GENOME-WIDE ASSOCIATION STUDY FOR MILK SOMATIC CELL SCORE IN HOLSTEIN CATTLE USING COPY NUMBER VARIATION AS MARKERS

Marina Durán Aguilar [1], Sergio I. Román Ponce [2], Felipe J. Ruiz López[2], Everardo González Padilla[3], Carlos G. Vásquez Peláez[3], Alessandro Bagnato[4], Maria Giuseppina Strillacci[4]

[1] Facultad de Estudios Superiores Cuautitlán. UNAM. Ave. 1o de Mayo S/N, Santa Maria las Torres, 54740 Cuautitlán Izcalli, México.

[2] Centro Nacional de Investigación en Fisiología y Mejoramiento Animal. INIFAP. Km. 1 Carretera a Colón, Auchitlán, Querétaro, México. CP 76280

[3] Departamento de Genética y Bioestadística, Facultad Medicina Veterinaria y Zootecnia, Universidad Nacional Autónoma de México, Av. Universidad 300, México DF, 04510

[4] Department of Health, Animal Science and Food Safety (VESPA), University of Milan, Via Celoria 10, 20133, Milan, Italy.

Corresponding author: Alessandro Bagnato, Università degli Studi di Milano, Via Celoria 10, 20133 Milano, Italy. Phone +39-0250315740. e-mail: alessandro.bagnato@unimi.it

Keywords: Holstein, copy number variants, GWAS, SNP, somatic cell score

Summary

Mastitis, the most common and expensive disease in dairy cows, implies significant losses in the dairy industry worldwide. Many efforts have been made to improve genetic mastitis resistance in dairy populations, but low heritability of this trait made this process not as effective as desired. The purpose of this study was to identify genomic regions explaining genetic variation of somatic cell count using copy number variations (CNVs) as markers in the Holstein population, genotyped with the Illumina BovineSNP777HD array. We found 24 and 47 copy number variation regions significantly associated with estimated breeding values for somatic cell score (SCS_EBVs) using SVS 8.3.1 and PennCNV-CNVRuler software's, respectively. The association analysis performed with these two software allowed the identification of 18 candidate genes (*TERT, NOTCH1, SLC6A3, CLPTM1L, PPARα, BCL-2, ABO, VAV2, CACNA1S, TRAF2, RELA, ELF3, DBH, CDK5, NF2, FASN, EWSR1,* and *MAP3K11*) that result classified in the same functional cluster. These genes are also part of two gene-networks, whose genes share the ‘‘stress’’, “cell death”, ‘‘inflammation’’ and “immune response” GO terms. Combining CNVs detection/association analysis based on two different algorithms helps towards a more complete identification of genes linked to phenotypic variation of the somatic cell count.

Introduction

The most common and expensive disease in dairy cows affecting the mammary gland is the mastitis, an inflammation caused by pathogens. Because of the increased milk production and veterinary treatments, clinical mastitis cases can cost up to $200 per case, with a total estimated cost to the U.S. dairy industry of approximately $1.7–2 million dollars/year (Cha *et al.,* 2011).

The susceptibility of the bovine mammary gland to mastitis largely depends on the involution process of the mammary gland tissue and on the exposure to different physiological, genetic and environmental factors (Sordillo and Streicher, 2002).

Many efforts have been made to improve the genetic immune resistance to mastitis in dairy cows. The results are still very limited and obtained mainly through correlated traits as somatic cells count (SCC). The milk SCC, in fact, can be used as a predictor of mastitis susceptibility, since a moderate to high genetic correlations have been reported between clinical mastitis and SCC or its log transformation in somatic cell score (SCS) (Hinrichs *et al.,* 2005).

In the recent past, a large number of studies have mapped QTL affecting mastitis SCC and SCS and are reported in the QTL database (http://www.animalgenome.org/cgi-bin/QTLdb/BT/index).

The availability of dense Single Nucleotide Polymorphism (SNP) arrays facilitates the identification of genomic regions associated with economically important traits in farm animals, thus allowing to better disclose QTLs and genetic variation for mastitis resistance. Recently a class of structural variants, the copy number variants (CNVs), have been suggested as markers of genomic variation in complex disease (Redon *et al.,* 2006). CNVs, in fact, are a genomic structural variation that, as SNP, is considered as an important marker of heritable genetic expression (Kijas *et al.,* 2011).

CNVs are distributed over the whole genome in humans, domestic animals and other species; they are defined as large-scale genome mutations ranging from 50bp to several Mb compared with a reference genome, which are presented as insertions, deletions and more complex changes (Mills *et al.,* 2011). Although SNPs are more frequent, CNVs involve larger genomic regions that may affect gene structure and possibly determining a change in its expression and regulation (Hou *et al.,* 2012a).

Several studies have shown that the CNVs are associated with residual feed intake variability in Holstein cows (Hou *et al.,* 2012b) and with fertility in Israeli Holsteins (Glick *et al.,* 2011). In addition, Xu *et al.,* (2014) reported a genome wide CNVs association analysis with milk production traits in Holstein, identifying thirty-four CNVs significantly associated with milk production traits, most of them overlapping known QTL.

Generally, QTL identifying studies are based on a very large number of individuals, as sample size is determinant to achieve reasonable power in a population wide experimental design. The selective genotyping (Darvasi, 1997) is an efficient method to identify chromosomal regions that harbor QTL by comparing marker allele frequencies from phenotypically extreme samples (samples that deviate the most from the mean of the phenotype). This approach allows maintaining the same statistical power for QTL detection, limiting the number of samples to genotype to those in the extreme high and low values, instead of genotyping the whole population as is done in population wide experimental designs. Several studies based on SNP markers, have addressed the feasibility and effectiveness of the selective genotyping method (also combined with the DNA pooling approach), to detect QTL associated with different traits (Strillacci *et al.,* 2014; Fontanesi *et al.,* 2007).

The use of CNVs as markers, to explain the genetic variation of SCS in milk, has not been explored so far. The purpose of this study was to identify genomic regions explaining the genetic variation of SCS using CNVs in the Holstein population using a selective genotyping approach. Two different algorithms were used to call CNVs in order to provide a cross integration and validation of results and a clearer indication on the size of the CNVs identified.

Materials and Methods

Sampling and genotyping

The SCS estimated breeding values (SCS_EBV) were obtained from the Mexican Holstein Association (http://www.holstein.com.mx/QueToro.aspx). In order to identify individuals with extreme high and low values, the entire database was ranked based on SCS_EBV values (1.38 mean ± 1.14 SD).

A total of 242 samples with available DNA, identified among individuals above and below 2 SD from the SCS_EBV values average, were selected and classified as following: i) high phenotypes. A total of 102 samples with SCS_EBV mean 3.37 ± 0.523; ii) low phenotype. A total of 140 samples with mean 1.67 ± 0.719.

SNP chip data, obtained from the Illumina BovineSNP777HD array (Illumina Inc., San Diego, CA), were provided by the Genomic Improvement project of INIFAP and the Mexican Holstein Association.

Data editing

The Log R Ratio (LRR) and the B allele frequency (BAF) values were extracted using the Illumina BeadStudio software V.2.0 (Illumina Inc.). Samples with a call rate below 98% were excluded from the subsequent analyses, which were performed for the 29 autosomes.

The overall distribution of derivative log ratio spread (DLRS) values were evaluated using the SVS 8.3.1 software (Golden Helix Inc.) to identify and filter outlier samples, as described by Pinto *et al.,* (2011). To normalize the LRR values and then exclude the samples with extreme wave factors from the analysis, we used the wave correction algorithm, which corrects for the waviness contributed by GC content. In addition, batch effects in the LRR were corrected via numeric Principal Component Analysis (PCA).

CNVs detection

As suggested by different authors (Pinto *et al.,* 2011) and applied in CNVs mapping studies (Bagnato *et al.,* 2015), because of intrinsic noisiness of CNV analysis, at least two algorithms should be used for the identification of CNVs. This strategy allows the integration and comparison of the CNV detection among different algorithms and may reduce the bias in detection (i.e. false negatives) proper of each algorithm. The possibility to identify false negatives may be relevant especially when running an association analysis between identified CNVs and traits of interest.

Two independent software based on different algorithms were here used to identify CNVs: i) the Copy Number Analysis Module (CNAM) provided by SVS 8.3.1 software (http://goldenhelix.com); ii) the Hidden Markov Model (HMM) by PennCNV software (http://penncnv.openbioinformatics.org/en/latest/).

For CNVs calling using SVS 8.3.1 software, LRR values were employed under the univariate approach: this approach segments each sample independently. As suggested in the software manual, the options used were the following: i) univariate outlier removal; ii) maximum number of segments: search for up to 10 per 10,000 markers; iii) a minimum of 3 markers per segment; iv) a significance level of p= 0.005 for pairwise permutations (n=2000). After segmentation analysis, all the segments were classified in three categories as losses, gains and neutral.

The individual-based CNV calling, based on LRR and BAF values for every SNP, was performed by PennCNV software using the default parameters of HMM (standard deviation (SD) of LRR <0.30 and BAF drift as 0.01).

CNV association with SCS_EBV

Linear regression in SVS 8.3.1 software was used to identify CNVs (detected by CNAM algorithm) associated with SCS_EBV with significance level of FDR >0.05, after classifying CNV calls in three state covariates, i.e. loss, neutral, gain (-1, 0, 1).

Instead, results of the CNV calling from PennCNV were utilized to perform an association analysis with SCS_EBV using the CNVRuler software (http://www.ircgp.com/CNVRuler/index.html), after the definition of CNV regions (CNVRs). In this study, the CNVRs were detected by merging overlapping CNVs by at least 1bp identified across all samples, as described by Redon *et al.,* (2006). The association analysis was performed between CNVRs and SCS_EBV, applying a linear regression model, with a minor allele frequency threshold value set to 0.02. The parameter of recurrence, set to 0.1 (default value), was applied to allow a more robust definition of regions. This option checks the density of regions of CNVs and trim the sparse area not satisfying the density threshold of 10%.

Additionally, the "Gain/Loss separated regions" option, which compiles the region based on the genotype (gain or loss of copy number), was applied.

Significant CNVs were detected when their false discovery rate adjusted p-values (FDR) had a value of $p < 0.05$.

For a graphical visualization of the results, two separated Manhattan plots of associated CNVRs with SCS_EBV were created using the -log10 of the p-values resulting from the association analyses performed by SVS 8.3.1 and PennCNV-CNVRuler software.

Annotation

The full Ensembl v83 gene set (bovine UMD 3.1 assembly) for the autosomes was downloaded (http://www.ensembl.org/biomart/martview/76d1cab099658c68bde77f7daf55117e/).

In order to identify the genes located within the CNVRs we created a consensus list (among CNVRs and the downloaded genes) using the BedTools software (Quinlan and Hall, 2010).

Gene Ontology (GO) and pathways analyses were performed using GenCLiP2.0, an online server for functional clustering of genes (http://ci.smu.edu.cn/GenCLiP2.0/analysis.php?random=new) accounting for false discovery rate.

RESULTS AND DISCUSSION

CNVs calling and association analysis with SVS 8.3.1

The CNVs detection was performed in 220 stringently quality filtered samples: i) 88 high phenotype samples (3.384 ± 0.511) and 132 low phenotype samples (1.670 ± 0.726).

We identified a total of 5194 CNVs (covered at least by 3 SNPs) (Table 1) distributed on all autosomes, mainly on the BTA 12 (n=881). Among the detected CNVs, the number of losses and gains were 5,088 (98%) and 106 (2%) gains, respectively. The number of CNVs in all samples ranges from 11 to 42 (average of 25.12).

Overlapping CNVs across samples were summarized at the population level into 252 CNVRs (11 gains, 236 losses and 5 complex), with 62 singletons and 128 CNVRs that comprise at least 5 CNVs

(Table S1). CNVRs cover a total of 39.29 Mb of sequence which corresponds to 1.5% of the Bovine UMD3.1 assembly.

Using a linear regression, 85 CNVs resulted to be associated with SCS_EBV (p-value <0.05 after FDR correction) (Figure 1A).

In order to better delineate the chromosomal regions resulting associated with the trait, the significant CNVs are grouped in 34 CNVR distributed on 17 autosomes, according to Redon *et al.*, (2006)'s approach, using the BedTools software. Among those CNVRs, only the ones with CNVs frequencies above 2% (CNVRs with at least one CNV identified in five samples) were retained and used to perform the annotation analysis. Based on UMB3.1 sequence assembly, 51 bovine genes were annotated within the significant CNVRs (Table 2).

CNVs calling and association analysis with PennCNV

The use of the "Filtering CNV calls by user-specified criteria" module of PennCNV allowed to identify low-quality samples and to eliminate them from further analysis. Out of the 220 stringently quality filtered samples we than obtained a subset of samples (n=124) with a maximum number of CNVs equal to 200: i) 49 high phenotype samples (3.307 ± 0.385); ii) 74 low phenotype samples (1.830 ± 0.080). This additional filtering is specifically required according to the PennCNV detection algorithm.

Overall, 12,070 CNVs distributed on all autosomes were then assessed, with an average per sample of 97.33 CNVs (ranging from 42 to 200) (Table 1).

CNVs overlapping by at least one nucleotide were summarized to 1,662 CNVRs (394 gains, 1,215 losses and 53 complex), with 844 singletons and 408 CNVRs that comprise at least 5 CNVs (Table S2). The defined CNVRs cover 82.67 Mb of autosomal genome sequences, corresponding to 3.3% of the Bovine UMD3.1 assembly.

After PennCNV-CNVRuler analysis, a total of 47 CNVRs distributed on 18 autosomes, were associated (p-value<0.05 after FDR correction) with SCS_EBV (Figure 1B). Table 3 reports the list of the 47 significant CNVRs and the 105 annotated genes.

Comparison of results obtained with SVS8.3.1 and PennCNV software

In order to identify the CNVRs that fully overlapped each other among those identified within the two software, the Wain *et al*. (2009)'s approach was used in a BedTool software routine. The consensus CNVR set contained 265 regions.

After association analysis, only six CNVRs resulted associated with SCS_EBV for both analyses. These common regions were located on BTA1 (at 93.95 Mb), on BTA5 (at 58.96 Mb), on BTA5 (at 117.28 Mb), on BTA7 (at 42.73 Mb), on BTA12 (at 74.84 Mb) and on BTA23 (at 28.82 Mb).

The CNVR_11SVS on BTA12 comprised two different associated regions identified by CNVRuler (CNVR_22P and CNVR_23P). In addition, the associated CNVR _9SVS on BTA11 at 103.64-104.19 Mb lies in the proximity (about 100Kb) of the associated CNVR_19P.

The overlapping CNVRs between PennCNV and SVS8.3.1 did not contain any functional gene, except for the CNVR located on BTA7. This may be due to incompleteness in the annotation of bovine genome compared to the human one; otherwise, as reported by Wieczorek *et al.*, (2010) some CNVs are located in gene poor regions or in noncoding regions.

Comparison with literature findings showed that 78.5% of significant CNVRs (a total of 65) here identified have been already reported in 9 studies (Table S3), providing evidence they are likely true CNVRs. Additionally many of the CNVRs reported by the 9 studies perfectly overlapped those found here and were found among different breeds suggesting that they are CNVRs conserved across populations. The remaining 21.5% (a total of 14 CNVR) of the identified CNVRs were not previously reported and may be thus population specific or not yet detected. A further evidence that significant CNVRs are true regions comes from the number of individuals defining them, spanning from 80 to 140 for SVS 8.3.1 and from 7 to 103 for PennCNV-CNVRuler.

We compared the identified associated CNVRs with the reported cattle QTL in the Animal QTL database (http://www.animalgenome.org/cgi-bin/QTLdb/BT/index). Among the associated CNVRs, seven (SVS 8.3.1) and ten (CNVRuler) are regions overlapping the mapped QTL for SCS or for

Mastitis, as reported for both software in Table 4 (Clinical Mastitis as CM; Somatic cell count as SCC).

The Literature Mining Gene Network tool (provided by GenCLiP2.0), that searches for genes linked to keywords based on up-to-date literature profiling, revealed that 14 genes included within the significant CNVRs and two flanking genes (*BCL-2, PPARα*) have been associated mainly with the keywords ''Stress'', "cell death", ''inflammation'', and "immune response", as reported in Figure 2. The GO analysis performed for the gene included in the Figure 2, revealed that they are clustered into 19 groups of genes that were involved in a variety of cellular functions such as cell death, programmed cell death, tissue and organ development, and so on (Table S4 and Figure 3).

KEGG Pathway analysis showed the involvement of several signal pathways, such as immune response, apoptosis and adipocytes signalling (Table S5 and Figure 4).

The annotation analyses has enabled the identification of genes encoding for proteins that may be involved in the phenotypic variation of the SCS_EBV and consequently in the mastitis resistance. In particular, the association analysis performed with the SVS 8.3.1 allowed the identification of 7 candidate genes (*TERT, NOTCH1, SLC6A3, CLPTM1L, CACNA1S, PPARα* and *BCL-2*), while 11 candidate genes were found associated with CNV identified with PennCNV-CNVRuler analysis (*ABO, TRAF2, RELA, ELF3, DBH, CDK5, NF2, FASN, EWSR1, VAV2* and *MAP3K11*). Details on genes included in the networks and their function are included in Supporting Information 1 file.

Conclusions

The selective genotyping approach here used revealed to be efficient in identifying CNVs in the population and in associating them to the SCS_EBVs. The strategy here adopted to report CNVs mapped through the use of two different algorithms (CNAM and HMM) successfully reduced the false negative (and positives) that may be identified by only one approach.

Finally, this study is the first GWAS for SCS based on CNVs in Holstein cattle breed. Combining the CNVs detection/association analysis using two software allows a more complete identification of genes linked to phenotypic variation of the SCS trait, compared to those revealed using only one software.

Competing interests

The authors declare they have no competing interests.

Acknowledgments: We acknowledge the Council of Dairy Cattle Breeding and Holstein de Mexico A.C. for sharing genomic data. This work received funding from CONACYT, México (project "IDENTIFICACIÓN DE VARIACIONES ESTRUCTURALES Y REGIONES GENÓMICAS ASOCIADAS A RESISTENCIA A MASTITIS EN BOVINOS PRODUCTORES DE LECHE", 176039).

References

Bagnato A., Strillacci M.G., Pellegrino L., Schiavini F., Frigo E., Rossoni A., Fontanesi L., Maltecca C., Prinsen R.T.M.M., Dolezal M.A. (2015) Identification and validation of copy number variants in Italian Brown Swiss dairy cattle using Illumina Bovine SNP50 Beadchip. *Ital J Anim Sci*., **14**:2015.

Cha E., Bar D., Hertl J.A., Tauer L.W., Bennett G., González R.N., Schukken Y.H., Welcome F.L., Gröhn Y.T. (2011) The cost and management of different types of clinical mastitis in dairy cows estimated by dynamic programming. *J Dairy Sci*., **94**, 4476-87.

Choi J.W., Chung W.H., Lim K.S., Lim W.J., Choi B.H., Lee S.H., Kim H.C., Lee S.S., Cho E.S., Lee K.T., Kim N., Kim J.D., Kim J.B., Chai H.H., Cho Y.M., Kim T.H., Lim D. (2016) Copy number variations in Hanwoo and Yanbian cattle genomes using the massively parallel sequencing data. *Gene,* **589**,36-42.

Darvasi A. (1997) The effect of selective genotyping on QTL mapping accuracy. *Mammalian Genome,* **8**, 67-68.

Fontanesi L., Scotti E., Dolezal M., Lipkin E., Dall'Olio S., Zambonelli P., Bigi D., Davoli R., Soller M., Russo V. (2007) Bovine chromosome 20: milk production QTL and candidate gene analysis in the Italian Holstein-Friesian breed. Proceedings of the 17th ASPA Congress, Alghero, May 29-June 1, 2007, 133-135.

Glick G., Shirak A., Seroussi E., Zeron Y., Ezra E., Weller J.I., Ron M. (2011) Fine Mapping of a QTL for Fertility on BTA7 and Its Association with a CNV in the Israeli Holsteins. *G3*, **1**, 65–74.

Hinrichs D., Stamer E., Junge W., Kalm E. (2005) Genetic Analyses of Mastitis data using animal Threshold models and genetic correlation with production traits. *J. Dairy Sci*., **88**, 2260-2268.

Kijas J.W., Barendse W., Barris W., Harrison B., McCulloch R., McWilliam S., Whan V. (2011) Analysis of copy number variants in the cattle genome. *Gene,* **482**, 73-7.

Hou Y., Liu G.E., Bickhart D.M., Matukumalli L.K., Li, C., Song, J., Gasbarre L.C., Van Tassell C.P., Sonstegard T.S. (2012) Genomic regions showing copy number variations associate with resistance or susceptibility to gastrointestinal nematodes in Angus cattle. *Funct. Integr. Genomics,* **12**, 81–92.

Hou Y., Bickhart D.M., Chung H., Hutchison J.L., Norman H.D., Connor E.E., Liu G.E. (2012). Analysis of copy number variations in Holstein cows identify potential mechanisms contributing to differences in residual feed intake. *Funct. Integr. Genomics*, **12**, 717–723.

Jiang L., Jiang J., Yang J., Liu X., Wang J., Wang H., Ding X., Liu J., Zhang Q. (2013) Genome-wide detection of copy number variations using high-density SNP genotyping platforms in Holsteins. *BMC Genomics,* **14**, 131.

Mills, R.E., Walter, K., Stewart, C., Handsaker, R.E., Chen, K., Alkan, C., et al. 1000 Genomes Project. (2001) Mapping copy number variation by population-scale genome sequencing. *Nature,* **470**, 59-65.

Quinlan A.R., Hall I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics,* **26**, 841-842.

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. Biotechnol*, **29**, 512-520.

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

Sordillo L.M., Streicher K.L. (2002) Mammary Gland Immunity and Mastitis Susceptibility. *J of Mammary Gland Biology and Neoplasia*, **7**, 135-146.

Sasaki S., Watanabe T., Nishimura S., Sugimoto Y. (2016) Genome-wide identification of copy number variation using high-density single-nucleotide polymorphism array in Japanese Black cattle. *BMC Genet*., **25**, 17(1):26.

Strillacci M.G., Frigo E., Canavesi F., Ungar Y., Schiavini F., Zaniboni L., Reghenzani L., Cozzi M.C., Samoré A.B., Kashi Y., Shimoni E., Tal-Stein R., Soller M., Lipkin E., Bagnato A. (2014) Quantitative trait loci mapping for conjugated linoleic acid, vaccenic acid and Δ(9)-desaturase in Italian Brown Swiss dairy cattle using selective DNA pooling. *Anim Genet*., **45**, 485-99.

Xu L., Cole J.B., Bickhart D.M., Hou Y., Song J., VanRaden P.M., Sonstegard T.S., Van Tassell C.P. Liu, G.E. (2014) Genome wide CNV analysis reveals additional variants associated with milk production traits in Holsteins. *BMC Genomics*, **15**, 683.

Xu L., Hou Y., Bickhart D.M., Zhou Y., Hay E.H.A, Song J., Sonstegard T.S., Van Tassell C.P, Liu G.E. (2016) Population-genetic properties of differentiated copy number variations in cattle. *Scientific Reports,* **6**,23161.

Wain L.V., Armour J.A..L, Tobin M.D. (2009) Genomic copy number variation, human health, and disease. *Lancet*, **374**, 340-350.

Wu Y., Fan H., Jing S., Xia J., Chen Y., Zhang L., Gao X, Li J., Gao H., Ren H. (2015) A genome-wide scan for copy number variations using high-density single nucleotide polymorphism array in Simmental cattle. *Anim Genet.,* **46,** 289-98.

Winchester L., Yau C., Ragoussis J. (2009) Comparing CNV detection methods for SNP arrays. Brief *Funct Genomic Proteomic.*, **8**, 353-366.

Zhang Q., Ma Y., Wang X., Zhang Y., Zhao X. (2015) Identification of copy number variations in Qinchuan cattle using BovineHD Genotyping Beadchip array. *Mol Genet Genomics*, **290**, 319–27.

Table 1 Descriptive statistics for CNVs identified with PennCNV and SVS 8.3.1 software

Copy number*	Number of CNVs	Mean Lenght	Min Lenght	Max Lenght
		SVS 8.3.1		
Loss	5088	318,123	1,245	2,760,295
Gain	106	718,633	9,245	2,805,791
Totale	*5194*	*518,378*	*1,245*	*2,805,791*
		PennCNV		
0	2354	64,47	1,229	602,303
1	8121	54,39	1,112	1,248,573
3	1566	94,328.5	998	1,185,515
4	29	147,364	4,044	724,916
Total	*12070*	*90,138*	*998*	*1,248,573*

*0 = homozygous deletion, 1 heterozygous deletion, 3 heterozygous duplication, and 4 homozygous duplication

Table 2 CNVRs significantly associated with SCS_EBV identified by SVS 8.3.1

CNVR_ID	CHR	START	END	LENGHT	STATE	FREQ	SNP Predictor	p-value	*Genes within the significant CNVRs*
CNVR_9_SVS	1	93957123	94357120	399997	loss	8	BovineHD0100026648	8,81E-03	
							BovineHD0100026649	7,11E-03	
							BovineHD0100026754	1,25E-02	
CNVR_23_SVS	2	46477034	46485436	8402	loss	30	BovineHD0200013459	1,44E-03	
CNVR_39_SVS	3	7957960	7964523	6563	loss	18	BTA-66943-no-rs	4,69E-02	
							BovineHD0300002600	1,27E-02	
CNVR_58_SVS	5	22514133	22563988	49855	loss	25	BovineHD0500006525	3,94E-02	
CNVR_60_SVS	5	58966295	59255853	289558	complex	51	BovineHD0500035991	4,94E-02	
							BovineHD0500036000	5,32E-02	
CNVR_63_SVS	5	117246007	117651752	405745	complex	179	BovineHD0500034077	4,79E-02	
							BovineHD0500034078	2,59E-02	
							BovineHD0500034128	2,50E-02	
							BovineHD0500034132	3,69E-02	
							BovineHD0500036296	3,63E-02	
							BovineHD0500034144	5,86E-04	
							BovineHD0500034145	8,88E-04	
							BovineHD0500034148	9,09E-04	
							BovineHD0500034150	8,98E-04	
							BovineHD0500034151	1,74E-03	
							BovineHD0500034152	1,80E-03	
							BovineHD0500034153	1,59E-03	
CNVR_83_SVS	7	42745346	42788788	43442	loss	31	BovineHD0700012438	1,28E-03	OR2AK2
							BovineHD0700012440	2,65E-03	
							BovineHD0700012441	2,10E-03	
							BovineHD0700012444	2,65E-03	
							ARS-BFGL-NGS-23938	7,27E-03	
CNVR_122_SVS	10	87873996	87878635	4639	loss	11	BovineHD1000024990	1,25E-02	
							ARS-BFGL-NGS-112168	1,13E-02	
CNVR_125_SVS	11	103644879	104195124	550245	loss	140	BovineHD1100031771	5,78E-02	C9orf69, LHX3, QSOX2, GPSM1, DNLZ, CARD9, SNAPC4, SDCCAG3, PMPCA, INPP5E, SEC16A, NOTCH1, EGFL7, bta-mir-126, AGPAT2, FAM69B
							BovineHD4100009284	5,39E-02	
							BovineHD1100030807	4,21E-02	
							BovineHD1100030845	5,43E-02	
CNVR_134_SVS	12	70363408	72077746	1714338	complex	159	BovineHD1200019362	9,96E-03	
							BovineHD1200028177	2,22E-02	
							BovineHD1200019797	4,63E-02	
CNVR_137_SVS	12	72411533	75238779	2827246	complex	481	BovineHD1200019975	5,20E-02	
							BovineHD1200019998	3,56E-02	
							BovineHD1200020000	2,35E-02	
							BovineHD1200020001	5,65E-03	
							BovineHD1200020003	4,06E-03	
							BovineHD1200020163	2,48E-02	
							BovineHD1200020442	3,88E-02	
							BovineHD1200020450	4,19E-02	
							BovineHD1200020457	3,05E-02	
							BovineHD1200020475	2,44E-02	
							BovineHD1200020488	2,20E-02	
							BovineHD1200020490	5,15E-02	
							BovineHD1200020495	5,15E-02	
							BovineHD1200020612	5,58E-02	
							BovineHD1200028386	5,14E-02	
							BovineHD1200020699	1,76E-02	
							BovineHD1200020725	1,04E-02	
							BovineHD1200020840	5,98E-02	
CNVR_138_SVS	12	75509770	76488279	978509	loss	39	BovineHD1200021096	1,91E-02	
CNVR_140_SVS	13	53858853	53862891	4038	loss	5	BovineHD1300015258	1,82E-03	
CNVR_175_SVS	16	81343003	81720984	377981	loss	80	BovineHD1600023844	3,95E-02	

							BovineHD1600023853	5,59E-02	C1orf106, KIF21B, CACNA1S, TMEM9, IGFN1, PKP1
							BovineHD1600023856	5,99E-02	
CNVR_186_SVS	18	65766249	65771834	5585	loss	10	BovineHD1800019186	2,81E-03	
CNVR_191_SVS	20	70913332	71571246	657914	loss	139	BovineHD2000020825	1,46E-02	IRX4, NDUFS6, MRPL36, LPCAT1, SLC6A3,CLPTM1L, TERT, SLC6A18, SLC6A19, SLC12A7, NKD2, TRIP13, BRD9, TPPP
							BTA-51318-no-rs	2,71E-02	
							BovineHD2000020835	1,17E-02	
							BovineHD2000020840	1,34E-02	
							BovineHD2000020849	1,62E-02	
							BovineHD2000020852	2,10E-02	
CNVR_198_SVS	21	66704964	66750757	45793	loss	5	BovineHD2100019578	4,06E-02	bta-mir-342, DEGS2
CNVR_208_SVS	22	60911345	60981720	70375	loss	17	BovineHD2200017757	8,88E-03	CHCHD6, TXNRD3, C3orf22, CHST13, UROC1, ZXDC, SLC41A3, ALDH1L1, KLF15, CCDC37
CNVR_213_SVS	23	25869447	26337243	467796	loss	9	BovineHD2300007174	2,28E-03	
CNVR_214_SVS	23	28448873	28469826	20953	loss	14	BovineHD2300008005	5,23E-05	
CNVR_217_SVS	23	28828468	28849820	21352	loss	47	BovineHD2300008182	4,41E-03	
							BovineHD2300008186	2,86E-03	
							BovineHD2300008188	6,64E-03	
CNVR_222_SVS	24	37553499	37581537	28038	loss	5	BovineHD2400010262	1,04E-02	LPIN2
CNVR_227_SVS	24	62411069	62431830	20761	complex	49	BTB-01625084	3,01E-02	
CNVR_240_SVS	28	10760635	10774825	14190	loss	5	BovineHD2800003298	3,65E-03	

Table 3 CNVRs significantly associated with SCS_EBV identified by PennCNV-CNVRuler

CNVR ID	CHR	START	END	LENGHT	STATE	FREQ	p-value	Genes within the significant CNVRs
CVNR_36_P	1	93954887	94357120	402234	loss	8	7,20E-03	
CVNR_66_P	1	146975308	147110229	134922	loss	86	5,50E-04	
CVNR_112_P	2	98480344	98490521	10178	gain	6	3,46E-04	
CVNR_171_P	3	50167465	50191213	23749	loss	10	3,70E-02	
CVNR_186_P	3	93310320	93315045	4726	loss	7	5,71E-03	
CNVR_232_P	4	28744454	28751390	6937	loss	7	1,14E-02	
CNVR_281_P	4	114419925	114514111	94187	loss	**7**	1,43E-02	ABCB8, ASIC3, CDK5, SLC4A2, FASTK, TMUB1, AGAP3
CNVR_324_P	5	58966295	59168921	202627	gain	6	3,68E-02	
CNVR_347_P	5	107628624	107660101	31478	loss	12	4,52E-02	IQSEC3, SLC6A12
CNVR_375_P	5	117281795	117639815	358021	loss	74	4,11E-02	
CNVR_379_P	5	118109413	118174364	64952	loss	8	3,66E-02	TBC1D22A
CNVR_395_P	5	121149647	121183174	33528	loss	9	3,60E-02	MOV10L1, bta-mir-2894, PANX2, TRABD
CNVR_471_P	7	15083922	15102276	18355	loss	10	4,05E-02	
CNVR_502_P	7	42736530	42788788	52259	loss	29	1,42E-02	OR2AK2
CNVR_503_P	7	42945525	43087430	141906	loss	13	6,71E-03	OR2AJ1
CNVR_508_P	7	45487894	45538477	50584	loss	21	1,57E-02	APC2, C19orf25, PCSK4, REEP6
CNVR_653_P	10	16816476	16844526	28051	loss	3	4,27E-02	
CNVR_658_P	10	23540925	23635452	94528	loss	3	3,94E-02	
CNVR_765_P	11	104295522	104764859	469338	loss	**76**	7,07E-03	ABO, SURF6, MED22, RPL7A, SURF1, SURF2, STKLD1, REXO4, ADAMTS13, CACFD1, SLC2A6, TMEM8C, ADAMTSL2, FAM163B, DBH, SARDH, VAV2
CNVR_770_P	11	106158972	106415916	256945	loss	**103**	3,21E-02	UAP1L1, SAPCD2, ENTPD2, NPDC1, FUT7, ABCA2, CLIC3, C9orf142, LCNL1, PTGDS, LCN12, C8G, FBXW5, TRAF2, EDF1, MAMDC4, PHPT1, C9orf172, RABL6, CCDC183, TMEM141, LCN8, LCN15, LCN10
CNVR_777_P	12	717729	731185	13457	gain	4	3,19E-04	
CNVR_824_P	12	72432362	73015638	583277	loss	57	3,69E-02	
CNVR_825_P	12	74840021	75238779	398759	loss	75	4,59E-02	
CNVR_1048_P	16	70814352	71165517	351166	loss	82	3,86E-02	SMYD2, RNPEP, ELF3, GPR37L1, ARL8A, PTPN7, LGR6, UBE2T, PPP1R12B, SYT2
CNVR_1062_P	17	15677009	15720425	43417	loss	4	2,54E-02	INPP4B
CNVR_1090_P	17	70714297	70748407	34111	loss	12	4,48E-02	EWSR1, GAS2L1, RASL10A, AP1B1
CNVR_1091_P	17	70794775	70817022	22248	loss	7	3,25E-02	
CNVR_1092_P	17	70963787	71024477	60691	loss	7	2,26E-02	NF2, CABP7, ZMAT5
CNVR_1134_P	18	27914135	28375996	461862	gain	12	1,86E-02	
CNVR_1192_P	19	24548362	24571149	22788	gain	4	1,75E-02	
CNVR_1210_P	19	37277118	37328651	51534	loss	8	1,15E-02	DLX3, DLX4
CNVR_1124_P	19	51028723	51073939	45217	loss	11	8,13E-03	
CNVR_1126_P	19	51365385	51514295	148911	loss	10	4,98E-02	FASN, DUS1L, GPS1, RFNG, DCXR, RAC3, LRRC45, STRA13
CNVR_1231_P	19	52776058	52903129	127072	gain	7	2,65E-02	
CNVR_1236_P	19	54639709	54687169	47461	loss	15	3,52E-02	TMC8, TMC6
CNVR_1321_P	21	54162719	54196002	33284	loss	9	1,44E-02	
CNVR_1345_P	22	20291128	20331448	40321	loss	7	5,32E-03	
CNVR_1398_P	23	21694996	21702537	7542	loss	5	3,40E-02	
CNVR_1399_P	23	25335659	25361041	25383	loss	12	1,01E-03	
CNVR_1408_P	23	28828468	28849820	21353	loss	22	4,64E-02	
CNVR_1417_P	23	34779270	34866601	87332	gain	3	1,76E-02	
CNVR_1519_P	26	23347145	23380565	33421	loss	7	1,00E-02	
CNVR_1549_P	26	51434163	51680135	245973	loss	38	3,30E-02	JAKMIP3, DPYSL4, STK32C, LRRC27, PWWP2B
CNVR_1584_P	28	2263677	2271424	7748	loss	4	1,21E-02	
CNVR_1617_P	29	27363231	27409510	46280	loss	14	3,78E-02	
CNVR_1640_P	29	44416282	44502548	86267	loss	14	4,69E-02	SSSCA1, FAM89B, EHBP1L1, KCNK7, MAP3K11, PCNXL3, SIPA1, RELA
CNVR_1648_P	29	47039694	47054342	14649	loss	3	3,66E-02	TPCN2

Table 4 QTL mapped within significant CNVRs

CNVR_ID	Chr	Start	End	Lenght	Start_QTL	End_QTL	QTL trait_id
Signifiacnt CNVR_SVS 8.3.1							
CNVR_23_SVS	2	46477034	46485436	8402	45424584	52384967	CM (DYD) QTL #19007, QTL #19004, QTL #19005, QTL #19006
CNVR_83_SVS	7	42745346	42788788	43442	27358606	42831622	SCS QTL #2667
CNVR_140_SVS	13	53858853	53862891	4038	51062875	56847265	SCS QTL #2775
CNVR_213_SVS	23	25869447	26337243	467796	23274081	31653997	SCS QTL #2688
CNVR_214_SVS	23	28448873	28469826	20953	23274081	31653997	SCS QTL #2688
					27452360	31104253	SCS QTL #4989
CNVR_217_SVS	23	28828468	28849820	21352	23274081	31653997	SCS QTL #2688
					27452360	31104253	SCS QTL #4989
CNVR_240_SVS	28	10760635	10774825	14190	10665897	11438802	SCS QTL #16056
Signifiacnt CNVR_PennCNV							
CVNR_502_P	7	42736530	42788788	52259	27358606	42831622	SCS QTL #2667
CNVR_503_P	7	42945525	43087430	141906	42834942	50547685	SCC QTL #2698
CNVR_508_P	7	45487894	45538477	50584	42834942	50547685	SCC QTL #2698
CNVR_658_P	10	23540925	23635452	94528	22939631	40797089	SCC QTL #2701
CVNR_1134_P	18	27914135	28375996	461862	27863715	33011652	SCC QTL #4638
CVNR_1226_P	19	51365385	51514295	148911	51395368	51495967	SCS (DYD) QTL #32265
CVNR_1398_P	23	21694996	21702537	7542	21554613	22522198	SCS QTL #19986, #19991
CVNR_1399_P	23	25329895	25417035	87140	23274081	31653997	SCS QTL #2688
					27452360	31104253	SCS QTL #4989
CVNR_1408_P	23	28828468	28849820	21353	23274081	31653997	SCS QTL #2688
					27452360	31104253	SCS QTL #4989
CVNR_1648_P	29	47039694	47054342	14649	46178647	52998234	CM (DYD) QTL #19031

Figure legends

Figure 1. Manhattan plots of associated CNVs for SCS_EBV using SVS 8.3.1. (A) and PennCNV-CNVRuler (B).

Figure 2. Candidate Genes Network

Figure 3. Cluster results of Go analysis for all genes included in significant CNVR (both software).

Figure 4. Cluster results for pathway analysis (both software).

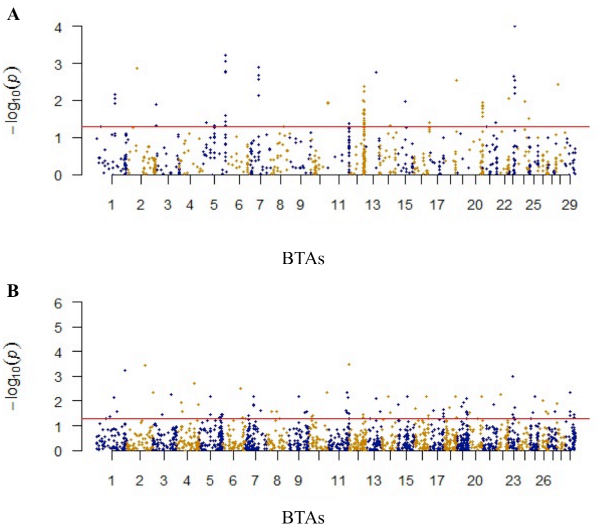
A
−log10(p)
0
1
2
3
4
1
2
3
4
5
6
7
8
9
11
13
15
17
20
22
25
29
BTAs
B
−log10(p)
0
1
2
3
4
5
6
1
2
3
4
5
6
7
8
9
11
13
15
17
20
23
26
BTAs

Figure 1

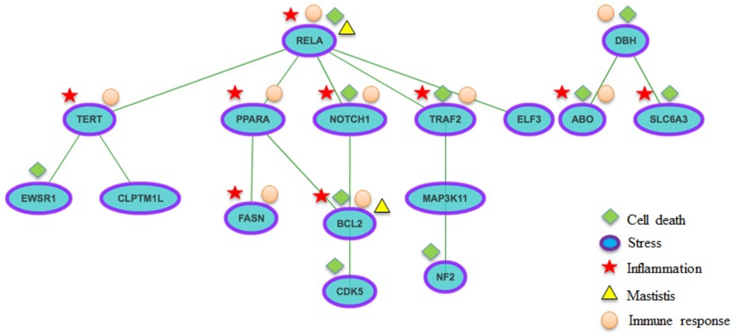
RELA
DBH
TERT
PPARA
NOTCH1
TRAF2
ELF3
ABO
SLC6A3
EWSR1
CLPTM1L
FASN
BCL2
MAP3K11
CDK5
NF2
Cell death
Stress
Inflammation
Mastistis
Immune response

Figure 2

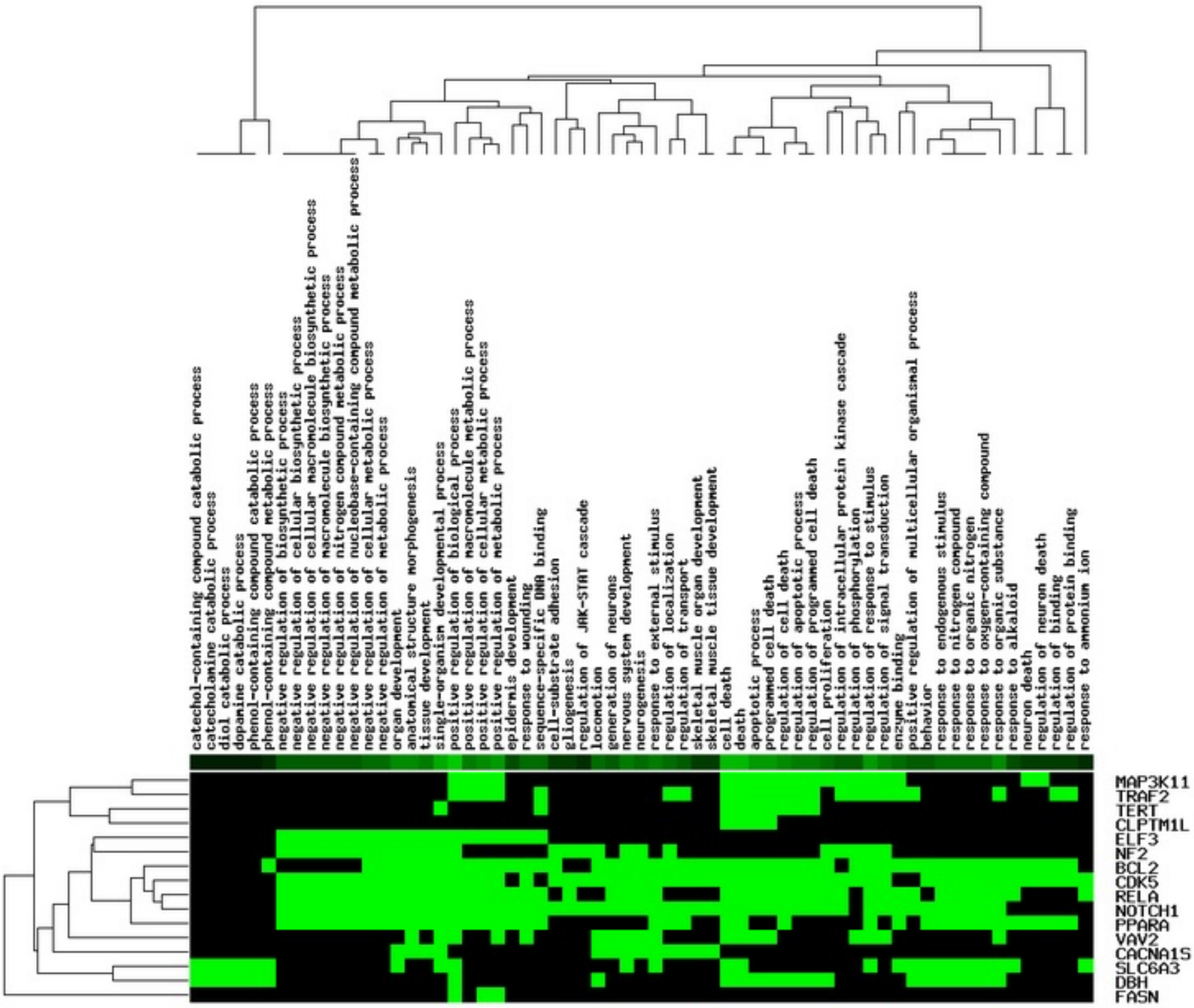
catechol-containing compound catabolic process
catecholamine catabolic process
diol catabolic process
dopamine catabolic process
phenol-containing compound catabolic process
phenol-containing compound metabolic process
negative regulation of biosynthetic process
negative regulation of cellular biosynthetic process
negative regulation of cellular macromolecule biosynthetic process
negative regulation of macromolecule biosynthetic process
negative regulation of nitrogen compound metabolic process
negative regulation of nucleobase-containing compound metabolic process
negative regulation of cellular metabolic process
negative regulation of metabolic process
organ development
anatomical structure morphogenesis
tissue development
single-organism developmental process
positive regulation of biological process
positive regulation of macromolecule metabolic process
positive regulation of cellular metabolic process
positive regulation of metabolic process
epidermis development
response to wounding
sequence-specific DNA binding
cell-substrate adhesion
gliogenesis
regulation of JAK-STAT cascade
locomotion
generation of neurons
nervous system development
neurogenesis
response to external stimulus
regulation of localization
regulation of transport
skeletal muscle organ development
skeletal muscle tissue development
cell death
death
apoptotic process
programmed cell death
regulation of cell death
regulation of apoptotic process
regulation of programmed cell death
cell proliferation
regulation of intracellular protein kinase cascade
regulation of phosphorylation
regulation of response to stimulus
regulation of signal transduction
enzyme binding
positive regulation of multicellular organismal process
behavior
response to endogenous stimulus
response to nitrogen compound
response to organic nitrogen
response to oxygen-containing compound
response to organic substance
response to alkaloid
neuron death
regulation of neuron death
regulation of binding
regulation of protein binding
response to ammonium ion
MAP3K11
TRAF2
TERT
CLPTM1L
ELF3
NF2
BCL2
CDK5
RELA
NOTCH1
PPARA
VAV2
CACNA1S
SLC6A3
DBH
FASN

Figure 3

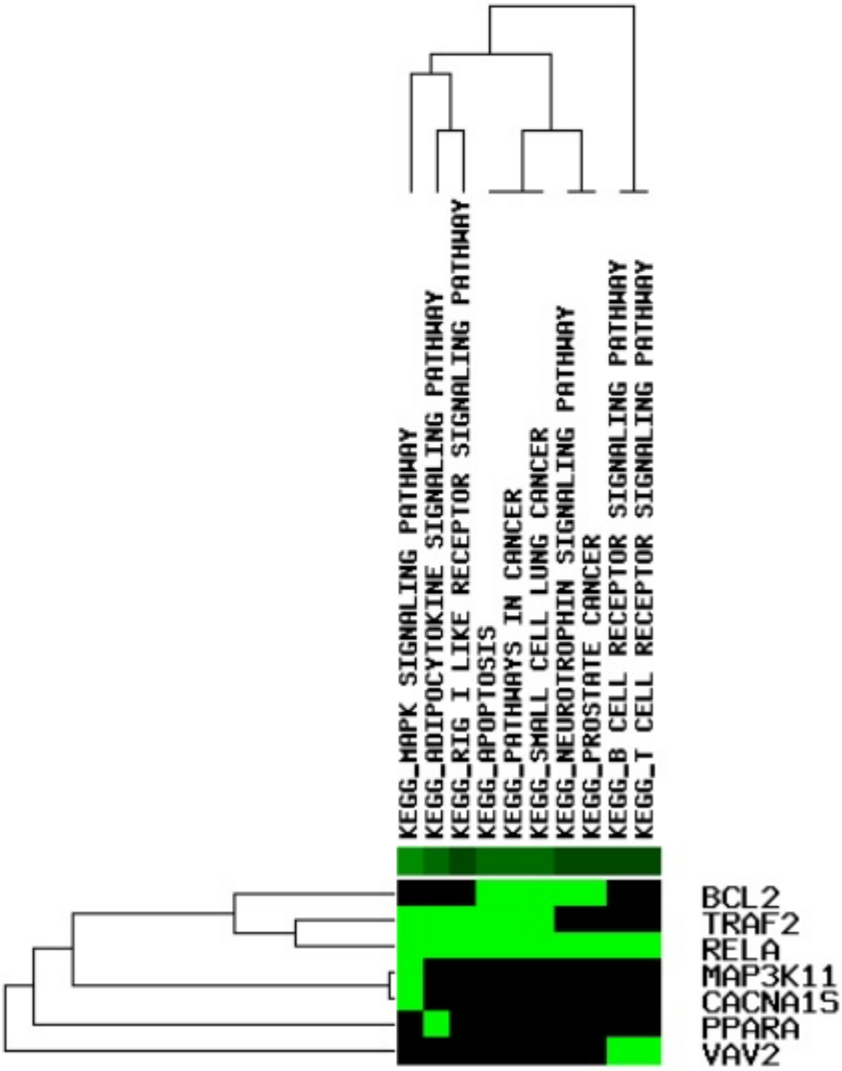
KEGG_MAPK SIGNALING PATHWAY
KEGG_ADIPOCYTOKINE SIGNALING PATHWAY
KEGG_RIG I LIKE RECEPTOR SIGNALING PATHWAY
KEGG_APOPTOSIS
KEGG_PATHWAYS IN CANCER
KEGG_SMALL CELL LUNG CANCER
KEGG_NEUROTROPHIN SIGNALING PATHWAY
KEGG_PROSTATE CANCER
KEGG_B CELL RECEPTOR SIGNALING PATHWAY
KEGG_T CELL RECEPTOR SIGNALING PATHWAY
BCL2
TRAF2
RELA
MAP3K11
CACNA1S
PPARA
VAV2

Figure 4

Supporting Information

Supporting Information 1 Candidate genes details and References

Table S1 CNVRs identified by SVS8.3.1 software

Table S2 CNVRs identified by PennCNV software

Table S3 Comparison of significantly associated CNVR found in this study with those identified in other researches.

Table S4 GO analysis results of candidate genes

Table S5 KEGG pathway results of candidate genes