DETECTION AND VALIDATION OF STRUCTURAL VARIATION IN CATTLE WHOLE-GENOME SEQUENCE DATA

L. Chen[1, 2, 3], A.J. Chamberlain[1,3], C.M. Reich[1, 3], H.D. Daetwyler[1, 2, 3] and B.J. Hayes[1, 2, 3]

[1]AgriBio, Centre for AgriBioscience, Biosciences Research, DEDJTR, 5 Ring Road, Bundoora,
VIC 3083, Australia
[2] La Trobe University, Victoria, Australia
[3] Dairy Futures Cooperative Research Centre, Victoria, Australia

SUMMARY

Several examples of structural variation (SV), or copy number variation (CNV) affecting traits exist in cattle. However the effect of SV on complex traits is largely unknown. The identification of SV suffers from high false positive and low overlapping rate when using different programs. We detected SV in dairy cattle whole-genome sequence from 308 Holsteins and 64 Jerseys with two SV detection programs - Breakdancer and Pindel. We constructed a set of validated SVs based on 28 individuals that were sequenced twice, and were transmitted sire to son. A total of 11,534candidate SVs covering 5.64 Mb were validated in the 28 twice-sequenced individuals, while 3.49 Mb and 0.67 Mb of SV were validated from Holstein and Jersey sire-son transmission.

INTRODUCTION

There are several categories of genome variation within a species. Single nucleotide polymorphisms (SNP) are the most frequent and have been widely utilized in association and genomic prediction studies. Another category is structural variation (SV) which refers to segments of 1 kilo bases (kb) to several mega bases (Mb) of deletions, duplications, inversions and translocations in the re-sequenced genome compared to a reference genome, of which copy number variation (CNV) only includes deletions and duplications.

In cattle, a number of studies have shown evidence that SVs spanning gene coding regions can affect a wide range of traits (Liu *et al.* 2010). In Angus cattle, 297 CNVs were found to be associated with parasite resistance or susceptibility (Hou *et al.* 2012). Recently a 660 kb deletion was found to be associated with fertility and milk production in Nordic red cattle (Kadri *et al.* 2014). In addition, SVs have been shown to be associated with the polled phenotype in cattle (Medugorac *et al.* 2012; Rothammer *et al.* 2014)

A number of genomic data types can be used to detect SV. PennCNV implements a hidden Markov model (HMM) to detect CNVs from SNP arrays (Wang *et al.* 2007). However, due to limited SNP density and high minor allele frequency of these SNP, the ability to identify rarer and/or smaller CNVs is limited. In addition, SNP chip methods cannot capture balanced SVs including inversions and translocations.

Whole-genome sequence data can potentially be used to recover the whole spectrum of SVs. Paired-end mapping (PEM) (Korbel *et al.* 2007), split read (SR) (Ye *et al.* 2009), read depth (RD) (Teo *et al.* 2012), and de novo assembly (Iqbal *et al.* 2012) are the current four basic strategies used to detect SVs from sequence data.

Here we detected SVs in whole-genome sequence data from Holstein and Jersey populations with a combination of Breakdancer (Chen *et al.* 2009) (PEM) and Pindel (Ye *et al.* 2009) (SR), combined with two novel validation strategies, to generate high quality SV sets. We also tested the hypothesis that highly conserved gene regions (between species) should have less SVs than in less conserved regions.

MATERIALS AND METHODS

Animal samples. The paired-end read whole-genome sequence data is described in (Daetwyler *et al.* 2014). A total of 308 Holstein and 64 Jersey were sequenced with Illumina sequencing platforms, with average coverage 10.76 and 10.92 respectively. All the short sequencing reads were then aligned to reference assembly UMD 3.1 with the Burrows-Wheeler Aligner (BWA). Our validation strategy included assessing how many SVs were detected in both replicates of a set of 28 Holstein individuals that were sequenced twice with different libraries, and whether we could observe sire-son transmission of the SV in 68 Holstein and 33 Jersey sire-son pairs.

Sequence population SV calls. We pooled the Holstein (not including twice-sequenced individuals) and Jersey populations and investigated the SV distribution differences between the two breeds. For each population, we first ran Breakdancer and Pindel to generate raw SV calls by each SV type (deletion, insertion, inversion and duplication). The default parameters were used for both programs. However, we enforced a threshold of a minimum of four supporting read pairs and observation in two individuals to classify higher quality SVs. We also filtered SVs that span chromosome gaps in the reference assembly. In the next step, we found the overlapping regions when merging the SV calls from Breakdancer and Pindel and considered these overlapped regions to be higher confidence SVs.

Validated SV calls. In the Holstein population, 28 individuals were sequenced twice. In theory for each individual the two sequences should convey exactly the same information. However due to random distribution of sequence reads, assembly error and different depth of coverage, the two sequences are not identical, and, thus, programs can report different SVs. We generated a high confidence SV set by only reporting SVs detected in both sequences. In addition, as most SVs should be inherited, we only report SVs that are inherited from sire to offspring. The validated sets were further compared between each other and with outputs from SNP chip.

Detecting SVs and CNVs from SNP chip genotype data. A total of 128 Holstein and 170 Jersey cattle were genotyped with the 800K HD SNP chip, which were afterwards converted to Log R Ratio (LRR) and B allele frequency (BAF) for further analysis. Individuals with standard deviation of LRR>0.35 and BAF >0.2 were discarded, as suggested by Wang *et al.* 2007. A total of 125 Holstein and 166 Jersey were kept after this filter. The genomic content (GC) model which incorporates the GC percentage information around each SNP was used to improve CNV outputs. SNP chip methods cannot detect inversions and therefore we eliminated inversion events when comparing to validated sets from sequence.

Conserved genes. To test the hypothesis that SV and CNV are less likely in genes that are highly conserved across species, 248 core eukaryotic genes were selected (Parra *et al.* 2007) that were likely to be found in a low number of paralogs in a wide range of species. We downloaded the protein file (fasta format) and put it into the BLAST program to search the most similar proteins and genes in cattle. The search results were further converted into coding nucleotides in bed format with chromosome, strand, start and end position that can be overlapped with our validated SV sets. We defined a minimum of 0.5% of the gene overlapped with validated SVs to be reported. A chi-squared test was performed to test whether these conserved genes contain less SVs than all the other reference genes downloaded from the UCSC genome browser.

RESULTS AND DISCUSSION

Population SV Calls. The overlapped region from the two programs dramatically shrunk the original SVs into a small set, as only about 2-10% of the calls (ranging from 25 to 44,412 bp) were kept after merging (p-value = 6.38448e-20). Overall, Holstein had more SV calls than Jersey, which may mainly be due to a larger sample size for Holstein. After filtering SVs less than 25 bp, the median length of deletion, insertion, inversion and duplication for Holstein was 1123, 72, 2533 and 857 and for Jersey was 1152, 0, 1337 and 1014 bp, respectively. Table 1 shows the total covered length of SVs shared by the two populations. A total of 4.62Mb SV events were detected in both population, occupying 16.89% in Holstein (27.36 Mb) and 53.47% in Jersey (8.64 Mb), of which deletions and duplications had a relatively high percentage.

Table 1. Covered region of SVs shared by Holstein and Jersey population

SV Covered Region Mb	DEL	INS	INV	DUP	Total
Holstein	8.49	0.639	13.84	4.40	27.36
Jersey	5.22	0	1.05	2.37	8.64
OVERLAP	3.18	0	0.22	1.23	4.62

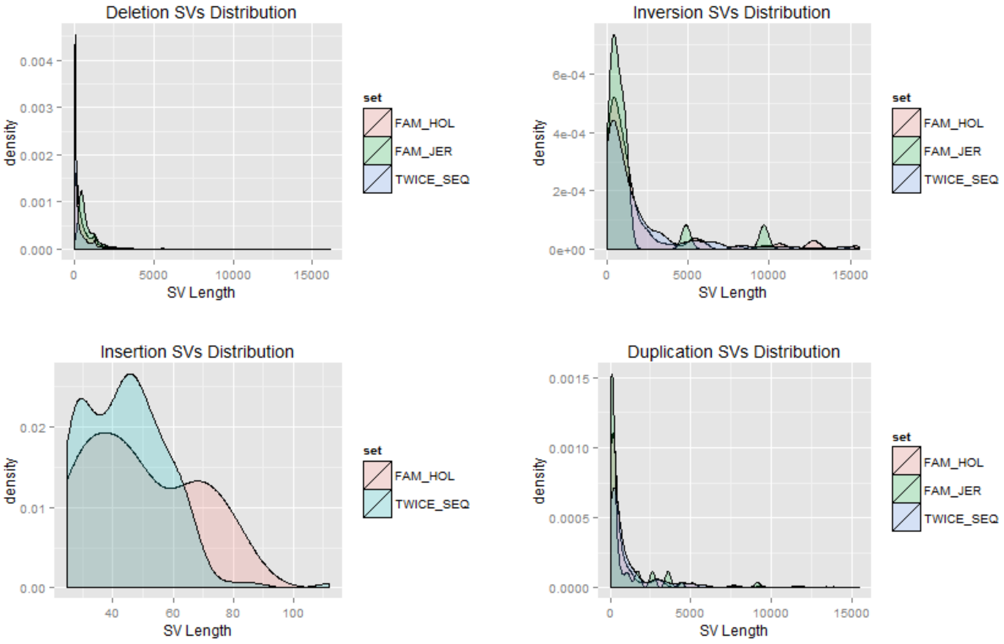

Figure 1. Size range distribution of four type of SVs in twice sequenced, Holstein and Jersey family validated sets.

Validated SV Calls. We generated three sets of validated SV calls: twice-sequenced, Holstein and Jersey family-level validated SV sets. A total of 5.64 Mb were validated from 28 twice-sequenced individuals, while 3.49 Mb and 0.67 Mb SVs were found in Holstein and Jersey families. We also compared the Holstein twice-sequenced set and Holstein family set. Overall 82.0% SVs in Holstein family were also found in the twice-sequenced set. This result illustrates less false positives and thus higher confidence SVs compared to population calls. Figure 1 demonstrates that the size distribution of SVs is similar across these validated sets. Most deletions and insertions are less than 100 bp; a large number of inversions are around 900 bp while duplications are around 350 bp. For inversions in Jersey family there are two small peaks at 5kb and 10 kb respectively. When looking into the sires with multiple sons, a total of about 80 kb deletions and 90 kb duplications on BTA1 were shared in Holstein and 27 kb inversions on BTA11 and 16 kb inversions and duplications on BTA14 in Jersey, suggesting these areas could be common CNV regions in both breeds.

The 800K SNP chip data results indicated a total of 2224 CNVs covering 250.5 Mb in Holstein (227 Mb deletions and 23.3 Mb insertions) and 2976 CNVs covering 357.4 Mb in Jersey (333 Mb deletions and 24.3 Mb insertions). As SNP platform resolution is limited, PennCNV cannot detect very small events. Therefore, we only compared this result with SVs larger than 5 kb detected from the sequence data. As a result, 12.33% deletions and 11.59% duplications in validated sets were also found in Holstein 800K outputs, while 14.95% deletions and 0% insertions overlapped in Jersey.

Conserved Genes Test. We found 293 identical genes according to core gene sets after searching by BLAST. Overall, there were not many conserved genes in our reported SV areas. Within the 293 genes only five genes were found in Holstein family, one in Jersey (*ETFDH* with 152 bps overlapped) and seven in twice-sequenced one. Among these genes, most harboured deletions, while two and one contained inversions and a duplication, respectively. All the five genes from the Holstein family set were confirmed in the twice-sequenced set. Compared to all the other reference sequence genes, however, no significant evidence was found to support that conserved genes regions contained less structural variants than all others (p-value >0.7). Our validated SV sets will assist genetic research in cattle such as genomic prediction and genome-wide association studies.

ACKNOWLEDGEMENTS

The authors thank all members of the 1000 Bull Genomes Consortium for provision of data.

REFERENCES

Chen K., Wallis J.W., McLellan M.D., Larson D.E., Kalicki J.M., *et al.*(2009) *Nat Methods* **6**(9): 677.

Daetwyler H.D., Capitan A., Pausch H., Stothard P., van Binsbergen R., *et al.* (2014) *Nat Genet* **46**(8): 858.

Hou Y., Liu G.E., Bickhart D.M., Matukumalli L.K., Li C., *et al.* (2012) *Funct Integr Genomics* **12**(1): 81.

Iqbal Z., Caccamo M., Turner I., Flicek P., McVean G., (2012) *Nat Genet* **44**(2): 226.

Kadri N.K., Sahana G., Charlier C., Iso-Touru T., Guldbrandtsen B., *et al.* (2014). *PLoS Genet* **10**(1): e1004049.

Korbel J.O., Urban A.E., Affourtit J.P., Godwin B., Grubert F., *et al.* (2007) *Science* **318**(5849): 420.

Liu G. E., Hou Y., Zhu B., Cardone M.F., Jiang L., *et al.* (2010) *Genome Res* **20**(5): 693.

Medugorac I., Seichter D., Graf A., Ingolf R., Helmut B., *et al.* (2012) *PLoS One* **7**(6): e39477.

Parra G., Bradnam K. and Korf I. (2007) *Bioinformatics* 23(9):1061

Rothammer S., Capitan A., Mullaart E., Seichter D., Russ I., *et al.*(2014) *Genet Sel Evol* **46**: 44.

Teo S. M., Pawitan Y., Ku C.S., Chia K.S. and Salim A. (2012) *Bioinformatics* **28**(21): 2711.

Wang K., Li M., Hadley D., Liu R., Glessner J., *et al.* (2007) *Genome Res* **17**(11): 1665.

Ye K., Schulz M. H., Long Q., Apweiler R. and Ning Z.M. (2009) *Bioinformatics* **25**(21): 2865.