UNIVERSITÀ DEGLI STUDI DI MILANO

SCUOLA DI DOTTORATO IN SANITÀ E PRODUZIONI ANIMALI: SCIENZA, TECNOLOGIA E BIOTECNOLOGIE

A Medium Resolution SNP Array Based Copy Number Variants Scan in

Tesi di: Laura PELLEGRINO

Docente guida: Prof. Alessandro BAGNATO

Ciclo: XXV

Anno Accademico 2011/2012

ABSTRACT

Recent reports indicate copy number variations (CNVs) to be functionally significant. This study presents a medium resolution map of CNV regions (CNVRs) in Brown Swiss dairy cattle, from to this day, the largest CNV genome scan in any cattle breed. We genotyped 1,342 bulls and after quality filtering on males we called CNVs with PennCNV and with "Copy Number Analysis Module" (CNAM) of SVS7 software (Goldenhelix) for a total of 46,728 loci anchored on the UMD3.1 assembly. We corrected for sequence composition flanking each SNP and employed principal component analysis for CNAM to correct for technical background noise to reduce false positive calls. PennCNV and SVS7 identified a total of 5,099 and 1,289 CNVs segregating in 632 and 651 bulls respectively. These were summarized at the population level into 1,101 (220 losses, 774 gains, 107 complex) and 277 (185 losses, 56 gains and 36 complex) CNVRs, covering 682 Mb (27.14%) and 33.7 Mb (1.35%) of the autosome, respectively. We then obtained the consensus between the two CNV scans using the approaches suggested by Redon *et al*. (2006), union set, and by Wain *et al*. (2009), intersection, covering 146 Mb (5.88%) and 17.1 Mb (0.68%), respectively. CNVRs were annotated with the bovine Ensembl gene set v69 and tested for enrichment of *GO* terms using DAVID database. Consensus CNVRs are enriched for protein-coding genes.
Go analysis identified genes (Bonferroni corrected) in the CNVRs related to cytoplasm, intercellular part, cellular processes, cytoplasmic part, and intracellular organelles.

Acknowledgement. This study funded by EC-FP7/2007-2013, agreement n°222664, "Quantomics".

TABLE OF CONTENTS

CHAPTER 1 – INTRODUCTION

1.1 GENOME AND STRUCTURAL VARIATIONS

Over the last few years, genomic studies have progressed rapidly and innovative High Throughput Technologies (HTT) have been carried out to investigate the structure and the characteristics of the human and bovine genome generating a number of valuable information. Since the similarity between human and bovine genome is 80% and it is higher than the analogy between human and mice, the bovine genome may be considered a model organism to better understand the genetic complexity and variation of human evolution (http://www.csiro.au/en/Outcomes/Food-and-Agriculture/Bovine-genome-decoded/Similarities-between-cow-and-human-DNA.aspx).

The human genome is comprised of 6 billion nucleotides, and the DNA is organized into two sets of 23 chromosomes, one set inherited from each parent (http://cnv.gene-quantification.info/), while the bovine genome is enriched of 3 billion nucleotides of DNA packaged into two sets of 30 chromosomes. The DNA encodes roughly 27,000 and 22,000 genes for the human and bovine genomes, respectively (Elsik *et al*., 2009).

It was generally thought that genes were almost always present in two copies in a genome; but recent discoveries have revealed that genes are sometimes present in one, three, or more than three copies; in a few rare instances genes are missing altogether (Wain L.V. *et al*., 2009).

Large segments of DNA, ranging in size from thousands to millions DNA bases, can vary in copy-number (Iafrate *et al*., 2004, Redon *et al*., 2006). Until now, the covering of copy number variants (CNVs) corresponds to 12% and 4.6% on human and bovine genomes, respectively (Redon *et al*., 2006; Hou *et al*., 2011). Such copy number variations can encompass genes leading to dosage imbalances, and differences in DNA sequence may contribute to the uniqueness of the genomes. These changes that can influence most traits including susceptibility to disease, are important sources of genetic and phenotypic variation (Nguyen *et al*., 2006, Xu *et al*., 2011). It was thought that single nucleotide polymorphisms (SNPs) in DNA were the most prevalent and important form of genetic variation (Van Tassell *et al*., 2008; Matukumalli *et al*., 2009). The current studies reveal that CNVs comprise at least three times the total nucleotide content of SNPs, (http://cnv.gene-quantification.info/).

The understanding of the mechanisms of CNV formation may help to better comprehend genome evolution and to create a more

accurate and complete genome reference sequence for different species including human and bovine. The creation of the CNV map may be adopted to improve scientific and genetic research in different fields that can be summarized as:

- Identification of genes responsible for common diseases;
- Study of familial genetic conditions;
- Detection of defects caused by chromosomal rearrangements.

The genomic studies and their use in selection programs are having a strong impact in dairy cattle selection (Hou *et al*., 2011). CNVs may represent an important source of information to integrate the genomic selection programs, because the availability of their knowledge is expected to improve the genomic breeding values (GEBVs) of animals (Bae *et al*., 2010).

1.1.1 REFERENCE GENOME

1.1.1.1 General Description

The reference genome is a nucleic acid sequence database, assembled by researches as a guide, on which new genomes are built and compared; it aims to provide, for a huge number of species (*Human, Mouse, Zebrafish, Primates, Rodents, Laurasiatheria, Afrotheria, Xenarthra, Other Mammals, Birds and Reptiles, Fish, Amphibians, Other Cordathes, Other Eukaryotes*), a comprehensive, integrated, non-redundant, well-

annotated set of sequences, including genomic DNA, transcripts, and proteins. The sequencing of DNA was obtained from different donors providing a good approximation, especially in the regions with high polymorphism, of the nucleic acid composition.

The human reference genome is maintained and improved by the Genome Reference Consortium (GRC) by building new alignments that contain fewer gaps, and fixing misrepresentations in the sequences. The human genome was assembled using the isolated mapped and sequenced bacterial artificial chromosomes (BACs). The disadvantage of BAC method is the high cost and to reduce that, the whole genome shotgun (WGS) method has been applied to improve the human assembly.

The bovine reference genome is obtained and improved by the Bovine Genome Sequencing and Analysis Consortium (2009). The WGS and the hierarchical (BAC clone) methods were used to create the bovine genome assembly (Zimin *et al*., 2009, Liu *et al*., 2009). The sequencing combines BAC shotgun and WGS reads from small insert libraries as well as BAC end sequences (BES). The DNA for WGS libraries was from white blood cells from a Hereford cow (L1 Dominette), and the DNA for BAC libraries was from a Hereford bull (L1 Domino), the sire of the ancestral animal (Liu *et al*., 2009, The bovine HapMap consortium 2009). In the last years, public browsers (Ensembl, NCBI, USCS

Genome Browser) including several reference assemblies, have been made available. The first *Bos Taurus* assembly (UMD2) was created from University of Meriland (Zimin *et al*., 2009). Table 1 shows the differences between the two most recent assemblies: Bos Taurus-UMD3.1 (2009) and Btau4.6.1 (2011) (http://www.ncbi.nlm.nih.gov/genome/82).

Organism	Assembly	Chr	Size (Mb)	GC %	Gene	Protein
Bos Taurus	Bos_taurus_UMD3.1	30	2670.42	41.9	27155	22070
Bos Taurus	Btau_4.6.1	31	2983.32	4.5	29754	23594

Table 1: Differences between the two most recent assemblies.

1.1.1.2 Database of elements annotated on Reference Genome

In the reference genome several elements are annotated: gene, proteins, trascriptome, non-coding RNA (nc-RNA) and structural variation (Table 2). The nc-RNA are represented by: transfer RNA (**tRNA),** transfer RNA located in the mitochondrial genome **(Mt-tRNA),** ribosomal RNA (**rRNA),** small cytoplasmic RNA **(scRNA),** small nuclear RNA **(snRNA),** small nucleolar RNA **(snoRNA),** microRNA precursors **(miRNA),** miscellaneous other RNA **(misc_RNA),** Long intergenic non-coding RNAs **(lincRNA)**

1.1.2 STRUCTURAL VARIATIONS

1.1.2.1 Structural Variation as a Source of Genetic Diversity

The existence of the genome structural variations (SVs) was described starting from the early years of the XX century in Drosophila (Bridge 1921, 1936), in 1950s in cattle (Knudsen 1958), and in humans from the 1980s (Goossens *et al*., 1980).

The progress in high-throughput genome scan technologies, has allowed the identification of sequence variation and several types of structural variation (genomic or chromosomal rearrangements) on the whole genome (Alkan *et al*., 2011). Genetic variation is mostly present in different forms and sizes ranging from the ubiquitous SNPs, to fine-scale copy number change such as small insertions and deletions (INDELs), microsatellite and minisatellite repeats, to larger scale structural variants from several kilobases to megabases such as inversions, translocations and CNVs (Feuk *et al*., 2006; Mills *et al*., 2006; Conrad and Hurles 2007). These variation and polymorphisms represent the dynamic genome architecture and underline the differences between subjects. A list of the most common genome variations is reported in Table 2; Figure 1 (Feuk *et al*., 2006) is a graphical scheme for deletion, insertion, inversion, copy number variant and segmental duplication.

Single Nucleotide Polymorphism (SNP): DNA sequence variation occurring when a single nucleotide (A, T, C, and G) in the genome differs respect to the reference genome.
Copy number variant (CNV): DNA segment of at least 1 kb to several megabases in size that differs in copy number respect to the reference genome.
Copy number polymorphism (CNP): CNV which appears in more than 1% in a population.
Inversion: DNA segment with reversed orientation respect to the major sequence of a chromosome.
Segmental Duplication (SD) or low copy repeats (LCRs): DNA very similar blocks that occur in more than one site within haploid genome (> 90%sequence identity). SDs can also be CNVs.
Traslocation: DNA segment with a modified position in the genome that has no gain or loss in DNA content. The traslocation can occur within a chromosome (intra-chromosomal) or between different chromosomes (inter-chromosomal).

Table 2: List of the genome variations (Feuk *et al*., 2006)

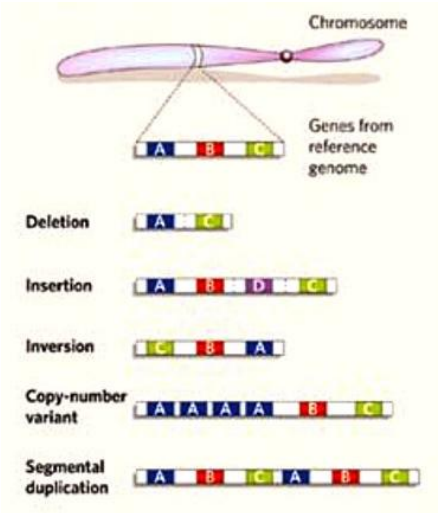


Figure 1: Graphical scheme for deletion, insertion, inversion, copy number variant and segmental duplication (Feuk *et al*. 2006).

1.1.2.2 Copy Number Variants (CNVs)

The copy number variants (CNVs) are a type of DNA variation including deletion, duplication and insertion of DNA fragments ranging from 1 kilobase (kb) to several megabase (Mb) and present variable copy number in comparison with a reference genome (Redon *et al*., 2006, Feuk *et al*., 2006). The first association of CNV with a phenotype was described by Bridge (1936), with the duplication of the Bar gene in *Drosophila melanogaster*, and recently CNV map for the *Drosophila melanogaster* was published (Dopman and Hartl 2007; Emerson *et al*., 2008; Zhou *et al*., 2008). In cattle in the 1950s Knudsen (1958) detected chromosomal translocation involvement in the reduced fertility of carrier bulls. In the human genome, this type of submicroscopic structural variation has been detected by Goossens *et al*., (1980) starting from the 1980s in a fraction of α-globin loci.

The CNVs can be inherited or may occur through *de novo* formation. Moreover, Gu *et al.*, (2008) have been shown that the CNVs can be manifested on germ line and somatic cells. The major mechanisms responsible for the CNV formation are the non-allelic homologous recombination (NAHR), the non-

homologous end-joining (NHEJ), the fork stalling and the template switching (FoSTeS).

The NAHR method requires, as recombination substrates, two segmental duplications (SDs) (Sharp *et al*., 2005) or low copy repeats (LCRs) (Lupski 2003) with a sufficient size (usually from 10 to 300 kb) and with high homology (major than 95-97%). Usually the frequency of deletions should be higher than the frequency of duplication because of biological reasons (Gu *et al*., 2008). However, Conrad *et al*., (2010) observed that the main rearrangement mechanism of large CNVs size was NAHR (Figure 2).

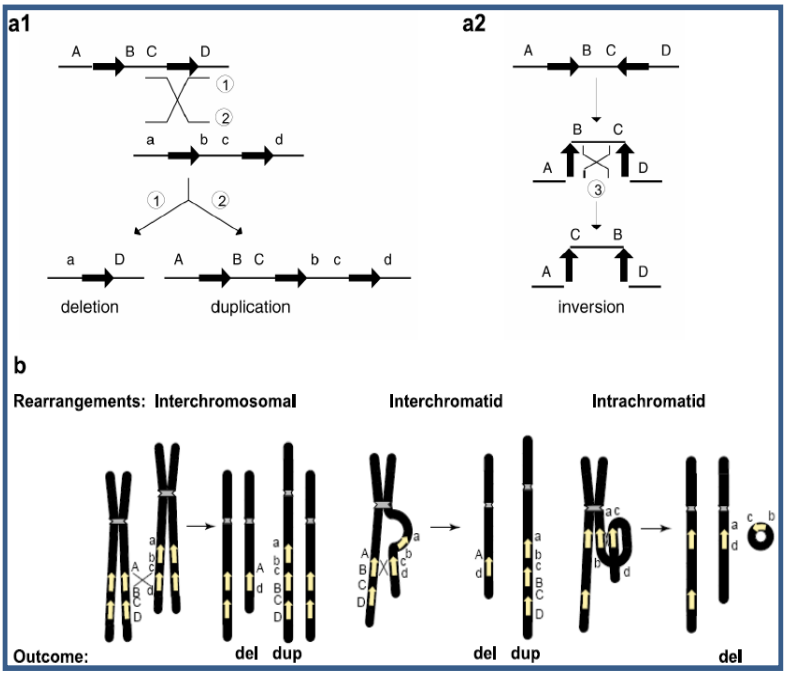


Figure 2: The genomic rearrangements (From Gu *et al.*, 2008). Genomic rearrangements resulting from recombination between low-copy repeats (LCRs). LCRs are depicted as black arrows with the orientation indicated by the direction of the arrowhead. Capital letters above the thin horizontal lines refer to the flanking unique sequences (for example, A). Homologues on the other strand (can be another chromatid or the homologous chromosome) are also shown (for example, a). Thin diagonal lines refer to a recombination event with the results shown by numbers 1, 2 and 3. a1 Recombination between direct repeats results in deletion and/or duplication. a2 Recombination between inverted repeats results in an inversion. b. Schematic representation of reciprocal duplications and deletions mediated by interchromosomal (left), interchromatid (middle) and intrachromatid (right) non-allelic homologous recombination (NAHR) using LCR pairs in direct orientation. Chromosomes are shown in black, with the centromere depicted by hashed lines. Yellow arrows depict LCRs. Letters adjacent to the chromatids refer

to the flanking unique sequence (for example, A, a). Interchromosomal and interchromatid NAHR between LCRs in direct orientation result in reciprocal duplication and deletion, whereas intrachromatid NAHR only creates deletion. Signatures of homologous recombination include the sequence identity of the substrates (LCRs) used for NAHR, recombination hotspots within the LCRs, and evidence for gene conversion at the crossovers within the LCRs.

The NHEJ is a pathway that repairs double-strand breaks in DNA and can affect some simple non-recurrent rearrangements. NHEJ proceeds in four steps: detection of DBS; molecular bridging of both broken DNA ends; modification of the ends to make them compatible and ligatable; and the final ligation step. This mechanism is more prevalent in unstable regions of the genome, as in the subtelomeric regions (Nguyen *et al*., 2006; Kim *et al*., 2008), and it has been implicated in different genomic disorders (Shaw and Lupski 2005).

The FoSTeS model is the main DNA replication-based mechanism for complex rearrangements that are induced by errors in the replication procedure (Figure 3).

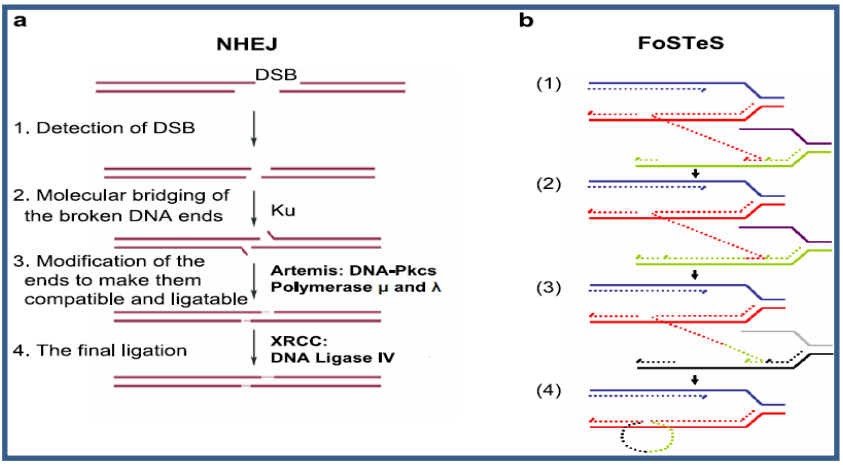


Figure 3: Genomic rearrangement mechanisms (From Gu *et al*.,2008). a. Non-homologous end-joining (NHEJ) in vertebrates. A double-stranded DNA break (DSB) occurs and is repaired via NHEJ mechanism. The two thick lines depict two DNA strands with DSB, the thin segments in the middle represent the modifications which the ends have gone through before the final ligation. At step 3 some addition or deletion of bases may be required, leaving behind a 'signature' of NHEJ. b. After the original stalling of the replication fork (dark blue and red, solid lines), the lagging strand (red, dotted line) disengages and anneals to a second fork (purple and green, solid lines) via micro homology (1), followed by (2) extension of the now 'primed' second fork and DNA synthesis (green, dotted line). After the fork disengages (3), the tethered original fork (dark blue and red, solid lines) with its lagging strand (red and green, dotted lines) could invade a third fork (gray and black, solid lines). Dotted lines represent newly synthesized DNA. Serial replication fork disengaging and lagging strand invasion could occur several times (e.g. FoSTeS x 2, FoSTeS x 3, ... etc.) before (4) resumption of replication on the original template.

The local genomic architecture is crucial for all the three models described above. In fact, it has been show that the CNVs are not

distributed uniformly in the genome, but the higher relative frequency of CNVs are correlated with the distance to the telomeres and centromeric regions, and to the simple tandem repeat sequences and segmental duplication (SDs) (Nguyen *et al*., 2006). Also, the genomic waves (variation in hybridization intensity) may affect the CNV formation as reported by Diskin *et al*., (2008).

Redon *et al*., (2006) defined the copy number variants regions (CNVRs) as the union or merge of the overlapping CNVs call. The CNVRs can be represented from duplication events, deletion events and even complex events, and their distribution is non-uniform along the genome but preferentially clustered SDs (Bailey *et al*., 2002). These CNVRs can encompass hundreds of genes, functional elements, disease loci, and segmental duplications (Redon *et al*., 2006).

The accuracy of CNV boundaries is affected by multiple factors as robustness of the statistical method, batch effects, population stratification, and differences between experiments (Dellinger *et al*., 2010).

1.1.3 COPY NUMBER VARIANTS IN HUMAN GENOME

1.1.3.1 CNV studies in human genome

Singleton *et al*., (2003), Iafrate *et al*., (2004), Sebat *et al*., (2004) and Redon *et al*., (2006) were the first researches who showed the distribution of CNVs in the human genome on genome-wide scan and reported several CNVs with a low size and breakpoint resolution.

Singleton *et al*. (2003) were the first to demonstrate mutations, more specific a triplication event, in the alpha-synuclein (SNCA), implicated in Parkinson's disease.

In the brief communication of Iafrate *et al*., (2004) it has been reported the detection of large-scale copy-number variations (LCVs) in the genome of unrelated individuals using the array-based comparative genomic hybridization (array CGH). They showed that the LCVs are not limited to the intergenic or intronic regions, and some of these variants were located close to loci associated with human genetic syndromes or with cancer. This suggests that the variations could lead to chromosomal rearrangements responsible for diseases and phenotypic variation by influencing the expression of genes.

Sebat *et al*., (2004) found a relationship between copy number polymorphism (CNPs) and susceptibility to health problems such as neurological disease, cancer, and obesity.

The first-generation CNV map of the human genome came in 2006 from Redon *et al*. In this study 270 individuals were genotyped from four populations with ancestry in Europe, Africa and Asia using different technologies and the importance of CNV in genetic diversity and evolution was underlined.

The identification of CNVs into whole-genome exploded in 2006 and 2007 with a supplementary issue of reviews on Nature Genetics journal (Volume 39, S1, 2007). These researches focused on how to integrate the CNVs datasets in genome wide association studies and in clinical diagnostics, and on the description of the different technologies to detect CNV, their impact on population and potential contributions on phenotypic variation.

1.1.3.2 Association with phenotypic variation and disease

The aim of different studies in humans was the inclusion of CNVs in the whole-genome association scans, and until now, several strong associations with neurodegenerative and neurodevelopmental diseases have been detected (De Cid *et al*., 2009; Glessner *et al*., 2010).

Recent publications have reviewed the effects of CNVs on gene expression and on simple genomic diseases as mental retardation and seizure (Sharp *et al*., 2008), and on several complex human diseases like Autism spectrum discorder (Pinto *et al*., 2010, Sebat *et al*., 2007), Schizophrenia (Xu *et al*., 2008, Walsh *et al*., 2008, Stefannson *et al*., 2008, 2009), Parkinson (Singleton *et al*., 2003), Crohn's disease (Aldhous *et al*., 2010), HIV/AIDS susceptibility (Gonzalez *et al*., 2005) and cancer (Campbell *et al*., 2008, Diskin *et al*., 2009).

1.1.4 COPY NUMBER VARIANTS IN BOVINE GENOME

1.1.4.1 CNV studies in bovine genome

In the bovine species (*Bos Taurus*, *Bos Indicus*) and Zebu several authors (Fadista *et al*., 2010, Liu *et al*., 2010, Hou *et al*., 2011, Bae *et al*., 2010, Matukumalli *et al*. 2009, Bickhart *et al*., 2012) have detected the CNVs using different methods and platforms. In the Table 3 a list of these studies is reported.

References	Breeds	Technologies	# CNVRs	% Coverage /(CNVR in Mb)
Bickhart et al. (2012)	Angus, Holstein, Hereford, Nelore	Read-depth analysis (Illumina GAIIx)	1265	2.1/(55.6)
Hou et al. (2011)	Holstein, Angus, Limousin, Hereford, Jersey, Charolais, BrownSwiss, Piedmontese, Ramagnola, Guernsey, Norwegian Red, Red Angus, Gelbvieh, Simmental, Gir, Nelore, Brahman, Beefmaster, Santa Geltridis, N'Dama, Sheko, Gaur, North American Bison, Lowland Anoa, Bantang, Yak, Cape Buffalo	BovineSNP50 BeadChip (Illumina)	743	4.6/-158
Fadista et al. (2010)	Holsteins, Red Danish, Simmental, Hereford	Bovine 2.1 M aCGH array (NimbleGen)	304	0.68/ (22)
Liu et al. (2010)	Angus, Bonsmara, Charolais, Gelbvieh, Hereford, Holstein, Limousin, N'Dama, Red Angus, Romosinuaro, Simmental, Brahman, Gir, Guzerat, Beefmaster, Brangus, Santa Gertrudis	Bovine 385k aCGH array	177	1.07/(28.1)
Bae et al. (2010)	*Bos taurus coreanae*	BovineSNP50 BeadChip (Illumina)	368	(63.1)
Matukumalli et al. (2009)	Hereford, Charolais, Holstein, Piemontese, Norwegian Red, Limousin, Romagnola, Angus, Red Angus, Guernsey, Jersey, Brown Swiss, Simmental, Gelbvieh, Beefmaster, Santa Gertrudis, Sheko, N'Dama, Brahman, Gir, Nelore	Infinium BovineSNP50 BeadChip (Illumina)	42	(49.1)

Table 3: List of CNVs studies on Bovine genome: the breeds, the platform and technologies applied, the total number of CNVRs identified, and the coverage on the genome.

The improvement of the SNP array permitted to detect CNVs by high-throughput genotyping on different set of breeds. More CNV loci were identified in African, composite, and indicine breeds and also in the taurine breeds (Matukumalli *et al*., 2009). Bae *et al*., (2010) and Fadista *et al*., (2010) created two CNV maps of bovine genome using SNP and CGH arrays (BovineSNP50 BeadChip by Illumina and Bovine 2.1 M aCGH array by NimbleGen, respectively). The size range of CNVRs in the Fadista *et al*., (2010) study was 1.7 Kb - 2 Mb, while in the Bae *et al*., (2010) study it was 50-200 Kb. The coverage of CNVRs on the bovine genome reported in these studies have a range from 0.68% (Fadista *et al*., 2010) to 4.6% by Hou *et al*., (2011) that corresponds to 22 Mb and 139.9 Mb, respectively.

These differences in the range size may have probably a methodological origin and can reflect the higher resolution of CGH array versus the SNP array platform.

1.1.4.2 Association with phenotypic variation and disease

Different authors found associations between CNV and genes in cattle genome. Liu *et al*., (2009, 2010), Matukumalli *et al*., (2009), Seroussi *et al*., (2010), Bae *et al*., (2010) and Hou *et al*., (2011) reported that these gene families include olfactory

receptors, ATP-binding cassette (ABC) transporters, Cytochrome P450, β-defensins, interleukins, the bovine MHC (BoLA) and multiple solute carrier family proteins.

Liu *et al*., (2010) found over 200 candidates CNVRs in total and 177 within known chromosomes. The CNVRs spanned about 400 annotated cattle genes with specific biological functions, such as immunity, lactation, and reproduction.

Bae *et al*., (2010) identified 368 CNVRs that contained 538 genes significantly enriched in multicellular organismal process, regulation of biological quality, and cell morphogenesis.

Fadista *et al*., (2010) detected 304 CNVRs in the genome of 20 bovine samples from 4 dairy and beef breeds. These CNVRs were found to be enriched for genes with functions related to environmental response, such as immune and sensory functions.

Hou *et al*., (2011) detected marked variation in copy number among individuals and across different cattle species (*Bos Taurus, Bos indicus*, and *Bubalus),* breeds (Table 3) and/or groups, detecting variations of TLR3 (toll-like receptor 3) and PPARA (peroxisome proliferator-activated receptor alpha) receptors. This CNV analysis have supported the hypothesis that the important differences of CNVRs frequencies among breeds, may be constant and have suggested one of the possible sources of variability than could have contributed to breed differentiation (Nei *et al*., 2005,

Matukumalli *et al*., 2009; Liu *et al*., 2010; Seroussi *et al*., 2010; Hou *et al*., 2011; Clop *et al*., 2012).

Bickhart *et al*., (2012) published the first study of sequenced-based CNV within cattle genomes analysing the genomic sequence of five *Bos Taurus Taurus* (Angus, Hereford, and Holstein) and one *Bos Taurus indicus* (Nelore) individuals. Total of 1,265 unique CNVRs were detected containing 413 genes. They reported breed-specific copy number differences in a Nelore individual as excellent candidate for pathogen and parasite resistance (CATHL4, ULBP7, and KRTAP9-2). In addition, copy number differences were detected for several lipid metabolism and transport genes in the taurine individuals.

1.2. METHODS OF CNVs DETECTION

The genome structural variation events can be discover by a modern technologies useful to genome-wide application as array comparative genomic hybridization, next generation sequencing, and SNP genotyping arrays. Alkan *et al*., (2011) described: i) the hybridization-based technologies that infer a copy number compared to a reference sample or population; ii) the next generation sequence that focuses on mapping sequence reads to the reference genome and subsequently identifies discordant

signature or patterns that are diagnostic of different classes of structural variation.

However, some CNVs were detected several years ago by standard methods of molecular genetic analysis at individual genetic loci, including cytogenetic methods and PCR-based approaches (Wain L.V *et al*., 2009).

1.2.1 CLASSIC CYTOGENETIC TECHNIQUES

Classic cytogenetic have identified structural variation as microscopically visible alteration using the fluorescent *in-situ* hybridization (FISH) (Wain L.V *et al*., 2009).

FISH is an in situ hybridization technique in which a labelled probe of specific DNA sequences is hybridized to a preparation of metaphase chromosome or interphase DNA, usually attached to a glass slide. The chromosomal DNA and probe mixture is then denatured, allowing the single-stranded probe and single stranded DNA to re-anneal, with the probe hybridizing to the complementary sequences on the DNA and reformed a double stranded molecule. The unbound probes are washed away after the hybridization step, while the hybridized probes are visualized directly if they are tagged with fluorochromes (Langer, 1981).

1.2.2 COMPARATIVE GENOME HYBRIDIZATION ARRAY (CGH-ARRAY)

The array CGH platform has opened new opportunities to assess copy number rearrangements associated with disease and genetic variation. A typical CGH array consists of mapped DNA sequences that can arise from different sources, which could be classified as genomic inserts (BAC or fosmid clones (Snijders *et al*., 2001), cDNA clones (Pollack *et al*., 1999), genomic Polymerase Chain Reaction (PCR) products (Dhami *et al*., 2005), or oligonucleotides (Urban *et al*., 2006). The first generation of BAC arrays was based on BAC, while an alternative of this type of DNA source was the oligonucleotide-arrays which are mostly supplied commercially. Array CGH platforms are based on the principle of comparative hybridization of two labelled samples (test and reference) to a set of hybridization targets (long oligonucleotides or bacterial artificial chromosome (BAC) clones), and the signal ratio is used for the detection of copy number. The test and reference samples are mixed together and applied to the chip and hybridisation takes place (the fragments of DNA hybridise with their matching probes on the array) (Figure 4). The chip is then scanned in a machine called microarray scanner which measures the amount of red (test sample) and green

(reference sample) fluorescence on each probe. The microarray scanner together with computer analytical software calculates the ratio of the red and green fluorescent dyes to determine whether, for the piece of DNA represented by each probe, the sample has the correct amount of DNA, too much DNA (duplication) which would be shown by too much red, or too little DNA (deletion), shown by too much green.

Currently, Roche NimbleGen and Agilent Technologies are the major suppliers of whole-genome array CGH platforms.

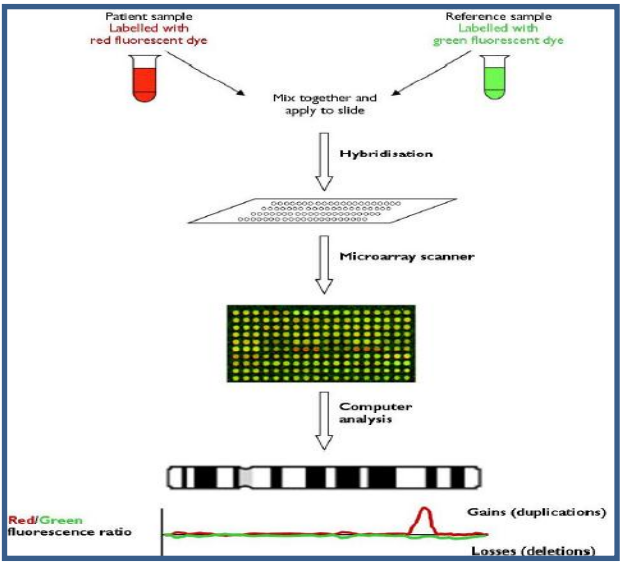


Figure 4: Graphical representation of the mechanism to detect CNV using CGH-array (http://www.rarechromo.org/information/other/array%20cgh%20ftnw.pdf).

1.2.3 SNP MICROARRAY TECHNOLOGY

Similarly to CGH platform, the SNP microarray platforms are based on hybridization; but for this array the hybridization is performed on a single sample per microarray and log-transformed ratios are obtained by clustering the intensities measured at each probe across multiple samples. Also, the probe are specific for SNPs that increase CNV sensitivity, distinguishing alleles and identifying regions of uniparental disomy through the calculation of the B allele frequency (BAF).

Recently SNP arrays incorporated finer SNP selection criteria for complex regions of the genome and non-polymorphic copy-number probes to improve the coverage of CNV regions.

1.2.3.1 Illumina Infinium II Whole Genome genotyping assay

The Infinium II Whole Genome genotyping assay was created for the interrogation of huge number of SNPs (from 10,000 to hundreds of thousands) at unlimited levels of multiple loci using a single bead type and dual color channel approach (http://www.illumina.com/). Genotypes are obtained comparing the two-color signal intensities produced by a BeadChip marker to the canonical genotype clusters. The values of canonical cluster (CC) are by 0 for the homozygous genotype (AA), 0.5 for the

heterozygous (AB), and 1 for the homozygous genotype (BB). The millions of calls are visualized in the Genome Studio software.

The standard protocol developed by Illumina is characterized from the following steps: denaturation and neutralization of 50 ng of genomic DNA (gDNA), amplification with PCR-free, enzymatic fragmentation of the amplified product, and precipitation using isopropanol. This product is resuspended in formamide-containing hybridization buffer. All beadchips are prepared for hybridization in a capillary flow-through chamber. The samples are applied to beadchips and the loaded beadchips are incubated overnight in the Illumina Hybridization Oven. Unhybridized and non-specifically hybridized DNA is washed away, and beadchips are prepared for staining and extension (Figure 5). The results in intensity values are obtained from the readout of the two colors channels (two alleles) for each SNP marker on the Infinium BeadChip.

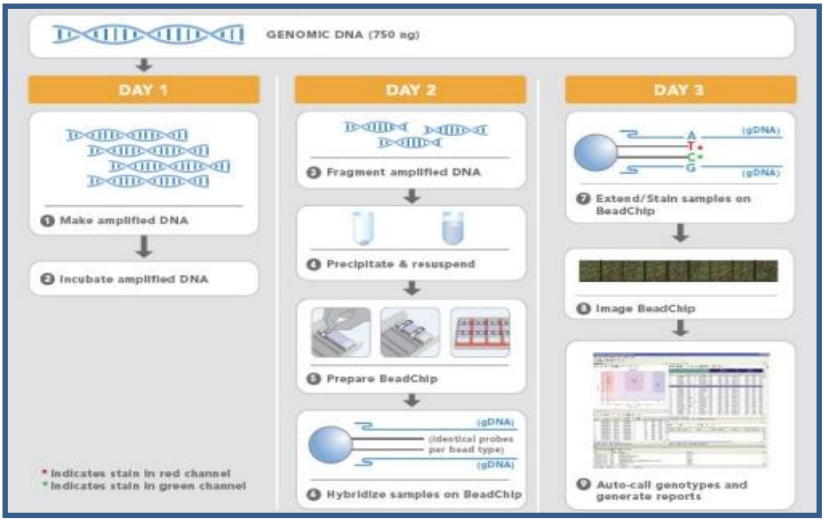


Figure 5: Illumina Infinium II Whole Genome genotyping assay.

The BeadChips are scanned on the Illumina BeadArray Reader using default settings, and intra-chip normalization is performed using the Illumina's Genome Studio software v.1.0.1 with a GenCall cut off of 0.1 and call rate cut off of 98%. Illumina BeadChips can detect several classes of variants: CNVs (amplification, duplication, deletion) and copy- neutral structural variants (copy neutral LOH).

1.2.3.2 Genome Studio

The GenomeStudio software is a powerful informatic tool, with a user-friendly graphical interface, that allows to visualize and to

analyse data generated by all Illumina's platforms. This software comprises the modules DNA/RNA sequencing, genotyping, gene expression, methylation, protein analysis and ChIP sequencing that enable to compare data from different applications to obtain an overall view of the genome.

Genotyping data generated using Illumina Infinium II Whole Genome genotyping assay are analysed in the Genome Studio Genotyping (GT) module. This module uses algorithms to perform primary data analysis, such as raw data normalization, clustering, and genotype calling. Many factors can influence the distribution of raw intensity values generated from each chip, and the normalization of raw values is necessary before making comparisons between subjects.

The specific algorithm applied to the raw data normalization includes five steps: outlier removal, translation correction, rotational correction, shear corrections and scaling correction (Peiffer *et al.*, 2006). With the normalization step the signal intensities from A and B alleles for a specific locus are transformed into X norm and Y norm. After normalization step, a polar coordinate plot of normalized intensity (R) is calculated with the following formula R= *X* norm + *Y* norm, while the allelic intensity ratio (theta) is calculated with the formula $\theta = (2/\pi) *$ arctan(*Y* norm/*X* norm) (Peiffer *et al.*, 2006).

R and θ values are used to create a canonical cluster for each SNPs marker per sample, and to determine SNP genotypes and copy number estimates.

In the GenomeStudio software the canonical clusters are three and they are visualized with three colours: red for the genotype AA, pink for the genotype AB and blue for the genotype BB. If the signal intensity per SNP is inside of one of these clusters, this SNP will be associated to the corresponding genotype of the cluster, but if the intensity value per SNP falls outside of a cluster, it is not possible to associate it to any genotype (Figure 6) (Peiffer *et al*., 2006).

To detect the copy number variants some software use two parameters from GenomeStudio, the log R ratio (LRR) and the B allele frequency (BAF). The LRR value represents the total signal intensity of the probe, and BAF value is the allelic balance.

The LRR of signal intensities is calculated as log2(*R*subject/*R*expected). The Rsubject is the observed total signal intensity for SNP for each individual, while the R expected is the interpolation of the midpoints of two neighbouring canonical clusters.

The BAF is calculated from the θ value of a sample, and it is an estimate of the relative frequency of allele B at a locus for an individual, ranging from 0 to 1. The BAF equal or close to "0"

corresponds to the AA genotype, "0.5" corresponds to the AB genotype, and "1" corresponds to the BB genotype (Figure 6). The allele frequency is calculated by linear interpolation with the lines D1 and D2 for an observed θ value of a sample that is localized between two clusters. The D1 is the difference of the distance of an observed θ value to the midpoint of the closest cluster solution, while the D2 is the difference of the distance between the θ values of the two canonical clusters. In other hand, the B allele frequency is calculated as BAF= [(D1/D2)* CC] (Peiffer *et al*., 2006).

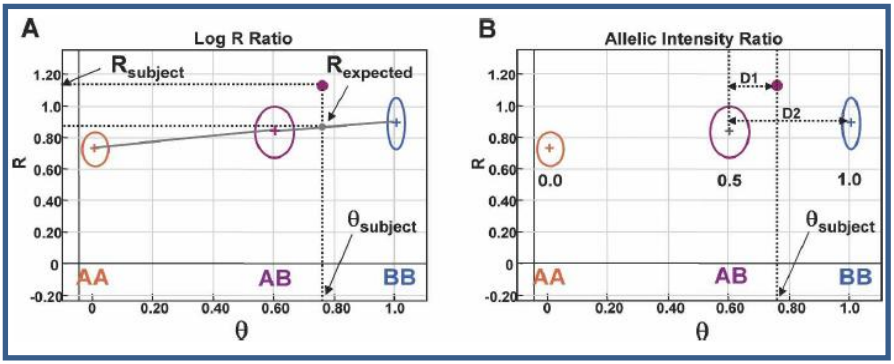


Figure 6: Analyzing SNP data (From Peiffer *et al*., 2006). (*A*) The log2 *R* ratio compares the observed normalized intensity (*R*subject) of the subject sample to the expected intensity (*R*expected; gray dot) based on the observed allelic ratio, θ subject, through a linear interpolation (gray lines) of the canonical clusters AA, AB, and BB (shown as circles) in the GenomeStudio. The normalized intensity value obtained from a single SNP is represented as a purple dot. The *R* and θ values for the subject are shown with thick black dotted lines. (*B*) The canonical clusters (shown as circles) are also used to convert θ values, that is, θ subject, to B allele frequency (allelic copy ratio). This is accomplished by a linear interpolation

of the known allele frequencies assigned to each cluster (0.0, 0.5, and 1.0). The allele frequency for an observed θ value falling between two clusters is also calculated by linear interpolation with lines D1 and D2.

1.2.4 ADVANTAGE AND DISADVANTAGE OF MICROARRAY PLATFORMS

The advantages offered from this type of technology are in terms of throughput and costs. Since the CNVs are very rare in the population, it is important to screen thousands of individuals to determine the significance of any structural variation. Considering the low cost and the large collection of SNP public data available for the genome-wide association studies, microarray data provide an opportunity to analysed the CNV landscape of large data sets (Peiffer *et al*., 2006). The disadvantages of this type of technology may be summarized as:

- low sensitivity in the detection of a single copy gains compared with the deletion;
- low sensitivity to detect events below 10 Kb;
- use of hybridization-based assays in repeat-rich and duplicated regions. For both array platforms it is assumed that each location is diploid in the reference genome, but this is not valid in duplicated sequence. This point is very crucial because CNVs have a strong correlation with the segmental

duplications and many breakpoints lie in duplicated regions. For these reasons, the accurate boundaries and copy number of these events will require additional technologies as next generation sequencing data (Alkan *et al*., 2011).

1.2.5 SEQUENCING

Next-generation sequencing (NGS) technologies replaced the microarrays technologies for discovery and genotyping. There are four types of strategy, all of them focusing on mapping sequence reads to the reference genome and identifying different classes of structural variation. The four strategies are: read-pair, read-depht, split-read, and sequence assembly. The read-pair, split-read and the assembly methods may be used to discover variants from all classes of structural variants, while the read-depth approach focuses on the detection of losses and gains and cannot discriminate between tandem and interspersed duplication (Alkan *et al*., 2011).

The read-pair method analyses the mapping information of paired-end reads and their discordancy from the expected span size and mapped strand properties. Sensitivity, specificity and breakpoint accuracy are dependent on the read length, insert size and physical coverage (Alkan *et al*., 20011).

The read-depth analysis examines the increase and decrease in sequence coverage to detect duplications and deletions, and to predict absolute copy numbers of genomic intervals (Alkan *et al*., 20011).

The split-read algorithms are capable of detecting exact breakpoints of all variant classes by analysing the sequence alignment of the reads and the reference genome; however, they usually require longer reads than the other methods and they have less power in repeat- and duplication-rich loci (Alkan *et al*., 20011).

Sequence assembly is the most powerful algorithm to detect SVs of all classes at the breakpoint resolution, even if the assembling of short sequences and inserts often result in contig/scaffold fragmentation in regions with high repeat and duplication content (Alkan *et al*., 20011).

1.2.5.1 Advantages and disadvantages of NGS technology

The main advantage of NGS technologies is that is possible to identify multiple of variant classes with a single sequencing experiment. Also, it is a power tool to understand the genetic variation. The disadvantage of NGS may be that each of the four methods has a different way to detect the structural variations; this

may be critical because validated variants remains unique to a particular approach.

1.2.6 VALIDATION AND DETECTION OF CNV USING PCR APPROACHES

PCR approaches were used to discover and verify CNV (Wain *et al*., 2009). These approaches are required for several reasons: i) as independent platforms to validate array-based CNV discoveries; ii) to detect map breakpoints of CNV regions; iii) to develop low cost, reliable assays for large-scale genotyping in large number of samples.

PCR approaches includes, the quantitative Fluoroscent Real-time PCR (qPCR) (Higuchi *et al*., 1992; Heid *et al*., 1996), the multiplex ligation-dependent probe amplification (MLPA) (Schouten *et al*., 2002), the multiplex amplicon quantification (MAQ) and the multiplex amplifiable probe hybridization (MAPH) (Armour *et al*., 2000; White *et al*., 2002).

Quantitative Fluoroscent Real-time PCR (qPCR) is still of particular use for molecular validation and studies of putative regions of variation (Wain *et al*., 2009).

The multiplex approaches facilitate parallel screening of a large number of samples across a large number of putative CNV loci with low costs and high reliability (Figure 7).

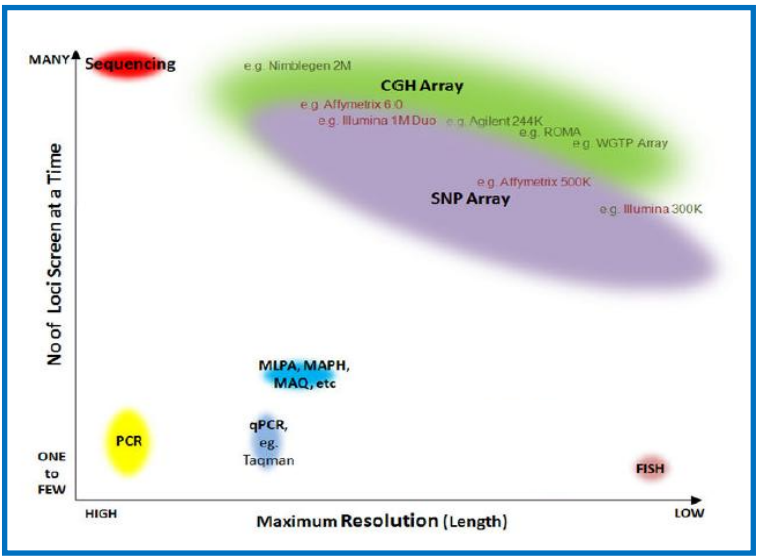


Figure 7: Sensitivity and throughput of several CNV detection techniques (From: ftp://ftp.sanger.ac.uk/). x-axis: maximum resolution of the technique (sensitivity); y-axis: number of loci that can be screened at a time. PCR: polymerase chain reaction; qPCR: quantitative PCR; MLPA: multiplex ligation-dependent probe amplification; MAPH: multiplex amplifiable probe hybridization; MAQ: Multiplex Amplicon Quantification; FISH: fluorescence in-situ hybridization; SNP Array: single nucleotide polymorphism genotyping array; CGH Array: Comparative Genome Hybridization array.

1.2.7 SOFTWARE ANALYSIS

The main function of CNV algorithms is to detect regions with a structural variation in which the mean of Log R Ratio (LRR) differs from the reference sample. Several CNV algorithms that use SNP data array are available.

In Table 4 some of the algorithms used to detect CNVs from array intensity data are reported: Circular binary Segmentation (CBS) approach is a nonparametric method, Hidden Markov Model uses the distribution of intensity data, and the Copy number analysis method uses the optimal segmenting algorithm.

The CBS model, developed by Olshen *et al*., (2004), is a modification of the binary segmentation approach (Sen and Srivastava, 1975), and the idea was to split each chromosome into regions of equal copy number that accounted for the noise in the array data. The CBS model was based on the fact that the CNVs are discrete gains or losses in contiguous regions of the chromosome that cover multiple array probes.

Algorithm	Software	Supplies
Circular Binary Segmentation (Olshen *et al*., 2004)	cnvPartition	Illumina Inc., San Diego, CA
	Nexus Copy Number	Biodiscovery Inc., El Segundo, CA
Hidden Markov Models (HMMs)	QuantiSNP	Wellcome Trust for Human Genetics, University of Oxford
	PennCNV	University of Pennsylvania
Golden Helix Optimal Segmentation (CNAM)		Golden Helix Inc.

Table 4: List of different algorithms and software available for the identification of CNVs.

1.2.7.1 PennCNV software

PennCNV is a free software for CNV detection from SNP genotyping arrays, and implements a hidden Markov model (HMM) that integrates several sources of information. It differs from segmentation-based algorithm in that it considered SNP allelic ratio distribution as well as other factors, in addition to signal intensity alone (http://www.openbioinformatics.org/penncnv/; Wang *et al*., 2007, 2008).

The HMM is a statistical technique that assumes that the distribution of an observed intensities data point depends on an unobserved (hidden) copy number state at each locus, where the elements of the hidden states follow a Markov process. CNV detection commonly uses aggregating information from multiple consecutive SNPs; for this reason, HMM provides a natural framework to model dependence structure between copy numbers at nearby markers. Transitions between copy number states are calculated by the probability of moving from one state to another state (Wang *et al*., 2007).

PennCNV software incorporates different information for each SNPs into the HMM in order to detect CNVs and to differentiate

copy number neutral (LOH) regions from normal state regions. These information are: the LRR, the B allele frequency (BAF), the population allele frequency and the distance between adjacent SNPs. Six hidden states are identified in the software: deletion of 2 copies, deletion of 1 copy, normal state, copy neutral with loss of heterozygosity (LOH), single copy duplication, and double copy duplication (Wang *et al*., 2007, Colella *et al*., 2007) (Table 5).

Hidden State	Total copy number	Description for autosome	CNV genotype
1	0	Deletion of two copies	Null
2	1	Deletion of one copy	A, B
3	2	Normal State	AA, AB, BB
4	2	Copy-neutral with LOH	AA, BB
5	3	Single copy duplication	AAA, AAB, ABB, BBB
6	4	Double copy duplication	AAAA, AAAB, AABB, ABBB, BBBB

Table 5: Hidden states, copy number, and their description by Colella *et al*., (2007) applied in PennCNV (Wang *et al*. 2007) and QuantiSNP (Colella *et al*., 2007).

Both the LRR and BAF values can be displayed and exported from GenomeStudio software given an appropriate clustering file with canonical cluster positions for each SNP. The distance between neighbouring SNPs determines the probability of having a copy number state change between them. Each SNP has two alleles referred to as the A and B alleles, thus the term "population

frequency of B allele" is used to differentiate it from the BAF term that measures allelic intensity ratio.

A flowchart outlining the procedure for CNV calling from genotyping data in PennCNV software is represented in Figure 8. The first step for LRR and BAF calculation can be alternatively performed by the GenomeStudio software, given a clustering file containing canonical genotype cluster positions. The HMM integrates several sources of information to give CNV calls. When genotype data are available for family members, the pedigree information can be incorporated to model CNV events more accurately.

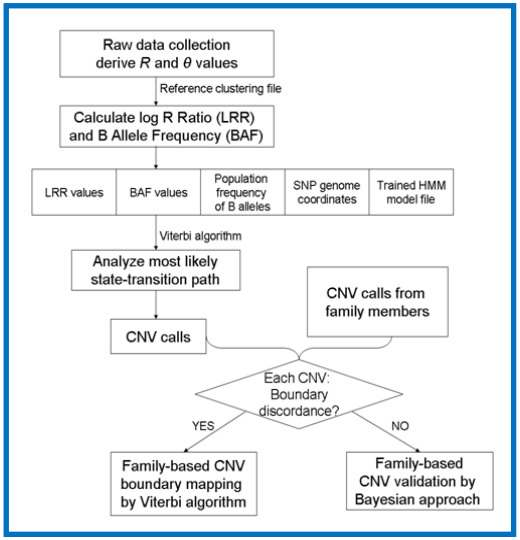


Figure 8: Flowchart that represents CNV calling from genotyping data by PennCNV software.

1.2.7.2 Copy Number Module (CNAM) of Golden Helix SNP and variation Suite 7.6.4

Another method to detect the CNVs is the powerful optimal segmenting algorithm called Copy Number Analysis Method (CNAM), implemented in SNP & Variation Suite c7.6.4 (SVS7) (Golden Helix, Bozeman, MT, www.goldenhelix.com) software. SVS7 improved two optimal segmentation algorithms that can detect inherited and de novo CNVs: the univariate analysis (that utilizes one sample at a time and is useful to detect rare and/or large CNV) and the multivariate analysis (that considers all samples and is useful for detecting small and common CNVs).
Respect to Hidden Markov Models, which assume that the means of different copy number states are consistent, the CNAM algorithm delineates CNV boundaries even at a single probe level, with controllable sensitivity and false discovery rate (www.goldenhelix.com). This algorithm can be applied to the data obtained by CGH and SNP arrays. A flowchart outlining the procedure for CNV calling from genotyping data in SVS7 software is represented in Figure 9.

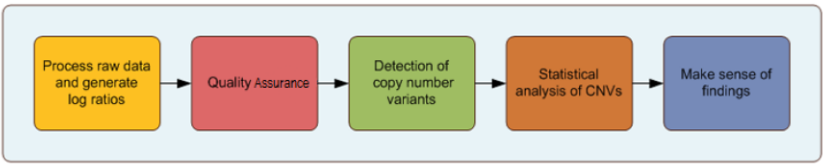


Figure 9: Flowchart that represents CNV calling from genotyping data by SVS7 software.

- Process raw data and generate log ratios. This step is necessary before the comparison among subjects. The measurement that is most commonly used to determine copy number status is the LRR, also called "Log2 ratio". The copy number variation corresponds to significant variation of Log2 ratios from segment to segment. CNAM module incorporates the Log2 ratio data and an appropriate genetic marker map that includes name, distance or position information for the genetic markers, and the chromosome information.
- Quality assurance. Significant sources of variation can affect analysis. This software offers several types of quality filters: the derivative log ratio spread (DLRS), the genomic wave detection and correction identified by Diskin *et al*., (2007) to control wave effects, and the principal component analysis (PCA) to correct for multiple batch effects. DLRS is a measurement of point to point consistency or noisiness in log R ratio data. This value is correlated with low quality SNP call rates and over/under abundance of identified copy number segments. Samples with higher values of DLRS tend to have poor signal-to-noise properties and accurate CNV

detection is often difficult for these samples.

Genomic waves are not platform-specific and Diskin *et al*., (2008) showed that they are an artefact caused by probe hybridization on an array. A signal intensity value is calculated for each probe, and these intensity measures are used to identified gain or losses of genomics segments. Several studies (Marinoni *et al*., 2007, Nannya *et al*., 2005) suggest a strong correlation between signal intensity of the probes and the local GC content. However, the GC content is correlated with many genomic features. Two quantitative measures (the wave factor and the GC wave factor) are used in SVS7 software to summarize the signal fluctuation. The first parameter summarises the total signal fluctuation of a genotyped sample across the genome, while the second parameter measures the fraction of signal fluctuation correlated with patterns of GC distribution.

PCA detects the presence of batch effects and other technical artefacts. The batch effects are represented from plate, machine, site variation, the sample extraction and preparation procedure, cell types, DNA source, array scanners, processing order and/or date and the temperature fluctuation. All these types of variation can lead to complications, which are represented from poorly defined segments and higher number

of false and non-replicable finding.

- Detect of CNV. The objective in this step is to determine regions in the genome where a given sample's mean LR value differs from the reference using univariate and/or multivariate analysis.
- Statistical analysis of CNVs. This can be done visually using advanced plotting techniques, or statistically using simple summary statistics or advanced association and regression analysis.
- Make sense of findings. This can be done looking at the genomic annotations, pathway analysis, and gene ontologies.

1.3. BIOLOGICAL IMPACT OF COPY NUMBER VARIATION

1.3.1 FUNCTIONAL CONSEQUENCES OF CNVs

Different phenotypic features can occur by genetic variants (for size and form) and can be explained by modification of gene expression (by action on transcription, splicing, or translation and stability) and/or by the alteration of protein structure. Usually, CNVs could encompass parts of genes, that reside entirely outside of genes or in the case of larger variants include several known genes (Wain *et al*., 2009).

Below are reported some examples of CNVs effects and their functional consequences (Lupski and Stankiewicz 2005; Feuk *et al*., 2006b; Eichler *et al*., 2007; Conrad & Antonarakis 2007; Zhang *et al*., 2009; Stankiewicz and Lupski 2010):

- Gene copy number changes (deletion and duplication) resulting in alteration of gene expression;
- Variation of gene expression (with long-range effect) can occurs by rearrangements that modify the regulatory elements for nearby gene (Kleinjan and van Heyningen 2005);
- CNVs may disrupt a gene (causing functional loss or modification);
- The generation of new fusion protein can occur by rearrangement (by fusing different protein domains);
- CNVs may affect and change microRNA level in the genome, which in turn may lead to alterations in gene expression. (Henrichsen *et al*., 2009). This alteration may propagate to other genes located in downstream pathways or regulatory networks (Henrichsen *et al*., 2009).

From an evolution's point of view, a duplicated gene could be modified for new functions, and facilitating a diversification and evolution of species (Lynch and Conery 2000; Dermitzakis and Clark 2001; Korbel *et al*., 2008). Some of these variations may have been favoured in positive selection processes, but also in

negative selection and can be predicted in deletion events where deleterious loss-of-function alleles could be generated.

1.3.1.1 CNVs effects on gene expression

The effect of CNVs on gene dosage is prominent in genomic disorders. These disorders are defined as diseases caused by genomic rearrangements affecting dosage sensitive genes (Stankiewicz and Lupski 2002; Lupski 2007). The gene dosage is the number of functioning gene copies and determines the amount of gene product (Wain *et al*., 2009). Because genomic disorders are caused by *de-novo* deletions, insertions, or other chromosomal rearrangements, these structural variants contribute to non-heritable components of disease risk, although evidence suggests that common CNVs are inherited and therefore caused by ancestral structural mutations (McCarroll *et al*., 2008).

Below are reported some examples of CNVs effects and the relationship with the dosage-sensitive and non-sensitive genes.

1) Dosage-sensitive gene:

- Positive relationship between the effects of CNVs and the gene expression: it suggests that higher CNVs increase the expression gene level;

- Negative relationship between the effects of CNVs and the gene expression: higher CNVs decrease the expression gene level;
- Complex relationship between the CNVs effects on gene expression: a small change in gene copy number could completely switch a biological network from one alternative steady state to another;
- Neutral relationship between the CNV that may not necessarily translate into gene expression. This situation may be explained by the genomic location of CNV, if the CNV's region was poor of genes, the insensitivity of the CNV genes to dosage, or the effect of dosage compensation are a mechanism which would balance out gene expression changes resulting from CNVs.

2) Non-dosage sensitive gene:

- Gene Disruption: the CNV might modify phenotypes and cause diseases by disrupting genes and modifying protein function.

1.3.1.2 CNVs and diseases

Several studies show that the CNVs play an important role in sporadic diseases, resulting from *de-novo* CNVs formation (Lupski 2007), and rare inherited copy number changes,

Mendelian disorders and on multifactorial and complex disease phenotypes.

The definition of these diseases are reported below:

- Genomic disorders are defined as diseases caused by genomic rearrangements affecting dosage sensitive genes (Stankiewicz and Lupski 2002; Lupski 2007). These types of disorders can be the results obtained by reciprocal deletions/duplications of the same loci, showing that the variation of the dosage at the same gene or genes could results in different phenotype (Lupski and Stankiewicz 2005);
- Rare inherited CNVs that act in the same way as Mendelian mutation (Estivill and Armengol 2007);
- Common and multi-allelic CNVs in multi-factorial or complex diseases: the frequently multi-allelic CNVs (several duplication of genomic segment could produce a wide range of diploid of copy number), play an important role in the complex diseases susceptibility, in conjunction with the context of population and ethnicity matters (Eichler *et al*., 2007).

1.3.2 CONSEQUENCES IN HUMAN

An example of genomic disorder with a positive relationship between CNV and the increase of the gene expression was

represented from the human salivary amylase gene (AMY1). In this case, the diploid copy number varies ranging from 2 to 15 depending of the quantity of starch in the diet. In other hand, the population with higher starch diet usually have higher AMY1 copy number and vice versa (Perry *et al*., 2007). Perry *et al*. (2007) also showed that this CNV locus affects gene expression at both the transcriptional and translational levels, and that copy numbers are correlated with mRNA and AMY1 protein level in human saliva samples. Other types of genomic disorder in human genome are Williams Beuren Syndrome (WBS) [del(7)(q11.23q11.23)], Velocardiofacial/ DiGeorge Syndrome (VCFS/DGS)[del(22)(q11.2q11.2)], Prader–Willi (PWS)[pat del(15)(q11.2q13)] or Angelman syndrome (AS) [mat del(15)(q11.2q13)], and Charcot Marie Tooth Disease (CMT) with reciprocal Hereditary Neuropathy with Liability to Pressure Palsies (HNPP) (Lupski 2006). Moreover, high number of genomic disorders are involved on nervous system disorders and neuropathies (example PW/AS with autism (Veltman *et al*., 2005). A negative relationship, in which higher copy number decreases expression levels, was observed by Lee *et al*., (2006) with the proteolipid protein gene (PLP1). In contrast with the other example of AMY1, in this case a small duplication downstream of

the PLP1 cause gene silencing by positional effect, thus lowering gene expression.

One example of gene disruption was represented by complete deletion of the RHD gene. A subject's (diploid) genome could contain two, one or zero copies of RHD, with the zero copies corresponding to rhesus-negative and absence of D antigen expression (Wain *et al*., 2009; Avent *et al*., 1997).

Two examples of disorders due to inherited CNVs were represented by the Parkison's diseases (Singleton *et al*., 2003; Chartier-Harlin *et al*., 2004) and the Alzheimer disease (ADEOAD) (Rovelet-Lecrux *et al*., 2006). Singleton *et al*., (2003) were the first that have demonstrated the triplication or the duplication of the alpha-synuclein (SNCA) locus that caused the hereditary Parkison's disease with dementia. Rovelet-Lecrux *et al*., (2006) confirmed that the duplication of the amyloid precursor protein (APP) on chromosome 21 caused the Alzheimer disease (ADEOAD) with amyloid angiopathy (CAA). The alteration in the APP gene probably lead to the accumulation of amyloid precursor protein which results in neurodegeneration and the APP gene is as a dosage sensitive gene.

In autism, large-scale of structural abnormalities have been recognised, including inherited duplication (15)(q11;q13). Strong evidence that rare CNVs are causally related to familial and

sporadic autism has come out from studies using a range of molecular approaches. Most of CNVs are rare and unique to an individual or family, and are difficult to replicate in independent studies. However, autism has been associated with uncommon de-novo and inherited deletions and duplications of genomic segments on chromosome 16 (Weiss *et al*., 2008). Also, in schizophrenia have been recognised several rare CNVs including del(1)(q21.1), del(15)(q11.2), del(15)(q13.3), and del(22)(q11.2) (Stefansson *et al*., 2008; The International Schizophrenia Consortium, 2008).

1.3.3 CONSEQUENCES IN LIVESTOCK

Fadista *et al*., (2010) and Hou *et al*., (2011) showed genes related to several functions in cattle: immunity and defence, for example macrophage, natural-killer- and T-cell-mediated immunity, major histocompatibility complex (MHC); sensory perception are related with the olfactory receptors, chemosensory perception; response to stimuli and neuro- logical system processes. Hou *et al*., (2011) also reported the genes that participated in gene transcription, cell cycling and nucleic acid binding and metabolism.

In the small ruminants, goats and sheep, Fontanesi *et al*., (2010, 2011) showed an over-representation of genes related to the MHC. Also other genes have shown an over-representation and

are related to metabolic pathway, for example the lipid binding, transport and localization; these results may have implication to identify loci that affect milk and carcass fat content and composition (Bickhart *et al.*, 2012).

1.3.3.1 CNVs and phenotypes

1.3.3.1.1 Coat Pigmentation in horse, pigs, and sheep

The mutation of coat pigmentation on horse species has been suggested by the duplication of the equine syntaxin 17 (STX17) gene, and a progressive hair depigmentation syndrome accompanied by an increased susceptibility to melanoma (Rosengren-Pielberg *et al*., 2008). It has been hypothesized that this mutation can affect a regulatory effect by upregulating STX17 and/or NR4A3 mRNA levels.

In swine, dominant white colour has been associated with two mutations, a duplication encompassing the whole gene (Giuffra *et al*., 2002), and a mutation causing the skipping of exon 17 (Giuffra *et al*., 1999), at the KIT gene. It is important to specify that changes in gene copy number do not always translate into differences in gene expression, in fact, in this case several factors might keep mRNA levels stable, as a Dosage compensation, differences in the chromatin environment (Henrichsen *et al*., 2009). The KIT gene has a role in the proliferation, survival and

migration of melanocytes (Wehrle- Haller 2003), and it also affects oocyte and follicle development (Hutt *et al*., 2006). For this reason, it would be interesting to test the impact of selection for white colour on reproductive traits, such as sow fertility and prolificacy.

In the case of sheep, dominant white coat is associated with a tandem genomic duplication encompassing three genes, the agouti signalling protein (ASIP) gene (which regulates melanin biosynthesis), the itchy E3 ubiquitin protein ligase homolog (ITCH) and the adenosylhomocysteinase (AHCY) loci (Norris & Whan 2008). This duplication causes an abnormal expression of the second copy of the ASIP gene because its transcription controlled by an ITCH promoter (Norris & Whan 2008). Similar situation in goats are reported by Fontanesi *et al*., (2009).

These mutations are expected to be relatively recent, subsequent to domestication, because they segregate in specific breeds and they are absent from genomes of wild ancestors.

1.3.3.1.2 Producion Traits

Recent studies have reported the relationships between CNVs and production traits of economic values.

Seroussi *et al*., (2010) identified the associations between CNVR (on BTA18), and index of genetic evaluations for protein and fat

production in Holstein cattle. These results are important because they constitute the demonstration that complex traits of livestock are modulated in part by CNVs and that genomic selection schemes might benefit from the incorporation of CNV data. However, there are still technical limitations that need to be overcome to implement such approaches. One possible way to solve many of these technical drawbacks, is represented by a systematic next-generation sequencing of bull genomes.

1.3.3.1.3 Diseases and development of abnormalities

In domestic animals, CNVs have been mostly related to the occurrence of Mendelian diseases, but their impact on complex traits is now starting to be understood. Some examples of diseases and development abnormalities are reported below.

Liu *et al*., (2011) reported the existence of associations between CNV and susceptibility to intestinal nematodes in cattle. A prevalent cause of disease is the genomic deletions interrupting genes and affecting their biological function.

In cattle three diseases such as, anhidrotic ectodermal displasia, osteopetrosis and renal tubular dysplasia are explained by rearrangements.

Bovine anhidrotic ectodermal dysplasia is characterized by hypotrichosis and dental defects, and the causal mutation consists

of a deletion (ranging from 2 kb to lager size) encompassing exon 3 of the ectodysplasin A (EDA) locus (Drogemuller *et al*., 2001). This rearrangement implies that the function of this gene becomes completely suppressed.

Bovine osteopetrosis is a skeletal disease involving the growth of extremely dense and fragile bones because of a deficiency in osteoclast activity. One study showed that in Angus red cattle, this disease is produced by a 2.8-kb deletion of part of intron 1, exon 2, intron 2 and half of exon 3 of the solute carrier family 4, anion exchanger, member 2 (erythrocyte membrane protein band 3-like 1) (SLC4A2) gene (Meyers *et al*., 2010). This gene plays a key role in osteoclast function by exchanging bicarbonate for chloride ions, a necessary step in the acidification of the resorption lacuna and in bone demineralization (Meyers *et al*., 2010).

Also the renal tubular dysplasia is a bovine disease caused by a genomic deletion. This disease is characterized by renal failure, owing to the malfunction of the epithelial cells of the renal tubules and growth retardation (Hirano *et al*., 2000; Ohba *et al*., 2000). Moreover, the occurrence of abortions and stillbirths in cattle has been recently linked to a 110-kb genomic deletion encompassing the 3′ end of the MER1 repeat containing imprinted transcript 1 (non-protein coding) (MIMT1) gene that shows a semi lethal pattern of inheritance (Flisikowski *et al*., 2010).

CHAPTER 2 - THE AIM OF THIS STUDY

The objective of this study was to produce a medium resolution map of CNVs in the Italian Brown Swiss breed using the genotyping array Illumina BovineSNP50 BeadChip.

In this study, two different algorithms were used. One from PennCNV software and one as in the module CNAM from SVS7 software by Golden Helix. CNV regions were defined with both analyses. Subsequently a consensus map was then created between the two maps generated according to the two different algorithms. Two different criteria were used for the consensus.

The detected common CNVRs, were annotated with genes and other Ensembl available elements (protein coding, pseudogene, retrotrasposed, miRNA, miscRNA, rRNA, snRNA, snoRNA).

This study was part of the European Quantomics project (*contract n. 222664-2*) with the collaboration of The Italian Brown Breeders Association (ANARB) that provided part of the used data.

CHAPTER 3 - MATERIAL AND METHODS

3.1 SAMPLES PREPARATION

A total of 1,342 bulls of the Italian Brown Swiss breed were sampled for this study. Genomic DNA was extracted from semen using the ZR Genomic DNA TM Tissue MiniPrep (Zymo), and from blood utilizing Macherey-Nagel NucleoSpin® Blood kit. Sample DNA was quantified using NanoQuant Infinite®m200 (Tecan, Switzerland) and diluted to 50ng/ul. The Quality Control (QC) was performed on each sample to verify the DNA integrity on Invitrogen E-Gel 1% Agarose Gel, needed to apply the Illumina Infinium II protocol. A total of 775 bulls were genotyped by Geneseek, USA (http://www.neogen.com/) while 578 bulls were processed at Kos Genetics in Milan, Italy (http://www.kosgenetic.com/).

DNA samples were genotyped using Illumina Bovine SNP50 BeadChip interrogating 54,001 polymorphic SNPs with an average probe spacing of 51.5 kb and a median spacing of 37.3 kb (Illumina Inc., USA). Nevertheless only 46,728 SNPs anchored on UMD3.1 assembly on autosomal chromosomes, were used to detect CNVs.

A call rate > 90% and a call frequency > 98% were used as a lower limit to classify high-quality SNPs.
UMD3.1 assembly was the reference genome used in this study.

3.2 SOFTWARE AND BIOINFORMATICS TOOLS USED

- BEDTools software was developed for a fast, flexible and easy manipulation of large and complex genomic datasets by Quinlan and Hall (2010). It is a genome arithmetic tool with an intuitive Python interface, and exploring complex genomic datasets in many common (BED, VCF, GFF, BEDGRAPH, SAM/BAM) formats.
- UNIX bash system
- R software for statistical analysis and graphics. Using R 2.15.0 version (http://www.cran.org)
- PennCNV (http://www.openbioinformatics.org/penncnv/) a free software tool for detection of CNV from Illumina and Affimetrix SNP array
- CNAM the copy number module implemented in SNP & Variation Suite v.7.6.4 (Golden Helix, Bozeman, MT, www.goldenhelix.com) were used to detect CNVs.

- The Database for Annotation, Visualization and Integrated Discovery (DAVID, http://david.abcc.ncifcrf.gov/) v6.7 was used to disclose biological interpretation behind large list of genes.
- Hotspot Detector for Copy Number Variants (HDCNV, http://daleylab.org) was used to obtain a graphical representation of the CNV calls from multiple samples.

3.3 DATA QUALITY ASSURANCE

Several filters were applied to data with PennCNV and CNAM.

3.3.1 CNAM: DERIVATIVE LOG RATIO SPREAD

The filter applied by Golden HelixSVS7 software was the derivative Log ratio spread (DLRS) parameter (an indicator of noise for samples). Pinto et al., (2011) describe the DLRS like the "absolute value of the log2 ratio variance from each probe to the next, averaged over the entire genome". They also describe the used interquartile range as the "measure of the dispersion of intensities in the centre of the distribution" less sensitive to outliers.

3.3.2 CNAM: WAVE FACTOR FILTER

The identification and correction for GC content was automatically generated by CNAM module of SVS7.

3.3.3 OUTLIERS REMOVE FROM PENNCNV

691 bulls were discarded from the PennCNV dataset. These 691 bulls were indicated as outliers from the DLRS and the WF filter criteria from SVS7 software. The DLRS filter was used instead of the standard deviation of Log R ratio filter, because for biological reason is here considered more suitable in identifying the signal intensity variation in CNV region. . In PennCNV software was used only the bulls considering by SVS7 software for a total of 651 animals.

3.3.4 PENNCNV SOFTWARE: WAVE FACTOR FILTER

The fasta sequence for UMD3.1 assembly (ftp://ftp.ncbi.nih.gov/genomes/Bos_taurus/Assembled_chromosomes/seq) was downloaded for each of the 29 autosomes chromosomes.

The wave factor filter was applied using the method as described by Diskin et al., (2008):

Using the –fastaFromBed command of BEDTools software (Quinlan and Hall, 2010) 1Mb window sequence (500 kb each side) for each SNP was extracted.

Using the nucBed command of BEDTools software to calculate GC content, equivalent to the sum of all occurrences of G and C, and dived by 1Mb window sequence for each SNP without 'NNs' (gaps in the sequence of the assembly).

3.3.5 PRINCIPAL COMPONENT ANALYSIS (PCA)

The principal component analysis (PCA) was used in order to detect the possible presence of batch effects and in this case to correct the signal intensity values.

The PCA was also applied to LRR values in order to detect clustering patterns related to handling processes of samples during genotyping. For this purpose the variables included in the PCA analysis were: lab project ID, dilution plate ID, dilution plate well ID, amplification plate ID, date, and amplification plate code, sample well ID, sentrix barcode, sentrix position, call rate, DNA extraction date, DNA 260/280 ratio, technitian , person genotyped, and scanner.

Two different approaches for the identification of the number of principal components necessary to correct the signal intensities values were. The first method is a graphical way based on a plot with principal component versus LRR eigenvalue and the second based on LRR Eigenvalues > 1.

3.3.6 FILTERING FOR THE CENTROMERE AND TELOMERE REGIONS

The telomere and centromere regions information were downloaded from www.bovinegenome.org for the autosomal chromosomes.

An overlapping of at least 10% among CNVs and telomere/centromere regions was chosen to filter out the CNVs, using the intersectBed command of BedTools software (Quinlan and Hall, 2010).

The same filtering criteria analysis for centromere and telomere regions here described for PennCNV software was also used for CNAM.

3.4 DETECTION OF CNVS

3.4.1 PENNCNV SOFTWARE

The CNV call was performed on autosomal chromosomes using –detect_cnv.pl feature which include also the correction for the wave factor, the expected signal intensity values and the expected transition probability for different copy number state.

3.4.2 SVS7 SOFTWARE

The identification of CNVs was performed using univariate analysis. The criteria considered for the analysis were: univariate outlier removal; maximum segments 10 per 10,000 markers; minimum number of markers per segment equal to 1; maximum pairwise permuted p-value was 0.005 for 2000 permutation per pair.

3.5 CNVRS DEFINTION

For both algorithms CNVRs were defined and created using the –mergeBed command by BedTools (Quinlan and Hall,2010) software, with criteria as suggested by Redon et al. (2006).

3.6 CONSENSUS BETWEEN PENNCNV AND CNAM

A consensus analysis was performed comparing the CNVRs obtained from both the algorithms. Two different approaches were applied (Redon et al, 2006, Wain et al., 2009), using the –mergeBed (Redon) and the interesectBed (Wain) commands of BedTools software (Quinlan and Hall, 2010).

3.7 ANNOTATION OF CNVRS

The gene list was downloaded, for the autosomal chromosomes from Ensembl v69 (http://www.ensembl.org/biomart/martview/76d1cab099658c68bde77f7daf55117e).

A catalogue of genes in the CNVRs was obtained using the intersectBed command by BedTools software (Quinlan and Hall, 2010).

The DAVID software (http://david.abcc.ncifcrf.gov/) was used to test the hypothesis that the molecular function, biological process, cellular component and pathway terms were under or over represented in CNVRs after Bonferroni correction.

3.8 COMPARISON WITH THE LITERATURE

CNVRs datasets available in literature were downloaded from Hou et al. (2011), Fadista et al. (2010), Liu et al. (2010), Bae et al. (2010), Bickhart et al. (2012). Using BedTools software, a list of the common regions among the results of the present study and those from literature was created.

CHAPTER 4 - RESULTS

The distribution of DLRS values on whole genome for each sample was used to determine which individuals to exclude. The inter-quartile range (IQR) of the distribution was used to identify the outlier and filter for them.

The Table 6 shows the distribution of this metric for each autosome chromosome.

Chr	Minimum	Q1	Median	Mean	Q3	Maximum	Threshold	IQR
1	0,192	0,254	0,283	0,294	0,318	0,838	0,414	0,064
2	0,173	0,226	0,255	0,269	0,294	0,897	0,396	0,068
3	0,170	0,230	0,259	0,271	0,298	0,808	0,399	0,068
4	0,161	0,214	0,247	0,261	0,291	0,898	0,408	0,078
5	0,167	0,232	0,263	0,272	0,298	0,899	0,396	0,065
6	0,174	0,236	0,268	0,281	0,311	0,878	0,423	0,075
7	0,185	0,254	0,287	0,297	0,326	0,841	0,434	0,072
8	0,168	0,234	0,265	0,275	0,304	0,866	0,409	0,070
9	0,190	0,266	0,301	0,311	0,347	0,898	0,468	0,081
10	0,166	0,245	0,279	0,287	0,316	0,827	0,423	0,071
11	0,163	0,233	0,266	0,275	0,304	0,900	0,409	0,070
12	0,181	0,255	0,296	0,304	0,338	0,915	0,462	0,083
13	0,181	0,247	0,272	0,282	0,304	0,848	0,391	0,058
14	0,170	0,224	0,254	0,266	0,292	0,825	0,395	0,069
15	0,181	0,260	0,296	0,304	0,333	0,879	0,442	0,073
16	0,171	0,239	0,270	0,282	0,312	0,814	0,423	0,074
17	0,189	0,259	0,289	0,298	0,325	0,843	0,425	0,066
18	0,166	0,219	0,250	0,261	0,291	0,948	0,399	0,072
19	0,154	0,225	0,255	0,266	0,298	0,794	0,408	0,073
20	0,171	0,228	0,262	0,273	0,301	0,838	0,409	0,072
21	0,166	0,255	0,287	0,295	0,323	0,804	0,424	0,068
22	0,157	0,212	0,244	0,257	0,288	0,923	0,403	0,077
23	0,180	0,270	0,321	0,327	0,372	0,894	0,526	0,102
24	0,166	0,230	0,262	0,274	0,303	0,772	0,413	0,073
25	0,153	0,204	0,230	0,240	0,261	0,902	0,347	0,057
26	0,198	0,259	0,289	0,300	0,328	0,751	0,430	0,068
27	0,155	0,219	0,258	0,271	0,308	1,094	0,442	0,089
28	0,182	0,256	0,292	0,301	0,337	0,850	0,458	0,081
29	0,163	0,229	0,261	0,271	0,301	0,871	0,409	0,072
All	0,206	0,245	0,272	0,283	0,304	0,864	0,393	0,059
Median	0,196	0,238	0,266	0,277	0,300	0,866	0,393	0,062

Table 6: Derivative log R ratio spread values obtained from 1342 bulls. Chr – chromosomes, Minimum, Q1 – first quartile, Median, Mean, Q3 – third quartile, Maximum, Threshold – upper outlier threshold, IQR – inter quartile range, All considering genome wide analysis, Median considering by chromosomes values analysis.

The upper Outliers threshold values are reported in the "threshold" column (Table 6). The threshold values are calculated as Q3 + IQR *1.5. The threshold cutoff value for excluding the outlier from the other analysis is the average value among all chromosomes reported: 0.393. Thus every individual with a median DLRS value above 0.393 have been discarded. According to this edit a total number of 54 animals were excluded from subsequent analysis. The median of the derivative Log R ratio per bulls is represented in Figure 10.

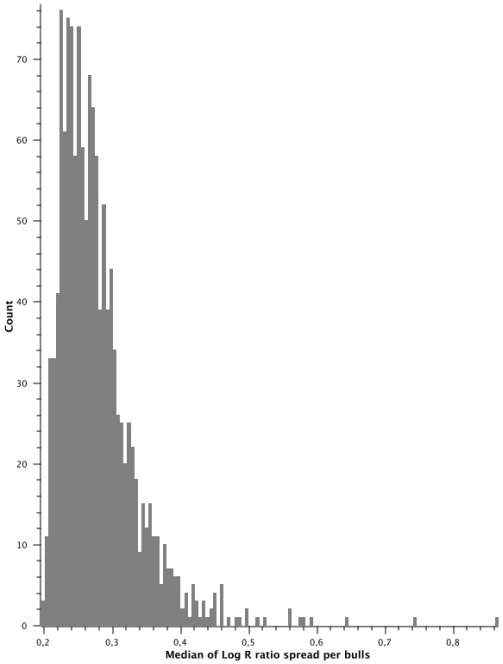


Figure 10: Histogram of Median of Log R ratio spread per bulls.

The Figure 11 is a graphical example of the difference between two individuals with a high (black) or low (grey) quality of the signal intensity noise of Log R ratio in whole genome space.

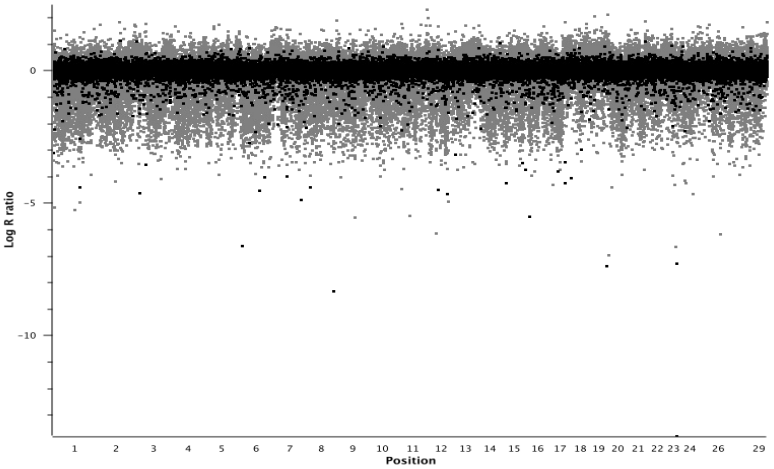


Figure 11: Plot of numeric values from Log R ratio values for one sample with a good signal to noise ratio (black) and with other sample with a bad signal to noise ratio (in grey).

4.2.2 CNAM: WAVE DETECTION AND CORRECTION

In the Table 7 some parameters relative to the wave detection and correction are reported. The threshold value for the Abs Wave Factor (median absolute deviation of signal intensities calculate for every 1Mb non-overlapping window in the genome) is 0.1124;

the threshold for the GC correlation (correlation between median signal intensity and local GC content in all 1Mb non-overlapping windows) results 0.7262; the threshold for the Wave Factor (summary of a total intensity signal fluctuation of a genotyped sample across the genome) is 0.1977. Finally the threshold for the GC Wave Factor (summary measure of the intensity signal fluctuation explained by local GC content) is 0.0401.

Column	Minimum	Q1	Median	Mean	Q3	Maximum	Threshold	IQR	Variance
AbsWave Factor	0.0234	0.0374	0.0497	0.0561	0.0674	0.2519	0.1124	0.0300	0.0007
GC Correlation	-0.5605	-0.0575	0.1343	0.0831	0.2559	0.6499	0.7262	0.3135	0.0547
WF	-0.1459	-0.0335	0.0391	0.0228	0.0589	0.2519	0.1977	0.0925	0.0033
GCWF	-0.0810	-0.0020	0.0051	0.0059	0.0148	0.1599	0.0401	0.0168	0.0004

Table 7: Summary statistic of waviness. Minimum, Q1 – first quartile, Median, Mean, Q3 – third quartile, Maximum, Threshold IQR – inter quartile range, Variance. The Abs Wave Factor - median absolute deviation of signal intensities calculate for every 1Mb, GC Correlation - correlation between median signal intensity and local GC content in all 1Mb, WF - summary of a total intensity signal fluctuation of a genotyped sample across the genome, GCWF - summary measure of the intensity signal fluctuation explained by local GC content.

The Figure 12 shows the distribution of the median absolute deviation of signal intensities across all samples.

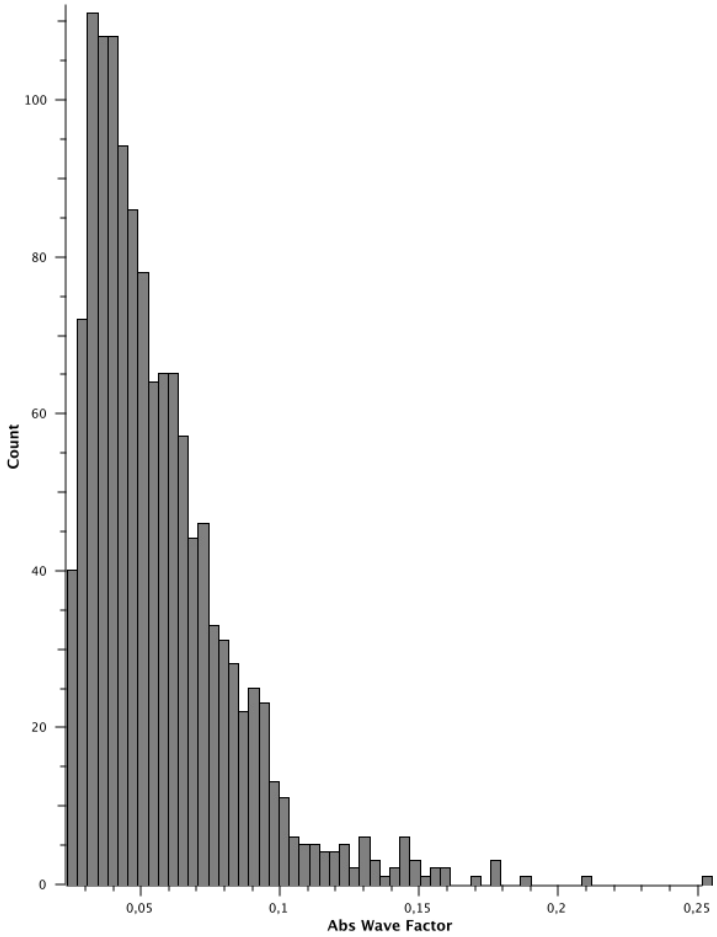


Figure 12: Distribution of Abs Wave Factor on sample population.

The Figure 13 shows the relationship between the Wave factor (WF) and the GC content wave factor (GCWF) explaining the

variability of wave factor caused by the local GC content. The GC-wave factor was determined by product of WF value and the absolute values of GC correlation.

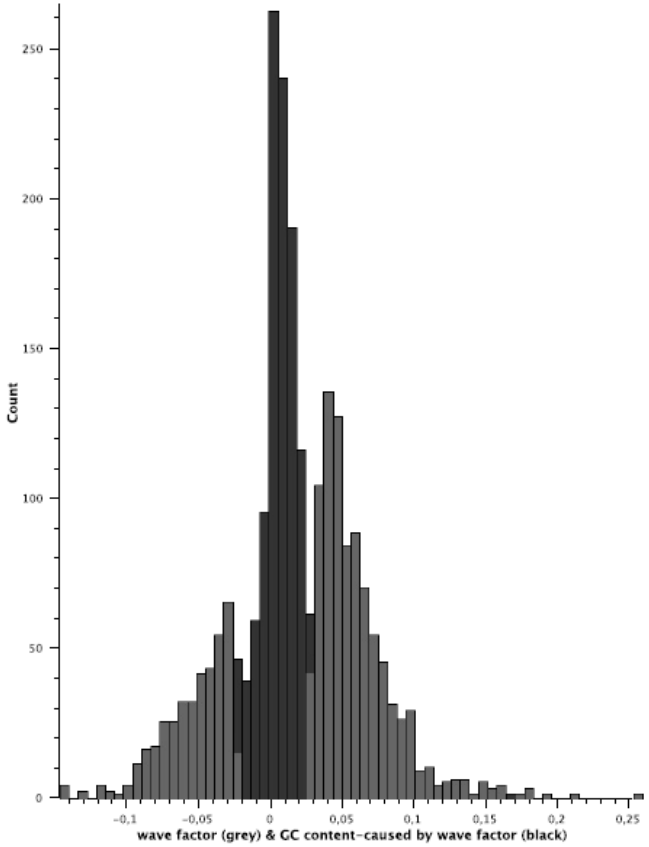


Figure 13: Histogram with wave factor (grey) and GC content-caused wave factor (black) before correction of Log R ratio measurements.

The Plot in Figure 14 represents an example for a low quality sample, (with high wave factor), while the plot in Figure 15 refers to a high quality sample (with low wave factor). In the x- axes the position of the probe for whole genome space, and in y-axes the Log R ratio values before and after the wave factor correction..

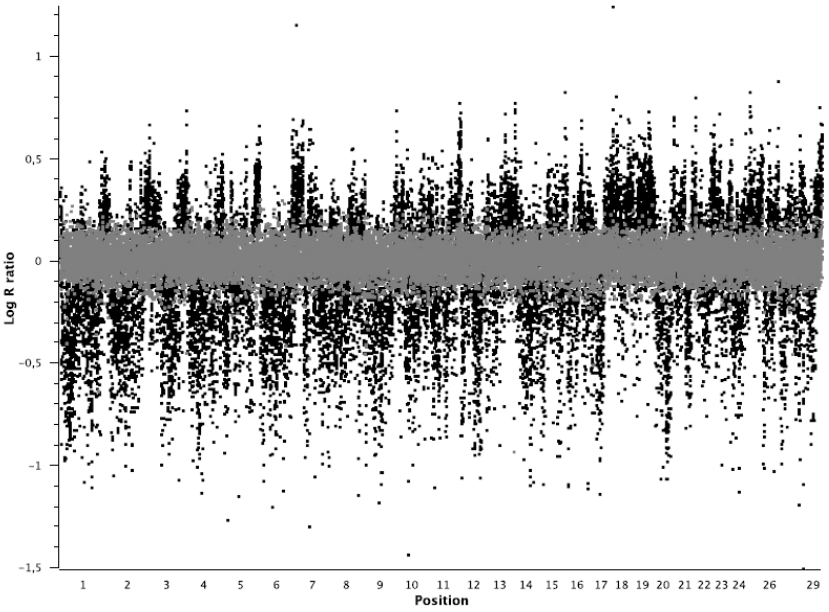


Figure 14: Plot of the total signal intensities (LRR) without wave factor correction versus whole genomic space for a low quality sample, with high wave factor (black) and with a high quality sample, with low wave factor (grey).

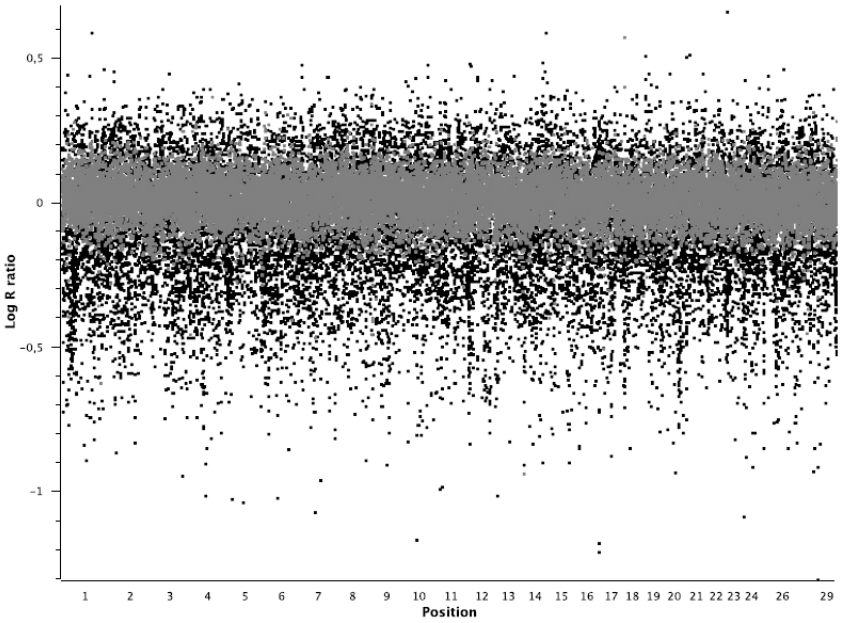


Figure 15: Plot of the total signal intensities (LRR) after the wave factor correction versus whole genomic space for a low quality sample, with high wave factor (black) and with a high quality sample, with low wave factor (grey).

A total number of 691 outlier bulls were identified after application of the DLRS and the wave factor corrections. These bulls were deleted in the subsequent analysis performed with PennCNV and SVS7.

4.2.3 CNAM: PRINCIPAL COMPONENT ANALYSIS

The principal component analysis was used to detect the presence of batch effects and/or technical variables. The Figures 16 and 17

are the plots of the second versus the third principal components to discover the batch effects that can influence the analysis. The Figure 16 is the representation of clustering according to the technician operating the genotyping of the samples. It shows a clear separation in three clusters. Figure 17 is the representation of clustering according to the scanner (Geneseek vs Kos) used in sample processing. These plots show a clear separation in two different cluster.

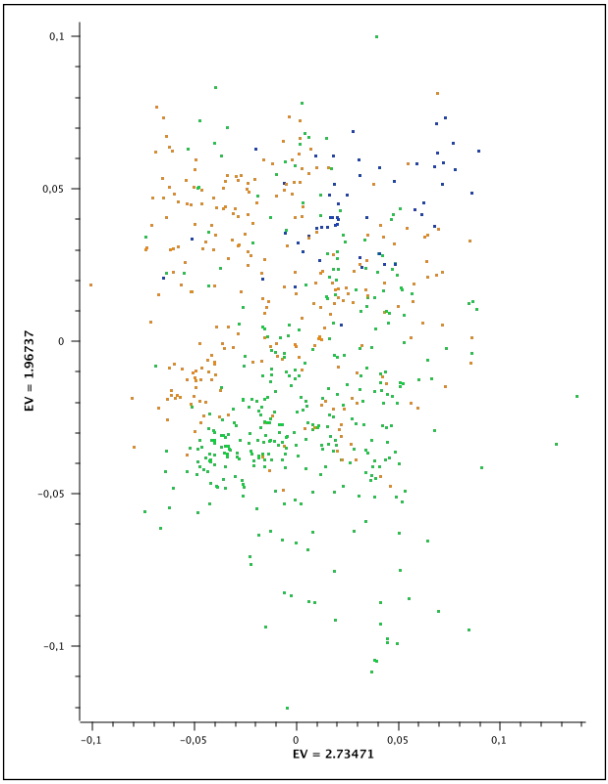


Figure 16: Plot of the second versus the third principal component that can explain the batch effects on different people that have genotyped the data.

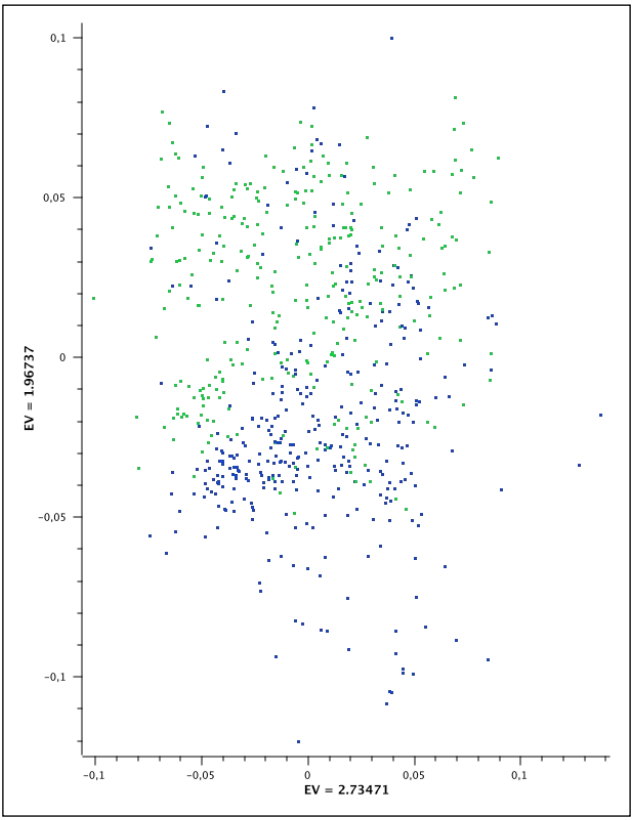


Figure 17: Plot of the second versus the third principal component that can explain the systematic difference between the two scanners used in the sample processing with a clear separation of the two clusters.

The first four PCs were used to correct the LRR values in the PCA because their value larger than 1 (Figure 18 and Table 8).

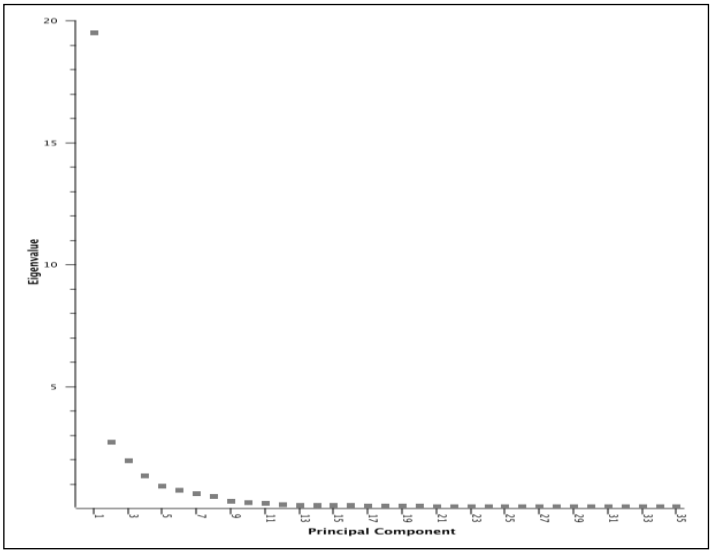


Figure 18: Plot of the Eigenvalues (y-axis) versus the principal component (x-axis).

Label	*Eigenvalue*
1	19.52
2	2.73
3	1.97
4	1.35
5	0.95
6	0.76

Table 8: The first six eigenvalues parameters were reported.

4.2.4 PENNCNV SOFTWARE: REMOVE OUTLIER BULLS

A total of 5,406 CNV events on 651 bulls were detected. In Table 9 some descriptive statistics are reported.

The overall threshold for WF and GCWF were 0.1364 and 0.0144, respectively.

Column	Minimum	Q1	Median	Mean	Q3	Maximum	Threshold	IQR
WFa*	-0.2249	-0.0419	-0.0288	-0.00186	0.0365	0.1677	0.1541	0.0784
GCWFa§	-0.0823	-0.0057	-0.0006	-0.0012	0.0037	0.0298	0.0178	0.0094
WFb*	-0.05860	-0.03175	0.0284	0.01064	0.0355	0.1588	0.1364	0.0673
GCWFb§	-0.0169	-0.0021	0.0016	0.001143	0.0045	0.0221	0.0144	0.0066

Table 9: Statistic summary of waviness.

***: Wave Factor**

§: GC Wave Factor

a: are the values obtained before the filtering of outlier bulls

b: are the values after the application filter

4.3 PENNCNV CNVS CALLING RESULTS

4.3.1 FILTERING FOR CENTROMERIC AND TELOMERIC REGIONS

According to the criteria hereinbefore exposed, i.e. using an overlap of at least 10% between the 5,406 CNV events and the centromeric and telomeric region, 307 CNV events were filtered out. Thus the final dataset used downstream analysis was composed to 5,099 CNVs.

4.3.2 DESCRIPTIVE STATISTICS OF CNVS RESULTS

PennCNV software detected 5,099 CNVs call, and as described in Table 10, these CNVs encompass 97 homozygous deletion, 2,086 heterozygous deletion, 2,915 heterozygous duplication and 1 homozygous duplication, that corresponds to copy number type equal to zero, one, three, and four (Colella et al., 2007) on 632 bulls.

Chr	type 0	type 1	type 3	type 4	# totali CNV
1	6	95	121	0	222
2	2	61	231	0	294
3	4	87	104	0	195
4	7	67	126	0	200
5	12	40	130	0	182
6	11	83	89	0	183
7	4	279	146	0	429
8	3	174	102	0	279
9	2	178	87	0	267
10	4	4	141	0	149
11	4	30	158	0	192
12	4	114	39	0	157
13	1	41	138	0	180
14	1	189	99	0	289
15	1	25	61	0	87
16	0	51	79	0	130
17	3	140	70	0	213
18	3	124	76	0	203
19	4	22	130	0	156
20	1	16	84	0	101

	Mimimum	Q1	Median	Mean	Q3	Maximum
21	5	29	122	0	156	
22	3	17	82	0	102	
23	1	68	88	0	157	
24	2	57	81	0	140	
25	4	30	68	1	103	
26	0	29	59	0	88	
27	1	5	47	0	53	
28	1	27	54	0	82	
29	3	4	103	0	110	
Total	97	2086	2915	1	5099	

Table 10: Frequency table of CNVs events identified by PennCNV software for each chromosomes, and with a different states (type 0, type1, type3, type4) and with the total of CNVs.

Table 11 shows the descriptive statistics of CNVs call identified by PennCNV software. Within a CNV, the minimum number of SNPs is 3. The length of CNV (expressed in base pairs) ranges from 40.4 kb to 4.46 Mb, with 350 and 230 kb as mean and median, respectively. The median of CNV number per bull is 4 as represented in the Figure 19.

# SNPs in CNV	3	4	6	9.451	12	111
Length CNV (bp)	40374	125551	229759	350297	446944	4457756
# CNV per bull	1	2	4	8.068	6	91

Table 11: Descriptive statistics of CNVs detected with PennCNV. Minimum, Q1 – first quartile, Median, Mean, Q3 – third quartile, Maximum, # SNPs in CNV – number of SNPs within CNV, Length CNV (bp), # CNV per bull – number of CNV for each bull.

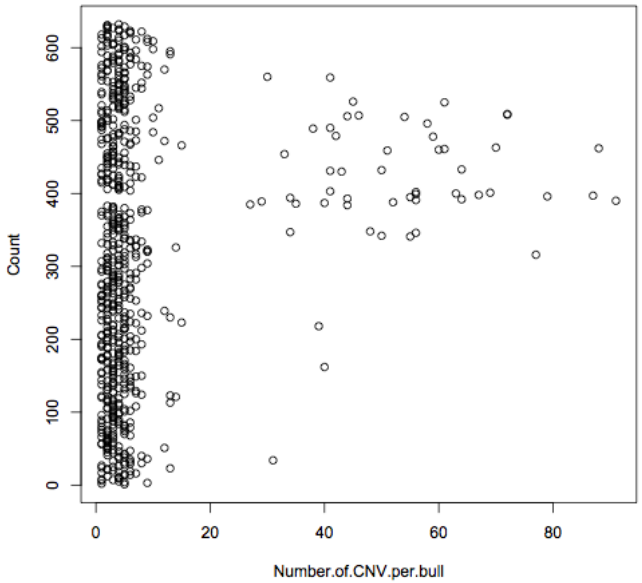


Figure 19: Plot of the distribution of the number of CNV per bull detected with PennCNV software.

More features on the CNV length are reported in the Table 12. The variable length was transformed in log10 and it was tested for the normality distribution with Kolmogorov-Smirnov test (D=0.059831, and p-value was < 0.01 thus reject the H0).

Copy number	Mean	Median	Sum	Min	Max
0	311345	245646	30200500	46665	1053143
1	159066	134534	331711379	40374	1688267
3	488559	385138	1423739019	41449	4457756
4	511301	511301	511301	511301	511301

Table 12: Descriptive statistics of CNV length separate for each copy number (0 = homozygous deletion, 1 heterozygous deletion, 3 heterozygous duplication, and 4 homozygous duplication). For the copy number 4 was detected only one event.

The effect of the CNV state (loss, gain) on the log10 transformation of the CNV length was tested with the correction of Tukey-Kramer of SAS.

The losses (homozygous and heterozygous deletion) are significantly (p-value < 0.0001) smaller respect to the gains (homozygous and heterozygous duplication) with the R-square of the model of 0.209.

All these results were tested and verified with a nonparametric test Kruskal-Wallis (p-value <0.0001).

Copy number	*lsmeans*	*Type*	*lsmeans*
0	*5.29*	*0 or 1 (loss)*	*4.99*
1	*4.98*		
3	*5.47*	*3 or 4 (gain)*	*5.47*
4	*5.71*		

Table 13: Least square means of log10 CNV length for each copy number states separately, and for the loss (homozygous and heterozygous deletion) and for the gain (homozygous and heterozygous duplication).

4.3.2.1 Genome Map of CNVs obtained by PennCNV software

Figure 20 shows the genome wide map of CNVs identified by PennCNV.

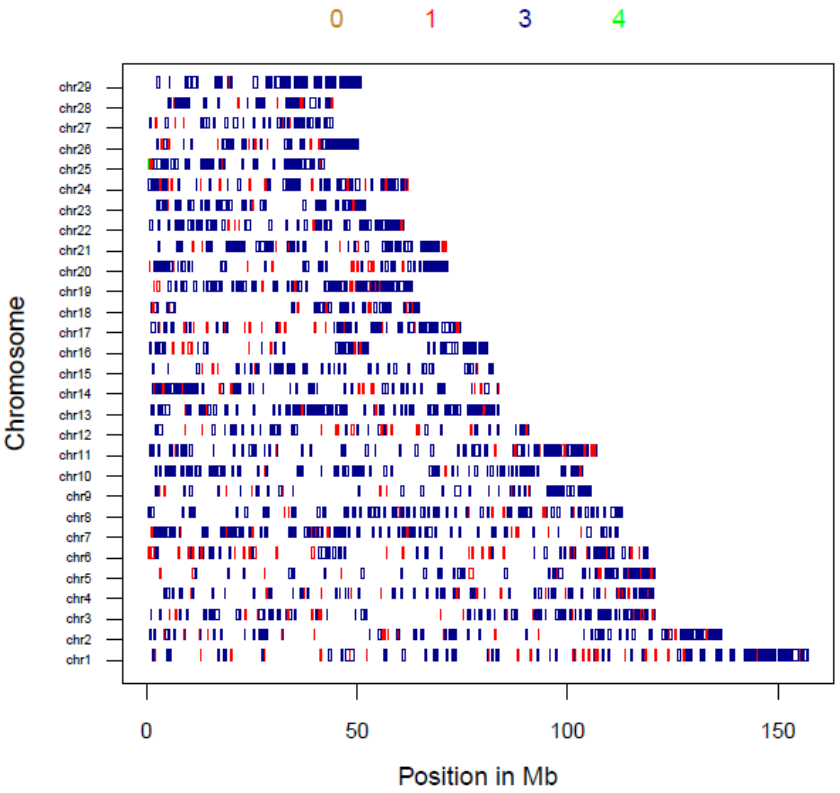


Figure 20: Genome map of CNVs identified on UMD3.1 autosomes by PennCNV software; x-axis represents the Position in Mb along each chromosome; y-axis represents the autosomes. Each lines identify CNV; 0 (Yellow): homozygous deletion, 1 (red): heterozygous deletion, 3 (blue): single copy duplication and 4 (green): homozygous duplication.

4.3.2.2 Graphical Representation of CNVs

A graphical representation of CNVs events obtained by PennCNV software for each chromosomes and visualized by HDCNV software (http://bioinformatics.oxfordjournals.org/). The red circle represents CNVs events with high number of overlap with the other events across all samples, ranging to blue events with no overlap (Figure 21).

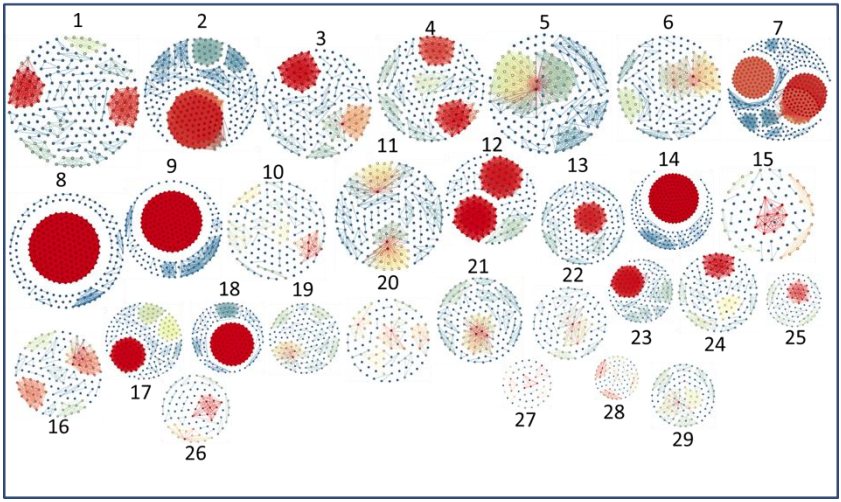


Figure 21: Karyotype of CNV events obtained by PennCNV software in bovine chromosome 1-29 and visualized by HD-CNV. Graph: each graph represents events for one chromosome. Node: a dot indicates a CNV event. The size for each chromosome depending from the total length of the chromosome. Edges: lines connect CNV events whose genomic regions overlap by at least 40%. Color: red indicates events with a high number of overlap with the other events, ranging to blue events with no overlap.

4.3.3 CNVRS IDENTIFIED WITH PENNCNV

A total of 1,101 CNVRs were mapped with PennCNV software. Among these: 220 are the homozygous or heterozygous deletion; 774 are the homozygous or heterozygous duplication; 107 represents complex regions. The total length of the sequence covered by the CNVRs is 682 Mb which correspond to the 27.14% of the bovine autosome genome in the Brown Swiss breed. The percentage of sequence covered by CNVRs ranges from 16.59 to 50.14% (Table 14).

Chr	Total length of CNVRs	Length of Chr (bp)	% Sequence covered by CNVRs	# Loss	# Gain	# Complex	Total # CNVRs
1	34697950	158337067	21,91	21	42	4	67
2	27559864	137060424	20,11	12	32	6	50
3	30977497	121430405	25,51	11	39	8	58
4	22868847	120829699	18,93	13	36	5	54
5	21613076	121191424	17,83	7	21	7	35
6	24642236	119458736	20,63	22	25	7	54
7	32333080	112638659	28,71	9	42	6	57
8	27607064	113384836	24,35	8	51	2	61
9	18355335	105708250	17,36	8	26	2	36
10	28653731	104305016	27,47	3	51	2	56
11	30510446	107310763	28,43	5	31	4	40
12	15125214	91163125	16,59	11	25	3	39
13	35001179	84240350	41,55	2	38	4	44
14	23677776	84648390	27,97	7	26	4	37
15	16578816	85296676	19,44	5	31	1	37
16	20477059	81724687	25,06	6	17	3	26
17	23143843	75158596	30,79	11	21	6	38
18	16540305	66004023	25,06	5	13	5	23
19	32119692	64057457	50,14	7	22	3	32
20	19307024	72042655	26,8	9	24	2	35
21	26064953	71599096	36,4	5	21	4	30
22	24984552	61435874	40,67	4	25	1	30
23	20627415	52530062	39,27	1	16	2	19
24	21250774	62714930	33,88	10	20	3	33
25	19126823	42904170	44,58	1	13	5	19
26	15594191	51681464	30,17	7	17	3	27
27	15333620	45407902	33,77	5	20	0	25
28	14613734	46312546	31,55	5	11	2	18
29	22312716	51505224	43,32	0	18	3	21
Sum	681698812	2512082506	27,14	220	774	107	1101

Table 14: Feature of CNVRs identified by PennCNV; Chr: number of autosome chromosome, length of chr: total length of the sequence for each autosome, # loss: number of loss events, # gain: number of gain events, # complex: number of complex events.

In Table 15 a descriptive statistics of CNVRs length (express in bp) for each type of CNVR (loss, gain and complex) obtained by PennCNV software is reported. The effect of CNVR type (complex, loss, and gain) on log10 transformation of the CNVR length was tested with the GLM procedure of SAS accounting for multiple comparison (Tukey-Kramer). The length is significantly different in losses respect to gains (p-value < 0.0001) with a R square of 0.299 (Table 16).

CNVRs Type	Mean	Median	Sum	Min	Max
Loss	210454,3773	148427,5	46299963	40754	977685
Gain	596255,2752	403827	461501583	45465	3873856
Complex	1625208,093	1068260	173897266	179707	6703707

Table 15: Descriptive statistics for each type of CNVRs (loss, gain and complex) obtained by PennCNV software with the Mean, Median, Sum, Min - Minimum and Max - Maximum values.

CNVR Type	*Lsmeans*
Complex	*6.07*

Gain	*5.63*
Loss	*5.21*

Table 16: Least square means for the log10 CNVR length obtained by PennCNV versus the independent variable the type (complex, gain, and loss).

4.4 CNAM CNVS CALLING RESULTS WITH UNIVARIATE ANALYSIS

4.4.1 DESCRIPTION OF THE INFLUENCE OF WAVE CORRECTION AND PRINCIPAL COMPONENT ANALYSIS

The histograms in Figure 22, 23, 24 represent the LRR means values segments calculated at the moment of run CNV detection. Figure 22 represents uncorrected LRR values, Figure 23 represents the LRR values corrected for wave effects only and Figure 24 represents the LRR values corrected for wave effects and cluster identified by PCA. All these Figures are scaled on same vale to make possible a direct comparison. Another important feature of all these Figures is the negative LRR values, corresponding to the homozygous deletions, being much more frequent respect to the positive ones.

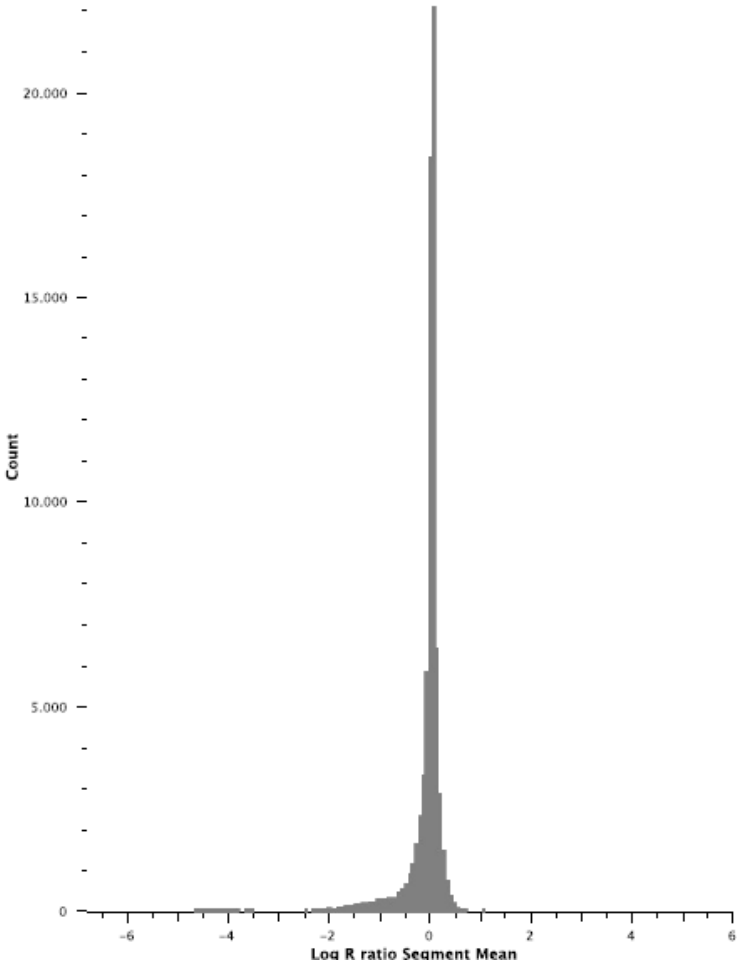


Figure 22: Histogram of Log R ratio segment means with uncorrected Log R ratio values.

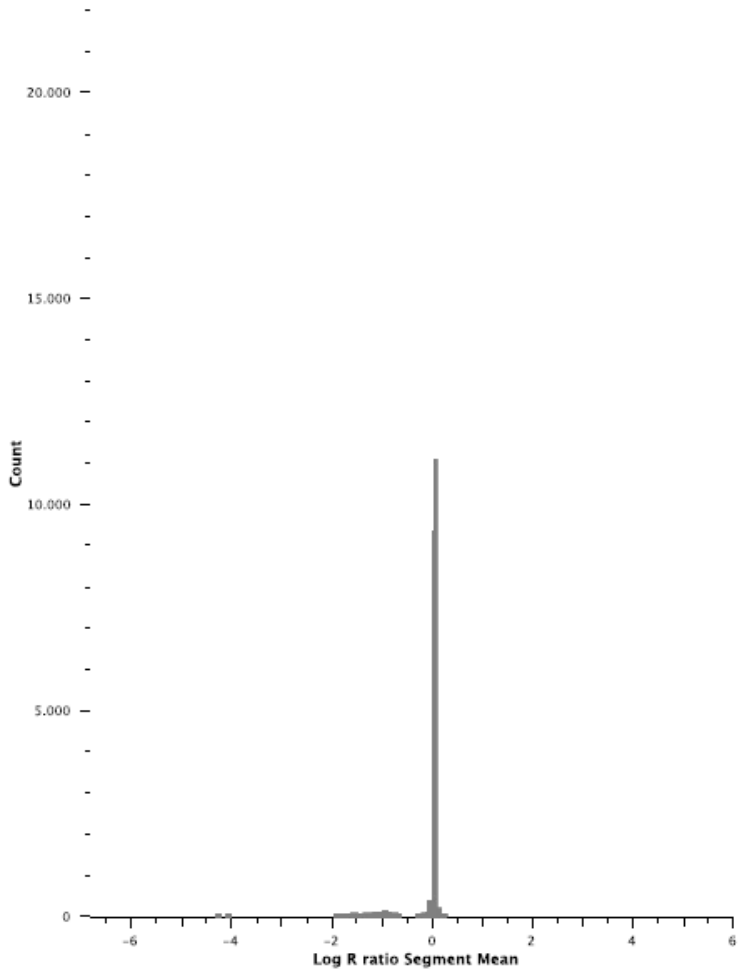


Figure 23: Histogram of Log R ratio segment means with wave corrected Log R ratio values.

Figure 24: Histogram of Log R ratio segment means based on wave corrected R ratio values that were also corrected for the first 4 PCS.

The comparison between the count of the segments obtained by SVS7 are reported in the Table 17 for the three groups of LRR values: the first is for uncorrected values, the second is for the wave corrected, and the third is for the wave and the PC corrected. In this Table is not reported the number of segments covering LRR means ranging between +/- 0.02 to +/- 0.36, that could be not a copy number or a single copy duplications and deletions.

	# Seg	#Seg avLRR <-.36 (loss)	#Seg avLRR > .36 (gain)	#Seg -.02 >avLRR <.02 (neutral)
Uncorrected LRR	74,093	8,239 (11.12%)	649 (0.88%)	17,338 (23.4%)
Wave corrected LRR	22,999	1,720 (7.48%)	69 (0.3%)	12,337 (53.64%)
Wave & PC corrected LRR	22,618	827 (3.66%)	538 (2.38%)	17,810 (78.74%)

Table 17: A comparison of the number of identified genomic segments based on uncorrected and corrected Log R ratio (LRR) data for WF and PCA, seg= segments, avLRR= LRR segment mean.

The wave and PCs correction are useful to obtain clearer signals to identify a breakpoints. In this case, it's possible to make the hypothesis that the total number of segments is smaller because the number of information contained in the chip (54,000 SNPs) were too spread to identify accurately segments breakpoints.

The overall large number of segments with an average value of LRR +/- 0.02 is possibly caused by the more appropriate wave and PCs joint correction. In the same time, the small number of

negative values (losses), could explain a possible reduction of the false positive segments.

4.4.2 FILTERING FOR CENTROMERIC AND TELOMERIC REGIONS

Using an overlap at least of 10% between the 1,365 CNV events and the centromeric and telomeric region, 76 CNV events were filtered. The final dataset used for the follow analysis was composed to 1,289 CNVs, which encompass 762 losses and 527 gains.

4.4.3 DESCRIPTIVE STATISTICS OF CNVS RESULTS

A total number of 1,289 CNVs call have been identified by CNAM using the univariate analysis. These CNVs encompass 762 losses and 527 gains, segregating in 651 bulls. Details are shown in Table 18.

Chr	Loss	Gain	# Total CNV
1	8	10	18
2	15	6	21
3	11	16	27
4	81	26	107
5	5	4	9
6	12	7	19
7	15	42	57
8	46	11	57
9	11	1	12
10	5	0	5
11	10	3	13
12	42	12	54
13	10	2	12
14	40	49	89
15	44	28	72
16	9	14	23
17	8	6	14
18	21	28	49
19	17	6	23
20	9	5	14
21	30	8	38
22	19	19	38
23	17	2	19
24	11	1	12
25	14	9	23
26	67	71	138
27	14	12	26
28	161	126	287
29	10	3	13
Total	762	527	1,289

Table 18: Frequency table of CNVs events identified by CNAM applied univariate analysis for each autosomes chromosomes, and with a different states (loss and gain) and the total of CNVs.

Table 19 shows the descriptive statistics of CNVs call identified by CNAM. The minimum number of SNPs within a CNV is 2. The length of CNV (express in base pairs) ranges from 11.3 kb to 1.4 Mb, with median 45 kb and average 88.9 kb. The median of the number of CNV per bull is 2 as represented in Figure 25.

	Mimimum	Q1	Median	Mean	Q3	Maximum
# SNPs in CNV	2	2	2	2.607	3	27
Length CNV (bp)	11315	30182	45200	88900	136783	1440751
# CNV per bull	1	1	2	2.306	3	9

Table 19: Descriptive statistics of CNVs detected with CNAM using univariate analysis, and are reported: #SNP in CNV – number of SNPs within CNV, Length CNV (bp), #CNV per bull – number of CNV per bull, Minimum, Q1 – first quartile, Median, Mean, Q3 – third quartile, Maximum

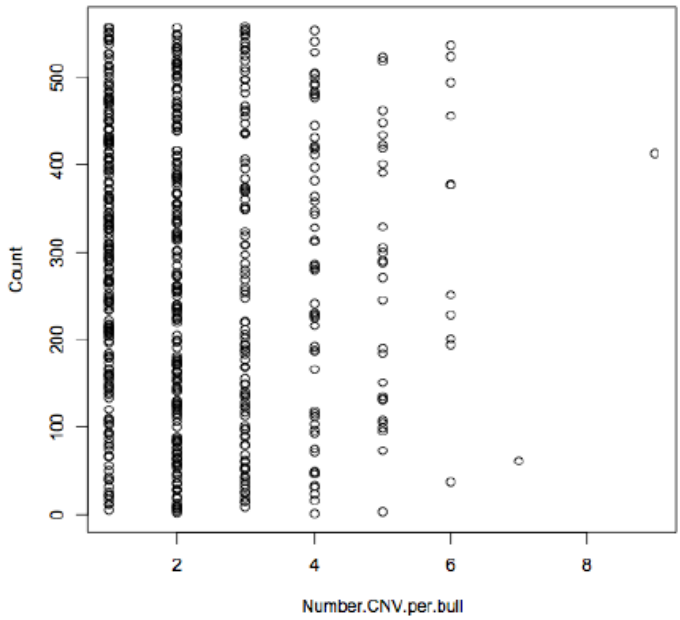


Figure 25: Histogram of the distribution of the number of CNV detected per bull using CNAM.

In Table 20 some descriptive statistics on the CNV length for deletion and duplication of copy number states are reported. The average variability of losses results larger respect to the variation of gains.

Copy number	Mean	Median	Sum	Min	Max
Loss	94830.4	57612,5	72260727	11315	1,440,751
Gain	80324.4	37591	42330968	20342	770,044

Table 20: Descriptive statistics of CNV length separate for each copy number (losses, and gains).

The variable length was transformed in log10 and it was tested for the normality distribution with Kolmogorov-Smirnov (D = 0.163303 and p-value < 0.01 thus reject the H0).

The effect of copy number state was tested on the log10 transformed length. The losses are significantly smaller respect to the gains (p-value 0.002) being the R-square of the model was 0.007391. (Table 21). This result was verified with a nonparametric test (Kruskal-Wallis) confirm the significant difference between gains and losses (p-value <0.0001).

Copy number states	*lsmeans*
0 or 1 (loss)	*4.74*
3 or 4 (gain)	*4.8*

Table 21: Least square means of log10 CNV length for CNV call for losses and gain copy number states separately.

4.4.3.1 Genome map of CNVs obtained by CNAM

Figure 26 shows a genome wide map of CNVs identified by CNAM.

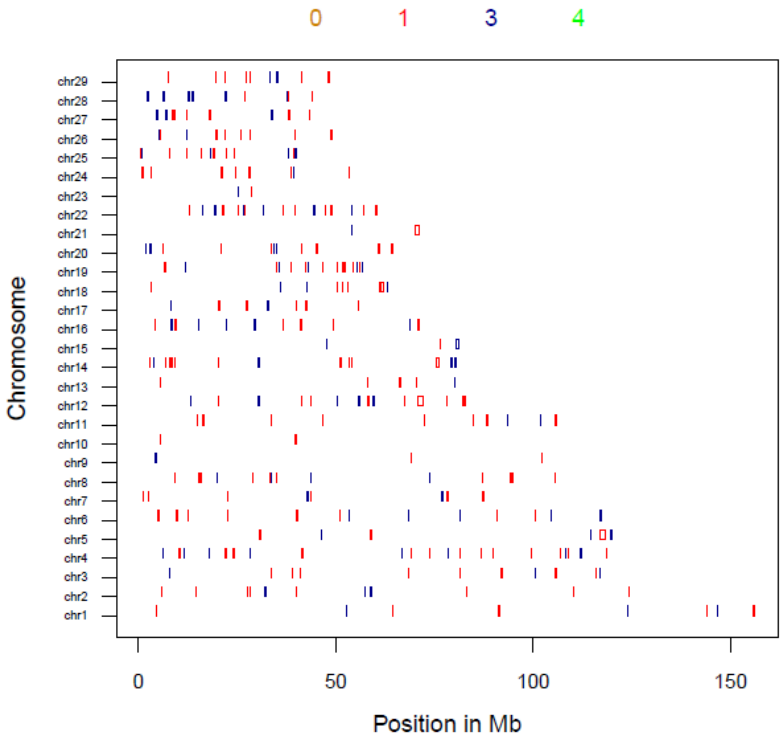


Figure 26: Genome map of CNVs identified on UMD3.1 autosomes by CNAM; x-axis represents the Position in Mb and y-axis represents the autosomes. Each y line identifies CNV; 0 (Yellow): homozygous deletion, 1 (red): heterozygous deletion, 3 (blue): single copy duplication and 4 (green): homozygous duplication.

4.4.3.2 Graphical Representation CNVRs

A graphical representation of CNVs events obtained by CNAM for each chromosomes and visualized by HDCNV software. The red circle represents CNVs events with high number of overlap with the other events across all samples, ranging to blue events with no overlap (Figure 27).

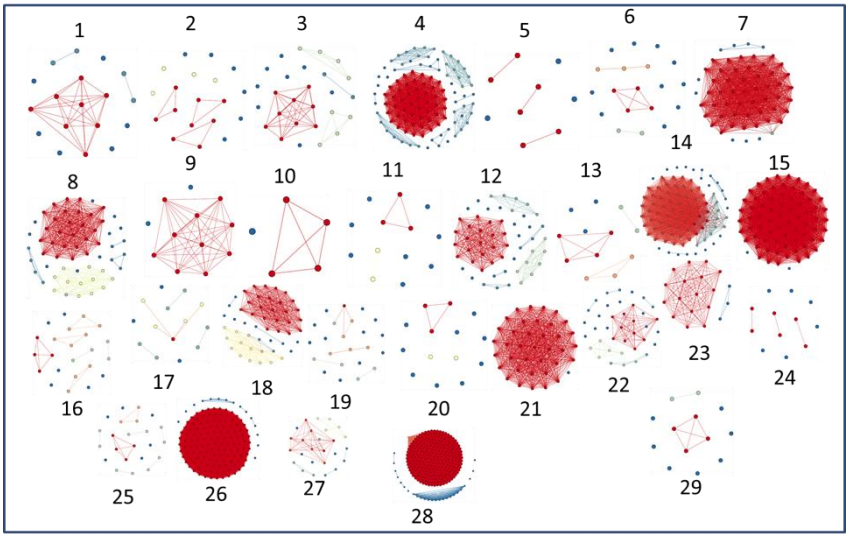


Figure 27: Karyotype of CNV events obtained by SVS7 software in bovine chromosome 1-29 and visualized by HD-CNV. Graph: each graph represents events for one chromosome. Node: a dot indicates a CNV event. The size for each chromosome depending from the total length of the chromosome. Edges: lines connect CNV events whose genomic regions overlap by at least 40%. Color: red indicates events with a high number of overlap with the other events, ranging to blue events with no overlap.

4.4.4 CNVRS IDENTIFIED WITH CNAM

The CNV calls were summarized at population level according to Redon's approach, resulting into 277 (185 losses, 56 gains and 36 complex) CNVRs. The total length of the sequence covered by the CNVRs is 33.71Mb (1.35%) of the bovine autosomes (Table 22). The percentage of sequence covered by CNVRs ranges between 0.12% to 3.5%.

Chr	Total length of CNVRs (bp)	Length of Chr (bp)	% Sequence covered by CNVRs	# Loss	# Gain	# Complex	Total # CNVRs
1	755248	158337067	0,477	5	2	1	8
2	964331	137060424	0,7036	8	2	1	11
3	835421	121430405	0,688	8	2	1	11
4	1617402	120829699	1,3386	13	4	4	21
5	2109507	121191424	1,7406	3	2	1	6
6	1001626	119458736	0,8385	8	4	1	13
7	1295002	112638659	1,1497	6	1	1	8
8	1885802	113384836	1,6632	8	1	3	12
9	242071	105708250	0,229	2	1	0	3
10	126202	104305016	0,121	2	0	0	2
11	935462	107310763	0,8717	8	1	1	10
12	3230778	91163125	3,544	9	3	2	14
13	572916	84240350	0,6801	4	0	1	5
14	2171830	84648390	2,5657	12	2	2	16
15	837756	85296676	0,9822	1	1	1	3
16	1437331	81724687	1,7587	7	5	0	12
17	964817	75158596	1,2837	5	1	1	7
18	1169431	66004023	1,7718	7	1	2	10
19	1351321	64057457	2,1095	10	3	2	15
20	1093115	72042655	1,5173	7	3	1	11
21	1142692	71599096	1,596	1	0	1	2
22	1941717	61435874	3,1606	11	5	1	17
23	149508	52530062	0,2846	1	0	1	2
24	518267	62714930	0,8264	8	1	0	9
25	1263893	42904170	2,9459	7	3	2	12
26	828044	51681464	1,6022	8	1	1	10
27	1127107	45407902	2,4822	6	2	1	9
28	1401323	46312546	3,0258	3	3	3	9
29	739246	51505224	1,4353	7	2	0	9
Sum	33709166	2512082506	1,3419	185	56	36	277

Table 22: Feature of CNVRs identified by CNAM; Chr=autosome number, length of chr – total length of the sequence for each autosome, # loss – number of losses events, # gains – number of gains events, # complex – number of complex (losses and gains) events.

In Table 23 descriptive statistics of CNVRs length (bp) for each type of CNVR (loss, gain and complex) obtained by CNAM algorithm are reported.

The effect of copy number type (complex, loss, and gain) was tested on the log10 transformed length accounting for for multiple comparison (Tukey-Kramer). The losses are significantly smaller respect to the gains (p-value 0.0214) with R-square of 0. (Table 24).

CNVRs Type	Mean	Median	Sum	Min	Max
Loss	116378,6162	61523	21530044	11314	1440750
Gain	115358,0179	83498,5	6460049	20341	460833
Complex	158863,1389	127525,5	5719073	21916	770043

Table 23: Descriptive statistic for each type of CNVRs (loss, gain and complex) obtained by CNAM algorithm.

CNVR Type	*Lsmeans*
Complex	5.03
Gain	4.91
Loss	4.84

Table 24: Least square means of dependent variable the log10 transformation of CNVR length obtained by CNAM versus the independent variable the type (complex, gain, and loss).

4.5 CONSENSUS CNVRS

Descriptive statistics of consensus CNVRs obtained by the two methods, Redon and Wain respectively are reported in Table 25. According to Redon et al., (2006) considering CNVRs overlapping for at least 1 bp of their sequence 139 consensus regions were obtained spanning a total length of 146 Mb (5.88 % autosome covered). The second method (Wain et al., 2009) considers only CNVRs that fully overlap one each other: a total of 151 consensus regions were identified with a total length of 17.1 Mb (0.68 % autosome covered).

Redon's method identify a smaller number of CNVRs, because being these regions longer, may contain multiple regions identify by Wain's approach.

Consensus	# CNVR	Total length (bp)	% autosome	Min (bp)	Max (bp)	Average (bp)
Redon	139	146611379	5.88	41600	6703707	1054758.12
Wain	151	17064978	0.68	11314	1047092	113013.099

Table 25: Descriptive statistic of consensus CNVRs obtained by the two methods: the union (Redon et al., 2006) , and the intersect method (Wain et al., 2009).

4.6 ANNOTATION OF CNVRS

The version 69 of the gene dataset downloaded from Ensembl contains a total of 26,740 annotated bovine elements. Within these elements 22,118 protein coding genes, 626 pseudogenes, 171 retro transposed, 24 MT-tRNA, 405 rRNA, 1,222 snRNA, 1,153 miRNA, 846 snoRNA and 175 miscellaneous RNA are listed.
After excluding, 1,267 elements on chromosome X and on UnChr (98 miRNA, 7 miscellaneous RNA, 24 MT-tRNA, 941 protein coding genes, 52 pseudogenes, 22 retro transposed, 14 rRNA, 39 snoRNA, 70 snRNA), and repeated information in the file, a total number of 23,431 elements on bovine autosome was considered.
In Table 26 the number and the proportion of annotated elements for each biotype classes out of the bovine autosome list is

presented. The annotation is reported for CNVRs detected by both algorithms, PennCNV and CNAM, and for the two consensus methods. The consensus is considered matching when an overlapping of at least 80% between the total length of CNVR and feature occurs. The elements of protein coding genes (PCGs) are the most represented in CNVRs.

transcript biotype	autosome #	autosome %	CNAM #	CNAM %	PennCNV #	PennCNV %	Wain #	Wain %	Redon #	Redon %
protein_coding	19135	81,67	296	87,57	6955	84,90	167	88,83	1990	88,92
pseudo gene	574	2,45	7	2,07	132	1,61	2	1,06	27	1,21
retrotransposed	149	0,64	1	0,30	33	0,40	1	0,53	7	0,31
rRNA	391	1,67	3	0,89	93	1,14	0	0,00	20	0,89
snRNA	1152	4,92	6	1,78	324	3,96	3	1,60	57	2,55
snoRNA	807	3,44	3	0,89	228	2,78	3	1,60	56	2,50
miRNA	1055	4,50	21	6,21	381	4,65	12	6,38	67	2,99
misc_RNA	168	0,72	1	0,30	46	0,56	0	0	14	0,63
Total	23431	100,00	338	100,00	8192	100,00	188	100,00	2238	100,00

Table 26: Overview of annotated ensemblv69 elements for each biotype on UMD3.1 autosomes, for the CNVRs obtained by both algorithms, and by the two consensus methods: the intersection and the union approach (Wain et al. 2009 and Redon et al. 2006), respectively. #: number of elements. The transcript biotype consists in several classes: rRNA – ribosomal RNA, snRNA – small nuclear RNA, snoRNA – small nucleolar RNA, miRNA – microRNA, misc_RNA – all other subspecies of RNA.

Table 27 shows the results for the hypothesis test that the molecular function, biological process, cellular component and patway terms were under-or overrepresented in CNVRs after Bonferroni correction.

The analysis was performed with all the four datasets used for the annotation with Ensembl elements.

BIOLOGICAL PROCESS		
GO Term	GO name	p-value
GO:0044237	cellular metabolism process	0.0419
GO:0009987	cellular process	2,06E-10
GO:0044260	Cellular macromolecule metabolic process	1,56E-01
GO:0008152	Metabolic process	3,79E-01
GO:0044238	Primary metabolic process	1.27e-4
GO:0032502	Developmental process	0.001
GO:0043170	Macromolecule metabolic process	0.001
GO:0009058	Biosynthetic process	0.001
GO:0044249	Cellular biosynthetic process	0.001
GO:0007275	Multicellular organismal development	0.0047
GO:0010467	Gene expression	0.0058
GO:0006807	Nitrogen compound metabolic process	0.019
GO:0016071	mRNA metabolic process	0.048
GO:0007166	Cell surface receptor linked signal transduction	0.01
CELLULAR COMPONENT		
GO:0005737	Cytoplasm	3.63e-4
GO:0005622	Intracellular	4.28e-4
GO:0044424	Intracellular part	8,60E-01
GO:0043226	Organelle	3,00E-08
GO:0043229	intracellular organelle	8,10E-13

GO:0043227	Membrane-bounded organelle	1,13E-06
GO:0043231	Intracellular Membrane-bounded organelle	2,18E-06
GO:0044444	Cytoplasmatic part	1,31E-01
GO:0044446	Intracellular ogganelle part	4,17E-01
GO:0044422	Organelle part	5,16E-01
GO:0005634	Nucleus	1,76E+00
GO:0044428	Nuclear part	0.0039
GO:0031974	Membrane enclosed lumen	0.0129
GO:0043233	Organelle lumen	0.02523
GO:0070013	Intracellular organelle lumen	0.0321
MOLECULAR FUNCTION		
GO:0004364	Glutathione transferase activity	7.59e-4
GO:0005515	Protein binding	3,91E-07
GO:0005488	Binding	1,87E-04
GO:0017076	Purine nucleotide binding	0.0058
GO:0003824	Catlytic activity	0.0091
GO:0032553	Ribonucletide binding	0.014
GO:0032555	Purine ribonucleotide binding	0.014
GO:0004871	Signal transducer activity	0.0203
GO:0060089	Molecular transducer activity	0.0203
KEGG PATWAY		
Bta00980	Metabolism of xenobiotics by cytochrome P450	0.0056
Bta03040	Spliceosome	0.0042
Bta03010	Ribosome	0.0076
Bta04740	Olfactory transduction	0.049

Table 27: Enriched GO terms associated with the CNVRs (Bonferroni p-value ≤ 0.05) obtained in this study.

CHAPTER 5 –DISCUSSION

Several authors have reported the large-scale variability across CNV-detection methods and also across platforms for genotyping. Additional substantial false positive and false negative rates can be associated with the methods used. For these reasons, the available copy number studies are frequently characterized by comparison of several calling algorithms and platforms.

Winchester et al. (2009) using two SNP array (Affimetrix and Illumina) discussed the comparison between CNV detection using the two different platforms.

Tsuang et al. (2010) studied the effect of four algorithms (PennCNV, QuantiSNP, HMMSeg, and cnvPartition) on copy number variant detection. In the same year, another author (Dellinger et al. 2010) suggested and emphasized a comparative analyses of seven algorithms for copy number variant identification from SNP array.

Pinto et al. (2011) compared CNV detection on eleven microarray platforms (CGH / SNP) to evaluate data quality and CNV calling, reproducibility, concordance across array platforms and laboratory sites, breakpoints accuracy and analysis tool variability. In this study, he found a low concordance < 50% between CNV calls using the same raw data but different analytic tools.

Haraksingh et al (2011) have quantified the abilities of twelve genome-wide CNV detection platforms, including CGH arrays (NimbleGen and Agilent Thecnologies) and SNP arrays (Illumina and Affymetrix), finding significant difference in performance. These differences take in account the sensitivity, the total number, the size range and the breakpoint resolution of CNV calls.

There are no available studies comparing CNV detection differences between the two algorithms here used: the PennCNV and the CNAM (copy number variation module of SVS7). In this context, additionally to the fact that this is the first mapping with CNAM in the Brown this study, bring to the knowledge of CNV mapping original results.

5.1 COMPARISON AND CONSENSUS BETWEEN PENNCNV AND CNAN CNVS

The total number of CNVs detected by PennCNV and CNAM are 5,099 and 1,289, respectively.

In Figure 21 and Figure 27 a graphical representation of CNVs events are reported. The differences between these two graphical representations can be explained by the following:

-) first of all by the datasets used to create the graphical representation had large differences in the total number of CNVs;

-) secondly the algorithms used to detect the CNVs are different.

This is in line with the hereinbefore just mentioned discussion on extreme variability in CNV detection according to algorithms used.

The median length of CNVs according to PennCNV (230 kb) was greater respect to the one detected by CNAM (45 kb). This could indicate a large variability in CNV breakpoints identified by PennCNV.

The number of CNVs detected for each bull ranging 1 to 91 is much larger in PennCNV, respect to CNAM where is ranging between 1 to 9. This could indicate the tendency to map a larger number of false positive in PennCNV.

The CNVs obtained by CNAM shows a major number of losses respect to the gains events. As was reported in the literature (Fadista et al. 2010, Gu et al. 2008, Turner et al. 2008), this over representation of deletion events can be explain by the type of chromosomal rearrangement (NAHR).

The CNVs obtained by PennCNV software shows a larger number of gains respect to the losses, but at best of our knowledge there is no biological reason to explain this phenomena.

The asymptotic Kolmogorov-Smirnov two-sample test 'proc npar1way' (SAS 9.2, SAS Institute) was used to test the differences in the empirical cumulative CNV frequency distribution, and it was significant (Figure 28).

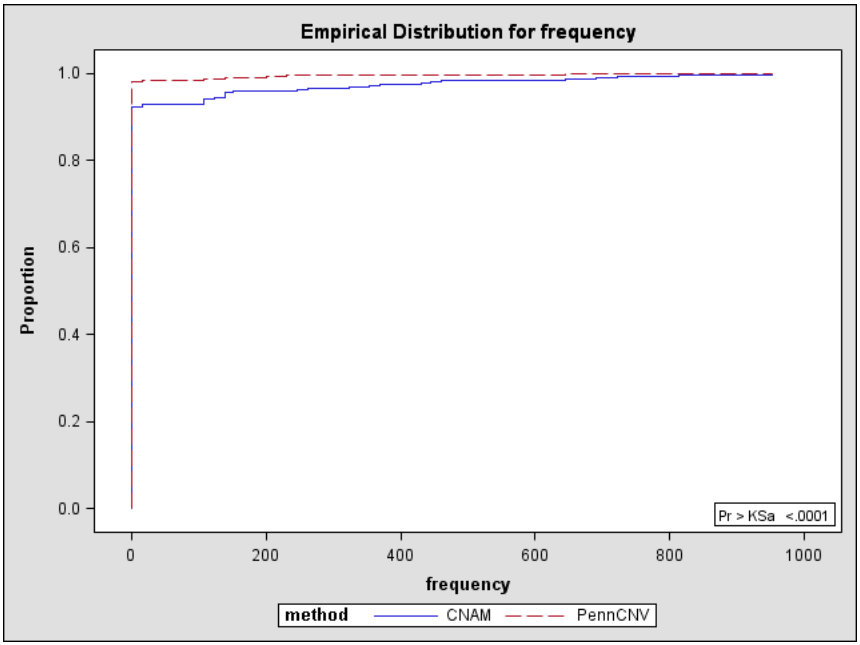


Figure 28: Empirical cumulative distribution of CNV frequencies for CNVs identified by PennCNV and CNAM.

5.2 COMPARISON AND CONSENSUS BETWEEN PENNCNV AND CNAM CNVRS

The total number of CNVRs identified by PennCNV is four time the number of CNVRs obtained by CNAM, and correspond to 1,101 and 277, respectively. The percentage on the bovine autosome of the Brown Swiss breed correspond to the 27.14% and 1.35%, respectively. This confirm the hypothesis that from same data set algorithms highly can affect CNV mapping.

5.3 COMPARISON TO LITERATURE

The comparison between the CNVRs here detected by Wain's approach (151) and the other five published CNV studies is reported in Table 28.

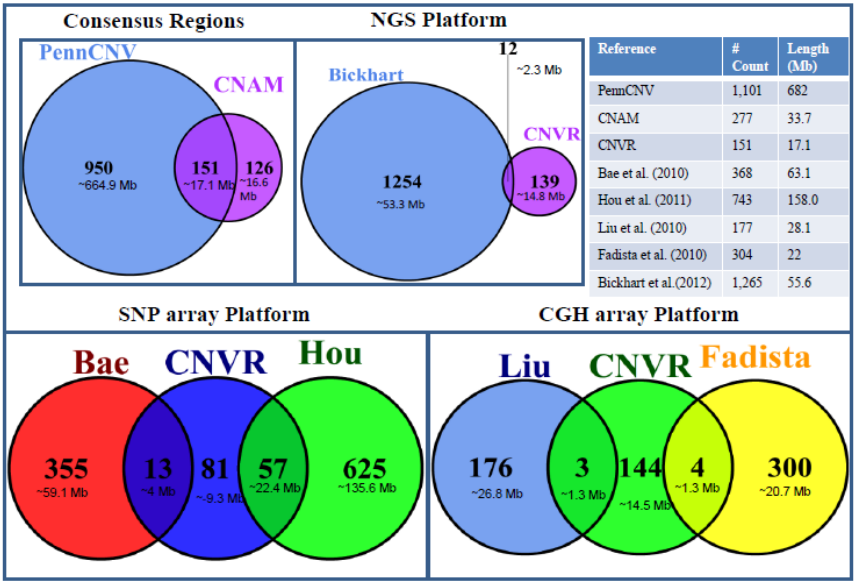
Hou et al. (2011) and Bae et al. (2010) used an SNP array platform, with Illumina BovineSNP50 BeadChip (Illumina) while Fadista et al. (2010) and Liu et al. (2010) used the comparative genomic hybridization array platform (NimbleGen). Bickhart et al. (2012) for the first time used a Next Generation Sequence to map CNV (Illumina GAIIx).

	This study	Length (Mb)	Hou et al. (2010)	Length (Mb)	Fadista et al. (2010)	Length (Mb)	Bae et al. (2010)	Length (Mb)	Liu et al. (2010)	Length (Mb)	Bickhart et al. (2012)	Length (Mb)
This study	151	17.1	57	22.4	4	1.3	13	4	3	1.3	12	2.3
Hou et al.(2010)	57	22.4	682	158	27	21.3	61	39.3	19	13.8	74	35.9
Fadista et al. (2010)	4	1.3	27	21.3	304	22	15	4.7	45	12.1	56	12.6
Bae et al. (2010)	13	4	61	39.3	15	4.7	368	63.1	5	1.7	31	7.4
Liu et al. (2010)	3	1.3	19	13.8	45	12.1	5	1.7	177	28.1	98	24.2
Bickhart et al. (2012)	12	2.3	74	35.9	56	12.6	31	7.4	98	24.2	1265	55.6

Table 28: Common CNVRs between my consensus regions and five different authors.

A graphical representation of the comparison, in term of count and length, between the results of this study and the other five datasets already published is represented in Figure 29. These graphs were created using the package Venn Diagram of R software.

The comparison with literature confirms the existence of the high variability across platform, methods breeds and populations used to detection of CNVs. In the Venn Diagram for the SNP array Platform negative values respect to the length are reported. This occurs because Hou's detected regions are larger respect to the regions obtained in this study.



Reference	# Count	Length (Mb)
PennCNV	1,101	682
CNAM	277	33.7
CNVR	151	17.1
Bae et al. (2010)	368	63.1
Hou et al. (2011)	743	158.0
Liu et al. (2010)	177	28.1
Fadista et al. (2010)	304	22
Bickhart et al.(2012)	1,265	55.6

Figure 29: Comparisons between 151 consensus regions and the other existing cattle CNVRs datasets in term of count and length. The first Venn Diagram: Consensus Region shows the two datasets obtained in this study using PennCNV software and CNAM. The intersection part shows the total number of the common regions and the length identified by both algorithms; Venn Diagram: SNP array Platform shows the comparison between the results of this study and CNVR derived from SNP array (Bae et al, 2010, Hou et al. 2011); the Venn Diagramm: CGH array Platofrom shows the comparison between the results of this study and the two CNVR datasets derived from array CGH studies (Liu et al,2010; Fadista et al, 2010); Venn Diagram: NGS Platform shows the comparison between the results of this study and the only public dataset obtained by NGS technology (Bickhart et al. 2012); the summaries and legends of existing cattle CNVR datasets.

5.3 FUNCTIONAL ANNOTATION

Another comparison with the literature was based on the functional annotation. From the results obtained by Gene Onthology analysis and KEGG Patway a similarity with other authors is found. Bae et al. (2010) reported genes significantly enriched in the identified bovine CNVs for the cytoplasm, intracellular part, cytoplasmic part, and intracellular organelle.

Hou et al. (2011) reported that several CNVs are important in drug detoxification, defense/innate and adaptive immunity and receptor and signal recognition. These gene families include

olfactory receptors, ATP-binding cassette (ABC) transporters, Cytochrome P450, β-defensins, interleukins, the bovine MHC (BoLA) and multiple solute carrier family proteins.

CHAPTER 6 - CONCLUSION AND COMPARISON WITH LITERATURE

In this study a genomic analysis of Brown Swiss dairy bulls using two different algorithms to detect CNVs based on whole genome SNP genotyping data was performed. A total of 139 and 151 CNVRs were identified with the consensus analysis, covering 146 Mb (~ 5.88%) and 17.1 Mb (~ 0.68%) of the bovine autosome, respectively.

The comparison between the other studies in cattle CNV shows us a high variability that may be due to the different platforms or

technologies used. However, a subset of CNVRs obtained by this study, overlap with other CNVRs dataset cattle studies.

Future analysis will be processed to confirm this genomic scan map using next generation sequencing data, and/or molecular technique.

The SNP data combined with available CNV data scan may help to identify genes undergoing artificial selection in domesticated animals and to improve the selection program.

CHAPTER 7 –REFERENCES

Aldhous M.C., Abu Bakar S., PrescottN.J., Palla R., Soo K., MansfieldJ.C., MathewC.G., Satsangi J., Armour J.A.L. (2010). Measurement methods and accuracy in copy number variation: failure to replicate associations of beta-defensin copy number with Crohn's disease. *Hum. Mol. Genet.*19 (24): 4930-4938 (doi: 10.1093/hmg/ddq411).

Alkan C., Coe B.P., Eichler E.E. (2011). Genome structural variation discovery and genotyping. *Nature reviews. Genetics*. Volume 12: 363:376. doi:10.1038/nrg2958.

Armour J.A., Sismani C., Patsalis P. C., and Cross G. (2000). Measurement of locus copy number by hybridisation with amplifiable probes. *Nucleic Acids Res*. **28**, 605–609.

Avent N.D., Martin P.G., Armstrong-Fisher S.S., Liu W., Finning K.M., Maddocks D., Urbaniak S.J. (1997). Evidence of genetic diversity underlying Rh D-, weak D (Du), and partial D phenotypes as determined by multiplex polymerase chain reaction analysis of the *RHD* gene. *Blood:* **89:** 2568–77.

Bae J.S., Cheong H.S., Kim L.H., Gung S.N., Park T.J., Chun J.Y., Kim J.Y., Pasaje C.F., Lee J.S., Shin H.D.(2010). Identification of copy number variations and common deletion polymorphisms in cattle. *BMC Genomics* , 11:232.

Bailey, J.A. Gu Z., Clark R.A., Reinert K., Samonte R.V., Schwartz S., Adams M.D., Myers E.W., Li P.W., Eichler E.E. (2002). Recent segmental duplications in the human genome. *Science* **297**, 1003–1007.

Bickhart D.M., Hou Y., Schroeder S.G., Alkan C., Cardone M.F.,Matukumalli L.K., Song J., Schnabel R.D., Ventura M., Taylor J.F., Garcia J.F., Van Tassell C.P., Sonstegard T.S., Eichler E.E.,and Liu G.E. (2012). Copy number variation of individual cattle genomes using next-generation sequencing. *Genome Research.* doi/10.1101/gr.133967.111.

Bridges C.B. (1921). Triploid Intersexs in Drosophila Melanogaster. *Science.* Vol 54:252-254.

Bridges C.B. (1936).The bar "gene" a duplication. Science 83(2148):210-1 (DOI: 10.1126/science.83.2148.210).

Butler J., Locke M.E.O., Hill K.A., and Daley M. (2012). HD-CNV: Hotspot Detector for Copy Number Variants. *Bioinformatics*. doi:10.1093/bioinformatics/bts650.

Campbell P.J., Pleasance E.D., Stephens P.J., Dicks E., Rance R., Goodhead I.,
Chartier-Harlin, M.C., Kachergus, J., Roumier, C., Mouroux, V., Douay, X., Lincoln, S., Levecque, C., Larvor, L., Andrieux, J., Hulihan, M., Waucquier, N., Defebvre, L., Amouyel, P., Farrer, M., and Destee, A. (2004). Alpha-synuclein locus duplication as a cause of familial Parkinson's disease. *Lancet* 364(9440): 1167-1169.

Chartier-Harlin, M.C., Kachergus, J., Roumier, C., Mouroux, V., Douay, X., Lincoln, S., Levecque, C., Larvor, L., Andrieux, J., Hulihan, M., Waucquier, N., Defebvre, L., Amouyel, P., Farrer, M., and Destee, A. (2004). Alpha-synuclein locus duplication as a cause of familial Parkinson's disease. *Lancet* 364(9440): 1167-1169.

Clop A., Vidal O., Amills M. (2012). Copy number variation in the genomes of domestic animals. *Anim Genet* doi: 10.1111/j.1365-2052.2012.02317.x.

Colella S., Yau C., Taylor J.M., Mirza G., Butler H., Clouston P., Bassett A.S., Seller A., Holmes C.C., Ragoussis J. (2007). QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res* 35:2013-25.

Conrad B., Antonarakis S.E. (2007). Gene duplication: A drive for phenotypic diversity and cause of human disease. *Annu Rev Genomics Hum Genet* 8: 17–35.

Conrad D.F. and Hurles M.E. (2007). The population genetics of structural variation. Nat Genet 39:S30-36 doi:10.1038/ng2042.

Conrad D.F., Bird C., Blackburne B., Lindsay S., Mamanova L., Lee C., Turner D.J., Hurles M.E. (2010). Mutation spectrum revealed by breakpoint sequencing of human germline CNVs. *Nat Genet* 42:385-391.

Conrad D.F., Pinto D., Redon R., Feuk L., Gokcumen O., Zhang Y., Aerts J., Andrews T.D., Barnes C., Campbell P., Fitzgerald T., Hu M., Ihm C.H., Kristiansson K., Macarthur D.G., Macdonald J.R., Onyiah I., Pang A.W., Robson S., Stirrups K., Valsesia A., Walter K., Wei J.; Welcome Trust Case Control Consortium, Tyler-Smith C., Carter N.P., Lee C., Scherer S.W., Hurles M.E. (2010). Origins and functional impact of copy number variation in the human genome. *Nature* 464:704-712 doi:10.1038/nature08516.

De Cid R., Riveira-Mun˜ oz E., Zeeuwen P.L. (2009) Deletion of the late cornified envelope LCE3B and LCE3C genes as a susceptibility factor for psoriasis. *Nature Genetics* 41, 211–5.

Dellinger A.E., Saw S.M., Goh L.K., Seielstad M., Terri L., Young T.L. and Li Y.J. (2010). Comparative analyses of seven algorithms for copy number variant identification from single nucleotide polymorphism arrays. *Nuc Acid Res* 38:9 e105, doi:10.1093/nar/gkq040.

Dermitzakis E., Clark A. (2001*).* Non-neutral diversification after duplication in mammalian developmental genes *Mol. Biol. Evol* 18:557-*562.*

Dhami, P., Coffey, A.J., Abbs, S., Vermeesch, J.R., Dumanski, J.P., Woodward, K.J., Andrews, R.M., Langford, C., and Vetrie, D. (2005). Exon array CGH: detection of copy-number changes at the resolution of individual exons in the human genome. *Am J Hum Genet* 76(5): 750-762.

Diskin S.J., Hou C., Glessner J.T., Attiyeh E.F., Laudenslager M., Bosse K., Cole K., Mosse Y.P., Wood A., Lynch J.E. (2009). Copy number variation at 1q21.1 associated with neuroblastoma. *Nature* 459: 987–991.

Diskin S.J., Li M., Hou C., Yang S., Glessner J., Hakonarson H., Bucan M., Maris J.M. and Wang K. (2008). Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nuc Acid Res* 36 19 e126, doi:10.1093/nar/gkn556.

Dopman E.B., Hartl D.L. (2007). A portrait of copy-number polymorphism in Drosophila melanogaster. *Proc Natl Acad Sci* 104:19920-19925.

Drogemuller C., Distl O., Leeb T. (2001). Partial deletion of the bovine ED1 gene causes anhidrotic ectodermal dysplasia in cattle. *Genome Res* 11: 1699–1705.

Eichler E.E. (2006). Widening the spectrum of human genetic variation. *Nat Genet* 38: 911.

Elsik C.G., Tellam R.L., Worley K.C., Gibbs R.A., Muzny D.M., Weinstock G.M., Adelson D.L., Eichler E.E., Elnitski L., Guigo R. *(2009).* The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science*, **324**(5926)**:**522-528.

Emerson J.J., Cardoso-Moreira M., Borevitz J.O., Long M. (2008). Natural selection shapes genome-wide patterns of copy-number polymorphism in Drosophila melanogaster. *Science* 320: 1629–1631.

Estivill X. and Armengol L. (2007). Copy number variants and common disorders: Filling the gaps and exploring complexity in genome-wide association studies. *PLoS Genet* 3:1787–1799. doi:10.1371/journal.pgen.0030190.

Fadista J., Thomsen B., Holm L.E. and Bendixen B.M.C. (2010). Copy Number Variation in the Bovine Genome. *Genomics* 11:284 doi:10.1186/1471-2164-11-84.

Feuk L., Carson A.R., Scherer S.W. (2006). Structural variation in the human genome. *Nat Rev Genet* 7:85-97.

Flisikowski K., Venhoranta H., Nowacka-Woszuk J. (2010) A novel mutation in the maternally imprinted PEG3 domain results in a loss of MIMT1 expression and causes abortions and stillbirths in cattle (Bos taurus). *PLoS ONE* 5, e15116.

Follows G.A., Green A.R., Futreal P.A., and Stratton M.R. (2008). Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing**.** *PNAS*: Vol. 105 no. 35 (doi_10.1073_pnas.0801523105).

Fontanesi L., Beretti F., Martelli P.L., Colombo M., Dall'olio S., Occidente M., Portolano B., Casadio R., Matassino D. and Russo V. (2011). A first comparative map of copy number variations in the sheep genome. *Genomics* 97:158-65.

Fontanesi L., Beretti F., Riggio V., Dall'Olio S., Occidente M., Incoronato C., Martelli P.L., Casadio R., Portolano B. (2009). A Comparative Analysis of Copy Number Variation of the Sheep and Goat Genomes.

Fontanesi L., Martelli P.L., Beretti F., Riggio V., Dall'olio S., Colombo M., Casadio R., Russo V. and Portolano B. (2010). An initial comparative map of copy number variations in the goat (Capra hircus) genome. *BMC Genomics* 11: 639.

Giuffra E., Evans G., To¨rnsten A., Wales R., Day A., Looft H., Plastow G. & Andersson L. (1999). The Belt mutation in pigs is an allele at the Dominant white (I/KIT) locus. *Mammalian Genome* 10, 1132–6.

Giuffra E., To¨rnsten A., Marklund S., Bongcam-Rudloff E., Chardon P., Kijas J.M., Anderson S.I., Archibald A.L. & Andersson L. (2002). A large duplication associated with dominant white color in pigs originated by homologous recombination between LINE elements flanking KIT. *Mammalian Genome* 13, 569–77.

Glessner J.T., Reilly M.P., Kim C.E. (2010) Strong synaptic transmission impact by copy number variations in schizophrenia. Proceedings of the National Academy of Sciences of the United States of America 107, 10584–9.

Gonzalez E., Kulkarni H., Bolivar H., Mangano A., Sanchez R., Catano G., Nibbs R.J., Freedman B.I., Quinones M.P., Bamshad M.J. (2005). The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science* 307: 1434–1440.

Goosens M., Dozy A. M., Embury S. H., Zachariades Z., Hadjiminas M. G., Stamatoyannopouloasn G. , Kan D.W. (1980). Triplicated a-globin loci in humans. *Proc. Natl. Acad. Sci.* 77:5 18-521.

Gu W., Zhang F. and Lupski J.R. (2008). Mechanisms for human genomic rearrangements. *PathoGenetics*, **1**:4 doi:10.1186/1755-8417-1-4.

Heid, C.A., Stevens, J., Livak, K.J., and Williams, P.M. (1996). Real time quantitative PCR. *Genome Res* 6(10): 986-994.

Henrichsen C.N., Chaignat E. & Reymond A. (2009). Copy number variants, diseases and gene expression. *Human Molecular Genetics* 18, R1–8.

Higuchi, R., Dollinger, G., Walsh, P.S., and Griffith, R. (1992). Simultaneous amplification and detection of specific DNA sequences. *Biotechnology (N Y)* 10(4): 413-417.

Hirano T., Kobayashi N., Itoh T., Takasuga A., Nakamaru T., Hirotsune S. & Sugimoto Y. (2000). Null mutation of PCLN-1/ Claudin-16 results in bovine chronic interstitial nephritis. *Genome Research* 10, 659–63.

Hou Y., Liu G.E., Bickhart D.M., Cardone M.F., Wang K., Kim E., Matukumalli L.K., Ventura M., Song J., VanRaden P.M., Sonstegard T.S., Van Tassell C.P. (2011). Genomic characteristics of cattle copy number variations. *BMC Genomics*, 12:127.

Hutt K.J., McLaughlin E.A. & Holland M.K. (2006) Kit ligand and c-Kit have diverse roles during mammalian oogenesis and folliculogenesis. *Molecular Human Reproduction* 12, 61–9.

Iafrate A.J., Feuk L., Rivera M.N., Listewnik M.L., Donahoe P.K., Qi Y., Scherer S.W., and Lee C. (2004). Detection of large-scale variation in the human genome. *Nat Genet* 36:949–951 doi:10.1038/ng1416.

Kim P.M., Lam H.Y., Urban A.E., Korbel J.O., Affourtit J., Grubert F., Chen X., Weissman S., Snyder M., Gerstein M.B. (2008). Analysis of copy number variants and segmental duplications in the human genome: Evidence for a change in the process of formation in recent evolutionary history. *Genome Res* 18: 1865–1874.

Kleinjan DA, van Heyningen V. (2005). Long-range control of gene expression: emerging mechanisms and disruption in disease. *Am J Hum Genet.* 76(1):8-32.

Knudsen O. (1958) Studies on spermiocytogenesis in the bull. *International Journal of Fertility* 3, 389–403.

Langer P.R., Waldrop A.A., Ward D.C. (1981). Enzymatic synthesis of biotin-labeled polynucleotides: Novel nucleic acid affinity probes. *Proc natl Acad Sci*, USA 78:6633–6637.

Lee A.S., Gutiérrez-Arcelus M., Perry G.H., Vallender E.J., Johnson W.E., Miller G.M., Korbel J.O., Lee C. (2008). Analysis of copy number variation in the rhesus macaque genome identifies candidate loci for evolutionary and human disease studies. *Hum Mol Genet* 17:1127-1136. doi: 10.1093/hmg/ddn002.

Lee C. and Scherer S.W. (2010). The clinical context of copy number variation in the human genome. *Expert Rev Mol Med* 12:e8, (doi:10.1017/S1462399410001390).

Lee J.A., Inoue K., Cheung S.W., Shaw C.A., Stankiewicz P., Lupski J.R. (2006). Role of genomic architecture in *PLP1* duplication causing Pelizaeus-Merzbacher disease. *Hum Mol Genet:* **15:**2250-2265.

Liu G.E., Brown T., Hebert D.A., Cardone M.F., Hou Y., Choudhary R.K., Shaffer J., Amazu C., Connor E.E., Ventura M., Gasbarre L.C. (2011). Initial analysis of copy number variations in cattle selected for resistance or susceptibility to intestinal nematodes. *Mamm Genome* 22:111-121.

Liu G.E., Hou Y., Zhu B., Cardone M.F., Jiang L., Cellamare A., Mitra A., Alexander L.J., Coutinho L.L., Dell'Aquila M.E., Gasbarre L.C., Lacalandra G., Li R.W., Matukumalli L.K., Nonneman D., Regitano L.C. de A, Smith T.P.L., Song J., Sonstegard T.S., Van Tassell C.P., Ventura M., Eichler E.E., McDaneld T.G., Keele J.W. (2010). Analysis of copy number variations among diverse cattle breeds. *Genome Res* 20:693-703. doi: 10.1101/gr.105403.110.

Liu G.E., Ventura M., Cellamare A., Chen L., Cheng Z., Zhu B., Li C., Song J., Eichler E.E. (2009). Analysis of recent segmental duplications in the bovine genome. *BMC Genomics* 10:571 doi:10.1186/1471-2164-10-571.

Lupski J.R. (2003). Genomic Disorders: Recombination-Based Disease Resulting from Genome Architecture. *Am J Hum Genet.* 72(2): 246–252.

Lupski J.R. (2004). Hotspots of homologous recombination in the human genome: not all homologous sequences are equal. *Genome Biol*, 5:242.

Lupski J.R. (2006). Genome structural variation and sporadic disease traits. *Nat Genet*, 38:974-976.

Lupski J.R. (2007). Genomic rearrangements and sporadic disease**.** *Nat Genet*, 39(Suppl 7):S43-47.

Lupski J.R., Stankiewicz P. (2005). Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes. *PLoS Genet*, **1:**e49.

Lynch M., Conery J.S. (2000). The Evolutionary Fate and Consequences of Duplicate Genes. *Science* Vol. 290: 1151-1155. (DOI: 10.1126/science.290.5494.1151).

Marioni J.C., Thorne N.P., Valsesia A., Fitzgerald T., Redon R., Fiegler H. (2007). Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization. *Genome Biol*.;8:R228. doi: 10.1186/gb-2007-8-10-r228.

Matukumalli L.K., Lawley C.T., Schnabel R.D., Taylor J.F., Allan M.F., Heaton M.P., Connell J., Moore S.S., Smith T.P., Sonstegard T.S., Van Tassell C.P. (2009). Development and characterization of a high density SNP genotyping assay for cattle. *PLoS ONE* 4:e5350.

McCarroll S.A., Kuruvilla F.G., Korn J.M., Cawley S., Nemesh J., Wysoker A., Shapero M.H., de Bakker P.I., Maller J.B., Kirby, Elliott A.L., Parkin M., Hubbell E., Webster T., Mei R., Veitch J., Collins P.J., Handsaker R., Lincoln S., Nizzari M., Blume J., Jnes K.W., Rava R., Daly M.J., Gabriel S.B. and Altshuler D. (2008). Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet* 40:1166–1174, doi:10.1038/ng.238.

Meyers S.N., McDaneld T.G., Swist S.L., Marron B.M., Steffen D.J., O'Toole D., O'Connell J.R., Beever J.E., Sonstegard S. & Smith T.P. (2010) A deletion mutation in bovine SLC4A2 is associated with osteopetrosis in Red Angus cattle. *BMC Genomics* 11, 337.

Mills, R.E, Luttig C.T., Larkins C.E., Beauchamp A., Tsui C., Pittard W.S., Devine S.E. (2006). An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res.* **16**, 1182–1190.

Nannya Y., Sanada M., Nakazaki K., Hosoya N., Wang L., Hangaishi A., Kurokawa M., Chiba S., Bailey D.K., Kennedy G.C., Ogawa S. (2005). A Robust Algorithm for Copy Number Detection Using High-Density Oligonucleotide Single Nucleotide Polymorphism Genotyping Arrays. *Cancer Res:* 65:6071-6079. doi:10.1158/0008-5472.CAN-05-0465.

Nei M., Rooney A.P. (2005). Concerted and birth-and-death evolution of multigene families. *Annu Rev Genet* 39: 121–152.

Nguyen D-Q., Webber C., Ponting C.P. (2006). Bias of Selection on Human Copy-Number Variants. *PLoS Genet* 2(2): e20.

Norris B.J. & Whan V.A. (2008) A gene duplication affecting expression of the ovine ASIP gene is responsible for white and black sheep. Genome Research 18, 1282–93.

Ohba Y., Kitagawa H., Kitoh K., Sasaki Y., Takami M., Shinkai Y., Kunieda T. (2000). A deletion of the paracellin-1 gene is responsible for renal tubular dysplasia in cattle. *Genomics* 68: 229–236.

Olshen A.B., Venkatraman E.S., Lucito R., Wigler M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* 5: 557–572.

Peiffer D.A., Le J.M., Steemers F.J., Chang W., Jenniges T., Garcia F., Haden K., Li J., Shaw C.A., Belmont J., Cheung S.W., Shen R.M., Barker D.L., Gunderson K.L. (2006). High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Res* 16:1136-48.

Perry G.H., Dominy N.J., Claw K.G. (2007). Diet and the evolution of human amylase gene copy number variation. *Nature Genetics* 39, 1256–60.

Perry G.H., Tchinda J., McGrath S.D., Zhang J., Picker S.R., Caceres A.M., Iafrate A.J., Tyler-Smith C., Scherer S.W., Eichler E.E. (2006). Hotspots for copy number variation in chimpanzees and humans. *Proc Natl Acad Sci* 103:8006-8011.

Pinto D., Darvishi K., Shi X., Rajan D., Rigler D., Fitzgerald T., Lionel A.C., Thiruvahindrapuram B., MacDonald J.R., Mills R., Prasad A., Noonan K., Gribble S., Prigmore E., Donahoe P.K., Smith R.S., Park J.H., Hurles M.E., Carter N.P., Lee C., Scherer S.W., Feuk L. (2011). Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nat Biotech* 29:512–520, doi:10.1038/nbt.1852.

Pollack, J.R., Perou, C.M., Alizadeh, A.A., Eisen, M.B., Pergamenschikov, A., Williams, C.F., Jeffrey, S.S., Botstein, D., and Brown, P.O. (1999). Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat Genet* 23(1): 41-46.

R 2.15.0 http://cran.r-project.org/

Redon R., Ishikawa S., Fitch K.R., Feuk L., Perry G.H., Andrews T.D., Fiegler H., Shapero M.H., Carson A.R., Chen W. (2006). Global variation in copy number in the human genome. *Nature* 444, 444–454.

Rosengren-Pielberg G., Golovko A., Sundstro¨m E. Curik I., Lennartsson J. (2008). A cis-acting regulatory mutation causes premature hair graying and susceptibility to melanoma in the horse. *Nature Genetics* 40, 1004–9.

Rovelet-Lecrux, A., Hannequin, D., Raux, G., Le Meur, N., Laquerriere, A., Vital, A., Dumanchin, C., Feuillette, S., Brice, A., Vercelletto, M., Dubas, F., Frebourg, T., and Campion, D. (2006). APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy. *Nat Genet* 38(1): 24-26.

SAS (2008): Software: Release 9.1.3 SAS Institute Inc., Cary NC, USA 27513.

Schook L.B. & Lamont S.J. (1996). The Major Histocompatibility Region of Domestic Animal Species. *CRC Press Inc.*, Boca Raton, FL.

Schouten J.P., McElgunn C.J., Waaijer R., Zwijnenburg D., Diepvens F. & Pals G. (2002). Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucleic Acids Research* 30, e57.

Sebat J., Lakshmi B., Malhotra D., Troge J., Lese-Martin C., Walsh T., Yamrom B., Yoon S., Krasnitz A., Kendall J. (2007). Strong association of de novo copy number mutations with autism. *Science* 316: 445–449.

Sebat J., Lakshmi B., Troge J., Alexander J., Young J., Lundin P., Maner S., Massa H., Walker M., Chi M. (2004). Large-scale copy number polymorphism in the human genome. *Science* 305: 525–528.

Sen A., and Srivastava M. S. (1975). On tests for detecting a change in mean. *Annals of Statistics* **3**, 98–108.

Seroussi E., Glick G., Shirak A., Yakobson E., Weller J.I, Ezra E. and Zeron Y.(2010). Analysis of copy loss and gain variations in Holstein cattle

autosomes using BeadChip SNPs. *BMC Genomics*, **11**:673 doi:10.1186/1471-2164-11-673.

Sharp A.J., Locke D.P., McGrath S.D., Cheng Z., Bailey J.A., Vallente R.U., Pertz L.M., Clark R.A., Schwartz S., Segraves R., Oseroff V.V., Albertson D.G., Pinkel D., Eichler E.E. (2005). Segmental duplications and copy number variation in the human genome. *Am J Hum Genet*, **77:**78-88.

Sharp A.J., Mefford H.C., Li K., Baker C., Skinner C., Stevenson R.E., Schroer R.J., Novara F., De Gregori M., Ciccone R., Broomer A., Casuga I., Wang Y., Xiao C., Barbacioru C., Gimelli G., Bernardina B.D., Torniero C., Giorda R., Regan R., Murday V., Mansour S., Fichera M., Castiglia L., Failla P., Ventura M., Jiang Z., Cooper G.M., Knight S.J., Romano C., Zuffardi O., Chen C., Schwartz C.E., Eichler E.E. (2008). A recurrent 15q13.3 microdeletion syndrome associated with mental retardation and seizures. *Nat Genet*, 40:322-328.

Shaw C.J., and Lupski J.R. (2005). Non-recurrent 17p11.2 deletions are generated by homologous and non-homologous mechanisms. *Hum. Genet*. 116, 1–7.

Singleton A.B., Farrer M., Johnson J., Singleton A., Hague M.S., Kachergus J., Hulihan M., Peuralinna T., Dutra A., Nussbaum R., Lincoln S., Crawley A., Hanson, Maraganore D., Adler C., Cookson M. R., Muenter M., Baptista M., Miller D., Blancato J., Hardy J., Gwinn-Hardy K. (2003). α-Synuclein Locus Triplication Causes Parkinson's Disease. *Science* Vol. 302 no. 5646 p. 841 (DOI: 10.1126/science.1090278).

Snijders, A.M., Nowak N., Segraves R., Blackwood S., Brown N., Conroy J., Hamilton G., Hindle A.K., Huey B., Kimura K., Law S., Myambo K., Palmer J., Ylstra B., Yue J.P., Gray J.W., Jain A.N., Pinkel D., Albertson D.G. (2001). Assembly of microarrays for genome-wide measurement of DNA copy number. *Nature Genet.* **29**, 263–264.

SNP & Variation Suite v7.6.4, (SVS7) Golden Helix, Bozeman, MT, www.goldenhelix.com

Stankiewicz P., Lupski J.R. (2002). Genome architecture, rearrangements and genomic disorders. *Trends Genet*, **18:**74-82.

Stankiewicz P., Lupski J.R. (2010). Structural Variation in the Human Genome and its Role in Disease. *Annual Review of Medicine***.** Vol. 61: 437-455 (DOI: 10.1146/annurev-med-100708-204735).

Stefansson H, Rujescu D, Cichon S, Pietiläinen OP, Ingason A, Steinberg S, Fossdal R, Sigurdsson E, Sigmundsson T, Buizer-Voskamp JE, Hansen T, Jakobsen KD, Muglia P, Francks C, Matthews PM, Gylfason A, Halldorsson BV, Gudbjartsson D, Thorgeirsson TE, Sigurdsson A, Jonasdottir A, Jonasdottir A, Bjornsson A, Mattiasdottir S, Blondal T, Haraldsson M, Magnusdottir BB, Giegling I, Möller HJ, Hartmann A, Shianna KV, Ge D, Need AC, Crombie C, Fraser G, Walker N, Lonnqvist J, Suvisaari J, Tuulio-Henriksson A, Paunio T, Toulopoulou T, Bramon E, Di Forti M, Murray R, Ruggeri M, Vassos E, Tosato S, Walshe M, Li T, Vasilescu C, Mühleisen TW, Wang AG, Ullum H, Djurovic S, Melle I, Olesen J, Kiemeney LA, Franke B; GROUP, Sabatti C, Freimer NB, Gulcher JR, Thorsteinsdottir U, Kong A, Andreassen OA, Ophoff RA, Georgi A, Rietschel M, Werge T, Petursson H, Goldstein DB, Nöthen MM, Peltonen L, Collier DA, St Clair D, Stefansson K. (2008). Large recurrent microdeletions associated with schizophrenia. *Nature.* Sep 11;455(7210):232-6 (doi: 10.1038/nature07229).

Stefansson H., Ophoff R.A., Steinberg S., Andreassen O.A., Cichon S., Rujescu D., Werge T., Pietilainen O.P., Mors O., Mortensen P.B. *(*2009). Common variants conferring risk of schizophrenia. *Nature* 460*:* 744–747.

The Bovine HapMap Consortium. (2009). Genome wide survey of SNP variation uncovers the genetic structure of cattle breeds. *Science* 324: 528–532.

The International Schizophrenia Consortium. (2008). Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature*; **455:** 237–41.

Tsuang D.W., Millard S.P., Ely B., Chi P., Wang K., Raskind W.H., Kim S., Brkanac Z., Yu C.E. (2010). The effect of algorithms on copy number variant detection. *PLoS One* 5:e14456.

Turner D.J., Miretti M., Rajan D., Fiegler H., Carter N.P., Blayney M.L., Beck S., Hurles M.E. (2008). Germline rates of *de novo* meiotic deletions and duplications causing several genomic disorders. *Nat Genet*, **40:**90-95.

Urban, A.E., Korbel, J.O., Selzer, R., Richmond, T., Hacker, A., Popescu, G.V., Cubells, J.F., Green, R., Emanuel, B.S., Gerstein, M.B., Weissman,

S.M., and Snyder, M. (2006). High-resolution mapping of DNA copy alterations in human chromosome 22 using high-density tiling oligonucleotide arrays. *Proc Natl Acad Sci U S A* 103(12): 4534-4539.

Van Tassell C.P., Smith T.P., Matukumalli L.K., Taylor J.F., Schnabel R.D., Lawley C.T. (2008). SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nat Methods*;5:247–252. doi: 10.1038/nmeth.1185.

Veltman, M.W., Craig, E.E., and Bolton, P.F. (2005). Autism spectrum disorders in Prader-Willi and Angelman syndromes: a systematic review. *Psychiatr Genet* 15(4): 243-254.

Wain L.V., Armour J.A.L., Tobin M.D. (2009). Genomic copy number variation, human health, and disease. *Lancet*; 374: 340–50.

Wain L.V., Pedroso I., Landers J.E., Breen G., Shaw C.E. (2009). The Role of Copy Number Variation in Susceptibility to Amyotrophic Lateral Sclerosis: Genome-Wide Association Study and Comparison with Published Loci. *PLoS ONE* 4(12): e8175. doi:10.1371/journal.pone.0008175.

Walsh T., McClellanJ.M., McCarthy S.E., Addington A.M., Pierce S.B., Cooper G.M., NordA.S., Kusenda M., Malhotra D., Bhandari A., Stray S.M., Rippey C.F., Roccanova P., Makarov V., Lakshmi B., Findling R.L., Sikich L., Stromberg T., Merriman B.N., Butler P., Eckstrand K., Noory L., Gochman P., Long R., Chen Z., Davis S., Baker C., Eichler E.E., Meltzer P.S., Nelson S.F., Singleton A.B., Lee M.K., RapoportJ.L., King M.C., Sebat J. (2008). Rare Structural Variants Disrupt Multiple Genes in Neurodevelopmental Pathways in Schizophrenia. *Science Vol. 320 no. 5875 pp. 539-543 (DOI: 10.1126/science.1155174).*

Wang K., Chen Z., Tadesse M.G., Glessner J., Grant S.F.A., Hakonarson H., Bucan M., Li M. (2008). Modeling genetic inheritance of copy number variations. *Nuc Acid Res* 36:e138.

Wang K., Li M., Hadley D., Liu R., Glessner J., Grant S.F.A., Hakonarson H.,and Bucan M. (2007). PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* 17: 1665-1674. doi:10.1101/gr.6861907.

Wehrle-Haller B. (2003) The role of kit-ligand in melanocyte development and epidermal homeostasis. *Pigment Cell Res*. 16,287-296.

Weiss L.A., Shen Y., Korn J.M., Arking D.E., Miller D.T., Fossdal R., Saemundsen E., Stefansson H., Ferreira M.A., Green T., Platt O.S., Ruderfer D.M., Walsh C.A., Altshuler D., Chakravarti A., Tanzi R.E., Stefansson K., Santangelo S.L., Gusella J.F., Sklar P., Wu B.L., Daly M.J. (2008). Autism Consortium. Association between microdeletion and microduplication at 16p11.2 and autism. *N Engl J Med* **358:** 667–75.

White, S., Kalf, M., Liu, Q., Villerius, M., Engelsma, D., Kriek, M., Vollebregt, E., Bakker, B., van Ommen,G.J., Breuning, M.H., and den Dunnen, J.T. (2002). Comprehensive detection of genomic duplications and deletions in the DMD gene, by use of multiplex amplifiable probe hybridization. *Am J Hum Genet* 71(2): 365-374.

Winchester L., Yau C., Ragoussis Comparing (2009). CNVdetection methods for SNP arrays. *J Briefings Funcl Genom Prot* 8:353-366, (doi:10.1093/bfgp/elp017).

Xu B., Roos J.L., Levy S., van Rensburg E.J., Gogos J.A., and Karayiorgou M. (2008). Strong association of *de novo* copy number mutations with sporadic schizophrenia. *Nature Genetics* **40**, 880 – (doi:10.1038/ng.162).

Xu Y., Peng B., Fu Y., Amos C.I. (2011). Genome-wide algorithm for detecting CNV association with diseases. *BMC Bioinformatics*, 12:331 (doi:10.1186/1471-2105-12-331).

Zhang F., Gu W., Hurles M.E., Lupski J.R. (2009).Copy number variation in human health, disease, and evolution. *Annu Rev Genomics Hum Genet*.;10: 451 481. doi:10.1146/annurev.genom.9.081307.164217.

Zhou, Q., Zhang, G., Zhang, Y., Xu, S., Zhao, R., Zhan, Z., Li, X., Ding, Y., Yang, S., and Wang, W. (2008). On the origin of new genes in Drosophila. *Genome Res* 18(9): 1446-1455.

Zimin A.V., Delcher A.L., Florea L., Kelley D.R., Schatz M.C., Puiu D., Hanrahan F., Pertea G., Van Tassell C.P., Sonstegard T.S., Marçais G., Roberts M., Subramanian P., Yorke J.A. and Salzberg S.L. (2009). A

whole-genome assembly of the domestic cow, *Bos Taurus*. *Genome Biology*, 10:R42 (doi:10.1186/gb-2009-10-4-r42).

WEB REFERENCES:

http://david.abcc.ncifcrf.gov/
http://code.google.com/p/bedtools/
ftp://ftp.sanger.ac.uk/
http://daleylab.org
http://cnv.gene-quantification.info/
http://www.csiro.au/en/Outcomes/Food-and-Agriculture/Bovine-genome-decoded/Similarities-between-cow-and-human-DNA.aspx
www.bovinegenome.org
www.goldenhelix.com
http://www.cran.org
http://www.ensembl.org/biomart/martview/76d1cab099658c68bde77f7daf55117e
http://www.illumina.com/
http://www.kosgenetic.com/
http://www.ncbi.nlm.nih.gov/genome/82
http://www.neogen.com/
http://www.openbioinformatics.org/penncnv/
http://www.rarechromo.org/information/other/array%20cgh%20ftnw.pdf
ftp://ftp.ncbi.nih.gov/genomes/Bos_taurus/Assembled_chromosomes/seq

P0535

A Medium Resolution SNP Array Based Copy Number Variants Scan in Brown Swiss Dairy Cattle

Laura Pellegrino[1], Marlies A. Dolezal[1], Christian Maltecca[2], Dinesh Velayutham[1], Fausta Schiavini[1], Enrico Santus[3], *Chris Warkup*[4], Alessandro Bagnato[1]
[1]Università degli Studi di Milano, Milano, Italy; [2]NC State University, Raleigh, NC; [3]ANARB, Italy; [4] Bioscience KTN

OBJECTIVE

To produce a medium resolution genome map of CNVRs obtained by the consensus of two scans performed with PennCNV and SVS7 in the Italian Brown Swiss

INTRODUCTION

- Copy number variations (CNV) have been identified as an important source of genomic structural variation
- CNVs consist of duplications, insertions or deletions of chromosomal segments in comparison with a reference genome of at least 1 kb in size
- Genomic regions harbouring CNVs in a population are summarized as copy number variable regions (CNVRs)
- CNV are not uniformly distributed throughout the genome, but enriched in centromeric and telomeric regions and in segmental duplication
- CNVs are increasingly recognized as functional elements in the genome acting through different mechanisms. Association between CNVs and some phenotypes, mainly human diseases, have been established
- Consensus analysis among different SNP array based CNV mapping algorithms has been proposed by several authors

RESULTS

Table 1. Summary statistics of PennCNV and SVS7 results respectively

Software	# CNVs	# CNVRs	# Loss	# Gain	# Complex	Total length(bp)	% Autosome covered	Min (bp)	Max (bp)	Average (bp)
PennCNV	5,099	1,101	220	774	107	681698812	27.14	40754	6703707	1348445.7
SVS7	1,289	277	185	56	36	33709166	13.419	11314	1440750	130199.9

Figure 5. Graphical visualization for BTA4, BTA12 and BTA18 of CNVRs across all samples using HD-CNV software[5]. **Node:** each dot represents a CNV event. **Edges:** lines connect CNV events whose genomic regions overlap by at least 40%. **Color:** Red shows events with a high number of overlap with other events, ranging to blue events with no overlap.

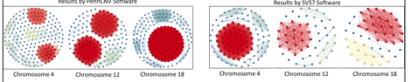


Table 2. Summary statistic of Consensus Regions

CONSENSUS	# CNVRs	Total length (bp)	% Autosome covered	Min (bp)	Max (bp)	Average (bp)
Redon[3]	139	146611379	5.88	41600	6703707	1054758.12
Wain[4]	151	17064978	0.68	11314	1047092	113013.1

Table 3. Overview of annotated Ensembl v69 elements per biotype on UMD3.1 autosome

Transcript biotype	Autosome #	Autosome %	Wain #	Wain %	Redon #	Redon %
protein coding gene	19135	81,67	167	88,83	1990	88,92
pseudo gene	574	2,45	2	1,06	27	1,21
retrotransposed	149	0,64	1	0,53	7	0,31
rRNA	391	1,67	0	0,00	20	0,89
snRNA	1152	4,92	3	1,60	57	2,55
snoRNA	807	3,44	3	1,60	56	2,50
miRNA	1055	4,50	12	6,38	67	2,99
misc. RNA	168	0,72	0	0	14	0,63
total	23431	100,00	188	100,00	2238	100,00

CNVRs are slightly enriched for protein coding genes. GO analysis[6] identified genes (Bonferroni corrected) in the CNVRs related to cytoplasm, intracellular part, cellular processes, cytoplasmic part, and intracellular organelles.

CONCLUSIONS

As expected the number of CNVRs identified here is smaller compared to a scan in the same population with Illumina's HD SNP chip (139/151-MD vs. 203-HD) (see POSTER P0533). A cross validation between 50k, HD, whole sequence data and qPCR is currently in progress within Quantomics project.

MATERIALS AND METHODS

EXTRACTION DNA & GENOTYPING:

•Sample population: 1342 bulls of Italian Brown Swiss
•Illumina BovineSNP50 BeadChip intensity signals were analyzed with Illumina's Genome Studio software producing data for a total of 46,728 SNP anchored on UMD3.1 assembly

DATA EDITING USING SVS7[1] SOFTWARE:

- Filtering of low-quality samples : derivative log ratio spread (DLRS)

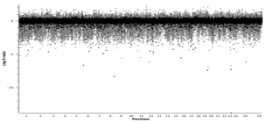


Figure 1. Plot of numeric values for Log R ratio (LRR) values for two samples
-) Black dots sample: high signal to noise ratio
-) Grey dots sample: low signal to noise ratio

- Genomic waves detection and correction

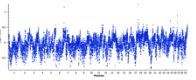


Figure 2. Plot of LRR vs. position for a low quality sample: high wave factor

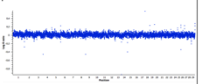


Figure 3. Plot of LRR vs. position for a high quality sample: low wave factor

- Principal Component Analysis (PCA) to correct for the effects due to XX different factors. The first 4 principal components were considered in data correction (with eigenvalues greater than 1)

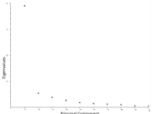


Figure 4. Plot of the eigenvalues (y axis) vs. the principal component (x axis)

- After quality filtering a total of 651 bulls were considered for CNV mapping

UNIVARIATE ANALYSIS:

- PennCNV[1]
- SVS7[2]

CONSENSUS OF CNVRs BETWEEN TWO ALGORITHMS :

- CNVs in centromeric and telomeric regions not considered
- Redon union set[3] (▯) of CNV regions (I/I) with at least 1 bp overlap
- Wain intersection set[4] (▯) of overlapping CNV regions (I/I)

ANNOTATION OF CNVRs:

- Using ENSEMBL v69 gene set
- Gene Ontology (GO) analysis using DAVID database

REFERENCES

1 Wang et al. (2007) *Genome Research* 17:1665-1674
2 SNP and Variation Suite SVS7 v.7.6.4 Golden Helix, Bozeman, MT (www.goldenhelix.com)
3 Redon et al. (2006) *Nature* doi:10.1038/nature05329
4 Wain et al. (2009)PLoS ONE 4(12):e8175 doi:10.1371/journal.pone.0008175
5 HD-CNV Software (http://dalevlab.org.)
6 DAVID database (http://david.abcc.ncifcrf.gov)

P0533

High resolution copy number variable regions in Brown Swiss dairy cattle and their value as markers

Marlies A Dolezal[1,2,3*], Karin Schlangen[1*], Laura Pellegrino[1], Morris Soller[4], Enrico Santus[5], Markus Jaritz[3], Alessandro Bagnato[1]

[1]Università degli Studi di Milano, [2]Vetmeduni Vienna, Austria, [3]FH-Campus Vienna, Austria, [4]Hebrew University of Jerusalem, Israel,
[5]Associazione Nazionale Allevatori Razza Bruna, Italy; *equally contributing

OBJECTIVES

- ✓ To produce a high resolution genome map of CNVRs in the Brown Swiss cattle breed.
- ✓ To assess the usefulness of CNV loci for population- and quantitative genetic analysis.

INTRODUCTION

- ✓ CNVs important source of genomic structural variation
- ✓ CNVs are duplications, insertions, deletions of chromosomal segments compared to a reference genome
- ✓ CNVs increasingly recognized as functional elements
- ✓ Genomic regions that contain CNVs in a population summarized to copy number variable regions (CNVRs)

MATERIAL & METHODS

- ✓ 192 Brown Swiss bulls
- ✓ Illumina HD chip; 735,238 SNPs on UMD3.1 autosomes
- ✓ Total signal intensity values (LRR) & allelic intensity ratio values (BAF)
- ✓ Stringent filtering of low-quality samples (waviness, derivative LRR spread) -> 164 bulls
- ✓ Correction for GC content and possible batch effects via PCA
- ✓ CNV detection with PennCNV[1], CNAM[2] & genoCN[3]
- ✓ Filtering in centromeric and telomeric regions
- ✓ Discretizing LRR segment means to copy number losses & gains for CNAM
- ✓ Identification of intersection consensus CNVRs[4]
- ✓ Annotation of consensus CNVRs with Ensembl v67 bovine gene set
- ✓ GO annotation with DAVID[5]
- ✓ Determination of total allelic content of CNV-SNPs with genoCN[3]
- ✓ Inference of haplotypic phases with polyHap v2[6]
- ✓ Determination of polymorphic information content (PIC) and linkage disequilibrium (LD) values measured with Wn metric [7,8]

CONCLUSIONS

- ✓ Low concordance for low frequency & short CNV calls between algorithms -> consensus calls important
- ✓ CNVRs enriched for protein coding genes, pseudogenes, retrotransposed genes & biological functions involved in immunity
- ✓ Gain breakpoints are less accurate
- ✓ As expected losses disrupt LD
- ✓ Higher LD among cn≠2 than cn=2 in gain regions
- ✓ As expected PIC in CNVRs of copy number variable bulls > PIC for copy number normal ones
- ✓ CNVs provide additional information for population and quantitative genetic analysis

RESULTS

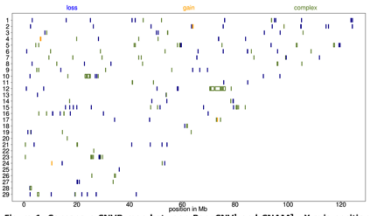


Figure 1: Consensus CNVR map between PennCNV[1] and CNAM[2] ; X-axis position on UMD3.1; Y-axis Bos Taurus autosomes, complex CNVRs contain gains & losses

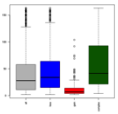


Fig 2: Boxplot of average number of cn ≠2 bulls in different classes of CNVRs; all: loss, gain & complex

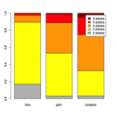


Fig 3: Number of alleles (0, A, B, AA, AB, BB) per CNVR considering cn=2 and cn ≠2 bulls

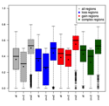


Fig 4: PIC content in CNVRs, cn=2: copy number normal, cn≠2: copy number variable, all: cn=2 & cn≠2 bulls

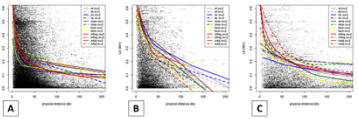


Figure 5: Decay of pairwise LD with physical distance in different classes of CNVRs (A) loss, (B) gain and (C) complex; cn=2: copy number normal bulls - dashed lines, cn≠2 copy number variable bulls - solid lines. Smoothed curves[9] were fitted for LD between: SNPs within CNVRs (red, inReg); 10 SNPs 3' & 5' of CNVRs (brown, inAdj); the 3' & 5' flanking SNP and all CNVR-SNPs (yellow, flank); tenth to sixth SNP 3' & 5' of CNVRs with all CNVR-SNPs (far, blue); fifth to first SNP 3' & 5' of CNVR with all CNVR-SNPs (close, green); 10 SNPs 3' & 5' of CNVR & all CNVR-SNPs (grey, all)

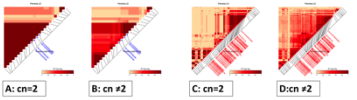


Fig. 6: Wn² LD heatmaps of CVNRs showing disruption of LD for a loss region (plot A vs. B) and formation of LD for a gain region (plot C vs. D) comparing cn=2 and cn ≠2 animals for the same set of SNPs.

REFERENCES

1 Wang et al. (2007) *Genome Research* 17:1665-1674
2 SNP & Variation Suite v7.6.4 (Golden Helix, Bozeman, MT, www.goldenhelix.com
3 Sun et al. (2009) Nucleic Acid Res. 37(16):3565-3577
4 Wain et al. (2009) PLoS ONE doi:10.1371/journal.pone.0008175
5 Huang et al. (2009) Nature Protoc. 4(1):44-57
6 Su et al. (2008) BMC Bioinformatics 9:513-521
7 Wineinger et al. (2011) Frontiers in Genetics 2(17) doi: 10.3389/fgene.2011.00017
8 Zhao JH (2007) J. Stat. Softw. 23:1-18
9 R package: Friedman's SuperSmoother. http://stat.ethz.ch/R-manual/R-patched/library/stats/html/supsmu.html

UNIVERSITÀ DEGLI STUDI DI MILANO

NC STATE UNIVERSITY — Animal Science Department — Genetics and Genomics

SNP array based identification of copy number variants in Italian Brown Swiss cattle

Laura Pellegrino[1], Marlies A. Dolezal[1], Christian Maltecca[2], Dinesh Velayutham[1], Fausta Schiavini[1], Attilio Rossoni[3], Alessandro Bagnato[1]
[1]Università degli Studi di Milano, Milano, Italy, [2]NC State University, Raleigh, NC, [3]Associazione Nazionale Allevatori Razza Bruna

OBJECTIVE

Production of a medium resolution genome map of CNVRs in the Italian Brown Swiss cattle breed.

INTRODUCTION

- **Copy number variations (CNV) have been identified as an important source of genomic structural variation**
- **CNVs consist of duplications, insertions or deletions of chromosomal segments in comparison with a reference genome of at least 1 kb in size**
- **Genomic regions harbouring CNVs in a population are summarized as copy number variable regions (CNVRs)**
- **CNVs are not uniformly distributed throughout the genome, but enriched in centromeric and telomeric regions and in segmental duplications**
- **CNVs are increasingly recognized as functional elements in the genome acting through different mechanisms. Associations between CNVs and various phenotypes mainly human diseases have been established**

MATERIALS AND METHODS

Sample population: 1353 Italian Brown Swiss bulls

Illumina BovineSNP50 BeadChip data for a total of 46,728 SNP anchored on UMD3.1 assembly were analyzed with Illumina's Genome Studio software

CNVs were detected via a hidden Markov model based on total signal intensity (Log R Ratio) and allelic intensity ratio (B Allele Frequency) using *PennCNV* software

Data editing:
- **Differences in hybridization efficiency due to the sequence composition flanking each SNP were accounted for in the model to reduce false positive calls - (correction of genomic waves)**
- **filtering of low-quality samples: SD of Log R Ratio > 0.3**
- **CNVs overlapping centromeric and telomeric regions were filtered**

CNV/CNVRs Detection

RESULTS

Table.1: Summary statistics

# CNVs	# CNVRs	# Loss	# Gain	# Both
1014	468	398	78	8
total length (bp)	**% autosome covered**	**min (bp)**	**max (bp)**	**average (bp)**
116,016,852	4.66%	38,236	7,965,447	249,498

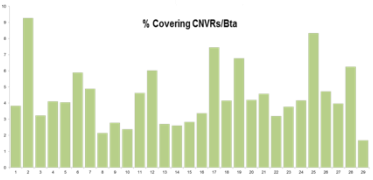


Figure.1: Percentage of assembled UMD3.1 autosomes covered by CNVRs

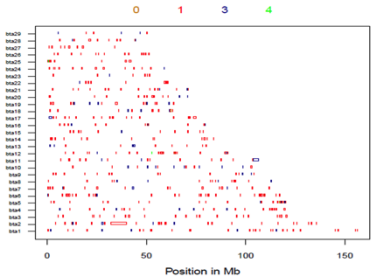


Figure.2: Distribution of CNVRs on UMD3.1 autosomes. Yellow: homozygous deletion, red: heterozygous deletion, blue: single copy duplication and green: homozygous duplication.

CONCLUSIONS

Here we present a genome map of CNVRs in the Italian Brown Swiss cattle breed identified from 50k Illumina SNP array data. Accurate CNVR maps are an important pre-requisite for population genetic and quantitative genetic analyses of copy number polymorphisms.

REFERENCE AND ACKNOWLEDGMENTS

Wang et al. (2007) ***Genome Research*** 17:1665-1674
Redon et al. (2006) ***Nature*** doi:10.1038/nature05329

The Italian Brown Breeders Association (ANARB) for providing the genotype data. This work was fundend by EU-QUANTOMICS *contract n. 222664-2*