RESEARCH REPORT
ANIMAL SCIENCE

Title: Genome wide identification and annotation of functional regulatory regions in livestock species - **NPB #**15-139

Investigator: Dr. Huaijun Zhou

Institution: University of California, Davis

Date Submitted: April 23, 2018

revised

Industry Summary
The objectives of this study were to identify previously unknown regions in the pig genome that act to regulate genes, and therefore play a role in the traits the pigs exhibit. Eight tissues from two adult male Yorkshire pigs were used for this study to profile a wide range of biological functions important to the industry. Biological assays were performed on these tissues to identify DNA-protein interactions that have been shown to be associated with promoters, enhancers, and insulators, which are important regulatory regions of the genome that exist outside of genes. A machine learning algorithm was used to integrate the data from these assays into predictions of regulatory elements for each tissue in the study. The results showed similarity with previous done in human and mouse, which supports the accuracy of the predictions. Identifying these regions will allow researchers to better understand the genetic basis for complex traits such as feed efficiency and disease resistance, which will in turn allow breeders to create new genetic lines of pigs to improve production efficiency, animal welfare, and food safety.

Key Findings:

- **Identify novel regulatory elements of genes that are potentially associated with economically important traits such as growth rate, feed efficiency and immune function.**
- **Functions of regulatory elements annotated on the genes provide novel targets for genetic selection and breeding.**
- **Assist in filtering SNPs on identification potential causative mutations associated with economically important traits.**

Keywords
Epigenetics, functional annotation, gene regulation

Scientific Abstract
Pigs are one of the major agricultural animal species as well as an important biomedical model. The domestic pig genome sequence was first reported in 2012 and has been regularly updated since, most recently in 2017. While the annotation of transcribed regions continues to improve, lately using PacBio whole-transcript sequencing, little has been reported in terms of non-coding regulatory regions of the pig genome. Functional non-coding portions of the genome are important contributors to phenotypes, as they regulate gene expression. As part of a FAANG pilot project, multiple tissues from two Yorkshire adult males were collected for chromatin profiling. Nuclei were isolated from fresh tissues and cryopreserved for assessing chromatin accessibility (ATAC-seq). Snap frozen tissue samples were also collected for transcriptomic (RNA-seq) and epigenomic (ChIP-seq) analysis. Datasets have already been generated for lung, liver and spleen including RNA-seq, ATAC-seq and ChIP-seq assays for H3K4me3, H3K27me3, H3K4me1, H3K27ac, and CTCF. Integration of these datasets using a multivariate Hidden Markov Model (ChromHMM) identified chromatin state maps for each tissue that correspond to active and inactive promoters, enhancers, and insulators. These chromatin states will be systematically analyzed and characterized to identify tissue specific and general regulatory elements. We expect that the functional annotations resulting from these study will help enhance our understanding of pig biology and assist in identification of functional genome variants and thus enlighten our interpretation of genome-wide association studies and allow fine-tuning of genomic selection approaches.

Introduction

Epigenetic regulatory elements are important functional regions of the genome that act to regulate genes and are a key factor in the genome to phenome problem. These regulatory elements can be identified using a variety of assays to profile open chromatin regions (ATAC-seq) associated with transcription factor binding and histone modifications (ChIP-seq) to locate specific chromatin states associated with active and repressed promoters, enhancers, and insulators. The histone modification H3K4me3 is associated with the promoters of actively transcribed genes, H3K27me3 is associated with polycomb repression, H3K27ac is associated with active promoter and enhancer elements, and H3K4me1 is associated with both active and inactive enhancers. CTCF is an insulator element that is helpful in predicting enhancer-promoter interactions. Creating a genome-wide catalog of these regulatory elements in adipose, cerebellum, cortex, hypothalamus, liver, lung, spleen, and skeletal muscle will provide a valuable tool for researchers to understand complex traits important to the industry, such as feed efficiency and disease resistance, and enable better selection methods for breeders to improve production efficiency, food safety, and animal welfare.

Objectives

1. Annotate chromatin states corresponding to DNase hypersensitivity, four histone modifications, and one transcription factor

2. Annotate promoters, enhancers, and silencers by integrating information from RNA-seq, DNAse-seq, and ChIP-seq against four histone modifications and one transcription factor.

3. Freely distribute all raw and annotated data via UCSC Genome Browser and Ensembl.

Materials & Methods

Tissue samples were collected from two castrated male Yorkshires at six months of age and flash frozen in liquid nitrogen. Nuclei were isolated from fresh tissue for use in ATAC-seq. Chromatin Immunoprecipitation (ChIP) was used to profile four histone modifications and the CTCF transcription factor across the pig genome. Single-end 50bp sequencing was done on an Illumina HiSeq-4000 machine to acquire a minimum of 20 million mapped and filtered reads for narrow marks (H3K4me3, H3K27ac, H3K4me1, CTCF) and 40 million mapped and filtered reads for broad marks (H3K27me3). Reads were mapped to the Sscrofa11.2 genome using the BWA program and then filtered to remove low quality alignments and multi-mapped reads. PCR duplicates were also removed. Next, peaks were called using the Macs2 peak caller on each replicate individually, and a final set of peaks was created for each tissue consisting of the peaks present in both replicates. Finally, integrative analysis of all five ChIP marks was done using the ChromHMM software developed as part of the ENCODE project. ChromHMM generated genome-wide chromatin state predictions that were assigned biologically meaningful names such as active promoter, enhancer, etc.

Results

The final numbers of aligned reads used for downstream analysis and the number of peaks called for each ChIP mark and tissue are shown in Table 1. Figure 1 shows the chromatin states of the ChromHMM model, with their biologically meaningful names, and a heatmap showing the presence of each ChIP mark in the state. Figure 2 shows the chromatin states predicted in each tissue in the region surrounding four genes.

Discussion

The genes shown in Figure 2 show examples of the type of information that these chromatin state predictions provide. For GAPDH, a housekeeping gene, active TSS chromatin states are predicted at the promoter in all three tissues. Similarly, the developmental HOX genes show repressive state predictions which is expected for adult animals. Two tissue-specific genes show prediction of active transcription only in their respective tissues. For researchers looking at a specific trait or gene of interest, these chromatin state predictions can be used along with GWAS or SNP data to identify candidate regulatory elements, such as enhancers, that may be the mechanistic cause of a phenotypic difference observed between genetic lines or under different experimental conditions. Functional studies can be designed to verify these mechanisms.

Work is on-going to extend the data to three brain tissues: cerebellum, cortex, and hypothalamus, as well as muscle and adipose. ATAC-seq is also being performed on these samples to profile chromatin accessibility.

Table 1: Number of aligned and filtered reads used in peak calling.

Tissue	Rep	H3K4me3	H3K27me3	H3K27ac	H3K4me1	CTCF
Liver	P348	24,023,377	49,355,967	39,709,894	23,660,577	16,035,277
Liver	P350	24,699,858	46,969,846	28,489,244	47,953,115	22,119,295
Lung	P348	30,797,614	50,482,161	13,506,676	12,746,016	17,446,813
Lung	P350	30,712,498	53,376,935	12,262,285	57,501,932	27,803,348
Spleen	P348	22,872,094	46,283,864	47,788,995	36,312,224	19,334,802
Spleen	P350	35,387,791	52,784,718	41,856,656	53,261,792	25,970,879

Table 2: Peaks common in both replicates for each tissue.

Tissue	H3K4me3	H3K27me3	H3K27ac	H3K4me1	CTCF
Liver	31,565	69,015	42,737	139,771	21,485
Lung	27,712	18,988	29,616	24,942	23,193
Spleen	26,752	36,578	41,559	70,112	35,082

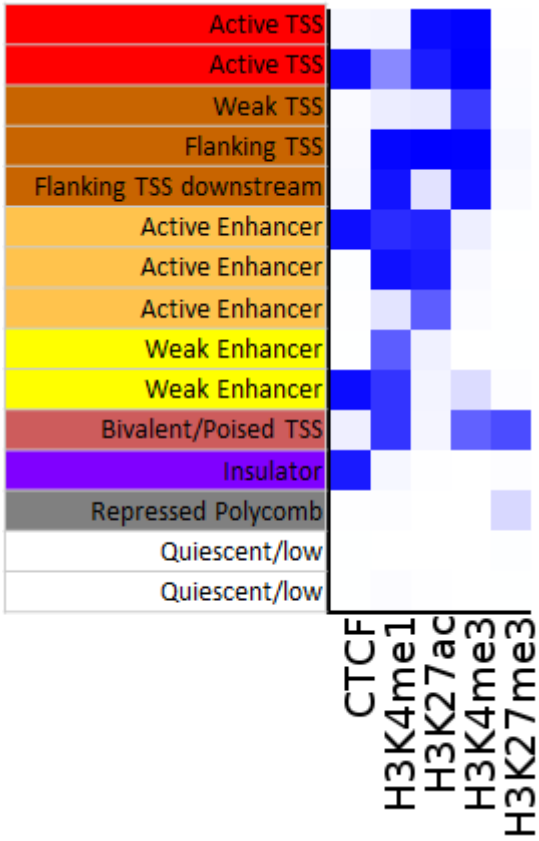

Figure 1: ChromHMM model trained using ChIP-seq alignments to Sscrofa11.2 genome. Biologically meaningful names are assigned based on presence of histone modifications and the genome locations that each state is predominantly found.

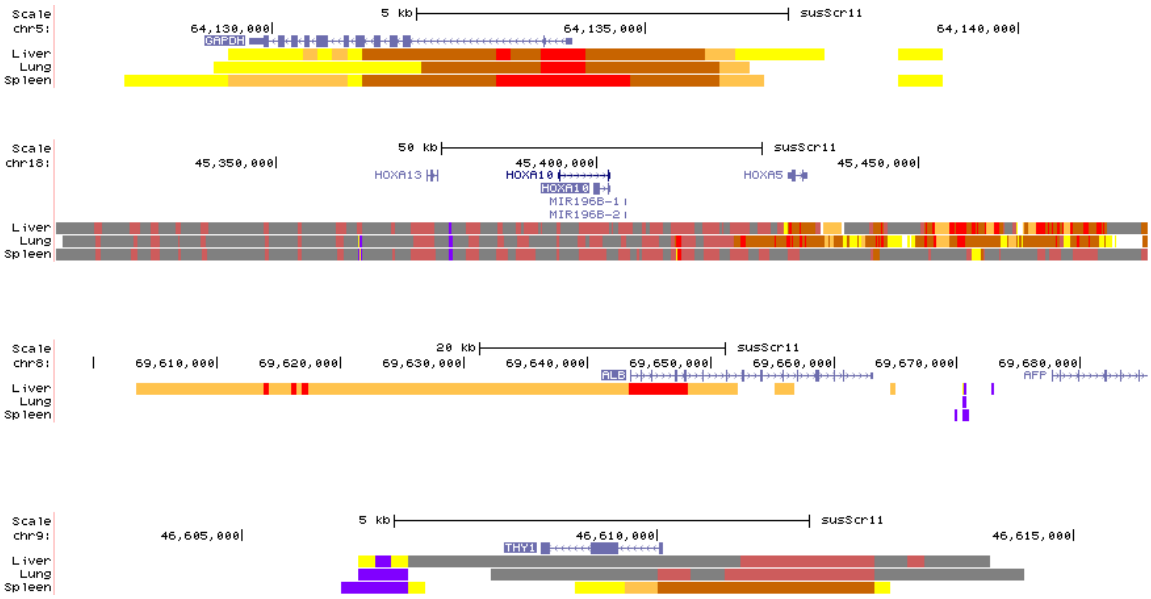

Figure 2: Examples of the chromatin state predictions in liver, lung, and spleen at four locations on the genome. Different colors indicate different predicted chromatin states (see Figure 2 for meaning of the colors). GAPDH is a housekeeping gene and the TSS is predicted as active in all three tissues. The HOX gene cluster shown in the second location are developmental genes and expected to be inactive in adults, which is supported by the repressed polycomb state predicted across this region. The third location is the liver-specific gene ALB, which is only predicted as an active TSS in liver. Finally, the THY1 gene is a spleen-specific gene which is supported by the repressed prediction in liver and lung and the weak TSS prediction in spleen.