Genomics Data 7 (2016) 97–100

Contents lists available at ScienceDirect

Genomics Data

journal homepage: www.elsevier.com/locate/gdata

Data in Brief

# Analysis of copy number variations in Mexican Holstein cattle using axiom genome-wide Bos 1 array

Ricardo Salomon-Torres [a,b,*,1], Rafael Villa-Angulo [c,1], Carlos Villa-Angulo [c]

[a] Sonora State University, S.L.R.C., Sonora, México
[b] Faculty of Engineering Mexicali, Autonomous University of Baja California, BC, Mexico
[c] Laboratory of Bioinformatics and Biofotonics, Engineering Institute, Autonomous University of Baja California, BC, Mexico

ARTICLE INFO

*Article history:*
Received 25 November 2015
Accepted 15 December 2015
Available online 17 December 2015

*Keywords:*
Holstein cattle
Copy number variation
SNP
Axiom genome-wide Bos 1 array
Bioinformatics
PennCNV

ABSTRACT

Recently, for copy number variation (CNV) analysis, bovine researchers have focused mainly on the use of genome-wide SNP genotyping arrays. One of the highest densities commercially available SNPchips for cattle is the Affymetrix axiom genome-wide Bos 1, which assays 648,315 informative SNPs across the whole bovine genome. Here, we describe the microarray data, quality controls and validation implemented in a study published in Genetics and Molecular Research Journal in 2015 [1]. The microarray raw data has been deposited into Gene Expression Omnibus under accession #GSE54813.



| Specifications | |
|---|---|
| Organism/cell line/tissue | Bos taurus |
| Sex | Female |
| Sequencer or array type | Affymetrix axiom genome-wide Bos 1 array |
| Data format | Raw data |
| Experimental factors | Any pretreatment of samples |
| Experimental features | DNA extraction from blood samples cows, genotyped with high density arrays, copy number variations detection and validation. |
| Consent | Not applicable. |
| Sample source location | Veterinary Science Research Institute of the Autonomous University of Baja California, Mexicali, Mexico. |

## 1. Direct link to deposited data

http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE54813.

## 2. Experimental design, materials and methods

### 2.1. Characteristics of the samples

The blood samples were collected by venipuncture of the coccygeal vein, from 12 Holstein dairy cows, registered in the Mexican Holstein Association. All were born after artificial insemination, and were between their first and fourth lactation; they were all clinically healthy and free of brucellosis and tuberculosis. All were selected such that they were not related in the last three generations.

### 2.2. DNA extraction and genotyping

DNA extraction and purification were performed using a QIAGEN kit. All DNA samples were analyzed by spectroscopy and agarose gel electrophoresis, and were genotyped with the axiom genome-wide Bos 1 array with an average call rate for each individual sample of 99.7%. The raw data of the SNPchip were submitted to the Gene Expression Omnibus under the accession number GSE54813.

### 2.3. Microarray data processing

We extracted signal intensity (SI) and B allele Frequency (BAF) values from CEL files (raw data) for each SNP. Values were generated by the Affymetrix Power Tools (APT) software, which implements a set of cross-platform command line algorithms for analyzing and working with Affymetrix arrays. APT documentation can be obtained from

* Corresponding author at: Laboratory of Bioinformatics and Biofotonics, Engineering Institute, Blvd. Benito Juárez and de la Normal Street S/N, Mexicali, Baja California C.P. 21280, México.
*E-mail address:* ricardo.salomon@uabc.edu.mx (R. Salomon-Torres).
[1] These authors contributed equally to the present manuscript.

http://dx.doi.org/10.1016/j.gdata.2015.12.007

(http://www.affymetrix.com/estore/ partners_programs/programs/developer/tools/powertools.affx). We also used the guideline defined in PennCNV-Affy Protocol for CNV detection in Affymetrix SNP arrays (http://penncnv.openbioinformatics.org/en/latest.user-guide/affy/).

### 2.4. Normalization and quality control

The intensity values from the two alleles are referred as the A and B alleles. These alleles are summarized signal intensity values obtained from "AxiomGT1.summary.txt" file, produced by the APT software. Finally, the two values of signal intensity for each SNP were normalized by expressing them as $Log_2$ ratio (LRR), using a Perl script which implements the following procedure: first, a reference is developed for each marker considering the formula $T = A + B$, where A and B are the values of the signal intensity of each allele. For each SNP, a reference to the value M = median is set (T_sample1, T_Sample2, …, T_sampleN). The second step is to estimate the intensity for each individual sample with the formula $\log_2$ (T/M) ratio, from which we get the normalized signal intensity for each SNP and each sample SNP [2].

We applied some quality control filters to the data, we eliminated all SNPs with genotyping errors (no call), based on the "AxiomGT1.calls" file, which contain genotype calls (−1 = NN, AA = 0, AB = 1, BB = 2). We also filtered all non-somatic SNPs. Our final working dataset was of 601,894 SNPs.

### 2.5. Genome-wide identification of CNVs

We used two algorithms: PennCNV [3] and QuantiSNP [4], for CNV detection. The PennCNV algorithm requires as input LRR and BAF values for each marker, and the distance between each SNP. PennCNV was executed using default values for the 29 autosomal chromosomes, and genomic waves were adjusted using the argument called GC model. The GC model file for this study was generated by a Perl script, which computes the GC content within 1 Mb around each marker (500 kb each side). QuantiSNP was executed with the options *-isaffy* and *-levels* enabled since we used an Affymetrix array. In the same way *-gcdir* option was enabled to perform the correction of the LRR, in markers affected by genomic waves [5] (Fig. 1).

For declaring a putative CNV, we considered at least three adjacent SNPs indicating a loss or gain, with a total length greater or equal to 1 kb, detected simultaneously by the two algorithms in the same animal, either in the same position or overlapping. Finally, CNV regions (CNVRs) were defined based on the criteria used in a study by Redon et al. [6].

PennCNV detected 155 CNVs, while QuantiSNP detected 302. The algorithms coincided for 77 putative CNVs, detected in the same position and the same sample (Fig. 1). Initially, we termed these variants as putative CNVs. We inspected the 77 CNVs for overlaps and defined 56 CNVRs. (Fig. 2).

## 3. Basic analysis

### 3.1. Functional analysis of genes

To identify gene contents and to obtain a description of each gene affected within the regions covered by CNVRs, we used the BioMart database (http://www.biomart.org) and the RefGen database (http://refgene.com). We found 103 genes, of which 96 encoded proteins, two were pseudogenes, three were snRNAs, and two were miRNAs. In order to analyze functional enrichment in the CNVRs, we searched the Gene Ontology (GO) database [7] and the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [8]. Both analyses were carried out using the bioinformatic tool DAVID [9]. The GO analysis showed common gene terms among mammals. KEGG pathway analysis showed that the genes were mainly represented in the pathway of olfactory transduction.

### 3.2. CNV validation by real-time PCR (qPCR)

For each target CNVR, two pairs of primers were designed considering the limits of each CNVR. PCR primers were designed using the NCBI Primer-BLAST (http://www.ncbi.nlm.nih.gov/tools/primer-blast). The PCR amplification program was 5 min at 95 °C, followed by 40 cycles at 95 °C for 10 s and 60 °C for 10 s. We used the Basic Transcription Factor (*BTF3*) as a control gene for comparing the number of copies in each CNVR [1].

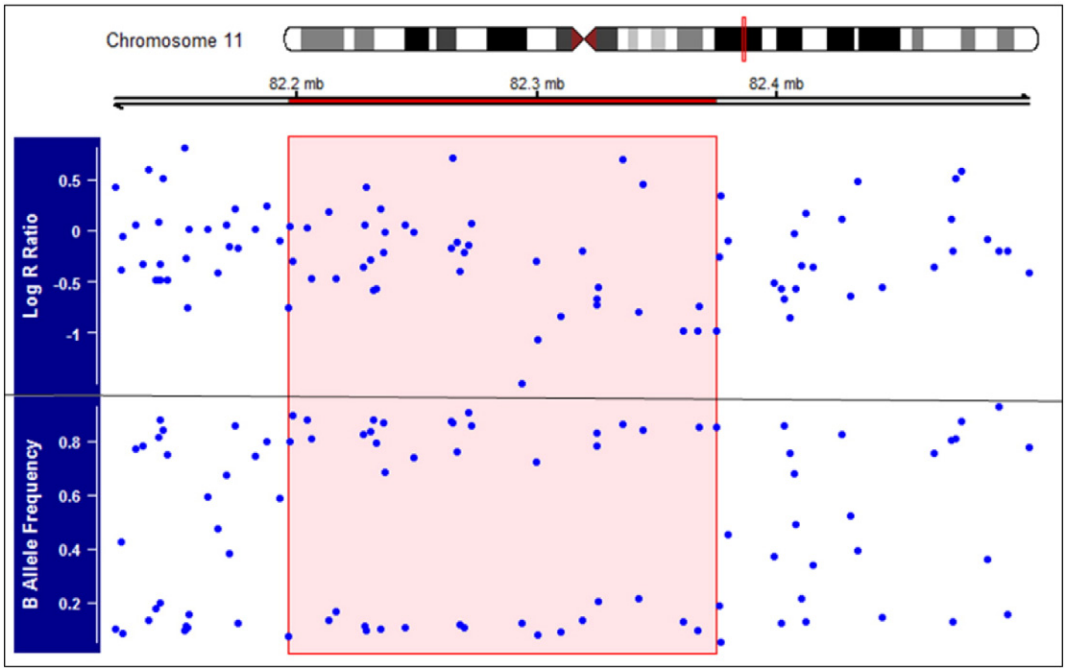


**Fig. 1.** Log R ratio (LRR) and B allele frequency (BAF) plot of one copy number variation region (CNVR). Inside the selected area, low values of LRR (less than −1) and no values in the 0.5 cluster indicate a single copy deletion in a region of chromosome 11.

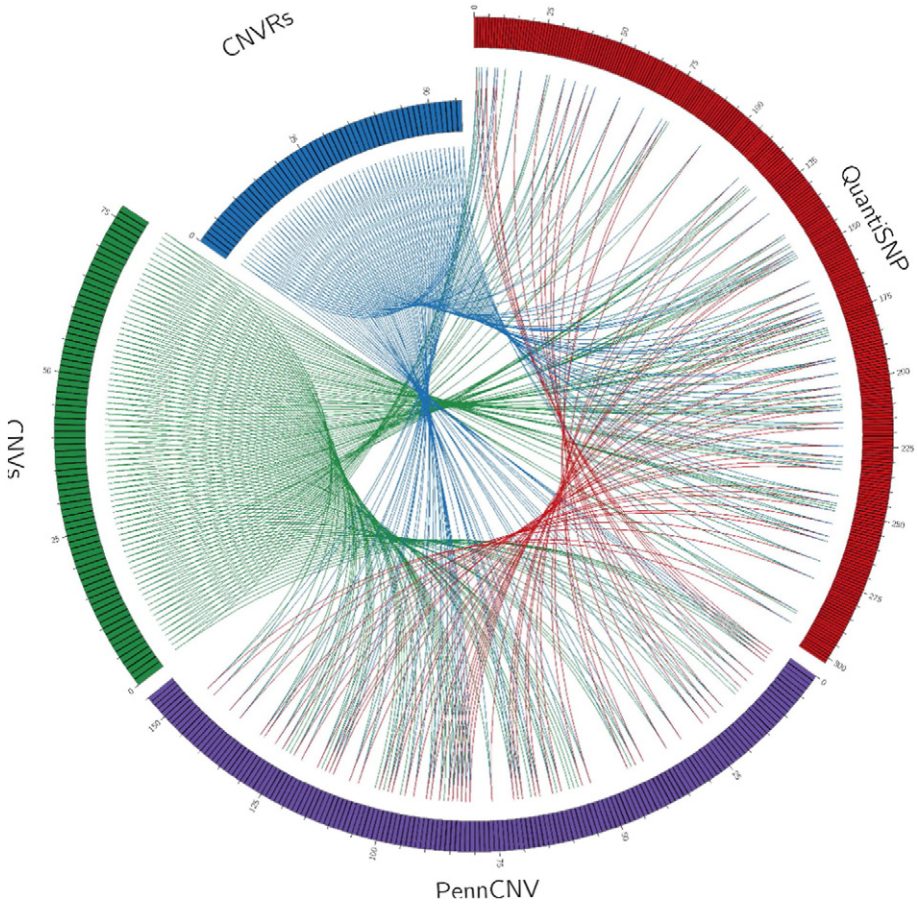


**Fig. 2.** The circle shows the coincidences of CNVs and CNVRs between the two algorithms. Links on blue and red represents the 56 CNVRs where PennCNV and QuantiSNP coincided. Links on green represents the 77 CNVs in which both algorithms coincided.

We use the method of comparative cycle threshold ($2^{-\Delta\Delta Ct}$) to quantify the number of changes of the copies by comparing the ΔCt value, from the samples with CNV to a ΔCt of a calibrator without CNV [10]. The average Ct value of three replicates for each sample was calculated, normalized, and compared against the control gene, with the assumption of the existence of two copies of the DNA segment in the control region.

For each CNVR to be validated, the value of $2 \times 2^{-\Delta\Delta Ct}$ was calculated for each individual. The obtained value was used to decide if a CNVR was normal (without CNVR, if the value was about two), or a gain (if the value was about three or above), or a deletion (if the value was near zero or one) [11] (Fig. 3).

## 4. Discussion

We describe here, in the best of our knowledge, the first publically available high-density SNP genotypes dataset from bovine genome. This dataset is composed of raw data from 12 Holstein cows. 56

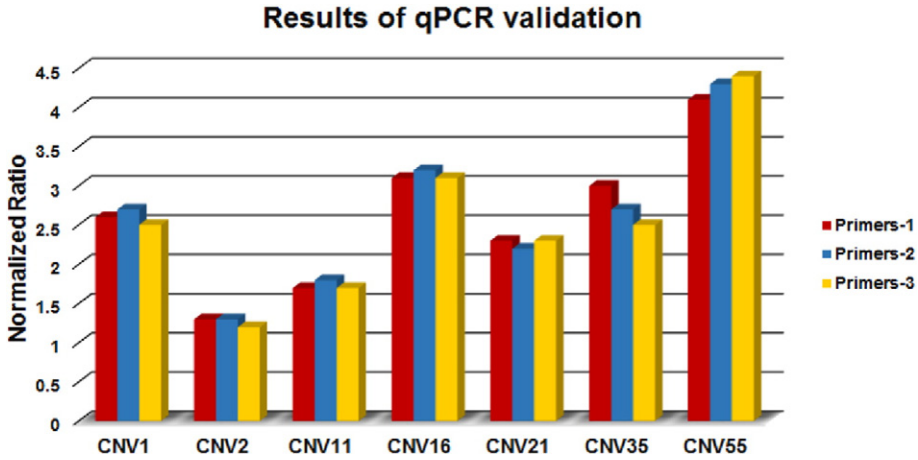


**Fig. 3.** The normalized ratio in two (normal state), assumes the existence of two copies of the DNA segment. Values around one indicate a single copy loss; values around three indicate a three copy gain, and around four indicate a four copies gain.

CNVRs genome-wide were identified in the analysis. In addition, five of the putative CNVRs were validated by qPCR. Finally, we showed that SNP data from Affymetrix axiom genome-wide Bos 1 array, allows achieving great accuracy in the identification of CNVRs and their candidate genes.

## Conflict of interest

The authors declare no conflict of interest.

## Acknowledgments

We are grateful to the Council for Science and Technology of Mexico (CONACYT) for supporting a scholarship for postdoctoral studies for Ricardo Salomon-Torres, number scholarship 362690.

## References

[1] R. Salomon-Torres, V.M. Gonzalez-Vizcarra, G.E. Medina-Basulto, M.F. Montano-Gomez, P. Mahadevan, V.H. Yaurima-Basaldua, 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. 14 (2015) 11848.

[2] G. Rincon, K.L. Weber, A.L. Eenennaam, B.L. Golden, J.F. Medrano, Hot topic: performance of bovine high-density genotyping platforms in Holsteins and Jerseys. J. Dairy Sci. 94 (2011) 6116.

[3] K. Wang, M. Li, D. Hadley, R. Liu, J. Glessner, S.F. Grant, et al., PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res. 17 (2007) 1665.

[4] S. Colella, C. Yau, J.M. Taylor, G. Mirza, H. Butler, P. Clouston, et al., QuantiSNP: an objective Bayes hidden-Markov model to detect and accurately map copy number variation using SNP genotyping data. Nucleic Acids Res. 35 (2007) 2013.

[5] S.J. Diskin, M. Li, C. Hou, S. Yang, J. Glessner, H. Hakonarson, et al., Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. Nucleic Acids Res. 36 (2008), e126.

[6] R. Redon, S. Ishikawa, K.R. Fitch, L. Feuk, G.H. Perry, T.D. Andrews, et al., Global variation in copy number in the human genome. Nature 444 (2006) 444.

[7] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, et al., Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat. Genet. 25 (2000) 25.

[8] M. Kanehisa, S. Goto, M. Furumichi, M. Tanabe, M. Hirakawa, KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res. 38 (2010) D355.

[9] W. Huang da, B.T. Sherman, R.A. Lempicki, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. 4 (2009) 44.

[10] K.J. Livak, T.D. Schmittgen, Analysis of relative gene expression data using real-time quantitative PCR and the 2(-delta delta C(T)) method. Methods 25 (2001) 402.

[11] L. Jiang, J. Jiang, J. Wang, X. Ding, J. Liu, Q. Zhang, Genome-wide identification of copy number variations in Chinese Holstein. PLoS One 7 (2012), e48732.