J. Dairy Sci. 100:5515–5525
https://doi.org/10.3168/jds.2016-11987

Analysis of copy number variations in Holstein-Friesian cow genomes based on whole-genome sequence data

M. Mielczarek,* M. Frąszczak,* R. Giannico,† G. Minozzi,†‡ John L. Williams,§ K. Wojdak-Maksymiec,# and J. Szyda*[1]

*Biostatistics group, Department of Genetics, Wroclaw University of Environmental and Life Sciences, Kozuchowska 7, 51-631 Wroclaw, Poland
†The Group of Molecular Epidemiology, Fondazione Parco Tecnologico Padano, Via Einstein Albert, Lodi, Lo 26900, Italy
‡Department of Veterinary Medicine, Università di Milano, Via Celoria 10, 20133 Milano, Italy
§The Davies Research Centre, School of Animal and Veterinary Sciences, University of Adelaide, Roseworthy SA 5371, South Australia
#Department of Genetics, Plant Breeding and Biotechnology, West Pomeranian University of Technology, Piastów 17, 70-310 Szczecin, Poland

ABSTRACT

Thirty-two whole genome DNA sequences of cows were analyzed to evaluate inter-individual variability in the distribution and length of copy number variations (CNV) and to functionally annotate CNV breakpoints. The total number of deletions per individual varied between 9,731 and 15,051, whereas the number of duplications was between 1,694 and 5,187. Most of the deletions (81%) and duplications (86%) were unique to a single cow. No relation between the pattern of variant sharing and a family relationship or disease status was found. The animal-averaged length of deletions was from 5,234 to 9,145 bp and the average length of duplications was between 7,254 and 8,843 bp. Highly significant inter-individual variation in length and number of CNV was detected for both deletions and duplications. The majority of deletion and duplication breakpoints were located in intergenic regions and introns, whereas fewer were identified in noncoding transcripts and splice regions. Only 1.35 and 0.79% of the deletion and duplication breakpoints were observed within coding regions. A gene with the highest number of deletion breakpoints codes for protein kinase cGMP-dependent type I, whereas the T-cell receptor α constant gene had the most duplication breakpoints. The functional annotation of genes with the largest incidence of deletion/duplication breakpoints identified 87/112 Kyoto Encyclopedia of Genes and Genomes pathways, but none of the pathways were significantly enriched or depleted with breakpoints. The analysis of Gene Ontology (GO) terms revealed that a cluster with the highest enrichment score among genes with many deletion breakpoints was represented by GO terms related to ion transport, whereas the GO term cluster mostly enriched among the genes with many duplication breakpoints was related to binding of macromolecules. Furthermore, when considering the number of deletion breakpoints per gene functional category, no significant differences were observed between the "housekeeping" and "strong selection" categories, but genes representing the "low selection pressure" group showed a significantly higher number of breakpoints.

Key words: copy number variation, Gene Ontology term, Kyoto Encyclopedia of Genes and Genomes pathway, next-generation sequencing

INTRODUCTION

Genomes contain various types of DNA variation that form the molecular basis of the phenotypic variation. Such polymorphisms range from single-nucleotide changes in DNA such as SNP, oligonucleotide insertions and deletions, multiplication of oligonucleotide fragments such as short tandem repeat polymorphisms and variable number tandem repeat polymorphisms, up to long-scale copy number polymorphisms involving thousands of nucleotides termed structural variations. Among structural variations, copy number variations (**CNV**), which are defined as the gains (duplications) and losses (deletions) of longer DNA fragments, are a major source of genetic diversity in mammals. The CNV sequence length ranges from 50 bp to several Mbp, enabling them to cover many functional elements of the genome including whole genes or regulatory sequences, and thus they may markedly affect phenotypes of individuals by changing gene structure, modifying gene expression by alterations in gene copy number, influencing gene regulation, and exposing recessive alleles (Zhang et al., 2009; Mills et al., 2011; Liu and Bickhart, 2012; Bickhart and Liu, 2014; Shin et al., 2014). It has been found that CNV often occur in gene-rich regions (Bickhart et al., 2012; Choi et al., 2013). Several CNV

Received September 12, 2016.
Accepted March 25, 2017.
[1] Corresponding author: joanna.szyda@upwr.edu.pl

have been shown to play a role in natural phenotypic variability and in disease susceptibility in humans (Aitman et al., 2006; Fellermann et al., 2006; Le Maréchal et al., 2006; Yang et al., 2007; Stankiewicz and Lupski, 2010) and in livestock. Cattle phenotypes affected by CNV include pigmentation and coat color, olfaction and immune response traits, pathogen and parasite resistance, lipid transport, and metabolism (Bickhart et al., 2012; Bickhart and Liu, 2014; Shin et al., 2014).

On a genome-wide scale, CNV have been mostly detected based on comparative genomic hybridization (aCGH) or oligonucleotide (i.e., SNP) arrays. Liu and Bickhart (2012) provide a list of array-based studies applied to bovine genomes, which has recently been expanded by Jiang et al. (2013) and Gurgul et al. (2015). However, the major limitation with SNP and CGH arrays is their low resolution, which is restricted by probe numbers and locations, which typically do not fully cover the whole genome. Copy number variation discovery based on whole genome sequence data, despite being computationally intensive, is becoming increasingly popular. Recent advances in next-generation sequencing (**NGS**) methods provide a more accurate approach to identify not only common, but also rare CNV. Furthermore, NGS provides CNV regions at a base-pair resolution (Bickhart et al., 2012). Studies based on NGS have discovered smaller, previously unknown fragments of structural variants not identified by array-based methods (Alkan et al., 2011).

Studies involving large groups of individuals to detect CNV based on NGS data for bovine genomes are still very uncommon (Bickhart et al., 2012; Shin et al., 2014; Boussaha et al., 2015). Therefore, the main goal of this study was the analysis of 32 cow genomes from the Polish Holstein-Friesian breed to increase information available on bovine CNV and to analyze their functional significance. The focus is on assessing the inter-individual variability in the distribution and length of CNV and genomic annotations of CNV breakpoints.

MATERIALS AND METHODS

Data Set

Thirty-two cows representing the Polish Holstein-Friesian breed were selected from a data set of 991 cows consisting of individuals diagnosed with clinical mastitis and their healthy herd-mates (Wojdak-Maksymiec et al., 2013). This experimental design included 16 paternal half-sibs matched by the number of parities, production level, and birth year, but differing in their mastitis resistance. Mastitis-resistant cows had no incidence of clinical mastitis through their production life, whereas mastitis-prone cows underwent multiple clinical mastitis cases. Whole-genome DNA sequences of the 32 cows were obtained using the Illumina HiSeq Next Generation Sequencing platform (Illumina, San Diego, CA). The total number of raw reads generated for a single animal varied between 164,984,147 and 472,265,620. The average coverage varied between 5× and 17× per cow. A detailed description of the data set and the sequencing procedure are given by Szyda et al. (2015). Sequence files corresponding to this data are publicly available through the National Center for Biotechnology Information BioProject database under the following accession ID: PRJNA359667.

Bioinformatics Pipeline

Raw fastq files from Szyda et al. (2015) were analyzed with the FastQC software (Andrews, 2010) for quality and were not trimmed before alignment. The following analysis pipeline consisted of the following steps: (1) alignment to the reference genome, (2) data processing after alignment, (3) CNV detection, and (4) CNV raw data set filtering. In the first step, BWA-MEM software (Li and Durbin, 2009) was used to align sequences with the reference genome (UMD 3.1; Zimin et al., 2009). In the second step, before further processing, each file generated during the alignment process (binary alignment map) was sorted and indexed, and PCR duplicates were removed using a combination of tools from the Picard (http://broadinstitute.github.io/picard/) and SAMtools (Li et al., 2009) packages. In the third step, the CNVnator software (Abyzov et al., 2011) was used for CNV detection that analyzes genome coverage and defined regions with high or low coverage as CNV (Alkan et al., 2009; Medvedev et al., 2009). This implies that CNV in form of duplications or deletions are defined in comparison to the UMD3.1 reference genome. More specifically, CNVnator divides the entire genome into nonoverlapping bins of identical size and counts the number of mapped reads within each bin as the RD signal. After that, the signal is partitioned into segments with presumably different underlying CNV. To predict true CNV, statistical significance tests are used for those segments. As recommended by Abyzov et al. (2011), for samples with coverage that ranges approximately from 20 to 30, the window size of 100 bp was used. As a consequence, CNV regions identified had a resolution of 200 bp in breakpoint prediction. In the last step, to exclude false positive (**FP**) variants being a consequence of artifacts of the reference genome, deletions shared by at least 15 cows were filtered out if they overlapped by at least 50% with gaps in the reference genome.

Statistical Analysis

The null hypothesis that the length sizes and number of deletions or duplications are normally distributed was tested using the Shapiro-Wilk test. Next, to check whether the number and the length of CNV was dependent on the coverage of the genome, different regression models were tested. Models with the best fit consisted of a linear-log model: $\left(Y_i^N\right) = \beta_0^N + \beta_1^N \log\left(X_i\right) + \varepsilon_i^N$, and a log-log model: $\log\left(Y_i^L\right) = \beta_0^L + \beta_1^L \log\left(X_i\right) + \varepsilon_i^L$, where Y_i^N denotes the number of CNV, Y_i^L is the total length of CNV in a genome, β_0^X is the intercept term, β_1^X is the slope, X_i is a genome-averaged coverage for an individual i, and ε_i^* is the corresponding residual. A Spearman correlation test was performed to test the null hypothesis assuming that deletions and duplications are independent $\left(H_0\colon r_S = 0\right)$ versus $\left(H_1\colon r_S \neq 0\text{:s}\right)$:

$$T = R_S \sqrt{\frac{n-2}{1-R_S^2}},$$

where

$$R_S = 1 - \frac{6\sum_{i=1}^{n}\left(R_i - S_i\right)^2}{n\left(n^2-1\right)},$$

with R_i and S_i denoting ranks of the number of deletions and duplications for ith cow and n representing the number of cows. The null hypothesis of the test can be approximated by the t-Student distribution with (n − 2) degrees of freedom. This approximation is possible for the condition $n > 10$, which is satisfied in this data set. Differences in the percentage of genome/autosomes covered by CNV were tested using the χ^2 test. The null hypothesis was based on the assumption that the same percentage of the genome covered by CNV is expected for all autosomes. Corresponding tests for multiple proportions were performed for each cow separately using the following formula:

$$F = \frac{\sum_{i=1}^{29} d_i \cdot \left(p_i - \overline{p}\right)^2}{\sum_{i=1}^{29} p_i \cdot \left(1 - p_i\right)} \cdot \frac{k}{k-1},$$

where

$$\overline{p} = \frac{\sum_{i=1}^{29} d_i \cdot p_i}{\sum_{i=1}^{29} d_i},$$

and p_i denotes the observed percentage of the ith given autosome covered by CNV, d_i is the length of ith autosome, and $k = 29$, the number of bovine autosomes. Under the null hypothesis, this test statistic follows the F distribution. Nominal P-values resulting from the test were subjected to the Bonferroni correction for multiple testing. The χ^2 test of goodness of fit was used to assess whether the number of CNV is uniformly distributed across the genome:

$$\chi^2 = \sum_{i=1}^{n} \frac{\left(O_i - E_i\right)^2}{E_i},$$

where O_i denoted the number of duplications or deletions for ith cow and $E_i = \frac{m}{29}$, where m was defined as the number of all possible deletions or duplications. Heat maps of deletions and duplications were generated by the R package (R Development Core Team, 2013) for the number of polymorphisms along the entire genomes of all 29 individuals for which CNV categorized as identical were defined based on the exact equity of breakpoint positions. To check whether the distribution of CNV lengths was the same for all animals $\left[H_0\colon \overline{R} = \left(n+1\right)/2\right]$, the Kruskal-Wallis test was applied:

$$H = \frac{12}{k\left(k+1\right)} \sum_{i=1}^{n} \frac{R_i^2}{k_i} - 3\left(n+1\right),$$

where k_i is the number of duplications or deletions for ith cow, and $k = \sum_{i=1}^{n} k_i$, n is the number of cows, and R_i denotes the sum of ranks for deletion/duplication length corresponding to ith cow. The test statistic is approximately χ^2 distributed with k − 1 degrees of freedom (Lehmann, 2006).

Functional Annotation of the CNV

Genomic position of breakpoints defined as start or end positions of CNV were annotated using the UMD3.1 reference genome by variant effect predictor (McLaren et al., 2010). Each position was assigned to 1 of the 28 Sequence Ontology (**SO**) terms (Eilbeck et al., 2005) characterizing functionally different regions of the genome. For the purpose of our study, SO terms were grouped into 8 more general categories consisting of (1) protein coding sequences, (2) noncoding transcript sequences, (3) intron sequences, (4) splice region sequences, (5) untranslated region (**UTR**) sequences, (6) noncoding upstream gene regions, (7) noncoding downstream gene regions, and (8) noncoding intergenic variants. Details of grouping the SO terms are given in Supplemental Table S1 (https://doi.org/10.3168/jds.2016-11987). The detailed analysis of breakpoint distribution in 16 genes representing 3 functional groups: (1) housekeeping, (2) under low selection pressure, and (3) strongly selected genes, was performed. The housekeeping category (1) included genes primarily important for basic metabolic functions. In this study, housekeeping genes considered were from the commercial bovine housekeeping gene array by Qiagen (RT2 Profiler PCR array cow housekeeping genes; Qiagen, Hilden, Germany). The "low selection pressure" category (2) consisted of genes proximal to short tandem repeat markers that do not have large effects on dairy cattle production traits (data not shown). Genes belonging to the "strongly selected genes" category (3) exhibit a very large effect on production traits in dairy cattle, and therefore are likely to be under strong unidirectional selection pressure over many generations. The list of genes in each category is given in Table 1. To check whether the average number of deletion breakpoints in genes is the same in the different functional categories, an empirical null hypothesis on distribution was constructed by permutation of the numbers of breakpoints in genes from given categories.

The logarithmic function of genes versus the total number of CNV breakpoints as well as transcripts versus the total number of CNV summed over all cows was fitted using the SAS software version 9.4 (SAS Institute Inc., Cary, NC). To identify genes/transcripts that exhibited a particularly high number of breakpoints/CNV overlap, a cutoff point was set for which the first derivative was equal to −1, meaning that the estimated rate of decline in the number of breakpoints was more than 1 breakpoint/CNV overlap per gene/transcript up to this point. Genes/transcripts with a large number of CNV breakpoints/CNV overlaps were assigned to Kyoto Encyclopedia of Genes and Genomes (**KEGG**) and GO terms using KOBAS software (Mao et al., 2005), which was also used to identify the total number of KEGG/GO terms represented by the whole *Bos taurus* genome. For each KEGG pathway, a binomial test was applied to assess whether it was under- or overrepresented among genes/transcripts characterized by a high breakpoint/CNV overlap count:

$$Z = \frac{p_b - p_g}{\sigma_{p_b}},\ \sim N\left(0,1\right),$$

where p_b represents the probability of observing a given KEGG pathway within the set of genes/transcripts with a high number of breakpoints/CNV overlaps, p_g is the corresponding probability within the set of *Bos taurus* genes defined by the UMD3.1 reference genome, σ_{p_b} is the standard error of p_b given by $\sqrt{\frac{p_b\left(1-p_b\right)}{N_g}}$. where $N_g = 456$ denotes the number of genes/transcripts with a high number of breakpoints/CNV over-

Table 1. The list of genes selected for comparison and the number of breakpoints located within them

Gene				Breakpoints of	
NCBI ID[1]	Acronym	Name	BTA	Deletions	Duplications
Housekeeping					
280979	*ACTB*	Actin, β	25	6	0
280729	*B2M*	Beta-2-microglobulin	10	0	21
281181	*G3PDH*	Glyceraldehyde-3-phosphate dehydrogenase	5	5	0
515614	*HMBS*	Hydroxymethylbilane synthase	15	1	0
767874	*HSP90AB1*	Heat shock 90kDa protein 1, β	23	25	0
444874	*UBC*	Ubiquitin C	17	1	0
Strong selection					
767906	*ARL4A*	ADP-ribosylation factor-like 4A	4	1	0
407216	*BMP4*	Bone morphogenetic protein 4	10	6	0
282609	*DGAT1*	Diacylglycerol O-acyltransferase 1	14	22	11
535043	*ITGA6*	Integrin, α 6	2	3	0
444881	*MYD88*	Myeloid differentiation primary response 88	22	0	0
Low selection[2]					
534958	*AGTPBP1* (*HEL9*)	ATP/GTP binding protein 1	8	30	4
520250	*ANKRD32* (*ILSTS006*)	Ankyrin repeat domain 32	7	37	1
533894	*LRP1* (*ETH10*)	Low density lipoprotein receptor-releated protein 1	5	8	0
540504	*SYNE2* (*INRA037*)	Spectrin repeat-containing, nuclear	10	48	1
515119	*URI1* (*INRA063*)	URI1, prefoldin-like chaperone	18	3	0

[1] National Center for Biotechnology Information identification number.

[2] A name of the short tandem repeat marker corresponding to a particular gene is given in parentheses.

laps. The GO terms were clustered using DAVID with medium classification stringency (Huang et al., 2009a,b) to identify enrichment of biological processes among the genes/transcripts exhibiting a large number of breakpoints/CNV overlaps.

RESULTS

Preliminary Analysis

A highly significant relation between genome averaged sequencing depth and the number of CNV detected per individual ($P = 1.96 \times 10^{-6}$) and the length of CNV ($P = 0.01$) was identified (Figures 1 and 2). As may be expected, higher sequence depth resulted in a significantly larger number of CNV being detected and the ability to identify shorter CNV. As a consequence, to balance between the number of analyzed genomes and CNV accuracy, we excluded 3 individuals with average genome coverage below 10 from further analyses. Additionally, 30.48% of deletions that had a 50% overlap with gap sequence in the reference genome were removed. Therefore, the final data set consisted of 29 animals for which 435,594 CNV were detected consisting of 373,805 deletions and 61,789 duplications. The lengths of deletions or duplications for each chromosome were not normally distributed and therefore nonparametric tests were incorporated throughout the study.

CNV Variability Across the Genome

The CNV lengths ranged between 200 to 724,000 bp for deletions and 200 to 439,300 bp for duplications. Note that variants shorter than 200 bp could not be detected by the CNVnator algorithm due to the parameters set for the analysis. Depending on the individual, deletions covered from 2.52 to 5.89% of the whole genome, whereas duplications accounted for 0.51 to 1.58%. A significant variation between autosomes was observed in the percentage of a genome covered by CNV.

CNV Variability Across Individuals

The total number of deletions identified per individual was between 9,731 and 15,051 and markedly exceeded the number of duplications, which varied between 1,694 and 5,187 (Figure 3A). Spearman correlation between the number of duplications and number of deletions was significantly negative ($P = 0.01$) and amounted to −0.5. In other words, for an individual genome, more deletions corresponded to fewer duplications. A highly significant inter-individual variation was observed both in the number of duplications ($P = 2.2 \times 10^{-16}$) and in the number of deletions ($P = 2.2 \times 10^{-16}$).

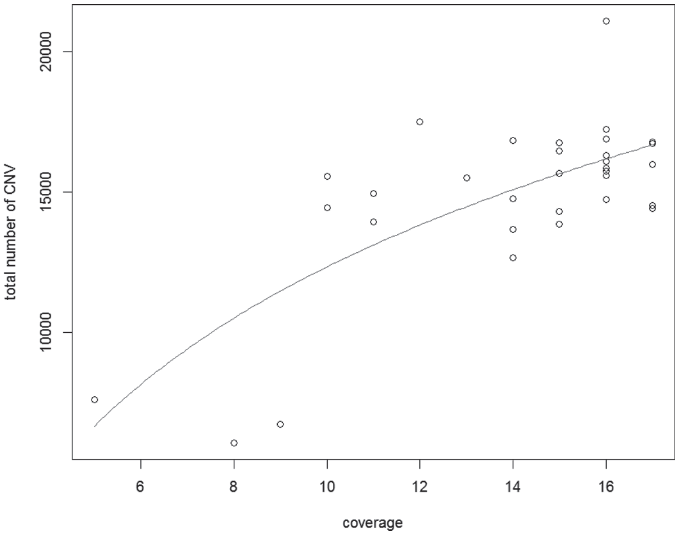

Figure 1. Dependency of the number of copy number variations (CNV) and the averaged coverage, explained by the linear-log regression model.

The estimated CNV frequencies varied from 0.034 (representing a variant unique for only 1 cow) to 1.000 (representing a variant present in all cows, but not in the reference genome). Most of CNV, consisting of 81% of all deletions and 86% of all duplications, were only found in 1 individual, whereas CNV identical for all

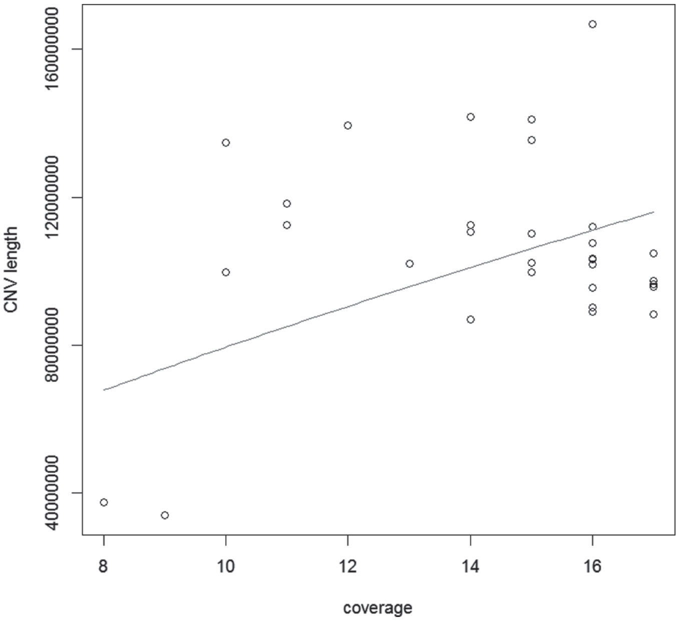

Figure 2. Dependency of the copy number variation (CNV) length and the averaged coverage, explained by the log-log regression.

analyzed animals were rare, and represented only 0.03% of all deletions. Thus, we found 5 deletions present in all 29 animals, located on BTA1, BTA10, and BTA19, respectively, with BTA9 harboring 2 common deletions. No duplications common to all 29 animals were found. Frequency plots of CNV identified for at least 2 animals are shown in Supplemental Figure S2 for deletions and Supplemental Figure S3 for duplications (https://doi.org/10.3168/jds.2016-11987).

The extent of shared variants along entire genomes for all 29 animals is summarized in Figure 4A (deletions) and Figure 4B (duplications). A subset of 14 cows with a large number of deletions in common shared an average of 2,057 pairwise deletions that varied from 1,819 to 2,336, depending on the animal pair compared. In the second subset, consisting of the remaining 15 animals, the number of common pairwise deletions was lower and varied from 724 to 1,372 with an average of 1,047. Moreover, 1 cow (denoted as H9 in Figure 4) shared a low number of CNV with all other individuals, and had a total number of CNV lower than all the others compared with the reference genome. In the case of duplications, the distinction between the subsets was not evident. Nevertheless, 2 groups, 1 of 14 animals and 1 of 5 animals, that shared a higher number of duplications within the groups than with the other individuals were identified. No visual correlation between the pattern of CNV sharing and family relationship or disease status was observed.

The average length of deletions per animal varied from 5,234 ± 16,086 bp to 9,145 ± 22,925 bp, whereas the median of deletion length was lower varying from 1,600 to 2,800 bp. For duplications, which generally represent longer DNA fragments, the average length varied between 7,254 ± 8,990 bp and 8,843 ± 12,409 bp. Median of duplication length ranged from 4,600 to 5,900 bp. Averages and medians of deletions and duplications calculated separately for each animal are summarized in Figure 3.

The inter-individual variation of CNV length across the whole genome was highly significant for both deletions and duplications. However, a more complex pattern emerges by separate comparison of each autosome. The variation of lengths of deleted regions was significant for all autosomes with *P*-values ranging from 7.50×10^{-56} to 1.66×10^{-15}. Seven autosomes (BTA1, BTA2, BTA5, BTA6, BTA10, BTA12, and BTA22) showed a significant variation of duplication length among cows with *P*-values ranging from 1.06×10^{-12} to 3.09×10^{-4}.

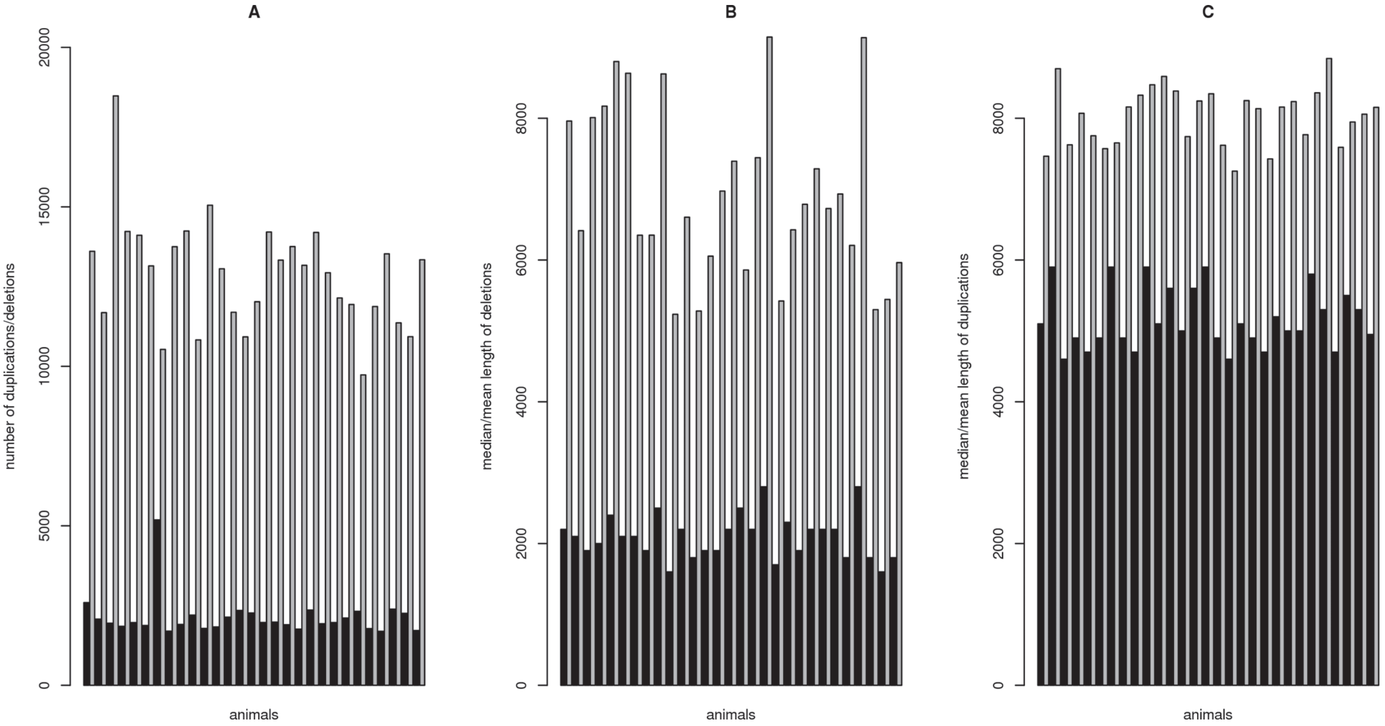

Figure 3. (A) The total number of detected autosomal duplications (black) and deletions (gray) for 29 cows. (B) The median (black) and mean (gray) lengths of deletions found on all autosomes for 29 cows. (C) The median (black) and mean (gray) lengths of duplications found on all autosomes for 29 cows.

Functional Annotation of the CNV

The CNV breakpoint positions (defined by a base pair corresponding to the beginning or end of a CNV) were mapped to the functional elements of the UMD3.1 reference genome. Breakpoints were assigned corresponding SO terms, which were further categorized as coding sequence, intron, splice region, noncoding transcript sequence, 5′ and 3′ UTR, upstream gene sequence, downstream gene sequence, and intergenic sequence (Supplemental Table S1; https://doi.org/10.3168/jds.2016-11987). The highest numbers of deletion breakpoints were located in intergenic regions and introns, which contained 613,006 (57.85%) and 261,570 (24.68%) breakpoints, respectively. The lowest numbers were reported for noncoding regions of gene transcripts: 316 (0.03%), and no breakpoints were located within splice regions. 13,150 (1.24%) of the breakpoints were found within coding regions. For duplications, the proportion of breakpoints in each functional group was similar to deletions. Most duplications were located in intergenic regions: 86,962 (70.37%) and introns: 26,740 (21.63%), whereas the fewest were detected in splice: 46 (0.37%) and noncoding transcript regions of genes: 137 (0.11%). Nine hundred eighty-two (0.79%) duplication breakpoints were found in coding regions. The numbers of all the annotated breakpoints within each functional category are summarized in Figure 5.

A detailed analysis of noncoding parts of transcripts based on the Ensembl noncoding gene cattle database (ftp.ensembl.org/pub/release-80/fasta/bos_taurus/ncrna/) and the miRBase repository for cattle (ftp://mirbase.org/pub/mirbase/CURRENT/genomes/bta.gff3) revealed that 87 duplication breakpoints and 189 deletion breakpoints were annotated to small noncoding RNA. The distribution of breakpoints across noncoding regions is highly nonrandom: 30% of those duplication breakpoints (24 breakpoints) were assigned to the same gene coding for small nucleolar RNA *SNORD116* located on BTA21. In total, this chromosome contains 68% of all duplication breakpoints observed in noncoding segments. Twelve percent of deletion breakpoints were located within *bta-mir-2887–1*, a gene encoding a microRNA molecule, located on BTA18.

The numbers of CNV breakpoints located within 16 selected genes representing different functional categories varied from 0 to 48, with no breakpoints in *B2M* and *MYD88* (Table 1). The number of duplication breakpoints ranged from 0 to 21. Duplications were only present within 4 genes: *B2M*, *DGAT1*, *ANKRD32*, and *SYNE2*. The number of deletion breakpoints per functional category was not significantly different between the "housekeeping" and "strong selection" categories, but genes representing the "low selection pressure" group showed a significantly higher number of breakpoints ($P = 0.03$).

To gain a better insight into the interplay between CNV formation and genome function, the logarithmic curve was fitted to gene ID versus the total number of deletion or duplication breakpoints summed over all cows (Figure 6). The highest number of deletion breakpoints (1,934) was observed within the gene coding for protein kinase cGMP-dependent type I (*PRKG1*; ENSBTAG00000018404), which is located on BTA26. Moreover, the 2 transcripts of this gene, ENSBTAT00000024490 and ENSBTAT00000030539, overlapped with the highest number of deletions amounting to 518 and 449 CNV, respectively. The

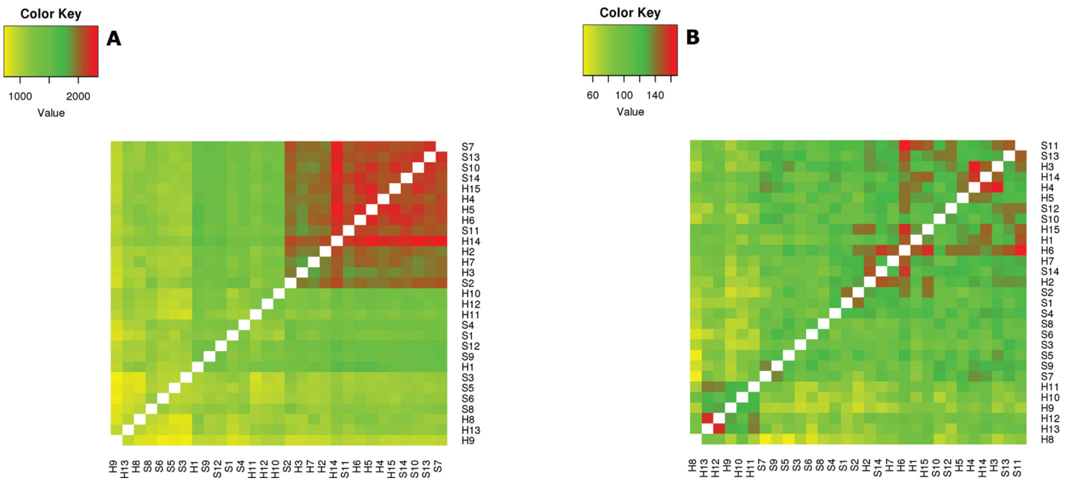

Figure 4. (A) The heat map of deletions for the number of shared polymorphisms along entire genomes of all 29 animals. Numbered S and H (e.g., H1 and S1, H2 and S2, and so on) denote half-sibs. (B) The heat map of duplications for the number of shared polymorphisms along entire genomes of all 29 animals. Numbered S and H (e.g., H1 and S1, H2 and S2, and so on) denote half-sibs. Color version available online.

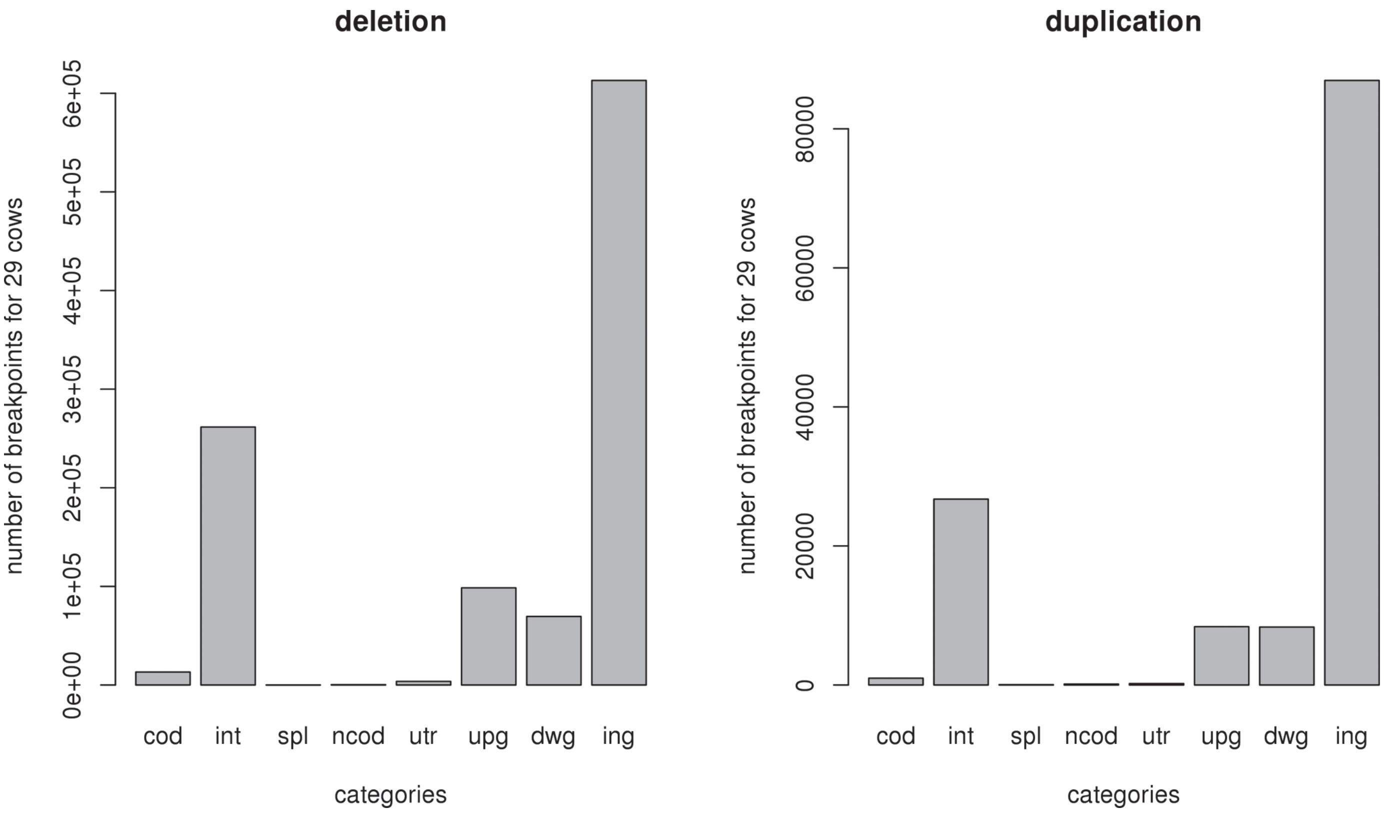

Figure 5. The number of annotated deletion and duplication breakpoints. The abbreviated names of 8 categories are (1) cod = coding sequences, (2) int = introns, (3) spl = splice regions, (4) ncod = noncoding transcripts, (5) utr = 5′ and 3′ untranslated regions, (6) upg = upstream gene regions, (7) dwg = downstream gene regions, and (8) ing = intergenic variants.

former transcript encodes a protein composed of 671 AA and the latter: a somewhat longer protein of 686 AA. The majority of duplication breakpoints (7,164) were located within T-cell receptor α constant (*TRAC*; ENSBTAG00000000432), located on BTA10, which also corresponds to the transcript (ENSBTAT00000002757) harboring the highest number of CNV duplications. We found 398 duplication CNV overlapping with this transcript, which encodes a 274-AA-long protein and is one of the 5 transcripts of this gene. It can be hypothesized that such a high number of genomic duplications may be an evolutionary tool for increasing the variability of transcripts due to the phenomenon of V(D)J recombination of immune-response-related genes.

The functional analysis revealed 45 genes with a large number of deletion breakpoints and 224 genes with a large number of duplication breakpoints. Neither the 86 KEGG pathways corresponding to genes

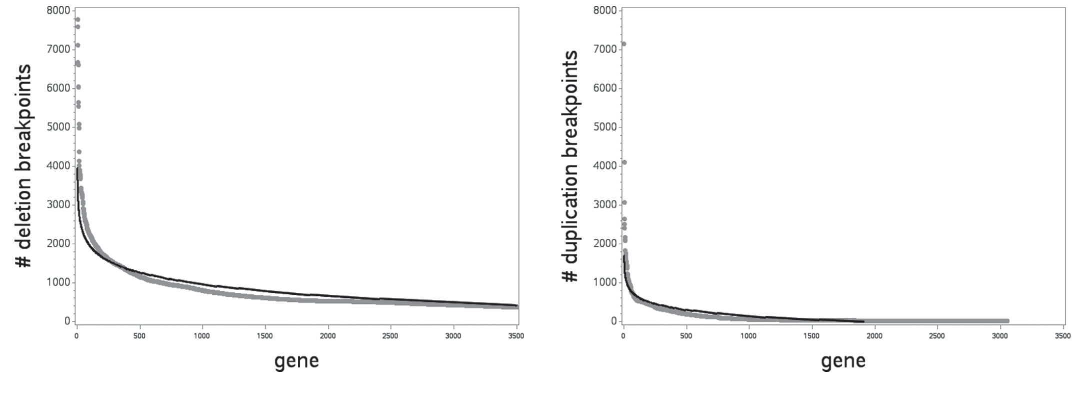

Figure 6. The total number of breakpoints located within a gene (gray) and the corresponding logarithmic function fitted (black).

with many deletion breakpoints, nor the 112 KEGG pathways corresponding to genes with many duplication breakpoints were significantly overrepresented or depleted as compared with the pathway representation underlying the entire bovine transcriptome. Similarly, no significant KEGG enrichment was detected for transcripts overlapping with CNV. The set of GO terms representing genes with many duplication breakpoints revealed 11 functional clusters, among which the cluster with the highest enrichment score was composed of 3 terms (GO:0030246, GO:0030247, and GO:0001871) related to binding of macromolecules. The gene set with many deletion breakpoints revealed 4 functional GO clusters, among which the cluster with the highest enrichment score was composed of 5 terms (GO:0006811, GO:0030001, GO:0006812, GO:0015672, and GO:0046873) related to the biological process of ion transport and the corresponding molecular function consisting of ion transmembrane transporter activity. No GO term clusters were identified for terms corresponding to transcripts with a high number of CNV deletions or duplications.

DISCUSSION

Genomic Landscape of CNV

Using whole-genome sequences of 29 Polish Holstein-Friesian cows, we systematically investigated the distribution and lengths of CNV. A very similar sample size was available to Shin et al. (2014), who reported 6,811 deletions, which is much lower than 373,805 deletions identified in the present study. The discrepancy is expected to arise mainly due to different breeds that were analyzed in both studies (Holstein Friesian and Hanwoo) and different CNV detection software (CNVnator and Genome STRiP). Previous reports observed that structural deletions are more common events than duplications (373,805 vs. 61,789), which is in agreement with the data reported here. A possible biological explanation provided by Turner et al. (2008) is that a nonallelic homologous recombination, one of the major sources of CNV, generates more deleted than duplicated regions. However, the difference in the number of deletions and duplications identified may also be an artifact of the CNV detection software algorithm, which applies more stringent criteria for calling duplications as they are susceptible to the systematic read mapping bias caused by unmapped regions in the reference genome (Abyzov et al., 2011).

The genome-wide CNV distribution is nonuniform. Previous studies have suggested that CNV are formed in hotspots along the genome (Bickhart and Liu, 2014). In the present study, nonuniform formation of CNV was investigated in a functional context along the whole bovine genome. We observed that, especially for deletions that have a potentially much higher effect than duplications, CNV breakpoints occur predominantly in nontranscribed regions such as introns and intergenic sequences. Moreover, when CNV breakpoints within genes with different potential functional effect on phenotypes were considered, genes under low selection pressure showed a significantly higher number of breakpoints. It can be hypothesized that the latter represent the true rate of variation, whereas housekeeping genes and genes under strong artificial selection in dairy cattle are attributed to selection against CNV.

A large variation was observed in the length of duplicated/deleted regions; the maximum length reported here was 724 Kbp for deletions, which is similar to the longest CNV observed by Bickhart et al. (2012). Nevertheless, the detection of CNV based on the NGS data is characterized by a relatively high number of FP results (Meacham et al., 2011; Li, 2014), revision of CNV lengths should be done as additional data become available.

Validation of CNV

A major problem in CNV detection is a low accuracy in determining the location of breakpoints. Zhan et al. (2011) compared CNV detected for the same individual using 3 different methods (NGS, oligonucleotide array, CGH array) and observed a maximum of 23% overlap. A validation of CNV by PCR was also attempted by Shin et al. (2014), who detected that ~20% of variants were wrongly determined by a NGS-based method. These findings emphasize the importance of applying stringent statistical methods to identify CNV to take account of sampling and technical errors present in the data. Common deletions may be artifacts of the reference genome or may be Hereford- or Dominette-specific real variants. A deletion common to all of 62 bulls was reported by Boussaha et al. (2015). In our data set, we found 5 deletions present in all 29 animals, whereas 2 of them also overlapped with deletions reported on BTA10 by Boussaha et al. (2015; DGVdatabase ID: esv3900619, www.ebi.ac.uk/dgva) and on BTA19 reported by Liu et al. (2010) and Boussaha et al. (2015; DGVdatabase ID: esv3900619 and esv3894430).

Another important aspect of CNV detection is the occurrence of FP calls. Although several factors influencing FP have been mentioned, all of them were linked to the structure of the reference genome. Based on the Illumina BovineHD Genotyping BeadChip, Zhou et al. (2016) demonstrated that in a data set consisting of a mixture of female and male animals, FP CNV were reported in genomic regions that in the UMD3.1 assembly correspond to sequences from sex chromosomes (mainly BTAY) misassembled to autosomes. Fadista et al. (2010) observed a significant overlap between CNV regions defined for cattle based on a CGH array and gaps in the BT4 reference genome. The latter problems were also observed in our study in which 30.48% of deletions were located in unsequenced regions of UMD3.1. These were categorized as FP and removed from further analyses. For example, 3 of the deletions excluded from our study as FP were reported by Boussaha et al. (2015) in the DGVa database. All of them are almost entirely located in gaps of the reference genome (Supplemental File S4; https://doi.org/10.3168/jds.2016-11987). Yet another problem for CNV detection is the presence of false duplications in reference genomes, which are artifacts resulting from assembling a haploid reference sequence from a diploid DNA in a heterozygous region (Kelley and Salzberg, 2010).

Relation of CNV to Genome Function

The very large number of deletion breakpoints identified within protein kinase, cGMP-dependent, type I is presumably due to the length of this gene, which is 1,441,876 bp and therefore may not have a clear biological basis. On the other hand, most duplication breakpoints were identified within a *TRAC* gene, which plays a role in the immune system because it encodes a protein located on the surface of type T lymphocytes. This observation is in accordance with the importance of the immune system and especially its genetic variability, which is here shown to be also promoted by frequent CNV formation.

An enrichment of duplications among genes responsible for molecule binding may promote a diversification of immune response. Another interesting finding is the high frequency of CNV duplications identified within small nucleolar RNA *SNORD116*. In knockout mice increased food intake accompanied with increased energy expenditure was demonstrated by Qi et al. (2016). When extrapolated to cattle it can be hypothesized that duplication of the gene results in an opposite effect of more food efficient energy utilization.

CONCLUSIONS

The analysis of data showed that the genomic landscape of CNV is very dynamic. Not only does a considerable variability exist between animals, but CNV breakpoints are also distributed nonuniformly along the genome. It is demonstrated that a different selection pressure exists for deleted and duplicated regions. A between-animal variability causes large sequence variations among animals, which is likely to have an effect on phenotypes. Therefore, a population-wide association analysis between complex phenotypes and CNV would be an interesting follow-up to the study. The nonuniform distribution of CNV breakpoints needs to be explored to understand in what extent it has arisen from functional genomics, evolutionary pressure, varying degree of DNA sequence complexity, or other causes.

ACKNOWLEDGMENTS

The research was supported by the European Union Seventh Framework Programme through The Network of Animal Disease Infectiology Research Facilities (FP7-228394) project, by the Polish National Science Centre (Kraków, Poland) grants 2014/13/B/NZ9/02016 and 2014/15/N/NZ9/03914, and by The Leading National Research Centre (KNOW) program for 2014–2018. Computations were carried out at the Poznan Supercomputing and Networking Centre (Poznań, Poland). We thank Neo Christopher Chung for fruitful discussions.

REFERENCES

Abyzov, A., A. E. Urban, M. Snyder, and M. Gerstein. 2011. CNVnator: An approach to discover, genotype, and characterize typical and atypical CNV from family and population genome sequencing. Genome Res. 21:974–984.

Aitman, T. J., R. Dong, T. J. Vyse, P. J. Norsworthy, M. D. Johnson, J. Smith, J. Mangion, C. Roberton-Lowe, A. J. Marshall, E. Petretto, M. D. Hodges, G. Bhangal, S. G. Patel, K. Sheehan-Rooney, M. Duda, P. R. Cook, D. J. Evans, J. Domin, J. Flint, J. J. Boyle, C. D. Pusey, and H. T. Cook.2006. Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. Nature 439:851–855.

Alkan, C., B. P. Coe, and E. E. Eichler. 2011. Genome structural variation discovery and genotyping. Nat. Rev. Genet. 12:363–376.

Alkan, C., J. M. Kidd, T. Marques-Bonet, G. Aksay, F. Antonacci, F. Hormozdiari, J. O. Kitzman, C. Baker, M. Malig, O. Mutlu, S. C. Sahinalp, R. A. Gibbs, and E. E. Eichler.2009. Personalized copy number and segmental duplication maps using next-generation sequencing. Nat. Genet. 41:1061–1067.

Andrews, S. 2010. FastQC: A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

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.

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.

Choi, J. W., K. T. Lee, X. Liao, P. Stothard, H. S. An, S. Ahn, S. Lee, S. Y. Lee, S. S. Moore, and T. H. Kim.2013. Genome-wide copy number variation in Hanwoo, Black Angus, and Holstein cattle. Mamm. Genome 24:151–163.

Eilbeck, K., S. E. Lewis, J. C. Mungall, M. Yandell, L. Stein, R. Durbin, and M. Ashburner. 2005. The Sequence Ontology: A tool for the unification of genome annotations. Genome Biol. 6:R44.

Fadista, J., B. Thomsen, L.-E. Holm, and C. Bendixen. 2010. Copy number variation in the bovine genome. BMC Genomics 11:284. https://doi.org/10.1186/1471-2164-11-284.

Fellermann, K., D. E. Stange, E. Schaeffeler, H. Schmalzl, J. Wehkamp, C. L. Bevins, W. Reinisch, A. Teml, M. Schwab, P. Lichter, B. Radlwimmer, and E. F. Stange. 2006. A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to Crohn disease of the colon. Am. J. Hum. Genet. 79:439–448.

Gurgul, A., I. Jasielczuk, T. Szmatoła, K. Pawlina, T. Ząbek, K. Żukowski, and M. Bugno-Poniewierska. 2015. Genome-wide characteristics of copy number variation in Polish Holstein and Polish Red cattle using SNP genotyping assay. Genetica 143:145–155.

Huang, W., B. T. Sherman, and R. A. Lempicki. 2009a. Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 37:1–13.

Huang, W., B. T. Sherman, and R. A. Lempicki. 2009b. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. Nat. Protoc. 4:44–57.

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.

Kelley, D. R., and S. L. Salzberg. 2010. Detection and correction of false segmental duplications caused by genome mis-assembly. Genome Biol. 11:R28. https://doi.org/10.1186/gb-2010-11-3-r28.

Lehmann, E. L. 2006. Nonparametrics Statistical Methods Based on Ranks. Rev. ed. Springer, New York, NY.

Le Maréchal, C., E. Masson, J. M. Chen, F. Morel, P. Ruszniewski, P. Levy, and C. Férec. 2006. Hereditary pancreatitis caused by triplication of the trypsinogen locus. Nat. Genet. 38:1372–1374.

Li, H. 2014. Towards better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics 30:2843–2851.

Li, H., and R. Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/map (SAM) format and SAMtools. Bioinformatics 25:2078–2079.

Liu, G. E., and D. M. Bickhart. 2012. Copy number variation in the cattle genome. Funct. Integr. Genomics 12:609–624.

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. McDaneld, and J. W. Keele. 2010. Analysis of copy number variations among diverse cattle breeds. Genome Res. 20:693–703. 10.1101/gr.105403.110.

Mao, X., T. Cai, J. G. Olyarchuk, and L. Wei. 2005. Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. Bioinformatics 21:3787–3793.

McLaren, W., B. Pritchard, D. Rios, Y. Chen, P. Flicek, and F. Cunningham. 2010. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics 26:2069–2070.

Meacham, F., D. Boffelli, J. Dhahbi, D. K. Martin, M. Singer, and L. Pachter. 2011. Identification and correction of systematic error in high-throughput sequence data. BMC Bioinformatics 12:451.

Medvedev, P., M. Stanciu, and M. Brudno. 2009. Computational methods for discovering structural variation with next-generation sequencing. Nat. Methods 6:S13–S20.

Mills, R. E., K. Walter, C. Stewart, R. E. Handsaker, K. Chen, C. Alkan, A. Abyzov, S. C. Yoon, K. Ye, R. K. Cheetham, A. Chinwalla, D. F. Conrad, Y. Fu, F. Grubert, I. Hajirasouliha, F. Hormozdiari, L. M. Iakoucheva, Z. Iqbal, S. Kang, J. M. Kidd, M. K. Konkel, J. Korn, E. Khurana, D. Kural, H. Y. Lam, J. Leng, R. Li, Y. Li, C. Y. Lin, R. Luo, X. J. Mu, J. Nemesh, H. E. Peckham, T. Rausch, A. Scally, X. Shi, M. P. Stromberg, A. M. Stütz, A. E. Urban, J. A. Walker, J. Wu, Y. Zhang, Z. D. Zhang, M. A. Batzer, L. Ding, G. T. Marth, G. McVean, J. Sebat, M. Snyder, J. Wang, K. Ye, E. E. Eichler, M. B. Gerstein, M. E. Hurles, C. Lee, S. A. McCarroll, J. O. Korbel, and 1000 Genomes Project. 2011. Mapping copy number variation by population-scale genome sequencing. Nature 470:59–65.

Qi, Y., L. Purtell, M. Fu, N. J. Lee, J. Aepler, L. Zhang, K. Loh, R. F. Enriquez, P. A. Baldock, S. Zolotukhin, L. V. Campbell, and H. Herzog. 2016. Snord116 is critical in the regulation of food intake and body weight. Sci. Rep. 6:18614.

R Development Core Team. 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Shin, D. H., H. J. Lee, S. Cho, H. J. Kim, Y. Jae Hwang, C. K. Lee, J. Jeong, D. Yoon, and H. Kim. 2014. Deleted copy number variation of Hanwoo and Holstein using next generation sequencing at the population level. BMC Genomics 15:240.

Stankiewicz, P., and J. R. Lupski. 2010. Structural variation in the human genome and its role in disease. Annu. Rev. Med. 61:437–455.

Szyda, J., M. Frąszczak, M. Mielczarek, R. Giannico, G. Minozzi, E. L. Nicolazzi, S. Kamiński, and K. Wojdak-Maksymiec. 2015. The assessment of inter-individual variation of whole-genome DNA sequence in 32 cows. Mamm. Genome 26:658–665.

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.

Wojdak-Maksymiec, K., J. Szyda, and T. Strabel. 2013. Parity-dependent association between TNF-α and LTF gene polymorphisms and clinical mastitis in dairy cattle. BMC Vet. Res. 9:114.

Yang, Y., E. K. Chung, Y. L. Wu, S. L. Savelli, H. N. Nagaraja, B. Zhou, M. Hebert, K. N. Jones, Y. Shu, K. Kitzmiller, C. A. Blanchong, K. L. McBride, G. C. Higgins, R. M. Rennebohm, R. R. Rice, K. V. Hackshaw, R. A. Roubey, J. M. Grossman, B. P. Tsao, D. J. Birmingham, B. H. Rovin, L. A. Hebert, and C. Y. Yu. 2007. Gene copy-number variation and associated polymorphisms of complement component C4 in human systemic lupus erythematosus (SLE): Low copy number is a risk factor for and high copy number is a protective factor against SLE susceptibility in European Americans. Am. J. Hum. Genet. 80:1037–1054.

Zhan, B., J. Fadista, B. Thomsen, J. Hedegaard, F. 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.

Zhou, Y., U. T. Utsunomiya, L. Xu, E. H. A. Hay, D. M. Bickhart, T. S. Sonstegard, C. P. Van Tassell, J. F. Garcia, and G. E. Liu. 2016. Comparative analyses across cattle genders and breeds reveal the pitfalls caused by false positive and lineage-differential copy number variations. Sci. Rep. 6:29219. https://doi.org/10.1038/srep29219.

Zimin, A. V., A. L. Delcher, L. Florea, D. R. Kelley, M. C. Schatz, D. Puiu, F. Hanrahan, G. Pertea, C. P. Van Tassell, T. S. Sonstegard, G. Marçais, M. Roberts, P. Subramanian, J. A. Yorke, and S. L. Salzberg. 2009. A whole-genome assembly of the domestic cow, *Bos taurus.* Genome Biol. 10:4 https://doi.org/10.1186/gb-2009-10-4-r42.