The FR-AgENCODE project

A French pilot project for the annotation of livestock genomes

Sylvain Foissac, INRA Toulouse
Elisabetta Giuffra, INRA Paris

IGGC Workshop - PAG XXIII, San Diego, California
Outline

- context
  - the ENCODE project (human)
  - the FAANG project (livestock)
- the FR-AgENCODE project
  - people
  - species
  - tissues
  - assays
- conclusion and perspectives
Outline

- context
  - the ENCODE project (human)
  - the FAANG project (livestock)
- the FR-AgENCODE project
  - people
  - species
  - tissues
  - assays
- conclusion and perspectives
Junk DNA' Debunked
Studies Find Human Genomic Makeup Is Vastly Messier; New Disease Links Seen

By GAUTAM NAIK and ROBERT LEE HOTZ
Updated Sept. 5, 2012 2:01 p.m. ET

Researchers describe DNA studies Wednesday that could point the way to new methods to detect and treat disease. From left, Tim Hubbard of Wellcome Trust Sanger Institute, Roderic Guigo of the Centre for Genomic Regulation and Ewan Birney of the European Bioinformatics Institute. Press Association/Associated Press

The deepest look into the human genome so far shows it to be a richer, messier and more intriguing place than was believed just a decade ago, scientists said Wednesday.
The ENCODE project

- Encyclopedia of DNA elements
- goal: characterize (annotate) all functional elements of the human genome
The ENCODE project

- Encyclopedia of DNA elements
- goal: characterize (annotate) all functional elements of the human genome
- mainly USA, UK, Spain, Singapore, Japan
The ENCODE project

- Encyclopedia of DNA elements
- goal: characterize (annotate) all functional elements of the human genome
- mainly USA, UK, Spain, Singapore, Japan
- 32 institutes
- 440 scientists
The ENCODE project
Encyclopedia of DNA elements
goal: characterize (annotate) all functional
elements of the human genome
mainly USA, UK, Spain, Singapore, Japan
32 institutes
440 scientists
The ENCODE project

- Encyclopedia of DNA elements
- goal: characterize (annotate) all functional elements of the human genome
- mainly USA, UK, Spain, Singapore, Japan
- 32 institutes
- 440 scientists
- 1600 experiments, 12000 files analyzed, 15TB of disk
- over $300M in total (NHGRI)
ENCODE experiments

Image credits: Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI)
ENCODE organization

The ENCODE project is structured into various groups and centers, each responsible for producing and analyzing different types of data. The main components include:

1. **Data Production Groups**
   - RNA
   - Histone Mods
   - DNase
   - DNAme
   - TF Binding
   - RBP Binding

2. **Data Coordination Center**
   - Data Analysis Center
   - Analysis Working Group

3. **Technology Development Groups**

4. **Computational Analysis Groups**
   - Gene Models
   - Chromatin States
   - ID Elements

The ENCODE project aims to produce a comprehensive encyclopedia of DNA accessibility in the human genome.
ENCODE
EN<e>说不出的 DNA Elements

GENCODE Known Genes
GENCODE Putative Genes
Pseudogenes
Consensus Pseudogenes Known+Pred RNA
RNA Secondary Structure
CpG Islands
cDNA Paired-End Tags
RNA Transcription

POLR3K
C16orf33
C16orf36
RH5DF1

GM06990
HeLa
HL-60
Neutrophil

Sylvain Foissac  The FR-AgENCODE project
IGGC Workshop - PAGXXIII, 2015 - San Diego, California
ENCODE data
## Showing 15 of 15

<table>
<thead>
<tr>
<th>Assay</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>ChIP-seq</td>
<td></td>
</tr>
<tr>
<td>RNA-seq</td>
<td>15</td>
</tr>
<tr>
<td>DNase-seq</td>
<td>10</td>
</tr>
<tr>
<td>CAGE</td>
<td>8</td>
</tr>
<tr>
<td>RNA profiling by array assay</td>
<td>8</td>
</tr>
</tbody>
</table>

### Experiment status
- released: 15

<table>
<thead>
<tr>
<th>Genome assembly</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>hg19</td>
<td>15</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Organ</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>skin of body</td>
<td>19</td>
</tr>
<tr>
<td>blood vessel</td>
<td>15</td>
</tr>
<tr>
<td>lung</td>
<td>6</td>
</tr>
<tr>
<td>extraembryonic structure</td>
<td>4</td>
</tr>
<tr>
<td>mammary gland</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Life stage</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>newborn</td>
<td>9</td>
</tr>
<tr>
<td>adult</td>
<td>6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Available data</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>bam</td>
<td>15</td>
</tr>
<tr>
<td>bigWig</td>
<td>15</td>
</tr>
<tr>
<td>fasta</td>
<td>15</td>
</tr>
</tbody>
</table>

### RNA-seq of vein endothelial cell (*Homo sapiens*, adult)
- **Lab**: Thomas Gingeras, CSHL
- **Project**: ENCODE
- **Reference**: ENCSR000CV1
- **Status**: released

### RNA-seq of thoracic aorta endothelial cell (*Homo sapiens*, adult)
- **Lab**: Thomas Gingeras, CSHL
- **Project**: ENCODE
- **Reference**: ENCSR000CVA
- **Status**: released

### RNA-seq of fibroblast of the aortic adventitia (*Homo sapiens*, adult)
- **Lab**: Thomas Gingeras, CSHL
- **Project**: ENCODE
- **Reference**: ENCSR000CLJ
- **Status**: released
ENCODE publications

2007: pilot project

Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project

2012: 30 articles

An integrated encyclopedia of DNA elements in the human genome
ENCODE publications

- using ENCODE data
- from the ENCODE consortium
Outline

- context
  - the ENCODE project (human)
  - the FAANG project (livestock)
- the FR-AgENCODE project
  - people
  - species
  - tissues
  - assays
- conclusion and perspectives
Outline

- context
  - the ENCODE project (human)
  - the FAANG project (livestock)
- the FR-AgENCODE project
  - people
  - species
  - tissues
  - assays
- conclusion and perspectives
The FAANG project

Functional Analysis of ANimal Genomes: coordinated international action to accelerate Genome to Phenome
The FAANG project

- Functional Analysis of ANimal Genomes: coordinated international action to accelerate Genome to Phenome
The FAANG project

- Functional Analysis of ANimal Genomes: coordinated international action to accelerate Genome to Phenome
- initial consortium: 43 scientists from 12 countries
The FAANG project

- Functional Analysis of ANimal Genomes: coordinated international action to accelerate Genome to Phenome
- initial consortium: 43 scientists from 12 countries
- expected white paper:

  *Clarke et al,* Poster 114
FAANG ≠ ENCODE
FAANG ≠ ENCODE

- many “livestock” species
  - quality of reference genome assembly
  - availability of phenotypic datasets
FAANG ≠ ENCODE

- many “livestock” species
  - quality of reference genome assembly
  - availability of phenotypic datasets

=> initial set: chicken, pig, cattle and sheep (+ goat, salmon, etc)

- less and disparate founding
  => coordination challenge
FAANG ≠ ENCODE

- many “livestock” species
  - quality of reference genome assembly
  - availability of phenotypic datasets
=> initial set: chicken, pig, cattle and sheep (+ goat, salmon, etc)
- less and disparate founding
=> coordination challenge
- biological targets: tissue samples (not cell lines)
=> muscle, repr. organs, immune cells, adipose, liver, gut, brain
FAANG ≠ ENCODE

- many “livestock” species
  - quality of reference genome assembly
  - availability of phenotypic datasets
  => initial set: chicken, pig, cattle and sheep (+ goat, salmon, etc)
- less and disparate founding
  => coordination challenge
- biological targets: tissue samples (not cell lines)
  => muscle, repr. organs, immune cells, adipose, liver, gut, brain
- molecular experiments
  => initial assays: RNA-seq, DNase-seq and ChIP-seq
    (4 histone modifications + CTCF)
Outline

- context
  - the ENCODE project (human)
  - the FAANG project (livestock)
- the FR-AgENCODE project
  - people
  - species
  - tissues
  - assays
- conclusion and perspectives
Outline

- context
  - the ENCODE project (human)
  - the FAANG project (livestock)
- the FR-AgENCODE project
  - people
  - species
  - tissues
  - assays
- conclusion and perspectives
FR-AgENCODE: overview

- contribution to FAANG (French pilot project)
- goal: improve functional annotation of livestock genomes
- founding: INRA (300KE)
- 4 INRA sites, 9 labs, 58 people
- animals: pig, chicken, cattle, goat
- primary targets: liver, sperm, CD4+ & CD8+ cells
- molecular assays: RNA-seq & Hi-C

project coordination

data analysis

tissue sampling

data production (+others)
FR-AgENCODE: organization

- Coordinators: Elisabetta Giuffra, Sylvain Foissac
- WP0: coordination (E. Giuffra, MH. Pinard-van der Land, S. Foissac)
- WP1: biorepository platform (M. Tixier-Boichard, S. Fabre)
- WP2: molecular assays (D. Esquerré, H. Acloque)
- WP3: data sharing and analysis (S. Foissac, C. Klopp)
- INRA labs:
  - GABI, GenPhySE, PEGASE, MIAT, URA, ISP, GeT-PlaGe
- non-INRA partners:
  - IGDR/CNRS/Rennes Univ., EMBL-EBI, Wageningen Univ., Roslin Inst.
FR-AgENCODE: sampling

- 34 somatic tissues + 13 reproductive tissues (8 female + 5 male):
  - liver, sperm, CD4+, CD8+, plasma, heart, lung, skin, fat, duodenum, ileum, jejunum, cerebellum, frontal lobe, olfactory bulb, trigeminal ganglia, hypothalamus, pancreas, adrenals, kidney, muscle, bone, joints, spleen, lymphatic nodes, peyer's patches, ovary, oocytes, oviduct, uterus, mammary gland, acini, testis, seminal vesicle, etc...
- total: 2,000 to 6,000 samples

INRA CRB-Anim biorepository
FR-AgENCODE: molecular assays

Image credits: Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI)
FR-AgENCODE: molecular assays

Image credits: Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI)
FR-AgENCODE: molecular assays

- RNA-seq
  directional protocol, Illumina Hi-Seq2000
  - polyA+ RNAs (mRNAs + IncRNAs)
    100M+ pairs (2x100bp) per sample (3/2 lanes)
  - sRNAs (miRNAs and others)
    35M reads (1x50bp) per sample (4/lane)
FR-AgENCODE: molecular assays

- RNA-seq
- Hi-C: chromosome conformation capture

J.-M. Belton et al. / Methods 58 (2012) 268–276
FR-AgENCODE: molecular assays

- RNA-seq
- Hi-C: chromosome conformation capture

Rao et al, Cell, 2014
FR-AgENCODE: molecular assays

- RNA-seq
- Hi-C: chromosome conformation capture

Sexton et al 2012
**FR-AgENCODE: molecular assays**

- **RNA-seq**
- **Hi-C: chromosome conformation capture**

**Kalhor et al.** (GM12878)

**Jin et al.** (IMR90)

**In situ Hi-C** (GM12878)

Rao et al, Cell, 2014
FR-AgENCODE: molecular assays

- RNA-seq
- Hi-C: chromosome conformation capture

Dixon et al 2012
**FR-AgENCODE: molecular assays**

- RNA-seq
- Hi-C: chromosome conformation capture

**Dixon et al. 2012**

**Rao et al., 2014**
FR-AgENCODE: molecular assays

- RNA-seq
- Hi-C: chromosome conformation capture

Rao et al, 2014
FR-AgENCODE: molecular assays

- RNA-seq
- Hi-C: chromosome conformation capture

**High-throughput genome scaffolding from in-vivo DNA interaction frequency**

Noam Kaplan¹ and Job Dekker¹
Noam Kaplan: noam.kaplan2@gmail.com; Job Dekker: job.dekker@umassmed.edu
¹Program in Systems Biology, Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, 368 Plantation Street, Worcester, MA, 01605-0103, USA

**Abstract**

Despite advances in DNA-sequencing technology, assembly of complex genomes remains a major challenge, particularly for genomes sequenced using short reads, which yield highly fragmented assemblies. Here we show that genome-wide in vivo chromatin interaction frequency data, which are measurable with chromosome conformation capture–based experiments, can be used as genomic distance proxies to accurately position individual contigs without requiring any sequence overlap. We also use these data to construct approximate genome scaffolds de novo. Applying our approach to incomplete variants of the human genome, we predict the positions of 65 previously...
FR-AgENCODE: molecular assays

- RNA-seq
  directional protocol, Illumina Hi-Seq2000
  - polyA+ RNAs (mRNAs + IncRNAs)
    100M+ pairs (2x100bp) per sample (3/2 lanes)
  - sRNAs (miRNAs and others)
    35M reads (1x50bp) per sample (4/lane)

- Hi-C: high-throughput chromosome conformation capture
  - 64 samples (4 replicates, 4 tissues, 4 species)
  - 70M+ read pairs per sample (2 lib/lane)

Comparative analysis of genome topology and expression
Outline

- context
  - the ENCODE project (human)
  - the FAANG project (livestock)
- the FR-AgENCODE project
  - people
  - species
  - tissues
  - assays
- conclusion and perspectives
Outline

- context
  - the ENCODE project (human)
  - the FAANG project (livestock)
- the FR-AgENCODE project
  - people
  - species
  - tissues
  - assays
- conclusion and perspectives
Summary / Conclusion

- livestock need better reference genomes and annotations
- FAANG: a worldwide coordinated effort
- FR-AgENCODE: a french pilot project
  - pig, cow, chicken and goat
  - biorepository of tissue sample
  - genome expression (RNA-seq) and topology (Hi-C)
  - generate and share genomic data and results
- expected contributions to the Genome to Phenome challenge
Acknowledgments

- FR-AgENCODE members
  (E. Giuffra, S. Lagarrigue, M. Tixier-Boichard, M.H. Pinard, etc)
- FAANG consortium
  (C. Tuggle, M. Groenen, A. Archibald, J. Silverstein, H. Zhou, etc)
- INRA GenPhySE “bisounours” lab
  (H. Acloque, D. Esquerré, F. Pitel, S. Fabre, T. Faraut, etc)
- ENCODE former colleagues
  (R. Guigo, T. Gingeras, P. Kapranov, J. Lagarde, etc)
- IGGC (G. Tosser-Klopp, etc)
JE SUIS CHARLIE