Functional annotation of livestock genomes
Chromatin structure and gene expression

Sylvain Foissac, INRA Toulouse, France
ASAS-ADSA Midwest Meeting, March 2019, Omaha, USA
Functional Annotation of ANimal Genomes

Genome → predict understand → Phenome
Functional Annotation of ANimal Genomes

Genome — predict — Phenome

raw DNA sequence — genomic annotation
Functional Annotation of ANimal Genomes

Genome predict understand Phenome

& =>

ENCODE
Functional Annotation of ANimal Genomes

Genome predict understand Phenome

& =>  ?
The FAANG pilot project FR-AgENCODE

FR-AgENCODE: the french pilot project of FAANG from INRA
The FAANG pilot project FR-AgENCODE

♀ x2
♂ x2

Bos taurus (Holstein)  Capra hircus (Alpine)  Gallus gallus (White Leghorn)  Sus scrofa (Large White)

CD4+ & CD8+ sorted primary T cells  Liver tissue dissociated cells

RNA-seq gene expression  ATAC-seq chromatin accessibility  Hi-C genome 3D topology

data analysis
Step 1: animal and tissue sampling

- 4 animals per species
- 40+ tissues per animal
  Liver, CD4+ T cells, CD8+ T cells, sperm, plasma, heart, lung, skin, fat, duodenum, ileum, jejenum, cerebellum, frontal lobe, olfactory bulb, trigeminal ganglia, hypotalamus, pancreas, andrenals, kidney, muscle, bone, joints, spleen, lymphatic nodes, peyer's patches, ovary, oocytes, oviduct, uterus, mammary gland, acini, testis, seminal vesicle, etc.
- 2,000+ frozen samples
- Protocols and BioSamples entries at the EMBL-EBI
- Samples available at the INRA Bridge/CRB-anim biobank

Index of ftp://ftp.faang.ebi.ac.uk/ftp/protocols/samples/

- Up to higher level directory

<table>
<thead>
<tr>
<th>Name</th>
<th>Size</th>
<th>Last Modified</th>
</tr>
</thead>
<tbody>
<tr>
<td>File: DEDJTR_SOP_CryofreezingTissue_20160317.pdf</td>
<td>69 KB</td>
<td>3/22/16 12:00:00 AM GMT+1</td>
</tr>
<tr>
<td>File: INRA_SOP_PBMC_purification_cattle_caprina_201...</td>
<td>60 KB</td>
<td>5/4/16 12:00:00 AM GMT+2</td>
</tr>
<tr>
<td>File: INRA_SOP_PBMC_seperation_swine_blood_2016...</td>
<td>89 KB</td>
<td>5/4/16 12:00:00 AM GMT+2</td>
</tr>
<tr>
<td>File: INRA_SOP_alveolar_macrophages_mammals_sam...</td>
<td>486 KB</td>
<td>7/29/16 12:00:00 AM GMT+2</td>
</tr>
<tr>
<td>File: INRA_SOP_chicken_splenocytes_sampling_2016...</td>
<td>131 KB</td>
<td>7/29/16 12:00:00 AM GMT+2</td>
</tr>
<tr>
<td>File: INRA_SOP_liver_spleen_mammarygland_forHiC_s...</td>
<td>327 KB</td>
<td>7/29/16 12:00:00 AM GMT+2</td>
</tr>
<tr>
<td>File: INRA_SOP_oocytes_granulosa_mammals_sampling_2016...</td>
<td>246 KB</td>
<td>7/29/16 12:00:00 AM GMT+2</td>
</tr>
</tbody>
</table>
Step 2: the molecular assays

- RNA-seq: transcriptome profiling
  - Poly-A selection, fragmentation and random priming
  - First and second strand cDNA synthesis
  - End-repair, phosphorylation and A-tailing
  - Adapter ligation, PCR amplification and sequencing

- ATAC-seq: chromatin accessibility
  - Assay for Transposase Accessible Chromatin
  - Open DNA
  - Tn5 Transposome
  - Insert in regions of open chromatin
  - Fragmented and primed
  - DNA purification
  - Amplification

- Hi-C: chromosome conformation
  - Crosslink DNA
  - Cut with restriction enzyme
  - Fill ends and mark with biotin
  - Ligate
  - Purify and shear DNA; pull down biotin
  - Sequence using paired-ends

From www.illumina.com

Diane Esquerré
Hervé Acloque

Rao et al, Cell, 2014
Step 2: the molecular assays

**Completed experiments**

<table>
<thead>
<tr>
<th>RNA-seq</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cattle</td>
<td>cd4</td>
<td>cd8</td>
<td>liver</td>
<td></td>
<td>Goat</td>
<td>goat1</td>
<td>goat2</td>
<td>goat3</td>
</tr>
<tr>
<td></td>
<td>cd4</td>
<td></td>
<td></td>
<td></td>
<td>cd4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>cd8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>liver</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chicken</td>
<td>chicken1</td>
<td>chicken2</td>
<td>chicken3</td>
<td>chicken4</td>
<td>Pig</td>
<td>pig1</td>
<td>pig2</td>
<td>pig3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>cd4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>cd8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>liver</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ATAC-seq</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cattle</td>
<td>cd4</td>
<td>cd8</td>
<td>liver</td>
<td></td>
<td>Goat</td>
<td>goat1</td>
<td>goat2</td>
<td>goat3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>cd4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>cd8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>liver</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chicken</td>
<td>chicken1</td>
<td>chicken2</td>
<td>chicken3</td>
<td>chicken4</td>
<td>Pig</td>
<td>pig1</td>
<td>pig2</td>
<td>pig3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>cd4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>cd8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>liver</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hi-C</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cattle</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Goat</td>
<td>goat1</td>
<td>goat2</td>
<td>goat3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>cd4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>cd8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>liver</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chicken</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Pig</td>
<td>pig1</td>
<td>pig2</td>
<td>pig3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>cd4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>cd8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>liver</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **RNA-seq:** ~5B read pairs
- **ATAC-seq:** ~3B read pairs
- **Hi-C:** ~2B read pairs

>80% of experiments completed
### Step 3: data analysis

#### Bioinformatics pipelines

<table>
<thead>
<tr>
<th>RNA-seq</th>
<th>ATAC-seq</th>
<th>HiC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read trimming (trimgalore)</td>
<td>Read trimming (trimgalore)</td>
<td>Read trimming (cutadapt)</td>
</tr>
<tr>
<td>Read mapping (STAR2)</td>
<td>Read mapping to genome (Bowtie2)</td>
<td>Read mapping to genome (Bowtie2)</td>
</tr>
<tr>
<td>Transcript modelling (Cufflinks2)</td>
<td>PCR duplicate removal (Samtools)</td>
<td>Inconsistent pairs filtering (Samtools)</td>
</tr>
<tr>
<td>New gene annotation (Cuffmerge2)</td>
<td>Mitochondrial read removal (Samtools)</td>
<td>Contact matrix generation and normalization (ICE)</td>
</tr>
<tr>
<td>Transcript/gene expression quantification (RSEM)</td>
<td>Peak calling (MACS2)</td>
<td>TAD calling (Armatus)</td>
</tr>
<tr>
<td>LncRNA calling (FEELnc)</td>
<td></td>
<td>Visualization (Juicebox)</td>
</tr>
</tbody>
</table>

#### Data integration and comparison

#### Statistical analyses
Step 3: data analysis

Bioinformatics pipelines

RNA-seq

- Read trimming (trimgalore)
- Read mapping (STAR2)
- Transcript modelling (Cufflinks2)
- New gene annotation (Cuffmerge2)
- Transcript/gene expression quantification (RSEM)
- LncRNA calling (FEELnc)
Hi-C detailed workflow
- Trim reads (ligation site)
- Map on reference genome
- Discard inconsistent pairs
- Count reads in pairs of genomic bins & generate contact matrix
- Normalize contact matrix (non parametric, matrix balancing)
- Generate html report
- Identify Topological Associated Domains, *cis* and *trans* interactions
- Identify A and B compartments

Software
- HiC-Pro pipeline (Servant et al 2015)
- Bowtie2 mapping (Langmead et al, 2009)
- ICE normalization (Imakaev et al, 2012)
- HiTC display and A/B comp. (Servant et al, 2012)
- HiFive pipeline (Sauria et al, 2015)
- Armatus TAD finding (Filippova et al, 2014)
- Juicebox browser (Durand et al, 2016)
Data overview: RNA-seq & ATAC-seq
Results: the transcriptomic landscape

- most of all the known transcripts are detected in liver and T cells
- reference annotation can be extended by a factor 2 to 3
- most of the new transcripts are alternative isoforms of coding genes

<table>
<thead>
<tr>
<th>Species</th>
<th>Reference transcripts</th>
<th>FR-AgENCODE transcripts</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All</td>
<td>Expressed</td>
</tr>
<tr>
<td>Cattle</td>
<td>26,740</td>
<td>16,100</td>
</tr>
<tr>
<td>Goat</td>
<td>53,266</td>
<td>34,442</td>
</tr>
<tr>
<td>Chicken</td>
<td>38,118</td>
<td>22,180</td>
</tr>
<tr>
<td>Pig</td>
<td>49,448</td>
<td>29,786</td>
</tr>
</tbody>
</table>
Results: the transcriptomic landscape

- most of all the known transcripts are detected in liver and T cells
- reference annotation can be extended by a factor 2 to 3
- most of the new transcripts are alternative isoforms of coding genes
- differential expression between liver and T cells:
  - most of the genes are Differentially Expressed (DE)

<table>
<thead>
<tr>
<th></th>
<th>Number of DE reference genes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cattle</td>
</tr>
<tr>
<td>liver &gt; T cells</td>
<td>4,992</td>
</tr>
<tr>
<td>T cells &gt; liver</td>
<td>3,943</td>
</tr>
</tbody>
</table>
Results: the transcriptomic landscape

- most of all the known transcripts are detected in liver and T cells
- reference annotation can be extended by a factor 2 to 3
- most of the new transcripts are alternative isoforms of coding genes
- differential expression between liver and T cells:
  - most of the genes are Differentially Expressed (DE)
  - DE genes have consistent GO functions
Results: the transcriptomic landscape

Hierarchical clustering using orthologous genes

Samples cluster first by tissue, then by species
Results: the chromatin accessibility landscape

- 75,000-150,000 accessibility sites by species (~2-4% of the genome)
- Most of them are intergenic & intronic
Results: the chromatin accessibility landscape

- 75,000-150,000 accessibility sites by species (~2-4% of the genome)
- Most of them are intergenic & intronic
- Promoter accessibility: max within 1Kb upstream of gene starts

Mean ATAC-seq score around and within genes
Results: the chromatin accessibility landscape

- 75,000-150,000 accessibility sites by species (~2-4% of the genome)
- Most of them are intergenic & intronic
- Promoter accessibility: max within 1Kb upstream of gene starts
- Comparative analysis between liver and T cells
  - 5,000 to 13,000 sites are Differentially Accessible (DA) by species
Results: the chromatin accessibility landscape

- 75,000-150,000 accessibility sites by species (~2-4% of the genome)
- Most of them are intergenic & intronic
- Promoter accessibility: max within 1Kb upstream of gene starts
- Comparative analysis between liver and T cells
  - 5,000 to 13,000 sites are Differentially Accessible (DA) by species
  - more Transcription Factor Binding Sites in DA vs. non-DA sites

=> support for a regulatory role
Chromatin structure & 3D genome

nuclear genome
(a few Gb)

open, active

A/B compartments
(\( \sim n \times Mb \))

closed, inactive

“A”

TADs
Topologically Associating Domains
(n x 0.1 Mb)

regulatory loops
(N x Kb)

regulators
(enhancers, TFBS… n x 100bp)

Giorgetti et al., 2013, Genome Biol.

Shlyueva et al, Nat. Rev. Genet., 2014
Data overview: the Hi-C interaction matrix

Sus scrofa
chrom #1
pig #1

Genomic position (Mb)
The A/B (epi)genome compartments
The A/B (epi)genome compartments
The A/B (epi)genome compartments

“A” compartments

“B” compartments
The A/B (epi)genome compartments

Sus scrofa
chrom #2
The A/B (epi)genome compartments

Links between genome structure & function

RNA-seq gene expression

ATAC-seq chromatin accessibility

A/B compartments

- capra_hircus
- gallus_gallus
- sus_scrofa
The A/B (epi)genome compartments

Gene distribution in A/B compartments

- Goat: 7844 genes in A (65.5%) vs. 4133 in B (34.5%)
- Chicken: 4571 genes in A (64.0%) vs. 2576 in B (36.0%)
- Pig: 6737 genes in A (63.4%) vs. 3883 in B (36.6%)
The A/B (epi)genome compartments

- Gene distribution in A/B compartments
  - Goat: 7844 genes in A (65.5%) vs. 4133 in B (34.5%)
  - Chicken: 4571 genes in A (64.0%) vs. 2576 in B (36.0%)
  - Pig: 6737 genes in A (63.4%) vs. 3883 in B (36.6%)

- Focusing on orthologous genes: compartments across species
  - “A” in the 3 species
    - expected: 1529 genes (26.7%)
    - observed: 2972 genes (51.9%)
  - “B” in the 3 species
    - expected: 259 genes (4.5%)
    - observed: 611 genes (10.7%)

(N=5728 orthologous genes with an assigned compartments in the 3 species)

**conservation of genome compartments** => **evidence of functional role**
Topologically Associating Domains (TADs)

Li et al, 2016, Scientific Reports

Sus scrofa
Topologically Associating Domains (TADs)

Sus scrofa
Topologically Associating Domains (TADs)

- Identify “orthologous” TAD boundaries (between 2 or 3 species):
  - N = 16,468 non ambiguous hits
    - 10,805 from 1 species => level zero
    - 5592 from 2 species => level one
    - 71 from 3 species => level two

- Compute interaction score across TAD boundaries of each level
  (the lower the score, the stronger the insulation)
Topologically Associating Domains (TADs)

- Identify “orthologous” TAD boundaries (between 2 or 3 species):
  N = 16,468 non ambiguous hits
  - 10,805 from 1 species => level zero
  - 5592 from 2 species  => level one
  - 71 from 3 species    => level two

- Compute interaction score across TAD boundaries of each level (the lower the score, the stronger the insulation)

  conserved TAD boundaries have stronger insulations => evidence of selective pressure
Conclusion

- The FR-AgENCODE project contributes to the FAANG action
- Substantial extension of the transcriptional map
- Identification of potential regulatory sites with accessible chromatin
- Integrative analysis across assays and/or across species: a powerful approach to investigate gene expression
- Chromatin conformation is under selective pressure at various organizational levels: accessibility sites, TADs & compartments

Conservation

Structure  ⇐  Function
# Acknowledgments

<table>
<thead>
<tr>
<th>Management</th>
<th>Sampling and assays</th>
<th>Data analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elisabetta Giuffra (coordination)</td>
<td>Hervé Acloque</td>
<td>Philippe Bardou</td>
</tr>
<tr>
<td>Sylvain Foissac (coordination)</td>
<td>Cécile Berri</td>
<td>Cédric Cabau</td>
</tr>
<tr>
<td>Sandrine Lagarrigue</td>
<td>Fany Blanc</td>
<td>Elisa Crisci</td>
</tr>
<tr>
<td>Marie-Hélène Pinard-Van der Laan</td>
<td>Sophie Dhome-Pollet</td>
<td>Thomas Derrien</td>
</tr>
<tr>
<td></td>
<td>Diane Esquerre</td>
<td>Sarah Djebali-Quelen</td>
</tr>
<tr>
<td></td>
<td>Stéphane Fabre</td>
<td>Sylvain Foissac</td>
</tr>
<tr>
<td></td>
<td>Joël Gautron</td>
<td>Christine Gaspin</td>
</tr>
<tr>
<td></td>
<td>Adeline Goubil</td>
<td>Ignacio Gonzalez</td>
</tr>
<tr>
<td></td>
<td>Sonia Lacroix-Lamandé</td>
<td>Christophe Klopp</td>
</tr>
<tr>
<td></td>
<td>Fabrice Laurent</td>
<td>Sandrine Lagarrigue</td>
</tr>
<tr>
<td></td>
<td>Florence Mompart</td>
<td>Sylvain Marthey</td>
</tr>
<tr>
<td></td>
<td>Pascale Quéré</td>
<td>Maria Marti-Marimon</td>
</tr>
<tr>
<td></td>
<td>Michèle Tixier-Boichard</td>
<td>Raphaëlle Momal-Leisenring</td>
</tr>
<tr>
<td></td>
<td>Gwenola Tosser-Klopp</td>
<td>Kylie Munyard</td>
</tr>
<tr>
<td></td>
<td>Silvia Vincent-Naulleau</td>
<td>Kévin Muret</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Andrea Rau</td>
</tr>
<tr>
<td></td>
<td></td>
<td>David Robelin</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Magali San Cristobal</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Nathalie Vialaneix</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Matthias Zytnicki</td>
</tr>
</tbody>
</table>

+ **Hi-C team @ Toulouse**

<table>
<thead>
<tr>
<th>Hervé Acloque</th>
<th>Martine Bouissou-Matet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sylvain Foissac</td>
<td>Sarah Djebali</td>
</tr>
<tr>
<td>Yvette Lahbib</td>
<td>Laurence Liaubet</td>
</tr>
<tr>
<td>Laurence Liaubet</td>
<td>Maria Marti</td>
</tr>
<tr>
<td>Florence Mompart</td>
<td>Pierre Neuvial</td>
</tr>
<tr>
<td>Pascale Quéré</td>
<td>Maria Marti-Marimon</td>
</tr>
<tr>
<td>Michèle Tixier-Boichard</td>
<td>Raphaëlle Momal-Leisenring</td>
</tr>
<tr>
<td>Gwenola Tosser-Klopp</td>
<td>Kylie Munyard</td>
</tr>
<tr>
<td>Silvia Vincent-Naulleau</td>
<td>Andrea Rau</td>
</tr>
<tr>
<td></td>
<td>David Robelin</td>
</tr>
<tr>
<td></td>
<td>Magali San Cristobal</td>
</tr>
<tr>
<td></td>
<td>Nathalie Vialaneix</td>
</tr>
<tr>
<td></td>
<td>Matthias Zytnicki</td>
</tr>
</tbody>
</table>

+ **INRA Platforms & Facilities**

@BRIDGe biorepository

GeT-PlaGe sequencing

GenoToul bioinformatics

GenoToul biostatistics

Experimental & animal facilities UE Le Pin, Bourges, GenESI, PEAT, PAO ; UR BOA, PRC CIRE

SelGen INRA metaprogram