An update on the FAANG pilot project
FR-AgENCODE

S. Foissac, S. Djebali, Andrea Rau, S. Lagarrigue, H. Acloque, E. Giuffra, and the FR-AgENCODE group

INRA - {Jouy en Josas, Toulouse, Rennes} - France

FR-AgENCODE data

Based on an image by Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI)
Overview of FR-AgENCODE analyses

Multi-species single data analyses:
- Expression (RNA-seq)
- Open chromatin (ATAC-seq)
- 3D chromatin conformation (Hi-C)

Quantification, normalization, differential analysis, clustering, ...

Multi-species integrative analyses:
- Expression + open chromatin + 3D chromatin conformation
- Expression + open chromatin
RNA-seq identifies many novel transcripts

<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 lncRNAs**</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 90</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 90</td>
<td>38,118</td>
<td>22,180</td>
<td>58.2</td>
</tr>
<tr>
<td>Sus scrofa</td>
<td>SScrofa 11.1 / Ensembl 90</td>
<td>49,448</td>
<td>29,786</td>
<td>60.2</td>
</tr>
</tbody>
</table>

* with TPM ≥ 0.1 in ≥ 2 samples / ** with at least 2 exons and classified by FEELnc

Directed, 2 x 150bp, 100 million read pairs / sample

Mapped reads:
- **Exonic**: 50-80%
- **Intronic**: 10-20%
- **Intergenic**: 20% for cattle/chicken, 10% for goat/pig
Differentially expressed genes reflect underlying biology

Two per-gene (TPM > 0.1 in at least 2 samples) differential models:

- **Tissue effect**, blocking on individuals:
  \[
  \text{expression} \sim \text{individual} + \text{tissue}
  \]

- **Tissue & tissue-specific sex effects**, blocking on (nested) individuals:
  \[
  \text{expression} \sim \text{sex} + \text{tissue:sex} + \text{individual:sex}
  \]

- Generally large differences for liver vs CD4/CD8, smaller differences for CD4 vs CD8 or tissue-specific ♂ vs ♀

- GO (BP) enrichment for genes consistently DE across species:

  Over-expressed in **liver**

  ![Bar chart showing over-expressed biological processes in liver]

  - o xoacid metabolic process
  - small molecule catabolic process
  - monocarboxylic acid catabolic process
  - carboxylic acid catabolic process
  - small molecule biosynthetic process
  - oxidation–reduction process
  - lipid oxidation
  - fatty acid beta-oxidation
  - cell adhesion
  - alpha–amino acid catabolic process

  Over-expressed in **immune cells**

  ![Bar chart showing over-expressed biological processes in immune cells]

  - immune response–activating cell surface receptor signaling pathway
  - activation of immune response
  - positive regulation of immune system process
  - immune response–regulating signaling pathway
  - hematopoietic or lymphoid organ development
  - T cell receptor signaling pathway
  - leukocyte differentiation
  - cellular response to stimulus
  - lymphocyte activation
  - regulation of leukocyte activation
Liver clearly separated from CD4/CD8 cells, and species cluster within cell types (liver vs T-cells)
ATAC-seq identifies pertinent regions of open chromatin

- 3 major peak classes: promoter regions (TSS +/- 1kb), intronic, intergenic
- Significant proportion of peaks in ATAC-seq reads found in promoter region (36-66%)
- For all species, peaks close to TSS (+/- 5kb) are most ubiquitous open chromatin

![Diagram of ATAC-seq process](image)

**Table: ATAC-seq peak coverage**

<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>Bos taurus</td>
<td>104,986</td>
<td>2,670,422,299</td>
<td></td>
</tr>
<tr>
<td>Capra hircus</td>
<td>74,806</td>
<td>2,922,813,246</td>
<td>57,043,999 1.95</td>
</tr>
<tr>
<td>Gallus gallus</td>
<td>119,894</td>
<td>1,230,258,557</td>
<td>51,000,066 4.15</td>
</tr>
<tr>
<td>Sus scrofa</td>
<td>149,334</td>
<td>2,501,912,388</td>
<td>106,645,814 4.26</td>
</tr>
</tbody>
</table>

* 50 million read pairs / sample
ATAC-seq sample clustering:
Liver vs immune cells, ♂ vs ♀ in immune cells

- PC1: liver versus CD4/CD8
- PC2: CD4 vs CD8

- PC1: males vs females
  (note: absence of liver samples)
- PC2: CD4 vs CD8
Differential ATAC-seq peaks are more likely to be regulatory

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

→ Differential ATAC-seq peaks are more likely to have a regulatory role
Hi-C for 3D genomic structures

Several levels for 3D conformation:
- TADs (Topologically associating domains)
- Compartments
- Enhancer-promoter interactions

- Liver x 4 animals x 3 species (all but cattle)
- 180 million read pairs / sample
- Predicted CTCF binding sites peak at TAD boundaries

[A diagram showing active and inactive states with TADs, histones, and DNA interactions.]
RNA-seq + ATAC-seq + Hi-C consistency

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

Global consistency between expression, chromatin accessibility, and chromatin conformation!
Exploring correlation between promoter accessibility and gene expression

- Pattern unique to **proximal promoter region** (TSS +/- 1kb)
- Promoter peaks negatively correlated with gene expression tend to be **wider**
Next steps & future work

- **More integrative analyses:**
  - Enhancer / gene relationships
  - Small RNAs vs long RNAs / open chromatin / HiC TADs

- **More comparative analyses:**
  - Evolution of functional elements, in particular regulatory
  - Human/mouse element projection to livestock species

- bioRxiv preprint to be submitted in coming weeks
- More tissues, functional validation and link to genotype/phenotype (G/P) data:
  - H2020 project proposal (co-coordinated by E. Giuffra & H. Acloque)
  ...
Fr-AgENCODE group

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

- **Sampling and assays:**
  - Hervé Acloque
  - Cécile Berri
  - Fany Blanc
  - Françoise Drouet
  - Diane Esquerre
  - Stéphane Fabre
  - Joel Gautron
  - Adeline Goubil
  - Fabrice Laurent
  - Florence Mompart
  - Sophie Pollet
  - Pascale Quéré
  - Michèle Tixier-Boichard
  - Gwenola Tosser-Klopp
  - Silvia Vincent-Nailleau

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

Thanks for your attention!