Continuing the Development of Sheep Genomics Resources

Noelle E. Cockett[1] on behalf of the International Sheep Genomics Consortium
[1]Utah State University, 4900 Old Main Hill, Logan, Utah 84322-4900 USA

Justification:

Worldwide, there are more than 1 billion sheep belonging to over 1,300 breeds (Scherf 2000), demonstrating the vast genetic variability that exists in sheep. This variation has been captured in global food and fiber markets through widely distributed meat, milk and wool production. Sheep commodities contribute US$50 billion (farm gate prices) annually, and they are particularly important to the economies of less developed nations.

In addition to the world-wide production of food and fiber, sheep serve an expanding role in biomedical research. The moderate size, low maintenance requirements, and docile temperament of sheep make them a practical large-animal model for human physiology and for studying genetic and acquired diseases. There are over 115,000 PubMed references for sheep, of which 7,500 are classified as sheep models. Currently there are 195 projects using sheep as a model in the National Institutes of Health (NIH) Research Portfolio Online Reporting Tool database (http://projectreporter.nih.gov/reporter.cfm), including 62 in the area of fetal development, 14 in metabolism, 129 in immunity and disease projects, 19 models for surgery techniques, and 16 studies on prion diseases. Similarly, the Online Mendelian Inheritance in Animals database (http://omia.angis.org.au) contains reports of at least 71 sheep models for human inherited traits, such as rickets (Thompson et al. 2007), polycystic kidney disease (Johnstone et al. 2005), muscular dystrophy (Johnsen et al. 1997), osteoarthritis (Scott 2001), and hemophilia A (Porada et al. 2010).

Sheep models have been developed to examine a wide variety of human medical conditions such as brain injury (Finnie et al. 2008), the effects of fetal alcohol exposure (Parnell et al. 2007), inter-vertebral disc degeneration in the spinal column (Zhou et al. 2007), the physiological effects of smoke inhalation (Sakurai et al. 2007), asthma (Snibson et al. 2005), and intrauterine and placental growth (Morrison 2008; Barry et al. 2008). In addition, tissue engineering (Roelofs et al. 2008; Kon et al. 2008), medical devices (Wang et al. 2010; Ahern et al. 2010), therapies for osteoporosis (Turner 2006), the feasibility of *in utero* gene therapy (Ersek et al. 2010), the safety of retroviral-based transmission systems for gene therapy (Van den Broeke and Burny 2003), vaccine delivery (Scheerlinck et al. 2008), and the possibility of stem cell transplantation (Mackenzie et al. 2001) have all been tested using sheep as the large-animal model for humans.

Thus, a better understanding of the genetic makeup of sheep will lead to improvements in the efficiency of meat, milk and wool production, as well as contribute to a better understanding of health and disease issues in humans.

International Sheep Genomics Consortium:

The International Sheep Genomics Consortium (ISGC), which includes active participation of scientists from Australia, Canada, China, France, New Zealand, the UK and the USA, has collectively developed and distributed resources that advance research in sheep genomics. Ovine genomic resources arising from ISGC collaborations include a high coverage BAC library, end-sequencing of 100% of the BAC library, high and moderate resolution radiation hybrid panels, an integrated ovine genetic map, a whole genome BAC physical map, development of a virtual sheep genome (http://www.livestockgenomics.csiro.au/vsheep/), an initial 3X skim assembly of the sheep genome covering 52% of the genome (Tables 1 and 2) and a more complete reference sequence

with 90% genome coverage (Tables 3 and 4), a 1536 SNP pilot chip, a high density 50K SNP array, and a global diversity panel of commercially important and rare sheep breeds.

Significant leveraging of funds, expertise and efforts has been a hallmark of ISGC, resulting in a highly effective pipeline for the development of resources necessary for exploration of the sheep genome. In addition, the collaborative nature of the ISGC has ensured that sheep genomics resources are developed in a well-coordinated and non-redundant manner. Bimonthly ISGC conference calls provide a regular venue for updates and discussions on ”next steps”. These calls are documented and minutes are circulated to all ISGC members for comment. In addition, members of the research team meet face-to-face once a year at the Plant and Animal Genome meeting in San Diego, CA for coordination of results and presentation of results to the greater scientific community.

All projects undertaken by the ISGC are conducted within the public domain. Regular updates from the ISGC are available through several established sheep genomics websites, including the International Sheep Genomics Consortium website (http://www.sheephapmap.org/), the NCBI Sheep Genome Resources website (http://www.ncbi.nlm.nih.gov/projects/genome/guide/sheep/), the CSIRO Sheep Genome website (http://www.livestockgenomics.csiro.au/sheep/), and the Australian Sheep Gene Mapping website (http://rubens.its.unimelb.edu.au/~jillm/jill.htm).

Sheep Genome Reference Sequence:

Substantial leveraging of funds and expertise from the International Sheep Genomics Consortium has led to the recent and ongoing development of an ovine whole genome reference sequence (ISGC et al. 2010). The resulting sequence will accelerate searches for genetic regions and genes influencing phenotypes in sheep, and combined with the bovine genome reference sequence, will serve as a backbone for other ruminant species. The assembly will also be an important genomic resource for ovine biomedical research models.

The current draft of the sheep reference genome is constructed with contigs and scaffolds built by the Beijing Genome Institute (BGI) from data generated using Illumina Genome Analyzer technology from male and female Texel animals sequenced at Roslin and BGI, respectively (Table 3). The total sequence coverage generated from the two animals is greater than 50 fold coverage (Table 4). In addition, a range of mate-pair libraries with up to 20 kb inserts and at a lower coverage were also sequenced from the female animal at BGI to generate long scaffolds (Table 3). The sheep BAC end sequences from the male Texel and the positions of SNPs included on the Ovine SNP50 BeadChip in the virtual sheep genome and in the sheep linkage and RH maps were used to link, order and orientate the scaffolds and scaffold pieces into super-scaffolds and chromosomes. The resulting genome assembly was inspected for outsize and incorrectly oriented sheep BACs and also combined with comparative mapping of contigs to the bovine genome. The position and/or orientation of a small number of scaffolds were modified to minimize anomalous sheep BACs. Cattle BACs and a set of long insert mate-pair reads from the male Texel generated at Baylor College of Medicine-Human Genome Sequence Center (BCM-HGSC) using 454 technology were also mapped to the revised genome reference sequence and manually inspected for assembly anomalies which were again minimized. This process will be repeated a number of times until anomalies are minimized. The current assembly contains ~2.6 Gb of scaffolds assigned to chromosomes.

During this process a number of issues with the reference sequence assembly have been identified, in part due to the variation between and within the sequences of the two Texel animals. More than 1000 small tandem duplications in the current assembly are probably incorrect and the size of large gaps may be over-estimated. In addition, regions of high GC are under-represented in the assembly, with regions >60% GC having around 50% of the average coverage of 40 fold. Although several other problematic regions in the assembly have been identified, the ground work has been laid for a detailed assessment of the strategy for completing the genome reference

sequence project. The ISGC is confident that high quality draft sequences of both Texels will be produced in the near future.

Annotation of the sheep genes is underway using the pipeline developed by BGI and transcriptomes from seven sheep tissues. Raw sequence will be made available through GeneBank/EMBL, the NCBI/Ensembl trace archives and the Short Read Archive/European Read Archive. Interim assemblies will be available through the ISGC consortium website. The final draft genome assembly will be processed through the NCBI, Ensembl, and UCSC genome pipelines and become available through their genome browsers. As outlined in a recent paper (ISGC 2010), the ISGC asks users of prepublication data to adhere to the framework for data sharing established at the 2009 Toronto International Data Release Workshop.

While the ISGC is willing and able to curate the reference sequence for the short term, ongoing funding will be required for periodically upgrading the genome sequence and providing annotations and variant information. Long term curation by NCBI, Ensembl, or UCSC may be the solution. Another option is to provide ISGC with funding to "purchase" the curation, although this will require a more formal structure of the Consortium than currently exists.

The 'next step' activities for the sheep genome reference sequence project are as follows:

- Identify gaps within scaffolds and problem regions through comparisons back to other assembled mammalian genomes. Targeted sequencing of these regions will be generated with a single or pooled BAC tiling approach and possibly array-based hybridization enrichment to fill short gaps.

- Identify regions for targeted BAC-to-BAC sequencing based on comparisons of the sheep assembly to other species. Regions of interest include segmental duplications, gene families, etc.

- Design arrays to investigate copy number variation (CNV), epigenetics and genome methylation in sheep. A 384 CGH array has been created and a 2.1M CGH array is under development.

- Expand information for the ovine X and Y chromosomes. This will likely require manual inspection of the sequence or perhaps a BAC-based sequencing approach to reduce the scale of assembly problems, caused in a large part by the repeat structures of these chromosomes.

- Develop bioinformatics tools, including genome annotation. A framework will be established to enable manual annotation that assists automatic annotation.

- Determine a long-term solution for the repository of the ovine genome sequence and its annotation, along with responsibility for updates and improvements to the sequence database.

Sheep Whole Genome Resequencing:

Due to significant advances in sequencing technology and dramatic cost reductions, it is now feasible to collect whole genome sequence on multiple individuals within a species. This resequencing approach holds considerable promise for livestock animals. One application is the ability to predict genetic values of individual animals within a breeding program. The availability of whole genome sequences for superior animals is expected to substantially improve accuracies and genetic gain beyond what is currently possible using SNP chips (Meuwissen and Goddard 2010). In addition, whole genome resequencing will advance the study of domestication during the development of specialized livestock breeds. Detection of genetic variation underlying meat, milk

and fiber production through resequencing promises to unlock the unique population history of domesticated sheep, leading to discoveries on the genetic cause of phenotypic variation. Finally, whole genome resequencing of case and control animals segregating for known disease states will enhance the identification of causal mutations. This genetic information will make the sheep increasingly valuable as a biomedical model for human diseases.

Two key resources developed by the ISGC over the last five years make whole genome resequencing a feasible research tool. First, the ISGC is nearing completion of a first draft of the sheep genome reference sequence (see above). This reference sequence can be used as a template for the alignment and assembly of whole genome resequences. Second, the ISGC has established a DNA repository which contains genetic material from over 3400 animals and 80 breeds of domestic sheep from across the globe, along with populations of wild sheep. This repository represents the largest collection of *Ovis aries* germplasm anywhere in the world. All of the animals have been genotyped using the Ovine SNP50 BeadChip, which will allow identification of a subset of animals that captures a maximum of genetic diversity and are therefore, suitable candidates for resequencing.

The 'next step' activities for the sheep whole genome resequencing project are as follows:

- Resequence 100 domestic sheep and 20 wild sheep from the ISGC DNA repository to at least 12-fold coverage per animal.

- Establish bioinformatics resources to hold and analyze the resulting data. This activity would incorporate software that is being developed across the human and livestock research communities that processes short read sequence data sets. Pipelines to automate annotation of genomic features will be identified.

- Release genome sequences for individual animals assembled against the available ovine reference sequence. The data will include the complete gene complement of each animal, including the variation present within the exons. High density resequencing microarrays could be designed to quickly and efficiently scan the exome of larger populations of sheep.

- Characterize the annotated SNP and structural variation present within each animal including insertions, deletions and CNVs. This analysis is expected to deliver tens of millions of SNP, which can be used for construction of a high-density SNP array. In addition, sequence from phenotypically diverse breeds and wild sheep allows identification of domestication genes and adaptive variation directly underpinning phenotypic differences.

Sheep Transcriptome:

In general, researchers using sheep as a medical model are not currently exploiting natural genetic variation that exists in sheep or knowledge of its evolutionary history, and therefore, they have made limited use of the publicly available ovine SNP arrays. In contrast, biomedical researchers are greatly interested in the transcriptomes of sheep tissue(s) they are studying but unfortunately, this information is not extensively developed. The advent of sequencing–based gene expression analysis has demonstrated that the transcriptome of mammals is vastly more complex than had previously been thought. It is highly likely that the use of primarily single transcripts to define a gene will disguise differences between the transcriptomes of human and sheep (and between sheep and cattle). Articulating these differences will allow better evaluation of the applicability and the deficiencies of sheep biomedical models. It will also allow a much greater understanding of the differences between sheep and cattle, two closely related ruminants. In addition, understanding the subtleties of species differences may increase our understanding of key biological processes in sheep, humans and cattle.

The ‘next step’ activities for the development of an ovine transcriptome are as follows:

- Collect key tissues, including the heart, joints, bone, muscle, lung, and other relevant tissues used in sheep models for human diseases, followed by very deep sequencing of the tissues.

- Construct a transcript organization map of key genes expressed in the tissues, including detailed annotation of the genes.

- Utilize a public domain database to store and make available the transcript structures that are built using ISGC and public domain data. These data will also be included on public sheep genome browsers.

- Compare orthologous genes between the sheep and human transcripts (for exploration of biomedical models) and between the sheep and cattle transcripts (for investigation of production traits).

Conclusion:

Because of significant leveraging of funding from a variety of sources, including recent USDA/AFRI and Australian International Science Linkages grants, the whole genome reference sequence for sheep is now a reality. While this is cause for significant celebration, the emerging challenge is how to continue development and annotation of the sequence, as well as identify a long-term solution for a data repository.

This paper also includes several ‘next step’ activities for resequencing and transcriptome projects in sheep. While interesting in their own right, these projects provide significant underpinning for the ovine reference sequence. Specifically, curation of the resequence data from multiple animals will provide a check on the integrity of the reference genome assembly. Refinement of the assembly, as well as construction of the sheep pangenome representing the extant variation of the "sheep" genome, will be possible using these data. Likewise, a detailed analysis of the transcriptome of sheep will contribute significantly to the correct assembly of the genome by providing another independent dataset containing information about the order of sequence features in the genome. It may also provide sequence data for features which are currently absent from the sheep genome assembly. In summary, these projects will enhance the development of a reference genome assembly for sheep.

APPENDIX A - REFERENCES

Ahern BJ, Harten RD, Gruskin EA, Schaer TP (2010) Evaluation of a fiber reinforced drillable bone cement for screw augmentation in a sheep model – mechanical testing. *Clin Transl Sci* 3:112-5.

Barry JS, Rozance PJ, Anthony RV (2008) An animal model of placental insufficiency-induced intrauterine growth restriction. *Semin Perinatol* 32:225-30.

Ersek A, Pixley JS, Goodrich AD, Porada CD, Almeida-Porada G, Thain DS, Zanjani ED (2010) Persistent circulating human insulin in sheep transplanted in utero with human mesenchymal stem cells. *Exp Hematol* 38:211-20.

Finnie JW, Manavis J, Blumbergs PC (2010) Diffuse neuronal perikaryal amyloid precursor protein immunoreactivity in an ovine model of non-accidental head injury (the shaken baby syndrome). *J clin Neurosci* 17:237-40.

International Sheep Genomics Consortium: Archibald A, Cockett NE, Dalrymple BP, Faraut T, Kijas JW, Maddox JF, McEwan JC, Oddy VH, Raadsma HW, Wade C, Wang Wang JW, Xun X (2010) The sheep genome reference sequence, a work in progress. *Anim Genet* (in print).

Johnsen RD, Laing NG, Huxtable CR, Kakulas BA (1997) Normal expression of adhalin and merosin in ovine congenital progressive muscular dystrophy. *Aust Vet J* 75:215-6.

Johnstone AC, Davidson BI, Roe AR, Eccles MR, Jolly RD (2005) Congenital polycystic kidney disease in lambs. *N Z Vet J* 53:307-14.

Kon E, Chiari C, Marcacci M, Delcogliano M, Salter DM, Martin I, Ambrosio LO, Fini M, Tschon M, Tognana E, Plasenzotti R, Nehrer S (2008) Tissue engineering for total meniscal substitution: animal study in a sheep model. *Tissue Eng Part A* 14:1067-80.

Mackenzie TC, Flake AW (2001) Human mesenchymal stem cells persist, demonstrate site-specific multipotential differentiation, and are present in sites of wound healing and tissue regeneration after transplantation into fetal sheep. *Blood Cells Mol Dis* 27:601-4.

Meuwissen T, Goddard M (2010) Accurate prediction of genetic values for complex traits by whole-genome resequencing. *Genetics* 185:623-31. Epub 2010 Mar 22.

Morrison JL (2008) Sheep models of intrauterine growth restriction: fetal adaptations and consequences. *Clin Exp Pharmacol Physiol* 35:730-43.

Parnell SE, Ramadoss J, Delp MD, Ramsey MW, Chen WJ, West JR, Cudd TA (2007) Chronic ethanol increases fetal cerebral blood flow specific to the ethanol sensitive cerebellum under normoxemic, hypercapnic and academic conditions: ovine model. *Exp Physiol* 92:933-43.

Poroda DC, Sanada C, Long CR, Wood JA, Desai J, Frederick N, Millsap L, Bomann C, Menges SL, Hanna C, Flores-Foxworth G, Shin T, Westhusin ME, Liu W, Glimp H, Zanjani ED, Lozier JN, Pliska V, Stranzinger G, Joerg H, Kraemer DC, Almeida-Porada G (2010) Clinical and molecular characterization of a re-established line of sheep exhibiting hemophilia A. *J Thromb Haemost* 8:276-85.

Sakurai H, Soejima K, Nozaki M, Traber LD, Traber DL (2007) Effect of ablated airway blood flow on systemic and pulmonary microvascular permeability after smoke inhalation in sheep. *Burns* 33:885-91/

Scheerlinck JP, Snibson KJ, Bowles VM, Sutton P (2008) Biomedical applications of sheep models: from asthma to vaccines. *Trends Biotechnol* 26:259-66.

Scherf BD (2000) World watch list for domestic animal diversity. FAO/UNEP, Domestic Animal Diversity Information System (http://www.fao.org/dad-is/index.asp).

Scott PR (2001) Osteoarthritis of the elbow joint in adult sheep. *Vet Rec* 149:652-4.

Snibson KJ, Bischof RJ, Slocombe RF, Meeusen EN (2005) Airway remodelling and inflammation in sheep lungs after chronic airway challenge with house dust mite. *Clin Exp Allergy* 35:146-52.

Thompson KG, Dittmer KE, Blair HT, Fairley RA, Sim DF (2007) An outbreak of rickets in Corriedale sheep: Evidence for a genetic aetiology. *N Z Vet J* 55:137-42.

Turner AS (2006) Seasonal changes in bone metabolism in sheep: further characterization of an animal model for human osteoporosis. *Vet J* 174:460-1.

Van den Broeke A, Burny A (2003) Retroviral vector biosafety: lessons from sheep. *J Biomed Biotech* 1:9-12.

Wang Q, McGoron AJ< Bianco R, Kato Y, PInchuk L, Schaephoerster RT (2010) In-vivo assessment of a novel polymer (SIBS) trileaflet heart valve. *J Heart Valve Dis* 19:499-505.

Zhou H, Hou S, Shang W, Wu W, Cheng Y, Mei F, Peng B (2007) A new in vivo animal model to create intervertebral disc degeneration characterized by MRI, radiography, CT/discogram, biochemistry, and histology. *Spine* 32:864-72.

APPENDIX B – TABLES

Table 1 Sequences included in the first draft of the sheep genome assembly

Sheep Breed	454 Reads	Average Length (bp)	Base Count
Awassi	7,113,075	235	1,675,721,394
Merino	9,004,167	220	1,995,873,002
Poll Dorset	7,917,802	238	1,890,589,115
Romney	6,008,805	219	1,330,683,710
Scottish Blackface	5,611,006	229	1,273,929,900
Texel	6,735,328	227	1,529,979,986
Total	41,000,192		9,696,777,107

Table 2 Statistics of the first draft of the sheep genome assembly

Assembly Statistics	Assembly v1.0	Assembly v1.5
Blast Results		
Sequences uniquely positioned (M)	21.657	23.747
Sequences uniquely positioned (%)	53	58
Contig Assembly Statistics		
Number of contigs and singletons (M)	2.725	3.581
Assembled bases (Gbp)	1.311	1.775
Average contig length (bp)	481	496
Meld Statistics		
Number of ovine melded contigs (M)	2.524	2.282
Genomic bases (nonN Gbp)	1.224	1.492
Average ovine melded contig length (bp)	485	654

Table 3 Sequences included in the first draft of the sheep genome reference sequence

Mate-pair Libraries (bp)	Library Insert Size	Range Length (bp)	Sequence Amount[1] Male	Sequence Amount[1] Female	Sequence Coverage Male	Sequence Coverage Female	Clone Coverage Male	Clone Coverage Female	Sequencing Method
180	150-210 bp	101		24.0 Gb		~8.0 X		~7.0 X	Illumina
200			77 Gb		~24.0 X				Illumina
350	380-420 bp	101		105 Gb		~35.0 X		~61 X	Illumina
500			72 Gb		~25.5 X				Illumina
800	650-950 bp	101		32.0 Gb		~10.5 X		~42 X	Illumina
2,000	1.6-2.4 kb	45		35.7 Gb		~12.0 X		~264 X	Illumina
2,000	1.5-3.0 kb						~36.5 X		Illumina
5,000	4.5-5.5 kb	45		18.5 Gb		~6.2 X		~342 X	Illumina
8,000							~15 X		454
10,000	8.5-10.5 kb	45		8.3 Gb		~2.5 X		~307 X	Illumina
20,000	15.0-22.0 kb	45		1.8 Gb		~0.6 X		~133 X	Illumina
20,000							~15 X		454
184,000		687	0.26 Gb		0.09 X		~13.5 X		Sanger

[1]Assuming a genome length of 3 Gb for simplicity.

Table 4 Statistics of the first draft of the sheep genome reference sequence

Contig Before Filling	Size (bp)	Number
N50	4,270	150,061
N60	3,353	208,908
N70	2,515	285,465
N80	1,729	391,793
N90	904	565,823
Total Number (>100bp)		1,441,059
Total Number (>2Kb)		350,131
Total Size	2,227,462,283	

Contig After Filling	Size (bp)	Number
N50	17,951	39,663
N60	14,205	55,011
N70	10,872	74,723
N80	7,718	101,384
N90	4,525	142,185
Total Number (>100bp)		794,490
Total Number (>2Kb)		194,049
Total Size	2,451,140,656	

Super Scaffold	Size (bp)	Number
N50	5,672,100	130
N60	4,228,953	184
N70	2,851,756	262
N80	1,661,292	384
N90	665,596	632
Total Number (>100bp)		8,049
Total Number (>2Kb)		5,390
Total Size	2,710,209,454	

Chromosome	Size (bp)
chr1	268,452,945
chr2	242,471,356
chr3	218,606,109
chr4	117,743,527
chr5	104,471,520
chr6	113,657,083
chr7	99,026,079
chr8	90,927,558
chr9	93,028,408
chr10	82,666,793
chr11	59,793,901
chr12	77,415,839
chr13	81,690,102
chr14	61,240,295
chr15	79,022,021
chr16	70,735,842
chr17	71,821,549
chr18	64,954,901
chr19	49,419,751
chr20	46,299,762
chr21	44,998,392
chr22	48,214,118
chr23	60,646,590
chr24	39,892,205
chr25	41,992,853
chr26	44,371,423
chrX	98,695,046
Total	2,472,255,968

APPENDIX C - CONTRIBUTING ISGC MEMBERS

Dr. Noelle E. Cockett
Utah State University
Department of Animal, Dairy and Veterinary Sciences
Logan, UT 84322-4900
USA

Dr. Brian Dalrymple
CSIRO Livestock Industries
306 Carmody Road
St Lucia
Queensland 4067
Australia

Dr. James Kijas
CSIRO Livestock Industries
306 Carmody Road
St Lucia
Queensland 4067
Australia

Dr. Jillian F. Maddox
16 Park Square
Port Melbourne
Victoria 3207
Australia

Dr. John E. McEwan
AgResearch
Invermay
Private Bag 50034
Mosgiel
New Zealand

Dr. Hutton Oddy
University of New England
Department of Animal Science
Armidale
NSW 2351
Australia

Dr. Herman Raadsma
University of Sydney
Department of Veterinary Science
Sydney
NSW 2006
Australia

Dr. Jiang Yu
Kunming Institute of Zoology, CAS
Jiaochang Road 32#
Kunming, Yunan 650223
China

Dr. Wenguang Zhang
Inner Mongolia Agricultural University
Hohhot,010018, China
Kunming Institute of Zoology
Chinese Academy of Sciences, Kunming,
China