Genetics and Molecular Research

Genome-wide identification of copy number variations in Holstein cattle from Baja California, Mexico, using high-density SNP genotyping arrays

R. Salomón-Torres[1], V.M. González-Vizcarra[2], G.E. Medina-Basulto[2], M.F. Montaño-Gómez[2], P. Mahadevan[3], V.H. Yaurima-Basaldúa[4], C. Villa-Angulo[1] and R.Villa-Angulo[1]

[1]Laboratory of Bioinformatics and Biofotonics, Engineering Institute, Autonomous University of Baja California, Mexicali, Baja California, México
[2]Veterinary Science Research Institute, Autonomous University of Baja California, Mexicali, Baja California, México
[3]Department of Biology, University of Tampa, Tampa, FL, USA
[4]Sonora State University, San Luis Rio Colorado, Sonora, México

Corresponding author: R. Salomón-Torres
E-mail: ricardo.salomon@uabc.edu.mx

Genet. Mol. Res. 14 (4): 11848-11859 (2015)
Received April 16, 2015
Accepted July 24, 2015
Published October 2, 2015
DOI http://dx.doi.org/10.4238/2015.October.2.18

ABSTRACT. Copy number variations (CNVs) are an important source of genomic structural variation, and can be used as markers to investigate phenotypic and economic traits. CNVs also have functional effects on gene expression and can contribute to disease susceptibility in mammals. Currently, single nucleotide polymorphism genotyping arrays (SNP chips) are the technology of choice for identifying CNV variations. Microarray technologies have recently been used to study the bovine genome. The objective of the present study was to develop CNVs in Holstein cows from the Northwest of Mexico using the Affymetrix Axiom Genome-Wide BOS 1 Array, which assays 648,315 SNPs and provides a wide coverage for

genome-wide studies. We applied the two most widely used algorithms for the discovery of CNVs (PennCNV and QuantiSNP) and found 56 CNV regions (CNVRs) representing 0.33% of the bovine genome (8.46 Mb). These CNVRs ranged from 1.5 to 970.8 kb with an average length of 151 kb. They involved 103 genes and showed a 28% overlap with CNVRs already reported. Of the 56 CNVRs found, 20 were novel. In this study we present the first genomic analysis of CNVs in Mexican cattle using high-density SNP data. Our results provide a new reference basis for future genomic variation and association studies between CNVs and phenotypes, especially in Mexican cattle.

Key words: *Bos taurus* Holstein cattle; Copy number variation; SNP; Axiom Genome-Wide BOS 1 Array; PennCNV; QuantiSNP

INTRODUCTION

Genetic variations in mammalian genomes have different forms ranging from single nucleotide changes to large genomic regions. One of the most important differences is copy number variation (CNV), which is formally defined as a genomic alteration involving DNA segments ≥1 kb, and may correspond to deletions, duplications, insertions, inversions, or translocations (Conrad and Hurles, 2007). CNVs can involve large regions of DNA and they have been found to cover between 12-15% of the human genome (Conrad et al., 2010). CNVs can be associated with changes in gene structure, alteration of gene regulation, and exposure of recessive alleles among other effects (Bovine Genome Sequencing and Analysis Consortium et al., 2009; Zhang et al., 2009). Phenotypic variations caused by CNVs are the subject of study in many domestic animals, such as cows (Fadista et al., 2010; Liu et al., 2010; Seroussi et al., 2010; Hou et al., 2011, 2012a,b; Jiang et al., 2012, 2013; Cicconardi et al., 2013), horses (Doan et al., 2012), pigs (Wang et al., 2012, 2013), sheep (Liu et al., 2013), chickens (Crooijmans et al., 2013), rabbits (Fontanesi et al., 2012), and dogs (Alvarez and Akey, 2012). There are several methods to identify CNVs genome-wide: comparative genomic hybridization arrays (aCGH); single nucleotide polymorphism (SNP) genotyping arrays; and high-throughput (next-generation) sequencing (Wang et al., 2013). SNP genotyping arrays are the most widely used because of their advantages of simultaneously measuring total signal intensity (Log R ratio, LRR) and allelic intensity ratio (B allele frequency, BAF), and of identifying the number of DNA copies and copy neutral loss of heterozygosity (LOH) (Peiffer et al., 2006). In the case of the cattle community, research has focused mainly on the use of SNPs for genome-wide variation studies, resulting in the SNP genotyping arrays becoming the main research tool for genetic variation in cattle (Hou et al., 2011). One of the highest density commercially available genotyping arrays for cattle is the Axiom Genome-Wide BOS 1 (Affymetrix, Inc., USA), which assays 648,315 informative SNPs across the whole bovine genome. In this study, we performed CNV detection using data from this array.

Initially in CNV detection studies, researchers used a single algorithm for the detection; recently, however, it has become common practice to use two or three algorithms in order to minimize detection of false positives, by assuring that they match, confirm the same location, or overlap a known CNV (Sanders et al., 2011; Jiang et al., 2012; Cicconardi et al., 2013; Zhao et al., 2013). Currently, several commercial and public algorithms are available for detecting CNVs; the

most widely used are PennCNV (Wang et al., 2007) and QuantiSNP (Colella et al., 2007). Both algorithms infer CNVs from the SNP intensity LRR and BAF, implementing a hidden Markov model (HMM). QuantiSNP implements an Objective Bayes framework. It uses a resampling method to set some hyper-parameters in the priors, and applies the maximum marginal likelihood method to the training data to set other parameters. In contrast, PennCNV writes the emission probabilities of LRR and BAF into the same likelihood function, and estimates the model parameters by maximizing the likelihood of observing the training data. Subsequently, parameters in transition and emission probabilities are fixed in the HMM when analyzing different data (Xu et al., 2011).

Currently, there are no reports on genome-wide structural variations in livestock in Mexico. Obtaining this information will be very important for future research in Mexican cattle improvement. The objective of this study was to develop CNVs in Holstein cows from the Northwest of Mexico using high-density genotyping arrays, and to compare the results with previously reported CNVs from cattle in other countries. We hypothesized that due to the environmental conditions in the geographic region in which the cows live, they might be characterized by unique CNVs. Therefore, this study will provide a new basis for reference for future genomic variation and association studies between CNVs and phenotypes in Mexican cattle.

MATERIAL AND METHODS

Ethics statement

The Autonomous University of Baja California (UABC) animal care and use committee deemed that it was unnecessary to obtain ethical clearance for the study as all blood samples used for DNA extraction were collected under the directives on animal research of the Institute for Research in Veterinary Science UABC (IICV-UABC) abiding by Mexican laws on animal studies (NOM-003-ZOO-1994 and NOM-062-ZOO-1999).

Location of the study

This study was performed at the dairy farm belonging to the Veterinary Science Research Institute of the UABC, located 3.5 km along the San Felipe road from Mexicali, Baja California, at 32°24'27"N and 115°23'03"W. The altitude of the farm is 8 m, and a desert-like temperature with an annual average temperature of 22°C prevails (http://www.inegi.org.mx).

Characteristics of the cows

The samples were obtained 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. The sampled animals were not related in the last three generations.

DNA extraction and genotyping

The blood samples were collected on January 2014 by venipuncture of the coccygeal vein using vacutainer tubes (Vacutainer, Hemogar, USA) with EDTA anticoagulant.

DNA extraction and purification was performed using a QIAGEN kit (QIAamp DNA

Blood, QIAGEN, Germany). All DNA samples were analyzed by spectroscopy and agarosegel 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 rawdata of the SNP chip were submitted to the Gene Expression Omnibus (Feb 10, 2014) under accession No. GSE54813 (http://www.ncbi.nlm.nih.gov/geo).

Identification of cattle CNVs

In order to increase the level of reliability for CNV detection and to decrease the rate of false positives we applied two of the most accurate algorithms for CNV prediction: PennCNV (Wang et al., 2007) and QuantiSNP (Colella et al., 2007). The PennCNV algorithm requires an input of signal intensity (normalized by the LRR), BAF for each marker, and the distance between each SNP (base-pair positions). LRR and BAF were obtained using the guideline established in PennCNV-Affy Protocol for CNV detection in Affymetrix SNP arrays (http://www.openbioinformatics.org/penncnv/penncnv_tutorial_affy_gw6.html) and using Affymetrix Power Tools (APT) (http://www.affymetrix.com/estore/partners_programs/programs/developer/tools/powertools.affx). PennCNV includes an argument called *GCmodel*, which uses a regression model for adjusting the high GC content and recovers samples affected by ‘‘genomic waves” (Diskin et al., 2008). The GC model file for this study was generated by a Perlscript, which computes the GC content within 1 Mb around each marker (500 kb each side), and genomic waves were adjusted using *-gcmodel* option. PennCNV was executed using the *-test* option, considering that there was no relationship between the samples and pedigree/trio information was not included. It was applied only to the 29 autosomal chromosomes with *-lastchr29* option, using default values (standard deviation LRR of 0.30, BAF drift 0.01, and a waviness factor of 0.05). 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. 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. (2006).

CNV validation by real-time PCR (qPCR)

In order to confirm the accuracy of our prediction of CNVs, we used qPCR to validate seven CNVRs selected from the 56 detected in this study. From the selected CNVRs, three were single copy duplication (CNVRs 1, 16 and 35), two were deletions of one copy (CNVRs 2 and 11), and two were double copy duplications (CNVR21 and 55). 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/index.cgi?LINK_LOC=BlastHome). All primers were designed based on the reference sequence *Bos taurus* UMD3.1 at NCBI.

Reactions were performed intriplicate in a volume of 20 µL in a Light Cycler CFXConnect, (Roche, Switzerland) using the following reagents: 10 µL iTaq Universal SYBR Green supermix (BIO-RAD, USA), 1 µL DNA (about 50 ng), 1 µL 20 pM/µL for both primers (forward and reverse), 10 µL 2x Master Mix, and water. 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. Primer efficiency was tested for each primer pair on three points of dissolution.

We used the Basic Transcription Factor (*BTF3*) as a control gene for comparing the number of copies in each CNVR (Jiang et al., 2012; Cicconardi et al., 2013). The method of comparative cycle threshold ($2^{-\Delta\Delta Ct}$) was used to quantify the number of changes of the copies by comparing the ΔCt value [cycle threshold (Ct) of the region of the target minus the region Ct control] from the samples with CNV to a ΔCt of a calibrator without CNV (Livak and Schmittgen, 2001; D'Haene et al., 2010; Jiang et al., 2012).

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 (Jiang et al., 2013). 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) (Jiang et al., 2012). See Table 1 for the results of testing the qPCRs.

Table 1. Quantitative real-time PCR analysis of the seven selected CNVRs.

CNVR No.	Chr	Start	End	Type	Average normalized value	Value expected	Validated by PCR
1	1	1,971,535	2,003,454	Gain	2.7	3	Yes
2	1	34,613,754	34,617,353	Loss	1.2	1	Yes
11	4	33,547,395	33,556,611	Loss	1.7	1	No
16	7	97,442,762	97,444,260	Gain	3.1	3	Yes
21	11	39,886,500	39,918,145	Gain	2.2	3	No
35	14	22,348,562	22,393,497	Gain	2.7	3	Yes
55	28	22,706,652	22,725,331	Gain	4.2	4	Yes

Average normalized values of approximately one indicate a single copy loss; values around three indicate a three-copy gain, and around four indicate a four copies gain.

RESULTS AND DISCUSSION

Genome-wide detection of CNVs

In this study, we analyzed genotype data in 12 Mexican Holstein cows using an array with 648,315 SNPs, and applied PennCNV and QuantiSNP algorithms for CNV detection. PennCNV detected 155 CNVs, while QuantiSNP detected 302. The algorithms coincided for 77 CNVs, detected in the same position and the same sample (Figure 1). Initially, we termed these variants as putative CNVs. The 77 putative CNVs were localized across 22 autosomal chromosomes in the 12 samples. The average number of CNVs per sample was 6.41 and the average number of CNVs per chromosome was 3.5. We inspected the 77 CNVs for overlaps and defined 56 CNVRs, covering 0.33% (8.46 Mb) of the bovine genome (see **Table S1**). The CNVRs lengths ranged from 1.5 to 970.81 kb with an average size of 151.11 kb and a median of 51.6 kb. Figure 2 shows the distribution of CNVR sizes; the most abundant size range was from 10 to 50 kb (see Table 2 for the statistics of the CNVRs).

Among the CNVRs were 24 loss, 30 gains, and 2 loss-gain types of variation; the CVNRs

were not evenly distributed across the genome. Among chromosomes, the proportion chromosome length covered by CNVRs ranged from 0.03 to 3.8%; chromosome 12 had the highest proportion (3.8%) and chromosome 2 the lowest proportion (0.03%). The largest CNVR loss was 0.30 Mb on chromosome 15 of cow number three, while the largest CNVR gain was 0.97 Mb on chromosome 12 of cow number one. The chromosome with the largest number of CNVRs was chromosome 12, which had eight, while chromosomes 9, 21, 23, 25, 28, and 29 all had only one CNVR. From the 56 CNVRs, 20 were novel CNV regions (not reported in previous studies), representing 35.71% of all detected CNVRs. From these novel CNVRs, 10 were losses and 10 were gains. Figure 3 shows two examples of LRR and BAF values, corresponding to 2 loss type CNVs (deletions).

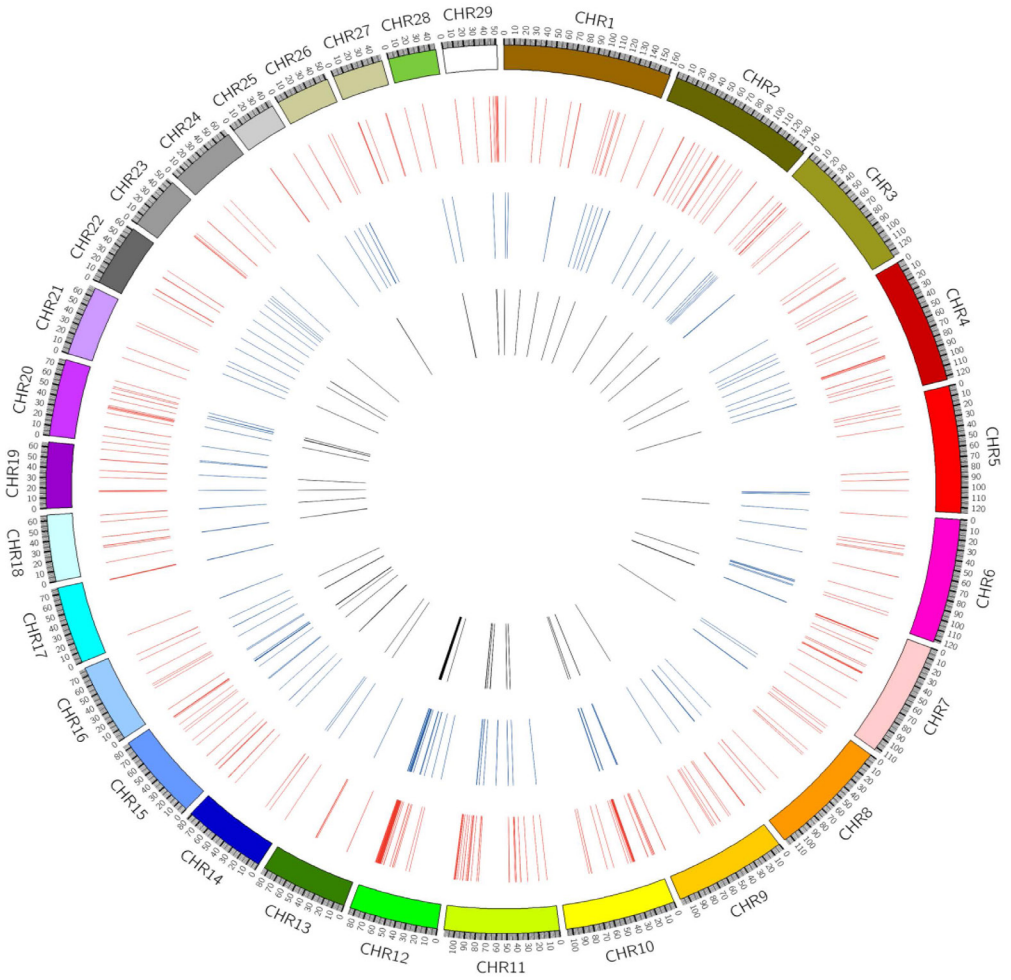

Figure 1. CNVs in autosomal chromosomes detected by PennCNV and QuantiSNP algorithms. The first inner circle (red lines) represents the 302 CNVs found by QuantiSNP. The second inner circle (blue lines) represents the 155 CNVs found by PennCNV. The third inner circle (black lines) represents the 77 CNVs in which both algorithms coincided.

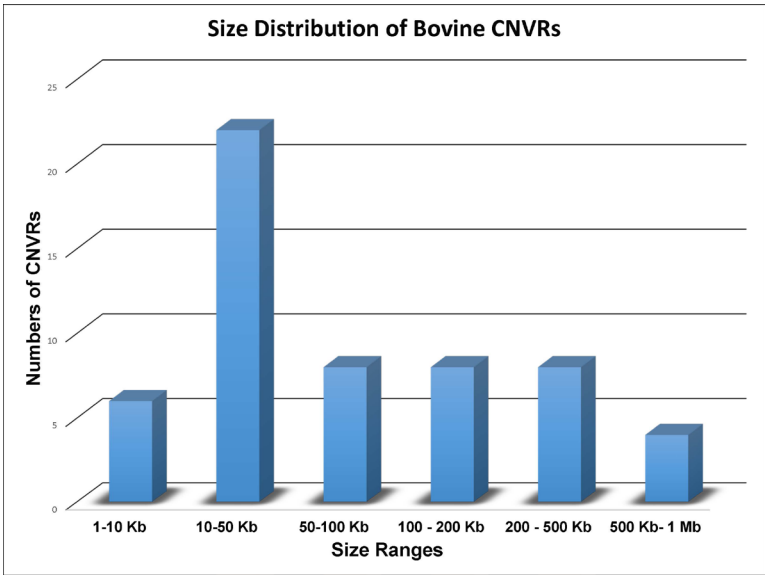

Figure 2. Size distribution of the detected CNVRs. The most frequent size category was 10-15 kb, while the least abundant were those larger than 500 kb.

Table 2. Characteristics and sizes (bp) of the CNV regions.

Type	CNVRs	Mean size	Median size	Size range	CNVR content	Sequence covered	% CNVR
Loss	25	71,458	37,694	343-302,434	1,789,724		0.071
Gain	29	207,314	59,268	1,498-970,815	6,012,125	2,512,082,506	0.239
Both	2	330,377	330,377	245,291-415,064	660,355		0.026
All	56	151,052	51,593	343-970,815	8,462,204		0.337

Gene content and functional analysis

We used the BioMart database (http://www.biomart.org, accessed May, 2014) to identify gene contents within the regions covered by CNVRs, and also used the RefGen database (http://refgene.com) to obtain a description of each gene affected. The CNVRs covered 103 genes, of which 96 encoded proteins, two were pseudogenes, three were snRNAs, and two were miRNAs. The genes were present within 37 CNVRs (66%); we could not find any gene annotations in the remaining 19 CNVRs (34%).

In order to analyze functional enrichment in the CNVRs, we searched the Gene Ontology (GO) database (Ashburner et al., 2000) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa et al., 2010). Both analyses were carried out using the bioinformatic tool DAVID (Huang et al., 2009). The GO analysis showed common gene terms among mammals, e.g., the sensory perception, cognition, olfactory receptor, neurological system process, G-protein coupled receptor protein signaling pathway, and cell surface receptor linked signal transduction. KEGG pathway analysis showed that the genes were mainly represented in the pathway of olfactory transduction (see **Table S2**).

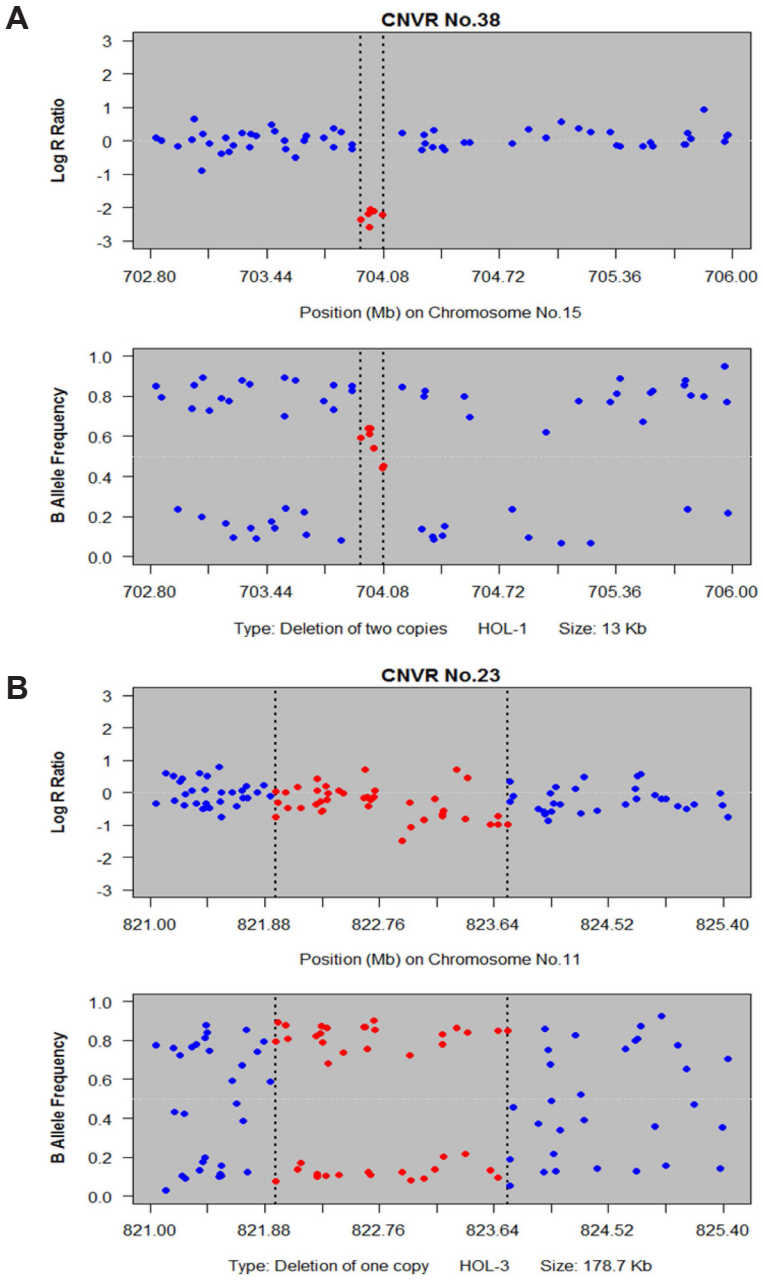

Figure 3. Log R ratio (LRR) and B allele frequency (BAF) plots of two copy number variation regions (CNVRs). The combination of BAF and LRR plots are used to generate CNV calls. Each pair of plots is for a different sample-chromosome combination and each point is a SNP. Points inside a CNVR are drawn in red, and outside are drawn in blue. The vertical dotted lines indicate the limit of the region. **A.** Low values of LRR (less than -1) and grouping of BAF values around 0.5 indicates a deletion of two copies in a region of chromosome 15. **B.** Some low values (below -1) in LRR and no values in the 0.5 cluster indicate a single copy deletion in a region of chromosome 11.

We compared the CNVRs identified in this study with QTLs from the bovine QTLdb (Hu et al., 2013) (http://www.animalgenome.org/cgi-bin/QTLdb/BT/index, accessed Sep. 2014) and found 34 overlaps. These QTLs are associated with a wide range of traits including body length, milk protein, milk fat, height, meat color, and some disease susceptibility traits, such as dystocia and clinical mastitis.

Finally, in order to confirm the accuracy of our predicted CNVRs, we selected seven of the 56 CNVRs and validated them using qPCR. Three of the seven CNVRs were single copy duplications (CNVRs 1, 16 and 35), two were deletions of one copy (CNVRs 2 and 11), and two were double copy duplications (CNVRs 21 and 55). 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/index.cgi?LINK_LOC=BlastHome). All primers were designed based on the reference sequence (*B. taurus* UMD3.1) from NCBI. We confirmed five CNVRs from the seven selected using the method of comparative cycle threshold ($2^{-\Delta\Delta Ct}$) to quantify the number of changes of the copies by comparing ΔCt values.

Analyzing overlapping CNVs from the algorithms, we found 49.67% overlap of PennCNV with QuantiSNP, but only 25.49% overlap for the reciprocal comparison. Next, we compared our results with reported CNVs detected with different technologies such as comparative genomic hybridization (aCGH) array (Fadista et al., 2010; Liu et al., 2010), next generation sequencing (Bickhart et al., 2012), and SNP Beadchip 54 kb (Hou et al., 2011, 2012b; Cicconardi et al., 2013) and high density SNP (777 kb) assay (Hou et al., 2012a; Jiang et al., 2013). Since the last studies used the Btau4.0 bovine genome assembly (all the rest including our study used UMD3.1 assembly), we translated their data to the UMD3.1 assembly using the UCSC liftOver tool (Rhead et al., 2010) and the NCBI genome remapping service (http://www.ncbi.nlm.nih.gov/genome/tools/remap, accessed Sep, 2014) and finally, to be consistent, we compared only CNVRs from autosomal chromosomes.

For aCGH-based studies, our detected CVNRs overlap with those reported by Fadista et al. (2010) by 16.07% (9 CNVRs) with a total overlap length of 1.83 Mb (21.63%). For those reported by Liu et al. (2010) the overlap was 25% (14 CNVRs) with a total overlap length of 3.93 Mb (46.45%). Bickhart et al. (2012) used next-generation sequencing to identify CNVs; our data here showed an overlap of 19.64% (11CNVRs) with a total overlap length of 2.82 Mb (33.33%) compared to the published data. Hou et al. (2011) reported an analysis based on SNP Beadchip 54 kb, and their data showed an overlap of 26.78% (15 CNVRs) with a total overlap length of 3.99 Mb (47.16%) compared to the present study. However, in a later study (Hou et al., 2012b), the overlap was 19.64% (11 CNVRs) with a total overlap length of 2.37 Mb (28.01%), while the data of Jiang et al. (2012) indicate an overlap of 3.19% (3 CNVRs) with a total overlap length of 0.66 Mb (7.80%), and those of Cicconardi et al. (2013) indicate an overlap of 26.78% (11 CNVRs) with a total overlap length of 1.20 Mb (14.18%). The data from high density SNP assays (777 kb) (Hou et al., 2012a) showed a 75% overlap (42 CNVRs) representing 7.59 Mb (89.72%) with our data. The data reported by Jiang et al. (2013) showed an overlap of 41.07% (23 CNVRs) representing an overlap length of 1.31 Mb (15.48%). Table 3 presents the results of the comparison between CNVRs detected in this study and those found in other studies.

The comparison of our results with other studies showed notable differences, mainly due to the use of different algorithms, different technology platforms, and different criteria in parameter adjustment. An additional factor that might contribute to the differences was the geographic region in which the animals lived. Sample sizes did not appear to influence the results since Jiang et al. (2012) and Cicconardi et al. (2013) used 2047 and 2654 animals, respectively; our data showed

matches of 3.19 and 26.78%, respectively. Hou et al. (2011) used 521 animals and their data showed 26.78% overlaps with those here; Fadista et al. (2010) sampled only 20 animals with aCGH technology and their data showed 16.07% overlaps with those here. Bickhart et al. (2012) used next-generation sequencing on a sample of six animals and their data showed 19.64% overlaping CNVs with those here. The largest number of overlaps between our results and those reported previously came with studies using high-density SNP genotyping technologies: 75% with Hu et al. (2013) and 41.07% with Jiang et al. (2013). One reasons for this is that high-density SNPs can detect smaller CNVs than is possible with lower density SNPs (Hou et al., 2011, 2012a; Jiang et al., 2012, 2013): by comparison, the 54k SNP density analysis yielded 682 and 94 CNVs in two studies, whereas use of 777k SNP density analysis yielded 3346 and 358 CNVs, respectively.

Table 3. Comparison between CNVRs detected in this study and those of other studies.

Study	Methods	Algorithm used	Other studies				CNVR overlaps with those identified here			
			Samples	Number of breeds	CNVR	Length (Mb)	Count	Pecentage of count (%)	Total length (Mb)	Percentage of length (%)
Fadista et al. (2010)	aCGH	seg-MNT	20	4	233	14.25	9	16.07	1.83	21.63
Liu et al. (2010)	aCGH	GIM; SW-ARRAY	90	17	142	35.83	14	25.00	3.93	46.45
Hou et al. (2011)	SNP chip (54 kb)	PennCNV	521	21	672	152.68	15	26.78	3.99	47.16
Bickhart et al. (2012)	Next-generation Seq	mrFAST	6	3	978	52.30	11	19.64	2.82	33.33
Hou et al. (2012a)	SNP chip (777 kb)	PennCNV	674	27	3438	146.90	42	75.00	7.59	89.72
Hou et al. (2012b)	SNP chip (54 kb)	PennCNV	472	1	462	86.05	11	19.64	2.37	28.01
Jiang et al. (2012)	SNP chip (54 kb)	PennCNV; cnvPartition; GADA	2047	1	94	21.30	3	3.19	0.66	7.80
Cicconardiet al. (2013)	SNP chip (54 kb)	PennCNV; QuantiSNP	2654	5	394	515.90	11	26.78	1.20	14.18
Jiang et al. (2013)	SNP chip (777 kb)	PennCNV	96	1	358	34.45	23	41.07	1.31	15.48
This study	SNP chip (648 kb)	PennCNV; QuantiSNP	12	1	56	8.46				

The GC content within each CNVR varied from 33.9 to 57.95%, in agreement with the proposal that CNVs are regions of high GC content (Fadista et al., 2010). CNVs are also associated with segmental duplications (SDs), which are defined as DNA sequences with lengths ≥1 kb with at least 90% sequence identity. Comparing our results with previously reported SDs (Liu et al., 2009) identified 12.5% overlaps.

CONCLUSIONS

In a genome-wide CNV study of 12 Holstein cows from the Northwest of Mexico, we identified 56 CNVRs distributed across the 29 autosomal chromosomes; 20 of these CNVRs were novel. Validation of seven of the putative CNVRs by qPCR showed that five were indeed CNVRs. High-density data permitted us to achieve greater accuracy in the identification of CNVRs and their candidate genes. Our study provides a new basis of reference for future investigations of the associations between CNVs and phenotypes, especially in Mexican cattle.

Conflicts of interest

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

We wish to thank MS Cristian Miguez and Fernando Nodari from Affymetrix, Inc. for their

support, Dr. Kai Wai from the University of Pennsylvania for resolving our doubts about the operation of PennCNV, Dr. Wang Jiying from the China Agricultural University for sharing his experience and knowledge in this type of study. We are grateful to The National Council for Science and Technology of Mexico (CONACyT) for supporting a scholarship for doctoral studies to R. Salomón-Torres, and to PIFI 2014 for supporting this research.

Supplementary material

REFERENCES

Alvarez CE and Akey JM (2012). Copy number variation in the domestic dog. *Mamm. Genome* 23: 144-163.

Ashburner M, Ball CA, Blake JA, Botstein D, et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25: 25-29.

Bickhart DM, Hou Y, Schroeder SG, Alkan C, et al. (2012). Copy number variation of individual cattle genomes using next-generation sequencing. *Genome Res.* 22: 778-790.

Bovine Genome Sequencing and Analysis Consortium, Elsik CG, Tellam RL, Worley KC, et al. (2009). The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science* 324: 522-528.

Cicconardi F, Chillemi G, Tramontano A, Marchitelli C, et al. (2013). Massive screening of copy number population-scale variation in *Bos taurus* genome. *BMC Genomics* 14: 124.

Colella S, Yau C, Taylor JM, Mirza G, et al. (2007). QuantiSNP: an objective Bayes hidden-Markov model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res.* 35: 2013-2025.

Conrad DF and Hurles ME (2007). The population genetics of structural variation. *Nat. Genet.* 39: S30-36.

Conrad DF, Pinto D, Redon R, Feuk L, et al. (2010). Origins and functional impact of copy number variation in the human genome. *Nature* 464: 704-712.

Crooijmans RP, Fife MS, Fitzgerald TW, Strickland S, et al. (2013). Large scale variation in DNA copy number in chicken breeds. *BMC Genomics* 14: 398.

Diskin SJ, Li M, Hou C, Yang S, et al. (2008). Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucleic Acids Res.* 36: e126.

D'Haene B, Vandesompele J and Hellemans J (2010). Accurate and objective copy number profiling using real-time quantitative PCR. *Methods* 50: 262-270.

Doan R, Cohen N, Harrington J, Veazey K, et al. (2012). Identification of copy number variants in horses. *Genome Res.* 22: 899-907.

Fadista J, Thomsen B, Holm LE and Bendixen C (2010). Copy number variation in the bovine genome. *BMC Genomics* 11: 284.

Fontanesi L, Martelli PL, Scotti E, Russo V, et al. (2012). Exploring copy number variation in the rabbit (*Oryctolagus cuniculus*) genome by array comparative genome hybridization. *Genomics* 100: 245-251.

Hou Y, Liu GE, Bickhart DM, Cardone MF, et al. (2011). Genomic characteristics of cattle copy number variations. *BMC Genomics* 12: 127.

Hou Y, Bickhart DM, Hvinden ML, Li C, et al. (2012a). Fine mapping of copy number variations on two cattle genome assemblies using high density SNP array. *BMC Genomics* 13: 376.

Hou Y, Liu GE, Bickhart DM, Matukumalli LK, et al. (2012b). Genomic regions showing copy number variations associate with resistance or susceptibility to gastrointestinal nematodes in Angus cattle. *Funct. Integr. Genomics* 12: 81-92.

Hu ZL, Park CA, Wu XL and Reecy JM (2013). Animal QTLdb: an improved database tool for livestock animal QTL/association data dissemination in the post-genome era. *Nucleic Acids Res.* 41: D871-D879.

Huang da W, Sherman BT and Lempicki RA (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4: 44-57.

Jiang L, Jiang J, Wang J, Ding X, et al. (2012). Genome-wide identification of copy number variations in Chinese Holstein. *PLoS One* 7: e48732.

Jiang L, Jiang J, Yang J, Liu X, et al. (2013). Genome-wide detection of copy number variations using high-density SNP genotyping platforms in Holsteins. *BMC Genomics* 14: 131.

Kanehisa M, Goto S, Furumichi M and Tanabe M (2010). KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.* 38: D355-D360.

Liu GE, Ventura M, Cellamare A, Chen L, et al. (2009). Analysis of recent segmental duplications in the bovine genome. *BMC Genomics* 10: 571.

Liu GE, Hou Y, Zhu B, Cardone MF, et al. (2010). Analysis of copy number variations among diverse cattle breeds. *Genome Res.* 20: 693-703.

Liu J, Zhang L, Xu L, Ren H, et al. (2013). Analysis of copy number variations in the sheep genome using 50K SNP BeadChip array. *BMC Genomics* 14: 229.

Livak KJ and Schmittgen TD (2001). Analysis of relative gene expression data using real-time quantitative PCR and the $2^{\Delta\Delta Ct}$ method. *Methods* 25: 402-408.

Peiffer DA, Le JM, Steemers FJ, Chang W, et al. (2006). High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Res.* 16: 1136-1148.

Redon R, Ishikawa S, Fitch KR, Feuk L, et al. (2006) Global variation in copy number in the human genome. *Nature* 444: 444-454.

Rhead B, Karolchik D, Kuhn RM, Hinrichs AS, et al. (2010). The UCSC Genome Browser database: update 2010. *Nucleic Acids Res.* 38: D613-D619.

Sanders SJ, Ercan-Sencicek AG, Hus V, Luo R, et al. (2011). Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron* 70: 863-885.

Seroussi E, Glick G, Shirak A, Yakobson E, et al. (2010). Analysis of copy loss and gain variations in Holstein cattle autosomes using BeadChip SNPs. *BMC Genomics* 11: 673.

Wang J, Jiang J, Fu W, Jiang L, et al. (2012). A genome-wide detection of copy number variations using SNP genotyping arrays in swine. *BMC Genomics* 13: 273.

Wang J, Wang H, Jiang J, Kang H, et al. (2013). Identification of genome-wide copy number variations among diverse pig breeds using SNP genotyping arrays. *PLoS One* 8: e68683.

Wang K, Li M, Hadley D, Liu R, et al. (2007). PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* 17: 1665-1674.

Xu Y, Peng B, Fu Y and Amos CI (2011). Genome-wide algorithm for detecting CNV associations with diseases. *BMC Bioinformatics* 12: 331.

Zhang F, Gu W, Hurles ME and Lupski JR (2009). Copy number variation in human health, disease, and evolution. *Annu. Rev. Genomics Hum. Genet.* 10: 451-481.

Zhao L, Bracken MB and Dewan AT (2013). Genome-wide association study of pre-eclampsia detects novel maternal single nucleotide polymorphisms and copy-number variants in subsets of the hyperglycemia and adverse pregnancy outcome (HAPO) study cohort. *Ann. Hum. Genet*. 77: 277-287.