Profiling the landscape of transcription, chromatin accessibility and chromosome conformation of cattle, pig, chicken and goat genomes

[FAANG pilot project “FR-AgENCODE”]

Sarah Djebali
INRA - Toulouse - France
Genetics, Physiology and Breeding Systems Laboratory

FR-AgENCODE data

Hypersensitive Sites

5C ChIA-PET Hi-C

DNase-seq FAIRE-seq ATAC-seq

ChIP-seq

WGBS RRBS methyl array

Computational predictions

RNA-seq

CLIP-seq RIP-seq

Long-range regulatory elements (enhancers, repressors/silencers, insulators)

Promoters

Transcripts

Bos taurus

Capra hircus

Gallus gallus

Sus scrofa

Holstein

Alpine

White Legorn

Large White

2 males

2 females

Liver

CD4

CD8

Based on an image by Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI)
## FR-AgENCODE data generation progress

<table>
<thead>
<tr>
<th>Species</th>
<th>ATAC-seq*</th>
<th>HiC (liver only)</th>
<th>Long RNA-seq</th>
<th>Small RNA-seq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bos taurus</td>
<td></td>
<td>**</td>
<td></td>
<td>**</td>
</tr>
<tr>
<td>Capra hircus</td>
<td></td>
<td>**</td>
<td></td>
<td>**</td>
</tr>
<tr>
<td>Gallus gallus</td>
<td></td>
<td>**</td>
<td></td>
<td>**</td>
</tr>
<tr>
<td>Sus scrofa</td>
<td></td>
<td>**</td>
<td></td>
<td>**</td>
</tr>
</tbody>
</table>

* No liver for bos_taurus ATAC-seq / ** Integrative analysis on-going

Our 16 animals are called cattle 1,2,3,4, goat 1,2,3,4, chicken 1,2,3,4 and pig 1,2,3,4:

- Animals 1 and 2 are **males**
- Animals 3 and 4 are **females**
RNA-seq for expression and annotation of long coding and non-coding RNAs

- **RNAs:**
  - longer than 200bp
  - Polya-selection before cDNA synthesis

- **Sequencing:**
  - Directed
  - 2x150bp
  - 100 million read pairs per sample
RNA-seq data processing and analysis pipeline

Sample 1 RNA-seq reads

Mapping
STAR 2.5.1b

Transcript modelling
 cufflinks 2.2.1

Mapping
STAR 2.5.1b

Ref Tr/Gn quantification
RSEM 1.3.0

Tr/Gn quantification
RSEM 1.3.0

Ref Tr/Gn quantification
RSEM 1.3.0

Sample N RNA-seq reads

Mapping
STAR 2.5.1b

Transcript modelling
 cufflinks 2.2.1

Ref Tr/Gn quantification
RSEM 1.3.0

Ref Tr/Gn quantification
RSEM 1.3.0

Sample N RNA-seq reads

Mapping
STAR 2.5.1b

Transcript modelling
 cufflinks 2.2.1

Tr/Gn quantification
RSEM 1.3.0

Mapping
STAR 2.5.1b

Ref Tr/Gn quantification
RSEM 1.3.0

Novel gene annotation

cuffmerge 2.2.1

Novel Tr/Gn quantif.
Sample N

Novel Tr/Gn quantif.
Sample 1

Ref. gene annotation

Tr/Gn quantification
RSEM 1.3.0

Ref Tr/Gn quantif. Sample 1

Novel Tr/Gn quantif.
Sample 1

Novel Tr/Gn quantif.
Sample N

PDF introduction

LncRNA set

Differential gene expression (DE) analysis
EdgeR 3.18.0

Nextflow implementation @ https://github.com/skptic/lncRNA-Annotation-nf
Many novel transcripts are found, which are globally less expressed than the annotated ones.

<table>
<thead>
<tr>
<th>Species</th>
<th>Genome / Gene annotation</th>
<th>Annotated Transcripts</th>
<th>Number of novel transcripts detected*</th>
<th>Number of novel IncRNAs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Total number</td>
<td>Detected*</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>#</td>
<td>% of total</td>
<td></td>
</tr>
<tr>
<td>Bos taurus</td>
<td>UMD 3.1 / Ensembl 84</td>
<td>26,740</td>
<td>16,100</td>
<td>60.2</td>
</tr>
<tr>
<td>Capra hircus</td>
<td>CHIR_ARS 1 / NCBI</td>
<td>53,266</td>
<td>34,442</td>
<td>64.7</td>
</tr>
<tr>
<td>Gallus gallus</td>
<td>GalGal 5 / Ensembl 87</td>
<td>38,118</td>
<td>22,898</td>
<td>60.1</td>
</tr>
<tr>
<td>Sus scrofa</td>
<td>SScrofa 10.2 / Ensembl 84</td>
<td>30,585</td>
<td>18,746</td>
<td>61.3</td>
</tr>
</tbody>
</table>

* with TPM ≥ 0.1 in ≥ 2 samples

See Kevin Muret’s talk just after this talk.
Biological process (BP), molecular function (MF), cell compartment (CC) GO term enrichment on DE genes is consistent with the underlying biology.

Liver over-expressed genes

T cell over-expressed genes
All species RNA-seq hierarchical clustering first separates liver from immune cells, and then species...
ATAC-seq for open chromatin regions

50 million read pairs per sample

<table>
<thead>
<tr>
<th>Species</th>
<th>Number of atac-seq peaks</th>
<th>Genome size (bp)</th>
<th>ATAC-seq peak coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bos taurus</td>
<td>58,384</td>
<td>2,670,422,299</td>
<td>42,288,741</td>
</tr>
<tr>
<td>Capra hircus</td>
<td>46,901</td>
<td>2,922,813,246</td>
<td>32,951,265</td>
</tr>
<tr>
<td>Gallus gallus</td>
<td>116,893</td>
<td>1,230,258,557</td>
<td>50,931,713</td>
</tr>
<tr>
<td>Sus scrofa</td>
<td>120,914</td>
<td>2,808,525,991</td>
<td>72,480,471</td>
</tr>
</tbody>
</table>

Buenrostro et al., Nature Methods, 2013

Buenrostro et al., Nature Methods, 2013

Buenrostro et al., Nature Methods, 2013

Buenrostro et al., Nature Methods, 2013
ATAC-seq peaks at Transcription Start Sites (TSS); those peaks are more ubiquitous

Open chromatin regions located close to TSSs are more ubiquitous than other open chromatin regions

ATAC-seq coverage peaks at annotated TSS in all species
ATAC-seq PCA first separates liver from immune cells; in absence of liver, ATAC-seq PCA first separates males from females
Differential ATAC-seq peaks are more likely to be regulatory.

Between tissue differential ATAC-seq peaks have a higher TFBS density than non differential ATAC-seq peaks (Wilcoxon test, p-value < $10^{-15}$)

→ Differential ATAC-seq peaks are more likely to have a regulatory role
HiC for 3D genomic structure

- **HiC data (liver of 4 animals of 3 species):**
  - 180 million read pairs per sample

- **HiC data analysis pipeline:**
  - HiC-Pro: Read Mapping/Filtering/Normalization
  - Armatus: TAD calling
  - HiTC: A/B compartment calling

---

Rao et al, Cell, 2014

Belton et al, Methods, 2012
Predicted CTCF binding sites peak at Topologically Associating Domain (TAD) boundaries

Li et al, Scientific Reports, 2016
There is global consistency between RNA-seq, ATAC-seq and HiC data

A compartments: open, expressed
B compartments: closed, repressed

See Sylvain Foissac’s talk at the *Pig Genetics and Genomics* session tomorrow Tuesday (talk # 144)
Acknowledgements

- **Management:**
  - Elisabetta Giuffra
  - Sylvain Foissac
  - Sandrine Lagarrigue
  - Marie-Hélène Pinard

- **Sampling:**
  - Michèle Tixier-Boichard
  - Stéphane Fabre
  - Gwenola Tosser-Klopp
  - Pascale Quéré
  - Fany Blanc
  - Fabrice Laurent

- **Assays:**
  - Hervé Acloque
  - Diane Esquerre
  - Sophie Pollet
  - Adeline Goubil
  - Florence Mompart
  - Françoise Drouet
  - Silvia Vincent-Nailleau

- **Data analysis:**
  - Kylie Munyard
  - Cédric Cabau
  - Nathalie Villa-Vialaneix
  - Matthias Zytnicki
  - Kévin Muret
  - Andrea Rau
  - Thomas Derrien
  - Christine Gaspin
  - Christophe Klopp
  - Ignacio Gonzalez
  - David Robelin
  - Magali San Cristobal
  - Maria Marti
  - Sylvain Marthey
  - Philippe Bardou

Thanks for your attention!
Additional slides
Transcripts detected and their expression

- **bos_taurus**
- **capra_hircus**
- **gallus_gallus**
- **sus_scrofa**

*class*
- 1_annotated
- 2_extension
- 3_antisense
- 4_intergenic

# experiments with TPM >= 0.1
All RNA-seq experiment PCA (~7300 orthologous genes)
Expression of the 13 T cell TF common to the 4 species (adding goat using gene name)
Expression of the 18 metabolism TF common to the 4 species (adding goat using gene name)
HiC results: interaction matrices
The normalized matrix shows many large blocks of enriched and depleted interactions, generating a plaid pattern (Fig. 3B). If two loci (here 1-Mb regions) are nearby in space, we reasoned that they will share neighbors and have correlated interaction profiles. We therefore defined a correlation matrix C in which $c_{ij}$ is the Pearson correlation between the $i$th row and $j$th column of $M^*$. This process dramatically sharpened the plaid pattern (Fig. 3C);

The plaid pattern suggests that each chromosome can be decomposed into two sets of loci (arbitrarily labeled A and B) such that contacts within each set are enriched and contacts between sets are depleted. We partitioned each chromosome in this way by using principal component analysis. For all but two chromosomes, the first principal component (PC) clearly corresponded to the plaid pattern (positive values defining one set, negative values the other) (fig. S1). For chromosomes 4 and 5, the first PC corresponded to the two chromosome arms, but the second PC corresponded to the plaid pattern. The entries of the PC vector reflected the sharp transitions from compartment to compartment observed within the plaid heatmaps.
Method

- Get contact matrix (raw counts)
- Extract intra-chromosomal sub-matrices (& regenerate corresponding indices)
- For each chrom independently:
  - ICE-normalize counts (matrix balancing)
  - Normalize by expected counts (scale by the distance factor)
  - Run a PCA of the bins using these counts => “direct” method
  - Generate pearson correlation matrix
  - Run a PCA on the bins using the correlations => “corr” method
  - Run a PCA on the bins using the correlations with HitC package (that does additional filtering) => “hitc” method
  - Extract from the PCA of each method the 3 first PCs
  - Choose the PC by comparing with the ICE-normalized counts on the diagonal (2 ways: PC vs. count correlation or t.test PC sign vs. counts)
- Segmentation by merging adjacents bins with same PC sign (+/-) and assign A/B compartment using the sign of the PC vs. count correlation.
HiC read density peaks at distal ATAC-seq peaks but is depleted at TSS ATAC-seq peaks