WHAT'S NEXT IN GENOMICS?
FUNCTIONAL ANNOTATION OF ANIMAL GENOMES

J. Kijas[1], B. Barendse[1], C. Bottema[2], R. Brauning[3], A. Chamberlain[4], S. Clarke[3], B. Dalrymple[1], R. Tellam[1], B. Hayes[4], J. McEwan[3], S. Moore[5]

[1] CSIRO Agriculture Flagship, Queensland, Australia
[2] University of Adelaide, South Australia, Australia
[3] Invermay Agricultural Center, AgResearch, Mosgiel, New Zealand
[4] Department of Economic Development, Jobs, Transport and Resources, Victoria, Australia
[5] QAAFI, University of Queensland, Queensland, Australia

SUMMARY

The last five years has witnessed the completion of reference genome projects for each of the major livestock species, along with the application of high throughput SNP genotyping to fast track gene discovery and genomic prediction. This paper explores one possible new direction in genomics and its possible impact on animal science. An international project has been initiated that aims to identify the genomic regions responsible for gene regulation, thereby providing functional annotation of animal genomes (FAANG). This seeks to increase our ability to interpret variation in genome sequence and predict the resulting phenotypic consequence. This has large implications for animal science and in particular animal breeding, given a key objective of genomic prediction is to use molecular data (currently SNP) to predict genetic merit. To successfully annotate the regulatory elements in genomic sequence, the FAANG Consortium has been created to provide coordination and standardisation in data collection, quality control and analysis. Aspects of the consortium are described, along with information on Australia's current and future contributions.

THE GENOME TO PHENOME CHALLENGE

A central goal in biological science involves understanding complex systems, so that accurate predictions about the behaviour of systems can be made. Predictions might involve the susceptibility of an individual to disease or the treatment response of a patient to the application of a particular drug. In the case of livestock a key goal involves predicting variation in production traits, particularly those that are economically important but currently hard to measure. Animals are exceedingly complex, which makes the task of predicting phenotype challenging. The last decade has seen tremendous progress, whereby quantitative genetics theory combined with technical advances in SNP genotyping have allowed statistic models to effectively predict genetic merit. The accuracy of these predictions has been increased through use of dense SNP arrays, and by increasing the number of animals with both genotype and phenotype data in training populations. Further, reductions in sequencing cost have meant it is now feasible to collect the whole genome sequence of hundreds of animals. This is being used to improve genomic prediction, primarily through the utilisation of additional SNP. The availability of whole genome sequence, however, opens much richer opportunities given these datasets directly contain the sequence level differences that control phenotypic variation. Currently functional mutations are indistinguishable from a sea of neutral variation. These difficulties capture the essence of the genome to phenome challenge, which is to successfully interpret genome sequence to predict it's consequence on phenotypic variation.

GENOME ANNOTATION

To meaningfully tackle the genome to phenome challenge, a richly annotated reference genome sequence is required for each of the farm yard animal species. An important milestone on route to this objective was reached in 2014 with the completion of the draft reference genome assembly for sheep (Jiang *et al*. 2014). This completed the collection of available reference genomes for each of the major livestock production species that includes pig (Groenen *et al*. 2012), cattle (Bovine Genome Sequencing and Analysis Consortium 2009), chicken (Burt 2005) and goat (Dong *et al.* 2013). While improvements to these assemblies are ongoing the task of annotating each genome has commenced. Annotation describes the process by which particular sequence characteristics and functional elements are identified in the genome. To date, the features annotated in detail extend only to variation (e.g. SNP and various repeat classes), protein coding genes (intron and exon location) and some aspects of gross sequence classification such as GC content. What is almost completely missing from animal genomes is the accurate identification and annotation of the gene regulatory machinery. The ENCODE project sought to rectify this in human by cataloging the full complement of gene transcripts, their isoforms and the hundreds of thousands of enhancers, transcription factor binding sites and promoter regions active across different cell types (ENCODE Project Consortium 2012). This large and costly international research effort has provided key advancements in our understanding of biology. For example, of the approximately 25,000 human protein coding genes only about 50% are expressed in any given cell type (Romanoski and Glass 2015). Further, it appears possible to identify the combinations of transcription factors responsible for directing the specialisation of precursor cells to differentiate into particular cell types. These fundamental observations represent the first steps towards a more sophisticated ability to understand how DNA sequence and gene regulation together serve to control complex traits.

FUNCTIONAL ANNOTATION OF ANIMAL GENOMES (FAANG)

Animal scientists with the shared goal of producing genome wide maps of functional elements held a planning workshop in January 2014 the Plant and Animal Genome Conference (PAG XXII). The meeting conceptualised the creation of a consortium to coordinate and execute the FAANG project. Subsequent discussion has defined the structure of the FAANG Consortium (i.e. working groups and their roles) and aspects of the FAANG Project (i.e. the operational plan for the science). Key aspects of both are described here, however additional considerations relating to the creation of a common data infrastructure, a centralised data analysis centre, pre-publication data release and the operational principles for participating scientists is available at the consortium website (http://www.animalgenome.org/community/FAANG/). In addition, the consortium recently published a white paper that describes the rationale for the science while providing details about the objectives (The FAANG Consortium 2015). It is important to note that any interested scientists are welcome to participate, and this can be initiated by signing up to the consortium on the website.

SPECIES AND DATA TYPES

Given the ENCODE project focussed on a single species (human) and cost in excess of $150 million dollars, the livestock community recognised early that clear prioritisation was needed to design a project broadly in line with the vastly diminished financial resources likely available to animal scientists. This planning sought to take advantage of i) the declining cost of the sequencing and ii) key lessons from ENCODE relating to the choice of core data and tissue types for investigation. A prerequisite for inclusion in the FAANG project is the availability of a draft reference genome assembly of sufficient quality to serve as the template for annotation. At present, this means the project is confined to cattle, sheep, chicken and pig however additional species are

likely to be included as their genome assemblies improve (e.g. salmon, goat and horse). The consortium has also defined the following set of core assays to be deployed in each species:

i) RNA Transcriptomics
Annotation requires detailed knowledge of the gene transcripts that are present within tissues, along with details of the transcriptional complexity many genes exhibit (e.g. tissue specific isoforms). RNA-sequencing will therefore be used to generate the transcriptome of each core tissue, from each species.

ii) Histone Modification Marks
To map the genomic location of putative promoters, enhancers and transcription start sites (TSSs), CHiP-seq assays will be used that identify DNA sequences that bind to modified histones. The project has prioritised four histone modification marks found to be most informative by the ENCODE projects. These are:

- H3K4me3 that correlates with promoters and transcription start sites
- H3K27me3 which marks silenced genes
- H3K27ac that indicates active regulatory elements
- H3K4me1 which is associated with enhancers and enriched downstream of TTSs

iii) Chromatin Accessibility and Architecture
To complement ii), methods are available that identify 'protein bound' DNA sequences due to chromatin accessibility and architecture. DNaseI footprinting was the first generation of such approaches, however more robust and sensitive approaches have been developed. One is ATAC-seq, and will be used to identify open chromatin. Importantly, the results will be co-analysed with histone modification information to decipher the location of specific protein-DNA binding events to base-pair resolution.

Beyond each of these core assays, the consortium has identified an additional set of data types considered non-essential but informative. These include DNA methylation, antibody dependant direct identification of transcription factor binding sites and genome conformation assays using methods such as Hi-C. Additional detail on all of these assays can be found elsewhere (Lane *et al*. 2014; The Mouse ENCODE Consortium 2014; The FAANG Consortium 2015).

AUSTRALIA'S CONTRIBUTION

To initiate an Australian contribution into the international FAANG consortium, the co-authors have commenced informal discussions to i) collate existing projects that might be included and ii) define the objectives for future projects and strategise how they might be funded. Table 1 shows details of four projects in cattle and sheep, and in each case the focus is largely on transcriptomics. The first two are underway, the third is pending grant approval and the final project is funded and data generation is likely to commence in the last quarter of 2015. It is worthwhile noting very large surveys of genomic variation (SNP and indels), identified by whole genome sequencing projects, will be an important dataset used by the FAANG data analysis teams. In the case of cattle and sheep, these projects are being lead by Australian researchers (e.g. Daetwyler *et al.* 2014). Currently no data generation is planned for *Bos indicus* by non-Australian FAANG partners. Given their importance to the Australian cattle industry, the co-authors have prioritised *Bos indicus* as the focus for joint project applications. Interested parties who would like to become involved should contact any of the authors.

Table 1. FAANG compliant datasets currently being generated by Australian Scientists

Species	Breed	Tissues (n) [1]	Assays	Status	Contacts
Cattle	Holstein	Various(38)	RNA-seq, CHiP-seq	Ongoing	Hayes, Chamberlain
Sheep	Various	GIT[1] (7)	RNA-seq of mRNA (coding only)	Ongoing	Dalrymple, Oddy
Sheep	Various	GIT[1] (7)	RNA-seq of lncRNA, microRNA	App. Pending	Dalrymple, Oddy
Sheep	Rambouillet	Various(20+)	PacBio Iso-seq	Funded	Kijas, Cockett

[1]The number of tissues collected is given in parenthesis, however in some cases only a subset will be used for data generation. GIT is an abbreviation for gastrointestinal tract.

IMPACT FOR ANIMAL SCIENCE

The completed FAANG project will provide a comprehensive data resource describing gene regulation and the genomic elements responsible. The manner in which this resource is used is likely to evolve over time. In the short term, the availability of a genome atlas of regulatory elements should greatly assist the interpretation of whole genome sequencing studies that aim to identify functional variants. Currently the hunt for functional variants is most often successful where a protein coding mutation is responsible, simply because the annotation of animal genomes is most complete for protein coding genes. Conversely regulatory mutations that underpin trait variation are far more difficult to identify, however they may be the most common. FAANG data should greatly assist in elucidating the consequence of variants that directly impact phenotype via alterations in gene action. In the field of genomic prediction, the outcomes from FAANG may propel the field beyond the use of SNP as the sole molecular input data-type for prediction. For example, it may be possible that transcription factor binding site networks, co-expressed gene sets or combinations of these along with SNP genotypes will become the input data that returns higher prediction accuracies than currently available for complex traits. In the short term FAANG data can be used to better annotate SNP and prioritise those likely to directly impact phenotypic variation for exploitation in reduced size SNP panels diagnostic of key traits. Regarding evolutionary questions, the availability of standardised data across at least four species (two ruminants, one monogastric mammal and one bird) should facilitate discoveries concerning those components of the gene regulatory machinery that are conserved and those that are lineage specific. We anticipate a completed FAANG project should provide a general resource for research into mammalian biology and variation in complex traits.

ACKNOWLEDGEMENTS

The FAANG consortium was facilitated by the EC-US Biotechnology Research Task Force.

REFERENCES

Bovine Genome Sequencing and Analysis Consortium (2009) *Science* **324:** 522.
Burt D. (2005) *Genome Res*. **15:** 1692.
Daetwyler H., Capitan A., Pausch H., Stothard P., *et al*. (2014) *Nat. Genet.* **46**: 858.
Dong Y., Xie M., Jiang Y., Xiao N., Du X., et al. (2013) *Nat. Biotechnol.* **31:** 135.
Groenen M., Archiblad A., Uenishi H., Tuggle C., Takeuchi Y. *et al*. (2012) *Nature* 491: 393.
Jiang Y., Xie M., Chen W., Talbot R., Maddox J.F. *et al*. (2014) *Science* **344**: 1168.
Lane M. Niederhuth C.E., Lexiang J., Schmitz R.J. (2014) *Annu.Rev.Genet.***48**:49.
Romanoski C. and Glass C. (2015) *Nature* **518:** 7539.
The ENCODE Project Consortium (2012) *Nature* **489:** 57.
The Mouse ENCODE Consortium. (2014) *Nature* **515**: 355.
The FAANG Consortium. (2015) *Genome Biol*. **16:**57.