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 re-

sults 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.

¹Corresponding 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 $7.5\times$ and $10.3\times$. 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

surrounding each SNP (500 kb each side) according to UMD3.1 annotation. Prior to CNV identification, a quality control of intensity signals was performed to reduce false-positive CNV that can originate from poor-quality DNA. Samples with a standard deviation of log R ratio >0.4, B allele frequency drift >0.1, or GC wave factor (after genomic waves were corrected by GC content) >0.1 were excluded for further analyses. A total of 176 samples did not pass the filtering criteria and were discarded, leading to 1,195 individuals left for the analyses. The identification of CNV was carried out using the individual option of PennCNV. The CNV were filtered for those (1) containing 3 or more consecutive SNP and (2) detected in at least 5 animals.

Here, we have used the simplest way to call CNV under the sequence-based control-FREE copy number caller analysis, but a cross-sample was conducted calling using the CNV call from the Bovine SNP chip. We considered as actual CNVR those coming from the next-generation sequencing (NGS) call but also identified as the same genotype by the array (not in the specific copy number, but considering the insertion or deletion state). If larger array-based CNV calls included several smaller intersecting sequence-based CNV, then the array-based CNV was split into corresponding sequence-based CNV. In such a case, the array-based CNV genotype for each animal was assigned to the corresponding intersecting CNVR. The resulting CNVR genotypes from the Bovine SNP chip mapping the actual CNVR detected from the sequence data were used subsequently as covariates in the association analyses.

Association Analyses

Deregressed proofs from progeny testing were used as phenotypes for the bulls (n ; 1,173) included in the analyses. The mean and variance of each trait are summarized in Table 1. A Bayesian regression approach with both linear and quadratic terms for the CNVR effect, one CNV at a time, was used for the association analyses. The linear model included a correction by the polygenic effect as follows:

$$\mathbf{y} = \mathbf{1}'\mu + \sum_{j=1}^{10} \mathbf{a}_j \mathbf{p}\mathbf{c}_j + \beta \mathbf{X} + \mathbf{e},$$

where $\mathbf{y} = \{y_i\}$ is a vector with the phenotypic observations ($i = 1 \dots n$), μ is the population mean, $\mathbf{1}'$ is a vector of ones, $\mathbf{p}\mathbf{c}_j$ ($j = 1, \dots, 10$) were the 10 first principal components of the genomic relationship matrix (Yang et al., 2010) and $\mathbf{a}$ was the respective coefficient, and $\mathbf{e}$ is the vector of residuals assumed to be

Table 1. Mean and SD of the deregressed proofs for the 1,195 sires included in the analyses

Trait	Mean	SD
Fat yield (kg)	3.03	25.22
Protein yield (kg)	5.17	22.27
SCC	99.04	12.67
Stature (score)	0.33	1.25
Foot angle (score)	0.20	1.27
Udder depth (score)	0.26	1.17

distributed as $N(0, \sigma_e^2)$, with σ_e^2 being the residual variance. The $\beta \mathbf{X}$ term in the linear model was $b_1 \mathbf{x}$, with $\mathbf{x}$ being the vector of corresponding genotypes for the CNVR of each sire coded as 0, 1, or 2 for deletion (less than 2 copies), normal (2 copies), and duplication (3 or more copies), respectively, and b_1 being its corresponding effect.

Then, in the quadratic model, $\beta \mathbf{X} = [b_1 \quad b_2] \begin{bmatrix} \mathbf{x} \\ \mathbf{x}^2 \end{bmatrix}$, with $\mathbf{x}^2$ being the square of the CNVR codification and b_2 the corresponding quadratic effect.

The hierarchical Bayesian model implemented to estimate the CNVR effects was

$$\text{Likelihood: } p(\mathbf{y} | \mu, \mathbf{a}, \beta, \sigma_e^2) = \prod_{i=1}^n N\left(y_i | \mathbf{1}'\mu + \sum_{j=1}^{10} a_{ij} p c_{ij} + \beta \mathbf{x}_i, \sigma_e^2\right), \text{ and}$$

$$\begin{aligned} \text{Prior: } p(\mu, \beta, \sigma_e^2) &= p(\mu | \sigma_e^2) p(\mathbf{a} | \sigma_e^2) p(\beta | \sigma_e^2) p(\sigma_e^2) \\ &= N(\mu | 0, \sigma_\mu^2) N(\mathbf{a} | 0, \mathbf{I} \sigma_a^2) N(\beta | 0, \sigma_\beta^2) \times \\ &\quad \chi^{-2}(\sigma_\mu^2 | d_\mu, S_\mu) \chi^{-2}(\sigma_a^2 | d_a, S_a) \chi^{-2}(\sigma_\beta^2 | d_\beta, S_\beta) \chi^{-2}(\sigma_e^2 | d_e, S_e). \end{aligned}$$

Conditional distributions of these effects were obtained using Monte-Carlo Markov chains, as

$$\begin{aligned} p(\mu | \text{else}) &\propto \left[\prod_{i=1}^n N\left(y_i | \mathbf{1}'\mu + \sum_{j=1}^{10} a_{ij} p c_{ij} + \beta \mathbf{x}_i, \sigma_e^2\right) \right] \\ N(\mu | 0, \sigma_\mu^2) &\propto \left[\prod_{i=1}^n N(y_i | \mu, \sigma_e^2) \right] N(\mu | 0, \sigma_\mu^2). \end{aligned}$$

This is recognized as the kernel of a normal distribution with mean $(n\sigma_e^{-2} + \sigma_\mu^{-2})^{-1} \sum_{i=1}^n (y_i - \beta \mathbf{x}_i) \sigma_e^{-2}$, and variance $(n\sigma_e^{-2} + \sigma_\mu^{-2})^{-1}$.

Then, assuming $\mathbf{I}$ to be an identity matrix of respective order,

$$p(\mathbf{a} | \text{else}) \propto \left[\prod_{i=1}^n N\left(y_i | \mathbf{1}'\mu + \sum_{j=1}^{10} a_{ij}pc_{ij} + \beta\mathbf{x}_i, \sigma_e^2\right) \right] N(\mathbf{a} | 0, \mathbf{I}\sigma_a^2) \\ \propto \left[\prod_{i=1}^n N\left(y_i^* | a_{ij}pc_{ij}, \sigma_e^2\right) \right] N(\mathbf{a} | 0, \mathbf{I}\sigma_a^2).$$

This is recognized as the kernel of a multinormal distribution. The distribution of each a_j would have mean and variance equal to the solution of

$$\left(\sigma_e^{-2} \sum_{i=1}^n \mathbf{p}\mathbf{c}_{ij}^2 + \sigma_a^{-2} \right)^{-1} \hat{a}_j = \sigma_e^{-2} \sum_{i=1}^n pc_{ij}y_i^*,$$

where

$$y_i^* = y_i - \mu - \sum_{k \neq j} a_{ik}pc_{ik} - \beta\mathbf{x}$$

and

$$p(\beta | \text{else}) \propto \left[\prod_{i=1}^n N\left(y_i | \mathbf{1}'\mu + \sum_{j=1}^{10} a_{ij}pc_{ij} + \beta\mathbf{x}_i, \sigma_e^2\right) \right] N(\beta | 0, \mathbf{I}\sigma_\beta^2) \\ \propto \left[\prod_{i=1}^n N\left(y_i^{**} | \beta\mathbf{x}_i, \sigma_\beta^2\right) \right] N(\beta | 0, \mathbf{I}\sigma_\beta^2).$$

This is recognized as the kernel of a multinormal distribution. The distribution of each β_j would have mean and variance equal to the solution of

$$\left(\sigma_e^{-2} \sum_{i=1}^n \mathbf{x}_{ij}^2 + \sigma_\beta^{-2} \right)^{-1} \hat{\beta}_j = \sigma_e^{-2} \sum_{i=1}^n x_{ij}y_i^{**},$$

where

$$y_i^{**} = y_i - \mu - \sum_{j=1}^{10} a_{ij}pc_{ij} - \beta_{k \neq m} \mathbf{x}_{k \neq m},$$

where $k \neq m$ stands for all other coefficients except the m on the equation to solve. In practice, σ_β^{-2} and σ_a^{-2} can be set large enough to assign an effectively flat prior before these coefficients:

$$p(\sigma_e^2 | \text{else}) \propto \left[\prod_{i=1}^n N\left(y_i | \mathbf{1}'\mu + \sum_{j=1}^{10} a_{ij}pc_{ij} + \beta\mathbf{x}_i, \sigma_e^2\right) \right] \chi^{-2}(\sigma_e^2 | d_e, S_e) \\ \propto \chi^{-2}(\sigma_e^2 | d.f. = d_e + n, S = S_e + \mathbf{e}'\mathbf{e}),$$

where $\mathbf{e} = \mathbf{y} - \mathbf{1}'\mu - \sum_{j=1}^{10} \mathbf{a}_j \mathbf{p}\mathbf{c}_j - \beta\mathbf{X}$, *d.f.* is the degrees of freedom, and S is the scale parameter of the inverted chi-squared distribution, with d_e and S_e are the respective hyperparameters of the a priori distribution.

These models were implemented in R software (www.r-project.org) using Markov chain Monte Carlo procedures with 50,000 Gibbs sampling cycles discarding the first 10,000 as burn-in. We rejected the null hypothesis (no association between the CNVR and the trait) if the posterior mean for the coefficient estimates did not fall within the 90% highest probability interval of the effect after a permutation test (i.e., the CNV effect estimate did not fall within the 90% confidence interval of the effect estimates after permutation). For the permutation test, reshuffled phenotypes at each cycle were used with 50,000 Gibbs sampling cycles and 10,000 burn-in iterations (Che and Xu, 2010). The CNVR copies were coded as 0, 1, and 2 for deletion (less than 2 copies), normal (2 copies), and duplication (3 or more copies), respectively, because the methodologies we used for CNVR genotyping (sequencing or SNP genotyping) are not accurate enough as to determine the exact copy number.

Gene Content Analysis

The gene content of the significant CNVR was examined using the bovine genome annotation in BioMart tool of Ensembl (www.ensembl.org/biomart/martview/980da003f03ba5eeca5c658caf5d38f) using the Ensembl genes 75 database. The National Center for Biotechnology Information database (www.ncbi.nlm.nih.gov/) and PubMed were employed to investigate the potential biological relation of the genes located within the CNVR and the analyzed traits to propose candidate genes that underlie the detected associations.

RESULTS AND DISCUSSION

CNV Discovery

In total, 1,737 CNV were detected in the 10 sequenced sires. 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 (Supplemental Table S1; <http://dx.doi.org/10.3168/jds.2015-10487>). Large differences in number and even more in sizes of the CNV identified using genome sequencing and Bovine SNP chip were found (Supplementary Figure S1; <http://dx.doi.org/10.3168/jds.2015-10487>), probably due to several limitations of

Figure 1. Distribution across autosomes of the copy number variant regions (CNVR) identified in the analyzed population. MAF = minor allele frequency.

the SNP chip for this purpose. First, the Bovine SNP50 BeadChip was not specifically designed for CNV detection and it does not cover the entire bovine genome, which biases results at a certain rate. The genome coverage biases can influence CNV call and exclude SNP in CNV-rich regions (Hou et al., 2011). Besides, a lack of nonpolymorphic probes designed for CNV identification reduces the power for detecting CNV. Besides, the CNV identified with the SNP array likely correspond to aggregated discrete CNV, due to the limited number of SNP interrogated along the genome (Fernández et al.,

2014). In fact, 330 (40%) out of the 823 actual CNVR were identified in single sequencing- and array-based calls, and the remaining 493 CNVR (60%) were identified by sequencing-based calls, which corresponded to more than 1 array-based CNV call, ranging from 2 to 31 calls (Supplemental Table S2; <http://dx.doi.org/10.3168/jds.2015-10487>).

Distribution across bovine autosomes and size ranges are shown in Figures 1 and 2, respectively. The CNV sizes ranged from 1 kb to 2.8 Mb (Figure 2). Most of the CNVR (82%) showed sizes that ranged between 1

Figure 2. Distribution of the sizes of the copy number variant regions (CNVR) identified in the analyzed population. MAF = minor allele frequency.

Table 2. Number of overlapping copy number variant (CNV) regions detected in our study and previous studies using comparative genomic hybridization-based (CGH-based), SNP array, and whole-genome sequencing platforms

Method	Study	No. of breeds	CNV count	Overlapping
CGH-based studies	Liu et al., 2010	17	177	14
	Fadista et al., 2010	4	266	9
BovineSNP50 Beadchip	Bae et al., 2010	1	368	17
	Hou et al., 2011	21	682	25
	Hou et al., 2012	1	811	17
	Seroussi et al., 2010	1	278	9
	Jiang et al., 2012	1	99	6
	Jiang et al., 2013	1	367	21
	Cicconardi et al., 2013	5	326	27
Next-generation sequencing	Bickhart et al., 2012	3	1,265	27
	Zhan et al., 2011	1	520	12
	Stothard et al., 2011	2	790	3
	Boussaha et al., 2015	3	957	30

and 50 kb. Those CNVR with minor allele frequencies over 1% were evaluated in association analyses. The number of CNVR with minor allele frequencies >0.01 was then largely reduced to 90 (Supplemental Table S3; <http://dx.doi.org/10.3168/jds.2015-10487>) and is overrepresented on chromosomes *Bos taurus* autosome 18 and 19, although these are not the largest bovine chromosomes. The CNVR on this chromosome expand across 66 and 64 Mb, respectively. Table 2 summarizes the overlapping CNVR with other studies grouped according to the applied genotyping technology. A total of 71 out of the 90 considered CNVR ($\approx 80\%$) matched with those previously described. Note that different studies may detect the same CNVR. The overlap observed across studies may support the reliability of the CNVR detection.

Association Analyses

Association analyses revealed 15 significant CNVR for 6 out of 7 traits under study. Coefficient estimates for significant CNVR for each model and trait are summarized in Table 3. Two different CNVR were associated with fat yield, 1 of them also associated with protein yield, 3 with stature, 1 with foot angle, 7 with udder depth, and only 1 with days open. Three CNVR showed quadratic effects (CNVR 493, 497, and 551). The most relevant associations detected corresponded to CNVR 497, which showed effects for protein and fat yield, and CNVR539 that revealed effects for foot angle and udder depth. The CNVR in this region seem to be associated with regulation of part of the genetic variability for growth and body conformation. The CNVR

Table 3. Significant linear and quadratic effects of the associated copy number variant regions (CNVR) and their frequencies of duplications (% Dup) or deletions (% Del) in the population for each analyzed trait¹

Trait ¹	CNVR_ID	Linear effect β (SD)	Quadratic effect β_2 (SD)	% Dup	% Del
FY	493	-16.56 (5.88)	3.69 (1.44)	12.5	11.9
	497	92.77 (44.19)	-18.34 (8.91)	0.1	3.2
PY	497	-56.90 (39.54)	11.13 (8.02)	0.1	3.2
STA	538	-0.99 (0.21)	—	1.0	1.4
	573	-0.75 (0.17)	—	2.6	2.0
	577	-0.64 (0.11)	—	1.5	5.4
FA	539	-0.39 (0.11)	—	3.5	5.4
UD	537	-0.31 (0.09)	—	10.1	2.7
	539	-0.40 (0.11)	—	3.5	5.4
	569	-0.21 (0.08)	—	8.4	5.3
	570	-0.27 (0.09)	—	6.2	4.0
	571	-0.32 (0.11)	—	5.2	3.4
	572	-0.35 (0.11)	—	4.2	2.7
	591	-0.66 (0.17)	—	2.2	0.4
DO	551	33.18 (8.78)	9.27 (4.23)	1.0	0.3

¹FY = fat yield; PY = protein yield; STA = stature; FA = foot angle; UD = udder depth; DO = days open.

Table 4. Overlap of significant copy number variant regions (CNVR) within previously reported QTL in cattle (QTL database: www.animalgenome.org/cgi-bin/QTLdb/BT/index)

CNVR_ID	QTL position (<i>Bos taurus</i> autosome: position)	QTL trait	QTL identification
493	17: 65–80 cM	C10:0 milk fatty acid C14:1 milk fatty acid C16:1 milk fatty acid	56634 56632 56633
497	18: 22 cM	Protein milk % C14:0 milk fatty acid	57278 56357
538	19: 2.7 cM	Stature	1618, 50233, 50094, 50172
573	19: 65 cM	Stature	15498
577	19: 75 cM	Stature	14109, 14110, 14111, 14112
539	18: 6 cM	Udder depth	49980, 50097, 4991, 50127
551	19: 19 cM	Conception rate	57170

497 and 539 are located on chromosome 18 and 19, respectively, with sizes of 8 (CNVR497) and 146 kb (CNVR539), and minor allele frequencies in the population around 3.5 and 10%, respectively.

Several of the CNVR identified in the current study overlap with bovine genome regions (QTL) for production, milk fatty acid composition, and conformation traits (Table 4), using an approximation of the linkage and physical maps in QTL database (<http://www.animalgenome.org/cgi-bin/QTLdb/index>). Moreover, CNVR551 was detected within 1 single gene, *NFI*, which is known to carry intragenic CNVR (Hsiao et

al., 2015), supporting the consistence of our CNV call result.

A total of 25 genes were annotated within the CNVR, showing significant association with the analyzed traits (Table 5). Few CNVR, such as 577, 539, 569, 571, and 572, do not overlap any annotated gene following the current bovine genome annotation, which may be caused by the incompleteness of the genome annotations (Connelley et al., 2014), or are composed of regulatory or undescribed structural elements. Some of the genes identified in our analysis represent strong biological candidates for their associated traits. The

Table 5. Genes contained within significant copy number variant regions (CNVR) for the analyzed traits¹ and their CNVR position

Trait	ID_CNVR	CNVR position	Gene; chromosome: position
FY	493	17: 75031000–75158596	<i>SNORA77</i> ; 17: 75037475: 75037596 <i>ZDHC8</i> ; 17: 75039314: 75050946 <i>CCDC188</i> ; 17: 75054452: 75056174 <i>RTN4R</i> ; 17: 75103302: 75104723
FY, PY	497	18: 12381000–12527000	<i>FOXF1</i> ; 18: 12379750: 12382722 <i>MTHFS</i> ; 18: 12397787: 12415356 <i>FOXC2</i> ; 18: 12422846: 12425239 <i>FOXL1</i> ; 18: 12433768: 12434808 <i>CA10</i> ; 19: 627558: 1472951 <i>GJC1</i> ; 19: 45154679: 45186611 <i>A1BG</i> ; 18: 65826618: 65830778 <i>RPS5</i> ; 18: 65845925: 65851516 <i>ZNF584</i> ; 18: 65866463: 65872294 <i>ZNF132</i> ; 18: 65878289: 65880106 <i>SNORA30</i> ; 18: 65899433: 65899580 <i>ZNF446</i> ; 18: 65906273: 65910529 <i>SLC27A5</i> ; 18: 65918286: 65927713 <i>TRIM28</i> ; 18: 65954144: 65960263 <i>CHMP2A</i> ; 18: 65960799: 65963496 <i>UBE2M</i> ; 18: 65964093: 65967276 <i>MZF1</i> ; 18: 65971580: 65980686 <i>CENPBD1</i> ; 18: 65988435: 65989328 <i>SNRPE</i> ; 19: 43828735–43829232 <i>U2</i> ; 19: 43871988: 43872177 <i>5S_rRNA</i> ; 19: 64048565: 64048684 <i>NFI</i> ; 19: 18972931: 19240417
STA	538	19: 972000–974000	
STA	573	19: 45179000–45189000	
UD	537	18: 65799000–66004023	
	569	19: 43827000–43833000	
	570	19: 43868000–439000000	
	591	19: 63987000–64057457	
DO	551	19: 19148000–19157000	

¹FY = fat yield; PY = protein yield; STA = stature; FA = foot angle; UD = udder depth; DO = days open.

CNVR 538 showed association with stature; it includes 1 gene, *GJC1*, which encodes a member of the connexin gene family, involved in bone development (Chaible et al., 2011). The CNVR 497 contains the *MTHFSD* gene that belongs to the folate metabolism genes, which play critical roles in regulating milk protein synthesis (Menzies et al., 2009). The CNVR 569 associated with stature and udder depth contains a single gene, *SNRPE*. This gene encodes a member of the small nuclear ribonucleoprotein family associated with several pathologies including hypotrichosis (Pasternack et al., 2013). The CNVR 551, associated with the days open trait, hosts the *NF1* gene, responsible for neurofibromatosis type 1 in human. Although this syndrome consists of cutaneous lesions, potential effects on fertility traits among others have been described (Samuelsson and Akesson, 1988).

Selection Trends in the Population

Changes of CNVR allele frequencies in the genotyped population were tested using linear regression of CNVR genotypes on the year of birth of the animals. Table 6 shows the slope coefficient estimates, *P*-value, and coefficient of determination of the model. The slope estimate for CNVR 538, 539, 551, 556, and 591 appeared to be statistically significant at an α -level of 0.05. However, the magnitude of both the slope estimate and the coefficient of determination of the model were lower than 0.01 (in absolute value) in all cases. These results suggest that CNVR have not changed their frequencies in the population yet. However, as they segregate in the populations, they could be selected in the future to

increase the frequencies of the favorable alleles in the population. Although genome sequencing data for CNV discovery through depth of coverage analysis was limited (mean coverage = 10 $\times$), the results of our study were supported by merging the CNV calls from 2 approaches: whole-genome sequencing and bovine SNP chip genotyping. Even though the strategy followed here may reduce the CNVR discovery probability due to the stringent criteria for their identification, it would considerably reduce the false-positive call. An important limitation of these methods is that they do not allow phasing the number of copy of CNVR in a chromosome pair. This can be a source of loss of information in the statistical models. Despite those different limitations, we have been able to detect relevant associations of CNVR containing genes for which copy number polymorphisms could affect economically important bovine traits.

CONCLUSIONS

Fifteen CNVR were observed to have significant effects on the traits analyzed. Some showed pleiotropic effects on several morphological traits. Most of the CNVR showed linear effects, suggesting that including CNVR effects as linear in the genetic evaluation models could be possible. The inference of the genomic phase of the CNVR to differentiate normal chromosomes versus deleted and duplicated alleles would probably improve the efficient incorporation of CNVR in genomic evaluations. The statistically significant CNVR showed to be segregating in the population, and their allele frequencies have not been changed during the last decades. Therefore, some potential exist to increase the frequencies of the favorable alleles in the population after independent validation of results in this study. However, little genetic variance was explained by the CNVR detected here.

Table 6. Estimate and *P*-value for the slope coefficient of the linear regression¹ model of year of birth on the copy number variant regions (CNVR) doses for the significant CNVR; R² are also shown

CNVR_ID	Slope	<i>P</i> -value	R ²
493	0.001	0.75	0.001
497	0.001	0.17	0.001
537	0.000	0.76	0.001
538	−0.002	0.04	0.003
539	−0.006	<0.01	0.010
551	−0.001	0.01	0.006
569	−0.003	0.04	0.004
570	−0.001	0.39	0.001
571	−0.001	0.49	0.001
572	−0.000	0.91	0.000
573	−0.002	0.64	0.001
577	−0.000	0.64	0.000
591	−0.001	0.04	0.003

¹ $y = \mu + \beta x + e$; where *y* is the vector with the CNVR doses (1, 2, or 3) of animals, *x* is the covariate vector of the year of birth, and *e* is the vector of residuals.

ACKNOWLEDGMENTS

We thank CONAFE for providing sequences, genotypes, phenotypes, and pedigree data used in this study. This work was funded by the agreement between Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (Madrid, Spain) and the Confederación de Asociaciones de Frisona Española (Conafe, Madrid, Spain; CC12-038). We also thank Aberekin S.A. (DE-RIO, Bizkaia, Spain), Cooperativa de ganaderos de Asturias (Ascol, Gijón, Asturias, Spain), Geneticale, Xenetica Fontao (Lugo, Galicia, Spain), Ministerio de Agricultura, Alimentación y Medio Ambiente de España (MAGRAMA, Madrid, Spain), Laboratorio Central

de Veterinaria (LCV, Algete, Madrid, Spain), and Centro Nacional de Análisis Genómico (CNAG, Barcelona, Spain) for their collaboration to the study. We thank Christa Kühn (Leibniz Institute for Farm Animal Biology, Dummerstorf, Germany) for the improvement and final editing of the manuscript.

REFERENCES

- Bae, J., H. Cheong, L. Kim, S. NamGung, T. Park, J. Y. Chun, J. Y. Kim, C. F. Pasaje, J. S. Lee, and H. D. Shin. 2010. Identification of copy number variations and common deletion polymorphisms in cattle. *BMC Genomics* 11:232.
- Bickhart, D. M., Y. Hou, S. G. Schroeder, C. Alkan, M. F. Cardone, L. K. Matukumalli, J. Song, R. D. Schnabel, M. Ventura, J. F. Taylor, J. F. Garcia, C. P. Van Tassell, T. S. Sonstegard, E. E. Eichler, and G. E. Liu. 2012. Copy number variation of individual cattle genomes using next-generation sequencing. *Genome Res.* 22:778–790.
- Bickhart, D. M., and G. E. Liu. 2014. The challenges and importance of structural variation detection in livestock. *Front. Genet.* 5:37.
- Boeva, V., T. Popova, K. Bleakley, P. Chiche, J. Cappo, G. Schleiermacher, I. Janoueix-Lerosey, O. Delattre, and E. Barillot. 2012. Control-FREEC: A tool for assessing copy number and allelic content using next generation sequencing data. *Bioinformatics* 28:423–425.
- Boeva, V., A. Zinovyev, K. Bleakley, J. P. Vert, I. Janoueix-Lerosey, I. Janoueix-Lerosey, O. Delattre, and E. Barillot. 2011. Control-free calling of copy number alterations in depth-sequencing data using GC-content normalization. *Bioinformatics* 27:268–269.
- Boussaha, M., D. Esquerré, J. Barbieri, A. Djari, A. Pinton, R. Letaief, G. Salin, F. Escudié, A. Roulet, S. Fritz, F. Samson, C. Grohs, M. Bernard, C. Klopp, D. Boichard, and D. Rocha. 2015. Genome-wide study of structural variants in bovine Holstein, Montbéliarde and Normande dairy breeds. *PLoS ONE* 10:e0135931.
- Calus, M. P., D. J. De Koning, and C. S. Haley. 2010. Including copy number variation in association studies to predict genotypic values. *Genet. Res.* 92:115–125.
- Chaible, L. M., D. S. Sanches, B. Cogliati, G. Mennecier, and M. L. Dagli. 2011. Delayed osteoblastic differentiation and bone development in Cx43 knockout mice. *Toxicol. Pathol.* 39:1046–1055.
- Che, X., and S. Xu. 2010. Significance test and genome selection in Bayesian shrinkage analysis. *Int. J. Plant Genomics* 2010:893206.
- Cicconardi, F., G. Chillemi, A. Tramontano, C. Marchitelli, A. Valentini, P. Ajmone-Marsan, and A. Nardone. 2013. Massive screening of copy number population-scale variation in *Bos taurus* genome. *BMC Genomics* 14:124.
- Clop, A., O. Vidal, and M. Amills. 2012. Copy number variation in the genomes of domestic animals. *Anim. Genet.* 43:503–517.
- Connelley, T. K., K. Degnan, C. W. Longhi, and W. I. Morrison. 2014. Genomic analysis offers insights into the evolution of the bovine TRA/TRD locus. *BMC Genomics* 15:994.
- Drögemüller, C., O. Distl, and T. Leeb. 2001. Partial deletion of the bovine *ED1* gene causes anhidrotic ectodermal dysplasia in cattle. *Genome Res.* 11:1699–1705.
- Fadista, J., B. Thomsen, L. E. Holm, and C. Bendixen. 2010. Copy number variation in the bovine genome. *BMC Genomics* 11:284.
- Fernández, A. I., C. Barragán, A. Fernández, M. C. Rodríguez, and B. Villanueva. 2014. Copy number variants in a highly inbred Iberian porcine strain. *Anim. Genet.* 45:357–366.
- Fontanesi, L., F. Beretti, V. Riggio, E. Gómez González, S. Dall'Olio, R. Davoli, V. Russo, and B. Portolano. 2009. Copy number variation and missense mutations of the agouti signaling protein (ASIP) gene in goat breeds with different coat colors. *Cytogenet. Genome Res.* 126:333–347.
- Frank, B., J. L. Bermejo, K. Hemminki, C. Sutter, B. Wappenschmidt, A. Meindl, M. Kiechle-Bahat, P. Bugert, R. K. Schmutzler, C. R. Bartram, and B. Burwinkel. 2007. Copy number variant in the candidate tumor suppressor gene *MTUS1* and familial breast cancer risk. *Carcinogenesis* 28:1442–1445.
- Henrichsen, C. N., E. Chaignat, and A. Reymond. 2009. Copy number variants, diseases and gene expression. *Hum. Mol. Genet.* 18(R1):R1–R8.
- Hou, Y., D. M. Bickhart, H. Chung, J. L. Hutchison, H. D. Norman, E. E. Connor, and G. E. Liu. 2012. Analysis of copy number variations in Holstein cows identify potential mechanisms contributing to differences in residual feed intake. *Funct. Integr. Genomics* 12:717–723.
- Hou, Y., G. E. Liu, D. M. Bickhart, M. F. Cardone, K. Wang, E. S. Kim, L. K. Matukumalli, M. Ventura, J. Song, P. M. VanRaden, T. S. Sonstegard, and C. P. Van Tassell. 2011. Genomic characteristics of cattle copy number variations. *BMC Genomics* 12:127.
- Hsiao, M. C., A. Piotrowski, T. Callens, C. Fum, K. Wimmer, K. B. Claes, and L. Messiaen. 2015. Decoding NF1 intragenic copy-number variations. *Am. J. Hum. Genet.* 97:238–249.
- Jiang, L., J. Jiang, J. Wang, X. Ding, J. Liu, and Q. Zhang. 2012. Genome-wide identification of copy number variations in Chinese Holstein. *PLoS ONE* 7:e48732.
- Jiang, L., J. Jiang, J. Yang, X. Liu, J. Wang, H. Wang, X. Ding, J. Liu, and Q. Zhang. 2013. Genome-wide detection of copy number variations using high-density SNP genotyping platforms in Holsteins. *BMC Genomics* 14:131.
- Klambauer, G., K. Schwarzbauer, A. Mayr, D.A. Clevert, A. Mitterecker, U. Bodenhofer, and S. Hochreiter. 2012. cn.MOPS: Mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Res.* 40:e69. <http://dx.doi.org/10.1093/nar/gks003>.
- Li, H., and R. Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754–1760.
- Liu, G. E., Y. Hou, B. Zhu, M. F. Cardone, L. Jiang, A. Cellamare, A. Mitra, L. J. Alexander, L. L. Coutinho, M. E. Dell'Aquila, L. C. Gasbarre, G. Lacalandra, R. W. Li, L. K. Matukumalli, D. Nonneman, L. C. Regitano, T. P. Smith, J. Song, T. S. Sonstegard, C. P. Van Tassell, M. Ventura, E. E. Eichler, T. G. McDanel, and J. W. Keele. 2010. Analysis of copy number variations among diverse cattle breeds. *Genome Res.* 20:693–703.
- Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorf, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttman, A. Kong, L. C. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. Mackay, S. A. McCarroll, and P. M. Visscher. 2009. Finding the missing heritability of complex diseases. *Nature* 461:747–753.
- Marklund, S., J. Kijas, H. Rodriguez-Martinez, L. Rönstrand, K. Funa, M. Moller, D. Lange, I. Edfors-Lilja, and L. Andersson. 1998. Molecular basis for the dominant white phenotype in the domestic pig. *Genome Res.* 8:826–833.
- Menzies, K. K., C. Lefevre, K. L. Macmillan, and K. R. Nicholas. 2009. Insulin regulates milk protein synthesis at multiple levels in the bovine mammary gland. *Funct. Integr. Genomics* 9:197–217.
- Pasternack, S. M., M. Refke, E. Paknia, H. C. Hennies, T. Franz, N. Schäfer, A. Fryer, M. van Steensel, E. Sweeney, M. Just, C. Grimm, R. Kruse, C. Ferrándiz, M. M. Nöthen, U. Fischer, and R. C. Betz. 2013. Mutations in SNRPE, which encodes a core protein of the spliceosome, cause autosomal-dominant hypotrichosis simplex. *Am. J. Hum. Genet.* 92:81–87.
- Samuelsson, B., and H. O. Akesson. 1988. Relative fertility and mutation rate in neurofibromatosis. *Hereditas* 108:169–171.
- Seroussi, E., G. Glick, A. Shirak, E. Yakobson, J. I. Weller, E. Ezra, and Y. Zeron. 2010. Analysis of copy loss and gain variations in Holstein cattle autosomes using BeadChip SNPs. *BMC Genomics* 11:673.
- Stefansson, H., A. Meyer-Lindenberg, S. Steinberg, B. Magnusdottir, K. Morgen, S. Arnarsdottir, G. Bjornsdottir, G. B. Walters, G. A. Jonsdottir, O. M. Doyle, H. Tost, O. Grimm, S. Kristjansdottir, H. Snorrason, S. R. Davidsdottir, L. J. Gudmundsson, G. F. Jonsen, B. Stefansson, I. Helgadóttir, M. Haraldsson, B. Jonsdottir,

- J. H. Thygesen, A. J. Schwarz, M. Didriksen, T. B. Stensbøl, M. Brammer, S. Kapur, J. G. Halldorsson, S. Hreidarsson, E. Sæmundsen, E. Sigurdsson, and K. Stefansson. 2014. CNV conferring risk of autism or schizophrenia affect cognition in controls. *Nature* 505:361–366.
- Stothard, P., J. W. Choi, U. Basu, J. M. Sumner-Thomson, Y. Meng, X. Liao, and S. S. Moore. 2011. Whole genome resequencing of Black Angus and Holstein cattle for SNP and CNV discovery. *BMC Genomics* 12:559.
- Venhoranta, H., H. Pausch, M. Wysocki, I. Szczerbal, R. Hänninen, J. Taponen, P. Uimari, K. Flisikowski, H. Lohi, R. Fries, M. Switon-ski, and M. Andersson. 2013. Ectopic KIT copy number variation underlies impaired migration of primordial germ cells associated with gonadal hypoplasia in cattle (*Bos taurus*). *PLoS ONE* 8:e75659.
- Wang, K., M. Li, D. Hadley, R. Liu, J. Glessner, S. F. Grant, H. Hakonarson, and M. Bucan. 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., L. Zhang, T. Shi, Y. Zhou, H. Cai, X. Lan, C. Zhang, C. Lei, and H. Chen. 2013. Copy number variations of *MICAL-L2* shaping gene expression contribute to different phenotypes of cattle. *Mamm. Genome* 24:508–516.
- Yang, J., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. E. Goddard, and P. M. Visscher. 2010. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* 42:565–569.
- Yue, X. P., T. C. Chang, J. M. DeJarnette, C. E. Marshall, C. Z. Lei, and W. S. Liu. 2013. Copy number variation of PRAMEY across breeds and its association with male fertility in Holstein sires. *J. Dairy Sci.* 96:8024–8034.
- Zhan, B., J. Fadista, B. Thomsen, J. Hedegaard, F. T. Panitz, and C. Bendixen. 2011. Global assessment of genomic variation in cattle by genome resequencing and high-throughput genotyping. *BMC Genomics* 12:557.
- Zhang, F., W. Gu, M. E. Hurles, and J. R. Lupski. 2009. Copy number variation in human health, disease, and evolution. *Annu. Rev. Genomics Hum. Genet.* 10:451–481.