

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

Danielle Cunha Cardoso

Universidade Federal de Minas Gerais Departamento Zootecnia

Eduardo Geraldo Alves Coelho

Universidade Federal de Minas Gerais Departamento Zootecnia

Brenda Neves Porto

Empresa Brasileira de Pesquisa Agropecuaria Recursos Geneticos e Biotecnologia

glacy silva (✉ glacyjaqueline@prof.unipar.br)

Universidade Paranaense <https://orcid.org/0000-0002-7088-1363>

Denea de Araújo Fernandes Pires

Instituto Federal de Educacao Ciencia e Tecnologia de Pernambuco



Denise Aparecida Andrade de Oliveira

Universidade Federal de Minas Gerais Departamento Zootecnia

Research

Keywords: InDels, Nordeste horse breed, semi-arid adaptative traits, SNPs, whole-genome sequencing.

DOI: <https://doi.org/10.21203/rs.3.rs-21956/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Abstract

Background: In this study, we perform a 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 conditions, 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.

Results: A total of 305,588,364 reads were mapped in the horse reference genome, 1,598,210 single nucleotide variations and 138,139 insertions/deletions were detected by FreeBayes, 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.

Conclusions: 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, 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 were overrepresented.

Background

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 (Mariante et al. 2011; Melo et al. 2013) 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, which requires 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 (Pires et al. 2014). To date, there are no studies about 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 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 (Andersson 2012). As a result, most equine breeds present populations with high phenotypic and genotypic uniformity within the breed, but there is a lot of variation between breeds.

The equine genome project (*Equus caballus*) has made publicly available a full and high quality genome data of a Thoroughbred female, representing a breakthrough in genomics and veterinary medicine (Wade et al. 2009). As 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 (Doan et al. 2012) and in Chinese horses (Lichuan and Kazakh breeds) (Zhang et al. 2018).

In early 2018, a new and improved equine genome assembly (EquCab3.0) was made available (Kalbfleisch et al. 2018), boosting future 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 (670 k), comprising genomic information from individual representatives of 24 different equine breeds. The study cataloged 23 million new genetic variants (Schaefer et al. 2017). 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 (McCue et al. 2012), sheep (Kijas et al. 2012), cattle (Zhan et al. 2011; Salomón-Torres et al. 2015; Valente et al. 2016) and other species of zootechnical interest.

When intending 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 (Das et al. 2015). There is a great interest in genetic variants in new equine breeds, especially SNVs, for the creation of a SNPs database and integration of quantitative and linkage maps, as performed for the Thoroughbred breed, in order to contribute to breeding strategies (Lee et al. 2014).

Although the whole genome re-sequencing has become an accessible and easy-to-perform technique for variant search, most of these studies focus on equine breeds aimed at 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 information 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 analysis. That annotation allows the identification, location and association of variations related to the complex resistance traits 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 from 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 \pm 8), hull type, characteristic stiffness and head size (ABCCN, 1987). The sample was kindly provided by our partner research group from the Instituto Federal de Educação, ciência e Tecnologia (IFPE) and Universidade Estadual do Sudoeste da Bahia (UESB).

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 a 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 \times 301 bp paired-end reads) in a 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) (Andrews, 2010). For the reads mapping, the horse reference genome (Ensembl EquCab3.0) was used. Clean sequencing reads were mapped to the reference assembly, using the Burrows-Wheeler Aligner tool (BWA, version 0.7.10-r789) with default parameters (Li & Durbin, 2009). 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 carried out to improve the mapping quality (McKenna *et al.* 2010). Downstream processing was carried out using typical GATK pipeline, according to parameters applied by Cornish and Guda (2015), for the GATK base quality score recalibration (BQSR) step.

Variant detection and annotation

Variant calling was conducted with two tools: FreeBayes (<https://github.com/ekg/freebayes>) (Garrison & Marth, 2012) and GATK (<https://software.broadinstitute.org/gatk>) (DePristo *et al.* 2011), 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 vcf filter (<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 (Cingolani *et al.* 2012), 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) (Huang *et al.* 2009). 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 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, a 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. Particularly in

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 (Table 1).

Table 1. Number of variants 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 the recommended quality metrics for each software, it was found that the total number of variants detected by FreeBayes was higher than GATK, which is expected, since GATK exhibits higher sensitivity while maintaining a lower number of false positive SNVs (Cornish & Guda, 2015).

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 future 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 with already known and present in SNPs arrays, since this is the first study about the breed and we initially searched 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 call tools, SNVs and InDels had very low occurrence in donor and acceptor splice-site 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

number of effects is greater than the number of variants because those found between genes positioned in close proximity can have their effects categorized as both downstream and upstream.

The percentage of SNVs effects per genotypic region of the Nordestino horse presented here was very similar to the average percentage found in native Chinese breeds (Lichuan and Kazakh breeds, small and rugged horses), as demonstrated by Zhang *et al.* (2018) 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 were transcript, intron and intergenic (29.07%, 28.12% and 27.02%, respectively), and the smaller effects were exactly the same as found here, with very similar percentages.

Using the SnpEff program [24], we also classified the effects of variants (SNVs and InDels) by impact as modifiers of mostly 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.

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 by both 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 checked which genes are present in the analysis by both variants calling tools. We found 28 Emsembl IDs from multigenic families containing at least one high impact SNVs, in accordance with both 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 accordance with both GATK and FreeBayes variant calling tools.

| GeneName | Transcript | Product | BioType | GATK SNVs impact HIGH | FreeBayes SNVs impact HIGH |
|--|------------------|------------------------------|--------------------|--------------------------------|----------------------------------|
| <i>LOC1021493</i> 42 | <i>gene32392</i> | | pseudogene | 1 | 1 |
| <i>LOC1001466</i> 99 (<i>id196201</i>) | <i>rna15620</i> | GTPase family member 7 | protein_codin g | 1 | 1 |
| <i>id196202</i> | <i>rna15621</i> | GTPase family member 7 | protein_codin g | 1 | 1 |
| <i>id196203</i> | <i>rna15622</i> | GTPase family member 7 | protein_codin g | 1 | 1 |
| <i>id196204</i> | <i>rna15623</i> | GTPase family member 7 | protein_codin g | 1 | 1 |
| <i>id196205</i> | <i>rna15624</i> | GTPase family member 7 | protein_codin g | 1 | 1 |
| <i>id196206</i> | <i>rna15625</i> | GTPase family member 7 | protein_codin g | 1 | 1 |
| <i>id196207</i> | <i>rna15626</i> | GTPase family member 7 | protein_codin g | 1 | 1 |
| <i>id196208</i> | <i>rna21264</i> | GTPase family member 7 | protein_codin g | 1 | 1 |
| <i>id196209</i> | <i>rna21265</i> | GTPase family member 7 | protein_codin g | 1 | 1 |
| <i>id196210</i> | <i>rna21266</i> | GTPase family member 7 | protein_codin g | 1 | 1 |
| <i>id196211</i> | <i>rna25375</i> | GTPase family member 7 | | 1 | 1 |

| | | | | | |
|--|-----------------|------------------------------|--------------------|---|---|
| <i>id196213</i> | <i>rna38835</i> | GTPase family member 7 | protein_codin g | 1 | 2 |
| <i>id196214</i> | <i>rna38989</i> | GTPase family member 7 | protein_codin g | 1 | 1 |
| <i>id196215</i> | <i>rna40155</i> | GTPase family member 7 | protein_codin g | 1 | 2 |
| <i>id196216</i> | <i>rna43275</i> | GTPase family member 7 | protein_codin g | 1 | 1 |
| <i>id196217</i> | <i>rna51848</i> | GTPase family member 7 | protein_codin g | 1 | 1 |
| <i>id196219</i> | <i>rna63380</i> | GTPase family member 7 | | 1 | 1 |
| <i>id196220</i> | <i>rna63381</i> | GTPase family member 7 | | 1 | 1 |
| <i>LOC1117731</i> 16 (<i>id196221</i>) | <i>rna71433</i> | GTPase family member 4 | protein_codin g | 1 | 1 |
| <i>id196222</i> | <i>rna71664</i> | GTPase family member 4 | | 1 | 1 |
| <i>id196223</i> | <i>rna73843</i> | GTPase family member 4 | protein_codin g | 1 | 1 |
| <i>id196224</i> | <i>rna73844</i> | GTPase family member 4 | protein_codin g | 1 | 1 |
| <i>LOC1000544</i> 58 (<i>id196225</i>) | <i>rna73845</i> | GTPase family member 2 | protein_codin g | 1 | 1 |
| <i>id196226</i> | <i>rna76474</i> | GTPase family | protein_codin g | 2 | 5 |

Considering also the relevance of moderate impact effects, we identified 392 Ensembl IDs containing SNVs with this effect, in accordance with both 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

| | | member 2 | | | |
|-------------------|-----------------|---------------|-----------------|---|---|
| <i>id196227</i> | <i>rna76774</i> | GTPase family | protein_codin g | 1 | 1 |
| | | member 2 | | | |
| <i>id196228</i> | <i>rna76804</i> | GTPase family | protein_codin g | 1 | 1 |
| | | member 2 | | | |
| <i>LOC1000637</i> | <i>rna76872</i> | GTPase family | protein_codin g | 1 | 2 |
| 77 | | | | | |
| <i>(id196229)</i> | | member 1 | | | |

Integrated Discovery (DAVID). The screening for 70 genes was performed 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 arid conditions resistance traits 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, NBP7, OR10D4, OR52L2, OR56A3, OR56A4, OR6B2, OR7G2, OR8S1, PDPR, RNASEL, RWDD3, SLC45A1, SWT1, WAPL, WD40*). We explored these genic 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 accordance with both GATK and FreeBayes

| | ID | Name | FDR | Genes with high and moderate impact SNVs |
|--------------------------|------------|---|----------|--|
| GO Molecular Function | | olfactory receptor activity | | <i>OR8S1,OR56A3,OR56A4,OR6B2,OR7G2,OR52L2P,OR10D4P</i> |
| | GO:0004984 | proton transmembrane transporter activity | 5,37E-01 | <i>MT-ATP6,MT-CO2,MT-CO3,MT-CYB</i> |
| | GO:0015078 | G protein-coupled receptor signaling | 1,81E+00 | <i>OR8S1,OR56A3,OR56A4,OR6B2,OR7G2,OR52L2P,OR10D4P</i> |
| | GO:0007186 | electron transfer activity | 1,48E+01 | <i>MT-CO2,MT-CO3,MT-CYB</i> |
| | GO:0009055 | transmembrane signaling receptor activity | 1,48E+01 | <i>KLRK1,OR8S1,OR56A3,OR56A4,OR6B2,OR7G2,OR52L2P,OR10D4P</i> |
| | GO:0004888 | oxidoreductase activity | 1,48E+01 | <i>MT-CO2,MT-CO3</i> |
| | GO:0016676 | cytochrome-c oxidase activity | 1,48E+01 | <i>MT-CO2,MT-CO3</i> |
| | GO:0004129 | heme-copper terminal oxidase activity | 1,48E+01 | <i>MT-CO2,MT-CO3</i> |
| | GO:0015002 | oxidoreductase activity | 1,48E+01 | <i>MT-CO2,MT-CO3</i> |
| | GO:0016675 | signaling receptor activity | 1,81E+01 | <i>KLRK1,OR8S1,OR56A3,OR56A4,OR6B2,OR7G2,OR52L2P,OR10D4P</i> |
| | GO:0038023 | squalene synthase activity | 1,81E+01 | <i>FDFT1</i> |
| | GO:0051996 | farnesyl-diphosphate farnesyltransferase activity | 1,81E+01 | <i>FDFT1</i> |
| | GO:0004310 | transmembrane transporter activity | 3,02E+01 | <i>MT-ATP6,MT-CO2,MT-CO3,MT-CYB</i> |
| | GO:0015077 | MHC class Ib receptor activity | 3,10E+01 | <i>KLRK1</i> |
| | GO:0032394 | molecular transducer activity | 3,89E+01 | <i>KLRK1,OR8S1,OR56A3,OR56A4,OR6B2,OR</i> |
| | GO:0060089 | | | |

| | | | | |
|------------------------------|------------|--|----------|--|
| | | [pyruvate dehydrogenase phosphatase activity | 3,89E+01 | <i>PDPR</i> |
| | GO:0004741 | | | <i>SLC45A1,MT-ATP6,MT-CO2,ABCD2,MT-CO3,MT-CYB</i> |
| | GO:0022857 | transmembrane transporter activity | 3,89E+01 | <i>PDPR,MT-CO2,MT-CO3,MT-CYB,FDFT1</i> |
| | GO:0016491 | oxidoreductase activity | 4,77E+01 | <i>OR8S1,OR56A3,OR56A4,OR6B2,OR7G2,OR52L2P,OR10D4P</i> |
| GO Biological Process | GO:0050911 | chemical stimulus in sensory perception of smell | 1,42E+00 | <i>OR8S1,OR56A3,OR56A4,OR6B2,OR7G2,OR52L2P,OR10D4P</i> |
| | GO:0007608 | sensory perception of smell | 1,42E+00 | <i>OR8S1,OR56A3,OR56A4,OR6B2,OR7G2,OR52L2P,OR10D4P</i> |
| | GO:0050907 | sensory perception | 1,42E+00 | <i>OR8S1,OR56A3,OR56A4,OR6B2,OR7G2,OR52L2P,OR10D4P</i> |
| | GO:0009593 | detection of chemical stimulus | 1,67E+00 | <i>OR8S1,OR56A3,OR56A4,OR6B2,OR7G2,OR52L2P,OR10D4P</i> |
| | GO:0007606 | sensory perception of chemical stimulus | 1,67E+00 | <i>OR8S1,OR56A3,OR56A4,OR6B2,OR7G2,OR52L2P,OR10D4P</i> |
| | GO:0050906 | detection of stimulus involved in sensory perception | 1,67E+00 | <i>OR8S1,OR56A3,OR56A4,OR6B2,OR7G2,OR52L2P,OR10D4P</i> |
| | GO:0051606 | detection of stimulus | 8,72E+00 | <i>MT-CO2,MT-CO3,MT-CYB</i> |
| | GO:0022900 | electron transport chain | 4,92E+01 | <i>OR8S1,OR56A3,OR56A4,OR6B2,OR7G2,OR52L2P,OR10D4P</i> |
| KEEG Pathway | 1269583 | Olfactory Signaling Pathway | 9,52E-01 | <i>PDPR,MT-ATP6,MT-CO2,MT-CO3,MT-CYB</i> |
| | 1270121 | The citric acid cycle and respiratory electron transport | 9,52E-01 | |

| | | | | |
|--------------------|---------|--|----------|--|
| | 1270127 | Respiratory electron transport, ATP synthesis | 3,36E+00 | <i>MT-ATP6,MT-CO2,MT-CO3,MT-CYB</i> |
| | | and heat production. | | |
| | 82942 | Oxidative phosphorylation | 3,36E+00 | <i>MT-ATP6,MT-CO2,MT-CO3,MT-CYB</i> |
| | 93344 | Cardiac muscle contraction | 7,00E+00 | <i>MT-CO2,MT-CO3,MT-CYB</i> |
| | 1270128 | Respiratory electron transport | 1,27E+01 | <i>MT-CO2,MT-CO3,MT-CYB</i> |
| | 83087 | Olfactory transduction | 1,27E+01 | <i>OR8S1,OR56A3,OR56A4,OR6B2,OR7G2</i> |
| | 1269574 | GPCR downstream signaling | 2,66E+01 | <i>OR8S1,OR56A3,OR56A4,OR6B2,OR7G2,OR52L2P,OR10D4P</i> |
| | 142287 | epoxysqualene biosynthesis | 3,56E+01 | <i>FDFT1</i> |
| Gene Family | 167 | Olfactory receptors, family 56 | 1,31E+00 | <i>OR56A3,OR56A4</i> |
| | 643 | Mitochondrial complex: cytochrome c oxidase subunits | 3,10E+00 | <i>MT-CO2,MT-CO3</i> |
| | 721 | Rho GTPase activating proteins | 1,44E+01 | <i>ARHGAP28,ARHGAP20</i> |
| | 808 | ATP binding cassette subfamily D | 3,03E+01 | <i>ABCD2</i> |
| | 580 | GTPases, IMAP | 4,31E+01 | <i>GIMAP7</i> |
| | 1055 | Exocyst complex | 4,31E+01 | <i>EXOC6</i> |
| | 642 | Mitochondrial complex III | 4,31E+01 | <i>MT-CYB</i> |
| | 621 | CD molecules Killer cell lectin like receptors | 4,52E+01 | <i>KLRK1</i> |

The *GTPase* 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 cell processes, including proliferation, differentiation, vesicle and organelle dynamics, transport and regulation of the cytoskeleton

(Bos *et al.* 2007). These are evolutionarily conserved proteins, associated with the activation of extracellular signals to various cellular responses, due to the ability to undergo conformational changes in response to the alternate binding of GDP and GTP. The GDP-bound “off” or “on” states recognize distinct effector proteins, thereby allowing these proteins to function as binary molecular switches (Bos *et al.* 2005). 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. A study conducted by Zhang *et al.* (2018) 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 representation, 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 (Trzaskowski *et al.* 2012).

When we analyzed moderate impact SNVs, the variety of gene families in which they occur is wide, drawing attention to the olfactory receptors family, which had high representation. These genes, involved in the Olfactory Signaling pathway, act in the perception of odor through olfact, interact with odorant molecules in the nasal epithelium, to initiate a neuronal response that triggers the perception of a smell (Antunes *et al.* 2014). The biochemical signaling events related to this (super pathway) act in food recognition and consequently food preference (Ma 2007), identification of sexual partners (Kang *et al.* 2015), mother-infant bonding (Doucet *et al.* 2009) 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.* (2016), 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.* (2013). The high representation 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, there are 390 OR genes and 465 pseudogenes (Olender *et al.* 2008), since these receptors recognize varied binders, from chemical compounds to peptides (Ma 2007).

In a cattle variant calling study (Stafuzza *et al.* 2017), the olfactory transduction pathway was over represented, in all four important cattle breeds in Brazil: Guzerat, Gyr, Girolando and Holstein. (Metzger *et al.* 2014) 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.* (2014) 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.* (2014), 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 housekeeping, 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.

Conclusions

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 containing 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 the development of resistance markers to arid climate conditions.

Declarations

Ethics approval

The blood sample was collected from a male horse, in a private property, with written consent from 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).

Consent for publication (Not applicable)

Availability of data and materials

The datasets generated and/or analysed during the current study are not publicly available due they belong to the germplasm bank of Embrapa Recursos Genéticos e Biotecnologia, but are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests

Funding

The authors thank Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Instituto Nacional de Ciência e Tecnologia Pecuária (INCT), grant CNPq 573899/2008 and FAPEMIG APQ-0084/08.

Author's contributions

DCC and EGAC Conceived of and designed the experiments. DCC, BNP and GJS analyzed the data. DAFP Contributed to sampling. DCC, GJS, DAAO contributed to the writing of the manuscript. DAAO contributed to the acquisition of reagents and materials. All authors read and approved the final manuscript.

Acknowledgements (Not applicable)

References

Andersson L. How selective sweeps in domestic animals provide new insight into biological mechanisms. *J. Intern. Med.* 2012; 271: 1-14.

Andrews S. FastQC: a quality control tool for high throughput sequencedata. 2010. <http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/>. Accessed 02 Feb 2020.

Antunes G, Sebastião AM, Souza FM. Mechanisms of Regulation of Olfactory Transduction and Adaptation in the Olfactory Cilium. *Plos One.* 2014; 9:e105531.

Bos JL, Rehmann H, Wittinghofer A. GEFs and GAPs: critical elements in the control of small G proteins. *Cell.* 2007; 129: 865-77.

Bos JL. Linking rap to cell adhesion. *Curr Opin Cell Biol.* 2005; 17: 123-128.

Cingolani P, Platts A, Wang L, Coon M, Nguyen T, Wang L, Ruden DM. 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: 80-92.

Cornish A, Guda C. A comparison of variant calling pipelines using genome in a bottle as a reference. *Biomed Res Int.* 2015;1

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.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, McKenna A. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet.* 2011; 43: 491-501.

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; 13:78-90.

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.

Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:2012; 1207.3907.*

Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nat. Protoc.* 2009; 4: 44-57.

Jun J, Cho YS, Hu H, Kim HM, Jho S, Gadhvi P, Manica A. Whole genome sequence and analysis of the Marwari horse breed and its genetic origin. *BMC Genomics.* 2014; 15: S4.

Kalbfleisch TS, Rice ES, DePriest Jr. MS, Waenz BP, Hestand MS, Vermeesch JR, Finno CJ. EquCab3, an updated reference genome for the domestic horse. *BioRxiv.* 2018: 306928.

Kang N, Kim H, Jae Y, Lee N, Ku CR, Margolis F, Koo J. Olfactory Marker Protein Expression Is an Indicator of Olfactory Receptor-Associated Events in Non-Olfactory Tissues. *Plos One.* 2015; 10: e0116097.

Kijas JW, Lenstra JA, Hayes B, Boitard S, Neto LRP, San Cristobal M, Paiva S. Genome-wide analysis of the world's sheep breeds reveals high levels of historic mixture and strong recent selection. *Plos Biol.* 2012; 10:e1001258.

Kupke A, Wensch S, Failing K, Herden C. Intranasal location and Immunohistochemical characterization of the equine olfactory epithelium. *Front. Neuroanat.* 2016; 10:97.

Lee JH, Lee T, Lee HK, Cho BW, Shin DH, Do KT, Cho S. horse single nucleotide polymorphism and expression joodatabase: HSDB. *Asian Austral J Anim.* 2014; 27: 1236-1243.

Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2009; 25: 1754-1760.

Ma M. Encoding olfactory signals via multiple chemosensory systems. *Crit Rev Biochem Mol.* 2007; 42: 463-480.

Mariante AS, Albuquerque MSM, Ramos AF. Criopreservação de recursos genéticos animais brasileiros. *RBRA.* 2011; 35: 64-68.

McCue ME, Bannasch DL, Petersen JL, Gurr J, Bailey E, Binns MM, Leeb T. 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; e1002451.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskins K, Kernytzky A, DePristo MA. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010; 20:1297-303.

Melo JB, Pires DAF, Ribeiro MN. Perfil fenotípico do remanescente do cavalo nordestino no nordeste do Brasil. *Archives de Zootecnia.* 2013; 62: 171-180.

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: 562-575.

Olender T, Lancet D, Nebert DW. Update on the olfactory receptor (OR) gene superfamily. *Hum. Genomics.* 2008; 3: 87-97.

Pires DAF, Coelho EGA, Melo JB, Oliveira DAA, Ribeiro MN, Gus Cothran E, Khanshour A. Genetic diversity and population structure in remnant subpopulations of nordestino horse breed. *Archivos de Zootecnia.* 2014; 63: 349-358.

Salomón-Torres R, González-Vizcarra VM, Medina-Basulto GE, Montañó-Gómez MF, Mahadevan P, Yaurima-Basaldúa VH, Villa-Ángulo R. 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: 11848-11859.

Schaefer RJ, Schubert M, Bailey E, Bannasch DL, Barrey E, Bar-Gal GK, Finno CJ. Developing a 670k genotyping array to tag ~2M SNPs across 24 horse breeds. *BMC Genomics.* 2017; 18: 565.

Stafuzza NB, Zerlotini A, Lobo FP, Yamagishi MEB, Chud TCS, Caetano AR, Carvalho MR. Single nucleotide variants and InDels identified from whole-genome re-sequencing of Guzerat, Gyr, Girolando and Holstein cattle breeds. *Plos One.* 2017; 12: e0173954.

Trzaskowski B, Latek D, Yuan S, Ghoshdastider U, Debinski A, Filipek S. Action of molecular switches in GPCRs-theoretical and experimental. *Curr Med Chem.* 2012; 19:109.

Valente TS, Baldi F, Sant'Anna AC, Albuquerque LG, Costa MJP. Genome-Wide Association Study between Single Nucleotide Polymorphisms and Flight Speed in Nellore Cattle. *Plos One* 2016; 14: 11- e0156956.

Wade CM, Giulotto E, Sigurdsson S, Zoli M, Gnerre S, Inslan F, Blocker H. Genome sequence, comparative analysis and population genetics of the domestic horse. *Science*. 2009; 326: 865-867.

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: 557.

Zhang C, Ni P, Ahmad HI, Gemingguli M, Baizilaitibei A, gulibaheti D, Chen J. Detecting the Population Structure and Scanning for Signatures of Selection in Horses (*Equus caballus*) from Whole-Genome Sequencing Data. *Evol Bioinform* 2018; 4, 1-9.

Figures

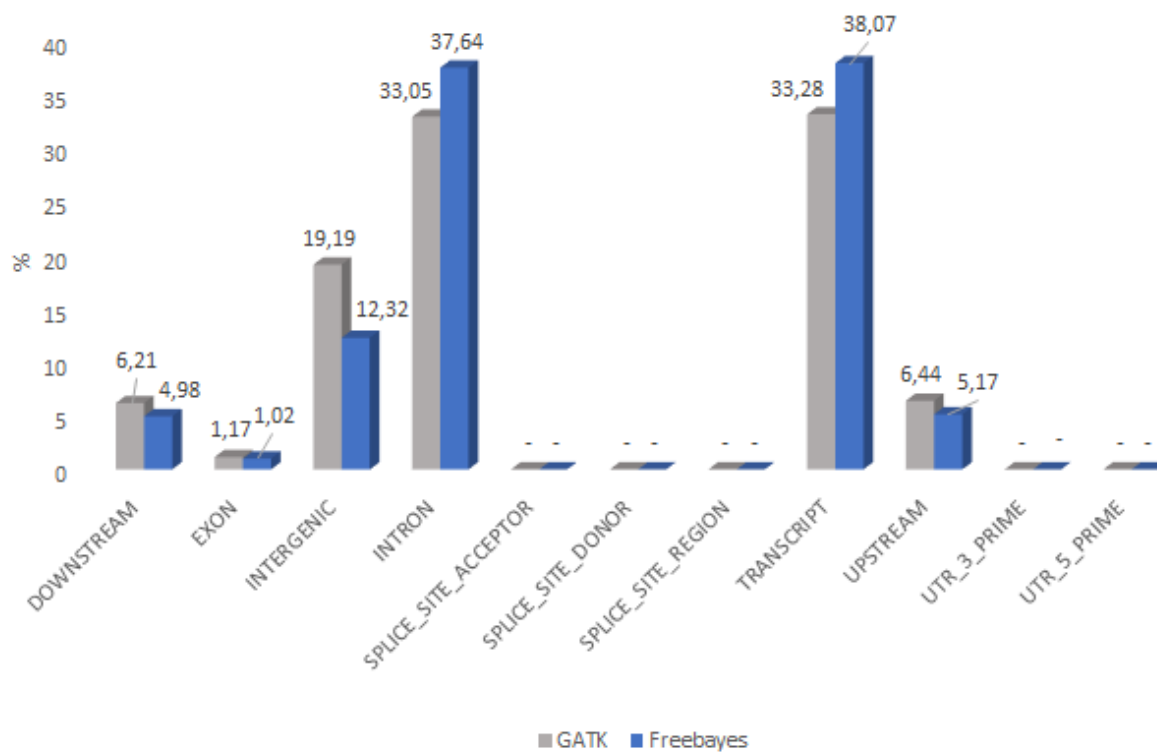


Figure 1

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 donor site, splice acceptor site, splice site region, transcripts and intergenic regions. "Downstream" and "Upstream" are 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 (Cingolani et al. 2012). 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 (Zhang et al. 2018).

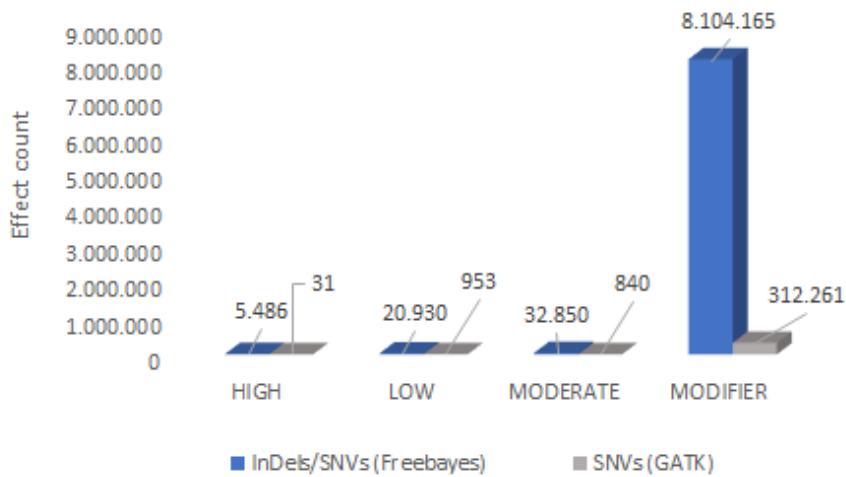


Figure 2

Number of variants effects by impact, according to FreeBayes (InDels + SNVs) and GATK(SNVs) softwares. 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)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [additionalfiles.xlsx](#)