FROM SHEEP SNP CHIPS, GENOME SEQUENCES AND TRANSCRIPTOMES VIA MECHANISMS TO IMPROVED SHEEP BREEDING AND MANAGEMENT

B.P. Dalrymple[1], V.H.Oddy[2], J.C. McEwan[3], J.W. Kijas[1], R. Xiang[1], J. Bond[2], N. Cockett[4], K. Worley[5], T. Smith[6] and P.E. Vercoe[7]

[1] CSIRO Agriculture Flagship, Queensland, Australia
[2] NSW Department of Primary Industries, Beef Industry Centre, New South Wales, Australia.
[3]AgResearch, Otago, New Zealand
[4] Utah State University, Utah, USA
[5] Human Genome Sequencing Center, Baylor College of Medicine, Texas, USA
[6] UDSA MARC, Nebraska, USA
[7] Institute of Agriculture, University of Western Australia, Western Australia, Australia

SUMMARY

SNP chips are transforming animal breeding; low cost “assay-by-sequencing” methodologies and high quality reference genome sequences provide the opportunity for further significant improvement in both breeding and management. The Functional Annotation of ANimal Genomes (FAANG) consortium is applying methods developed by the human ENCODE project to annotate the genomes of livestock (sheep, cattle, pigs, etc.) with functional information including the probability that variation at a particular nucleotide has a causal role in any phenotype. We will contribute the detailed annotation of the transcriptome of the gastrointestinal tract of sheep to FAANG. We will undertake an integrated analysis of the variation in: genome sequence, transcription, gastrointestinal tract phenotypes and the environment across ~100 animals. This will be combined with analysis of a developmental time course of the gastrointestinal tract transcriptome from 30 days post conception to weaning, and an in-depth analysis of the gastrointestinal tract transcriptome from the new reference sheep, a North American Rambouillet. From this, and public FAANG data, we will estimate the probability that variation in a particular nucleotide has an impact on gastrointestinal phenotypes of interest (methane, nutrition, infection, microbial population) and identify the biological processes underlying the phenotype. This information will inform breeding schemes, identify management options and define phenotypes more precisely.

INTRODUCTION

Over the last ten years there has been a paradigm shift in the use of genetic markers in animal breeding. Microsatellites have been very quickly replaced by Single Nucleotide Polymorphisms (SNPs). For sheep the first whole genome SNP data was generated using a 1536 SNP platform (Kijas *et al.* 2009). This was soon followed by the Ovine SNP50 BeadChip with 50K SNPs (Kijas *et al.* 2012) and more recently a high density SNP-Chip with more than 600K SNPs (Kijas *et al.* 2014). In addition, targeted small SNP chips have been designed for specific purposes, such as parentage testing and use in industry breeding programs (Heaton *et al.* 2014). SNP genotyping information from large numbers of individuals is being applied in sheep breeding programs (Auvray *et al.* 2014, Moghaddar *et al.* 2014). However, the vast majority of the SNPs are still markers in linkage disequilibrium with the causative variation, not the causative variation themselves. Accurate identification of the causative variants would increase the accuracy of the prediction equations, by removing the linkage uncertainty. In addition, SNPs are not the only variations in the genome with a causal role in phenotype variation, for example duplications and rearrangements are involved in agouti (Norris and Whan 2008) and weight of lamb weaned

(Gonzalez *et al.* 2013). Copy number variations (CNVs) of a range of sizes are common in the sheep genome (Jiang *et al.* 2014). Some, but not all, of this variation is captured by linked SNPs (Gonzalez *et al.* 2013).

The search for causative mutations has identified a small number of large effect in sheep, including myostatin (Clop *et al.* 2006) and Callipyge (Smit *et al.* 2003). In both cases the causative mutations are not in the coding region of genes, but are in associated regulatory sequences, a new micro RNA-binding site (myostatin) and a methylated control region (Callipyge). Whilst the effect of variations in coding sequences on the function of the gene products can be predicted fairly reliably, this is not the case for variations in non-coding sequences such as long non-coding RNAs (lncRNAs) and regulatory sequences to which transcription factors bind. Experimental validation of causative mutations can only be justified for mutations of large effect. For most phenotypes many genes of small effect are involved and high throughput data generation followed by computational analysis is the only realistic way to approach the genome-wide identification of causal mutations. The first major barrier to effective prediction is that the role of the majority of individual nucleotides in the genome of production animals is not known. One of the major goals of the human and model organism ENCODE projects is to identify the role, or not, of each nucleotide in the genome (Dunham *et al.* 2012). To do this these projects have focussed on a small number of approaches, generally “assay-by-sequencing” methodologies including: in-depth transcriptomics, methylation, chromatin accessibility and conformation, transcription factor binding sites etc. Thus, across the vast majority of the genome each nucleotide can be annotated for a number of attributes: in a transcription factor binding site, transcribed, in an exon, in a splice site, in open or closed chromatin, etc. For each of these attributes the effect of variation on the role of the nucleotide in the functional element can be estimated (Gulko *et al.* 2015). This prediction is phenotype independent. Subsequently in studies of the association between variation in the genome and variation in the phenotype and the calculation of predictive equations, the probability that variation in a particular nucleotide will affect a downstream process can be included into the equations, and genetic relationship matrices (GRM) can be built using causative sites. The utility of this approach has been demonstrated in preliminary analyses (Gusev *et al.* 2014, Koufariotis *et al.* 2014). The FAANG consortium has been established to coordinate the international projects for the annotation of the roles of the nucleotides within the genomes of the major production animal species using the methodologies validated in the ENCODE projects (Andersson *et al.* 2015) (Figure 1).

However, association studies using the FAANG generated datasets will also inform our understanding of biological processes underlying a phenotype, by providing an estimate of the probability of a particular variation in the genome sequence affecting the phenotype of interest. The increased understanding of the biological processes will also be used to improve the management of the animals to reach their genetic potential. In addition, understanding the biological processes underlying the phenotype may enable us to define phenotypes better, reducing complex phenotypes to a series of simple phenotypes based on different biological processes (Figure 1).

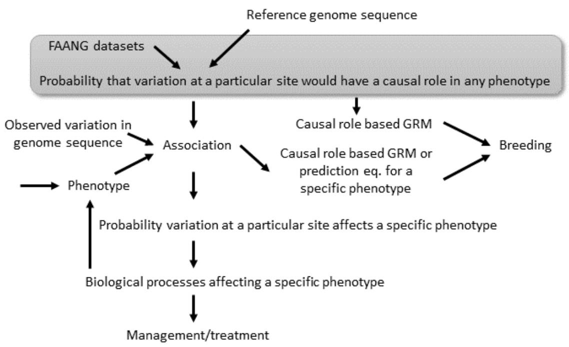
Reference genome sequence
FAANG datasets
Probability that variation at a particular site would have a causal role in any phenotype
Observed variation in genome sequence
Association
Causal role based GRM
Causal role based GRM or prediction eq. for a specific phenotype
Breeding
Phenotype
Probability variation at a particular site affects a specific phenotype
Biological processes affecting a specific phenotype
Management/treatment

Figure 1. Pipeline for FAANG annotation (highlighted) and delivery.

The gastrointestinal tract (GIT) is the

major source of nutrients in animals and an important source of their waste, such as methane (Johnson and Johnson 1995); appropriate function of the GIT is essential for the efficient production of animals and their products. A more detailed understanding of the genes and gene products (and their regulation) contributing to the development and function of the GIT, and hence the biological processes involved, will facilitate the development of new breeding strategies, methodologies for feeding animals, and managing the function of the GIT. Successful implementation of these will increase production efficiency and reduce waste products/kg of meat, milk, wool etc.

MATERIALS AND METHODS

In preparation for FAANG a new version of the sheep reference genome sequence is being assembled (Oar v4) using the long read PacBio sequencing technology. In addition, we will initially build scaffolds *de novo* using Hi-C (Burton *et al.* 2013), followed by using the ovine BAC library (Dalrymple *et al.* 2007), and the SNP-based sheep linkage and RH maps (Jiang *et al.* 2014). The individual being sequenced, a North American Rambouillet, will also be the source of reference tissue samples for the FAANG consortium assays. The Oar v3.1 assembly (Jiang *et al.* 2014), based on a male and a female Texel, will be gap-filled and errors corrected using a low coverage of PacBio sequencing of the male Texel to generate Oar v3.2.

Samples along the GIT, salivary gland, reticulum, sacs of the rumen, omasum, abomasum, duodenum, cecum, colon and rectum, have been collected from 63 Australian sheep (ewes) with a diversity of origins (Merino and Suffolk, Border Leicester, Dorset cross bred animals) and from 48 NZ sheep (ewes and rams) from a high/low methane selection line, also with a diversity of origins (Pinares-Patino *et al.* 2013). Samples from a time course of the development of the whole GIT of Merino sheep from 30 days post conception to weaning have been collected. Sampling times were selected based on the development of the sheep GIT (Franco *et al.* 1992). We will develop a detailed description of the transcriptome of the GIT of sheep by undertaking Illumina RNA-Seq and PacBio Iso-Seq on mRNA, lncRNA and small RNA (miRNA, snoRNA etc.) of the sheep GIT tissues. Gene and transcript models of protein and non-coding RNAs will be built using both assembly guided and *de novo* methodologies. The significantly expanded GIT-relevant transcriptome will identify transcript isoforms currently poorly represented in the sheep transcript models. The focus of the targeted manual annotation of transcripts will be gene products identified as likely to play key roles in the development and function of the GIT. We will use the correlation of the expression of transcripts with each other and with development and productivity GIT phenotypes measured (and microbial samples from the GIT collected) as part of the Australian Department of Agriculture funded “Host control of methane emissions from sheep” project to create gene/transcript networks, and to identify sets of genes and their transcripts informative of key biological processes in the sheep GIT. We will also map the changes in the processes during development of the GIT. We will use the sets of transcripts to identify transcription factors and other regulatory molecules likely to be involved in the regulation of key processes in the sheep GIT. All the data will be generated following the FAANG standard operating protocols and will be contributed to the FAANG consortium. We will promote and facilitate the use of the data by making it publicly available prior to publication in accordance with standard data sharing protocols and by providing online support for users.

RESULTS AND DISCUSSION

Analysis of the sheep genome and transcriptome has identified a strong relationship between the rumen and cornified and keratinized tissues such as the skin and ruminant specific genes encoding proteins predicted to be involved in the cornification of the rumen epithelium (Jiang *et al.* 2014). A first generation sheep GIT transcriptome atlas is currently being constructed using data generated as part of the sheep reference genome project. Initial analysis of the rumen gene expression from 24 NZ ewes has been undertaken demonstrating that the expression of genes relating to the cornified epithelium is dynamic, probably due to dietary effects (Ruidong et al., in preparation).The initial output of this work will be an in-depth and detailed catalogue of the transcriptome of the sheep GIT tissues increasing the number of alternate splice variants from an average of one per gene to more than four. Annotation accuracy and coverage of transcript isoforms is expected to be high for mRNAs and slightly lower for short RNAs, such as miRNAs. Conversely the discovery rate of new lncRNAs is expected to be high and of new protein coding genes is expected to be low.The developmental gene expression atlas of the GIT of sheep and the integration of other FAANG datasets with the atlas will support global research activities into the development and function of the GIT and how putative causative SNPs and other genomic variation affects the efficiency of animal growth, methane production, parasite resistance and other traits critical to the profitability and sustainability of livestock production. For example it is expected that reticulo-rumen, and possibly salivary gland gene expression will be associated with differences in rumen morphology and digesta flow and in turn methane production, and that abomasal and duodenal gene expression patterns will inform how sheep differ in parasite resistance.

CONCLUSIONS

From the research described above causative SNPs and CNVs weighted for phenotypic impact on GIT function will be available for inclusion in GRMs and prediction equations in breeding schemes. It is likely that these analyses will also identify biological pathway based phenotypes which may enable new approaches to the selection of animals for production traits to be identified. In contrast the pathway to the utilization of the outputs for improved management of sheep is much more poorly defined and likely to be more challenging.

REFERENCES

Andersson L., Archibald A., *et al.* (2015). *Genome Biol.* **16**: 57.
Auvray B., McEwan J.C., *et al.* (2014). *J. Anim. Sci.* **92**: 4375.
Burton J.N., Adey A., *et al.* (2013). *Nat. Biotechnol.* **31**: 1119.
Clop A., Marcq F., *et al.* (2006). *Nature Genet.* **38**: 813.
Dalrymple B.P., Kirkness E.F., *et al.* (2007). *Genome Biol.* **8**: 20.
Dunham I., Kundaje A., *et al.* (2012). *Nature* **489**: 57.
Franco A., Regodon S., Robina A. and Redondo E. (1992). *Am. J. Vet. Res.* **53**: 1209.
Gonzalez M.V., Mousel M.R., *et al.* (2013). *PLoS One* **8**: 9.
Gulko B., Hubisz M.J., Gronau I. and Siepel A. (2015). *Nature Genet.* **47**: 276.
Gusev A., Lee S.H., *et al.* (2014). *Am. J. Hum. Genet.* **95**: 535.
Heaton M.P., Leymaster K.A., *et al.* (2014). *PLoS One* **9**: 10.
Jiang Y., Xie M., *et al.* (2014). *Science* **344**: 1168.
Johnson K.A. and Johnson D.E. (1995). *J. Anim. Sci.* **73**: 2483.
Kijas J.W., Townley D., *et al.* (2009). *PLoS One* **4**: 13.
Kijas J.W., Lenstra J.A., *et al.* (2012). *PLoS. Biol.* **10**: 14.
Kijas J.W., Porto-Neto L., *et al.* (2014). *Anim. Genet.* **45**: 754.
Koufariotis L., Chen Y.P.P., Bolormaa S. and Hayes B.J. (2014). *BMC Genomics* **15**: 16.
Moghaddar N., Swan A.A. and van der Werf J.H.J. (2014). *Anim. Prod. Sci.* **54**: 544.
Norris B.J. and Whan V.A. (2008). *Genome Res.* **18**: 1282.
Pinares-Patino C.S., Hickey S.M., *et al.* (2013). *Animal* **7**: 316.
Smit M., Segers K., *et al.* (2003). *Genetics* **163**: 453.