Indian Journal of Animal Sciences **93** (8): 802–805, August 2023/Article
https://doi.org/10.56093/ijans.v93i8.136161

Genome-wide copy number variations in Bhutia equine breed using SNP genotyping data

NITESH KUMAR SHARMA[1], PRASHANT SINGH[2], BIBEK SAHA[1], ANURADHA BHARDWAJ[2], MIR ASIF IQUEBAL[1], YASH PAL[2], VARIJ NAYAN[3], SARIKA JAISWAL[1*], SHIV KUMAR GIRI[4], RAM AVATAR LEGHA[2], T K BHATTACHARYA[2], DINESH KUMAR[1], ANIL RAI[1] and BHUPENDRA NATH TRIPATHI[5]

ICAR-Indian Agricultural Statistics Research Institute, Pusa, New Delhi 110 012 India

Received: 9 May 2023; Accepted: 20 June 2023

ABSTRACT

Copy number variants (CNVs) have dynamic potential and evolutionary significance like other genetic variants, namely, single nucleotide polymorphisms, InDels, short tandem repeat polymorphisms, inversion variants, etc. Discovering CNVs leads to further speculation that the genomic DNA contains more changes than previously thought and contributes to the phenotypic variation. CNVs are big DNA fragments (> 1 kb) being duplicated or deleted. A bridge between CNVs and phenotypic variations supports CNVs to be utilized in GWAS, which are currently mostly based on SNPs. CNV, which refers to the structural differences, influence gene expression and can be an indicator of numerous traits for improvement. There is a severe dearth of research on CNVs in animals, especially equine. The present study investigates the genomes of the Bhutia Equine breed for genome-wide discovery of CNVs using the Axiom™ Equine Genotyping Array chip for a better understanding of its traits which had been unexplored till date. A total of 619 CNVs from 20 Bhutia equines were identified with the median and average size as 49.394 kb and 114.955 kb, respectively. Total 225 frequent CNVRs with > 1% CNV frequency were identified among them along with singleton type. These CNVRs contained 361 genes in all. The information obtained on genomic variation could be utilized to identify economically advantageous genetic features in Bhutia equine breed.

Keywords: Bhutia, Copy number variation (CNV), Copy number variation region (CNVR), Equine, Genes

The domesticated horse (*Equus ferus caballus*) is a hoofed, one-toed mammal. It is one of the two existing subspecies of *Equus ferus* and a member of the taxonomic family Equidae (Wilson and Reeder 2005). Eohippus, a small animal with several toes, is the ancestor of the horse, which developed 45 to 55 million years ago to become the huge animal, as it is today. Humans began domesticating horses around 4000 BC. The animal is adapted to run fast and possesses a good sense of balance and a strong fight response.

There exist four separate pony breeds in India, namely, Bhutia, Spiti, Zanskari, and Manipuri found in different regions (Gupta *et al.* 2012). The Bhutia breed is a small mountain horse, native to Sikkim and Darjeeling in India, sharing similarities with Mongolian and Tibetan breeds. Phenotypically, it has a large head with smaller eyes and ears. It has straight shoulder, deep chest, and strong legs with powerful hindquarters, mainly used to carry loads as well as riding. Due to low populations in their home tracts of fewer than 5,000 in India, the Bhutia (2,230) pony's an endangered breed (Provisional Livestock Census, India 2008). This represents the total number of horses in these regions, which also includes hybrids of the native Tibetan and Spiti breeds. Breeds are regarded as endangered if they are less than 5,000 in number (Alderson 1999).

There are numerous lines of evidence that at an intra-species level, DNA copy number variations (CNVs) contribute to the phenotypic variation of humans, animals, and plants along with single-nucleotide polymorphisms and short insertion-deletion polymorphisms (Handsaker *et al*. 2015, Xu *et al.* 2016). The occurrence of a big DNA fragment (> 1 kb) being duplicated or deleted is known as CNV. A bridge between CNVs and phenotypic variations also suggest that CNVs can be utilized in GWAS, which is now based mostly on SNPs (Atwell *et al.* 2010). Compared to SNPs, CNVs show more drastic effects on gene expression and function such as altering gene dosage, disrupting coding sequence, or perturbing long-range gene regulation (Zhou *et al.* 2018).

Recent research suggests that structural differences,

Present address: [1]Division of Agricultural Bioinformatics, ICAR-Indian Agricultural Statistics Research Institute, New Delhi. [2]ICAR-National Research Centre on Equines, Sirsa Road, Hisar, Haryana. [3]ICAR-Central Institute for Research on Buffaloes, Hisar, Haryana. [4]Department o f Biotechnology, Maharaja Agrasen University, Baddi, Himachal Pradesh. [5]ICAR, New Delhi. ✉Corresponding author email: sarika@icar.gov.in

such as CNV, influencing gene expression, are connected to the beginning of numerous illnesses (Lee and Redon *et al.* 2006, Scherer *et al.* 2007, Morton 2008, Glessner *et al.* 2009). These researches, however, are typically exclusively examining the human genome. Such investigations into the genomes of other economically important animal breeds, such as in Equines would pave the way ahead for a better understanding of its exclusive traits. In this work, we aim at discovering the genome-wide CNVs from the Bhutia equine genome, which is unexplored domain to date. The results/information obtained on genomic variation could be utilized to identify economically advantageous genetic features in this significant Bhutia equine breed.

MATERIALS AND METHODS

Sample collection, DNA extraction, and genotyping: Peripheral blood samples from the Bhutia pony breed, covering 20 genotypes, were aseptically taken from the jugular vein into a vacutainer with K2 EDTA. The blood sampling of animals was performed following the proper guidelines and regulations as approved by the Institutional Animal Ethics Committee (IAEC) of the National Research Centre on Equines (ICAR-NRCE), Hisar, Haryana, India. DNA isolation from the blood samples was performed from these samples stored at 4°C. Genomic DNA was extracted using a ReliaPrepTM Blood gDNA Miniprep System (Promega, USA) as per the manufacturer's protocol and isolated DNA was stored at -20°C to avoid DNA deterioration. DNA was analyzed qualitatively and quantitatively using agarose gel electrophoresis (0.8% agarose gel) and a Qubit4 fluorometer (at 260 nm/280 nm absorbance). The Axiom™ Equine Genotyping Array was used to genotype the samples, with an average call rate of 92.6% for each sample.

Microarray data processing: The signal intensity (SI) and B allele Frequency (BAF) values were derived using the CEL files (raw data) for each SNP. Affymetrix Power Tools (APT) program which implements a suite of cross-platform command line techniques for analyzing and interacting with Affymetrix arrays, was used to produce the values. For CNV identification in Affymetrix SNP arrays, the rules outlined in the PennCNV-Affy Protocol (http://penncnv.openbioinformatics.org/en/latest.user-guide/affy/) we also followed.

The two alleles' intensity values known as the A and B alleles are a compilation of signal intensity values taken from the APT software's "AxiomGT1.summary.txt" file. The Perl script that implements the following process was used to normalize the two signal intensity levels for each SNP by expressing them as the Log2 ratio (LRR): (a) first step is to create a reference for each marker using the equation T = A + B, where A and B represent the signal intensity levels for each allele. A reference to M = median is set for each SNP (T sample1, T sample2,..., T sampleN); (b) normalized signal intensity for each SNP and each sample SNP was obtained by estimating the intensity for each sample using the expression log2 (T/M) ratio (Rincon *et al*. 2011). Based on the "AxiomGT1.calls" file, which contains genotype calls (labeled as -1 = NN, AA = 0, AB = 1, BB = 2), and other quality control filters to the data, SNPs with genotyping mistakes (no call) along with any non-somatic SNPs, if present, were removed. This led to a total of final 601894 SNPs taken forward for further analysis.

Genome-wide identification of CNVs and gene content within CNVRs: For CNV identification, PennCNV algorithm was utilized (Wang *et al.* 2007). The LRR, BAF, and separation between each SNP value for each marker are inputs needed by the PennCNV algorithm. The 31 autosomal chromosomal default values were used to run PennCNV, and the GC model option was used to modify the genomic waves. A Perl script was used to create the GC model file for this investigation, and it calculates the GC content within 1 Mb of each marker (500 kb on each side). At least three contiguous SNPs with a total length larger than 3 kb were taken into account to identify putative CNV. The criteria used in a study by Redon *et al.*(2006) and team were utilized to designate CNV regions (CNVRs).

The ensemble BioMart database (https://asia.ensembl.org/info/data/biomart/index.html) was used to identify gene contents and obtain a description of each gene affected within the regions encompassed by CNVRs (Cunningham *et al.* 2015).

RESULTS AND DISCUSSION

In this study, PennCNV and Axiom™ Equine Genotyping Array chips were employed to find CNVs in the Bhutia Equine breed. The median size and average length of the 31 CNVs in one sample were 49.394 kb and 114.955 kb, respectively (Table 1).

A total of 225 CNV regions (CNVRs) were discovered after combining all the CNVs. The median size and the average length of the average 11 CNVRs in each sample were 43.08 kb and 115.5 kb, respectively. In addition, this study also inferred 55 CNVRs with > 10% frequency, 19 CNVRs with > 25% frequency, and 10 CNVRs with >

Table 1. Summary of identified copy number variations in the Bhutia breed of equine (*n* = 20)

	Total number	Average no. of CNVs per sample	Average size of CNVs (kb)	Median size of CNVs (kb)	No. of gain	No. of loss	No. of both (gain + loss)	Ratio (loss/ gain)	No. of common CNVs (freq. >10%)	No. of common CNVs (freq. >25%)	No. of common CNVs (freq. >50%)	Genes
Individual CNV	619	30.95	114.95	49.39	268	351	_	1.31	_	_	_	_
CNV region	225	11.25	115.5	43.08	74	127	24	_	55	19	10	361

Table 2. Summary of common copy number variation regions in Bhutia breed of equine (freq. >50%)

CNVR ID	Chromosome	Start position	End position	Length	No. of CNVs	Frequency (%)
CNVR12	chr1	155349336	155732506	383171	18	90
CNVR15	chr1	158768332	160176963	1408632	14	70
CNVR71	chr6	72012114	72548957	536844	31	155
CNVR72	chr6	72826374	73416876	590503	11	55
CNVR90	chr8	4251195	4636875	385681	13	65
CNVR113	chr12	12321830	14937463	2615634	51	255
CNVR156	chr20	30598562	30688857	90296	12	60
CNVR158	chr20	30853546	31048896	195351	11	55
CNVR162	chr20	32113332	32477946	364615	23	115
CNVR209	chr29	6371	807818	801448	13	65

50% frequency. The detected CNVR includes 361 genes in total. Table 2 lists the common CNVRs that have a CNV frequency>50%, with the chr12:12321830-14937463 CNVR having the highest frequency. The circular map of CNVRs was constructed using the CNVs and CNVRs (Fig. 1). It was observed that 225 CNVRs were uniformly spaced throughout the chromosomes. Of about 24 CNVRs (both) that had similar values, 127 were with only loss (deletion), and 74 were with only gain (duplication) among them. Out of a total of 361 genes, of which 244 were encoded for proteins, 16 were pseudogenes, 6 were snoRNAs, 6 were miRNAs, 3 were TR_V_genes, and one was snRNA.

A meager information/ research on equine CNV is available in the literature, despite it being a significant source of genetic diversity alongside SNPs for understanding how CNVs affect specific traits, including vulnerability to disease as explored in the human genome (Zhang *et al.* 2009). There are some economically advantageous traits in animal genomes too and future studies may be enabled

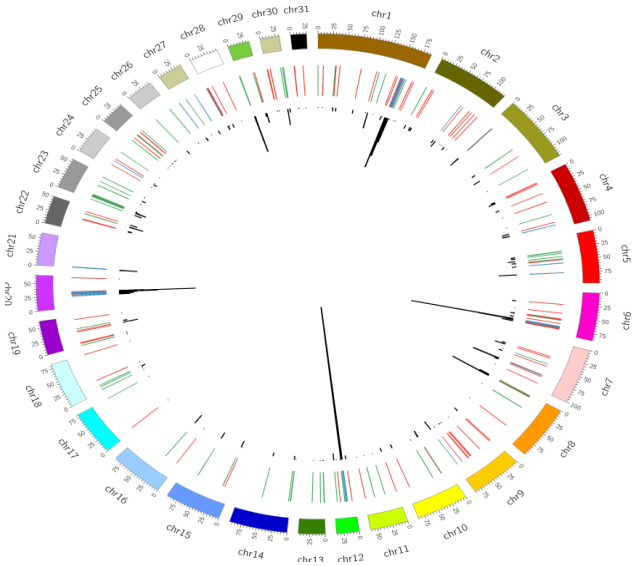

Fig. 1. Circular map of CNVRs identified using PennCNV distributed across 31 chromosomes of Bhutia equine. Chromosome sizes are displayed on the outer band. Chromosome CNVRs are visible in the center band. The frequency of samples in a CNVR is displayed in the inner band.

to investigate the genetic influences of CNV on numerous economic variables based on the results of this study. Till now, the majority of CNV-related research has been conducted on their association analyses solely based on genotyping changes in signal intensity. CNV, however, is typically connected to SNPs in the area. The concept of CNV and SNP combination analysis has recently been applied. The primary benefit of this approach is the simultaneous analysis of allelic differences and signal intensity. In other words, CNVs with multiple SNP markers can be genotyped using multi-allelic CNV/SNP techniques. The genotyped multi-allelic CNV markers in the human genome's deletion region, and CNV/SNP combination analysis leading to higher accuracy in association results as compared to SNP or CNV genotyping alone is reported in the literature (Bae *et al.* 2008).

In this investigation, 225 CNVRs and 619 CNVs in Bhutia equine genome were found. The common CNVRs detected will help examine specific correlations between phenotypes to apply the study's findings. For instance, a typical CNVR offers crucial genome data for identifying genes related to temperaments, such as spirited hot bloods with speed and endurance; cold bloods, like draft horses and some ponies, suitable for slow, heavy work; and warmbloods, developed from crossing hot blood and cold blood, frequently with a focus on developing breeds for specific riding purposes. The Axiom™ Equine Genotyping Array chip and EquCab3.0 algorithm were used to identify CNVs. More precise CNV identification results would have been anticipated if better or specific horse genome assembly had been utilized.

According to recent reports, the advancements achieved in the chip platform caused the most recent discoveries of CNV sizes to be significantly less than the prior findings (McCarroll *et al.* 2008, Perry *et al.* 2008). The threshold value of 1 kb was criticized by Zhang and colleagues, who suggested instead using an average exon size (100 bp) to define CNV (Zhang *et al.* 2009). Venter and Watson's research using whole-genome shotgun sequencing and DNA sequencing technologies reveled a notable concentration of CNV sizes within the range of 300 to 350bp, as highlighted in studies by Wheeler *et. al.* (2008) and Levy *et. al.* (2007). Even though a 60K chip was used for this study, high-

resolution approaches, like high-density chips or next-generation sequencing, should be used to examine animal genomes, including equine genomes. The current study would be useful as a preliminary report offering whole-genome CNV distribution resources for the horse genome in studies defining the precise CNV border.

This work is the first of its kind in the Bhutia breed that offers the genetic resources necessary for examining the potential economic impacts that phenotypic and horse CNVs may have. The high-resolution CNV mapping follow-up investigations have been under process. However, CNVs associated with economically valuable traits have not yet been comprehensively investigated, therefore we anticipate that the findings of this work will contribute significant information on chromosomal variation to related studies. The magnitude of horse CNVs should be accurately estimated in future studies, and then an association study of equine phenotypes should be conducted. Such a study will help to understand how CNVs affect specific traits, in the endeavour of trait improvement and disease management of this economically important, less explored Equine breed.

A total of 619 CNVs were discovered in the Bhutia breed of Equines along with 225 frequent CNVRs that contained 361 genes in all. We anticipate that this outcome/ information will offer significant resources for further equine genomic research.

ACKNOWLEDGEMENTS

The financial grants, ICAR-CABin, and IARI Merit scholarship to NKS are duly acknowledged. The authors further acknowledge the supportive role of the Director, ICAR-IASRI, New Delhi, India.

REFERENCES

Alderson L. 1999. Criteria for the recognition and prioritisation of breeds of special genetic importance. *Animal Genetic Resources* **33**: 1–9.

Atwell S, Huang Y S, Vilhjálmsson B J, Willems G, Horton M, Li Y and Nordborg M. 2010. Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature* **465**(7298): 627–31.

Bae J S, Cheong H S, Kim J O, Lee S O, Kim E M, Lee H W, *et al*. 2008. Identification of SNP markers for common CNV regions and association analysis of risk of subarachnoid aneurysmal hemorrhage in Japanese population. *Biochemical and Biophysical Research Communications* **373**(4): 593–96.

Cunningham F, Amode M R, Barrell D, Beal K, Billis K, Brent S, *et al*. 2015. Ensembl 2015. *Nucleic Acids Research* **43**(D1): D662–D669.

Glessner J T, Wang K, Cai G, Korvatska O, Kim C E, Wood S, *et al*. 2009. Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature* **459**(7246): 569–73.

Gupta A K, Tandon S N,Pal, Y, Bhardwaj A and Chauhan M. 2012. Phenotypic characterization of Indian equine breeds: A comparative study. *Animal Genetic Resources* **50**: 49–58.

Handsaker R E Van Doren, V, Berman J R, Genovese G, Kashin S, Boettger L M and McCarroll S A. 2015. Large multiallelic copy number variations in humans. *Nature Genetics* **47**(3): 296–303.

Lee C and Morton C C. 2008. Structural genomic variation and personalized medicine. *New England Journal of Medicine* **358**(7): 740–41.

Levy S, Sutton G, Ng P C, Feuk L, Halpern A L, Walenz B P, *et al*. 2007. The diploid genome sequence of an individual human. *PLoS Biology* **5**(10): e254.

McCarroll S A, Kuruvilla F G, Korn J M, Cawley S Nemesh J, Wysoker A, Altshuler D, *et al.* 2008. Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nature Genetics* **40**(10): 1166–74.

Perry G H, Ben-Dor A, Tsalenko A, Sampas N, Rodriguez-Revenga L, Tran C W, Lee C, *et al.* 2008. The fine-scale and complex architecture of human copy-number variation. *The American Journal of Human Genetics* **82**(3): 685–95.

Redon R, Ishikawa S, Fitch K R, Feuk L, Perry G H, Andrews T D, Hurles M E, *et al.* 2006. Global variation in copy number in the human genome. *Nature* **444**(7118): 444–54.

Rincon G, Weber K L, Van Eenennaam A L, Golden B L, Medrano J F, *et al.* 2011. Hot topic: Performance of bovine high-density genotyping platforms in Holsteins and Jerseys. *Journal of Dairy Science* **94**(12): 6116–21.

Scherer S W, Lee C, Birney E, Altshuler D M, Eichler E E, Carter N P, Feuk L, *et al.* 2007. Challenges and standards in integrating surveys of structural variation. *Nature Genetics* **39**(Suppl 7): S7–S15.

Wang K, Li M, Hadley D, Liu R, Glessner J, Grant S F, Bucan M, *et al.* 2007. PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Research* **17**(11): 1665–74.

Wheeler D A, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, Rothberg J M, *et al.* 2008. The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**(7189): 872–76.

Wilson D E and Reeder D M. 2005. Mammal species of the world: A taxonomic and geographic reference (Vol. 1). 3rd edition. Johns Hopkins University Press, Baltimore, Maryland.

Xu L, Hou Y, Bickhart D M, Zhou Y, Hay E H A, Song J, *et al.* 2016. Population-genetic properties of differentiated copy number variations in cattle. *Scientific Reports* **6**(1): 23161.

Zhang F, Gu W, Hurles M E and Lupski J R. 2009. Copy number variation in human health, disease, and evolution. *Annual Review of Genomics and Human Genetics* **10**: 451–81.

Zhou Y, Connor E E, Wiggans G R, Lu Y, Tempelman R J, Schroeder S G, Liu G E, *et al.* 2018. Genome-wide copy number variant analysis reveals variants associated with 10 diverse production traits in Holstein cattle. *BMC Genomics* **19**: 1–9.