Copy Number Variation in the Ovine Genome

G.M. Payne[*], J.C. McEwan[*], J. Kijas[‡], R. Brauning[*], M.A. Black[#], R. McCulloch[‡] and M.E. Goddard[†]

Introduction

With the advent of new sequencing technologies, large amounts of ovine DNA sequence has become available and an Illumina 50K ovine SNP chip developed. Much research is being done on utilising this as a tool to enhance breeding value (BV) prediction. However, currently there are some major limitations with this technology, including the reliance on the relationship between SNPs and trait holding in target populations. This is dependent on the level of LD between SNP and causal polymorphism, which can vary between populations. Thus, the accuracy of BV prediction is a function of the level of LD and the genetic distance between the reference and target populations.

In addition to SNPs there are other DNA variants that underlie phenotypic variation, such as copy number variants (CNV). CNVs are large stretches of DNA (>1000 base pairs) that are duplicated or deleted in the genome (Cooper et al, 2007). A study by Stranger et al (2007) showed that CNVs may explain up to 20% of variation in gene expression levels, with little overlap with SNP associations. CNVs have been shown to be associated with complex traits, such as HIV susceptibility in humans (Gonzalez et al, 2005) and can be the direct cause of the disease/trait, as is the case with agouti coat colour in sheep (Norris et al, 2008).

If indeed CNVs are responsible for an untapped source of phenotypic variation in complex traits, their formal incorporation into BV predictions can only increase the accuracy of prediction. This study focuses on detecting CNVs in the ovine genome and validating them across platforms.

Material and methods

Three platforms were used to detect CNVs in the ovine genome: 384K CGH NimbleGen array; Illumina 50K ovine SNP chip; and raw Roche 454FLX sequence reads. The CGH array was the main method for detecting CNVs, with SNP and sequence data used to confirm results where possible. As platforms were based on different ovine genome builds, the CGH probes, SNP probes and 454 raw reads were mapped onto the cattle genome (UMD3) to provide a common framework for comparing results.

[*] Invermay Agricultural Centre, AgResearch, Private Bag 50034, Mosgiel 9053, New Zealand
[†] Department of Primary Industries and University of Melbourne, Australia
[#] University of Otago, P.O. Box 56, Dunedin 9054, New Zealand
[‡] CSIRO Livestock Industries, Queensland Bioscience Precinct, QLD, Australia

Animal resources included 6 sheep (Awassi, Merino, Poll Dorset, Romney, Scottish Blackface and Texel), sequenced to 0.5X coverage each and run on both the 384K CGH and 50K SNP arrays (less one animal that showed contamination on the 50K SNP chip); and 2 trios (sire, dam and progeny) – 1 Merino and 1 Romney – assayed with the 384K CGH and 50K SNP arrays.

384K CGH array

Probes were designed using the unique fraction of the sheep genome (BTA_OAR_ver.2) with an average spacing of 1 probe every 6000bps. Roche-NimbleGen SegMNT software (BTA_OAR_ver.2 probe positions) was used to detect CNVs in this dataset. SegMNT calls regions of DNA where two or more consecutive probes show higher/lower dye intensities in the sample compared to the reference animal, measured as $\log_2$ratios. The sequenced Poll Dorset ewe was used as the reference animal. Outputs from the software were restricted to $\log_2$ratios < -0.25 or > 0.25. Trios were used to validate duplicated/deleted regions –a region seen as duplicated/deleted in one or both parents and the progeny was considered to be a true CNV. Data was restricted further to include only simple CNV loci (same start and stop position in all trio members in which the CNV was observed). Probe UMD3 positions were obtained and DNAcopy (Olshen, 2004) was used to confirm CNVs on UMD3.

50K ovine SNP chip

Of the validated CNV regions (from the 384K data), 3 had corresponding SNPs on the 50K chip. These SNPs appeared to have many no calls (NCs) across the 19 animals, which likely relates to the Illumina GenomeStudio software’s inability to cluster an individual’s genotype due to aberrant intensities of the SNP alleles. This could be due to duplication or deletion of the surrounding sequence so that more or less than the usual 3 genotypes were apparent. Chi-squared tests were performed to determine if there were significantly more NCs at CNV loci compared to “normal” loci.

454 sequence

Raw 454 sequence reads with UMD3 positions for the 6 sequenced animals were available. Reads were portioned into 50kb bins excluding repetitive sequence, on a per individual basis. Read depth at each base pair was determined and the average depth reported for each bin. Average read size (240bps) was used to estimate the number of reads per bin ((avg depth*bin size)/avg read size). The average number of reads/bin (and standard deviation) was determined on a per chromosome basis for each individual. The average number of reads/bin in regions corresponding to the CNV loci detected in the 384K CGH data was calculated on an individual basis. Regions with average reads per bin greater/less than the chromosomal average +/- two standard deviations were considered to represent duplications/deletions respectively. Regions with average reads/bin counts falling within this confidence range were considered to be normal. The number of animals concordant between sequence analysis and 384K CGH data was determined. An animal was considered concordant if it was consistently duplicated/deleted/normal in both datasets. The Poll Dorset animal was not included in these comparisons, as it was a reference animal on the 384K CGH platform and hence results are expected to differ between platforms for this animal.

Results and discussion

In total, SegMNT called 5,081 CNV loci, of which 14 were validated using our criteria and genome build BTA_OAR_ver.2. Based on inheritance, maximum likelihood prediction estimates the error rate of the SegMNT software is ~65-70%. This suggests ~ 1,500 of the 5,081 CNV loci were true CNVs but we could only validate ~1% using our criteria. Validated CNVs were, on average, much shorter and had $\log_2$ratios 3-6 times the magnitude of non-validated CNVs (Table 1). However, the range of both CNV length and $\log_2$ratios overlapped between validated and non-validated CNVs, which means these characteristics cannot be used alone to detect true CNVs in the SegMNT output. Given this and the high error rate of the SegMNT software, trios provide an invaluable means of validating CNVs on this platform. Five of the 14 validated CNVs could be mapped on to UMD3 autosomal chromosomes as distinct loci for across platform comparisons.

Chi-squared tests of 50K Illumina results showed there were significantly more no calls for SNPs in regions corresponding to CNV loci identified with the CGH data (Table 2). Concordance between sequence and CGH results varied across CNVs. One of the CNVs with no concordant animals across these platforms corresponded to the published agouti duplication. The duplication was present in all 6 animals, hence did not show up in the CGH data as these animals did not differ from the reference Poll Dorset. This highlights that caution must be used when comparing results across platforms as results depend on the reference animal used and the comparisons made. Disparity between the analyses may also be the result of the power of the sequence analysis to detect CNVs, as it depends on CNV size, bin size and coverage. Here, we had a low coverage of 0.5X for each animal; therefore the power to detect small CNVs on this platform was also low. Nevertheless, the other 4 CNV loci found in the CGH data were confirmed by the 454 sequence data for some animals (Table 2).

Conclusion

In summary, the use of trios to detect CNVs initially is vitally important, as the characteristics of true CNVs overlap with potential false positives in our data. While CNV detection methods lack power to detect true CNVs when used alone, their use in combination with other platforms provides a means to validate true CNVs. Future work will include genotyping trios on denser CGH arrays to detect smaller CNVs and sequencing animals for imputation of CNV genotypes for association studies.

Table 1: Characteristics of the 14 validated CNV loci versus non-validated CNV loci in all animals. CNV calls that met criteria for validation at each of the trio validated loci were included in the validated dataset. CNV calls not fitting criteria for trio validated loci and CNV calls not corresponding to trio validated loci are included in the non-validated dataset.

	Duplications		Deletions	
	Validated	Non-validated	Validated	Non-validated
Num CNV calls	15	6,000	42	5,916
Length (bp) -mean	36,409	2,996,841	251,966	2,466,137
Length (bp) -min	390	165	1,710	165
Length (bp) -max	236,042	161,040,387	1,655,902	161,040,387
Log2ratio –mean	0.61	0.10	-0.41	-0.13
Log2ratio –min	0.30	0	-0.80	-1.89
Log2ratio –max	1.08	2.69	-0.25	0

Table 2: Cross platform validation of the 5 CNV loci detected in the 384K CGH data. NC = "no call" for SNP genotype.

384K CGH CNV loci	Illumina 50K ovine SNP chip	Sequence analysis – number of concordant animals (max=5)
1	No SNP matches	2
2	Significant number NCs at 0.001 level (chi-squared)	5
3	Significant number NCs at 0.001 level (chi-squared)	0
4	Significant number NCs at 0.001 level (chi-squared)	2
5	No SNP matches	4

References

Cooper, G.M., Nickerson, D.A., and Eichler, E.E. (2007). *Nat Genet.*, 39:S22-9.

Gonzalez, E., Kulkarni, H., Bolivar, H., *et al* (2005). *Science,* 307(5714): 1434-40.

Norris, B.J., and Whan, V.A. (2008). *Genome Res*, 18(8):1282-93.

Olshen, A.B., Venkatraman, E.S., Lucito, R., *et al* (2004). *Biostatistics*, 5:557-572.

Stranger, B.E., Forrest, M.S., Dunning, M., *et al* (2007). *Science*, 315(5813):848-853.