ACCEPTED FOR PUBLICATION IN *ANIMAL GENETICS*

The Sheep Genome Reference Sequence: A Work in Progress

The International Sheep Genomics Consortium[1], *Alan L. Archibald*[2], *Noelle E. Cockett*[3], *Brian P. Dalrymple*[4], *Thomas Faraut*[5], *James W. Kijas*[4§], *Jillian F Maddox*[6], *John C. McEwan*[7], *V. Hutton Oddy*[8], *Herman W. Raadsma*[9], *Claire Wade*[9], *Jun Wang*[10], *Wen Wang*[11] *and Xu Xun*[10]

[1] www.sheephapmap.org

[2] The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Roslin, EH25 9PS, UK

[3] Utah State University, Logan Utah, UT 84321-4900, USA

[4] CSIRO Livestock Industries, Brisbane, Queensland 4067, Australia

[5] INRA, Toulouse BP 52627, France

[6] Department of Veterinary Science, The University of Melbourne, Victoria 3010, Australia

[7] AgResearch Invermay Agricultural Centre, Mosgiel 9053, New Zealand

[8] School of Environmental and Rural Science, University of New England, New South Wales 2351, Australia

[9] Faculty of Veterinary Science, University of Sydney, New South Wales 2006, Australia

[10] BGI Shenzhen, 518083, China

[11] Kunming Institute of Zoology, Kunming 650223, Yunnan, China

[§] Address for correspondence, James Kijas, ISGC Secretary

Phone, +61 7 32142442; Fax, +61 7 32142440; e-mail, James.Kijas@csiro.au

Summary

Until recently, the construction of a reference genome was performed using Sanger sequencing alone. The emergence of next-generation sequencing platforms now means reference genomes may incorporate sequence data generated from a range of sequencing platforms, each of which have different read length, systematic biases and mate pair characteristics. The objective of this review is to inform the mammalian genomics community about the experimental strategy being pursued by the International Sheep Genomics Consortium (ISGC) to construct the draft reference genome of sheep (*Ovis aries*). Component activities such as data generation, sequence assembly and annotation are described, along with information concerning the key researchers performing the work. This aims to foster future participation from across the research community through the coordinated activities of the consortium. The review also serves as a 'marker paper' by providing information concerning the pre-publication release of the reference genome. This ensures the ISGC adheres to the framework for data sharing established at the recent Toronto International Data Release Workshop and provides guidelines for data users.

Keywords

Reference, genome, sequence, *Ovis aries*, sheep

Introduction

Since initial domestication approximately 11,000 years ago (Colledge *et al.,* 2005; Zeder, 2008), man has selected sheep (*Ovis aries*) for the specialised production of fibre, meat and milk within a diversity of production systems. The result is a spectrum of phenotypically diverse breeds which are of vital global economic and social importance. In order to accelerate genetic gain, understand the etiology of disease and obtain insights into the genetic control of milk and wool production, the International Sheep Genomics Consortium has developed a range of resources for the ovine research community. These are listed in Table 1 and include genome maps, SNP repositories and sequence. By comparison with the cattle genome which was prioritised by the USDA, NIH and others for sequencing (Bovine Genome Sequencing and Analysis Consortium *et al*., 2009), the funding resources available to the ovine genomics community have been limited. The opportunity to obtain vast amounts of sequence at dramatically reduced cost arrived with the advent of next-generation sequencing (NGS) technology. The suitability of using short sequence reads (less than 100 bp) for assembly of complex eukaryotic genomes was initially unproven. However, this changed in early 2009 with the publication of the draft sequence of the giant panda genome constructed solely from paired-end NGS reads (Li *et al*., 2009). Thus, the consortium agreed to initiate the sequence of the reference sheep genome at the ISGC workshop held in January 2009 at the Plant and Animal Genome Conference in San Diego California.

Sequencing, Assembly and a Physical Map

The data generation phase of the sheep reference genome project commenced at two sequencing facilities in late 2009 (Table 2). Kunming Institute of Zoology and BGI Shenzhen (lead contacts, Wen Wang, Jun Wang and Xu Xun) commissioned whole-genome shotgun sequencing of a single Texel ewe using the Illumina Genome Analyzer. Libraries with insert

sizes ranging from 170 bp up to 40 Kb have been constructed and used to generate approximately 220 Gb (75-fold coverage) of paired-end sequence. Simultaneously, the ARK-Genomics Center for Comparative and Functional Genomics at The Roslin Institute (lead contact, Alan Archibald) has produced an additional 140 Gb of Illumina paired-end sequence data from libraries with insert sizes ranging from 200 to 500 bp (45-fold coverage) derived from a single Texel ram which was used previously as the DNA source for CHORI-243 BAC library (Table 1, Dalrymple *et al.*, 2007). Additional sequence data are being generated from a mate-pair library constructed from fragments of ~3-8 Kb from the same Texel ram. A primary *de novo* assembly will be constructed using sequence derived from the female Texel. Contigs will be constructed from small insert-size libraries using read overlaps, before contigs are joined into scaffolds using an iterative procedure reliant on the mate pair characteristics of paired-end reads. Preliminary analysis indicates this is likely to generate contigs and scaffolds with an N50 in excess of 20 Kb and 2 Mb respectively. Once complete, sequence from the Texel ram will be added to fill gaps, increase the number of SNP identified and facilitate analysis of the Y chromosome. The final assembly parameters are currently unknown, meaning the possibility remains that independent *de novo* assemblies of sufficient quality may be constructed for both animals and compared.

To assist in the assembly process, the consortium has constructed a revised version (VSG2) of the virtual sheep genome (Dalrymple *et al.*, 2007). VSG2 has more than 190,000 data points based on the mapping of sheep BAC end sequences to cattle and other mammalian genomes, providing a dense virtual physical map (lead contact, Brian Dalrymple). This map is being refined using a combined analysis of two radiation hybrid (RH) panels constructed using either 5000 rad (Wu *et al.*, 2009) or 12,000 rad of radiation (Laurent *et al.*, 2007). Both RH panels have been typed for approximately 50,000 evenly spaced SNP formatted onto the *ovine*

SNP50 BeadChip (Illumina, USA). Preliminary analysis suggests that the resulting RH map will contain in excess of 35,000 markers (lead contacts, Thomas Faraut and Noelle Cockett). The consortium plans to increase map density and resolve discrepancies using locus ordering based on linkage disequilibrium (LODE) mapping (lead contact, Herman Raadsma) and haplotype blocks (lead contact, Hutton Oddy) to obtain a final integrated physical map. Sequence assembly and physical map construction will independently generate positional data for tens of thousands of loci. This allows the consortium to (i) inspect the concordance between these positional predictions to assess the quality of the draft genome assembly and (ii) merge the information during production of an enhanced final assembly. The later point is an essential component of the assembly strategy, as the final integrated physical map will be used to order and orient scaffolds, with the final product being a pseudomolecule for each sheep chromosome.

Gene Prediction and Annotation

To assist in characterisation of the transcribed component of the genome, BGI Shenzhen has completed generation of approximately 15 Gb of sequence derived from seven tissues (lead contacts, Wen Wang, Jun Wang and Xu Xun). These sequences complement existing ovine EST collections and emerging NGS datasets derived from an expanded set of tissues. Together, the data will be used to predict the number of sheep genes and to annotate their position and structure in the genome (lead contacts, Wen Wang, Jun Wang and Xu Xun).

Prepublication Data Access and Usage

All projects undertaken by the ISGC are conducted within the public domain. In order to promote public benefit arising from the reference genome project, the consortium plans to share the information by submitting datasets into public repositories soon after data

generation. Raw sequence will be made available through GeneBank/EMBL, the NCBI/Enseml trace archive and the Short Read Archive / European Read Archive. Interim assemblies will be available at 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. Importantly, the data will be made available prior to the publication of an ISGC paper describing global analysis of the sheep genome. This prepublication data release strategy is in accordance with the Bermuda and Fort Lauderdale agreements and the more recent Toronto Statement (Toronto International Data Release Workshop 2009) which provides guidelines for both data generators and data users. The ISGC asks that users adhere to each of these policies and refrain from publishing any global analysis of the sheep genome before the ISGC using consortium data. Global analysis includes the identification of complete (whole genome) sets of genomic features such as genes, gene families, regulatory elements, repeat structures and GC content. Global analysis also includes chromosome wide or whole-genome scale comparisons with other species and reassemblies of the sheep data. Further, users are asked to cite this paper as the source of all ISGC prepublication data. The consortium plans to develop a series of detailed companion papers which describe the discoveries enabled by the production of the draft reference assembly. To assist, the consortium would welcome participation from any research group with an aligned interest and who would like to contribute (lead contact Brian.Dalrymple@csiro.au). In addition, the lead contacts for each component of the work are listed in Table 2.

Looking Forward

The ISGC anticipates that the availability of a sheep reference genome assembly, coupled with continued improvement to NGS throughput, will mean hundreds of individual sheep will be subjected to whole genome sequencing in the coming year. In preparation, the consortium

plans to merge the reference genome with sequence derived from unrelated animals to create an “integrated genome” which describes SNP and the distribution of structural variation (lead contacts, John McEwan and James Kijas). In the first instance, both Roche 454 (3-fold coverage) and Illumina GA (0.8-fold coverage) reads produced during construction of the *ovine* SNP50 BeadChip will be used (Table 1). As more domestic sheep and their wild relatives are re-sequenced, iterative builds of the integrated genome will be performed to provide a resource for evolutionary studies and the identification of genes which underpin phenotypic change.

Acknowledgements

The International Sheep Genomics Sequencing Consortium is grateful to the following for funding support for the sheep genome sequencing project, The Roslin Institute, University of Edinburgh and Biotechnology and Biological Sciences Research Council, U.K.; The Scottish Government, U.K.; Defra/HEFC/SHEFC Veterinary Training and Research Initiative, U.K., USDA-ARS, USA, USDA-NRICGP, USA (grant numbers 2008-03923 and 2009-03305); USDA-NRSP-8, USA; Meat and Livestock Australia and Australian Wool Innovation Limited through sheepGENOMICS, Australia; Australian Government International Science Linkages Grant (CG090143), Australia; University of Sydney, Australia; CSIRO, Australia; AgResearch, NZ, Meat and Wool NZ though Ovita, NZ; INRA and ANR project SheepSNPQTL, France; European Union through FP7 Quantomics and 3-SR projects; a 973 Program (No. 2007CB815700), 100 Talents Program of Chinese Academy of Sciences and a CAS-Max Planck Society Fellowship to Kunming Institute of Zoology, China; The National Natural Science Foundation of China (30725008), Shenzhen (ZYC200903240077A ,

ZYC200903240078A), the Ole Rømer grant from Danish Natural Science Research Council, and the Solexa project (272-07-0196) to BGI Shenzhen, China.

References

Bovine Genome Sequencing and Analysis Consortium (2009) The genome sequence of taurine cattle: A window to ruminant biology and evolution. *Science* **324,** 522-528.

Colledge S., Conolly J. & Shennan S. (2005) The Evolution of Neolithic Farming from SW Asian Origins to NW European Limits. *Eur. J. Archaeol.* **8,** 137-156.

Dalrymple B.P., Kirkness E.F., Nefedov M., McWilliam S., Ratnakumar A., Barris W., Zhao S., Shetty J., Maddox J.F., O'Grady M., Nicholas F., Crawford A.M., Smith T., de Jong .PJ., McEwan J., Oddy V.H., Cockett N.E. & International Sheep Genomics Consortium. (2007) Using comparative genomics to reorder the human genome sequence into a virtual sheep genome. *Genome Biol.* **8,** R152.

Faraut T., de Givry S., Hitte C., Lahbib-Mansais Y., Morisson M., Milan D., Schiex T., Servin B., Vignal A., Galibert F. & Yerle M. (2009) Contribution of radiation hybrids to genome mapping in domestic animals. *Cytogenet Genome Res.* **126,** 21-33.

Kijas J.W., Townley D., Dalrymple B.P., Heaton M.P., Maddox J.F., McGrath A., Wilson P., Ingersoll R.G., McCulloch R., McWilliam S., Tang D., McEwan J., Cockett N., Oddy V.H., Nicholas F.W., Raadsma H. & International Sheep Genomics Consortium. (2009) A genome wide survey of SNP variation reveals the genetic structure of sheep breeds. *PLoS One* **4,** e4668.

Laurent P., Schibler L., Vaiman A., Laubier J., Delcros C., Cosseddu G., Vaiman D., Cribiu E.P. & Yerle M. (2007) A 12 000-rad whole-genome radiation hybrid panel in sheep,

application to the study of the ovine chromosome 18 region containing a QTL for scrapie susceptibility. *Anim Genet.* **38,** 358-563.

Li R., Fan W., Tian G., Zhu H., He L. *et al*., (2010) The sequence and de novo assembly of the giant panda genome. *Nature* **463,** 311-317.

Maddox J.F., Davies K.P., Crawford A.M., Hulme D.J., Vaiman D., *et al.*, (2001) An enhanced linkage map of the sheep genome comprising more than 1000 loci. *Genome Res.* **11,** 1275-1289.

Wu C.H., Jin W., Nomura K., Goldammer T., Hadfield T., Dalrymple B.P., McWilliam S., Maddox J.F. & Cockett N.E. (2009) A radiation hybrid comparative map of ovine chromosome 1 aligned to the virtual sheep genome. *Anim Genet.* **40,** 435-438.

Toronto International Data Release Workshop (2009) Prepublication data sharing. *Nature* **461,** 168-170.

Zeder M.A. (2008) Domestication and early agriculture in the Mediterranean Basin: Origins, diffusion, and impact. *Proc. Natl. Acad. Sci. U.S.A.* **105,** 11597-11604.

Table 1, Existing Genomic Resources Developed by the International Sheep Genomics Consortium

Resource	Lead contacts	Description and Reference
Linkage Map	Jillian Maddox jillm@rubens.its.unimelb.edu.au	Microsatellite based linkage map produced through genotyping of the International Mapping Flock. See Maddox *et al.*, 2001 and http, //rubens.its.unimelb.edu.au/~jillm/jill.htm
1536 SNP chip	James Kijas James.Kijas@csiro.au	Sanger re-sequencing of 2700 loci, 6021 SNP and a 1536 SNP chip genotyped across a range of breeds. See Kijas *et al.*, 2009
Virtual Sheep Genome	Brian Dalrymple Brian.Dalrymple@csiro.au	Comparative analysis of BAC end sequences to generate a virtual assembly http, //www.livestockgenomics.csiro.au/vsheep/
CHORI-243 BAC library	Pieter J. de Jong, Noelle Cockett pdejong@chori.org Noelle.Cockett@usu.edu.au	A 12 fold coverage BAC library constructed from a single male Texel. End sequencing of the library formed the basis of the virtual sheep genome.
Illumina GA sequence and76,000 SNP	James Kijas and Brian Dalrymple see above for e-mail	Deep re-sequencing of pooled genomic DNA from 60 animals identified approximately 76,000 SNP.
Roche 454 sequence and 595,000 SNP	John McEwan John.McEwan@agresearch.co.nz	Three-fold coverage of 454 sequence derived from six animals. Alignment identified approximately 595,000 SNP. https, //isgcdata.agresearch.co.nz/
Ovine SNP50 BeadChip	Brian Dalrymple, James Kijas and John McEwan see above for e-mail	Illumina Infinium based platform for genotyping 50,000 SNP distributed across the sheep genome.
HapMap and Breed Diversity Experiment	James Kijas See above for e-mail	Analysis of over 70 breeds of domestic sheep to investigate diversity and domestication.

INRA 1200-rad RH Panel and genome maps	Thomas Faraut Thomas.Faraut@toulouse.inra.fr	RH panel used for the assignment of sequence tagged sites. See Laurent *et al.*, 2007.
USDA 5000-rad RH Panel and genome maps	Noelle Cockett Noelle.Cockett@usu.edu.au	RH panel used for the assignment of sequence tagged sites.

Table 2, Sheep Reference Genome Project Groups

Analysis Group	Lead contacts	Notes
Sequence generation	Alan Archibald alan.archibald@roslin.ed.ac.uk	Produce approximately 140+ Gb of paired-end and mate-pair Illumina GA sequence from a single Texel ram.
Sequence generation	Wen Wang, Jun Wang and Xu Xun wwang@mail.kiz.ac.cn wangj@genomics.org.cn xuxun@genomics.org.cn	Produce approximately 220 Gb of paired-end Illumina GA sequence from single Texel ewe.
Assembly	Wen Wang and Xu Xun xuxun@genomics.org.cn wwang@mail.kiz.ac.cn	Short read assembler used to build contigs. Contigs arranged into scaffolds using the mate pair relationships derived each insert library. Gap filling and higher order scaffolds (scaftigs) produced.
Integrated Physical Map Construction	Brian Dalrymple Brian.Dalrymple@csiro.au	Comparison made between ovine, bovine, canine and human genome assemblies to refine RH map and build integrated physical map. Comparative genomics used to evaluate quality of contigs and scaffolds.
RH Mapping	Thomas Farut and Noelle Cockett Thomas.Faraut@toulouse.inra.fr Noelle.Cockett@usu.edu.au	The *ovine* SNP50 BeadChip and two radiation hybrid (RH) panels will be used to generate a consensus RH map (see Faraut *et al*., 2009). This will be assessed and improved using both LODE and comparative data. Once complete, a genome wide physical map will be used to order and orient contigs and scaffolds.
LODE mapping	Herman Raadsma Herman.Raadsma@sydney.edu.au	Locus ordering based on linkage disequilibrium (LODE) mapping will be used to position SNP and assist in ordering contigs and scaffolds.
Comparative Genomics	Claire Wade claire.wade@sydney.edu.au	Exploration of mammalian conserved elements in the sheep sequence. Analysis of transcriptomic data for evidence of species specific gene function and exon usage.

Haplotype Mapping	Hutton Oddy hoddy2@une.edu.au	Locus ordering based on haplotype mapping will be used to position SNP and assist in ordering contigs and scaffolds.
Polymorphism	John McEwan john.mcewan@agresearch.nz	Analysis of the sequence for the position and frequency of polymorphism including SNP, indels and copy number variants.
Repetitive Elements	David Adelson david.adelson@adelaide.edu.au	The genome sequence will be explored for the type and distribution of repetitive elements.
Gene Prediction	Wen Wang, Jun Wang and Xu Xun wwang@mail.kiz.ac.cn wangj@genomics.org.cn xuxun@genomics.org.cn	Generation of approximately 15 Gb of transcribed sequence collected from seven tissues will be used, along with available EST collections, for gene prediction. Predictions will be used to generate a gene catalog and for quality assessment of scaffolds.
Evolution	James Kijas James.Kijas@csiro.au	Comparison of the domestic sheep with wild sheep sequence and other mammals for analysis of genome evolution and identification of domestication genes.
Data Submission	Deanna Church church@ncbi.nlm.nih.gov	Guide NCBI genome submission process and subsequent annotation.