Characterization of copy number variants in a large multibreed population of beef and dairy cattle using high-density single nucleotide polymorphism genotype data[1]

Pierce Rafter,[†,‡] Deirdre C. Purfield,[†] Donagh P. Berry,[†,2] Andrew C. Parnell,[‡] I. Claire Gormley,[‡] J. Francis Kearney,[‖] Mike P. Coffey,[§] and Tara R. Carthy[†]

[†]Teagasc, Animal and Grassland Research and Innovation Centre, Moorepark, Fermoy, Co. Cork, Ireland; [‡]UCD School of Mathematics and Statistics, Insight Centre for Data Analytics, University College Dublin, Belfield, Dublin 4, Ireland; [‖]ICBF, Highfield House, Shinagh, Bandon, Co.Cork, Ireland; and [§]Animal and Veterinary Sciences, SRUC, Roslin Institute Building, Easter Bush, Midlothian EH25 9RG

ABSTRACT: Copy number variants (**CNV**s) are a form of genomic variation that changes the structure of the genome through deletion or duplication of stretches of DNA. The objective of the present study was to characterize CNVs in a large multibreed population of beef and dairy bulls. The CNVs were called on the autosomes of 5,551 cattle from 22 different beef and dairy breeds, using 2 freely available software suites, QuantiSNP and PennCNV. All CNVs were classified into either deletions or duplications. The median concordance between PennCNV and QuantiSNP, per animal, was 18.5% for deletions and 0% for duplications. The low concordance rate between PennCNV and QuantiSNP indicated that neither algorithm, by itself, could identify all CNVs in the population. In total, PennCNV and QuantiSNP collectively identified 747,129 deletions and 432,523 duplications; 80.2% of all duplications and 69.1% of all deletions were present only once in the population. Only 0.154% of all CNVs identified were present in more than 50 animals in the population. The distribution of the percentage of the autosomes that were composed of deletions, per animal, was positively skewed, as was the distribution for the percentage of the autosomes that were composed of duplications, per animal. The first quartile, median, and third quartile of the distribution of the percentage of the autosomes that were composed of deletions were 0.019%, 0.037%, and 0.201%, respectively. The first quartile, median, and third quartile of the distribution of the percentage of the autosomes that were composed of duplications were 0.013%, 0.028%, and 0.076%, respectively. The distributions of the number of deletions and duplications per animal were both positively skewed. The interquartile range for the number of deletions per animal in the population was between 16 and 117, whereas for duplications it was between 8 and 23. Per animal, there tended to be twice as many deletions as duplications. The distribution of the length of deletions was positively skewed, as was the distribution of the length of duplications. The interquartile range for the length of deletions in the population was between 25 and 101 kb, and for duplications the interquartile range was between 46 and 235 kb. Per animal, duplications tended to be twice as long as deletions. This study provides a description of the characteristics and distribution of CNVs in a large multibreed population of beef and dairy cattle.

Key words: bovine, BovineHD, copy number variant, PennCNV, QuantiSNP, structural variant

J. Anim. Sci. 2018.96:4112–4124
doi: 10.1093/jas/sky302

[1]This work was supported by Science Foundation Ireland (SFI) principal investigator award grant number 14/IA/2576.

[2]Corresponding author: Donagh.Berry@teagasc.ie

Received May 23, 2018.
Accepted August 7, 2018.

INTRODUCTION

Two broad classes of genetic variation exist, single nucleotide variants (SNPs) and structural

variants (Feuk et al., 2006). A SNP is where one nucleotide in the genome differs between 2 members of the same species or between homologous chromosomes in an individual (The International HapMap Consortium, 2003); a structural variant is a modification of the structure of the DNA sequence that varies from the usual structure of the DNA (Freeman et al., 2006). A structural variant formed by the deletion or duplication of a stretch of DNA is termed a copy number variant (**CNV**) (Feuk et al., 2006). Typically, a CNV is considered to have a minimum length of 1 kb (Werdyani et al., 2017). Daetwyler et al. (2014) estimated from 234 bulls that, on average, 0.14% of the bovine genome is composed of SNPs. Due to the small percentage of SNPs detected within cattle, it has been hypothesized that a greater percentage of the genome might be composed of CNVs rather than SNPs (Fadista et al., 2010; Seroussi et al., 2010; Hou et al., 2011). Jiang et al. (2012), using Illumina Bovine SNP50K BeadChip data (i.e., 54,001 SNPs) from 2,047 Chinese Holstein cattle, estimated that 0.91% of the bovine genome was composed of CNVs. As well as contributing to genomic variation, CNVs have also been associated with phenotypic variation in beef and dairy cattle. Xu et al. (2014a) demonstrated an association between milk production and CNVs in 26,362 Holsteins from the United States. In a separate study, Xu et al. (2014b) documented that a deletion on chromosome 7 between 40,692,412 and 40,797,251 bp was associated with resistance to gastrointestinal nematodes in 575 Angus cattle. The objective of the present study was to characterize CNVs in a population of 5,551 natural mating and artificial insemination purebred beef and dairy sires from 22 breeds. The CNVs were called from high-density SNP genotype data using the 2 CNV calling platforms of QuantiSNP (Colella et al., 2007) and PennCNV (Wang et al., 2007).

MATERIALS AND METHODS

The genotype data were obtained from the Irish national Cattle Breeding Federation; some of these data were received through exchange with Scotland's Rural College (**SRUC**).

Genotypes

Bulls from an assortment of dairy and beef cattle breeds were genotyped using the BovineHD BeadChip (Illumina Inc., San Diego, CA). The BovineHD BeadChip is a high-density SNP platform that comprises 777,962 SNPs. Single nucleotide polymorphisms on either sex chromosome, without a reported chromosome, or without a reported position were removed from the dataset. Within the study population of 5,931, 2,291 parent–progeny relationships existed. A transition from a homozygous SNP (AA) in the parent to an opposing homozygous SNP (BB) in the offspring is inconsistent with Mendelian inheritance. The SNPs that were inconsistent with Mendelian inheritance in more than 0.2% of parent–progeny pairs were removed from all animals in the dataset; in total, 1,945 SNPs were removed due to inconsistency with patterns of Mendelian inheritance. Animals with a call rate of less than 95%, and SNPs with a call rate of less than 95% were also discarded. After edits, 5,551 bulls remained with called genotypes on 713,162 SNPs; the population consisted of 1,394 Limousin, 1,015 Charolais, 991 Holstein-Friesian, 536 Aberdeen Angus, 397 Simmental, 353 Hereford, 348 Belgian Blue, 87 Jersey, 75 Blonde D'Aquitaine, 66 Aubrac, 66 Parthenasie, 51 Salers, 50 Piedmontese, 36 Montbeliarde, 11 Norwegian Red, 11 Meuse-Rhine-Issel, 5 Swedish Red, 4 Ayrshire, 3 Danish Red, and 2 Shorthorn bulls.

CNV calling protocol

PennCNV (Wang et al., 2007) and QuantiSNP (Colella et al., 2007) are CNV calling platforms and both software suites were used to call CNVs from the autosomes of each animal in the population. Both PennCNV and QuantiSNP account for the distance between SNPs when a CNV is called. The probability that there is a change in copy number between adjacent SNPs is dependent on the distance between the SNPs; the greater the distance between SNPs the greater the chance a change in copy number can occur between adjacent SNPs. PennCNV did not call CNVs that were less than 3 SNPs long. For comparability between QuantiSNP and PennCNV, all CNVs identified by QuantiSNP with less than 3 SNPs were removed. No upper threshold for the number of SNPs in a CNV was specified. Both software suites called the CNVs from the high-density SNP genotype data, based on the Log R ratio (**LRR**) and B allele frequency (**BAF**) value of each SNP using a Hidden Markov Model. The LRR and BAF values were available for all called SNPs in the dataset. The LRR is the log of the observed probe intensity divided by the expected probe intensity. The expected probe intensity is the probe intensity observed in a reference sample and it is a measure of the fluorescence intensity produced by hybridization of a

probe to the SNP array. The BAF is the percentage of B alleles at the locus. Diskin et al. (2008) reported that the median LRR for a 1Mb region of the genome was correlated with the percentage of the DNA that was composed of guanine or cytosine (**GC**) nucleotides. PennCNV was run with and without the GC adjustment option. The GC correction was applied to correct for signal intensity waves associated with the GC content in the 500 kb flanking SNPs as specified by the UCSC GC annotation file (UMD_3.1.1 / bosTau8 June2014). The GC content of the genome was calculated from the UMD_3.1.1 / bosTau8 genome, compiled as of June 2014.

Software Comparison

It is known that the endpoints of CNVs called by PennCNV and QuantiSNP do not always match the true endpoints of the CNV (Colella et al., 2007; Wang et al., 2007). When determining if PennCNV and QuantiSNP called the same CNV, a difference of 1 SNP between the start points and a difference of 1 SNP between the end points was allowed. Dellinger et al. (2010), using simulated SNP data, found that PennCNV and QuantiSNP typically call CNV endpoints to within 1 SNP of the true endpoint of the CNV. If the same CNV was called by PennCNV and QuantiSNP, then this CNV was considered concordant. This criterion for concordance was used to identify concordant CNVs in the final dataset.

The concordance rate per animal was calculated separately for deletions and for duplications as follows:

$$\frac{Number\ of\ Concordant\ CNVs}{Number\ of\ PennCNV\ CNVs + Number\ of\ QuantiSNP\ CNVs - Number\ of\ Concordant\ CNVs}$$

The concordance rate between PennCNV with the GC adjustment and PennCNV without the GC adjustment was calculated per animal in the same way.

The concordance rate between PennCNV and QuantiSNP was also calculated with different overlap thresholds to qualify overlapping CNVs as concordant. Eleven different thresholds of overlap between base pair coordinates were set to qualify a CNV as concordant between PennCNV and QuantiSNP. Those overlap thresholds were as follows: any overlap, at least 10% overlap, at least 20% overlap, and so on in increments of 10% up to direct overlap between CNVs.

Population Characteristics

The final set of CNVs used to characterize the population was the union of all CNVs identified by PennCNV, without the GC adjustment option, and all the CNVs identified by QuantiSNP. The proportion of the autosomes that were composed of CNVs was calculated per animal as the sum of the length of all the CNVs in the animal, divided by the total length of all the autosomes. The length of each autosome was determined as the distance between the outermost SNPs for each autosome on the Illumina BovineHD BeadChip. The calculated total length of all the covered autosomes was 2.51 Gb. A CNV region (**CNVR**) was defined as the combined region covered by overlapping CNVs that were present in at least 2 animals in the population. A CNV hotspot was defined as the intersection of common CNVs in the population such that the intersecting region was a CNV in at least 15% of animals in the population. Seven breeds (i.e., Aberdeen Angus, Belgian Blue, Charolais, Hereford, Holstein-Friesian, Limousin, and Simmental) were represented by >300 individuals in the study population. Using a subset of the sample population used in the present study, Kelleher et al (2017) reported a mean fixation index (i.e., F_{st}) value between the 7 major breeds of 0.098, the minimum F_{st} value being 0.049 (Limousin and Charolais), and the maximum F_{st} value being 0.136 (Hereford and Simmental). Differences in the mean number of CNVs per breed, the mean CNV length per breed, and the mean percentage of the autosomes that were composed of CNVs per breed were evaluated using ANOVA with a Tukey's honest significant difference (**HSD**) adjustment.

A gene cluster enrichment analysis was performed on all the genes that overlapped with CNV hotspots using the online DAVID algorithm (Huang et al., 2009). Using the ensembl bioMart genome browser tool for the UMD3.1 bovine build (Zerbino et al., 2018), the ensembl gene IDs were obtained for each of the genes that overlapped with CNV hotspots. The list of ensembl gene IDs were submitted to the DAVID algorithm, which identified enrichment of genes by biological function. In humans, frequently occurring CNVs are often flanked by homologous stretches of DNA (Sasaki et al., 2010). To determine whether the CNVs hotspots were flanked by homologous DNA sequence, the nucleotide sequence 20 kb upstream and downstream of each CNV hotspot was obtained from the ensembl database and examined for homology. The BLAST nucleotide alignment tool

(Altschul et al., 1990) was used to detect DNA sequences that were homologous between the upstream and downstream regions of each CNV hotspot in the population. The BLAST search was of the upstream region against the downstream region using a nucleotide comparison (blastn) to identify highly similar sequences (megablast). The BLAST tool reports the length of the aligned regions, the percentage match between the 2 aligned regions, an expectation score, and an alignment score. The alignment score is a function of the length of the aligned region and the proportion of matching bases in the aligned sequence. The expectation score is a measure of how likely it would be for the homologous stretch of DNA, in the flanking genomic regions, to occur by chance. The chi-square test was used to determine whether more CNV hotspots were flanked with stretches of homologous DNA than was expected by chance.

RESULTS

Comparison of Copy Number Calling Platforms and Protocols

PennCNV without the GC adjustment called 415,037 deletions and 176,942 duplications from the whole population. When the GC adjustment was applied, 176,910 deletions and 105,197 duplications were called. Of the 176,910 deletions called with the GC adjustment, 105,287 were also called by PennCNV without GC adjustment; the remaining 71,623 were unique to PennCNV with GC adjustment. Of the 105,197 duplications called with the GC adjustment, 82,701 were also called by PennCNV without the GC adjustment, whereas 22,496 were only called by PennCNV with GC adjustment. When the GC adjustment was applied, 730 animals which had previously had CNVs called when the GC adjustment was not applied, no longer had any CNVs called. The loss of CNVs in these 730 animals accounted for 41.3% of the reduction in the number of deletions called from the whole population and 19.6% of the reduction in the number of duplications called from the whole population. The percentage of CNVs per animal that were concordant between PennCNV with or without GC adjustment was positively skewed for both deletions and duplications. The median percentage of deletions that were concordant per animal was 45.7%. The median percentage of duplications that were concordant per animal was 55.6%.

A total of 425,279 deletions and 261,432 duplications were called by QuantiSNP, whereas 418,225 deletions and 176,827 duplications were called by PennCNV. The percentage of CNVs per animal, that were concordant between PennCNV and QuantiSNP, without the GC adjustment, was positively skewed. The first quartile, median, and third quartile for the percentage of concordant deletions per animal was 9.6%, 18.5%, and 38.5%, respectively. The first quartile, median, and third quartile for the percentage of concordant duplications, per animal, was 0%, 0%, and 2.5%, respectively. The concordance rate between PennCNV and QuantiSNP, for deletions and duplications, increased with each relaxation of the overlap criterion for concordant CNVs (Figure 1)

Frequency of CNVs

PennCNV and QuantiSNP collectively identified 432,523 duplications and 747,129 deletions from all animals in the whole population. Per animal, there were more deletions than duplications ($P < 0.05$). The distribution of the number of deletions per animal was positively skewed, as was the distribution of the number of duplications per animal (Figure 2). Per animal, there tended to be more than twice as many deletions as duplications (Table 1). The majority of CNVs were rare (Figure 3); 69.1% of deletions and 80.2% of duplications were present in only one animal in the population. Only 2.9% of deletions, and just over 1.1% of duplications, were present in more than 10 animals in the population. The most common deletion was present in 11.7% of the population (650 animals); it was located between 72,432,362 and 72,503,466 bp on chromosome 12. The most common duplication was present in 11.2% of the population (623 animals); it was located between 83,429,553 and 83,452,740 bp on chromosome 4.

Of the 7 common breeds in the study population, the Holstein-Friesians had a higher mean number of deletions per animal (204) than any of the 6 beef breeds ($P < 0.05$). There was no difference in the mean number of deletions per animal between the beef breeds, which ranged from 102 (Limousin) to 142 (Belgian Blue) (Table 2). For the 7 common breeds in the study population, the mean number of duplications per animal ranged from 66 (Belgian Blue) to 109 (Simmental) (Table 2). There was no difference in the mean number of duplications per animal, except for the following pairs: the

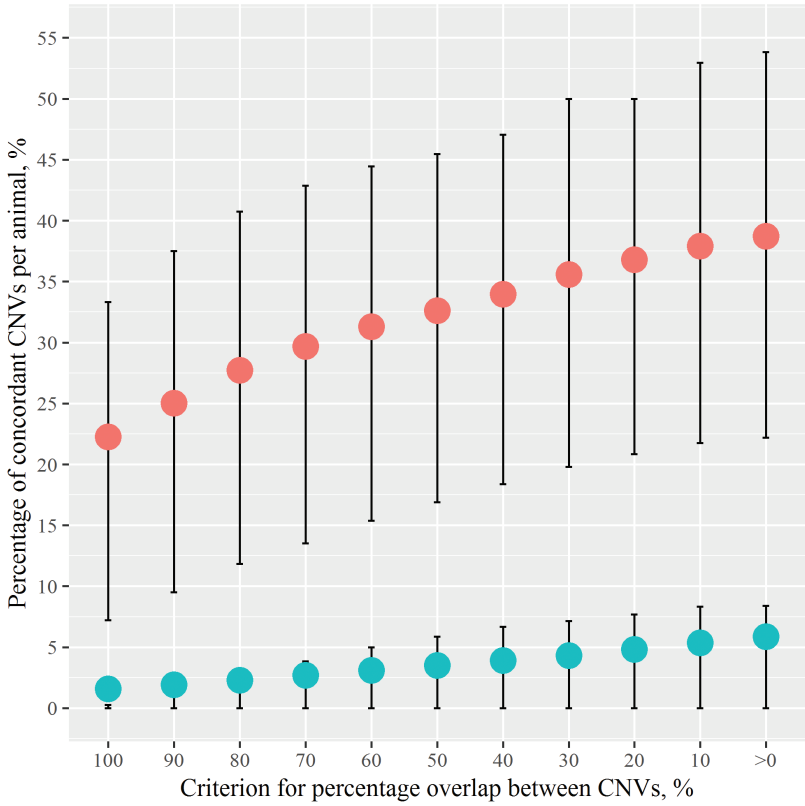

Figure 1. The mean concordance rate (as a percentage) between CNVs called with PennCNV and QuantiSNP with different overlap criterion for concordant CNVs. The CNVs called by PennCNV and QuantiSNP were considered concordant when a certain percentage of their lengths were overlapping. The black bars represent the interquartile range for the rate of concordance. Deletions are represented by red circles and duplications are represented by blue circles.

Holstein-Friesian v Belgian Blue, Charolais, Hereford, and Limousin ($P < 0.05$) and the Simmental v Limousin ($P < 0.05$).

Length of CNVs

The distribution of the length of deletions in the population was positively skewed, as was the distribution of the length of duplications (Figure 4). Duplications were on average more than twice as long as deletions ($P < 0.05$). As well as being longer, the interquartile range for the length of duplications was more than twice that of deletions. The majority (i.e., 69.8%) of CNVs < 175 kb were deletions, whereas 64.3% of CNVs > 175 kb were duplications. On average, deletions that were present only once in the whole population were 38.6% longer than deletions that were present twice in the population. On average, duplications that were present only once in the population were 29.3% longer than duplications that were present twice in the population (Figure 5). For the 7 common breeds in the study population, the mean length of deletions per breed ranged from 78 kb (Limousin) and 104 kb (Holstein-Friesian) (Table 2). The mean length of duplications per breed ranged from 166 kb (Belgian Blue) to 248 kb (Aberdeen Angus) (Table 2). The mean length of deletions per breed was different between all of the breeds ($P < 0.05$), except for the following pairs: the Aberdeen Angus v Belgian Blue, the Hereford v Aberdeen Angus, the Hereford v Belgian Blue, and the Simmental v Hereford. The mean length of duplications per breed was different between all of the breeds ($P < 0.05$), except for the Hereford v Charolais, the Limousin v Charolais, and the Simmental v Limousin.

For the whole population, the distribution of the percentage of the autosomes that were composed of deletions, per animal, was positively skewed, as was the distribution of the percentage of the autosomes that were composed of duplications per animal (Figure 6). Most animals tended to have more of their genome composed of deletions rather than duplications (Table 1).

For the 7 most common breeds in the population, on average, more of the Holstein-Friesian autosomes were composed of deletions than any of the 6 beef breeds ($P < 0.05$). On average, 0.71% of the autosomes in the Holstein-Friesian were composed of deletions (Table 2). There was no difference in the mean percentage of the autosomes that were composed of deletions between any of the 6

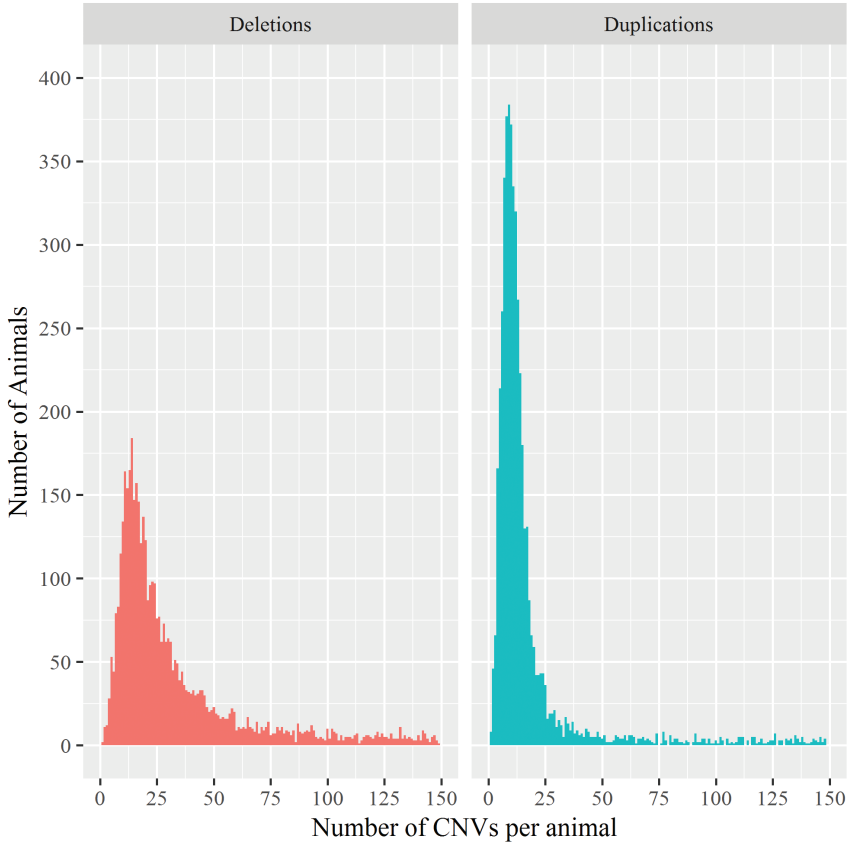

Figure 2. Frequency distribution of the number of copy number variants (CNVs) per animal. The distributions are shown separately for deletions and duplications; animals with >150 (n = 751) duplications or >150 deletions (n = 1,219) not included.

Table 1. Summary statistics on the number of copy number variants (CNVs) per animal, the percentage of the autosomes that were composed of CNVs per animal, and the length of CNVs in the population

	Count—deletions	Count—duplications	Length—deletions (kb)	Length—duplications (kb)	Autosomes percentage—deletions (%)	Autosomes percentage—duplications (%)
25th percentile	16	8	25	46	0.020	0.014
Median	29	12	52	109	0.039	0.029
Mean	135	78	87	202	0.468	0.627
75th percentile	117	23	101	235	0.211	0.080

beef breeds, which ranged from 0.27% (Limousin) to 0.41% (Belgian Blue). The mean percentage of the autosomes that were composed of duplications was between 0.81% (Holstein Friesian) and 0.37% Belgian Blue (Table 2). There was a difference in the mean percentage of the autosomes that were composed of duplications for Limousins v Aberdeen Angus ($P < 0.05$), and the Holstein-Friesian v Belgian Blue, Charolais, Hereford, and Limousin ($P < 0.05$).

Copy Number Variable Regions

There were 7,004 deletion CNVRs discovered in the entire population. The distribution of the length of deletion CNVRs was positively skewed. The first quartile, median, and third quartile were 15, 40, and 103 kb, respectively. Per autosome, the percentage of the chromosome that was a deletion CNVR ranged from 21.8% (chromosome 22) to 68.2% (chromosome 19). In total, 34.3% of all the autosomes were composed of deletion CNVRs. The average distance between deletion CNVRs was 221 kb. In the whole population, there were 4,866 duplication CNVRs. The distribution of the length of CNVR duplications was also positively skewed. The first quartile, median, and third quartile were 19, 59, and 162 kb, respectively. The average distance between duplication CNVRs was 344 kb. Per autosome, the percentage of the chromosome

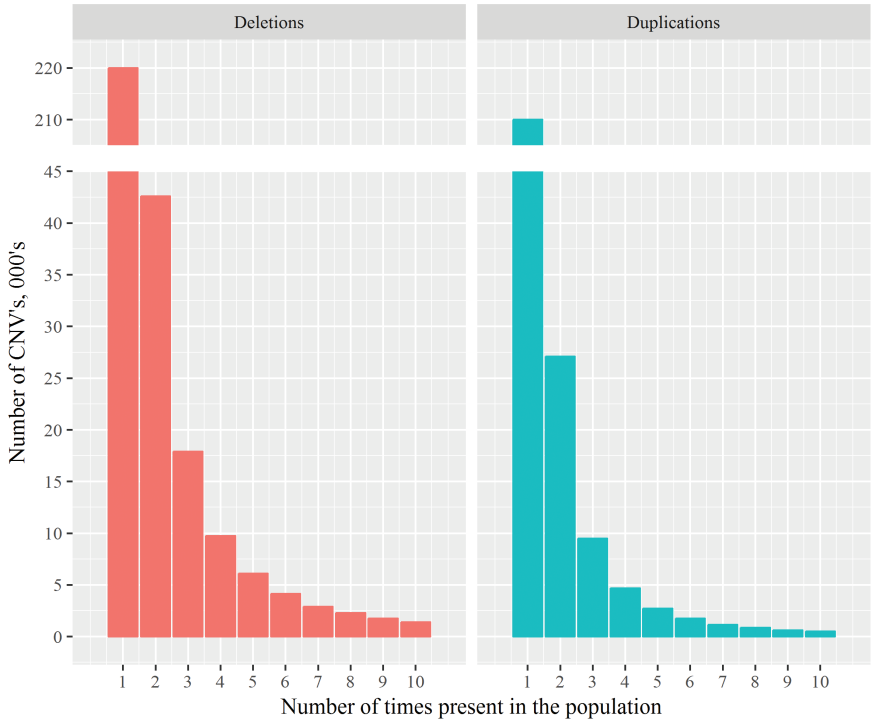

Figure 3. Distribution of the number of times a given copy number variant (CNV) was present in the population. The distributions are shown-separately for deletions and duplications. Deletions present in more than 10 animals in the population (n = 9,358) and duplications present in more than 10 animals in population (n = 3,096) were not included.

Table 2. Summary statistics on the number of copy number variants (CNVs) per animal within breed, the percentage of the autosomes that were composed of CNVs per animal within breed, and the length of CNVs within the breed

Breed	Number of animals	Mean number of deletions	Mean number of duplications	Mean length of deletions (kb)	Mean length of duplications (kb)	Mean autosomes percentage—deletions	Mean autosomes percentage—duplications
Aberdeen Angus	536	125	95	87	248	0.36	0.78
Belgian Blue	342	142	66	87	166	0.41	0.37
Charolais	1,015	117	78	80	187	0.31	0.49
Hereford	353	131	69	85	182	0.37	0.42
Holstein-Friesian	991	204	108	104	224	0.71	0.81
Limousin	1,392	102	67	78	191	0.27	0.43
Simmental	395	139	109	83	196	0.38	0.71

that was a duplication CNVR ranged from 14.6% (chromosome 9) to 72.3% (chromosome 19). In total, 32.1% of the autosomes were composed of duplication CNVRs. Overlap existed between some of the deletion and duplication CNVRs; in total, 51.7% of the autosomes were composed of deletion or duplication CNVRs. After adjusting for the length of the chromosome, chromosome 12 had the greatest excess of deletions over duplications and chromosome 19 had the greatest excess of duplications over deletions.

The most common deletion hotspot in the whole population was present in 21% of animals; it was located on chromosome 23 between 7,675,451 and 7,684,782 bp. The most common duplication hotspot was present in 31% of the population, and it was located on chromosome 29 between 37,319,160 and 37,340,369 bp. The most common CNV hotspot that existed as both a deletion and duplication, in different animals of the population, was present in 27% of animals. This hotspot was a deletion in 12% of animals and it was duplication in 15% of animals; it was located on chromosome 17 between 74,878,327 and 74,891,836 bp. Within the entire population, there were 52 CNV hotspots, 32 of these CNV hotspots overlapped with 129 known genes. Within these 129 genes, there were 13 enriched gene clusters, i.e., genes clustered by biological function. There were 2 clusters ranked joint first in terms of gene enrichment and these

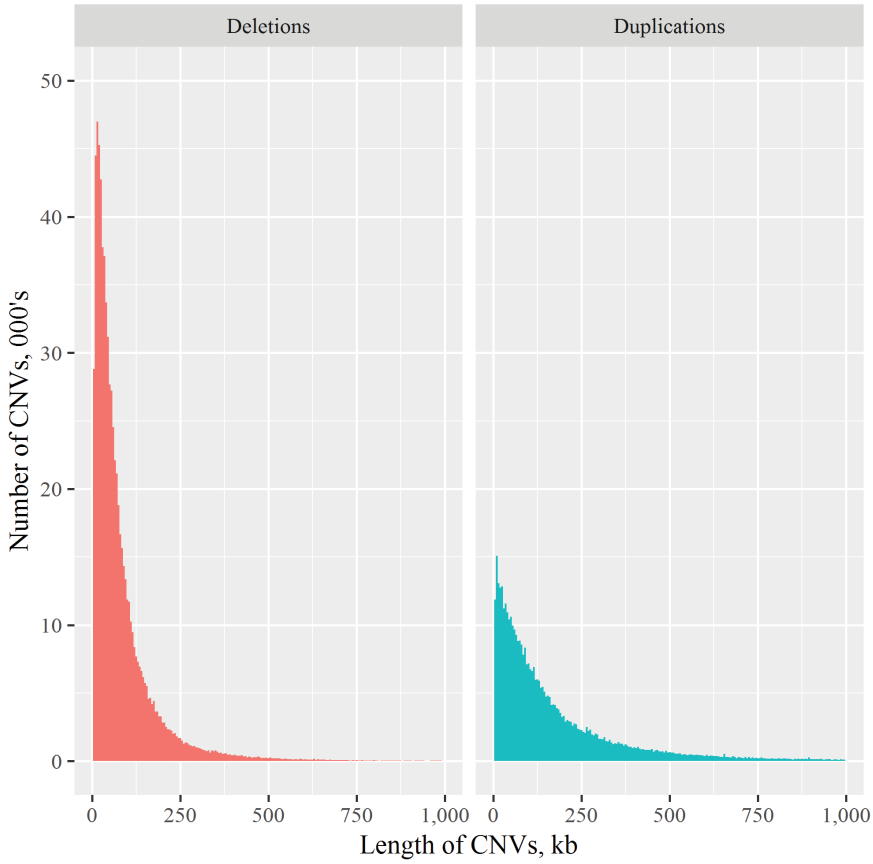

Figure 4. Frequency distribution of the number of copy number variants (CNVs) by length. The distributions of deletions and duplications are shown separately. Duplications longer than 1,000 kb (n = 2,341) and deletions longer than 1,000 kb (n = 10,611) were not included.

were genes encoding leucine-rich repeats and genes encoding the WD40 amino acid motif, which are short ~40 amino acid motifs, often terminating in a Trp-Asp (W-D) dipeptide.

Of the 52 CNV hotspots in the population, 42 were flanked by stretches of homologous DNA sequence. For each CNV hotspot that was flanked by homologous DNA, the homologous region was longer than that was expected by chance ($P < 0.05$). More CNV hotspots were flanked by homologous regions than that was expected by chance ($P < 0.05$). However, the alignment score between the flanking regions did not correlate with the population frequency of the CNV hotspots ($r = -0.199$, $P = 0.1615$). The average length of the flanking homologous regions was 806 bp and the average percentage match between the homologous regions in the upstream flanking region and the downstream flanking region was 89%.

DISCUSSION

The objective of the present study was to characterize CNVs in a large multibreed population of beef and dairy cattle. The CNVs were called from high-density SNP genotypes of 5,551 animals using 2 CNV calling platforms, PennCNV and QuantiSNP. This is the largest and most diverse published scientific study on the characterization of CNVs in cattle using high-density SNP genotype data. Previous studies have documented a poor concordance between PennCNV and QuantiSNP in calling CNVs (Pinto et al., 2011; Metzger et al., 2013). Consistent with those studies, poor concordance between PennCNV and QuantiSNP was also evident in the present study. Although PennCNV and QuantiSNP both use a hidden Markov model to identify CNVs, the underlying algorithms differ in how they determine a transition in copy number from SNP to SNP (Colella et al., 2007; Wang et al., 2007; Xu et al., 2013). PennCNV uses a transition matrix to model changes in copy number that incorporates the LRR and BAF values of the SNP, the distance between adjacent SNPs, and the population frequency of the genotype for the SNP (Wang et al., 2007). QuantiSNP uses an objective Bayesian approach in conjunction with the hidden Markov model to determine transitions in copy number state of adjacent SNPs based on the respective LRR and BAF (Colella et al., 2007). The differences in the CNVs called by PennCNV and QuantiSNP may be a result of the different methods employed by each algorithm to transition from copy number state

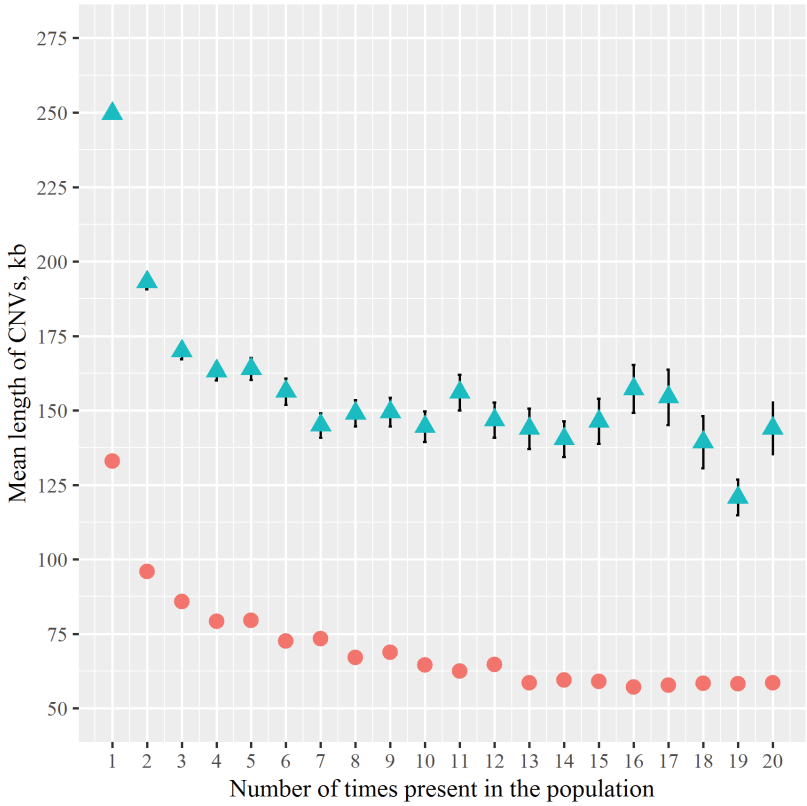

Figure 5. Interaction plot of the number of times copy number variants (CNVs) appear in the population against mean length of CNVs. Duplications are represented by blue triangles and deletions are represented by the red circles. The black error bars give the 95% confidence interval for each point. Deletions (n = 3,427) and duplications (1,051) present in more than 20 animals in the population were not included.

between adjacent SNPs. Nonetheless, PennCNV has a reported low false positive rate of between 1.0% (Wang et al., 2007) and 2.0% (Dellinger et al., 2010); QuantiSNP also has a similarly low false positive rate (Colella et al., 2007; Dellinger et al., 2010). Therefore, given that the reported false positive rate for both CNV calling suites was low, and the concordance between the 2 calling algorithms was low (especially for duplications), it suggested that either CNV calling algorithm, by itself, was not capable of calling the full complement of CNVs in a genome using high-density genotype data. As recommended by Winchester et al. (2009), 2 CNV calling platforms were used to call CNVs, as this would reduce the false negative rate. However, the use of more than 1 CNV calling platform probably also increased the false positive rate. The GC adjustment to PennCNV probably increased the false negative rate of CNV detection. When the GC adjustment was applied, the number of CNVs called by PennCNV decreased by 68%; this is much larger than the false positive rate of 1% to 2% for PennCNV without GC adjustment, as reported by both Wang et al. (2007) and Dellinger et al. (2010). So, even if the GC adjustment eliminates all the false positive CNVs, the majority of the CNVs not called as a result of the GC adjustment could be true CNVs.

Frequency of CNVs

Previous studies that have called CNVs from the BovineHD chip using PennCNV have identified an average of between 10 (Wu et al., 2015) and 74 (Xu et al., 2013) CNVs per animal. In the present study, the average number of CNVs per animal was 213 and the median number of CNVs per animal was 43. More CNVs were called in the present study because 2 CNV calling algorithms were used to call CNVs without adjusting for GC content. Similar to previous studies that called CNVs from the BovineHD chip using PennCNV (Jiang et al., 2013; Xu et al., 2013; Sasaki et al., 2016), more deletions than duplications were identified. In contrast, however, da Silva et al. (2016) identified approximately 5 times as many duplications as deletions in a population of 1,717 Nelore cattle, genotyped with the BovineHD SNP array. In the study of da Silva et al. (2016), the minimum size of a CNV was 20 SNP long, whereas the default minimum length for a CNV is 3 SNPs in PennCNV. Therefore, the study by da Silva et al. (2016) was biased towards longer CNVs. In the present study, it was observed that the majority of longer CNVs were duplications, and duplications were actually more frequent than deletions at

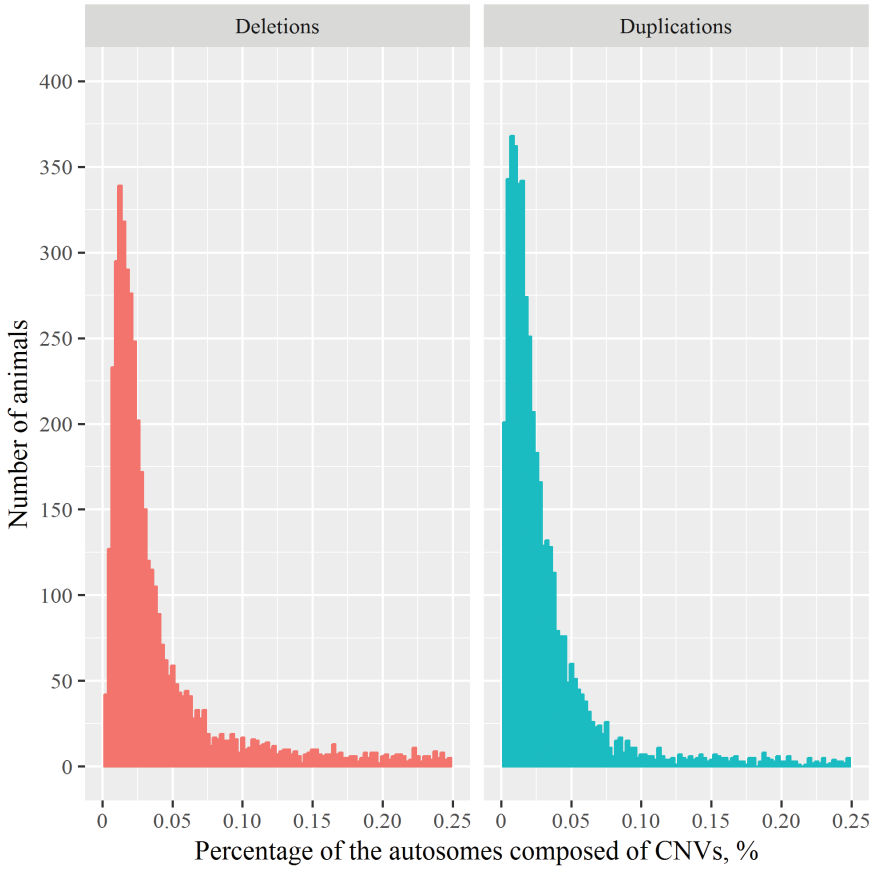

Figure 6. Frequency distribution of the percentage of the autosomes that were composed of copy number variants (CNVs) per animal in the population. The distributions are shown separately for deletions and duplications. Animals whose a percentage of the autosomes that were composed of deletions was > 0.25% (n = 1,218), or animals whose percentage of the autosomes that were composed of duplications was > 0.25% (n = 953) were not included.

lengths > 175 kb. This could be the reason why da Silva et al. (2016) observed more duplications than deletions. There is both a biological and a technical reason why more deletions than duplications may be observed in a population. Firstly, one of the modes of CNV formation is nonallelic homologous recombination; nonallelic homologous recombination has a bias towards deletion formation rather than duplication formation (Turner et al., 2008). Secondly, it has been suggested that deletions are more readily detectable than duplications because a deletion represents a larger proportional change in the LRR value of SNPs than a duplication (Fadista et al., 2010; Eckel-Passow et al., 2011). As PennCNV and QuantiSNP call CNVs based on the LRR values of SNPs, deletions might be more recognizable to the CNV calling algorithms.

After adjusting for chromosome length, chromosome 12 had the greatest frequency of deletions in the present study. Daetwyler et al. (2014) examined the whole genome sequence of 234 bulls and found that chromosome 12 had the greatest proportion of Mendelian inconsistencies. Deletions can result in apparent Mendelian inconsistencies (Winchester et al., 2009). For example, a single copy deletion (A−) has a reported genotype of a homozygote (AA). If a mating occurs between the single copy deletion (A−) and an opposing homozygote (BB), the resulting progeny will have the genotypes (AB) or (B−). The single copy deletion parent will have a reported genotype of AA and the single copy deletion offspring will have a reported genotype of BB. Hence, deletions can result in an apparent Mendelian inconsistency where none actually exists. This may explain why chromosome 12 was reported to have the highest proportion of Mendelian inconsistencies by Daetwyler et al. (2014).

Length of CNVs

Similar to other studies in cattle (Fadista et al., 2010; Jiang et al., 2012; Wang et al., 2015), it was observed that the distributions of length of deletions and duplications in the present study were positively skewed. Xu et al. (2013) observed an average length of 49.9 kb for CNVs called with

PennCNV from the BovineHD genotypes of 630 cattle from 27 different breeds, and da Silva et al. (2016) reported an average CNV length of 320 kb (±413kb) for a population of 1,717 Nelore cattle. By comparison, the average length of CNVs called in the present study was 129 kb. The longer average length of CNVs observed by da Silva et al. (2016) may be due to the fact that, in their study, the minimum number of SNPs per CNV was 20, whereas the default minimum number of SNPs per CNV called by PennCNV is 3. In the present study, the minimum number of SNPs per CNV was set to 3 for both PennCNV and QuantiSNP. The reported average length of CNVRs for cattle in previous studies ranged from 50 kb (Sasaki et al., 2016) to 320 kb (da Silva et al., 2016). In the present study, the average length of CNVRs in the whole population was 177 kb. However, as each of these studies had different population sizes, the length of CNVRs is not directly comparable between studies. In the present study, duplications tended to be twice as long as deletions. This trend was also noted in cattle studies by Fadista et al. (2010) and Boussaha et al. (2015). In the present study, the peaks for both distributions were at or near the minimum length that a deletion or duplication could be detected (Figure 4). This suggested that shorter CNVs may exist, but could not be detected. The fact that rare CNVs were longer than common CNVs indicates that the overall length of the CNV might contribute a negative selection pressure on the CNV (Figure 5). This could be because longer CNVs were more likely to contain a region of the genome that was sensitive to copy number variation.

CNV hotspots

In the gene enrichment analysis using the DAVID algorithm, CNVs could not be weighted by their frequency in the population. To avoid any bias that could result from the absence of weighting by frequency, only the genes that overlapped with CNV hotspots were considered for gene enrichment analysis. Similar to other studies in cattle, there was an enrichment of genes with gene ontology categories associated with biological regulation, cellular component organization, immune system process, development process, localization, and metabolic process (Wang et al., 2015; Sasaki et al., 2016; da Silva et al., 2016).

It is thought that the primary mechanism responsible for CNV formation is nonallelic homologous recombination (Carvalho and Lupski, 2016). Studies in yeast have shown that dispersed repetitive DNA sequences in the genome DNA can cause chromosomal aberrations through nonallelic homologous recombination (Argueso et al., 2008; Hoang et al., 2010). In humans, there are over 30 genomic disorders that are caused by CNVs; these CNVs are known to be produced as a result of nonallelic homologous recombination between dispersed repetitive DNA (Sasaki et al., 2010). Nonallelic homologous recombination is a crossing-over event that occurs between homologous chromosomes at nonallelic positions, i.e., the crossing-over event does not occur at the same position for each chromosome. This causes an unequal crossing-over event, where one of the chromosomes receives more DNA (duplication) than it exchanged, and the other chromosome receives less DNA (deletion) (Carvalho and Lupski, 2016). For nonallelic homologous recombination to occur, the CNV region must be flanked by homologous DNA so that misalignment of the chromosomes can occur (Sasaki et al., 2010). In the present study, it was noted that 42 of the 52 CNV hotspots in the population were flanked by homologous sequence. This result was consistent with the observation in previous studies, in that, CNVs are associated with segmental duplications (Fadista et al., 2010; Hou et al., 2011; Bickhart et al., 2012) and this suggests that nonallelic homologous recombination may be a major mechanism of CNV formation in cattle. However, as there was no relationship between the population frequency of CNV hotspots and the degree of homology between the 20-kb stretches of DNA that flanked the CNV hotspot, there are other factors which may also be affecting the frequency of CNVs in the population.

In conclusion, CNVs are a common feature of the bovine genome; the median number of deletions per animal was 29 and the median number of duplications per animal was 12. Per animal, deletions tended to be twice as frequent as duplications, but duplications tended to be twice as long as deletions. Per animal, the median proportion of the autosomes that was composed of CNVs was 0.077%; for 58.5% of animals in the population, less than 0.1% of their genome was composed of CNVs. The majority of CNVs were rare, with only 0.132% of CNVs present in more than 1% of the population. Just over 80% of the CNV hotspots in the population were flanked by stretches of homologous DNA.

LITERATURE CITED

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. J. Mol. Biol. 215:403–410. doi:10.1016/S0022-2836(05)80360-2

Argueso, J. L., J. Westmoreland, P. A. Mieczkowski, M. Gawel, T. D. Petes, and M. A. Resnick. 2008. Double-strand breaks associated with repetitive DNA can reshape the genome. Proc. Natl. Acad. Sci. U. S. A. 105:11845–11850. doi:10.1073/pnas.0804529105

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, et al. 2012. Copy number variation of individual cattle genomes using next-generation sequencing. Genome Res. 22:778–790. doi:10.1101/gr.133967.111

Boussaha, M., D. Esquerré, J. Barbieri, A. Djari, A. Pinton, R. Letaief, G. Salin, F. Escudié, A. Roulet, S. Fritz, et al. 2015. Genome-wide study of structural variants in bovine holstein, montbéliarde and normande dairy breeds. PLoS One 10:e0135931. doi:10.1371/journal.pone.0135931

Carvalho, C. M., and J. R. Lupski. 2016. Mechanisms underlying structural variant formation in genomic disorders. Nat. Rev. Genet. 17:224–238. doi:10.1038/nrg.2015.25

Colella, S., C. Yau, J. M. Taylor, G. Mirza, H. Butler, P. Clouston, A. S. Bassett, A. Seller, C. C. Holmes, and J. Ragoussis. 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. doi:10.1093/nar/gkm076

Daetwyler, H. D., A. Capitan, H. Pausch, P. Stothard, R. van Binsbergen, R. F. Brøndum, X. Liao, A. Djari, S. C. Rodriguez, C. Grohs, et al. 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat. Genet. 46:858–865. doi:10.1038/ng.3034

Dellinger, A. E., S. M. Saw, L. K. Goh, M. Seielstad, T. L. Young, and Y. J. Li. 2010. Comparative analyses of seven algorithms for copy number variant identification from single nucleotide polymorphism arrays. Nucleic Acids Res. 38:e105. doi:10.1093/nar/gkq040

Diskin, S. J., M. Li, C. Hou, S. Yang, J. Glessner, H. Hakonarson, M. Bucan, J. M. Maris, and K. Wang. 2008. Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. Nucleic Acids Res. 36:e126. doi:10.1093/nar/gkn556

Eckel-Passow, J. E., E. J. Atkinson, S. Maharjan, S. L. R. Kardia, and M. de Andrade. 2011. Software comparison for evaluating genomic copy number variation for Affymetix 6.0 SNP array platform. BMC Bioinformatics. 12:220. doi:10.1186/1471-2105-12-220

Fadista, J., B. Thomsen, L. E. Holm, and C. Bendixen. 2010. Copy number variation in the bovine genome. BMC Genomics. 11:284. doi:http://www.biomedcentral.com/1471–2164/11/284

Feuk, L., A. R. Carson, and S. W. Scherer. 2006. Structural variation in the human genome. Nat. Rev. Genet. 7:85–97. doi:10.1038/nrg1767

Freeman, J. L., G. H. Perry, L. Feuk, R. Redon, S. A. McCarroll, D. M. Altshuler, H. Aburatani, K. W. Jones, C. Tyler-Smith, M. E. Hurles, et al. 2006. Copy number variation: new insights in genome diversity. Genome Res. 16:949–961. doi:10.1101/gr.3677206

Hoang, M. L., F. J. Tan, D. C. Lai, S. E. Celniker, R. A. Hoskins, M. J. Dunham, Y. Zheng, and D. Koshland. 2010. Competitive repair by naturally dispersed repetitive DNA during non-allelic homologous recombination. PLoS Genet. 6:e1001228. doi:10.1371/journal.pgen.1001228

Hou, Y., G. E. Liu, D. M. Bickhart, M. F. Cardone, K. Wang, E. Kim, L. K. Matukumalli, M. Ventura, J. Song, P. M. VanRaden, et al. 2011. Genomic characteristics of cattle copy number variations. BMC Genomics. 12:127. doi:http://biomedcentral.com/1471–2164/12/127

Huang, d. a. W., B. T. Sherman, and R. A. Lempicki. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. 4:44–57. doi:10.1038/nprot.2008.211

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. doi:10.1371/journal.pone.0048732

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. doi:10.1186/1471-2164-14-131

Kelleher, M. M., D. P. Berry, J. F. Kearney, S. McParland, F. Buckley, and D. C. Purfield. 2017. Inference of population structure of purebred dairy and beef cattle using high-density genotype data. Animal 11:15–23. doi:10.1017/S1751731116001099

Metzger, J., U. Philipp, M. S. Lopes, A. C. Machado, M. Felicetti, M. Silvestrelli, O. Distl. 2013. Analysis of copy number variants by three detection algorithms and their association with body sizes in horses. BMC Genomics 14:487. doi:http://biomedcentral.com/1471–2164/14/487

Pinto, D., K. Darvishi, X. Shi, D. Rajan, D. Rigler, T. Fitzgerald, A. C. Lionel, B. Thiruvahindrapuram, J. R. Macdonald, R. Mills, et al. 2011. Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. Nat. Biotechnol. 29:512–520. doi:10.1038/nbt.1852

Sasaki, S., T. Ibi, T. Akiyama, M. Fukushima, and Y. Sugimoto. 2016. Loss of maternal ANNEXIN A10 via a 34-kb deleted-type copy number variation is associated with embryonic mortality in Japanese black cattle. BMC Genomics 17:968. doi:10.1186/s12864-016-3312-z

Sasaki, M., J. Lange, and S. Keeney. 2010. Genome destabilization by homologous recombination in the germ line. Nat. Rev. Mol. Cell Biol. 11:182–195. doi:10.1038/nrm2849

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. doi:http://www.biomedcentral.com/1471–2164/11/673

da Silva, J. M., P. F. Giachetto, L. O. da Silva, L. C. Cintra, S. R. Paiva, M. E. Yamagishi, and A. R. Caetano. 2016. Genome-wide copy number variation (CNV) detection in nelore cattle reveals highly frequent variants in genome regions harboring qtls affecting production traits. BMC Genomics 17:454. doi:10.1186/s12864-016-2752-9

The Internal HapMap Consortium. 2003. The international hapmap project. Nature 426:789–796. doi:10.1038/nature02168

Turner, D. J., M. Miretti, D. Rajan, H. Fiegler, N. P. Carter, M. L. Blayney, S. Beck, and M. E. Hurles. 2008. Germline rates of de novo meiotic deletions and duplications causing several genomic disorders. Nat. Genet. 40:90–95. doi:10.1038/ng.2007.40

Wang, M. D., K. Dzama, C. A. Hefer, and F. C. Muchadeyi. 2015. Genomic population structure and prevalence of

copy number variations in South African nguni cattle. BMC Genomics 16:894. doi:10.1186/s12864-015-2122-z

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. doi:10.1101/gr.6861907

Werdyani, S., Y. Yu, G. Skardasi, J. Xu, K. Shestopaloff, W. Xu, E. Dicks, J. Green, P. Parfrey, Y. E. Yilmaz, et al. 2017. Germline indels and cnvs in a cohort of colorectal cancer patients: their characteristics, associations with relapse-free survival time, and potential time-varying effects on the risk of relapse. Cancer Med. 6:1220–1232. doi:10.1002/cam4.1074

Winchester, L., C. Yau, and J. Ragoussis. 2009. Comparing CNV detection methods for SNP arrays. Brief. Funct. Genomic. Proteomic. 8:353–366. doi:10.1093/bfgp/elp017

Wu, Y., H. Fan, S. Jing, J. Xia, Y. Chen, L. Zhang, X. Gao, J. Li, H. Gao, and H. Ren. 2015. A genome-wide scan for copy number variations using high-density single nucleotide polymorphism array in simmental cattle. Anim. Genet. 46:289–298. doi:10.1111/age.12288

Xu, L., J. B. Cole, D. M. Bickhart, Y. Hou, J. Song, P. M. VanRaden, T. S. Sonstegard, C. P. Van Tassell, and G. E. Liu. 2014b. Genome wide CNV analysis reveals additional variants associated with milk production traits in Holsteins. BMC Genomics 16:683. doi:10.11186/1471-2164-15-683

Xu, L., Y. Hon, B. M. Bickhart, Y. Hou, J. Song, C. P. Van Tassell, T. S. Sonstegard, and G. E. Liu. 2014a. A genome-wide survey reveals a deletion polymorphism associated with resistance to gastrointestinal nematodes in Angus cattle. Funct Integr. Genomics 14:683. doi:10.1007/s10142-014-0371-6

Xu, L., Y. Hou, D. M. Bickhart, J. Song, and G. E. Liu. 2013. Comparative analysis of CNV calling algorithms: literature survey and a case study using bovine high-density SNP data. Microarrays (Basel) 2:171–185. doi:10.3390/microarrays2030171

Zerbino, D. R., P. Achuthan, W. Akanni, M. R. Amode, D. Barrell, J. Bhai, K. Billis, C. Cummins, A. Gall, C. G. Girón, et al. 2018. Ensembl 2018. Nucleic Acids Res. 46(D1):D754–D761. doi:10.1093/nar/gkx1098