# Analysis of copy number variants in Spanish authoctonous beef cattle breeds

**T.B R. da Silva[1,2], A. González-Rodríguez[3], C. Avilés[4], E. Mouresan[3], J.J. Cañas-Álvarez[5], L. Varona[3], M.J. Carabaño[2], P. Martínez-Camblor[6], C. Díaz [2]**

[1]UNESP-Jaboticabal, Brazil, [2]INIA, Spain ,[3]Universidad de Zaragoza, Spain, [4]Universidad de Córdoba, Spain, [5]Universitat Autònoma de Barcelona, Spain, [6]Universidad de Valladolid, Spain

**ABSTRACT:** Copy number variants (CNV) are an important source of genetic and phenotypic information. They have been associated with diseases and production traits in cattle. In this work, the objective is to identify CNVs and CNV regions (CNVR) in five Autochtonous Spanish beef cattle breeds accounting 366 individuals in trios genotyped with BovineHD Beadchip. From those, we extracted the corresponding subset relative to 50K SNP array. After detection and quality filtering been applied, we identified 3,965 CNVRs in the HD array and 201 with the 50K. The total length of CNVRs covered 8.6% and 1.2% of the autosomal cattle genome, respectively. Was found segregating CNV present in more than one breed varying from 33 to 268 for the HD information and from 0 and 10 for the 50K. The interest and applicability of these results needs further investigation.

**Keywords:** beef cattle, genomic, copy number variants, single nucleotide polymorphism

## Introduction

A copy-number variant (CNV) is a segment of DNA that is 1 kb or larger and is present at a variable copy number in comparison with a reference genome. Classes of CNVs include insertions, deletions and duplications, being deletions more frequent than duplications (Fu et al. (2010); Bickhart et al. (2012)). This definition also includes large-scale copy number variants, which are variants that involve segments of DNA ≥50 kb, allowing them to be detected by clone-based array comparative genome hybridization (array-CGH). Furthermore, a CNV region (CNVR) can be defined as a region that comprises more than one CNV event (Feuk et al. (2006); Butler et al. (2012)).

CNVs are part of the genomic structural variation. They are involved in changes in gene expression by the deletions or duplications, normal phenotypic variation, disrupt and gene dosage alterations. In this regard they have been identified as cause of diseases or the increase of the susceptibility of complex diseases (Redon et al. (2006); Meyers et al. (2010); Flisikowski et al. (2010)), as well as associated to production traits (Seroussi et al. (2010)). Thus, CNV are an important source of genetic and phenotypic variation, which highlights the role of CNVs in genetic diversity and evolution.

Autochtonous Spanish breeds have an important role in the beef production in Spain. One strategic approach for the implementation of genomic selection in these populations is to consider a metapopulation as a reference population (Mouresan et al. (2014)). In that respect, knowledge of similarities or differences in causal elements of genetic variation could be of potential interest. Thus, the objective of this paper is to identify and describe the genome-wide characteristic of CNVs and CNVRs in five Spanish autochthonous breeds using bovine HD SNP (777K) arrays and compare results with the 50K SNP arrays

**Materials and Methods** A total of 366 individual from five Spanish beef cattle populations were used in the analysis. Animals belonged to Asturiana de los Valles (AV, 75 individuals; 25 trios), Avileña-Negra Ibérica (ANI, 74 individuals; 24 full trios), Bruna dels Pirineus (BP, 75 individuals; 25 trios), Pirenaica (Pi, 74 individuals; 24 full trios) and Retinta (Re, 68 individuals; 18 distinct trios) breeds. Trios were chosen to capture the existing variability in the population.

Samples were genotyped with the Illumina Bovine beadchip High-density with 777,962 SNPs SNP markers. SNPs kept for the study belonged to autosomal chromosomes and those at repeated positions were excluded. Additional requirements were Mendelian error rate lower than 0.05 and genotyping rates over 95% for both, individuals and SNPs. The quality control was made using PLINK software (Purcell et al. (2007)). Finally, 706,978 SNPs were retained, covering 2,510,395 kb that represent an average density of one marker every 3.551 kb. To study the behavior of the 50K chip, the markers present in that chip were extracted from the HD. A total of 49,345 SNP markers were in common.

To identify CNV in these samples PennCNV was used (Wang, et al. (2007)). PennCNV considers information of Log R ratio (LRR), the frequency of B allele (BAF) at each SNP marker, the distance between neighboring markers, the population frequency of B allele (PFB, and the pedigree information when available, into a hidden Markov model (HMM). Both LRR and BAF were exported from Illumina high-density bovine beadchip. The PFB file was calculated based on the BAF of each marker in the population.

We used the -trio option of PennCNV which considers pedigree information after CNV calling and increases the CNV call rates as compared to –test option which does not consider the relationship between samples (Hou, et al. (2011)). After detection of CNVs, filtering of low quality samples was carried out with the following cutoffs: standard deviation (STD) of LRR < 0.30, BAF drift < 0.01 and waviness factor of LRR < 0.05. The final CNVs set was the set of non-redundant CNVs within breeds.

## Results and Discussion

In the analysis with the HD SNP array, a total of 9,839 CNVs were detected. Thus, an average of a 25.6

events was obtained for each sample (Table 1) in the whole dataset when all five breeds we put together. The range of number of CNVs across breeds varied from 3,005 in BP to 1,827 in Pi, which corresponds with a range of average CNVs count by sample within breed between 40.06 and 24.69, respectively.

**Table 1. CNV events and comparison within whole dataset and breeds in the 777K SNP array**

| Breed | CNV | | | | | |
|---|---|---|---|---|---|---|
| | N | Count | Unique | Gain | Loss | Total length (kb) |
| AV | 75 | 1979 (26.38) | 929 (12.38) | 753 (10.04) | 1,226 (16.34) | 124,720.228 (63.02) |
| ANI | 74 | 2506 (33.86) | 1296 (17.51) | 584 (7.89) | 1,922 (25.97) | 141,106.86 (56.30) |
| BP | 75 | 3005 (40.06) | 2616 (34.88) | 300 (4) | 2,705 (36.06) | 207,589.24 (69.08) |
| Pi | 74 | 1827 (24.69) | 820 (11.08) | 635 (8.58) | 1,192 (16.10) | 99,392.23 (54.40) |
| Re | 68 | 2569 (37.78) | 1508 (22.17) | 748 (11) | 1,821 (26.78) | 131,431.012 (51.16) |
| All | 366 | 9389 (25.65) | 5592 (15.27) | 2416 (6.60) | 6,973 (19.05) | 564,259.86 (60.09) |

The numbers in parentheses are the average by sample counts except that the lengths in parentheses are average lengths normalized by CNV counts.

CNVRs were generated by aggregating overlapping CNVs identified across all samples and within samples. A total of 3,965 autosomal CNVRs (Table 2) were identified in the whole dataset, spanning 218.52 Mb of polymorphic sequence. This corresponds to 8.6% of the autosomal genome sequence (218.52/2,545.9Mb) and 7.2% of the whole cattle genome (218.52/3,036.6 Mb). The numbers of CNVRs were 1034, 1272, 1592, 945 and 1426 for AV, ANI, BP, Pi, Re, respectively. As observed in other populations, the number of gains (insertions) was lower than the number of losses (deletions). Thus, the 3,965 CNVRs included 1,173 gain events, 2,987 loss, and 195 both (loss and gain within the same region), ranging from 1.015 to 4345.36 Kb with mean and median of 55.1 and 26.951 Kb, respectively. CNVRs included 412 gain events, 698 loss and 76 both for AV, 310 gain, 998 loss and 36 both events for ANI, 229, 1402 and 39 (gain, loss and both) for BP, 359, 643 and 57 (gain, loss and both) for Pi and 473 gain, 1014 loss, and 61 both events for Re breed. The mean and median length for the CNVRs for the five breeds were 46.38 and 17.9; 47.18 and 21.76; 78.87 and 44.72; 43.04 and 16.6; 42.76 and 22.28. Furthermore, 1,531 CNVRs were found in only one sample (Unique), 824 CNVRs were present in 2 animals, and 1610 CNVRs were shared by 3 or more animals.

A relevant number of unique CNV and CNVR were found in the total number of samples and in each population (Tables 1 and 2). However, 285, 267, 321, 250 and 290 CNVs were present in at least 2 animals in AV, ANI, BP, Pi and Re, respectively and 388, 518, 410, 393 and 498 were shared by 3 or more animals in each breed. These results confirm that segregating CNVs exist among these 366 samples within population. In Table 2 we show the number of segregating CNVs. We found 980 CNVs segregating in AV breed, through the 25 trios, which means that these CNVs were present at least in one of the parents and in the offspring. For the other breeds, the corresponding numbers are 1,168 for ANI breed, 221 in BP, 968 in Pi and 916 CNVs segregating in Re. The common segregating CNVs are shown in Table2. One interesting results is that some of these segregating CNVs are shared among breeds. The values ranged from 33 between BP and Re to 268 between AV and ANI.

We compared the 3,965 CNV regions found with the literature (Liu et al. (2010); Hou et al. (2011); Hou et al. (2012); Bickhart et al. (2012)) and we got 71, 318, 193 and 205 CNVRs overlapped, respectively, which means, approximately 1.8%; 8%; 4.8% and 5.2% of our CNV calls.

When we extract the information relative to 50K from the HD chip the number of CNVs and CNVRs decreased (Table 3 and Table 4), as expected. However, when we obtained that information, 5,264 SNP markers present in the current 50K Illumina chip were missing in the HD chip. To what extent the results are influenced by such event is under investigation. Thus, we found 309 CNVs considering the whole dataset (366 samples), with an average of 0.84 CNVs event by sample (Table 3). The range across breeds was from 109 with an average of 1.47 per sample in ANI to 15 CNVs, 0.2 per sample in BP (Table 3). As well as in the HD, segregating CNVs were found in all of the five breeds, being 25, 37, 4, 22 and 25 for AV, ANI, BP, Pi and Re, respectively. We found common CNVs ranging from 0 to 10, between AV-BP and AV-ANI, respectively (Table 5).

**Table 2. CNVR events and comparison within whole dataset and breeds**

| Breed | CNVR | | | | | |
|---|---|---|---|---|---|---|
| | N | Count | Unique | Gain | Loss | Total Length(kb) |
| AV | 75 | 1034 | 361 | 412 | 698 | 47,959.904(46.38) |
| ANI | 74 | 1272 | 457 | 310 | 998 | 60,013.626(47.18) |
| BP | 75 | 1592 | 861 | 229 | 1402 | 125,646.993(78.92) |
| Pi | 74 | 945 | 302 | 359 | 643 | 40,675.635(43.04) |
| Re | 68 | 1426 | 638 | 473 | 1014 | 60,982.128(42.76) |
| All | 366 | 3965 | 1531 | 1173 | 2987 | 218,528.381(55.11) |

The numbers in parentheses are the average by sample counts except that the lengths in parentheses are average lengths normalized by CNVR counts.

A total of 201 CNVRs were found in the whole dataset with a total length of 32.61 Mb corresponding to 1.2% of the autosomal genome sequence. 61, 74, 14, 53 and 88 CNVRs were found for the AV, ANI, BR, Pi and Re breeds, respectively. As expected, the 50k array provided larger CNVs when compared to the HD array, with the size of the CNVRs ranging from 40.755 kb to 965.371 kb with a mean and median size of 149.261 and 121.085 kb respectively. The CNVRs from whole dataset included 77 gain, 140 loss and 16 both events. Considering the breeds separately, AV had 23 gain, 38 loss and none both events; ANI – 17 gain, 58 loss and 1 CNVR with loss and gain event simultaneously; for BP breed, we found 4 gain event and 10 loss and no both events; for Pi and Re breeds we found 12 gain, 51 loss, no both event for the first and 35, 57 and 4, gain, loss and both events for the latter. Once more assenting with the literature, there were more loss events than gain events (Hou et al. (2011)).

Compared with the four earlier cited studies, approximately 2.5%, 6.4%, 34.32% and 30.34% overlapped with Liu et al. (2010), Bickhart et al. (2012), Hou et al. (2011), and Hou et al. (2012), respectively.

When we compared the results of the high-density and the 50k SNP array, we detected 127 CNVRs overlapping.

## Conclusion

Our analyses have identified a total of 3,965 and 201CNVRs with information from the HD and 50K arrays from Illumina, respectively. These regions covered a total of 218,528 and 32,610 Mb (8.58 and 1.2% of the autosomal genome) for the HD and 50k chips, respectively. Some of these CNVs showed Mendelian inheritance. Common CNVs were found across the local Spanish breeds. The interest and applicability of these results needs further investigation.


## Acknowledgements

The financial support of Spanish AGL2010-15903 and UE FP7-289592-GENE2FARM grants made possible this research. T.B.R. da Silva thanks Capes-Brazil for making available the Sandwich scholarship. A. González-Rodríguez and J. J. Cañas-Álvarez acknowledges the financial support given by the fellowship BES-2011-045434 and COLCIENCIAS Francisco José de Caldas 497/2009 fellowships.


## Literature Cited


Bickhart, D. M., Hou, Y., Schroeder, S. G., et al. (2012). *Genom. Res.*, 22:778-790.

Butler, J., Locke, M. E. O., Hill, K. A., et al. (2012). *Bioinf. Adv. Acc.*, 29 (2):262-263.

Feuk, L., Carson, A. R., and Scherer, S. W. (2006). *Nat. Re. Genet.*, 7:85-97.

Flisikowski, K, Venhoranta, H., Nowacka-Woszuk J., et al. (2010). *PLoS One*, 5(11):e15116.

Fu, W., Zhang, F., Wang, Y., et al. (2010). *Am. J. Hum. Genet.*, 87:494–504.

Hou, Y., Liu, G. E., Bickhart, D. M., et al. (2011). *BMC Genomics*, 12:127.

Hou, Y., Liu, G. E., Bickhart, D. M., et al. (2012). *Funct. Integr. Genomics*, 12:81–92.

Liu, G. E., Zhu, B., Cardone, M. F., et al. (2010). *Genome Res.*, 20:693-703.

Meyers, S. N., McDaneld, T. G., Swist, S. L., et al. (2010). *BMC Genomics*, 11:337.

Mouresan, E. F., Munilla, S., Díaz, C., et al. (2014). 10th WCGALP (Submitted)

Redon, R., Ishikawa, S., Fitch, K. R., et al. (2006). *Nature*, 23: 444–454.

Seroussi, E., Glick, G., Shirak, A., et al. (2010). *BMC Genomics*, 11:673.

Wang, K., Li, M., Hadley, D., et al. (2007). *Genome Res.*, 17:1665-1674.


**Table 3. CNV events and comparison within whole dataset and breeds in the 50K SNP array**

| Breed | CNV | | | | | |
|---|---|---|---|---|---|---|
| | N | Count | Unique | Gain | Loss | Total Length (kb) |
| AV | 75 | 72(0.96) | 45(0.6) | 27(0.36) | 45(0.6) | 11,374.32(157.97) |
| ANI | 74 | 109(1.47) | 67(0.9) | 20(0.27) | 89(1.20) | 17,506.02(160.60) |
| BP | 75 | 15(0.2) | 11(0.14) | 4(0.05) | 11(0.14) | 2,302.42(153.49) |
| Pi | 74 | 68(0.92) | 40(0.54) | 13(0.17) | 55(0.74) | 9,418.59(138.50) |
| Re | 68 | 108(1.58) | 81(1.19) | 37(0.544) | 71(1.04) | 15,116.42(139.96) |
| All | 366 | 309(0.84) | 195(0.53) | 91(0.24) | 218(0.60) | 47,780.71(154.63) |

The numbers in parentheses are the average by sample counts except that the lengths in parentheses are average lengths normalized by CNV counts

**Table 4. CNVR events and comparison within whole dataset and breeds in the 50K SNP array**

| Breed | CNVR | | | | | |
|---|---|---|---|---|---|---|
| | N | Count | Unique | Gain | Loss | Total Length(kb) |
| AV | 75 | 61 | 37 | 23 | 38 | 9254.237(151.70) |
| ANI | 74 | 74 | 35 | 17 | 58 | 12825.696(173.32) |
| BP | 75 | 14 | 11 | 4 | 10 | 2302.426(164.46) |
| Pi | 74 | 53 | 24 | 12 | 41 | 7625.902(143.88) |
| Re | 68 | 88 | 59 | 35 | 57 | 12890.418(146.48) |
| All | 366 | 201 | 100 | 77 | 140 | 32610.934(162.24) |

The numbers in parentheses are the average by sample counts except that the lengths in parentheses are average lengths normalized by CNVR counts.

**Table 5. Common segregating CNVs among breeds - Above diagonal CNVs common among breeds with CNV calling in the high-density SNP array and below diagonal with the 50k SNP Array. The values in diagonal represent the segregating CNV count for each breed with the HD SNP chip (the values in parentheses are the accounting of CNV with the 50k SNP array)**

| | AV | ANI | BP | Pi | Ret |
|---|---|---|---|---|---|
| AV | 980(25) | 268 | 49 | 258 | 174 |
| ANI | 10 | 1075(33) | 54 | 241 | 87 |
| BP | 0 | 0 | 203(4) | 34 | 33 |
| Pi | 0 | 3 | 1 | 876(22) | 152 |
| Re | 1 | 4 | 1 | 0 | 819(23) |