J. Dairy Sci. 99:6371–6380
http://dx.doi.org/10.3168/jds.2015-10487


# Associated effects of copy number variants on economically important traits in Spanish Holstein dairy cattle

**Neila Ben Sassi,* Óscar González-Recio,* Raquel de Paz-del Río,* Silvia T. Rodríguez-Ramilo,*† and Ana I. Fernández*[1]**
*Departamento Mejora Genética Animal, Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria-INIA. Ctra La Coruña km 7.5, 28040 Madrid, Spain
†Departamento Técnico Conafe, Ctra. Andalucía, Km. 23.6, 28340 Valdemoro, Madrid, Spain

## ABSTRACT

Copy number variants (CNV) are structural variants consisting of duplications or deletions of genomic fragments longer than 1 kb that present variability in the population and are heritable. The objective of this study was to identify CNV regions (CNVR) associated with 7 economically important traits (production, functional, and type traits) in Holstein cattle: fat yield, protein yield, somatic cell count, days open, stature, foot angle, and udder depth. Copy number variants were detected by using deep-sequencing data from 10 sequenced bulls and the Bovine SNP chip array hybridization signals. To reduce the number of false-positive calls, only CNV identified by both sequencing and Bovine SNP chip assays were kept in the final data set. This resulted in 823 CNVR. After filtering by minor allele frequency >0.01, a total of 90 CNVR appeared segregating in the bulls that had phenotypic data. Linear and quadratic CNVR effects were estimated using Bayesian approaches. A total of 15 CNVR were associated with the traits included in the analysis. One CNVR was associated with fat and protein yield, another 1 with fat yield, 3 with stature, 1 with foot angle, 7 with udder depth, and only 1 with days open. Among the genes located within these regions, highlighted were the *MTHFSD* gene that belongs to the folate metabolism genes, which play critical roles in regulating milk protein synthesis; the *SNRPE* gene that is related to several morphological pathologies; and the *NF1* gene, which is associated with potential effects on fertility traits. The results obtained in the current study revealed that these CNVR segregate in the Holstein population, and therefore some potential exists to increase the frequencies of the favorable alleles in the population after independent validation of results in this study. However, genetic variance explained by the variants reported in this study was small.
**Key words:** copy number variants, dairy cattle, whole-genome sequencing, Bovine SNP50 BeadChip

## INTRODUCTION

Currently, most of the genetic and genome-wide association studies have used SNP as genetic variants to test associations with complex traits (Manolio et al., 2009). However, genome-wide association studies have not succeeded in recovering all of the heritability estimated with traditional methods. One of the hypotheses for this missing heritability is the existence of other genetic variants in the genome, such as copy number variants (**CNV**). These are genomic fragments longer than 1 kb in size, display variable copy number across individuals, and are heritable. These variants may potentially have some effect on gene structure or gene dosage (increases or decreases); they may also have a global influence on the transcriptome (e.g., alterations in gene regulation or exposition of recessive alleles on phenotypic variability; Henrichsen et al., 2009; Zhang et al., 2009). Although CNV have been widely associated with complex traits and phenotypic variability in humans (Frank et al., 2007; Zhang et al., 2009; Stefansson et al., 2014), few studies have been focused on CNV associations to phenotypic traits in domestic species, probably due to lack of specific tools to genotype CNV in these species (Clop et al., 2012; Bickhart and Liu, 2014). Some examples are the CNV containing the *ED1* gene, which is responsible for ectodermal dysplasia in cattle (Drögemüller et al., 2001), *ASIP* gene associated with coat colors in ruminants (Fontanesi et al., 2009), or the CNV containing the *KIT* gene, responsible for different patterns of coat color in pigs (Marklund et al., 1998) and likely responsible for gonadal hypoplasia in cattle (Venhoranta et al., 2013).

Recently developed sequencing and genotyping technologies are expected to further contribute to fine map-

Received October 5, 2015.
Accepted April 15, 2016.
[1]Corresponding author: avila@inia.es

ping and detection of CNV in livestock studies. In dairy cattle, CNV were identified mainly using the commercial high- and medium-density SNP bead-arrays due to the lack of specific technology to genotype CNV genome-wide before next-generation sequencing was available (Stothard et al., 2011; Bickhart et al., 2012; Jiang et al., 2013). Nonetheless, many CNV regions have already been detected in Holstein cattle using massive genotyping from SNP chips (Bae et al., 2010; Zhan et al., 2011; Hou et al., 2012; Jiang et al., 2012, 2013); however, few studies have reported association of these CNV with economically important traits (Xu et al., 2013; Yue et al., 2013). Copy number variation of the *PRAMEY* gene is associated with male fertility in Holstein cattle (Yue et al., 2013), and copy number variation of the *MICAL-L2* gene is associated with BW, body height, and body length of Chinese cattle breeds. Although Calus et al. (2010) suggested that using CNV could assist in explaining part of the phenotypic variation on simulated data, to the best of our knowledge the percentage of missing heritability captured by CNV in real data have not previously been reported, except in the study conducted by Seroussi et al. (2010), and no CNV have been used in whole-genome prediction.

Association analyses on CNV pose 2 main challenges: determining the number of copy variants is not straightforward with current technology, and the number of copies may not have a linear effect on the phenotypic expression of the trait. For instance, having 2 copies may translate in twice the expression compared with 1 copy; however, having more than 2 copies might imply a proportional decrease or increase in the expression, but it could also result in saturation in the biological dose of proteins, with no further change in the phenotypic expression. Further, suppression of the CNV (zero copies) might imply important biological changes, with no linearity compared with having 1 or 2 copies. The aim of the current study was to discover CNV regions (**CNVR**) using sequencing and bovine SNP genotyping data to identify association with economically important traits in the Spanish Holstein population.

## MATERIALS AND METHODS

### Phenotypic and Genotypic Data

Phenotypes for 7 traits from 1,371 Holstein sires, born between 1962 and 2006, were provided by the Spanish Holstein Association (CONAFE), as well as respective SNP genotypes from the Bovine SNP50 BeadChip (Illumina, San Diego, CA). Traits in this study were fat and protein yields as production traits, SCC and days open as functional traits, and stature, foot angle, and udder depth as morphological traits.

### CNV Using Whole-Genome Sequence Data

Ten Holstein sires with genotypes for the Bovine SNP50 BeadChip, were sequenced with the Illumina Hiseq2000 platform (Fasteris SA, Plan-les-Ouates, Switzerland) at Centro Nacional de Análisis Genómico (Barcelona, Spain) generating paired-end reads of 100 bp with coverage per animal ranging between 7.5× and 10.3×. Copy number variants for these sires were detected using deep-sequencing data with the control-FREE copy number caller (Boeva et al., 2011, 2012). Reads were aligned to the UMD3.1 version of the bovine genome using BWA (Li and Durbin, 2009), allowing 7 mismatches and filtering by mapping quality of 20. The mapped paired-end read files were then used in the control-FREE copy number caller, which employs a sliding window (1 kb) approach to calculate read count in nonoverlapping windows. The program calculates GC content in the same set of windows and performs normalization by GC content. Then, segmentation is applied on the resulting normalized profile. This was followed by the analysis of predicted regions of gains and losses to assign copy numbers to these regions (Boeva et al., 2011). The program was run using the default parameters without control sample. It is known that even after correction for both GC bias and mappability, depth of-coverage methods are vulnerable to false positives and cross-sample calling is required to reduce this effect (Klambauer et al., 2012).

### CNV Genotypes from Bovine SNP50 BeadChip

Raw SNP data were visualized using the GenomeStudio software (Illumina). Intensity signals were exported at each SNP position using the parameters of total signal intensity, log R ratio, allelic intensity ratio, and B allele frequency. Then, the detection of CNV was implemented with the PennCNV software (Wang et al., 2007). The SNP annotations, chromosome, and physical position were from the UMD3.1 build version of the bovine genome. Additionally, the allelic intensity ratio of the whole population (population B allele frequency) was also calculated from the signal files using the compile_pbf.pl command in PennCNV. Wave adjustment for GC content was performed applying the gcmodel option of PennCNV software, as GC-rich and -poor regions of the genome differ during amplification and labeling reactions. The bovine gcmodel was generated by calculating the GC content of a 1-Mb genomic region