Proc. Assoc. Advmt. Anim. Breed. Genet. 22:209-212

PREDICTION OF GENOME-WIDE REGULATORY REGIONS IN SHEEP

M. Naval-Sanchez, Q. Nguyen, B.P*. Dalrymple, T. Vuocolo, R. L. Tellam, L.R. Porto-Neto, S. McWilliam, A. Reverter and J. Kijas**

CSIRO Agriculture & Food, 306 Carmody Road, St. Lucia, 4067, QLD, Australia; *Current address: Institute for Molecular Bioscience, 306 Carmody Road, St. Lucia, 4067, QLD, Australia ; **The UWA Institute of Agriculture, The University of Western Australia, 35 Stirling Highway, Crawley, WA 6009, Australia

SUMMARY

The annotation of the regulatory genome is essential to investigate the link between genotype and phenotype. In this study, we applied the computational method "Human Projection of Regulatory Sequences" (HPRS) (Nguyen *et al.* 2016) to project human regulatory information to sheep coordinates and provide a predictive sheep regulatory genome.

Firstly, we selected human large-scale publicly available human datasets as a reference for promoter and enhancer regions. Secondly, we converted the human regulatory information into sheep coordinates. We successfully mapped 70% and 65% of human promoter and enhancers regions into the sheep genome. Finally, we evaluated whether the predicted sheep regulatory genome captures sheep-regulatory information by assessing its overlap with in-house H3K27ac and H3K4me3 ChIP-seq data from sheep brown adipose tissue. We find that our predicted regulatory elements are enriched for sheep regulatory regions and present high sensitivity and specificity to discern between promoters and enhancers.

INTRODUCTION

The human regulatory genome has been extensively characterized by large-scale genomic Consortiums such as the ENCODE (ENCODE Project Consortium 2012) and Epigenomics Roadmap (Roadmap Epigenomics Consortium *et al.* 2015). Meanwhile, the functional annotation of livestock species, specifically sheep, is lagging behind. Projects such as the Functional Annotation of Animal Genomes (FAANG) (Consortium *et al.* 2015) aim to resolve this issue in the near future. Alternatively, computational approaches, in particular the "Human Projection of Regulatory Sequences" (HPRS) (Nguyen *et al.* 2016) pipeline has been successfully used to project human regulatory information into cattle coordinates providing high confidence regulatory information at the promoter and enhancer level.

Here, we apply the HPRS pipeline to predict the sheep regulatory genome and use sheep-specific regulatory data to show that the method captures sheep regulatory information with high sensitivity and specificity.

MATERIALS AND METHODS

Human genomic databases. Human regulatory information was obtained from three distinct databases:

1) FANTOM5 promoters and enhancer atlas detected by CAGE (Forrest *et al.* 2014; Andersson *et al.* 2014).
URL:http://fantom.gsc.riken.jp/5/datafiles/latest/extra/CAGE_peaks/hg19.cage_peak_phase1and2combined_coord.bed.gz; http://enhancer.binf.ku.dk/presets/permissive_enhancers.bed.

2) Epigenomics Roadmap enhancers from 88 human primary tissues (Roadmap Epigenomics Consortium *et al.* 2015). We use the chromatin states defined as enhancer, enhancer genic and enhancer bivalent.URL: http://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html.

3) ENCODE Transcription Factor Binding Sites (TFBSs): ENCODE proximal and distal TFBSs by ChIP-seq for 163 TFs. URL: https://www.encodeproject.org/data/annotations/v2/.

Human Projection of Regulatory Regions pipeline (HPRS). We followed the same procedure as (Nguyen *et al.* 2016) (https://bitbucket.csiro.au/users/ngu121/repos/hprs/browse/). In brief, the program liftOver (minMatch=0.2) (Hinrichs *et al.* 2006) was applied to convert human regions into sheep coordinates. Unmapped regions or not reciprocally mapped were allowed multiple mapping (liftOver, minMatchMulti >=0.80). The results from different datasets were then combined into a single dataset with non-overlapping regions.

Sheep Experimental ChIP-seq. Chromatin immunoprecipitation followed by next generation sequencing (ChIP-Seq) of the histone chromatin modification H3K4me3 and H3K27ac was performed on perirenal brown adipose tissue at 130 days post conception from three and two animals respectively. Sequence reads were mapped to the unmasked ovine genome sequence (*Ovis aries* Oar_v3.1.74) using the NGS core tool mapping application in CLCBIO (Peak calling comparing the H3K4me3 or H3K27ac ChiP-Seq versus the input control was performed using MACS (Zhang *et al.* 2008). Only peaks found in both replicates per chromatin mark, either H3K4me3 or H3K27ac, were further considered.

Validation of the sheep regulatory information. We produced 1,000 randomizations for each genomic feature using bedtools shuffle (-noOverlapping) (Quinlan 2014) set. Next, we calculated an empirical p-value per feature and overlap by counting how many times an equal or greater overlap observed in the original features was observed in the 1,000 randomizations.

RESULTS AND DISCUSSION

In order to annotate the sheep regulatory genome we selected human promoter and enhancer information from large-sequencing international consortiums such as FANTOM5, RoadMap Epigenomics and ENCODE (Table 1). Data from different databases differ in the biochemical process used to define enhancers and promoters, number of detected features, feature length and genome coverage (Table1). For example, RoadMap chromatin marks provide larger genome coverage (4.93% and 35.79% for promoters and enhancers, respectively) due to the capture of regulatory information from a larger number of conditions, namely 88 distinct human primary tissues.

To depict a potential sheep regulatory genome we converted human regulatory information into sheep coordinates. Table 1 shows that for each database we were able to successfully recover from 58.28% to 72.56% of their human regulatory information. It also shows that the recovery of proximal or promoter elements is higher (70%) compared to distal or enhancer elements (62%) in agreement with higher sequence conservation at the promoter than at the enhancer level. Based on these steps we captured 21.35% of the sheep genome as potentially regulatory (4.40% promoter and 16.95 % enhancer-like) (Table 2).

Next, we performed H3K27ac and H3K4me3 ChIP-seq in sheep late gestation perirenal brown adipose tissue. These chromatin marks indicate active chromatin and promoter regions respectively. A total of 35,366 regions were identified by H3K27ac, whereas 16,098 regions were identified as promoters using H3K4me3. 26,496 regions only enriched with H3K27ac and not H3K4me3 were defined as enhancers.

We assessed the recovery of sheep brown adipose H3K27ac for each converted dataset (Figure 1A) ranging from 12% recovery from FANTOM enhancers to 93% recovery from RoadMap Enhancers. Finally, to evaluate if the converted datasets were enriched for sheep regulatory information we performed 1,000 randomizations per dataset and compared their H3K27ac recovery with the original features (Figure 1B). Promoter and enhancer databases showed a clear enrichment

for sheep brown adipose regulatory regions compared to random (Figure 1B). However, Enhancers Roadmap dataset presented a much lower enrichment probably caused by presenting a higher number of features from multiple tissues that appear as false positives once compared to a single tissue, namely, brown adipose regulatory information. Next, ENCODE TFBSs (proximal and distal) are depleted for general brown active chromatin. This can be explained because these datatypes present higher sensitivity and specificity for promoters and enhancers (Figure 1 C-D) rather than general open-chromatin (H3K27ac). Thus, although depleted for the overlap with the ensemble of H3K27ac signal they are enriched for H3K4me3 and enhancer signal respectively (data not shown).

Table 1. Summary statistics of regulatory sequences

	Human					Sheep			
Database	# Features	# Merged	Avg bp	% Genome	% Mapped	# Features	# Merged	Avg bp	% Genome
Promoters FANTOM	201,802	198,710	20	0.13	70.43	142,140	137,779	18	0.09
Promoters RoadMap	1,771,836	146,860	1053	4.93	68.15	1,207,522	85,378	1099	3.62
Enhancers FANTOM	43,011	43,011	288	0.39	64.13	27,583	27,532	288	1.63
Enhancers RoadMap	9,928,635	494,583	2270	35.79	58.28	5,786,318	3,80,785	2180	32.09
ENCODE Proximal TFBSs	384,343	384,343	150	1.84	72.56	278,883	271,308	153	1.61
ENCODE Distal TFBSs	1,122,364	1,122,364	150	5.37	65.05	730,112	723,645	155	4.34

Table 2. Predicted sheep regulatory sequences

	# Features	Average bp	Total bp	% Sheep genome
Promoters	258472	441	113987479	4.40
Enhancers	387613	1131	438586881	16.95

To assess the sensitivity and specificity of each converted datatype we calculated their overlap with sheep promoters and enhancers (Figure 1 C-D). In this case we showed that promoters and ENCODE proximal TFBSs databases clearly recover most sheep adipose H3K4me3 peaks (Figure 1C). Thus, concluding that promoter datatypes recover mostly promoter regions rather than enhancers.

Alternatively, the same analysis at the enhancer level clearly showed that ENCODE Distal TFBSs is specific for enhancers recovering 71% of enhancers and only 33% of sheep promoters. However, the rest of enhancer databases do not only recovery enhancer regions but promoters as

well (Figure 1D). For example, RoadMap enhancers recovered 90% of sheep enhancers and 92% of sheep promoters. To solve that issue we only considered enhancers with no overlap to converted promoter datasets. This resulted in 62% of sheep enhancer recovery and only 13% promoter recovery.

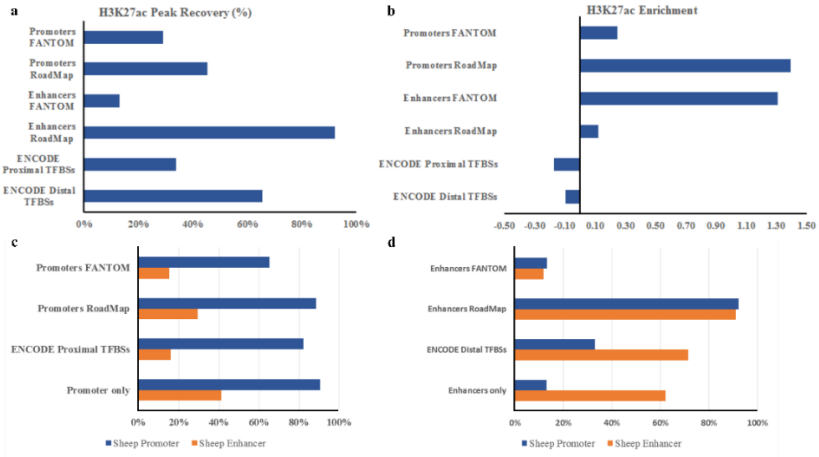

Figure 1. Recovery of experimentally defined sheep active chromatin, promoters and enhancers. (a) Percentage of recovery of sheep H3K27ac peaks from sheep brown adipose by the distinct sheep converted datatypes. (b) Fold enrichment compared to 1000 randomizations. Recovery of sheep promoters and enhancers, as a measure of specificity and sensitivity, by Promoter (c) and enhancer datasets (d).

CONCLUSIONS

Altogether, we show that the application of the HPRS pipeline successfully converts human regulatory information into sheep coordinates with potential regulatory function. This predicted regulatory map will allow the prioritization of trait-associated genetic variants, as well as further investigation and understanding between genetic variants, functional impact and phenotype. Further filtering of the dataset will be performed to increase the signal-to-noise ratio as performed in the original study (Nguyen *et al*. 2016) and then we will make this resource available to the sheep community.

REFERENCES

Andersson R., Gebhard C., Miguel-Escalada I., Hoof I., Bornholdt J., *et al.* (2014) *Nature* 507: 455.

Consortium T.F., Andersson L., Archibald A.L., Bottema C.D., Brauning R., *et al.* 2015. (2016) *Genome Biol* 16.

ENCODE Project Consortium. (2012) *Nature* 489: 57.

Forrest A.R.R, Kawaji H., Rehli M., Baillie J.K., de Hoon M.J.L., *et al.* (2014) Nature 507:462.

Hinrichs A.S., Karolchik D., Baertsch R., Barber G.P., Bejerano G., *et al.* (2006) *Nucleic Acids Res* 34: D590.

Nguyen, Q., Tellam, R.L., Kijas, J., Barendse, W. and Dalrymple B.P. (2016) *35th Proc. Int. Society for Animal Genetics Conference*, P1037

Quinlan A.R. (2014) *Curr Protoc Bioinforma Ed Board Andreas Baxevanis Al* 47: 11.12.1.

Roadmap Epigenomics Consortium, Kundaje A., Meuleman W., Ernst J., Bilenky M., *et al.* (2015). *Nature* 518: 317.

Zhang Y., Liu T., Meyer C.A., Eeckhoute J., Johnson D.S., *et al.* (2008) *Genome Biol* 9: R137.