Copy Number Variation in Brown Swiss Dairy Cattle

M. A. Dolezal[1,2], A. Bagnato[1], F. Schiavini[1], E. Santus[3], L-E. Holm[4], C. Bendixen[4], and *F. Panitz*[4]

[1]Università degli Studi di Milano, Milano, Italy, [2]University of Veterinary Medicine Vienna, Vienna, Austria, [3]Associazione Nazionale Allevatori Razza Bruna, Bussolengo, Italy, [4]Aarhus University, Molecular Biology and Genetics, Arhus, Denmark.

ABSTRACT: CNVs are increasingly recognized as substantial source of genetic variation, fueling studies that assess their impact on complex traits. In particular rare CNVs have been suggested to potentially explain part of the missing heritability problem in genome wide association studies for complex traits. The objective of this study was to perform a high resolution genome scan for CNV, in a sample of 20 Brown Swiss dairy cattle bulls based on ~20x Illumina whole genome resequencing data. Employing CNVnator for variant discovery, we present descriptive statistics for the CNVs detected and define consensus CNV regions at the population level. We identified 29,975 deletion-, 1,489 duplication- and 365 complex CNVRs, respectively, which cover 3.3% of the UMD3.1 autosome. We further compared NGS based CNV calls to CNV calls detected by PennCNV based on Illumina HD chip data for 17 bulls with high quality data for both platforms.

Keywords:
dairy cattle
copy number variation

Introduction

One of the central questions in biological sciences is to understand how genetic variation shapes phenotypes, elucidating adaptation, selection, general physiology and disease. Copy number variants (CNVs) are DNA segments, originally for technical reasons defined as bigger than 1 kb, differing in copy number between individuals. They are sub-classified as deletions, insertions and duplications as compared to a reference (Lee and Scherer, (2010)). Early estimates of the extent of copy number variable sequence in the human genome ranged from 5 to 12% (Redon et al., (2006); McCarroll et al., (2008)) depending on sample size and resolution of technology employed for their detection. CNVs are far less abundant than SNPs. However, as they can affect up to several Mb of sequence, inter individual variability in humans based on CNVs is much higher (Redon et al. (2006)).

The aim of this study was to perform a genome scan for CNVs based on high coverage Illumina next generation sequencing data to expand on the catalogue of CNVRs in the bovine genome and to provide a high resolution map of CNVRs specific to Brown Swiss dairy cattle, an important prerequisite for any population genetic and quantitative genetic analysis of copy number polymorphisms (CNPs) in this breed. We further set out to compare NGS based CNV calls to CNVs called from the Illumina bovine HD Infinium Bead Chip with a median marker spacing of < 3 kb in the same individuals.

Materials and Methods

Data. Twenty Brown Swiss (BSW) bulls were re-sequenced at about 20x coverage with 100 bp paired-end reads derived from 300 bp and 800 bp insert Illumina libraries. After aligning reads to the UMD3.1 reference assembly (Zimin et al., (2009)) with BWA (Li and Durbin, (2009)), resulting BAM files are marked for duplicates (Picard; http://picard.sourceforge.net) and filtered for mapping quality, realigned and recalibrated (GATK v2.4.9; McKenna et al. (2010)). 192 BSW bulls were genotyped with Illumina's HD chip featuring ~777k SNPs on UMD3.1 assembly.

Variation discovery. CNVs were called with CNVnator v0.3 (Abyzov et al. (2011)) chromosome-wise per individual with a bin size of 100. We removed all CNV events that were within 100 bp of 63,358 assembly gaps greater than 11 bp. We further filtered for deletions with average read depth (RD) <0.5 and duplications with 1.5 < RD < 10. Bovine HD chip data was analyzed with PennCNV (Wang et al. (2007)) with option -gcmodel to correct for genomic waves in 164 bulls with high signal to noise ratio after stringent quality control and filtering based on derivative log R ratio (LRR) spread values, waviness patterns and manual inspection of LRR values plotted in genomic space.

Statistical analyses. Descriptive statistics were calculated for identified CNVs and CNVRs for both datasets. We employed the CNVR definition of Redon et al. (2006). CNVs identified across 20 bulls that were overlapping by at least 1 nucleotide were summarized to CNVRs using BEDTools (Quinlan and Hall (2010)) 'mergeBed' command at the Brown Swiss population level as overlapping CNVs of the same copy number state are likely to represent the same mutational event. CNVRs were then classified as loss, gain or complex CNVRs, the latter comprising loss and gain events in a particular region in the population.

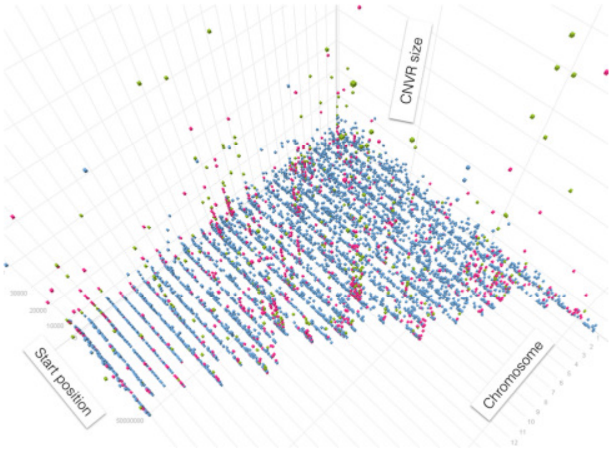
Consensus between NGS and HD chip data. Seventeen of the re-sequenced bulls had a high signal to noise ratio for the HD chip data. For these we extracted all CNVs called by PennCNV and CNVnator and calculated the fraction of overlapping CNVs per bull and CNV type (deletion or duplication) with BEDTools (Quinlan and Hall (2010)) intersectBed with option –u to account for multiple NGS based hits within CNVs called from the HD chip data

Comparison of NGS-CNV calls to literature. We downloaded CNV/CNVR information from cattle CNV scans based on NGS data (Zhan et al. (2011); Stothard et al. (2011); Bickhart et al. (2012); Shin et al. (2014)). For scans based on Btau4.0 (bosTau4) we used the lift over tool at UCSC genome browser (http://genome.ucsc.edu/cgi-bin/hgLiftOver) to convert autosomal CNVRs to UMD3.1 (bosTau6). We then used BEDTools (Quinlan and Hall (2010)) intersectBed command (with default minimum overlap of 1 nucleotide) to compare CNVRs identified in our study to those reported in literature.

Results and Discussion

Results from NGS data. Applying the RD thresholds yielded total of 181,951 CNVs comprising 170,537 deletions (RD<0.5) and 11,414 duplications (1.5<RD<10), respectively. Table 1 shows descriptive statistics of CNV length (in kb) for deletion and duplications. The mean CNV coverage of autosomes across bulls was 0.73 % +/- 0.13 and 0.17 % +/- 0.03 for deletions and duplications, respectively. The genome-wide distribution of the identified CNVRs is shown in Figure 1 where deletion-, duplication- or complex-type CNVRs with lengths between 3kb and 70kb are displayed. All CNVRs combined cover 82.92 Mb of the UMD3.1 autosome; (64.88 Mb by deletions, 9.85 Mb duplications and 8.19Mb by complex CNVRs, respectively). We observe excessive coverage in difficult to assemble regions. CNVnator performs GC content correction. We nevertheless observe a considerable fraction of deletions in regions of extremely unbiased nucleotide content. This is a known technical artefact of Illumina sequencing data that is not consistent between samples and hard to remove (Benjamini and Speed, (2012)).

Figure 1: Genome map of CNVRs identified with CNVnator. [&]



[&] blue – Deletion-, red – Duplication- , green Complex- CNVRs

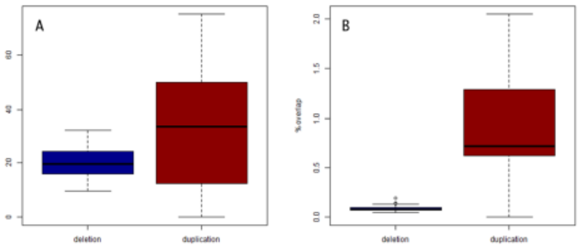
Results from HD chip data. PennCNV identified a total of 9,203 CNVs (6,217 deletions and 2,986 duplications) in 164 bulls. 1,060 of these (614 deletions and 446 duplications) were identified to segregate in the 17 BSW bulls that were also re-sequenced. Table 1 shows descriptive statistics of CNV length (in kb) for deletion and duplications. Mean autosome coverage per bull was 0.07 % +/- 0.02 and 0.04 %+/- 0.02 for deletions and duplications, respectively.

Comparison between NGS and HD chip data. Figure 2 shows a comparison of CNV calls from HD chip data and NGS based calls for 17 bulls. Only less than 2% of NGS based calls were identified in HD chip data. However on average 20% of deletions and almost 40% of duplications of HD chip based calls could be verified by at least one overlapping NGS based call. Despite the fact that NGS based calls are likely to contain false positive calls, CNV scans from HD chip data have as expected a high false negative rate, particularly for small events. CNV breakpoints were much less variable in the NGS dataset.

Comparison of NGS-CNV calls to literature. Table 3 shows the overlap between CNVRs found in BSW in this study and CNV or CNVRs described in recent literature which could be successfully converted to UMD3.1 coordinates. Note that such a comparison can potentially be biased by the smaller number of animals in older studies. Furthermore the study at population level (Shin et al. (2014)) reports on deletions only.

CNVR hotspots. Individual chromosomes show clear hotspots for CNVRs (Figure 1 and Table 2) in agreement in both datasets (data not shown) such as the MHC region on BTA23. Another interesting complex CNVR locates to 70-75 Mb on BTA12 harboring nine paralogous protein coding genes. Furthermore the orthologous regions in *Capra hircus* (Fontanesi et al., (2010)), *Ovis aries* (Fontanesi et al., (2011)) and *Sus scrofa* (Luca Fontanesi, pers. comm.) all show strong evidence for CNV. ENSBTAG00000032603 located in this region is a 1:many orthologue of human ENSG00000125257 (ABCC4) that has been shown to play a role in Kawasaki disease pathogenesis with effects on immune activation and vascular response to injury in humans (Khor et al., (2011)). There are multiple QTLs annotated in this region at cattleQTLdb (Hu et al. (2013)) for both functional traits (e.g. Schuman et al., (2008); Seidenspinner et al., (2009)) and production traits (Ashwell et al., (1998)) in several cattle populations.

Figure 2: Boxplot of percentage overlapping CNVs identified for each of 17 bulls (A) with respect to calls made with HD chip data; (B) with respect to NGS based CNV calls.



Conclusion

At least one additional independent CNV calling tool per dataset to decrease false positive calls, at the cost of increased false negative rates should be used. We will further attempt to filter in regions of GC/AT content bias to reduce false positive deletion calls.

Acknowledgements

M.A.D. and F.P. contributed equally to data analysis and writing. Data was generated as part of the FP7 project QUANTOMICS contract n. 222664-2.

Literature Cited

Abyzov, A., Urban, A., E., Snyder, M., et al. (2011). Genome Res., 21:974-84.

Ashwell, M. S., Da, Y., Van Tassell, C. P., et al. (1998). J. Dairy Sci. 81:3309-3314.

Benjamini, Y. and Speed, T. P. (2012). Nucl. Acids Res. (2012) 40 (10): e72.

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

Fontanesi, L., Martelli, P. L., Beretti, F., et al. (2010). BMC Genomics 11: 639.

Fontanesi, L,. Beretti, F., Martelli, P. L., et al. (2011). Genomics 97:158-65.

Hu, Z. L., Park, C. A., Wu, X. L. et al. (2013). Nucl. Acid. Res. 41 (D1): D871-D879.

Khor, C. C., Davila, S., Shimizu, C., et al. (2001). J. Med. Genet. 48:467-472.

Lee, C. and Scherer, S. W. (2010). Expert. Rev. Mol. Med. 12:e8.

Li, H. and Durbin, R. (2009). Bioinformatics, 25:1754-60.

McCarroll, S. A., Kuruvilla, F. G., Korn, J. M., et al. (2008). Nat. Genet. 40:1166–1174.

McKenna, A., Hanna, M., Banks, E. et al. (2010). Genome Res. 20:1297-1303.

Quinlan, A. R. and Hall, I. M. (2010). Bioinformatics 26:841–842.

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

Seidenspinner, T., Bennewitz, J., Reinhardt, F., et al. (2009). J. Anim. Breed. Genet. 126:455-62.

Schulman, N. F., Sahana, G., Lund. M. S., et al. (2008). Genet. Sel. Evol. 40:195-214.

Shin, D., Lee, H., Cho, S., et al. (2014). BMC Genomics 15:240.

Stothard, P., Choi, J. W., Basu, U., et al. (2011). BMC Genomics 15:2:559.

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

Zhan, B., Fadista, J., Thomsen, B., et al. (2011). BMC Genomics, 12:557.

Zimin, A., Delcher, A., Florea, L. et al. (2009). Genome Biol. 10:R42.

Table 1: Descriptive Statistics for CNVs (in kb) identified by PennCNV for HD chip data and CNVnator for NGS data.

	PennCNV		CNVnator	
	deletions	**duplications**	**deletions**	**duplications**
minimum	1.17	1.01	0.2	0.3
maximum	581.43	623.25	210.7	317.9
mean	45.02	36.46	2.16	7.29
median	21.90	13.78	1.6	4.40
sum	27643.9	16260.4	368419.2	83185.2

Table 2: CNVR summary table; Coverage in percentage by autosome

BTA	all CNVRs	duplication CNVRs	deletion CNVRs	complex CNVRs
1	2.78	0.18	2.52	0.08
2	3.21	0.25	2.30	0.66
3	3.11	0.42	2.48	0.21
4	3.69	0.48	2.48	0.72
5	3.45	0.36	2.82	0.26
6	3.13	0.25	2.78	0.10
7	3.36	0.50	2.67	0.18
8	3.23	0.25	2.78	0.20
9	2.96	0.24	2.59	0.14
10	3.35	0.47	2.64	0.24
11	2.66	0.18	2.37	0.10
12	4.80	0.78	2.69	1.33
13	3.14	0.48	2.23	0.43
14	3.33	0.65	2.48	0.20
15	4.17	0.70	2.96	0.52
16	3.28	0.40	2.43	0.45
17	3.47	0.33	2.81	0.33
18	3.98	0.85	2.51	0.61
19	2.97	0.43	2.38	0.16
20	2.87	0.22	2.59	0.07
21	3.14	0.38	2.59	0.16
22	2.51	0.13	2.26	0.12
23	4.56	0.37	3.20	0.98
24	3.03	0.18	2.71	0.14
25	2.97	0.31	2.63	0.03
26	2.99	0.35	2.51	0.13
27	3.65	0.85	2.31	0.49
28	2.74	0.33	2.33	0.08
29	3.77	0.52	2.98	0.27
autosome	3.30	0.39	2.58	0.33

Table 3: Comparison of CNVRs in this study and cattle CNV scans based on NGS data reported in literature.

	this study	[1]		[2]		[3]		[4]	
	20 BSW	**1 Holstein**		**1 Holstein, 1 Black Angus**		**1 Nelore, 1 Holstein, 3 Angus**		**22 Hanwoo, 10 Holstein**	
#CNV/CNVR		#	#∩	#	#∩	#	#∩	#	#∩
all	31820	520	84	565	152	760	503	6811	2471
deletions	29975	345	43	NA	NA	697	73	6811	2471
duplication	1489	175	3	NA	NA	62	36	NA	NA
complex	365	NA	NA	NA	NA	1	1	NA	NA

[1] Zhan et al. (2011) [2] Stothard et al. (2011) [3] Bickhard et al. (2012) [4] Shin et al. (2014); # number of CNVs/CNVRs; number of #∩ overlapping CNVs/CNVRs between studies