Nordestino é considerado um patrimônio genético que deve ser conservado. Todos os dez *loci* contendo SNVs específicas da raça, validados no presente estudo em nível populacional, apresentaram alelos polimórficos nas três populações, com frequências genotípicas elevadas de indivíduos heterozigotos e homozigotos contendo polimorfismo alélico. As frequências genotípicas de oito *loci* analisados foram significativamente diferentes entre as três populações, exceto para dois SNVs associadas ao *locus* de GIMPAP 1(C/T) e de OR(T/G), sugerindo a fixação dos alelos polimórficos nas populações e diversidade genética. A presença de variantes nesses genes relacionados a mecanismos adaptativos representa importantes informações genéticas para compreender a excelente adaptação da raça ao semiárido do nordeste brasileiro e para associá-las ao fenótipo peculiar exibido pelo cavalo Nordestino.

Palavras-chave

Cavalo Nordestino, SNVs, validação populacional, GTPases, receptores olfatórios, alelos polimórficos, características adaptativas.

Abstract

Characteristic SNVs (single nucleotide variants) of the Nordestino horse breed of high impact (affecting slice-sites, stop and start codons) and moderate impact (non-synonymous SNVs), present in OR genes (Olfactory Receptor) and GTPases family genes (GIMAP, members 1, 4 and 7), identified in a previous study by our research group, from variant calling in whole genome sequencing data was validated in 60 horses of three remaining subpopulations of the breed, from the states of Bahia, Pernambuco and Piauí in the brazilian Northeastern. These genes act in the signaling pathways related to the perception capacity of the environment and, consequently, to the adaptation mechanism to new environments and domestication. The breed has odd adaptive characteristics extremely important for adaptation in semiarid conditions of Brazilian northeastern, such as low forage availability, high temperatures and excessive sunlight, rare rains and stony soils. Such conditions do not generate damages to the perpetuation and performance of the breed, which presents physical resistance phenotype, specific to the racial pattern. Considering these particularities, the Nordestino horse is considered an important genetic resource that must be conserved. All 10 *loci* that contain breed-specific SNVs, validated in the present study at the population level have polymorphic alleles in three populations, with high genotypic frequencies of heterozygotes and homozygotes carrying allelic polymorphism. Genotypic frequencies for 8 analyzed *loci* were different between three populations, except to SNVs in GIMPAP 1 (C / T) and OR (T / G) *locus*, suggesting allelic fixation in the populations and genetical diversity. The presence of variants in the genes related to adaptive mechanisms includes important genetic insights to understand the breed's excellent adaptation to the semiarid region of Brazilian northeastern and to associate the specific phenotype exhibited by Nordestino horse breed.

Keywords

Nordestino horse breed, SNVs, populational validation, GTPases, Olfactory Receptor, polymorphic alleles, adaptive characteristics.

I. Introdução

A raça equina Nordestina descende diretamente do cavalo Barbo-árabe, originário da península Ibérica, os quais foram introduzidos no Brasil com o advento da colonização. Desse modo, os tipos morfológicos são semelhantes, destacando-se o porte pequeno, inserção baixa da cauda, perfil retilíneo para subconvexo, ossatura forte, orelhas mal dirigidas, entre outras características, que permanecem sendo transmitidas aos descendentes. A raça exibe visíveis daptações singulares, que a permitiram sobreviver com sucesso sob as condições inóspitas do semiárido do nordeste brasileiro (Costa *et al*., 1974; Beck, 1985).

Entre as adversidades dessa região, destaca-se a elevada insolação ao longo de quase todo o ano, baixa nebulosidade, pouca disponibilidade de forragem, solos irregulares e pedrosos e chuvas escassas e irregulares. A sobrevivência da raça sob essas condições ocorre sem quaisquer perdas de performance nas atividades de força motriz ou lida no campo. Raças equinas melhoradas dificilmente apresentariam desempenho satisfatório diante dessas condições.

Os exemplares da raça equina Nordestina exibem resistência à doenças e rusticidade fenotípica no padrão racial, destacando-se narinas dilatadas, membros delgados, corpo parcialmente arqueado, conformação corporal média (até 145 cm para machos e 140 cm para fêmeas), os cascos rígidos com raras alterações patológicas, com ranilhas bem desenvolvidas e conformadas, as quais atuam como forte elemento amortecedor de impacto e auxiliar para a irrigação sanguínea para o interior do casco (Stashak, 1994; Travassos, 2004; Melo, 2011).

Devido às características adaptativas dessa raça considerada “naturalizada” brasileira, seus traços genéticos peculiares, bem como sua importância para a região nordeste do Brasil ao longo dos últimos séculos, o cavalo Nordestino é considerado um recurso genético valioso, que necessita ser preservado e estudado, especialmente devido à situação real de redução de populações desde as duas últimas décadas, quando núcleos de preservação e seleção da raça, bem como a Associação Brasileira dos Criadores do Cavalo Nordestino (ABCCN) vem sendo desativados (Mariante *et al*., 2011; Melo *et al*., 2013). As principais populações remanescentes encontram-se em regiões específicas dos estados da Bahia, Pernambuco, Piauí e Ceará (Nóbrega *et al*., 2010).

A maior parte dos estudos sobre a raça relaciona-se à caracterização fenotípica e morfométrica (Travassos, 2004; Melo *et al*., 2006; Pires *et al*., 2008; Melo *et al*., 2008; Nobrega *et al*., 2010; Melo, 2011; Melo *et al*., 2013). Entretanto, Pires *et al*. (2014) realizaram um estudo de diversidade genética da raça, em subpopulações de três regiões (as mesmas escolhidas para

o presente estudo) e observaram que as subpopulações apresentam elevada diversidade genética, mesmo diante das ameaças de declínio populacional. O baixo coeficiente de endogamia encontrado confirmou que as populações não passaram por gargalo genético recente, reforçando o potencial genético da raça para uso em programas de melhoramento genético e a necessidade de estudos subsequentes.

Com base nessas informações, em estudo prévio do grupo de pesquisa do laboratório de genética animal da Escola de Veterinária da Universidade Federal de Minas Gerais (LGEV-EV/UFMG), capítulo 1 da presente tese, realizou-se o sequenciamento completo do genoma de um animal macho, representante clássico da raça equina Nordestina, via plataforma Illumina™ e efetuou-se a chamada de variantes de nucleotídeo único (SNVs) e InDELs, em comparação com o genoma referência equino (Ensembl EquCab3.0), da raça Puro Sangue Inglês.

O mapeamento das *reads* e a chamada de variantes pelos softwares GATK (DePristo *et al*., 2011) e FreeBayes (Garrison e Marth, 2012) revelou ocorrência de SNVs de alto impacto em 31 genes, incluindo quatro famílias de GTPases (membros 1, 2, 4 e 7) e um pseudogene.

No mecanismo de hidrólise do GTP, as hidrolases GTPases atuam em processos biológicos essenciais, como na transdução de sinais, via receptores transmembrana; permitindo reconhecimento de gosto, cheiro, luminosidade e reconhecimento de patógenos, atuam na tradução, no transporte de vesículas celulares, no controle da divisão celular e na translocação de proteínas via membrana celular, sendo proteínas altamente conservadas (Bos *et al*.,2007; Bos, 2005).

No mesmo estudo, outros 392 genes contendo SNVs de impacto moderado foram identificados em várias famílias gênicas, sendo que maior atenção foi dada à família de genes Receptores Olfativos, os quais são expressos em neurônios desses receptores, em vertebrados, no epitélio das vias respiratórias e fazem a detecção de compostos que contém odor, por meio da ativação de interações moleculares complexas envolvendo proteínas tipo G, por meio da fosforilação da proteína adenilato ciclase, ativa vias de transdução de sinal, que culminam com a transmissão do olfato ao cérebro e o reconhecimento do cheiro (Trzaskowski *et al*., 2012; Antunes *et al*., 2014).

O impacto de variantes é de extrema importância para processos adaptativos, uma vez que o impacto alto afeta diretamente a síntese proteica, podendo causar perda de função da proteína, por quebra ou alteração da cadeia proteica. Para o presente estudo, SNVs de alto impacto, em três famílias de GTPases foram selecionados para validação populacional.

Por sua vez, o fato de eventos de sinalização bioquímica da superfamília de genes de Receptores Olfativos estarem relacionados diretamente com o sucesso adaptativo, por atuarem

no reconhecimento do meio, do alimento, identificação de parceiros sexuais e da cria, defesa contra patógenos, entre outros aspectos fundamentais da sobrevivência animal (Kang *et al*., 2015; Doucet *et al*., 2009), os SNVs presentes nessa superfamília gênica também foram selecionados para validação no presente estudo, nas três populações escolhidas de remanescentes da raça, visando associá-los à excelente adaptação desta, às condições inóspitas do semiárido Nordestino.

Análises genômicas comparativas, a fim de elucidar a origem genética de diversas raças equinas tem confirmado que a seleção natural, os processos de domesticação e adaptações à ambientes locais causam modificações fenotípicas por meio de mutações e fixação de alelos benéficos para determinada condição, sendo possível utilizar as ferramentas mais recentes da genômica para identificar pontualmente os *loci* que foram sujeitos à seleção (Zhang *et al*., 2018).

Variações de base única (SNVs), uma vez compreendidas como as variantes mais frequentes observadas no DNA, comumente envolvem bases nitrogenadas de mesma característica estrutural, ou seja, são trocas entre duas purinas (Adenina/Guanina ou G/A) ou duas pirimidinas (Citosina/Timina ou T/C) e são denominadas transições. As transversões são substituições de purina por pirimidina ou o contrário e, portanto, menos frequentes. Essas alterações podem ser provocadas por erros de incorporação de bases durante a replicação do DNA ou em outros casos, são causadas por agentes ambientais (Vignal *et al*., 2002).

Caso essas mutações ocorram em células germinativas e sejam transmitidas às gerações seguintes e se fixem na população a uma frequência mínima de 1%, passam a ser denominadas de SNPs (Single Nucleotide Polymorphisms). Esses polimorfismos segregam de acordo com as leis mendelianas para características monogênicas, ou apresentam distribuições compatíveis com as esperadas para características poligênicas (Ferreira e Grattapaglia, 1998) e geralmente ocorrem a cada 600 pares de base ao se comparar regiões correspondentes do mesmo genoma (Kwok e Gu, 1999; Kim e Misra, 2007).

Nesse sentido, a identificação de SNPs novos, em raças como a estudada no presente estudo é de suma importância para a compreensão dos seus mecanismos adaptativos e para uso da informação genética em programas de melhoramento animal, bem como em estratégias de conservação da raça.

Esse estudo teve como objetivo validar SNVs identificados em estudo prévio (capítulo 1 da presente tese) e característicos da raça equina Nordestina, como possíveis SNPs de impacto alto ou moderado, presentes em três famílias gênicas de GTPases (1, 4 e 7) e na família de genes de Receptores Olfativos (a qual apresentou a maior representatividade de SNVs na

análise funcional e de enriquecimento de vias metabólicas) em três populações remanescentes da raça, dos estados do Piauí, Bahia e Pernambuco, totalizando 60 animais e dez SNVs avalidados em 4 regiões gênicas.

II. Materiais e métodos

1. Declaração de ética e obtenção das amostras animais

Coletou-se pelos com bulbos de um total de 60 animais, sendo 38 pertencentes à criadores das mesorregiões norte e centro norte do estado do Piauí, 11 animais do sertão pernambucano e 11 animais da mesorregião Vale do São Francisco, do estado da Bahia. Todos os animais (machos ou fêmeas, proporcionalmente selecionados) tiveram amostras de pelo cedidas por seus criadores, por intermédio da professora Denea de Araújo Fernandes Pires (IFPE), sem que fosse necessário, nesse caso, uma aprovação ética específica (Lei brasileira 11.497/Outubro de 2008, Capítulo 1, Art. 3, Parágrafo III).

Todos esses animais foram considerados remanescentes da raça equina Nordestina, por possuírem características fenotípicas semelhantes ao antigo padrão da raça, como: porte e cabeça pequenos, pele de pigmentação escura, cascos fortes com ranilhas profundas, variação de altura entre 127 e 146 cm e eficiência no trabalho de campo no semiárido nordestino (ABCCN, 1987; Melo, 2011).

Uma vez que SNVs específicas da raça foram comparadas com a mais recente montagem genômica equina disponível (EquCab3.0 da base de dados Ensembl, raça PSI, Puro Sangue Inglês), a partir de dados genômicos em larga escala e alta cobertura, por tecnologias de sequenciamento de próxima geração, NGS (Next Generation Sequencing), não foi necessário utilizar *outgroup* de animais da raça PSI para geração de sequências para as análises comparativas.

2. Extração de DNA

Para a extração de DNA das amostras de pelos, foram usados oito bulbos de cada animal colocados em tubos de microcentrífuga e em seguida adicionou-se 1ul de proteinase K (5 mg/ml) e passo de incubação a 41^{0}C por 15 minutos. Após a incubação, adicionou-se 20 ul de solução de NaOH 0,8%, seguido de incubação à 97^{0}C por 15 minutos. Para neutralização,

adicionou-se 20 µl de solução Tris-HCl 1M. Nas etapas seguintes seguiram de centrifugação de 15 minutos a 14.000 rpm e lavagem do sobrenadante com 50ul de isopropanol absoluto, centrifugação à 14.000 rpm por 15 minutos e adição ao precipitado de 100 ul de solução de etanol à 70%. Após a centrifugação e secagem do DNA, o mesmo foi ressuspendido em água Milli-Q (Millipore, MA, EUA). As amostras de DNA foram quantificadas em espectrofotômetro Nanodrop (Thermo Fisher, CA, EUA).

3. Triagem de variantes e escolha do método para validação populacional de variantes de nucleotídeo único

Após seleção minuciosa dos SNVs identificados no estudo anterior, com base no impacto e na análise de enriquecimento de vias metabólicas (False Discovery rate (FDR)<0,10), o método de validação escolhido foi a amplificação das regiões do genoma, que contém as variantes e posteriormente sequenciar os *amplicons*, por sequenciamento capilar automático, para identificação pontual das SNVs nas sequências geradas.

A viabilidade de geração de sequências da população a ser avaliada contendo as variantes foi o fator crucial para a escolha do método de validação, diante da diversidade de metodologias atualmente disponíveis para genotipagem.

Apesar de a chamada de variantes do estudo prévio ter, teoricamente, identificado variantes específicas da raça equina Nordestina com relação à raça PSI, foi utilizada no presente estudo a ferramenta *SnpSift* (Cingolani *et al*., 2012), o qual é um *software* que contém um banco de dados de polimorfismos (SNPs e inserções e deleções/InDELs) de diversas espécies, a fim de certificar se as variantes selecionadas são, de fato, novas. Uma vez presentes no banco de dados de variantes de equinos, não podem ser consideradas novas e, como consequência, nem específicas do cavalo nordestino. As variantes selecionadas não estavam presentes no banco de dados *SnpSift.*

4. Desenho de *primers* específicos de fragmentos genômicos contendo SNVs e reação de PCR

Para cada SNV previamente selecionada, dois conjuntos de primers, *foward* e *reverse* foram desenhados sobre as regiões gênicas das famílias de GTPases (GIMAP), (membros 1, 4 e 7) e de ORs (membros 2 e 4), usando-se o *software* Primer-Blast, uma ferramenta disponibilizada pelo NCBI (National Center for Biotechnology Information), disponível em: https://www.ncbi.nlm.nih.gov/tools/primer-blast/, totalizando dez conjuntos de primers.

As sequências FASTA das regiões gênicas utilizadas e a localização pontual das SNVs foram obtidas de dois modos, a fim de comparação e certificação de identidade: diretamente nos dados gerados pelo sequenciamento completo e chamada de variantes no genoma do cavalo Nordestino, visualizando as variantes pela ferramenta IGV viewer (http://www.broadinstitute.org/igv) e diretamente no genoma referência, através da busca das sequências FASTA, por meio da posição dos SNVs no genoma referência (Ensembl EquCab3.0) (https://www.ensembl.org/Equus_caballus/Info/Index) (tabela 1).

As reações de amplificação por PCR foram conduzidas em volume final de 12,5 ul, incluindo 2,5 ul de água Milli-Q (Millipore, MA, EUA), 7,5 ul de Platinum™ PCR SuperMix High Fidelity (Invitrogen/ThermoFisher Scientific, CA, EUA), 0,4 ul de cada primer (10 uM), 1,7 ul de DNA molde (50 ng/ul). O protocolo de amplificação utilizado foi: 94ºC por 2 minutos, 35 ciclos de 94ºC por 30 s, 55ºC 30s e 72ºC por 1 minuto e um ciclo final de extensão de 72ºC por 12 minutos.

5. Sequenciamento das regiões genômicas contendo os SNVs

As sequências foram determinadas em ambas as direções (*Foward* e *Reverse*) para todos os *amplicons* de todas as amostras. Aquelas sequências com qualidade não recomendável foram descartas e repetidas até a obtenção de eletroferogramas contendo picos uniformes, em pelo menos, dois *amplicons* distintos para cada amostra, utilizando-se o kit BigDye Terminator v.3.1 Cycle Sequencing Kit (Thermo Fisher, CA, EUA) e os primers descritos na tabela 1.

As reações de sequenciamento e os controles internos da reação (pGEM Control DNA, Thermo Fisher, CA, EUA) foram realizados com volume final de 10ul por amostra, sendo aplicados 0,25 ul de BigDye® Terminator v3.1 Ready Reaction Mix, 2.35 ul de 5X Sequencing Buffer, 5,5 ul de água ultra pura (UltraPure™ DNase/RNase-Free Distilled Water, Thermo Fisher, CA, EUA), 1 ul de cada primer, *foward* ou *reverse* e 1ul de DNA amplificado (200ng/ul). As reações de sequenciamento em termociclador foram realizadas nas seguintes condições: 1 ciclo de 96 ºC por 1 minuto, 25 ciclos de 96 ºC por 10 s, 59 ºC por 5 s, 60 ºC por 4 minutos e 1 ciclo final de 60 ºC por 5 minutos.

Após a reação de sequenciamento, as amostras foram submetidas à precipitação com EDTA e etanol, adicionando-se 2,5 ul de EDTA (125mM) e 30 ul de etanol absoluto e incubação por 15 minutos a 25 ºC. As amostras foram centrifugadas por 45 minutos a 4400 rpm e, após o descarte do sobrenadante, adicionou-se 30 ul de solução de etanol 70%. Uma nova

centrifugação por 15 minutos a 2880 rpm foi realizada, seguida do descarte do sobrenadante. Após a completa evaporação do etanol, as amostras foram eluidas em 10 ul de formamida.

Em seguida, a eletroforese capilar foi feita em sequenciador automático de DNA ABI 3500 (Thermo Fisher, CA, EUA).

6. Análise das sequências geradas

As sequências geradas, senso e anti-senso (*Foward* e *Reverse*) foram analisadas quanto aos parâmetros de controle de qualidade e discriminação alélica da SNV pelo software Sequencing Analysis v7.0 (Thermo Fisher, CA, EUA). As sequências de todos os animais foram alinhadas, utilizando-se o software Molecular Evolutionary Genetics Analysis (MEGA X) (Kumar *et al*., 2018), com atenção à identificação pontual da presença ou ausência da SNV, feita para cada animal, de cada uma das três populações avaliadas (tabela 2).

7. Frequência populacional das SNVs e análise estatística

Os cálculos das frequências genotípicas dos *loci* contendo as dez variantes validadas nas três populações e a análise de associação das frequências entre as três populações de cavalos da raça Nordestina foram calculadas usando-se o teste Exato de Fisher, com 95% de intervalo de confiança, por meio da interface de estatística para genética do *software* RStudio (Ryman e Jorde, 2001; Ryman *et al*., 2006).

III. Resultados e Discussão

1. Sequências de primers gerados para amplificação de regiões genômicas contendo SNVs exclusivas e de alta relevância às características adaptativas exibidas pela raça equina Nordestina

A tabela 1 abaixo, apresenta os conjuntos de primers desenhados especificamente para ancorar os SNVs de impacto alto encontradas na família gênica de GTPAses, membros 1, 4 e 7 (GIMAP1, GIMAP4 e GIMAP7) e os SNPs de impacto moderado encontrados na família de genes OR. Conforme mencionado, OR l apresentou maior representatividade (5,37E-01 FDR, False Discovery Rate) na análise de enriquecimento funcional nos termos do Gene Onthology

(GO) e na análise de vias metabólicas da base de dados KEEG (Kyoto Encyclopedia of Genes and Genomes).

Tabela 1. Sequências de primers desenhados para sequenciamento das regiões genômicas contendo SNVs específicos da raça equina Nordestina, em genes de GTPAses (GIMAP), membros 7, 4 e 1 e em genes de Receptores Olfativos (OR). SNVs de genes GIMAP contém impacto alto e de genes OR contém impacto moderado.

Gene/Id Ensembl	Cromossomo	Primer 5`-3`	Posição da SNV no genoma	Tamanho do fragmento (Pb)	Ref/Alt	N^0 acesso ao gene (NCBI)
GIMAP7/id 196203	4	F: CCTGGTACTCCTCTGTGGGG	102.315.497	699	C/A	LOC100146699
		R: CGCCTCTGCTATAGCGTCAC				
		F: ACTCCTCTGTGGGGAAACTCT		542		
		R: AGAACCACGATGATGGCGTG				
GIMAP4/id 196221	4	F: CAG GAGGTGTGATGTGGCTT	102.355.651	952	G/A	LOC111773116
		R: GCAAGTCAGCTGGCTCAGA				
		F: ATGTGGCTTTGCCTTTCCCT		925		
		R: AGAGCACCTCCTGCAGCTA				
GIMAP1/id 196229		F: ACAGGCCAGCATGCAATTAAA	102.482. 546 / 102.482. 549 102.482. 551 / 102.482.487	868	C/T (102482546)	LOC100063777
		R: AGCGTGGCTTCTGATGAGTG			G/C (102482549)	
		F: GACAGGCCAGCATGCAATTA		937	G/C (102482551)	
		R: AACTTAACCGCTACGTCCCC			C/T (102482487)	
OR/Id 273421	6	F: TTGGTCTCTTGATCACCTGGC	67227304/67236887 67236895/67236898	581	A/T (67236887)	LOC100058068
		R: GGGATCAATCCATCCCCCAC			A/G (67227304)	
		F:TGGTCTCTTGATCACCTGGCTA		495	T/G (67236895)	
		R:TATGCTCTGTTCCAGCCTCC			G/A (67236895)	

Conforme observado, primers GIMAP1 e OR foram desenhados estrategicamente para ancorar mais de uma variante, em regiões genômicas onde genes estão total ou parcialmente sobrepostos. Nesse caso, SNVs geralmente tem impactos distintos na expressão de cada um desses genes, uma vez que promotores diferentes determinam o local de início da síntese de RNA. Nesses locais, em uma mesma região gênica, são gerados mais de um transcrito, que podem inclusive atuar na regulação da expressão gênica (Selbach *et al.*, 2008). Conforme representado na figura 1, observa-se a região de sobreposição gênica de anelamento dos primers GIMAP1, com base no genoma referência EquCab3 (Emsembl database), que contém quatro SNVs exclusivamente encontrados no genoma do Cavalo Nordestino.

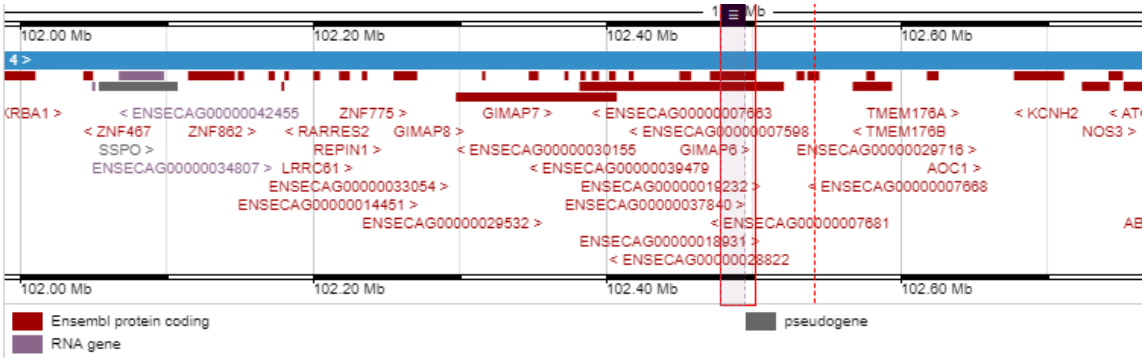

Figura 1. Região de sobreposição gênica, no cromossomo 4, localização 102.478.098 à 102.500.831 (marcada em vermelho), que contém as quatro variantes de nucleotídeo único exclusivas da raça equina Nordestina, onde ancoram ambas as sequências dos primers GIMAP1 (tabela 1).

https://www.ensembl.org/Equus_caballus/Location

2. Validação de variantes gênicas de nucleotídeo único, específicas da raça equina Nordestina em três subpopulações remanescentes

A genotipagem de variantes gênicas de nucleotídeo único, aplicando sequenciamento capilar automático de DNA também foi a estratégia utilizada por Pereira e colaboradores (2016) para avaliar o efeito de variantes em genes candidatos, sobre o fenótipo e comparação de frequências de SNPs entre grupos de equinos da raça Quarto de Milha.

A identificação da presença das SNVs selecionadas pela chamada de variantes (capítulo 1) com base na relevância do impacto (alto para genes de GTPases e moderado para genes OR) e na análise de enriquecimento funcional, nas três populações avaliadas confirma a robustez da chamada de variantes via três ferramentas, a partir de dados de NGS e a sensibilidade do método de validação escolhido no presente estudo.

A análise da diversidade genética entre indivíduos de uma mesma população e entre quatro populações de remanescentes da raça equina Nordestina, também dos estados do Piauí, Bahia e Pernambuco foi avaliada por Pires *et al*. (2014). Esses autores utilizaram marcadores microssatélites do painel atual da análise de paternidade equina, o qual revelou a existência de elevada variação alélica entre populações remanescentes (heterozigosidade esperada elevada) e não detecção de efeito “gargalo” recente para as populações.

Embora não seja objetivo do presente estudo realizar análise de diversidade genética e sim, identificar SNVs, como possíveis SNPs associados ao fenótipo peculiar exibido pela raça, estudar populações geneticamente diversas, possibilita comparar frequências genotípicas e alélicas entre as mesmas e implicações mais robustas de associação da variante com o fenótipo exibido.

Através da análise criteriosa dos resultados de alinhamento das sequências geradas por sequenciamento automático de DNA para realização da genotipagem, foi possível identificar a presença de pelo menos um alelo polimórfico de todas as dez variantes de nucleotídeo único em animais pertencentes às três populações remanescentes avaliadas (estados de Piauí, Bahia e Pernambuco).

Um número maior de animais homozigotos para os alelos mutantes ocorreu na população remanescente do estado da Bahia, em comparação com as outras duas populações (tabelas 2 e 3), sendo um possível indicativo de ocorrencia de maior fixação alélica dos polimorfismos nessa população. É importante considerar que, embora isso represente uma menor diversidade genética que as outras populações avaliadas, a fixação alélica observada representa uma contribuição mais profunda na perpetuação das características de resistência exibidas pela raça.

A tabela 2 apresenta a genotipagem de 60 animais das três subpopulações remanescentes, para as dez SNVs selecionadas e a tabela 3, por sua vez, apresenta as frequências genotípicas das SNVs avaliadas. Na coluna relativa aos genótipos (tabela 3), 0 representa ausência do alelo mutante, 1 representa presença de um alelo mutante (indivíduos heterozigotos) e 2 representa presença dos 2 alelos mutantes no referente *locus* da variante.

Além das frequências genotípicas (tabela 3), também foram obtidas as frequências alélicas das SNVs dos *Loci* avaliados (tabela 4), reforçando tanto a observação de maior frequência dos alelos alterados na subpopulação do estado da Bahia, em comparação com as outras duas, quanto a ideia de fixação destes nas três populações.

Tabela 2. Genótipo de animais provenientes das subpopulações remanescentes da raça equina Nordestina, dos estados da Bahia, Pernambuco e Piauí, para as dez variantes de nucleotídeo único (SNVs) nos genes de GTPAses (GIMAP), membros 7, 4 e 1 e Receptor Olfativo (OR).

*** pop: população.**

 # Ref/Alt: nucleotídeo referência/nucleotídeo alterado.

(BA) estado da Bahia, (PE) estado de Pernambuco, (PI) estado do Piauí

POP.*	ANIMAL	GIMAP7 Ref/Alt# (C/A)	GIMAP4 (G/A)	GIMAP1 (C/T)	GIMAP1 (G/C)	GIMAP1 (G/C)	GIMAP1 (C/T)	OR (A/T)	OR (A/G)	OR (T/G)	OR (G/A)
BA	1	AA	AA	CT	CC	CC	TT	TT	AG	TG	AA
	2	CA	AA	CT	GG	CC	TT	TT	AG	TG	AA
	3	CA	AA	TT	GG	GC	TT	TT	GG	GG	AA
	4	CC	GA	CT	GC	GC	CT	AT	GG	GG	GA
	5	AA	GG	TT	CC	CC	CT	TT	GG	GG	AA
	6	AA	GG	TT	GC	CC	CC	AT	GG	TG	AA
	7	CA	AA	CC	GC	GG	CC	AT	GG	TT	GA
	8	AA	AA	CT	CC	GG	CC	AT	AG	TG	AA
	9	CA	AA	TT	CC	GG	CC	TT	AA	TG	GA
	10	AA	AA	TT	CC	CC	TT	TT	AA	TT	AA
	11	AA	GA	TT	GC	CC	TT	AT	AG	TT	AA
PE	12	CC	GA	CT	GC	GC	CT	TT	AA	TG	GA
	13	CC	GA	CT	GC	GC	CC	TT	AG	TT	GA
	14	CC	GA	CC	CC	GC	CT	TT	AG	TG	GA
	15	CC	GG	CT	GC	GC	CT	TT	AA	TG	GA
	16	CA	AA	CT	GC	GC	CC	TT	AG	TG	GA
	17	CC	GA	CC	GC	CC	CC	AT	AG	TT	AA
	18	CA	AA	CC	CC	CC	TT	TT	GG	TT	GA
	19	CA	GA	CC	CC	GC	CT	AT	AG	TG	GA
	20	CA	GA	CC	GC	GC	CT	AT	AG	TG	GA
	21	CA	GA	CT	GG	GG	CT	AT	AG	GG	AA
	22	CA	GG	CT	GG	GG	CT	AT	GG	TG	GA
PI	23	CC	AA	CC	GG	GC	TT	AA	AA	TG	GA
	24	CA	AA	CT	GG	GG	CC	AA	AA	TG	GA
	25	CA	GA	CT	GG	CC	CT	AT	AA	GG	GA
	26	CA	AA	CC	GG	GC	CT	AA	AA	GG	GA
	27	AA	GA	CT	GC	GC	CC	AT	AA	TG	GA
	28	CA	GA	TT	GC	GC	CT	AT	AA	GG	GA
	29	CA	GA	CT	GC	CC	TT	AT	AG	GG	GA
	30	CC	GA	TT	GC	CC	TT	TT	AG	TT	GA
	31	CC	GA	TT	GC	GG	TT	AT	AG	TT	AA
	32	CC	GA	CC	GG	GC	CC	TT	AA	TT	AA
	33	CA	GG	CT	CC	GC	CT	TT	AA	TT	GA
	34	AA	GG	TT	GG	GG	CC	AT	GG	TT	GA
	35	CA	GA	CT	GC	GG	CC	AT	AG	GG	GA
	36	AA	GG	CC	GG	GG	CC	AA	AA	TG	GG
	37	CC	GG	CC	CC	GG	CC	AA	AA	TT	GA

38	CC	AA	CC	GG	GC	CC	AT	AA	TG	GG
39	CA	GA	CT	GC	GG	CT	AT	AG	TG	GG
40	CA	GA	CT	GG	GC	CT	AT	GG	TG	GG
41	CA	GA	CC	GC	GG	CT	AT	AA	TT	AA
42	CC	GA	CT	GC	GG	TT	TT	AA	GG	GG
43	AA	GG	CC	GC	GC	CC	AA	AA	TT	GG
44	CC	GG	CT	GG	GC	CC	AA	AA	TT	GA
45	CA	GA	CT	GC	GG	CC	AT	AG	TG	AA
46	CC	GA	TT	GG	GG	CT	AT	AA	TG	GA
47	CC	GG	CT	GC	GG	CC	TT	AG	TG	GG
48	CC	GA	CT	GC	GC	CT	AA	AG	TT	GG
49	CA	GG	CT	GC	CC	CT	AA	AA	TT	GG
50	CA	GG	CC	GG	GC	CC	AT	AG	TT	GA
51	CC	GG	CC	GG	GG	CC	AA	GG	TT	GG
52	CC	AA	TT	CC	CC	CC	AT	AG	TT	GA
53	CC	AA	CT	GC	GC	CC	AT	AA	TT	GA
54	CA	GA	CT	CC	GC	TT	TT	AG	TG	GA
55	CA	GG	CT	GG	GG	CT	TT	AG	TG	GG
56	CA	GG	CC	GG	GG	CC	AT	AG	TT	GA
57	CC	GA	CC	GG	CC	CC	AT	GG	TT	GA
58	CA	GG	CC	GG	GG	CT	AA	AA	TT	AA
59	CA	GG	TT	GG	GG	CT	AA	AA	TG	GG
60	CA	GA	CC	GG	GG	CC	AT	AG	TT	GG

A análise estatística de Fisher confirma que as frequências genotípicas dos *loci* analisados, contendo dez SNVs específicas da raça são significativamente diferentes entre as três populações, exceto para dois SNVs associadas ao *locus* de GIMPAP 1 (C/T) e de genes OR (T/G) (tabela 3). Isso sugere que a fixação da maioria dos alelos polimórficos ocorreu nas três populações e de forma diversificada. Observa-se portanto, que as populações mantém certo grau de diversidade genética, uma vez que esta corresponde não apenas à variedade de alelos presentes em um grupo, mas também à heterozigosidade observada nas populações (Glowatzki-mullis *et al*., 2005).

É importante ressaltar que as duas variantes para as quais a frequência genotípica não variou significativamente entre as populações são justamente as SNVs localizadas em regiões de genes sobrepostos (tabela 1 e figura 1), cujo impacto é variável de um gene para outro e não necessariamente atuam sobre a sequência do gene em que houve a identificação, no caso, GIMPA1 e OR. Nesses casos, deve-se considerar os mecanismos de regulação gênica, especialmente os envolvidos na remodelagem da cromatina. Desse modo, é possível que não tenham se fixado na população e consequentemente, não tenham, de fato, efeito associado ao fenótipo peculiar exibido pela raça.

De acordo com os resultados observados, as frequências de indivíduos heterozigotos para a maioria dos *loci* contendo as variantes foi elevada nas três populações. Uma vez que o déficit de heterozigoto é indício de acasalamento frequente entre indivíduos aparentados (Luikart *et al*., 2006), o que é comum em populações com número de animais reduzido, as frequências genotípicas observadas indicam que os criadores da raça têm essa preocupação com o planejamento dos cruzamentos. Conforme mencionado, a população do estado da Bahia, particularmente, apresentou maior frequência de homozigose para o alelo polimórfico, na maioria dos *loci* avaliados, que as outras duas subpopulações, possivelmente devido à uma maior taxa de cruzamentos entre indivíduos aparentados. Segundo Pires *et al*. (2014), embora a raça mantenha parâmetros indicativos de diversidade genética, que inferem indícios consistentes de que o cavalo Nordestino não tenha sido submetido ao efeito gargalo genético recente, o fechamento da Associação Brasileira dos Criadores do Cavalo Nordestino (ABCCN) pode ter contribuído para esses cruzamentos consanguíneos.

Tabela 3. Comparação das frequências genotípicas de dez variantes de nucleotídeo único (SNVs), específicas da raça equina Nordestina, em genes de GTPAses e Receptor Olfativo em três populações remanescentes dos estados da Bahia, Pernambuco e Piauí.

SNV *locus*/ (ref/alt)	Genótipo[#]	Grupos (n)			*P Value**
		BAHIA (11)	PERNAMBUCO (11)	PIAUÍ (38)	
		Frequência genotípica	Frequência genotípica	Frequência genotípica	
GIMAP7	0	0.0900	0.4545	0.3947	0.01089
(C/A)	1	0.3636	0.5454	0.5000	
	2	0.5454	0	0.1052	
GIMAP4	0	0.1818	0.1818	0.3684	0.02563
(G/A)	1	0.1818	0.6363	0.4736	
	2	0.6363	0.1818	0.1578	
GIMAP1	0	0.0900	0.4545	0.3684	0.03151
(C/T)	1	0.3636	0.5454	0.4473	
	2	0.5454	0	0.1842	
GIMAP1	0	0.1818	0.1818	0.5000	0.04306
(G/C)	1	0.3636	0.5454	0.3947	
	2	0.4545	0.2727	0.1052	
GIMAP1	0	0.2727	0.1818	0.4736	0.04127
(G/C)	1	0.1818	0.6363	0.3684	
	2	0.5454	0.2727	0.1578	
GIMAP1	0	0.3636	0.2727	0.4473	0.0928
(C/T)	1	0.1818	0.6363	0.3421	
	2	0.4545	0.0900	0.1578	
OR	0	0	0	0.3157	0.009087
(A/T)	1	0.4545	0.4545	0.5000	
	2	0.5454	0.5454	0.1842	
OR	0	0.1818	0.1818	0.5263	0.02617
(A/G)	1	0.3636	0.6363	0.3684	
	2	0.4545	0.1818	0.1052	
OR	0	0.2727	0.2727	0.5000	0.3305
(T/G)	1	0.4545	0.6363	0.3421	
	2	0.2727	0.0900	0.1578	
0R	0	0	0	0.3421	0.0002084
(G/A)	1	0.2727	0.8181	0.5263	
	2	0.7272	0.1818	0.1315	

Abreviações: SNV *Locus* (*Locus* da variante de nucleotídeo único); GIMAP7, GIMAP4, GIMAP1 (gene GTPase, membros 7, 4,1), OR (gene Receptor Olfativo); Ref/alt (nucleotídeo referência/nucleotídeo alterado).

genótipo: 0 (Homozigoto com ausência do alelo mutante), 1 (Heterozigoto, um alelo mutante), 2 (Homozigoto com ambos alelos mutantes).

* P value, significância do teste exato de Fisher (Fisher's Exact Test), 95% de intervalo de confiança, análise por variante, comparação das três populações.

[a] Frequência genotípica em cada população de equinos da raça Nordestina, provenientes dos estados da Bahia, Pernambuco e Piauí.

Tabela 4. Frequências alélicas de dez variantes de nucleotídeo único (SNVs), específicas da raça equina Nordestina, em genes de GTPAses e Receptor Olfativo em três populações remanescentes dos estados da Bahia, Pernambuco e Piauí.

		POPULAÇÕES (n)		
		BA (11)	PE (11)	PI (38)
Locus	Alelo	Frequências alélicas*		
GIMAP7	C	0.273	0.727	0.645
	A	0.727	0.273	0.355
GIMAP4	G	0.273	0.500	0.605
	A	0.737	0.500	0.395
GIMAP1	C	0.273	0.727	0.592
	T	0.727	0.273	0.408
GIMAP1	G	0.364	0.455	0.697
	C	0.636	0.545	0.303
GIMAP1	G	0.364	0.500	0.658
	C	0.636	0.500	0.342
GIMAP1	C	0.455	0.591	0.671
	T	0.545	0.409	0.329
OR	A	0.227	0.227	0.566
	T	0.773	0.773	0.434
OR	A	0.364	0.500	0.711
	G	0.636	0.500	0.289
OR	T	0.500	0.591	0.671
	G	0.500	0.409	0.329
OR	G	0.136	0.409	0.605
	A	0.864	0.591	0.395

Abreviações: SNV *Locus* (*Locus* da variante de nucleotídeo único); GIMAP7, GIMAP4, GIMAP1 (gene GTPase, membros 7, 4,1), OR (gene Receptor Olfativo).

* Frequências alélicas (alelo referância/ alterado) de cada l*ocus* avaliado, nas população de equinos da raça Nordestina, provenientes dos estados da Bahia, Pernambuco e Piauí.

Uma vez que os resultados desse estudo confirmaram a presença de alelos polimórficos das variantes nas três populações avaliadas e provável fixação populacional destes, parece consistente a ideia de que essas variantes de impacto alto em GTPses, afetando sítios de *splicing,* ou códons de início e fim da transcrição, tenham efetiva contribuição sobre as notáveis características adaptativas desenvolvidas pela raça. Essa família gênica de proteínas hidrolases encontram-se ativas quando ligadas ao GTP (trifosfato de guanosina) atuando na ativação de vias de sinalização e como reguladores essenciais da maioria dos processos celulares, como diferenciação, proliferação celular e transporte de vesículas e organelas (Bos *et al*., 2007; Cherfils e Zeghouf, 2013). Portanto, embora GTPases sejam genes envolvidos no metabolismo celular basal, em processos adaptativos, tais como os sofridos pelo cavalo Nordestino, esses genes moduladores de vias de sinalização apresentam papel crucial.

A busca de variantes exclusivas de raças equinas realizada por Zhang *et al*. (2018), visando compreender mecanismos moleculares associados à domesticação de duas raças nativas chinesas (Kazakh e Lichuan) também identificou variantes de nucleotídeo único de alto impacto em genes ativadores de GTPases, família gênica altamente representada na análise de enriquecimento funcional, resultados que confirmam o quão importantes são as vias metabólicas com envolvimento de GTPases para os processos de domesticação e mecanismos adaptativos.

Níveis elevados de radiação ultravioleta, temperaturas altas e baixa humidade são fatores que geram alta pressão adaptativa e a elucidação de mecanismos genéticos envolvidos na modulação de vias bioquímicas é fundamental para compreender a adaptação dos mamíferos a ambientes extremos (Hendrickson, 2013), condições estas presentes no semiárido nordestino, com as quais o cavalo Nordestino convive sem quaisquer prejuízos à raça.

Hendrickson (2013) realizou estudo genômico de busca de SNPs usando array equino de ~50k, com foco em 130 genes candidatos para estudar a adaptação de equinos introduzidos na região dos Andes por espanhóis no século XVI. A análise funcional revelou presença de SNPs, principalmente em famílias de genes associados ao sistema nervoso, envolvidos na percepção sensorial do olfato, como sete membros da família de genes OR, corroborando com o estudo prévio do nosso grupo de pesquisa, cuja família de genes OR teve a maior representatividade na análise funcional de genes que continham variantes de impacto alto e moderado no cavalo nordestino. Esses resultados confirmam a importância das variantes em genes OR para o sucesso adaptativo, especialmente de equinos, em ambientes inóspitos.

Diversos estudos em outros mamíferos também associaram genes OR, à partir de dados de NGS, à domesticação e à adaptação a novos ambientes, como em gatos domésticos (Montague *et al*., 2014), raças bovinas zebuínas (Stafuzza *et al*., 2017), cão e morcego (Hughes *et al*., 2018), entre outros; evidenciando que essa ampla família gênica tem sofrido modificações estruturais associadas à adaptações dos mamíferos a novos nichos ecológicos, apresentando papel crucial nos mecanismos adaptativos.

Jun *et al*. (2014) identificou SNVs e InDELs em genes OR, com efeito de impacto alto e moderado sobre a expressão desses genes, em uma raça equina indiana resistente ao clima desértico, originária principalmente de cavalos árabes (cavalo Marwari).

As vias metabólicas nas quais genes OR estão envolvidos atuam em diversos aspectos da sobrevivência dos mamíferos, entre eles no reconhecimento do alimento, de parceiros sexuais, de ameaças, identificação de feromônios e reconhecimento da prole (Kang *et al*., 2015; Kupke *et al*., 2016). A presença de variantes em genes relacionados a esses mecanismos

adaptativos nas três populações remanescentes do cavalo Nordestino, com frequências alélicas e genotípicas elevadas, indicam que esses alelos polimórficos, que apresentam impacto relevante sobre a expressão dos genes onde se localizam, encontram-se fixados nas populações e constituem mecanismos genéticos essenciais para compreender a excelente adaptação da raça ao semiárido do nordeste brasileiro e para associa-las ao fenótipo peculiar exibido por essa raça.

IV. Conclusão

Este estudo apresenta dados genéticos efetivos, à partir da validação populacional de variantes exclusivas da raça equina Nordestina, previamente identificadas por meio do primeiro dado genômico, em larga escala dessa raça. A comprovada presença e possível fixação nas populações remanescentes, devido às frequências elevadas dos alelos polimórficos das variantes de impacto alto à moderado nos *loci* avaliados, e significativamente distintas entre as populações, representam dados genéticos consistentes, que ajudam a explicar a excelente adaptação do cavalo Nordestino ao semiárido brasileiro e o fenótipo exclusivo exibido pela raça.

GTPases, na função de moduladores de vias bioquímicas do metabolismo celular e genes OR, diretamente relacionados às funções olfativas, através da percepção sensorial do olfato parecem estar envolvidos na excelente adaptação exibida pela raça às condições inóspitas do nordeste brasileiro. Portanto, este estudo fornece a base para classificar pelo menos oito das dez SNVs estudadas, como SNPs característicos e associados ao fenótipo do cavalo Nordestino, fornecendo dados genéticos para ampliação de estudos populacionais, por meio, inclusive, de pesquisas de associação de genômica ampla (GWAS, *Genome-Wide Association Study*). Os "*insights*" genômicos apresentados pelo presente trabalho podem ajudar na conservação da raça e no desenvolvimento de marcadores moleculares para resistência equina à condições inóspitas, os quais podem ser introduzidos em programas de melhoramento genético equino.

V. Referências Bibliográficas

ABCCN. Associação Brasileira dos Criadores do Cavalo Nordestino. Regulamento do registro genealógico do cavalo Nordestino. ABCCN. Recife; 1987. pp.33-34

ANTUNES G, SEBASTIÃO AM, SIMOES DE SOUZA FM. Mechanisms of Regulation of Olfactory Transduction and Adaptation in the Olfactory Cilium. PLoS One. 2014; 9(8): e105531. https://doi.org/10.1371/journal.pone.0105531.

BECK, SL. Pantaneiro Nordestino e Marajoara, raças brasileiras pouco conhecidas. In: Equinos: raças, manejo e equitação. São Paulo: Criadores, 1985. 179-190p.

BOS JL, REHMANN H, WITTINGHOFER A. GEFs and GAPs: critical elements in the control of small G proteins. Cell. 2007; 129, 865-877.

Bos JL. Linking rap to cell adhesion. Current Opinion in Cell Biology. 2005; 17(2): 123-128 https://doi.org/10.1016/j.ceb.2005.02.009.

CHERFILS J AND ZEGHOUF M. Regulation of small GTPases by GEFs, GAPs, and GDIs. Physiol. Rev. 2013; 93(1):269-309. [PMID: 23303910].

CINGOLANI PI, PATEL VM, COON M, NGUYEN T, LAND SJ, RUDEN DM, LU X. Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. Front Genet. 2012; 15(3). doi: 10.3389/fgene.2012.00035.

COSTA N, VAL LJ, LEITE GU. Estudo da preservação do cavalo Nordestino. Recife: Departamento de Produção Animal, 1974. 38p.

DEPRISTO MA, BANKS E, POPLIN R, GARIMELLA KV, HARTL C, PHILIPPAKIS AA et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics. 2011; 43(5):491-501. doi: 10.1038/ng.806.

DOUCET S, SOUSSIGNAN R, SAGOT P, SCHAAL B. The secretion of areolar (Montgomery's) glands from lactating women elicits selective, unconditional responses in neonates. PLoS One. 2009; 4: e7579 doi: 10.1371/journal.pone.0007579.

FERREIRA, M. E.; GRATTAPALIA, D. Introdução ao uso de marcadores moleculares em análise genética. Embrapa Cenargen: Brasília, 1998. p. 15-67.

GARRISON E AND MARTH G. Haplotype-based variant detection from short-read sequencing. http://arxiv.org/abs/1207.3907. (on line).

GLOWATZKI-MULLIS ML, MUNTWYLER J, PFISTER W, MARTI E, RIEDER S, PONCET PA, GAILLARD C. Genetic diversity among horse populations with a special focus on the Franches-Montagnes breed. Animal Genetics. 2006; 37(1). doi:10.1111/j.1365-2052.2005.01376.x.

HENDRICKSON S L. A genome wide study of genetic adaptation to high altitude in feral Andean Horses of the paramo. BMC Evolutionary Biology. 2013; 13(273).

HUGHES MG, BOSTON ESM, FINARELLI JA, MURPHY WJ, HIGGINS DG, TEELING EC. The Birth and Death of Olfactory Receptor Gene Families in Mammalian Niche Adaptation. Molecular Biology and Evolution. 2018; 35(6). doi: 10.1093/molbev/msy028.

JUN J, CHO YS, HU H, KIM H, JHO S, GADHVI P, et al. Whole genome sequence and analysis of the Marwari horse breed and its genetic origin. BMC Genomics. 2014; 15(9): S4. doi:10.1186/1471-2164-15-S9-S4.

KANG N, KIM H, JAE Y, LEE N, KU CR, MARGOLIS F, et al. Olfactory Marker Protein Expression Is an Indicator of Olfactory Receptor-Associated Events in Non-Olfactory Tissues. PLoSOne. 2015; 10(1): e0116097. doi: 10.1371/journal.pone.0116097.

KIM, S.; MISRA, A. SNP Genotyping: Technologies and Biomedical Applications. Annu.Rev. Biomed. 2007; 9(28).

KUMAR S, STECHER G, LI M, KNYAZ C, TAMURA K. 2018. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Mol. Biol. Evol. 35:1547–1549.

KUPKE A, WENISCH S, FAILING K AND HERDEN C. Intranasal location and lmmunohistochemical characterization of the equine olfactory epithelium. Front. Neuroanat. 2016; 10:97. doi: 10.3389/fnana.2016.00097.

KWOK, P. Y.; GU, Z. Single nucleotide polymorphism libraries: why and how are we building them? Molecular Medicine Today. 1999; 5(2).

LUIKART G, ALLENDORF FW, CORNUET JM AND SHERWIN WB. Distortion of allele frequency distributions provides a test for recent population bottlenecks. J Hered. 2006; 6(89).

MARIANTE AS, ALBUQUERQUE MSM, RAMOS AF. Criopreservação de recursos genéticos animais brasileiros. Rev. Bras. Reprod. Anim. 2011; 35(2).

MELO U, FERREIRA C, SANTIAGO RM, PALHARES M, MARANHÃO R. Equilíbrio do casco equino – uma revisão. Ciência Animal Brasileira. 2006;7(4)389-398.

MELO JB, RIBEIRO MN, ANDRADE JÚNIOR AM, PIRES DAF, OLIVEIRA JVC, ROCHA LL. Estudo morfométrico de éguas adultas do cavalo Nordestino no município de Altinho, Pernambuco, Brasil. In: REUNIÃO ANUAL DA SOCIEDADE BRASILEIRA DE ZOOTECNIA, 45., 2008, Lavras. Anais… Lavras: UFLA, 2008.

MELO JB. Caracterização zoométrica do remanescente da raça equina Nordestina nos estados de Pernambuco e Piauí. 2011. 118f. Tese (Doutorado em Zootecnia) - Universidade Federal Rural de Pernambuco. Departamento de Zootecnia, Recife.

MELO JB, PIRES DAF, RIBEIRO MN. Perfil fenotípico do remanescente do cavalo Nordestino no nordeste do Brasil. Archivos de Zootecnia. 2013; 62(6)171-180.

MONTAGUE MJ, LI G, GANDOLFI B, KHAN R, AKEN BL, SEARLE SMJ et al. Comparative analysis of the domestic cat genome reveals genetic signatures underlying feline biology and domestication. PNAS. 2014, 11:48. doi: 10.1073/pnas.1410083111.

NÓBREGA SMD, FILHO ECP, CRUZ GRB, ALMEIDA MJO, COSTA TP, MOURA RS. Caracterização morfológica de equinos da raça cavalo Nordestino criados na grande região de Campo Maior – Piauí: medidas lineares. In: REUNIÃO ANUAL DA SOCIEDADE BRASILEIRA DE ZOOTECNIA, 47., 2010, Salvador. Anais… Salvador: SBZ, 2010.

PEREIRA, GL, MATTEIS R, REGITANO LCA, CHARDULO LAL, CURI RA. MSTN, CKM, and DMRT3 Gene Variants in Different Lines of Quarter Horses. Journal of Equine Veterinary Science. 2016 39(4) 33-37.

PIRES DAF, COELHO EGA, MELO JB, OLIVEIRA DAA; RIBEIRO MN, GUS COTHRAN E et al. Genetic diversity and population structure in remnant subpopulations of Nordestino horse breed. Arch. Zootec. 2014; 63(242) doi: 10.4321/S0004-05922014000200013.

PIRES DAF, COELHO EGA, MELO JB, OLIVEIRA DAA; RIBEIRO MN, GUS COTHRAN E, JURAS R. Genetic relationship between the Nordestino horse and national and international horse breeds. Genetics and Molecular Research. 2016; 15(2). doi: 10.4238/gmr.15027881.

RYMAN N and JORDE PE. Statistical power when testing for genetic differentiation. Molecular Ecology. 2001; 10 (23).

RYMAN N, PALM S, ANDRÉ C, CARVALHO GR, DAHLGREN TG, JORDE PE et al. Power for detecting genetic divergence: differences between statistical methods and marker loci. Molecular Ecology. 2006; 15(31). doi: 10.1111/j.1365-294X.2006.02839.x.

STAFUZZA NB, ZERLOTINI A, LOBO FP, YAMAGISHI MEB, CHUD TCS, CAETANO AR et al. Single nucleotide variants and InDels identified from whole-genome re-sequencing of Guzerat, Gyr, Girolando and Holstein cattle breeds. PLoS One. 2017; 12(3): e0173954. https://doi.org/10.1371/journal.pone.0173954.

SELBACH, M et al. Widespread changes in protein synthesis induced by microrna s. Nature, 455, p. 58-63, 2008.

STASHAK, T R. Claudicação em equinos segundo Adam´s, 4. ed. São Paulo: Roca, 1994. 923p.

TRAVASSOS, AEV. Caracterização fenotípica do cavalo Nordestino no estado de Pernambuco. 2004. 59f. Dissertação (Mestrado em Zootecnia) - Universidade Federal Rural de Pernambuco, Recife.

TRZASKOWSKI B, LATEK D, YUAN S, GHOSHDASTIDER U, DEBINSKI A, FILIPEK S. Action of molecular switches in GPCRs-theoretical and experimental. CurrentMedicinalChemistry.2012; 19 (8):1090109. doi:10.2174/092986712799320556.

VIGNAL A, MILAN D, SANCRISTOBAL M, EGGEN A. A reviw on snp and other types of molecular markers and their use in animal genetics. GENETICS SELECT EVOLUTION. 2002; 34(27).

ZHANG C, NI P, AHMAD HI, GEMINGGULI M, BAIZILAITIBEI A, GULIBAHETI D et al. Detecting the Population Structure and Scanning for Signatures of Selection in Horses (Equus caballus) from Whole-Genome Sequencing Data. Evol Bioinform. 2018; 4:14. doi: 10.1177/1176934318775106.

VI CONSIDERAÇÕES FINAIS

A adaptação animal a novos ambientes, com ocupação de novos nichos ecológicos é extremamente complexa e tem sido motivo de estudos envolvendo diversas áreas do conhecimento biológico. Sabe-se que as bases genéticas dessa adaptação envolvem a participação de múltiplos genes, vias bioquímicas, mecanismos de evolução/alteração genômica em termos estruturais e funcionais, sendo estas por meio da regulação da expressão gênica e da síntese proteica, em seus diversos níveis. Com o desenvolvimento das tecnologias de geração de dados genômicos em larga escala, a ciência tem obtido informações valiosas e há poucos anos, inimagináveis. Especificamente, o desenvolvimento das tecnologias de sequenciamento de próxima geração tem permitido estudar profundamente o genoma e identificar mecanismos genéticos e genes envolvidos nos processos adaptativos.

Particularmente no caso de equinos, a domesticação e posteriormente as adaptações ocorridas após a introdução de raças em outros continentes, com o advento da colonização, muitas raças sofreram pressões seletivas, perpetuaram, e obtiveram sucesso adaptativo. Dessa forma, apresentam padrão racial distinto das raças que as originaram, como é o caso do cavalo Nordestino, que exibe padrão racial que lhe é peculiar, permitindo-o sobreviver e desenvolver atividades que exigem força e resistência no semiárido do Nordeste brasileiro. Nesse ambiente, sobrevivem naturalmente, diante de altas temperaturas, insolação excessiva, escassez de forragem, regime pluvial irregular e escasso, entre outras caraterísticas que seriam inviáveis para sobrevivência de outras raças geneticamente melhoradas.

O presente trabalho identificou variantes genômicas específicas da raça, que exercem impacto direto sobre a expressão gênica e estão situados em genes envolvidos diretamente na ativação de vias bioquímicas relacionadas aos processos básicos do metabolismo celular, como variantes de alto impacto presentes em genes de GTPases, mas principalmente variantes de nucleotídeo único associadas à genes Receptores Olfativos, relacionados à percepção do cheiro através do olfato, cujos eventos de sinalização atuam no reconhecimento do alimento, da prole, de parceiros sexuais e outros aspectos essenciais da sobrevivência animal, sendo portanto, a principal família gênica relacionada ao reconhecimento do meio e aos processos adaptativos animais.

A presença de alelos polimórficos nos *loci* desses importantes genes, para todas as variantes testadas nas três populações estudadas de remanescentes da raça, indicam possível fixação dos alelos polimórficos nas populações e fortes indícios de que o efeito dessas variantes está, de fato, relacionado ao sucesso adaptativo da raça e ao seu padrão racial. Esse é um dos

diversos artifícios genéticos provavelmente relacionados à rusticidade e resistência da raça.

É importante considerar, ainda, que os dados do presente estudo, uma vez disponibilizados e divulgados incentivam a inclusão da raça equina Nordestina em estratégias de conservação, incluindo bancos de germoplasma de raças naturalizadas; estratégias muito necessárias para evitar o desaparecimento do cavalo Nordestino, que representa um patrimônio genético de grande importância.

VII REFERÊNCIAS BIBLIOGRAGICAS

ABCCN. Associação Brasileira dos Criadores do Cavalo Nordestino. Regulamento do registro genealógico do cavalo Nordestino. ABCCN. Recife; 1987. pp.33-34

ALMEIDA, L. D. Diversidade genética de raças asininas criadas no Brasil, baseada na análise de locos microssatélites e DNA mitocondrial. 2009. 83 f. Dissertação de Mestrado. Faculdade de Agronomia e Medicina Veterinária, Universidade de Brasília, Brasília.

ALFONSO, L. Use of meta-analysis to combine candidate gene association studies: application to study the relationship between the ESR PvuII polymorphism and sow litter size. *Genetics, Selection and Evolution*, v.37, p.417-435, 2005.

ANDERSSON, L. How selective sweeps in domestic animals provide new insight into biological mechanisms. *J Intern Med*., v. 271, p.1-14, 2012.

ANDERSSON, L.; GEORGES, M. Domestic-animal genomics: deciphering the genetics of complex traits. *Nature Reviews Genetics*, v.5, p.202-212, 2004.

BRAGA, R.M. Os cavalos trazidos para o Brasil. In: Cavalo Lavradeiro em Roraima: Aspectos históricos, ecologia e de Conservação. Brasília: EMBRAPA, 2000. p.26-32.

CHAN, E. K. F.; HAWKEN, R.; REVERTER, A. The combined effect of SNPmarker and phenotype attributes in genome-wide association studies. *Anim. Genet*, v.40, p.149-156, 2009

CHURCH, D.M.; SCHNEIDER, V.A.; STEINBERG, K.M.; SCHATZ, M.C.; QUINLAN, A.R.; CHIN, C.S et al. Extending reference assembly models. *Genome biology*, v.16, n.1, p.13-20, 2015.

COSTA, N.; VAL, L.J.; LEITE, G.U. Estudo da Preservação do Cavalo Nordestino. Recife: Departamento de Produção Animal, 1974. 38p.

COSTA, H.E.; MANSO FILHO, H.; FERREIRA, L. Exterior e treinamento do cavalo. Recife: UFRPE – Imprensa Universitária, 2001. 169p.

CUNNINGHAM, J. G.Tratado fisiologia veterinária. 3ed Rio de Janeiro: Guanabara Koogan, 2004. 579p.

DAS A.; PANITZ F.; GREGERSEN V.R.; BENDIXEN C.; HOLM L.E. Deep sequencing of Danish Holstein dairy cattle for variant detection and insight into potential loss-of-function variants in protein coding genes. *BMC Genomics*, v.16, n.1043, 2015.

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

DOAN, R.; COHEN, N. D.; SAWYER, J.; GHAFFARI, N.; JOHNSON, C. D.; DINDOT, S. V. Whole-Genome Sequencing and Genetic Variant Analysis of a Quarter Horse Mare. *BMC Genomics*, v.13, n.78, p.1471-2164, 2012.

DALRYMPLE, B. P.; KIRKNESS, E. F.; NEFEDOV, M.; MCWILLIAM,, S.; RATNAKUMAR, A.; BARRIS, W.; ZHAO, S.; SHETTY, J.; MADDOX, J. F.; O'GRADY, M.; NICHOLAS, F.; CROWFORD, A. M.; SMITH, T.; JONG, P. J.; McEVAN, J.; ODDY, V. H.; COCKETT, N. E. and INTERNATIONAL SHEEP GENOMICS CONSORTIUM. Using comparative genomics to reorder the human genome sequence into a virtual sheep genome. *Genome Biology*, v.8, n.7, p.152.1-152.20, 2007.

EGITO, A.A.; MARIANTE A.S.; ALBUQUERQUE, M.S.M. Programa brasileiro de conservação de recursos Genéticos animais. *Archivos de zootecnia*, v.51, n. 193-194, p.39-52, 2002.

GARRISON, E. & MARTH G. Haplotype-based variant detection from short-read sequencing. 2012. Preprint. Available from: arXiv:1207.3907. 2012.

GRIFFIN, TJ, SMITH, L.M. Single nucleotide polymorfisman analysis by maldi-tof mass spectrometry. *Trends in biotecnology*, V.18, n. 2, p.77-84, 2000.

GROENEN, M. A. M.; MEGENS, H. J.; ZARE, Y.; WARREN, W. C.; HILLIER, L. W.; CROOIJMANS, R. P. M. A.; VEREIJKEN, A.; OKIMOTO, R.; MUIR, W. M.; CHENG,H.H.

The development and characterization of a 60K SNP chip for chicken. *BMC genomics*, v.12, p.1-9, 2011.

HORSE GENOME PROJECT (HGP). 2011. Disponível em: < http://www.uky.edu/Ag/Horsemap/abthgp.html> acesso em: 02/06/2018.

ILLUMINA. Introduction to Whole-Genome Sequencing. Disponível em: < http://www.illumina.com/techniques/sequencing/dna-sequencing/whole-genome-sequencing.html > acesso em: 06/11/2019.

JOON-HO L.; TAEHEON L.; HAK-KYO L.; BYUNGWOOK C.; DONG-HYUN S.; KYOUNG-TAG D.; et al. Thoroughbred horse single nucleotide polymorphism and expression joodatabase: HSDB. *Asian-Australs J. Anim. Sci*; v.27, p. 1236-1243, 2014.

KALBFLEISCH T.S.; RICE E.S.; DEPRIEST JR M.S.; WALENZ B.P.; HESTAND M.S.; VERMEESCH J.R.; et al. EquCab3, an updated reference genome for the domestic horse. *BioRxiv*, 2018. https://doi.org/10.1101/306928 (on line).

KIJAS J.W.; LENSTRA J.A.; HAYES B.; BOITARD S.; PORTO NETO L.R.; et al. Genome-wide analysis of the world's sheep breeds reveals high levels of historic mixture and strong recent selection. *PLoS Biol.*, v.10, 2012.

KHATKAR, M.S.; THOMSON, P.C.; TAMMEN, I.; RAADSMA, H.W. Quantitative trait loci mapping in dairy cattle: review and meta-analysis. *Genetics, Selection, Evolution*, v.36, p.163–190, 2004.

MARIANTE A.S.; ALBUQUERQUE M.S.M.; RAMOS A.F.. Criopreservação de recursos genéticos animais brasileiros. *Rev. Bras. Reprod. Anim*., v.35, n.2, p.64-68, 2011.

MARAI, I.F.M.; EL- DARAWANY, A.A.; ABOU-FANDOUD, E.I.; ABDEL-HAFEZ, M.A.M. Serun blood components during pre-oestrus, oestru and pregnancy phases in Egyptian Suffolk as affected by heat stress, under the conditions of Egypt. *Anim. Sci*., v.1, n.4, p.47–62, 2007.

MARAI, I.F.M.; HABEEB A.A.M. Buffalo's biological functions as affected by heat stress-a review. *Livest Sci*., v.127, n.3, p.89-109, 2010.

MCLAREN W.; PRITCHARD B.; RIOS D.; CHEN Y.; FLICEK P.; CUNNINGHAM F. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics*, v.26, n.16, p.2069-2070, 2010.

MCCUE M.E.; BANNASCH D.L.; PETERSEN J.L.; GURR J.; BAILEY E.; BINNS M.M.; et al. A High Density SNP Array for the Domestic Horse and Extant Perissodactyla: Utility for Association Mapping, Genetic Diversity, and Phylogeny Studies. *PLoS Genet*., v.8, n.1, p.36-42, 2012.

METZGER J.; TONDA R.; BELTRAN S.; ÁGUEDA L.; GUT M.; DISTL O. Next generation sequencing gives an insight into the characteristics of highly selected breeds versus non-breed horses in the course of domestication. *BMC Genomics*, v.15, n.562, 2014.

MELO, J.B.; RIBEIRO, M.N.; JÚNIOR, A.M.A.; TRAVASSOS, A.E.V. Caracterização fenotípica do Cavalo Nordestino, na mesorregião agreste do Estado de Pernambuco.(Relatório Final de Licença Sabática), 2006. 32p.

MELO, J. B. Caracterização zoométrica do remanescente da raça equina Nordestina nos estados de Pernambuco e Piauí. 2011. 118f. Tese (Doutorado em Zootecnia) - Universidade Federal Rural de Pernambuco. Departamento de Zootecnia, Recife.

MELO, J.B. ; PIRES, D. A. F. ; RIBEIRO, M. N. . Perfil fenotípico do remanescente do cavalo nordestino no nordeste do brasil. *Archivos de Zootecnia*. 62:171-180, 2013.

MODREK, B. & LEE, C. A genomic view of alternative splicing. *Nature Genetics*, v. 30, p. 13-19, 2002.

NEOGEN, Equine SNP70 BeadChip Whole Genome SNP profiling, 2013. Disponível em: acesso em 21/12/2019.

PATEN, B.; NOVAK, A.; HAUSSLER, D. Mapping to a reference genome structure. arXiv preprint arXiv:1404.5010, 2014.

PEREIRA, J.C.C. Iniciação à bioestatística médica. Benvinda, 3a edição, 2016. 336p

PIRES, D.A.F.; RIBEIRO, M.N.; MELO, J.B. et al. Estudo zoométrico do remanescente de Cavalos Nordestinos no município de Agrestina – Pernambuco. In: Jornada de ensino, pesquisa e extensão da UFRPE, 2008. Recife. **Anais...** Recife: 2008. (CD-ROM).

PIRES, D. A. F. ; COELHO, E. G. A. ; MELO, J. B. ; OLIVEIRA, D. A. A. ; RIBEIRO, M. N. ; GUS COTHRAN, E. ; JURAS, R. ; KHANSHOUR, A. . Genetic diversity and population structure in remnant subpopulations of nordestino horse breed. *Archivos de Zootecnia* (Internet), v. 63, p. 349-358, 2014.

PIRES, D. A. F. ; MELO, J. B. ; RIBEIRO, M. N. ; ARANDAS, J. K. G. ; NASCIMENTO, R. B. Avaliação de Componentes Principais em medidas morfométricas de cavalos remanescentes da raça equina Nordestina do município de Agrestina-PE, Brasil. In: *49ª Reunião Anual da Sociedade Brasileira de Zootecnia*, 2012, Brasília. 49ª Reunião Anual da Sociedade Brasileira de Zootecnia. Brasília - DF: Sociedade Brasileira de Zootecnia, 2012.

PIRES, D. A. F. Caracterização genética de remanescentes da raça equina nordestina em mesorregiões dos estados da bahia, pernambuco e piauí através de marcadores microssatélites. 2012a. 101f. Dissertação (Mestrado em Zootecnia) - Universidade Federal Rural de Pernambuco. Departamento de Zootecnia, Recife.

RAFALSKI, A. Applications of single nucleotide polymorphisms in crop genetics. *Current Opinion in Plant Biology*, v.5, n.2, p. 94-100, 2002.

REMINGTON, D. L., THORNSBERRY, J. M., MATSUOKA, Y., WILSON, L.M., WHITT, S. R., DOEBLEY, J et al. Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proceedings of the National Academy of Science*, v. 20, p.11479-11484, 2001.

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*, v.56, p. 63-77, 2008.

SALOMÓN-TORRES R.; GONZÁLEZ-VIZCARRA V.M.; MEDINA-BASULTO G.E.; MONTAÑO-GÓMEZ M.F.; MAHADEVAN P.; YAURIMA-BASALDÚA V.H.; et al. Genome-wide identification of copy number variations in Holstein cattle from Baja California, Mexico, using high-density SNP genotyping arrays. *Genet Mol Res*, v.14, p.11848-11859, 2015.

SAMPAIO, I.B.M. Estatística aplicada à experimentação animal. FEPMVZ, Escola de Veterinária da UFMG, 4a edição, reimpressão, 2015, 265p.

SCHAEFER R.J.; SCHUBERT M.; BAILEY E.; BANNASCH D.L.; BARREY E.; BAR-GAL G.K.; et al. Developing a 670k genotyping array to tag ~2M SNPs across 24 horse breeds. *BMC Genomics*, n.18, v.1, p.565-576, 2017.

SCHNEIDER S, D ROESSLI, L EXCOFFIER. Arlequin v.2.000: software for population genetics data analysis. User manual: ver. 2.0. Geneva, Switzerland: Genetics and Biometry Laboratory, University of Geneva; 2000.

SILVA, R. G. Introdução à bioclimatologia animal. São Paulo: Nobel, 2000, 286p.

SOUZA W.; CARVALHO B. Rqc: Quality Control Tool for High-Throughput Sequencing Data. R package version 1.2.0, 2014.

STARLING, J. M. C.; SILVA, R. G.; CERÓN-MUÑOZ, M.; BARBOSA, G. S. S. C.; PARANHOS DA COSTA, M. J. R. Análise de algumas variáveis fisiológicas para avaliação do grau de adaptação de ovinos submetidos ao estresse por calor. *Revista Brasileira de Zootecnia*, v. 31, n.5, p. 2070-2077, 2002.

TRAVASSOS, A.E.V. Caracterização fenotípica do cavalo nordestino no estado de Pernambuco. 2004. 59f. Dissertação (Mestrado em Zootecnia) - Universidade Federal Rural de Pernambuco, Recife.

VALENTE T.S.; BALDI F.; SANT'ANNA A.C.; ALBUQUERQUE L.G.; PARANHOS D.A.; COSTA M.J. Genome-Wide Association Study between Single Nucleotide Polymorphisms and Flight Speed in Nellore Cattle. PLoS One 14: 11(6), 2016 doi: 10.1371/journal.pone.0156956. PMID:27300296.

WADE, C. M.; GIULOTTO, E.; SIGURDSSON, S.; ZOLI, M.; GNERRE, S.; IMSLAND, F.; et al. Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science*, v. 326, n. 5954, p. 865-867, 2009.

WRAY, G.A. The evolutionary significance of cis-regulatory mutations. *Natature Reviews Genetics*, v.8, n.3, p.206-16, 2007.

WOMACK, J.E. Advances in livestock genomics: opening the barn door. Genome Research, 15:1699-1705, 2005.

ZHAN B.; FADISTA. J; THOMSEN B.; HEDEGAARD J.; PANITZ F.; BENDIXEN C. Global assessment of genomic variation in cattle by genome resequencing and high-throughput genotyping. *BMC Genomics*., v.12, n. 1, p. 557-569, 2011.

ZHANG C.; NI P.; AHMAD H.I., GEMINGGULI M., BAIZILAITIBEI A., GULIBAHETI D.; et al. Detecting the Population Structure and Scanning for Signatures of Selection in Horses (*Equus caballus*) from Whole-Genome Sequencing Data. *Evol Bioinform*, v. 4, n.14, 2018.