PLOS ONE

RESEARCH ARTICLE

Copy number variant scan in more than four thousand Holstein cows bred in Lombardy, Italy

Andrea Delledonne, Chiara Punturiero, Carlotta Ferrari, Francesca Bernini, Raffaella Milanesi, Alessandro Bagnato, Maria G. Strillacci *

Department of Veterinary Medicine and Animal Science, Università degli Studi di Milano, Lodi, Italy

* maria.strillacci@unimi.it

Abstract

Copy Number Variants (CNV) are modifications affecting the genome sequence of DNA, for instance, they can be duplications or deletions of a considerable number of base pairs (i.e., greater than 1000 bp and up to millions of bp). Their impact on the variation of the phenotypic traits has been widely demonstrated. In addition, CNVs are a class of markers useful to identify the genetic biodiversity among populations related to adaptation to the environment. The aim of this study was to detect CNVs in more than four thousand Holstein cows, using information derived by a genotyping done with the GGP (GeneSeek Genomic Profiler) bovine 100K SNP chip. To detect CNV the SVS 8.9 software was used, then CNV regions (CNVRs) were detected. A total of 123,814 CNVs (4,150 non redundant) were called and aggregated into 1,397 CNVRs. The PCA results obtained using the CNVs information, showed that there is some variability among animals. For many genes annotated within the CNVRs, the role in immune response is well known, as well as their association with important and economic traits object of selection in Holstein, such as milk production and quality, udder conformation and body morphology. Comparison with reference revealed unique CNVRs of the Holstein breed, and others in common with Jersey and Brown. The information regarding CNVs represents a valuable resource to understand how this class of markers may improve the accuracy in prediction of genomic value, nowadays solely based on SNPs markers.

Introduction

For millennia, humans have established a profound relationship with cattle domesticating them to exploit their resources, obtain food as milk, and meat, and meet various needs [1]. Since the 20th century, the selection to improve production traits in animal species, such as the Holstein cattle breed, represents a fundamental step in the development of modern animal husbandry. The Holstein breed, nowadays worldwide recognized for its milk production, has undergone a strong selection effort aimed at improving milk yield, quality, and in the last two decades in enhancing overall functionality and health [2]. In recent years, the evolution of

nanotechnology made available the SNP genotyping platforms that made possible the genomic selection revolution in cattle breeding theorized by Meuwissen et al. [3]. The utilization of SNP chips in genotyping has proven to be a potent tool in animal selection, empowering breeders to make well-informed decisions based on the collective genetic information [4]. SNP genotyping data also enable the detection of Copy Number Variants (CNVs) through the computation of the Log R Ratio (LRR) and B Allele Frequency (BAF). LRR represents a normalized measures of the total signal intensity for two alleles of a SNP, and the B allele frequency (BAF), the one measuring the allelic intensity ratio at marker level [5]. The LRR and BAF facilitate the assessment of CNV status (loss vs gain, LRR; homozygote vs heterozygote, BAF). CNVs represent a category of genomic structural variants recognized to influence phenotypic diversity through the deletion (loss status) or duplication (gain status) of DNA segments, potentially affecting gene structure and regulating expression [6, 7]. These variations typically range in size from 1 kilobase (kb) to 5 megabase (Mb) [8].

The functional impact of CNVs has been studied across various animal species, highlighting their role in influencing a range of phenotypic traits [9–13]. The fact that CNVs affect a multitude of traits across different animal species underlines their role also in adaptive responses to various environmental conditions [14–17]. In several studies on Holstein cattle, CNVRs have been identified to impact economically important traits as milk production, residual feed intake, fertility and somatic cell score [18–21].

Although CNVRs cover a small part of bovine genome length (about 2–10%), as reported by [22], these structural variants can be integrated with SNP information in genomic prediction, offering new insights to explain complex traits and understand the proportion of missing heritability not explained by SNP.

Thus, taking into account all information related to Copy Number Variations (CNVs), the objectives of this study were to examine a substantial population comprising 4,282 Holstein cows from seven distinct farms in Italy, with the purpose of map CNVs across the autosomal genome. Additionally, within the more frequent CNVRs, the goal encompassed the annotation of genes and of quantitative trait loci (QTL) associated with relevant traits in this breed. To validate our findings, we conducted a comparative analysis both within and across different cattle breeds, drawing on insights from prior research studies.

Materials and methods

Animal sampling, genotyping and ethics statement

All cows of 7 herds of the Lombardy region were genotyped with the Illumina GGP Bovine 100K (GeneSeek®) from 2019 to 2023 for a total of 4,282 individuals. These 7 herds are representative of the possible farming systems and selection objectives of Holstien farmers: they in fact spans from a small family run farm (110 cows in lactation) with historically low selection, to a large farm with Automatic Milking System and with more than 3 decades of directional selection to improve production and functionality (about 550 lactating cows) and a medium size farm producing Parmigiano Reggiano cheese and thus, requiring specific nutritional practices (no silage) and selection for milk quality. Log R Ratio (LRR) available from the SNP chip processing were used to map CNVs. The quality assessment of LRR and the mapping of CNVs was performed with the Golden Helix Inc. SVS 8.9 software (SVS).

The sampling of individual was approved by the OPBA (*i.e.*, Animal Welfare Organisation) of the University of Milan (Protocol number 160_2019), by Directive 2010/63/EU of the European Parliament and the Council of 22 September 2010, updating Directive 86/609/EEC on the protection of animals used for scientific purposes.

Quality control of genotyping data

The quality assessment of LRR values was performed considering the Derivative Log Ratio Spread (DLRS) as described by Pinto et al. [23] and the GC Wave Factor (GCWF) [24], both affecting signal intensity and possible cause of bias in CNVs mapping. A total of 47 samples were excluded due to their high DLRS values, while other 135 samples were excluded because of the elevated GCWF values. The detection of CNVs was then conducted on a dataset of 4,100 samples.

CNVs and CNVRs detection

CNVs detection was obtained on autosomes with SNPs mapped on the ARS UCD1.2 assembly reference genome. The detection was performed using the Copy Number Analysis Module (CNAM) of SVS by means of the univariate analysis based on LRR values. Default parameters for CNVs calling in CNAM were set as follows: i) a maximum of 100 segments per 10,000 markers; ii) a minimum of 3 markers per segment; iii) 2000 permutations per pair with a p-value cut-off of 0.005.

To identify animals with outliers CNVs frequencies and length, their distributions were analysed using QQ plots (R routine in ggplot2 library [25]). Outliers were identified as samples having CNV length greater than 7.5 Mbp. After the identification and exclusion of the individuals considered outliers (3,809 subjects were left), the individual frequency of gain and loss in relation to each sample mean CNVs length was plotted with the ggplot2 library of R.

Using the Bedtools -mergeBed command [26], CNVs that overlapped by at least one bp and were shared by a minimum of two animals were combined to generate CNV regions (CNVRs). Then, CNVRs were classified as gain, loss, or complex if comprising both deletions (loss) and duplications (gain). A CNV found in a single individual was classified as a singleton CNVR.

To be representative, only CNVRs shared by at least 2% of the population were selected for descriptive statistics as well as for downstream analyses.

The R package HandyCNV [27] was used to visualized the physical distribution of CNVRs on autosomes.

Genes and QTL annotations

The genes list with official "gene name ID" was downloaded from NCBI online Database. Genes were then annotated within the detected CNVRs using the Bedtools "-intersectBed" command [26], while the QTL associated with the genes found in the CNVRs were identified thanks to the cattle QTL database (https://www.animalgenome.org/cgi-bin/QTLdb/BT/search) by gene name, using the "Search by associated gene" option of QTLdb.

The Cytoscape plugin ClueGo was used to identify potential biological connections among candidate genes identified in the CNVRs [28, 29]. The network construction relied on information from GO and KEGG database. This analysis utilized the bovine databases integrated into the ClueGO app. Only connections with a p-value lower than 0.05 were considered.

Diversity at the population level

To study the diversity within the breed we recoded CNVs defining a CNVR for each cow as follows: i) '1' for loss state; ii) '0' for normal state; iii) '2' for gain state. We used the Past 4.03 software to perform a principal component analysis (PCA).

Comparison with results from the literature

Our identified CNVRs were compared with the results reported in recent literature studies using the HandyCNV library of R-Studio software (compare_cnvr() function).

As reported in Table 3, two distinct comparisons were performed in order to validate Holstein specific CNVRs (comparison within breed), and to identify genomic regions shared by different breeds (comparison among breeds), i.e. Jersey (JER) and Brown Swiss (BSW). For studies with CNVRs using a different genome assembly from ARS-UCD1.2, the positions were remapped using the UCSC Lift Genome Annotations tool (https://genome.ucsc.edu/cgi-bin/hgLiftOver). A graphical visualization of overlapped CNVRs was realized through a Venn diagram built using an online tool (http://bioinformatics.psb.ugent.be/webtools/Venn/).

Results

CNVs and CNVRs detections

According to the number of CNVs per cow and their total length (sum of each CNV length), 291 samples were identified as outliers and subsequently removed to avoid the introduction of possible false positive CNVs; the final dataset comprising 123,814 CNVs was obtained in 3,809 cows (S1 Table); with a total of 4,150 non-redundant CNVs.

As reported in Table 1, CNVs have a maximum, minimum, and average length of 1,860,579, 1,005 and 86,166 bp, respectively. The frequency of loss CNVs doubles the frequency of gain CNVs and the mean length of losses (90,439.4) is longer than the mean length of gains (77,785.5).

Fig 1A shows the different distribution of gain and loss CNVs according to the relationship between the CNV mean length and their number per samples. Furthermore, as shown in Fig 1B, the majority of CNVs falls into the first three classes of length. Over 30,000 loss state CNVs exhibited a length below 0.05, falling in the first length class. Conversely, the majority of gain CNVs had a length ranging between 0.05 and 1 Mb. The longest CNVs were low represented for both of CNV states.

The 123,814 CNVs were aggregated into 1,397 CNVRs (Table 2 and S2 Table), covering 9.18% (228 Mbp) of the total autosomal length (2,489 Mbp). After removing singletons and CNVRs shared by less than 2% of the population, 267 CNVRs remained (CNVRs_2% in Table 2 and S2 Table): 76 in gain state, 129 in loss and 62 categorized as complex. CNVs in CNVR_2% are listed in the S2 Table. These latter CNVRs cover 2.92% of the autosomal genome length and their physical distribution on autosomes is shown according to their states in Fig 2. Values (%) on this graph represent the genomic proportion covered by CNVRs with respect to each chromosome length. CNVRs on chromosomes 12, 18 and 23 covered more than 5% of chromosomal length, 9.5%, 7.4% and 5.1% respectively, while all other chromosomes were impacted by a lower proportion of CNVRs. The CNVRs shared by the largest number of cows were on BTA 10 at 22,676,353 bp (n. 3,528 cows, loss) and on BTA 2 at 93,926,090 (n. 3,107 cows, loss). Instead, CNVRs shared by the lowest number of cows, i.e. 76 animals, were found in gain state within chromosome 20 (at 66,818,777 bp).

Table 1. Descriptive statistics of identified CNVs.

N.	N. CNVs	N. Gain	N. Loss	Loss/Gain	Min-Max CNV per ID (mean)	Min-Max (mean) length*	Min-Max (mean) coverage per ID*
3,809	123,814	41,556	82,258	1.98	13–51 (32.5)	1–1,860 (86)	947–7,483 (2,792)

*Value expressed in Mbp.

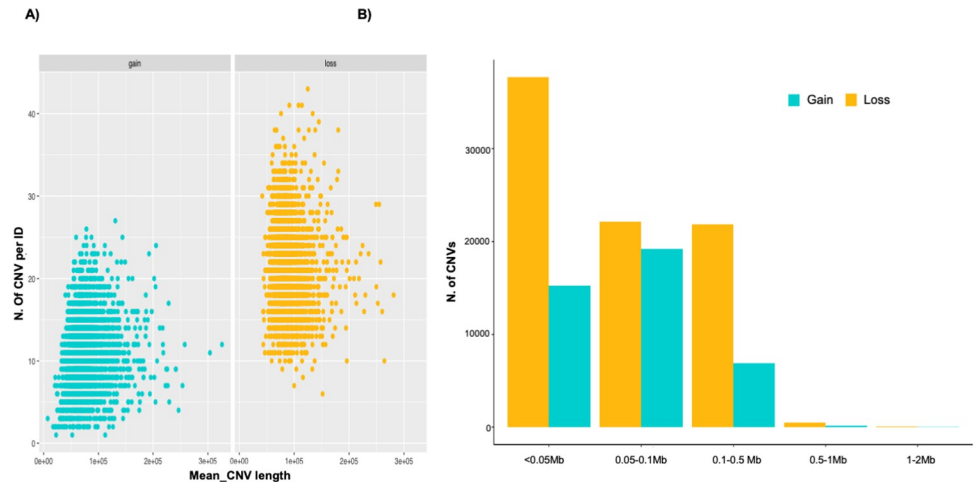https://doi.org/10.1371/journal.pone.0303044.t001

Fig 1. Summary of statistic for detected CNVs. A) Relationship between number and mean total length (bp) of CNVs identified in each sample by state (gain vs loss); B) Number of CNVs for five classes of length.

https://doi.org/10.1371/journal.pone.0303044.g001

S1 Fig shows the genome-wide distribution of the 267 CNVRs across the chromosomes together with the mean CNVRs coverage length. The maximum number of CNVRs are on BTA 1 and BTA 9. The mean CNVRs length is not uniform along all chromosomes, and the maximum mean CNVR length was on BTA 12 (717,015.8 bp).

Principal component analysis results (Fig 3A and 3B) depict the genetic variability in the 3,809 cows analyzed, according to the presence or absence of CNVs in the identified CNVRs, considering their state. Each point in the scatter plots represents an individual animal, coloured as unique population (Fig 3A) or taking into account the herd from which it was sampled (Fig 3B).

Gene content and annotation

A total of 996 genes were annotated within 194 Holstein CNVRs (72.6% of the CNVRS_2%). Their functional classification, according to the David database, is reported in the S3 Table (recognized gene IDs = 942).

In S2 Fig (ClueGo network) it's possible to observe the presence of five macro-groups of genes associated with the following categories: troponin complex, sensory perception of smell, nervous system process, tuberculosis, and MHC class II protein complex. The KEGG pathway comprising the majority of genes is the one connected to tuberculosis, the same result has been obtained with David analysis.

After consulting the Cattle QTLdb, 142 genes were associated with a total of 122 different "Trait Name", grouped into 24 "Trait Types" corresponding to 6 "Trait Classes" (Exterior, Healthy, Meat and Carcass, Milk, Production, and Reproduction Traits), in concordance with

Table 2. Descriptive statistics of identified CNVRs.

CNVRs	Tot n. CNVRs	Tot n. Singleton	CNVRs State			CNVRs length		
			Loss	Gain	Complex	Min	Max	Mean
CNVRs	1,397	329	714	513	170	1,005	2,286,232	163,678
CNVRs_2%	267	-	129	76	62	1,716	2,286,232	272,307

https://doi.org/10.1371/journal.pone.0303044.t002
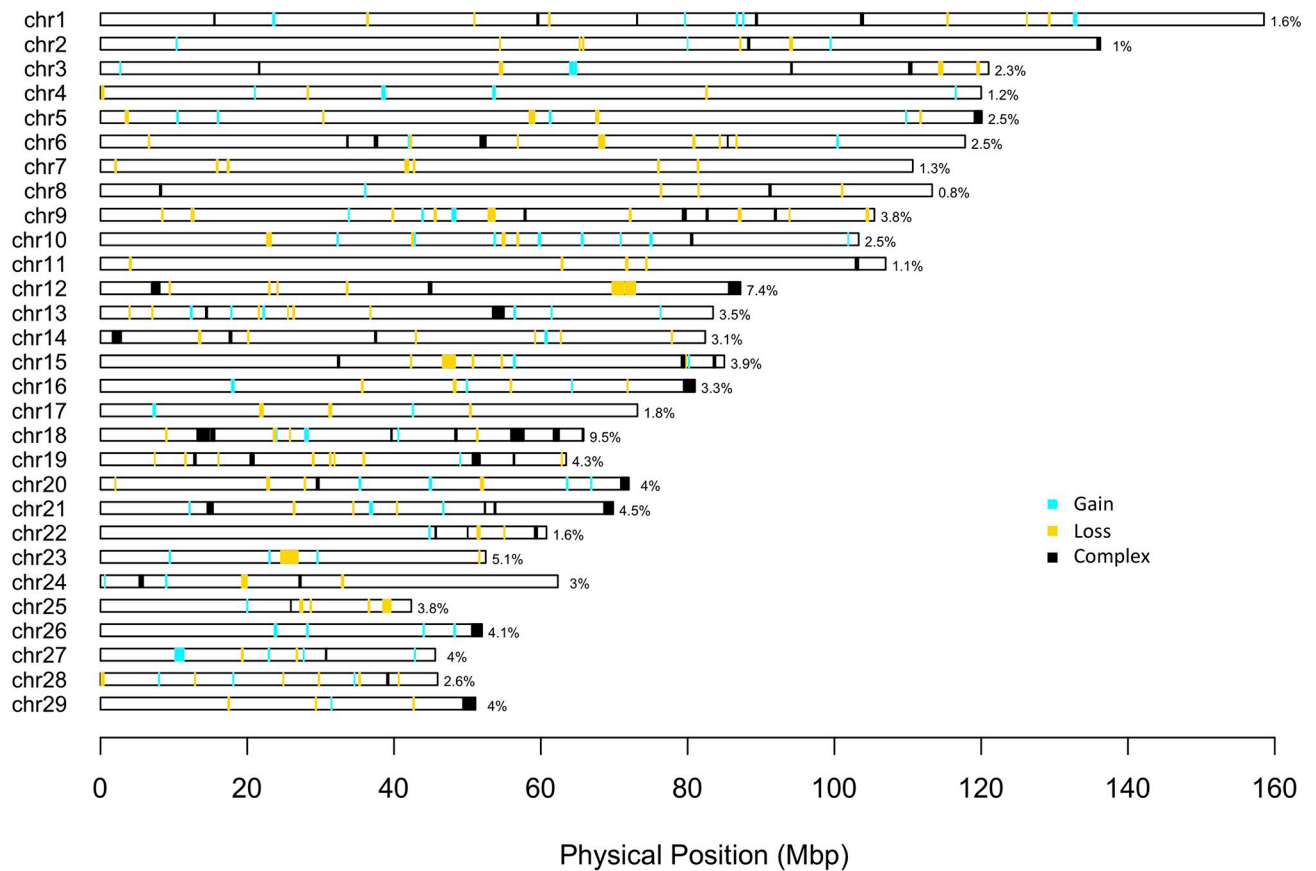
Fig 2. Physical distribution of the Copy Number Variants Regions (CNVRs) according to states (complex, gain and loss) on the *Bos taurus* ARS-UCD 1.2 assembly. Plotted CNVRs are those shared by at least 2% of individuals. Percentage values refer to the genomic proportion covered by CNVRs respect to the BTA length.

the database nomenclature (Fig 4). As Fig 4 shows, the most of traits associated with the genes annotated in the CNVRs are related to the phenotypes for which the Holstein population has been selected for years.

Fig 3. Principal component analysis results. A) Samples are coloured in black as unique Holstein breed; B) Samples are coloured according to the herds in which the cows were sampled.

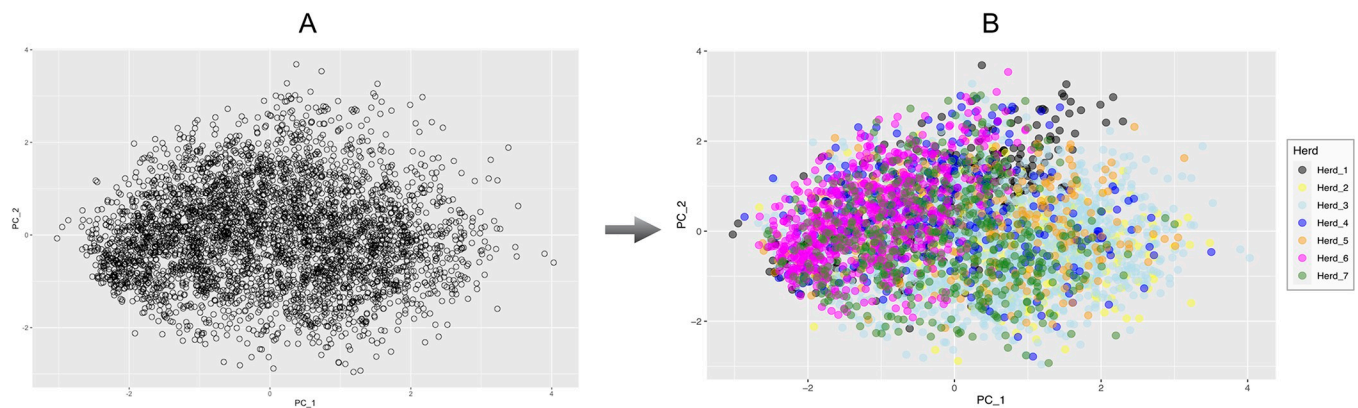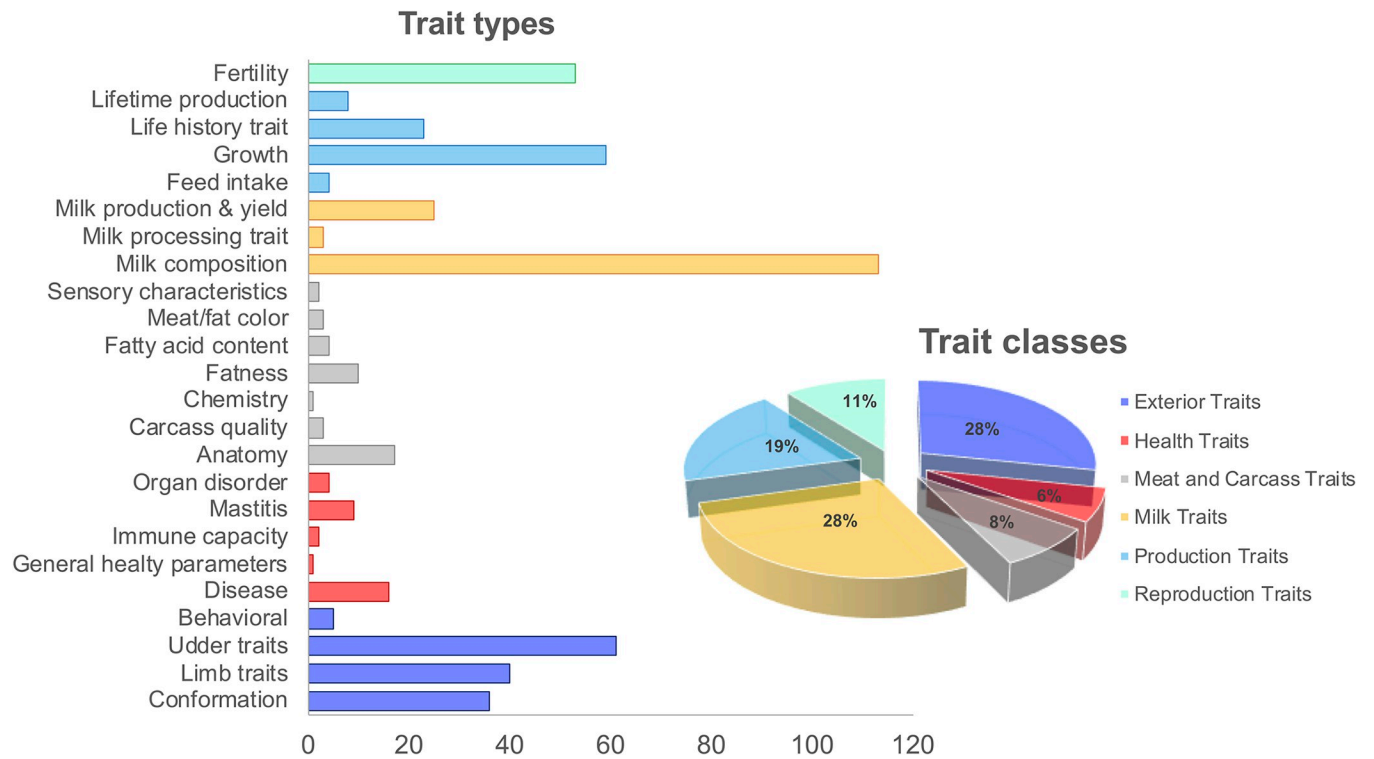**Fig 4. Graphical representation of QTL_terms (Trait types and Trait classes) associated with genes mapped in CNVRs.** Colours of Trait types corresponded to the ones in Trait classes.

Comparison with references

CNVRs here identified were compared with those identified in three other Holstein populations (comparison within breed) and in two different breeds (comparison among breed; one dairy cattle–Jersey; one dual-purpose cattle–Brown Swiss) (Table 3 and Fig 5 and S2 Fig and S4 Table). As reported in Table 3, the minimum and the maximum number of overlapping regions were 7 and 27, respectively.

The 48 CNVRs resulted overlapping regions (S4 Table) included 32 regions identified in others Holstein samples, i.e. CNVRs mapped in at least two studies (shared_HOL) as shown in Fig 5A. When the comparison was performed with the JER and BSW cattle, the 32

Table 3. Comparison with literature.

Platform	Software	Breed (N. of IDs)	Reference Genome	N. CNVRs[a]	N. overlapped CNVRs[b]	Overlapping length (bp)	Ref
			Comparison within breed				
Illumina HD; 50K; GGP150K	PennCNV	Holstein (96)	ARS-UCD1.2	36	7 (2.6%; 19.4%;)	1,239,370	[30]
Ilumina HD	PennCNV	Holstein (315)	ARS-UCD1.2	135	14 (5.2%; 10.4%)	1,374,082	[31]
Illumina HD	CNAM (SVS)	Holstein (242)	UMD3.1 remapped	112	23 (8.6%; 20.5%)	12,017,083	[21]
			Comparison among breeds with different aptitude				
Illumina HD	PennCNV	Jersey (107)	ARS-UCD1.2	142	15 (5.6%; 10.5%)	1,915,749	[31]
Illumina HD	CNAM (SVS)	Brown Swiss (1,116)	UMD3.1 remapped	233	27 (8.6%; 11.6%)	10,651,728	[32]

[a]When remapped, this number refers to the CNVRs resulting after the positions remapping
[b]Proportion of overlapping: calculated as n. overlapped CNVRs/n. CNVR in this study; n. overlapped CNVRs/n. CNVR in other study.

A

Lee_HOL

121

Duran_HOL

This study_HOL

91   0

235

15   6   4

3   0   2   2

0   0   0

29   0

Buffy_HOL

B

Lee_JER

127

Prinsen_BWS

This study_HOL

206   0

219

9   1   6

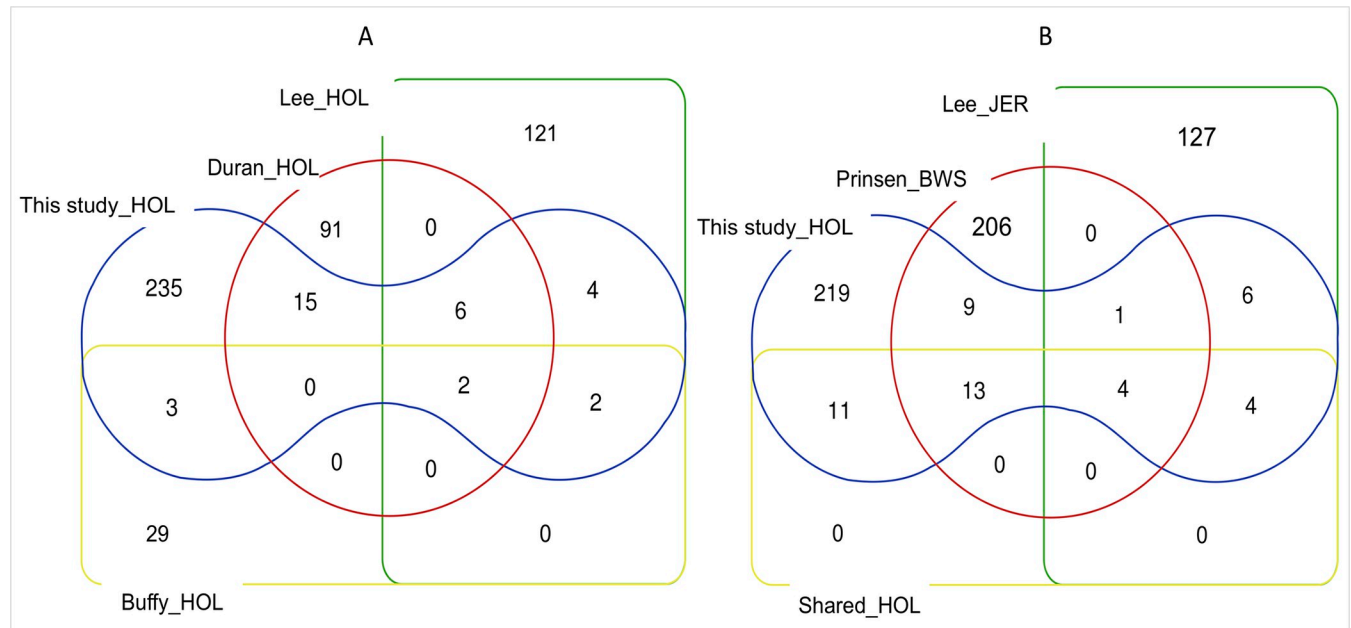11   13   4   4

0   0   0

0   0

Shared_HOL

Fig 5. Comparison of CNVRs identified in different Holstein populations (A) and in others two breeds (B). Shared_HOL are those CNVRs (n.32) identified in at least two studies (part A of this Fig).

shared_HOL regions in Fig 5B, resulted in 11 Holstein proprietary CNVRs and 4 ones found in all breeds. As in Fig 5B, the BSW breed shared the largest number of overlapping regions CNVRs. The total overlapping CNVR length was similar for those studies in which CNVs were identified with the same software (< 2 Mb–PennCNV and > 10 Mb–SVS, Table 3).

Discussion

In the literature there are several studies investigating genetic variability of Holstein' population using SNPs, and to increase knowledge on this breed, a large set of Italian Holstein cows has been inhere analyzed through CNVs detection. CNVs, a class of structural variation, can inform about population variability and are known to occur in the genome in response to environmental stressors, including positive selection, as a consequence of farming strategies [33].

This study, based on a medium density SNP chip, i.e. the Illumina GGP Bovine 100K, allowed the identification of a high number of CNVs in a substantial number of Holstein cows. The number of CNVs per sample (32, on average), is relatively higher compared to studies that rely on non-dense SNP chips, but lower compared to studies that rely on dense SNP chips or use sequences to call CNVs [34–36]. As reported in the majority of CNV mapping studies performed with Illumina SNP chips, the number of deletions calls was approximately 1.98 more recurrent than duplications [21, 31, 37]. The mean length of deletion calls inhere (90,439.4 bp) is bigger than the mean length of found gains (77,785.5 bp). Interestingly Lee *et al.* [31], using the Illumina BovineHD BeadChip, found that duplications are longer then deletions.

Overlapping CNVs resulted in 1,397 CNVRs covering 9.18% of the cattle genome. This value is much higher than the ones reported in the literature for Holsteins, which range from 0.5% to 2.8% [31, 38], but in line with the coverage found by Butty et al. [30], depending on the density of the SNP chip and the detection algorithm used [30, 39]. When CNV regions shared by at least 2% of the population were selected, the percentage of genome covered by

CNVRs decreased (2.9% of the autosomal genome length, Fig 2), a value similar to those reported by other authors [31, 38].

As shown in S1 Fig, CNVRs are not uniformly distributed on the autosomes, and the distribution of CNVRs according to their length class (Fig 1C) shows that the majority are short to medium in length and only a few are observed in the long classes, consistently with previous findings [31].

To visualize the genomic variability related to CNVs detected in our study population, we performed a Principal Component Analysis and the results in Fig 3A, at first glance, show that all animals are spread in the graph without any clustering tendency.

The homogeneous grouping in this study appears to be related to the fact that all the cows, although bred on different farms, undergone similar intensive farming system. Nevertheless, the genetic selection performed by the farmers seems to produce an effect: when the grouping animals by herd (Fig 3B) a slight clustering can be observed, mainly for animals in Herd_6 (magenta colour). In Herd_6, mating plans have been based on bulls from a unique AI center for years, while all other herds use sires from different semen providers [40]. When the gain/loss ratio was calculated in each herd to explain our findings, it was equal to 0.40 in Herd_6 (this value correspond to a loss/gain ratio = 2.40) and up to 0.49 in all the others herds (maximum value was 0.70 in Herd_5; loss/gain ratio = 1.41). The lower proportion of gain CNVs found in Herd_6 may be linked to the highest number of daughters for sire in Herd_6, with a reduction of variability in specific genomic regions. The lower number of common bulls across all herds (as reported by Punturiero et al. [40]) can explain the cows' distribution of Herd_6 respect to the ones belong to all other farms. In Herd_5, the number of daughters per sire is one of the lower.

Gene content and annotation

According to the David database (S3 Table), the genes annotated within the CNVRs were classified in 91 Go-Terms. The KEGG pathway analysis revealed that among the genes under analysis 56 are mainly represented in the pathway of immune system, namely, in the classes "Tuberculosis" and "Staphylococcus aureus infection", and in the pathway of thermogenesis. Disease resistance (or susceptibility) is a complex trait and interestingly it could be affected by genomic variations, as found by different authors reporting a substantial immune gene enhancement within CNV regions [21, 41–43].

The network constructed with ClueGO (S2 Fig) aligns with the results found with the David analysis. It's possible to see genes connected to different GO categories linked to nervous system, troponin complex, sensory perception of smell, nervous system process, together with the KEGG category of susceptibility to tuberculosis. Some genes are connected with more than one category, for example BOLA genes.

Variation in gene copy number leads to phenotypic variation among animals. After consulting the AnimalQTLdb for cattle we grouped the QTL in 24 trait types. As listed in the S3 Table and showed in Fig 4, the most common trait type is milk composition, for which 102 QTL were found. This result is in line with the expectations, being the animals part of commercial farms that sell milk for the dairy industry. Milk composition, together with udder conformation, fertility, and growth (more representative trait types in Fig 4) are all object traits of selection for high-productive breeds, such as the Holstein.

Noteworthy CNVRs and comparison with references

Nine CNVRs resulted over-represented due to a high number (> 2,000) of CNV defining these regions: 4 CNVRs do not harbor genes, and most of them are in loss state. The only

Table 4. CNVRs in common between our study and the ones found in Holstein and in different cattle breeds.

CNVR_ID this study	Chr	Start	End	State	Genes	QTL
					Common CNVRs (HOL, JER, BSW)	
cnvr_075	8	76336567	76348332	loss		
cnvr_121	12	71701903	71765886	complex		
cnvr_133	13	53463194	53511604	complex	SIRPB1	SIRPB1: Milk protein percentage (QTL: 174904)
cnvr_225	23	25953514	26064642	complex		
cnvr_225	23	26113327	26350925	complex		
					Only HOL CNVRs	
cnvr_034	3	119662571	119718948	loss	COPS9, OTOS	
cnvr_035	4	182210	217902	loss		
cnvr_055	6	37695352	37736960	complex		
cnvr_058	6	51884459	52200066	complex		
cnvr_068	7	17374656	17409367	complex		
cnvr_072	7	81385397	81392696	loss		
cnvr_137	14	1645654	2064157	complex	LY6D, LYNX1, LYPD2, SLURP1,THEM6, PSCA, TSNARE1, ARC, ADGRB1, JRK	LY6D: Milk fat percentage (QTL:33308; 166962; 161706), Milk protein percentage (QTL:161824)
cnvr_176	18	27856333	28303561	complex		
cnvr_181	18	57234258	57254890	complex		
cnvr_216	21	69040055	69788216	complex	C21H14orf180, TMEM179, INF2, ADSSL1, SIVA1, AKT1, ZBTB42, CEP170B, PLD4, AHNAK2, CLBA1, CDCA4, GPR132, JAG2, NUDT14, BRF1, BTBD6, PACS2, TEX22, MTA1, CRIP2, CRIP1, TEDC1, TMEM121	AKT1: Bovine respiratory disease susceptibility (QTL: 160320; 160321); BRF1: Conception rate (QTL: 123998)
cnvr_245	26	50598736	51990348	complex	KNDC1, ADGRA1, CFAP46, NKX6-2, INPP5A, BNIP3, JAKMIP3, DPYSL4, STK32C, LRRC27, PWWP2B	

https://doi.org/10.1371/journal.pone.0303044.t004

duplication region is the cnvr_234 identified on BTA 25 (in 2,400 cows) (S2 Table). In this CNVR, map the EEF2K and POLR3E genes that are involved in the cellular response to oxidative stress [44] and the host innate immune defense against viruses [45], respectively. Even for the genes mapped in the cnvr_024 (3,107 cows) on BTA 2 (PARD3B, NRP2) a roles in immune response was reported [46, 47]. Finally, the cnvr_069 located on BTA 7 (2,156 cows) overlaps the CNVR20 (complex state) identified by [30]. This region harbors five genes belonging to the family 2 of olfactory receptor genes (OR). CNVs are frequently found within OR genes and this variability may contribute to individual or breed-specific differences in olfactory capacity [48], which is also associated with feed intake and efficiency [49]. This aligns with the findings in our research; indeed, conducting gene ontology analysis with ClueGO (S2 Fig) yielded results for 35 genes in a copy number variation state linked with the following functional categories: sensory perception of smell, detection of stimulus involved in sensory perception, detection of chemical stimulus involved in sensory perception, olfactory perception activity and sensory perception of chemical stimulus. Nonetheless, these results only contribute to a small portion of our understanding given the size and complexity of this gene family comprising more than 1,000 known OR genes.

Regarding the comparison with references, as reported in Table 4, among the 267 CNVRs, 11 overlapped with the ones identified only in Holstein populations and 4 in all the considered breeds (Holstein, Jersey, and Brown). It is important to note that the size of the CNVRs identified in this study decrease after comparison (we reported only regions perfectly overlapping). This is particular evident for cnvr_225, splitted in two small regions as listed in Table 4. The

entire cnvr_225 harbour genes belonging to the BOLA family, a well known gene implicated in host immune response. In the cnvr_133, located on BTA 13 (both in loss and complex states, according to breeds, see S4 Table), lied the SIRPB1 gene, also involved in the immune response [50].

Across the identified CNVRs proper of the Holstein cows, a wider variability in the regions state can be observed, more than 70% are in fact in complex state. Only 4 CNVRs harbour genes. Among them, cnvr_137 contains genes such as LY6D, LYNX1, LYPD2, SLURP1, THEM6, PSCA, TSNARE1, and ARC associated to clinical mastitis in US Holstein dairy cows [51]. While the cnvr_245 includes the BNIP3 gene, that plays a critical role in inducting autophagy during heat stress and was associated with the immune response phenotype [52]. The same region partially overlaps the CNVR_1549_P (the region comprising the JAKMIP3, DPYSL4, STK32C, LRRC27, PWWP2B) resulted associated with clinical mastitis in Mexican Holstein Cattle [21].

Conclusions

The study provides novel insights into CNVs mapped within the Italian Holstein cows. To date, this is the only study that conducted a CNV analysis on such a large number of animals within this breed. Based on CNVs, the Principal Component Analysis (PCA) revealed a homogeneous distribution of cows, indicating a shared effect of the intensive farming system on these animals. The slight clustering observed among cows from the same farm implies that genetic selection may influence CNV distribution, underscoring the potential impact of selective breeding practices.

The functional analysis of genes annotated in the more common CNVRs revealed biological mechanism related to immune resistance to infection and adaptability. QTL linked with the main traits object of directional selection overlapped with many CNVRs here identified. Genes involved in immune response and defense against oxidative stress were identified within CNVRs, suggesting that genetic variability could affect the animals' ability to respond to environmental stressors.

The analysis of CNVs not only provides an additional dimension of genetic information, but also represents a valuable resource to optimise (new prospective) genomic selection in a more complete and accurate way.

Supporting information

S1 Table. CNVs identified in Holstein breed.
(XLSX)

S2 Table. List of the total CNVR (sheet_1); CNVRs identified in at least 2% of cows (sheet_2), ad list of CNV defining CNVRs identified in at least 2% of cows.
(XLSX)

S3 Table. Gene functional annotation from David database.
(XLSX)

S4 Table. CNVRs comparison with references.
(XLSX)

S1 Fig. Graphical representation of CNVRs number and mean CNVR coverage length on autosomes.
(TIF)

S2 Fig. ClueGo network of genes annotated in CNVRs identified in at least 2% of cows. (TIF)

Acknowledgments

Author Contributions

Conceptualization: Maria G. Strillacci.

Data curation: Francesca Bernini, Raffaella Milanesi.

Formal analysis: Andrea Delledonne, Chiara Punturiero, Carlotta Ferrari.

Funding acquisition: Alessandro Bagnato.

Investigation: Alessandro Bagnato.

Methodology: Maria G. Strillacci.

Project administration: Alessandro Bagnato.

Supervision: Alessandro Bagnato.

Writing – original draft: Andrea Delledonne, Chiara Punturiero, Carlotta Ferrari, Raffaella Milanesi, Maria G. Strillacci.

Writing – review & editing: Andrea Delledonne, Chiara Punturiero, Carlotta Ferrari, Francesca Bernini, Raffaella Milanesi, Alessandro Bagnato, Maria G. Strillacci.

References

1. Frantz LAF, Bradley DG, Larson G, Orlando L. Animal domestication in the era of ancient genomics. Nat Rev Genet 2020; 21:449–60. https://doi.org/10.1038/s41576-020-0225-0 PMID: 32265525

2. Egger-Danner C, Cole JB, Pryce JE, Gengler N, Heringstad B, Bradley A, et al. Invited review: overview of new traits and phenotyping strategies in dairy cattle with a focus on functional traits. Animal 2015; 9:191–207. https://doi.org/10.1017/S1751731114002614 PMID: 25387784

3. Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics 2001; 157:1819–29. https://doi.org/10.1093/genetics/157.4.1819 PMID: 11290733

4. Wiggans GR, Cole JB, Hubbard SM, Sonstegard TS. Genomic selection in dairy cattle: the USDA experience. Annu Rev Anim Biosci 2017; 5:309–27. https://doi.org/10.1146/annurev-animal-021815-111422 PMID: 27860491

5. Wang K, Bucan M. Copy number variation detection via high-density SNP genotyping. Cold Spring Harb Protoc 2008; 2008:pdb–top46. https://doi.org/10.1101/pdb.top46 PMID: 21356857

6. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazlsy C, Thorne N, et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 2007; 315:848–53. https://doi.org/10.1126/science.1136678 PMID: 17289997

7. Margareto J, Leis O, Larrarte E, Pomposo IC, Garibi JM, Lafuente JV. DNA copy number variation and gene expression analyses reveal the implication of specific oncogenes and genes in GBM. Cancer Invest 2009; 27:541–8. https://doi.org/10.1080/07357900802563044 PMID: 19219654

8. Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, et al. Mapping copy number variation by population-scale genome sequencing. Nature 2011; 470:59–65. https://doi.org/10.1038/nature09708 PMID: 21293372

9. Wright D, Boije H, Meadows JRS, Bed'Hom B, Gourichon D, Vieaud A, et al. Copy number variation in intron 1 of SOX5 causes the Pea-comb phenotype in chickens. PLoS Genet 2009; 5:e1000512. https://doi.org/10.1371/journal.pgen.1000512 PMID: 19521496

10. Olsson M, Meadows JRS, Truve K, Pielberg GR, Puppo F, Mauceli E, et al. A novel unstable duplication upstream of HAS2 predisposes to a breed-defining skin phenotype and a periodic fever syndrome in Chinese Shar-Pei dogs. PLoS Genet 2011; 7:e1001332. https://doi.org/10.1371/journal.pgen.1001332 PMID: 21437276

11. Kropatsch R, Dekomien G, Akkad DA, Gerding WM, Petrasch-Parwez E, Young ND, et al. SOX9 duplication linked to intersex in deer. PLoS One 2013; 8:e73734. https://doi.org/10.1371/journal.pone.0073734 PMID: 24040047

12. Venhoranta H, Pausch H, Wysocki M, Szczerbal I, Hänninen R, Taponen J, et al. Ectopic KIT copy number variation underlies impaired migration of primordial germ cells associated with gonadal hypoplasia in cattle (Bos taurus). PLoS One 2013; 8:e75659. https://doi.org/10.1371/journal.pone.0075659 PMID: 24086604

13. Awasthi Mishra N, Drögemüller C, Jagannathan V, Keller I, Wüthrich D, Bruggmann R, et al. A structural variant in the 5'-flanking region of the TWIST2 gene affects melanocyte development in belted cattle. PLoS One 2017; 12:e0180170. https://doi.org/10.1371/journal.pone.0180170 PMID: 28658273

14. Pierce MD, Dzama K, Muchadeyi FC. Genetic diversity of seven cattle breeds inferred using copy number variations. Front Genet 2018; 9:163. https://doi.org/10.3389/fgene.2018.00163 PMID: 29868114

15. Xu L, Hou Y, Bickhart DM, Zhou Y, Hay EHA, Song J, et al. Population-genetic properties of differentiated copy number variations in cattle. Sci Rep 2016; 6:23161. https://doi.org/10.1038/srep23161 PMID: 27005566

16. Arendt M, Cairns KM, Ballard JWO, Savolainen P, Axelsson E. Diet adaptation in dog reflects spread of prehistoric agriculture. Heredity (Edinb) 2016; 117:301–6. https://doi.org/10.1038/hdy.2016.48 PMID: 27406651

17. Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, et al. Diet and the evolution of human amylase gene copy number variation. Nat Genet 2007; 39:1256–60. https://doi.org/10.1038/ng2123 PMID: 17828263

18. Hou Y, Bickhart DM, Chung H, Hutchison JL, Norman HD, Connor EE, et al. Analysis of copy number variations in Holstein cows identify potential mechanisms contributing to differences in residual feed intake. Funct Integr Genomics 2012; 12:717–23. https://doi.org/10.1007/s10142-012-0295-y PMID: 22991089

19. Glick G, Shirak A, Seroussi E, Zeron Y, Ezra E, Weller JI, et al. Fine Mapping of a QTL for Fertility on BTA7 and Its Association With a CNV in the Israeli Holsteins. G3: Genes| Genomes| Genetics 2011; 1:65–74. https://doi.org/10.1534/g3.111.000299 PMID: 22384319

20. Xu L, Cole JB, Bickhart DM, Hou Y, Song J, VanRaden PM, et al. Genome wide CNV analysis reveals additional variants associated with milk production traits in Holsteins. BMC Genomics 2014; 15:1–10.

21. Durán Aguilar M, Román Ponce SI, Ruiz López FJ, González Padilla E, Vásquez Peláez CG, Bagnato A, et al. Genome-wide association study for milk somatic cell score in holstein cattle using copy number variation as markers. J Anim Breed Genet 2017; 134:49–59. https://doi.org/10.1111/jbg.12238 PMID: 27578198

22. Hay EHA, Utsunomiya YT, Xu L, Zhou Y, Neves HHR, Carvalheiro R, et al. Genomic predictions combining SNP markers and copy number variations in Nellore cattle. BMC Genomics 2018; 19:1–8.

23. Pinto D, Darvishi K, Shi X, Rajan D, Rigler D, Fitzgerald T, et al. Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. Nat Biotechnol 2011; 29:512–20. https://doi.org/10.1038/nbt.1852 PMID: 21552272

24. Diskin SJ, Li M, Hou C, Yang S, Glessner J, Hakonarson H, et al. Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. Nucleic Acids Res 2008; 36:e126–e126. https://doi.org/10.1093/nar/gkn556 PMID: 18784189

25. Wickham H. ggplot2: Elegant Graphics for Data Analysis. New York: 2016.

26. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010; 26:841–842. https://doi.org/10.1093/bioinformatics/btq033 PMID: 20110278

27. Zhou J, Liu L, Lopdell TJ, Garrick DJ, Shi Y. HandyCNV: Standardized Summary, Annotation, Comparison, and Visualization of Copy Number Variant, Copy Number Variation Region, and Runs of Homozygosity. Front Genet 2021; 12:731355. https://doi.org/10.3389/fgene.2021.731355 PMID: 34603390

28. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003; 13:2498–504. https://doi.org/10.1101/gr.1239303 PMID: 14597658

29. Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, et al. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics 2009; 25:1091–3. https://doi.org/10.1093/bioinformatics/btp101 PMID: 19237447

30. Butty AM, Chud TCS, Miglior F, Schenkel FS, Kommadath A, Krivushin K, et al. High confidence copy number variants identified in Holstein dairy cattle from whole genome sequence and genotype array data. Scientific Reports 2020 10:1 2020; 10:1–13. https://doi.org/10.1038/s41598-020-64680-3 PMID: 32415111

31. Lee Y-L, Bosse M, Mullaart E, Groenen MAM, Veerkamp RF, Bouwman AC. Functional and population genetic features of copy number variations in two dairy cattle populations. BMC Genomics 2020; 21:89. https://doi.org/10.1186/s12864-020-6496-1 PMID: 31992181

32. Prinsen RTMM, Rossoni A, Gredler B, Bieber A Bagnato A, Strillacci MG. A genome wide association study between CNVs and quantitative traits in Brown Swiss cattle. Livest Sci 2017;202. https://doi.org/10.1016/j.livsci.2017.05.011.

33. Strillacci MG, Gorla E, Cozzi MC, Vevey M, Genova F, Scienski K, et al. A copy number variant scan in the autochthonous Valdostana Red Pied cattle breed and comparison with specialized dairy populations. PLoS One 2018; 13:1–18. https://doi.org/10.1371/journal.pone.0204669 PMID: 30261013

34. Ahmad SF, Singh A, Panda S, Malla WA, Kumar A, Dutt T. Genome-wide elucidation of CNV regions and their association with production and reproduction traits in composite Vrindavani cattle. Gene 2022; 830:146510. https://doi.org/10.1016/j.gene.2022.146510 PMID: 35447249

35. Sassi N Ben, González-Recio Ó, de Paz-Del Río R, Rodríguez-Ramilo ST, Fernández AI. Associated effects of copy number variants on economically important traits in Spanish Holstein dairy cattle. J Dairy Sci 2016; 99:6371–80. https://doi.org/10.3168/jds.2015-10487 PMID: 27209136

36. Hou Y, Liu GE, Bickhart DM, Cardone MF, Wang K, Kim E, et al. Genomic characteristics of cattle copy number variations. BMC Genomics 2011; 12:1–11. https://doi.org/10.1186/1471-2164-12-127 PMID: 21345189

37. Prinsen RTMM Strillacci MG, Schiavini F, Santus E, Rossoni A, Maurer V, et al. A genome-wide scan of copy number variants using high-density SNPs in Brown Swiss dairy cattle. Livest Sci 2016;191. https://doi.org/10.1016/j.livsci.2016.08.006.

38. Jiang L, Jiang J, Yang J, Liu X, Wang J, Wang H, et al. Genome-wide detection of copy number variations using high-density SNP genotyping platforms in Holsteins. BMC Genomics 2013; 14:1–10. https://doi.org/10.1186/1471-2164-14-131/TABLES/2.

39. Xu L, Hou Y, Bickhart DM, Song J, Liu GE. Comparative analysis of CNV calling algorithms: literature survey and a case study using bovine high-density SNP data. Microarrays 2013; 2:171–85. https://doi.org/10.3390/microarrays2030171 PMID: 27605188

40. Punturiero C, Milanesi R, Bernini F, Delledonne A, Bagnato A, Strillacci MG. Genomic approach to manage genetic variability in dairy farms. Ital J Anim Sci 2023; 22:769–83. https://doi.org/10.1080/1828051X.2023.2243977.

41. Suchocki T, Szyda J. Genome-wide association study for semen production traits in Holstein-Friesian bulls. J Dairy Sci 2015; 98:5774–80. https://doi.org/10.3168/jds.2014-8951 PMID: 26051317

42. Szyda J, Mielczarek M, Frąszczak M, Minozzi G, Williams JL, Wojdak-Maksymiec K. The genetic background of clinical mastitis in Holstein-Friesian cattle. Animal 2019; 13:2156–63. https://doi.org/10.1017/S1751731119000338 PMID: 30835192

43. Lee Y-L, Takeda H, Costa Monteiro Moreira G, Karim L, Mullaart E, Coppieters W, et al. A 12 kb multi-allelic copy number variation encompassing a GC gene enhancer is associated with mastitis resistance in dairy cattle. PLoS Genet 2021; 17:e1009331. https://doi.org/10.1371/journal.pgen.1009331 PMID: 34288907

44. Sanchez M, Lin Y, Yang C-C, McQuary P, Campos AR, Blanc PA, et al. Cross talk between eIF2α and eEF2 phosphorylation pathways optimizes translational arrest in response to oxidative stress. IScience 2019; 20:466–80.

45. Ramanathan A, Weintraub M, Orlovetskie N, Serruya R, Mani D, Marcu O, et al. A mutation in POLR3E impairs antiviral immune response and RNA polymerase III. Proceedings of the National Academy of Sciences 2020; 117:22113–21. https://doi.org/10.1073/pnas.2009947117 PMID: 32843346

46. Schramek H, Sarközi R, Lauterberg C, Kronbichler A, Pirklbauer M, Albrecht R, et al. Neuropilin-1 and neuropilin-2 are differentially expressed in human proteinuric nephropathies and cytokine-stimulated proximal tubular cells. Laboratory Investigation 2009; 89:1304–16. https://doi.org/10.1038/labinvest.2009.96 PMID: 19736548

47. Raphaka K, Matika O, Sánchez-Molano E, Mrode R, Coffey MP, Riggio V, et al. Genomic regions underlying susceptibility to bovine tuberculosis in Holstein-Friesian cattle. BMC Genet 2017; 18:1–10.

48. Hasin Y, Olender T, Khen M, Gonzaga-Jauregui C, Kim PM, Urban AE, et al. High-resolution copy-number variation map reflects human olfactory receptor diversity and evolution. PLoS Genet 2008; 4: e1000249. https://doi.org/10.1371/journal.pgen.1000249 PMID: 18989455

49. Connor EE, Zhou Y, Liu GE. The essence of appetite: does olfactory receptor variation play a role? J Anim Sci 2018; 96:1551–8. https://doi.org/10.1093/jas/sky068 PMID: 29534194

50. Van Beek EM, Cochrane F, Barclay AN, van den Berg TK. Signal regulatory proteins in the immune system. The Journal of Immunology 2005; 175:7781–7. https://doi.org/10.4049/jimmunol.175.12.7781 PMID: 16339510

51. Tiezzi F, Parker-Gaddis KL, Cole JB, Clay JS, Maltecca C. A genome-wide association study for clinical mastitis in first parity US Holstein cows using single-step approach and genomic matrix re-weighting procedure. PLoS One 2015; 10:e0114919. https://doi.org/10.1371/journal.pone.0114919 PMID: 25658712

52. Livernois AM, Mallard BA, Cartwright SL, Cánovas A. Heat stress and immune response phenotype affect DNA methylation in blood mononuclear cells from Holstein dairy cows. Sci Rep 2021; 11:11371. https://doi.org/10.1038/s41598-021-89951-5 PMID: 34059695