BUILDING A DEER SNP CHIP

M.J. Bixley, J.F. Ward, R. Brauning, J.A. Archer and P.J. Fisher

AgResearch, Invermay Agricultural Centre, PB 50034, Mosgiel, New Zealand

SUMMARY

Advances in genomic sequencing programmes in livestock species such as cattle and sheep have enabled the building and application of large SNP (single nucleotide polymorphism) chips containing more than 50,000 markers. Applications include improved parentage and pedigree assignment as well as more accurate diversity and breed composition analysis. Further, genomewide selection (GWS) has been used to predict performance with the potential of increasing genetic gain. Other livestock species, including deer do not have a large scale genomic sequence, nor do they have other adequate supporting tools to enable trait to marker associations to be established. We have produced a reduced representational sequence of >160 million base pairs (Mbp), of which we mapped 44 Mbp to unique positions on the bovine genome. From this we selected 768 SNPs to be included in a Golden Gate (Illumina^{TM}) SNP chip. Further, we have assembled a mapping pedigree in order to quality control check these and other SNPs and to produce a genetic map. This mapping population will also be used to assess recombination rates and to reorder the deer sequence from bovine physical order to deer order. Other immediate outputs from this SNP chip will be new parentage assignment and breed composition panels. And we will investigate whether the chip will be informative for assessing within vs. across farm LD.

INTRODUCTION

This last year has seen the international dairy communities, including the dairy industry in New Zealand, adopting genome wide selection (Harris and Montgomery 2009). It is expected that this will lead to 50-70% greater genetic gain than was previously possible whilst maintaining flexibility and a multi-trait interest. One reason why this technology is valuable is because the generation interval for proving heavily used sires is dramatically reduced. Additionally there is the potential for early prediction and selection of commercial animals. Similarly, the sheep industry has developed a 60K ovine SNP chip and multiple research groups around the world have taken advantage of this new tool. The deer industry in New Zealand, with over a million animals in production, is keen to explore the options of improved genetic selection for desirable traits. We have begun phenotyping animal collections for seasonality (with conception date scanning), Johnes disease and some carcass and meat traits. To generate phenotype-marker associations, the upgrading of genomic tools (including a genetic mapping herd and larger marker sets) is also needed. The DNA from our previous (Pere-David x red) deer mapping resource (Slate et al., 2002) is nearly depleted and is not representative of the commercially important sub-species. A new 440-deer genetic mapping herd (halfsib and fullsib 3-generation families) was sourced but will not be discussed further in this paper. Here, we describe the process of sequencing for identification of SNPs and the use of bioinformatics to select markers for inclusion in a 768-SNP chip.

MATERIALS AND METHODS

Construction of Reduced Representation Libraries. The following procedure which selects a defined portion of the genome for sequencing is an adaptation of the method of Van Tassell *et al.* (2008). Breed composition using STRUCTURE^{TM} (Pritchard *et al.* 2000) was determined for Eastern Red Deer, Western Red Deer populations within New Zealand. Here, these populations will be referred to as different "breeds" although the term "sub-species" may be more appropriate.

From this analysis, pools of 25 animals, summarised in Table 1, were created by selecting those that above 70% for their respective breed. Wapiti were selected from animals that met the registration criteria of the New Zealand Elk and Wapiti Society (Asher *et al.* 2005).

Table 1. Summary of the 3 Deer breed-pools

Pool	n	Breed composition (%)	SD
Eastern	25	76.7	3.7
Western	25	85.7	3.0
Wapiti	25	N/A	

DNA was isolated from blood (Montgomery and Sise 1990) and quantified; 4µg of each sample was pooled (~100µg total). The pools were digested overnight with 2U/µg of *Rsa*I (New England Biolabs, Beverley, MA, USA) to ensure complete digestion of the genome and create blunt end fragments. Aliquots (150µl) of the digest were fractionated in 1.5% agarose gels (TBE Buffer) at 120V for 3 hours, then stained with Ethidium Bromide. Fragments in the range of 350 bp to 600 bp were excised from the gel (Figure 1) and purified using a Gel Purification kit (Qiagen Inc., Montgomery County, MD, USA).

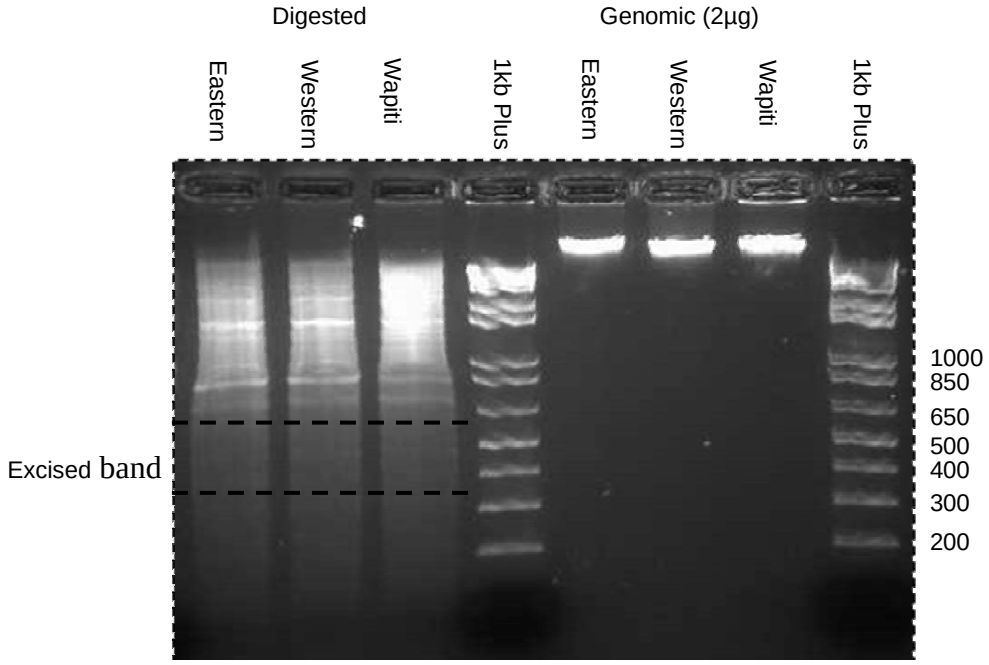

Figure 1. Digestion of pooled genomic DNA and region excised from gel

DNA Pyrosequencing. Each pool was run on a separate sequence plate section to distinguish which sequence reads belonged to which sub-species. Fragment libraries were sequenced at Otago University Sequencing facility using Roche-454 GS FLX Titanium sequence technology. This platform was used because it promised to yield long sequence reads (300-500 base pairs in length).

Genome Assembly. Pipelines originally built for assembling the ovine genomic sequence (http://www.sheephapmap.org/publications.php) were adapted for data cleaning. These procedures included masking of repetitive sequences (reads with <150 bp repeat-free DNA were removed from further analysis). The deer sequence reads were assembled, ordered and orientated using the bovine framework (Btau4 build). The assembly used the Newbler algorithm provided with the sequencing system. Alignment to the bovine genome was carried out using NCBI's BLAST tools (http://blast.ncbi.nlm.nih.gov/Blast.cgi). A GBrowse track was created of this "bovine-order deer genome", placing contigs and singleton reads on bovine chromosome and position (bp) from the chromosome start. Using this assembly, SNPs were identified and characterised based on the deer sequence reads taking into account the breed origin, sequence depth, and variability.

Deer Genome Annotation. SNP-based variants were observed and assigned each to a GBrowse track. SNP classification was based at two levels: (i) the quality of the SNP itself and (ii) the SNPs and flanking sequences were ranked technically for quality as (Golden Gate) SNP chip markers.

(i) Identification of SNP-based sequence variants

We identified variants for each sequence pool, forming six "allele x breed" categories per contig: allele 1 or 2 for each of the three (Western red, Eastern red or Wapiti) sequence runs. We selected putative SNPs from contigs where at least four reads were present but excluded SNPs where more than five reads were observed in two or more categories. We classified SNPs as "class A" if both alleles were present in each of two or more breed-pools or "class B" if we observed two alleles in one pool and if at least one other breed-pool had two or more reads of either allele.

(ii) SNP sequences and marker generation

SNP-containing sequences were sent to Illumina only if at least 50 bp of SNP-free flanking DNA was present on both sides of the SNP of interest. Further, at least 20 bp of the DNA immediately adjacent to either side of the targeted SNP had to be completely free of sequence variation.

Validation of SNPs. A panel of 70 SNPs were selected from across the genome and they have been assembled into 3 Sequenom™ mass-spectrometry (SNP) multiplexes; they will be analysed (against a 48-deer validation set) to confirm that the SNP sequence predictions are accurate.

RESULTS AND DISCUSSION

A total of 207 Mbp of raw genome sequence was produced (average read length = 267 bp, Table 2), but the distribution was bimodal (not shown). A sequencing artefact might have caused the smaller of the modes to peak at ~80 bp resulting in 41 Mbp of the reads (20%) being less than 120 bp. The largest modal peak was 340 bp or greater for all three sequence pools (Table 2); this long read length is very good for marker generation even after masking for repetitive motifs.

Table 2. Summary of the sequencing data

Breed pool	Reads (1000)	Million base pairs (Mbp)	Mean Read Length (bp)	Mode Read Length (bp)
Eastern Red	235	65.5	279	340
Western Red	356	95.8	269	340
Wapiti	179	45.5	254	380

A total of 43.9 Mbp was mapped to unique locations on Btau4 (where it was $>e^{-20}$ superior to the next best match) using BLAST tools. There were 127,266 singletons and 25,592 contigs (the term "contig" here refers to loci with more than one overlapping sequence read). The contigs'

upper size limits were seldom above 600 bp as the contigs were derived from 350-600 bp fragments that began or ended at the same enzyme cut-site. The average read depth was 1.44 (for all three breed-pools). From the exponential distribution (not shown) we calculated that >99% of the loci were expected to have fewer than six reads per allele x breed category. Loci with more than five reads in two or more categories were assumed to represent repetitive loci and were excluded. However if only one group was represented by more than five reads, we postulated that this might be due either to chance or to a run-specific sequence artefact; the contig was kept. Whilst the class A SNPs may be more likely to be genuine (i.e. not a sequence error), we also required class B SNPs because we intended to utilise breed-specific markers. We ended up with 2,170 class A or B SNPs, located on 830 contigs. The SNPs and flanking sequences were sent to Illumina for creation of a cervine 768-SNP chip. We will finish validating a 70-SNP validation set prior to final chip purchase; this testing is underway.

Our initial expectation was that we would be able to select 768 SNPs from >2,000 different contigs using a single 454 GS FLX (Titanium) sequencing run assuming a sequence yield of 350 Mbp. Further, we assumed that we could remove >97% of each genome from the gel. We were conservative in band excision, opting for a wider gel width than originally desired to ensure that the three pools contained similar loci. It was difficult to precisely determine the proportion of each genome and the sequence output was insufficient to estimate this subsequently, so it may be that > 3% was excised. Therefore it remains possible that a single Titanium sequencing run could yield >2,000 unique SNP-containing contigs (each represented by four or more reads), provided that measures are taken to increase the overall read depth from 1.44 to above 2.5.

The 768 SNPs are aligned to all 30 bovine chromosomes but their distribution is not uniform; the lowest density is on Bta24 (6 SNPs) and the highest density is on Bta28 (44 SNPs). They will be suitable for for parentage testing and breed composition panels as well as for diversity studies. We will also use the chip to genotype our new genetic mapping population for quality control of the SNPs, to determine linkage and to enable the reordering of the markers into deer-chromosomal order. And we will combine the SNPs with other markers to determine whether we can assess LD and/or haplotype blocks among samples sourced from multiple families and farms. The non-uniform SNP distribution might improve our chances of detecting LD because, for example there are 20 loci where >10 SNPs are present in less than 10 Mbp of bovine homologous sequence.

We have begun a large scale genomic sequence and SNP discovery programme to progress towards GWS in deer; whilst we expect to reveal many more than 50,000 SNPs from this project, we will optimise future work based on information and resources gathered in the described project.

ACKNOWLEDGEMENTS

This work was funded from the Foundation for Research Science and Technology, DEEResearch Ltd and Livestock Improvement Ltd.

REFERENCES

Asher, G.W., Archer, J.A., Scott, I.C., O'Neill, K.T., Ward, J.F. and Littlejohn, R.P. (2005) *Animal Reproduction Science* **90**:287.

Harris, B.L. and Montgomery, W.A. (2009) Interbull bulletin no. 39, www.interbull.org/bulletins

Montgomery G.W. & Sise J.A. (1990) *New Zealand Journal of Agricultural Research* **33**:437

Pritchard J.K., Stephens M., Donnelly P. (2000) *Genetics* **155**:945.

Slate J., Van Stijn T., Anderson R., McEwan K., Maqbool N., Mathias H., Bixley M., Stevens D., Molenaar A., Beever J., Galloway S. and Tate M. (2002) *Genetics* **160**:1587.

Van Tassell, C.P., Smith, T.P.L., Matukumalli, L.K., Taylor, J.F., Schnabel, R.D., Lawley, C.T., Haudenschild, C.D., Moore, S.S., Warren, W.C. and Sonstegard, T.S. (2008) *Nature Methods* **5:** 247.