Final Scientific Report

Cover Page

BARD Project Number: IS-4201-09

Date of Submission of the report: 1/9/2010

Project Title: Genome-Wide Association Study of Copy Number Variation and QTL for Economic Traits in Holstein Cattle

Investigators

Institutions

Principal Investigator (PI):

Eyal Seroussi

ARO

Co-Principal Investigator (Co-PI):

George Liu

USDA

Keywords *not* appearing in the title and in order of importance. Avoid abbreviations.
milk production, tiling array, beadchip array, fertility

Abbreviations commonly used in the report, in alphabetical order:

CNV- copy number variation; **SNP-** single nucleotide polymorphism; **DPR-** daughter pregnancy rate; **QTL-** quantitative trait locus; **qPCR-** quantitative PCR

Budget: IS: \$ 28,000

US: \$ 82,000

Total: \$ 110,000

Signature
Principal Investigator

Signature
Authorizing Official, Principal Institution

Final Scientific Report

Publication Summary (numbers)

	Joint IS/US authorship	US Authors only	Israeli Authors only	Total
Refereed (published, in press, accepted) BARD support acknowledged	-	-	-	-
Submitted, in review, in preparation	-	1	1	2
Invited review papers	-	-	-	-
Book chapters	-	-	-	-
Books	-	-	-	-
Master theses	-	-	-	-
Ph.D. theses	-	-	-	-
Abstracts	-	-	3	3
Not refereed (proceedings, reports, etc.)	-	-	-	-

Postdoctoral Training: List the names and social security/identity numbers of all postdocs who received more than 50% of their funding by the grant.

Cooperation Summary (numbers)

	From US to Israel	From Israel to US	Together, elsewhere	Total
Short Visits & Meetings	-	-	-	-
Longer Visits (Sabbaticals)	-	-	-	-

Description Cooperation:

Within the feasibility study, we performed the activities of preparation of DNA samples, Comparative Genomic Hybridization (CGH), initial association study between CNVs and production traits and characterization of CNV-SNP associations. Most of the funding was dedicated to the CGH experiment, which was performed using NimbleGen chips. Since the USA side negotiated a reduced price for these chips and since it was scientifically important that samples from both countries would be hybridized using the same reference and experiment setting, we decided that the USA side will be also responsible for the hybridization of the Israeli samples, which were sent to the USA. This also facilitated the bureaucracy associated with organizing this experiment and thus ensured that the results were produced in a timely fashion. Consequently, the relative portion of the USA budget was increased to 75% of the total budget. USA samples were sent to the Israeli side to be analyzed by sequencing and qPCR of specific loci. The results were made available for both side, exchange of opinions and the result interpretation was frequently conducted using e. mail correspondence.

Patent Summary (numbers)

	Israeli inventor only	US inventor only	Joint IS/US inventors	Total
Submitted	-	-	-	-
Issued (allowed)	-	-	-	-
Licensed	-	-	-	-

Abstract

Copy number variation (CNV) has been recently identified in human and other mammalian genomes and increasing awareness that CNV might be a major source for heritable variation in complex traits has emerged. Despite this, little has been published on CNVs in Holsteins. In order to fill this knowledge-gap, we proposed a genome-wide association study between quantitative trait loci (QTL) for economic traits and CNV in the Holstein cattle. The approved feasibility study was aimed at the genome-wide characterization of CNVs in Holstein cattle and at the demonstrating of their possible association with economic traits by performing the activities of preparation of DNA samples, Comparative Genomic Hybridization (CGH), initial association study between CNVs and production traits and characterization of CNV-SNP associations. For both countries, 40 genomic DNA samples of bulls representing the extreme sub-populations for economically important traits were CGH analyzed using the same reference genome on a NimbleGen tiling array. We designed this array based on the latest build of the bovine genome (UMD3) with average probe spacing of 1150 bases (total number of probes was 2,166,672).

Two CNV gene clusters, *PLA2G2D* on BTA2 and *KIAA1683* on BTA7 revealed significant association with milk percentage and cow fertility, respectively, and were chosen for further characterization and verification in a larger sample using other methodologies including sequencing, tag SNPs and real time PCR (qPCR). Comparison between these four methods indicated that there is under estimation of the number of CNV loci in Holstein cattle and their complexity. The variation in sequence between different copies seemed to affect their functionality and thus the hybridization based methods were less informative than the methods that are based on sequencing. We thus conclude that large scale sequencing effort complemented by array CGH should be considered to better detect and characterize CNVs in order to effectively employ them in marker-assisted selection.

Achievements

The Holstein-Friesian breed is the world's highest-producing dairy cattle; much of its outstanding milk production was gained by selection of elite artificial insemination bulls based on breeding values that were estimated by progeny testing. Genomic selection is a newly developed tool for the estimation of breeding values through the use of genome-wide genotyping of single nucleotide polymorphisms (SNPs). The BovineSNP50 bead-chip provides fundamental tool for genomic selection for most of the genome as it was produced with inter-SNP interval of as close to 60 kb as possible and linkage disequilibrium (LD) in cattle disappears at a range of about 500 kb. Nevertheless, in the segmental duplication regions of the genome, where SNPs are under-represented, this tool is likely to fail. Moreover, copy number variation (CNV) has been recently identified in human and other mammalian genomes and increasing awareness that CNV might be a major source for heritable variation in complex traits has emerged. Despite this, little has been published on CNVs in Holsteins. In order to fill this knowledge-gap, we proposed a genome-wide association study between QTL for economic traits and CNV in the Holstein cattle. The approved feasibility study was aimed at the genome-wide characterization of CNVs in Holstein cattle and at the demonstrating of their possible association with economic traits by performing the activities of preparation of DNA samples, Comparative Genomic Hybridization (CGH), initial association study between CNVs and production traits and characterization of CNV-SNP associations.

For both countries, 40 genomic DNA samples of bulls representing the extreme sub-populations for economically important traits were CGH analyzed using the same reference genome on a NimbleGen tiling array. We designed this array based on the latest build of the bovine genome (UMD3) with average probe spacing of 1150 bases (total number of probes was 2,166,672). Results (see appendix, unpublished data) indicated that technically the CGH experiments were effectively performed. Comparison of CNVR distribution in the genomes of Holstein sires with high and low daughter pregnancy rate (DPR) evaluations pointed to notable variation in CNVR distributions in middle of chromosome 1, in the telomeric end of chromosome 2, and in the centromeric end of chromosome 7. These variations may indicate presence of CNVRs associated with this fertility related trait.

The CGH experiments as well as analysis of SNP data obtained from the BovineSNP50 bead-chip indicated that there is under estimation of the number of CNV loci in Holstein cattle. This prompted both the USA and Israeli sides to expand CNV discovery stage, to better describe cattle CNV repertoire and to improve the CNV discovery methodology. These achievements were detailed in two publications by Hou et al., and by Seroussi et al., submitted to the journals NAR and BMC bioinformatics, respectively (a copy of each paper is included in the appendix). It should be noted that two of the reviewers ranked the level of interest of the article by Seroussi et al., as an article of outstanding merit and interest in its field or as an article of importance in its field.

Following the indication for association of CNV with cow fertility in centromeric end, we further characterized the cluster of *KIAA1683* genes (see appendix, unpublished data; data was also presented in abstract number P4041, ISAG2010, Edinburgh, UK). In a sample of 326 bulls, allele specific qPCR for testing the association of the copy number with the fertility trait indicated that this association was significant ($p=0.0002$) only for a certain variation of the gene copies. Similarly, we could demonstrate association of the gene clusters, *PLA2G2D* on BTA2 with milk fat and protein percentages only for a certain variation of this phospholipase like gene (see appendix, unpublished data). Hence, comparison between these four methods (CGH, beadchip SNP analysis, qPCR and direct sequencing) showed that there is under estimation of the number of CNV loci in Holstein cattle and their complexity. The variation in sequence between different copies seemed to affect their functionality and thus the hybridization based methods were less informative than the methods that are based on sequencing. We thus conclude that large scale sequencing effort complemented by array CGH should be considered to better detect and characterize CNVs in order to effectively employ them in marker-assisted selection.

Details of cooperation

Within the feasibility study, we performed the activities of preparation of DNA samples, Comparative Genomic Hybridization (CGH), initial association study between CNVs and production traits and characterization of CNV-SNP associations. Most of the funding was dedicated to the CGH experiment, which was performed using NimbleGen chips. Since the USA side negotiated a reduced price for these chips

and since it was scientifically important that samples from both countries would be hybridized using the same reference and experiment setting, we decided that the USA side will be also responsible for the hybridization of the Israeli samples, which were sent to the USA. This also facilitated the bureaucracy associated with organizing this experiment and thus ensured that the results were produced in a timely fashion. Consequently, the relative portion of the USA budget was increased to 75% of the total budget. USA samples were sent to the Israeli side to be analyzed by sequencing and qPCR of specific loci. The results were made available for both side, exchange of opinions and the result interpretation was frequently conducted using e. mail correspondence. The project objectives were promoted as a result of the cooperation: the Israeli side gained excess to important DNA samples of USA bulls, which hold key information for Holstein genetics. The USA side gained access to the Israeli well organized herd book, which allowed testing of CNV association with production traits.

Project Title: Genome-Wide Association Study of Copy Number Variation and QTL for Economic Traits in Holstein Cattle

Principal Investigator (PI): Eyal Seroussi, ARO

Co-Principal Investigator (Co-PI): George Liu, USDA

Unpublished data briefly summarized

Forty array Comparative Genome Hybridization (CGH) experiments were performed as previously described (Selzer et al., 2005) using NimbleGen technology (<http://www.nimblegen.com>). Whole-genome CGH arrays contained 2,166,672 oligonucleotide probes that were designed and fabricated on a single slide to provide an evenly distributed coverage with an average interval of 1150 bases using UMD3 genome assembly. These types of arrays utilize synthetic probes 50-75 bp in length with similar melting temperatures and do not require sample amplification or reduced representation. Standard genomic DNA labeling (Cy3 for samples and Cy5 for references), hybridizations, array scanning, data qspline normalization, and segmentation were performed as described before (Selzer et al., 2005). In all experiments the reference DNA was from L1 Dominette, the female Hereford cow whose DNA was sequenced. As control (Fig 1C), one of the experiments was dedicated to analyze background levels by hybridizing this cow's DNA as the sample. The rest of the sample indicated the presence of above background level signals indicating the existence of multiple copy number variation regions (CNVR) that are typical to Holstein cattle tested (Figure 1A and 1B). The segment in which CNVRs were most abundant was the designated chromosome U of contigs with unknown genomic location. This enrichment is expected as the presence of CNVs is a factor known to complicate and to hamper the genomic assignment. CNVRs were also detected along the X chromosome whose base line was lower, with log2ratio around -0.5 for all bulls, which as males have only one X chromosome comparing to the two X chromosome copies of the female cow reference. This and the relatively flat baseline in the control hybridization indicate that technically the CGH experiments were effectively performed. To demonstrate the potential of this CNV data to explain variation in quantitative economical traits, we present a comparison of CNVR distribution in the genomes of Holstein sires with high and low daughter pregnancy rate (DPR) evaluations (Figure 1A and 1B, respectively). Since that in these plots each individual contributed only ~5% of the overall opacity, CNVRs that were unique are virtually unnoticed. Hence, notable variation in CNVR distributions between these panels (e.g. near the middle of chromosome 1, in the telomeric end of chromosome 2, and in centromeric end of chromosome 7) may indicate presence of CNVRs associated with this trait.

References

Selzer, R. R., T. A. Richmond, N. J. Pofahl, R. D. Green, P. S. Eis, P. Nair, A. R. Brothman, and R. L. Stallings. 2005. Analysis of chromosome breakpoints in neuroblastoma at sub-kilobase resolution using fine tiling oligonucleotide array CGH. *Genes Chromosomes. Cancer* 44:305-319.

Figure 1. Qualitative representation of CNVR distribution in genomes of Holstein bulls with high and low DPR evaluations.

Rainbow plots show the log2 ratio for all probes plotted versus genomic position for all chromosomes, with each chromosome differentiated by color and vertical lines. These plots are based on CGH-segMNT analyses that identify DNA copy number differences between test and reference genome (L1 Dominette) using a set of transformations that is applied to the signal data for 10 probes within a given window (~12,000 bp) and reduces the data to a single data point. Group plots (A and B) represent plots of multiple individuals superimposed using Adobe Photoshop, with each individual layer opacity ~5%. **A.** Superposition of 21 plots derived from analyses of sires with high DPR (>1.5). **B.** Superposition of 18 plots of sires with low DPR (<-1.4). **C.** Control plot showing the result of the hybridization for which both test and reference genomes were of L1 Dominette. Arrows point to chromosomal segments of notable variation in the distribution of CNVRs between panels A and B.

Association between a CNV in the bovine *KIAA1683* gene and fertility QTL on BTA7 in Israeli Holstein dairy cattle

GIORA GLICK^{1,2}, ANDREY SHIRAK¹, EYAL SEROUSSI¹, YOEL ZERON³, EFRAIM EZRA³, JOEL I. WELLER¹ and MICHA RON¹

¹ Agricultural Research Organization, The Volcani Center, Bet Dagan, ISRAEL

² The Hebrew University of Jerusalem, The Robert H. Smith Faculty of Agriculture, Rehovot, ISRAEL

³ Sion, AI Institute, Shikmim 79800, ISRAEL

Isolation of *KIAA1683* copy-variants.

Sequencing analysis enabled us to isolate and identify four different variants of the bovine *KIAA1683* exon3 (Figure 2).

Figure 2. Sequencing traces of the 4 different variants of the *KIAA1683* exon 3 region produced by the selective primers amplification. Arrows point to the specific nucleotides that differ among the gene variants.

Quantitative PCR (qPCR) for testing copy-number association with the fertility trait

Relative copy numbers based on allele specific qPCR of *KIAA1683* variants are shown in Figure 3. CNV was found in all gene variants tested. The qPCR results are supported by the genotyping results as 3258 which lacks V2 and 2283, 3241 and 3274 which lacks V1 showed only residual copy number of their lacking gene variation.

An association analysis between the V1 copy number and female fertility trait was conducted in a sample of 326 bulls with estimated breeding values for fertility. The regression of the bulls' gene copy number on fertility breeding values was significant ($p = 0.0002$, Figure 4.) Identical probability was obtained for the SNP NGS-58779 in the same bull sample. The correlation between the V1 copy number and the NGS-58779 was 0.564, which is also significant ($p < 0.001$).

The variant V2, also showed differential distribution among the bulls. The qPCR analysis of V2 using version specific primers revealed copy number variation among the bulls. However, no association was found between V2 and fertility in the analysis of the 326 bulls (Figure 4).

Figure 3. Using allele specific qPCR, we determined the normalized relative copy numbers of *KIAA1683* variants in 10 bulls that also were analyzed for the QTL segregation by the daughter design. Sample quantities were normalized using an autosomal non-CNV reference gene (*RPP30*). Colors of bars correspond to the following scheme: green: bulls which are homozygous for the NGS-58779 "A" allele and homozygous for the fertility QTL; red: bulls heterozygous for the NGS-58779 and heterozygous for the fertility QTL; blue- bars: Bulls which are homozygous for the NGS-58779 "G" allele and homozygous for the fertility QTL. **A.** Total copy number of all four *KIAA1683* variants. **B.** Copy number of the V1 variant. **C.** Copy number of the V2 variant.

Figure 4. Statistical regression analyses for association of genotypes from 326 bulls with the bulls' breeding values for daughter fertility: **A.** SNP NGS-58779 genotypes, **B.** V1 variant copy number. **C.** V2 variant copy number. $P = 0.0002$ for both regressions. NGS-58779 genotypes are 0 or 2 for homozygous and 1 for heterozygous.

Association between a CNV in the bovine *PLA2G2D* gene and milk concentration QTL on BTA2 in Israeli Holstein dairy cattle

MARINA GOLIK¹, EYAL SEROUSSI¹

¹ Agricultural Research Organization, The Volcani Center, Bet Dagan, ISRAEL

(These results were presented as part of the preliminary results for the original proposal)

Isolation of *PLA2G2D* copy-variants.

PLA2G2D gene variant were previously described (Golik et al., 2006).

Testing for association of copy number with milk production traits

We tested for association between one of the variants of this gene and milk production traits using single nucleotide polymorphism (SNP) in the third intron of *PLA2G2D4* by linkage disequilibrium (LD) analysis (Figure 5). Population wide LD was tested for 380 bulls with genetic evaluations for all the economical traits. Genotyping was performed by mass spectrometry.

Figure 5. Population-wide genotyping of *PLA2G2D4*-intron3 SNP using mass spectrometer.

Genotyping of SNP markers was performed using DNA MassArray technology. External and extension primers were designed using Sequenom's assay-design software using the *IPLEX* module. MALDI-TOF mass spectrometry analysis was performed using the MassArray genotype analyzer and two amplification steps. For the *PLA2* marker the primer pair for the first PCR was: acgttgatgAGCCATTCAGCTTCTAGCC and acgttgatgAAATGAGAAAGCCTGGGCTG. For the second PCR the extension primer was tcGCCCCGGGACTTGCTGCCTTTC. Lower case bases in primers indicate sequences that were added by the software to modify the mass and do not match the binding sites. Genotypes were assigned to 375 independent DNA samples extracted from semen of Holstein bulls that represent the Israeli population of artificial-insemination sires. The distribution of ratios between high and low mass heights was obtained using MassARRAY TYPER 4.0 genotyping software. This output for the genetic marker in *PLA2G2D4* (*PLA2*, panel A) was compare with that of a typical SNP marker (rs31255197, panel B)

The “C” allele was rare and specific to the *PLA2G2D4* variant. The analysis (Figure 5A) indicated that this SNP is polymorphic within the tested sire population. As presented in Figure 5B, in typical SNP (RS41255197) the calls fall within four groups: blue (T homozygotes) - near the horizontal axis and the 16.5° dashed guide line; green (C- homozygotes)- near the vertical axis and the 73.5° dashed guide line; yellow (CT heterozygotes)- near the 45° dashed guide line; and red (no call) mostly near the origin of the axes including no DNA controls, results with low signal and ambiguous results that are located between the above mentioned groups. The three groups, two of homozygotes and one of heterozygotes are distributed in accordance to the Hardy-Weinberg equilibrium (Fig 5B). However, the calls of the *PLA2G2D4*-intron3 SNP (Figure 5A) were differently distributed within these groups testify that genotyping was affected by more than one gene. Table 1 presents the expected high and low mass-height ratios in MALDI-TOF analysis. The results in figure 2A are in accordance with parallel detection of two gene copies.

Table 1. Expected high and low mass-height ratios in MALDI-TOF analysis

	One gene	Two genes	Three genes
Homozygous type A	2:0	2:2	2:4
Heterozygous	1:1	1:3	1:5
Homozygous type B	0:2	0:4	0:6

Association between *PLA2G2D4*-intron3 SNP and production traits.

The sire population was divided into two groups of genotypes GC (yellow) and G (blue) (Figure 5A). A substantial portion of the sample was excluded (red group) since the signal intensities were considered ambiguous by the detection algorithm. LD analysis following Cohen-Zinder et al., 2005 revealed the GC group was most significantly associated with increase in milk protein percentage (Table 2). Hence the gene copy carrying the rare “C” allele would have been beneficial if it was not associated with other negative traits. However this allele was also associated with decrease in milk production and tendency to increase the somatic cell count. This may be related to general influence on the health condition of the cows mediated by the *PLA2G2D* genes, which are thought to be involved in inflammation or by other unknown genes that display tight linkage to the analyzed gene variation.

Table 2. Effects of the polymorphism on the bulls' breeding values for the quantitative traits.

Trait	Milk (kg)	Fat (kg)	Protein (kg)	% Fat	% Protein	SCC*	Index PD07
GC group	-169	3	0.4	0.09	0.05	0.08	-15
Probability	0.0568	0.3574	0.8695	0.0134	0.0096	0.1909	0.8728

* Breeding value for somatic cell count, logarithmic scale.

References

- Cohen-Zinder, M., Seroussi, E., Larkin D.M., Loor, J.J., Everts-van der Wind, A., Heon-Lee, J., Drackley, J.K., Band, M.R., Hernandez, A.G., Shani, M., Lewin, H.A., Weller, J.I., and Ron, M., (2005) Identification of a missense mutation in the bovine ABCG2 gene with a major effect on the QTL on chromosome 6 affecting milk yield and composition in Holstein cattle. *Genome Research* **15**, 936-944.
- Golik, M., Cohen-Zinder, M., Loor, J.J., Drackley, J.K., Band, M.R., Lewin H.A., Weller, J.I., Ron, M., Seroussi E. (2006) Accelerated expansion of group IID-like phospholipase A2 genes in *Bos taurus*. *Genomics* **87**, 527-533.

Data submitted for publication

Two publications by Hou et al., and by Seroussi et al., were submitted to the journals NAR and BMC bioinformatics, respectively. These manuscripts include web appendixes, which are too large to print. We therefore present only the manuscript title pages, abstracts and one key figure from each of the manuscripts. Complete submission is available in the electronic format.

Analysis of copy loss and gain variations in Holstein cattle autosomes using BeadChip SNPs

**Eyal Seroussi^{1*}, Giora Glick¹, Andrey Shirak¹, Emanuel Yakobson¹, Joel I. Weller¹, Ephraim Ezra²,
Yoel Zeron³**

¹ Institute of Animal Sciences, ARO, The Volcani Center, Bet Dagan 50250, Israel

² Israel Cattle Breeders Association, Caesaria Industrial Park 38900, Israel

³ Sion, AI Institute, Shikmim 79800, Israel

*Corresponding author

ABSTRACT

Background: Copy number variation (CNV) has been recently identified in human and other mammalian genomes and there is a growing awareness of CNV's potential as a major source for heritable variation in complex traits. Genomic selection is a newly developed tool for the estimation of breeding values for quantitative traits through the use of genome-wide genotyping of SNPs. The Illumina BovineSNP50 BeadChip has been used worldwide to genotype over 30,000 Holstein cattle. On this chip, 54,001 SNPs are represented (~SNP/50,000 bp), and some of them fall within CNV regions. **Results:** We used the BeadChip data obtained for 912 Israeli bulls to investigate the effects of CNV on SNP calls. For each of the SNPs, we estimated the frequencies of occurrence of loss of heterozygosity (LOH) and of gain based either on deviation from the expected Hardy-Weinberg equilibrium (HWE) or on signal intensity (SI) using the PennCNV “detect” option. Correlations between LOH/CNV frequencies predicted by the two methods were low (up to $r = 0.08$). Nevertheless, 418 locations displayed significantly high frequencies by both methods. Efficiency of designating large genomic clusters of olfactory receptors as CNVs was 29%. Frequency values for copy loss were distinguishable in non-autosomal regions, indicating misplacement of a region in the current BTA7 map. Analysis of BTA18 placed important net merit QTLs in regions rich in segmental duplications and CNVs (Figure 6). Enrichment of transporters in CNV loci suggested their potential effect on milk-production traits. **Conclusions:** Expansion of HWE and PennCNV analyses allowed estimating LOH/CNV frequencies, and combining the two methods yielded more sensitive detection of inherited CNVs and better estimation of their possible effects on cattle genetics. Although this approach was more effective than methodologies previously applied in cattle, it has severe limitations. Thus the number of CNVs reported here for the Holstein breed may represent as little as one-tenth of inherited common structural variation.

Genomic and evolutionary characteristics of cattle copy number variations

Yali Hou^{1,2}, George E. Liu^{1*}, Maria Francesca Cardone³, Kai Wang⁴, Eui-soo Kim¹, Lakshmi K. Matukumalli^{1,5}, Mario Ventura³, Paul M. VanRadon⁶, Tad S. Sonstegard¹, and Curt P. Van Tassell¹

¹Bovine Functional Genomics Laboratory, ANRI, USDA-ARS, Beltsville, Maryland 20705, USA;

²Department of Animal and Avian Sciences, University of Maryland, College Park, Maryland 20742, USA;

³Department of Genetics and Microbiology, University of Bari, Bari 70126, Italy;

⁴Center for Applied Genomics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA;

⁵Bioinformatics and Computational Biology, George Mason University, Manassas, VA 20110, USA;

⁶Animal Improvement Programs Laboratory, ANRI, USDA-ARS, Beltsville, Maryland 20705, USA.

*Corresponding Author

ABSTRACT

Background: CNV represents another important source of genetic variation complementary to SNP data, which has been confirmed to have significantly functional effects on gene expression and human diseases. With the development of high-density SNP arrays, it has been routinely used for human CNV detection. In dairy industry, sufficient SNP genotyping results are becoming available and potentially used for CNV discovery to understand and accelerate genetic improvement for complex traits. **Results:** We performed a systematic analysis of CNV using the Bovine HapMap SNP genotyping data, including 539 animals of 21 modern cattle breeds and 6 outgroups. After correcting genomic waves and considering the trio information, we identified 682 candidate CNV regions (CNVRs), which cover 139.8 megabases (~4.60%) of the genome (Figure 2). Selected CNVs were further validated by FISH or qPCR, and we found that duplication CNVs are predominantly tandem clustered repeats rather than interspersed duplications. Approximately 56% of the CNVRs overlap with 1,263 cattle genes, which are significantly enriched for immunity, lactation, reproduction and rumination. We compared this dataset with other published CNVRs datasets and the overlap was usually <40%; while considering large CNVRs with frequency>5%, the overlap increased to 90%, highlighting the differences and commonalities between technical platforms. A cluster analysis based on the copy number distributions of 3,456 polymorphic SNPs from 15 representative breeds showed that Taurine, Indicine and African breeds were clearly separated, and the distribution of Taurine breeds also reflected their known historical and geographical origins. **Conclusions:** We presented a comprehensive genomic and evolutionary analysis of cattle CNVs by SNP arrays, provided a valuable genomic variation resource, supported three cattle domesticated origins in Fertile Crescent, Indus Valley and Africa, respectively. Compared to SNPs, gene containing CNVRs may help identify genes undergoing artificial selection (such as for milk and meat production) in domesticated animals.

Figure 6. Marker effect on net merit and CNV on BTA18. (A) The largest effect of all the BeadChip SNPs on production traits important for net merit was previously mapped to BTA18 and marker effects for net merit within this chromosome were previously calculated (Cole et al., 2009). (B) Segmental duplications (Liu et al., 2010) were annotated in green. Common CNVs (frequency > 2.5%) previously detected using genomic and CGH analyses (Liu et al., 2009; Bae et al., 2010; Fadista et al., 2010) were marked by letters “L”, “B” and “F”, respectively. (C) LOH and copy gain frequencies were calculated using the trinomial expansion of HWE and 912 BeadChip samples. For each marker, only positive values for copy loss (rl) and gain (rg) are presented. (D) Loss and gain frequencies (lsi and Gsi) were also calculated based on the PennCNV analysis for these samples.

Figure 7. Cattle copy number variation and segmental duplication regions display a local tandem distribution pattern. CNV regions (682 events, 139 Mb, ~4.60 % of the bovine genome) reported by 521 SNP genotyped individuals are shown above the chromosomes in green (gain), red (loss) and dark blue (both), while below are the CNV regions (177 events, 28 Mb, ~1% of the bovine genome) reported by 90 array CGH experiments by Liu et al. The bar height represents their frequencies: short (appeared in 1 sample), median (≥ 2 samples) and tall (≥ 5 samples). Segmental duplications (94.4 Mb, 3.1% of the bovine genome) predicted by two independent computational approaches are illustrated on the chromosomes in red (WSSD), blue (WGAC) or purple (both). The patterns are depicted for all duplications for ≥ 5 kb in length and $\geq 90\%$ sequence identity. The gaps in the assembly are represented on the chromosomes as white ticks.