SHEEPGENOMICS AND THE INTERNATIONAL SHEEP GENOMICS CONSORTIUM

Hutton Oddy[1], Brian Dalrymple[2], John McEwan[3], James Kijas[2], Ben Hayes[5], Julius van der Werf[1], David Emery[4], Phil Hynd[6], Terry Longhurst[7], Troy Fischer[8], Duncan Ferguson[7] , Rob Forage[7,8], Noelle Cockett[9] and Frank Nicholas[4]

[1.] University of New England, Armidale, NSW 2351 [2]. CSIRO Livestock Industries, St Lucia, QLD 4067 [3.] AgResearch, Invermay, NZ [4.] University of Sydney, Sydney NSW 2006 [5.] Department of Primary Industries Victoria, Attwood, Vic 3049, [6.] University of Adelaide, Roseworthy, SA 5371 [7.] Meat & Livestock Australia, North Sydney, NSW 2160 [8.] Australian Wool Innovation Limited, Sydney, NSW 2000 [9.] Utah State University, Logan, Utah 84322-4800 USA

SUMMARY

SheepGenomics is a strategic investment by Meat & Livestock Australia and Australian Wool Innovation Limited and 11 Australian and New Zealand research organizations to deliver tangible outcomes from genomics research to the sheep industry. The overall strategy of SheepGenomics is to "find useful genes and put them to work". To achieve this has required the development of significant resources including a half-sib design mapping flock using over 16 industry sires and 4 sires from previous QTL studies to generate and extensively phenotype from 200 to 400 progeny / sire. The original intent was to genotype the progeny using a limited number of microsatellite markers and then fine-map selected progeny to discover genes for use in industry breeding programs and for further study. Development of genomic resources for sheep has proceeded to the stage where it is now becoming practical to genotype the progeny of the SheepGenomics flock with tens of thousands of SNPs and use the outputs to derive genome selection derived breeding values (in addition to many new QTL). This change in strategy and deliverables would not have been possible without a substantial contribution from the International Sheep Genomics Consortium (ISGC) to develop sheep-specific genomic information in the public domain. The ISGC has been instrumental in developing a sheep BAC library, its end sequencing and alignment against other genomes. This has resulted in development of a virtual sheep genome, which in turn underpins current activities to discover and use tens of thousands of ordered sheep SNPs.

INTRODUCTION

SheepGenomics is a strategic investment to translate the technological and scientific insights obtained from the human, bovine and other mammalian genome projects into tangible outcomes for the Australian sheep industry (Oddy *et al.* 2005). SheepGenomics was seen as a vehicle to address previously intractable sheep industry problems by discovering and applying new insights into mechanism at the genetic and molecular level. SheepGenomics targets include improvements in host resistance to parasites (including new insights into gut physiology and immunology) and production and product quality of meat and wool. It was anticipated that development of tools required to conduct these studies would enable further applications of knowledge of genes and their functions to be explored, and eventually used in the sheep industries.

SheepGenomics commenced in late 2003 and is a work in progress. Results achieved to date include:

- Over 5000 phenotyped progeny of selected sires for subsequent genotyping, gene discovery and estimation of breeding values using genomic selection. More than 250 separate traits have been measured on over 4600 progeny at Falkiner Memorial Field Station and more than 400 at CSIRO Chiswick.
- Development of tools and reagents for description of new traits associated with muscling, wool growth and most importantly, resistance to helminth infection.
- Establishment of business models for delivery of gene marker information together with breeding values through Sheep Genetics Australia
- Identification of gene pathways for muscle development, wool growth and host-resistance to internal parasites through gene expression and proteomics analyses
- Identification of a number of DNA marker tests for muscling and fatness traits and evaluation of these in industry flocks
- Development of sheep specific genomic resources including a virtual sheep genome (www.livestockgenomics.csiro.au/vsheep) in conjunction with the ISGC

Establishment of a sheep focused bioinformatics/biostatistics capacity throughout the program
Resources for genomics research in sheep are limited compared to those available for man, mouse, dog and cattle. A loose international coalition of those interested in applying genomics technologies for sheep improvement, and / or in generating public-domain genomics resources, has been formalized somewhat into the International Sheep Genomics Consortium (ISGC) established over the last two years. This consortium is currently developing further resources (www.sheephapmap.org). This paper will briefly describe progress with development of genomic resources for sheep from the ISGC, and the implications of this resource development for the research strategies pursued in SheepGenomics.

ISGC AND THE DEVELOPMENT OF OVINE RESOURCES

Since our report to AAABG in 2005 (Oddy *et al.*, 2005) there has been considerable progress in development of sheep genomic resources. End sequencing of the majority of the clones in the previously described CHORI-243 ovine BAC library (http://bacpac.chori.org/library.php?id=162) has been completed, using funds from the USDA, Utah State University and SheepGenomics. http://www.tigr.org/mammalianGenomics.shtml The BAC end sequencing was originally carried out to allow positioning of ovine BACs against other related genomes to assist in the selection of BACs containing a particular gene(s) of interest for further study. However, it was soon realized that a virtual sheep genome could be constructed from the ordered end sequences of the BACs. Dalrymple *et al* (2007) developed the virtual sheep genome as follows. The sheep BAC-end sequences were independently mapped to the cow (bosTau2), dog (canFam2) and human (hg17) genome sequences. All BAC-end position information from the three different genome was combined together on the framework of the human, genome, and sheep BACs with their two ends correctly orientated and separated by the correct gene distance were identified. This identified regions of local conserved synteny between the sheep and human genome. Where the sheep BACs overlapped, longer regions of conserved synteny were constructed, resulting in a set of 1172 sheep BAC-Comparative Genome Contigs (BAC-CGCs) covering most of the predicted length of the sheep genome. The sheep linkage map v4.6 (http://rubens.its.unimelb.edu.au/~jillm/jill.htm) was then used to order and orientate the

BAC-CGCs into their most likely order in the sheep genome. To enable us to utilize the extensive annotation of the human genome, the UCSC utility http://genome.ucsc.edu/cgi-bin/hgLiftOver was used to transfer the human genome features to the virtual sheep genome. Sheep researchers now have access to a comprehensive prediction of gene order in the sheep genome and the framework for overlaying other genetic and sequence-based information, such as QTLs and SNPs.In this process ~84,000 (48%) of the sheep BACs were mapped with high confidence (both BAC-end sequences positioned and in the tail-to-tail arrangement) and an additional ~55,000 BACs were mapped with lower confidence (one BAC-end sequence positioned). The virtual sheep genome therefore also provides a prediction of the relative order in the sheep genome of a large number of short segments of actual sheep sequence, the BAC-end sequences. These positioned segments of sheep sequence provide a set of sequences suitable for the identification of sheep SNPs by targeted resequencing of sheep BAC end-sequences in panels of sheep.

The virtual sheep genome also provides a starting point for assembly of sheep genomic sequence, and provides the opportunity to use less costly technologies for subsequent sheep genome sequencing and SNP discovery. The ISGC is actively seeking funds to sequence the sheep genome. In the meantime, a major discovery project funded in part with Australian Government International Science Linkage Grant funds is underway to develop ovine SNP resources for high-density mapping. Polymorphic positions (SNP) in the bovine genome are unlikely to be polymorphic (and thereby useful) in sheep. Available estimates suggest less that 1% bovine SNPs convert to ovine SNPs (S. Moore pers comm). A pilot study using Sanger sequencing of regions of the genome corresponding to selected BAC ends positioned on the virtual sheep genome and some ESTs in a panel of diverse sheep and in pools of DNA from diverse sheep commenced in mid-2006. Results of that study (Table 1) indicate that, if we were to continue with the originally planned strategy, approximately 15,000 useable SNPs would be discovered with the available budget. Although this is substantially less than the planned 80,000 it is still more than the ~10,000 SNPs we originally envisaged would be required for SNP genotyping (based on the simulations of Meuwissen *et al.,* 2001). However, as SNP genotyping experiments have been conducted in other species, the extent of LD has become better known, and more rather than fewer SNPs are required for initial research studies, particularly those that use animals of different breeds. The extent of LD in sheep is not well known, although it is anticipated that sheep will show less LD than cattle. Initial estimates of LD in Romney sheep (Roberts-Thompson *et al.,* 2005) indicate that r^2=0.25 at 114kb, and suggest that to obtain r^2>0.5 there is a need for at least 60,000 SNPs (over a 3Gb genome). During the course of the work, alternative sequencing technologies based on automated pyrosequencing (454 Technologies http://www.454.com/enabling-technology/the-system.asp, Margulies *et al.,* 2005) have been developed. When released just 18 months ago, 454 technology could deliver read lengths of 100 bases. Improvements in the process now make it possible for 230+ base reads to be routinely achieved, with the promise of longer reads within 6 months. Changes in technology lead to changes in approach to resource development, but usually do not lead to major changes in research strategy. For example, to develop the required number of SNPs we are now planning to obtain whole genome shotgun sequence from 6 individuals using 454 technology (Table 1) to generate approximately 3x coverage of the sheep genome. The anticipated outcome of this work is >170k "useable" SNPs by the end of 2007 and a research "chip" with at least 50k SNPs in early 2008. The increase in SNPs is not linearly related to the number of individual sheep sequenced: above a threshold of 4 sheep, there is a rapid increase in power to distinguish real SNPs from sequencing artifacts.

Table 1. Comparison of the options for discovery of useable ovine SNPs. Options 1a through 1d are results from ISGC / ISL grant pilot resequencing. Options 2 through 5 have been generated by simulation based on experience with both Sanger and 454 (GS20) sequencing

		Probable SNPs Detected	Positioned SNPs	Useable SNPs[1]	Cost[2]	Cost/ useable SNP	Useable and informative SNPs[1] with MAF[4] >0.2
1a	Sanger pilot BES bidirectional 9 breeds at ~1X	6,067	6,067	1,548			1,068
1b	Sanger pilot BES MEP pool	815	815	333			239
1c	Sanger pilot ESTs	413	413	153			17
1d	Sanger pilot total			1,701	217k	~A$128	1,085
2	Sanger reseq BES unidirectional	~215k	~215k	~55k	3,875k	~A$70	~38k
2a	Sanger reseq BES pooled unidirectional	~29k	~29k	~12k	1,730k	~A$144	~8k
3	Pooled BACs 454[3]	~477k	~286k	~180k	2,740k	~A$15[3]	~44k
4	4 breeds at 0.5X each 454	~160k	~95k	~60k	1,200k	~A$20	~60k
5	6 breeds at 0.5X each 454	~446k	~268k	~170k	1,800k	~A$10	~170k

[1]For targeted resequencing number of targets containing at least one SNP likely to convert to a SNP assay on the Illumina platform

[2]Based on known costs (Sanger) or best estimate (454).

[3]lower conversion rate of SNPs than other methods of discovery due to the inclusion of only 4 chromosomes.

[4]Minor allele frequency

IMPLICATIONS FOR RESEARCH STRATEGY IN SHEEPGENOMICS

The mantra that drives SheepGenomics is “find useful genes and put them to work”. This is derived from the two objectives of the program: to discover new genes for important and hard-to-measure traits, and to use specific knowledge of the function of some of these genes to develop specific products (principally diagnostics and modulators). The experimental strategy required to underpin these objectives is to identify causative mutations of at least moderate effect and validate the effect of these mutations in another population. This information then provides a basis for experiments that can be replicated as required to understand function of the genes, and to develop materials and methods to mimic that function.

With such objectives in mind it is not surprising that the original intent was to collate all the QTL studies conducted in sheep and find common genetic regions for further study. However, in 2002 a review of resources (M. Goddard, J. Henshall and J. van der Werf personal communication) indicated that the power of extant QTL studies was low relative to that required. They recommended that a new and considerably larger study be conducted. The design for such a study was developed, debated and eventually implemented at Australian Wool Innovation's Falkiner Memorial Research Station. This study involves the use of 16 industry sires chosen to represent the key traits under investigation (parasite resistance, wool quality, meat production and reproduction), and 4 sires from past QTL studies that had a reasonable chance of being heterozygous for a QTL of reasonable size. These sires have been joined to Merino ewes (in the case of Merino sires) and to Border Leicester / Merino cross and terminal sire ewes (in the case of composite sires). The study has generated over 200 progeny per sire for the industry sires and 400 / sire from each of 4 sires previously used in QTL studies. The progeny generated have now been phenotyped for over 250 traits and await genotyping.The design enables both linkage and LD mapping.

Given the projected limitations of the genotyping budget, one of the strategies under consideration was to genotype, using microsatellites, only 6 chromosomal regions (chosen as having high probability of containing QTL for the traits of interest). This was considered a practical solution relative to available resources. If a QTL were confirmed, it would be fine-mapped and sufficient offspring with contrasting QTL genotypes would be bred to enable detailed study of the QTL. Following from this, a diagnostic haplotype could be validated and the causative mutation identified.

The major challenge inherent in this strategy was which 6 chromosomes to chose. The intended solution was to consider all available evidence from published QTL studies in sheep (e.g. Beh *et a.l,* 2002; Walling *et al.,* 2004) and from unpublished results available to SheepGenomics. As outlined below, this strategy has been superseded.

The imminent availability of tens of thousands of ovine SNPs allows a new strategy to be used for detecting QTL. If dense markers are available, for example >60,000 SNPs, it becomes possible to detect QTL based on linkage disequilibrium rather than linkage methods. Meuwissen *et al.* (2001) proposed a method (called genomic selection) for estimating breeding values by exploiting linkage disequilibrium from many SNPs and phenotype measures. Using information from a "training" population in which associations between SNPs and phenotypes have been determined, one can determine breeding values of animals which have no progeny or measured performance. The great value of this technology is that it will reduce the interval required for making breeding decisions for hard-to-measure traits, thus speeding genetic progress.

Many SNPs were known for man and mouse (e.g. www.hapmap.org) at the time of publication of Meuwissen *et al.* (2001) but nothing like the density of SNPs required was available for livestock. It was not until 2005 that a bovine SNP chip with ~10,000 features became available: at that point, it became immediately obvious that the genotyping strategy for SheepGenomics needed to follow suit. Specifically, there was a need to develop sheep-specific SNPs that would be sufficiently numerous and cost-effective to enable all regions of all chromosomes to be genotyped. Perhaps more importantly, the approach also provided a logical and simple commercialization strategy, via use of the chips themselves or selected SNP subsets to aid breeding value prediction.

Schaeffer (2006) calculated that in dairy breeding programs, the cost of bull testing could be drastically reduced and the rate of genetic improvement could be doubled. The situation is different in sheep as more traits can be measured on both sexes, whereas dairy traits can generally not be measured in bulls. However, also in sheep breeding programs, there are many traits that are not easy to measure in young breeding animals, e.g. carcass traits and meat quality, reproduction, feed efficiency and disease resistance are currently difficult to improve due to the lack of information to select young breeding animals. For such traits, the rate of gain could easily be doubled using GWS, although this would to some extent be at the expense of other traits that are currently favoured (e.g. weight). In a multiple trait breeding objective, the overall benefit of MAS or GWS is relatively less then what can be additionally achieved for some traits. One of the main benefits of molecular information in breeding programs is that it helps shift genetic improvement towards traits that are currently hard to improve. GWS offers the opportunity to include new difficult to measure traits in breeding programs provided that the phenotypes are measured in an appropriate genotyped population.

While the initial commercial product from genomic selection may be panels of small numbers of SNPs (100-1000) which predict breeding values from the desired phenotype, it should also be noted that the genomic selection approach permits mapping of QTL to much smaller confidence intervals (0.5-1Mb) than was possible with the linkage approach (if the LD structure of the genotyped population allows such resolution). Thus the time required to achieve the goal of "find useful genes and put them to work" can be greatly reduced.

SheepGenomics commenced 4 years ago with a dual strategy of discovering gene variants and their functions, and exploiting them via genetic selection and development of mimetics of their function. It is now clear that technological changes have made it more likely that the stated goals can be achieved. To maintain an effective gene discovery program and ensure that the original goals be adhered to as technology changes, has required continual modification of methods to maximize chances of success. The primary changes have been in the rapid increase in knowledge of the structure of the sheep genome, and, in particular, the numbers and locations of markers (microsatellites and SNPs). The identity and location of these markers have been deposited in the public domain, as a joint activity of ISGC and SheepGenomics. Application of these resources to the phenotyped populations developed in SheepGenomics and elsewhere (e.g. previous studies, the Sheep CRC Information Nucleus Flocks (Banks *et al.,* 2006) and industry flocks) should yield many more outputs than were originally envisaged.

Another dimension to the adaptive response of SheepGenomics is exemplified by its approach to "validation" and application of DNA markers. Kijas *et al.* (2007) report size of allele effects for a number of muscle and fat phenotypes and their allelic frequency in Australian sheep for a Myostatin mutation (Clop *et al.,* 2006). To apply these results to industry requires not only use of extant SheepGenomics resources, but rapid engagement with the sheep industry to access and genotype animals, and develop robust procedures to utilize the results in "enhanced" breeding values. A close association between SheepGenomics and Sheep Genetics Australia has facilitated this engagement and developed relationships to facilitate incorporation of gene-specific information into breeding values.

The tools that are essential for SheepGenomics to deliver on its promise of "find useful genes and put them to work" have also become essential for the next phase of genetic improvement in the Australian Sheep Industry. The Sheep CRC has explicitly designed its Information Nucleus flocks to exploit SNPs as tools for delivery of hard-to-measure new traits to industry (see Fogarty *et al.,* these proceedings). The SNPs, and the delivery / testing platform that will eventually be used by industry will be largely derived as a consequence of the close collaboration between the ISGC, Sheep CRC, Sheep Genetics Australia and SheepGenomics.

CONCLUSION

In summary, we can now see a clear path for translation into industry application of the first 5 years of SheepGenomics investment. This has been via setting ambitious long term goals, creating critical underpinning resources, and adapting rapidly to changes in technology. However, at best, the current work will only deliver the comparatively easy-to-obtain results. Perhaps more importantly we also see how an extension of this approach will deliver additional benefits: both in enhancement of genetic gain and better intervention based tools.

ACKNOWLEDGEMENTS

SheepGenomics is a joint initiative of Meat & Livestock Australia and Australian Wool Innovation Ltd supported by AgResearch Ltd, CSIRO Livestock Industries, Monash University, Department of Primary Industries Victoria, South Australian Research and Development Institute, and the Universities of Adelaide, Melbourne, New England, Sydney, Western Australia and Western Sydney. Members of the ISGC can be found at www.sheephapmap.org. This is SheepGenomics publication number 84.

REFERENCES

Beh, K.J., Hulme, D.J., Callaghan, M.J., Leish, Z., Lenane, I., Windon, R.G., Maddox, J.F. (2002) *Animal Genetics.* **33(2)**:97.

Banks, R.G., van der Werf, J., and Gibson. J.P. (2006) *8th World Congress on Genetics Applied to Livestock Production August 13-18, 2006, Belo Horizonte, MG, Brazil,* paper 30.

Clop, A., Marcq, F., Takeda, H., Pirottin, D., Tordoir, X., Bibé, B., Bouix, J., Caiment, F., Elsen, J-M., Eychenne, F., Larzul, C., Laville, E., Meish, F., Milenkovic, D., Tobin, J., Charlier C., and Georges, M. (2006) *Nature Genetics* **38**:813.

Dalrymple, B.P., Kirkness, E.F, Nefodov, M., McWilliam, S., Ratnakumar, A., Barris, W., Zhao, S., Shetty, J., Maddox, J.F., O'Grady, M., Nicholas, F., Crawfoprd, A.M., Smith, T., de Jong, P., McEwan, J.C., Oddy, V.H. and Cockett, N.E. (2007) *Submitted.* http://www.livestockgenomics.csiro.au/vsheep

Fogarty, N. *et al*. (2007) *Proc. Assoc. Advmt. Anim. Breed. Genet.* ***17****:417.*

Kijas, J., McCulloch, R., Hocking-Edwards, J., Oddy, V.H., Lee, S.H., van der Werf, J. (2007) *Mammalian Genome (submitted)*

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., *et al.* (2005) *Nature.* **437:**376.

Meuwissen, T.H., Hayes, B.J. and Goddard, M.E. (2001) *Genetics.* **157:**1819.

Oddy, V.H., T.J. Longhurst, F.W. Nicholas, J.F. Maddox and M. B. McDonagh (2005) *Proc AAABG* **16:**209.

Schaeffer, L.R. (2006) *J. Anim. Breed. Genet.* 123:218.

Roberts-Thomson, M., Paterson, K., Dodds, K., Lee, M., Crawford, A. and McEwan, J. (2005) *Plant Animal Genome* **XIII**: P540 http://www.intl-pag.org/pag/13/abstracts/PAG13_P540.html

Walling, G.A., Visscher, P.M., Wilson, A.D., McTeir, B.L., Simm G. and. Bishop S.C. (2004) *J. Anim. Sci.***82**:2234.