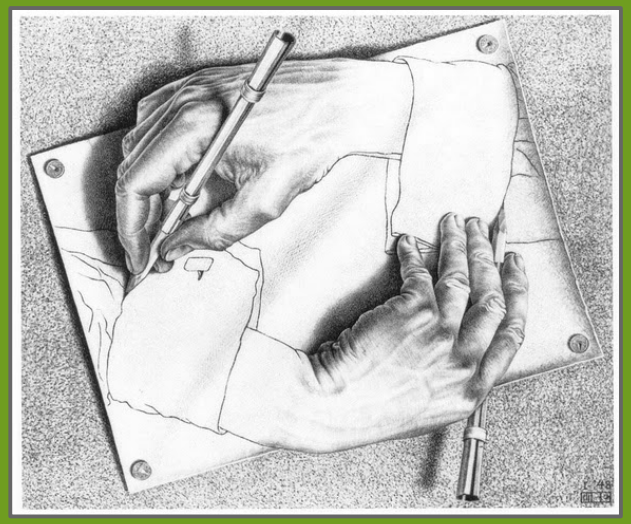
# Functional annotation of livestock genomes
# Chromatin structure and gene expression

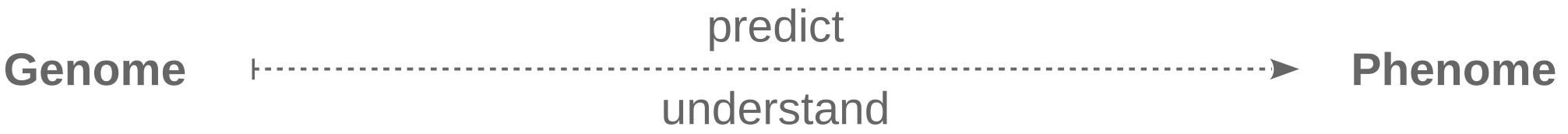


**Sylvain Foissac, INRA Toulouse, France**

**ASAS-ADSA Midwest Meeting, March 2019, Omaha, USA**

INRA
SCIENCE & IMPACT

# Functional Annotation of ANimal Genomes

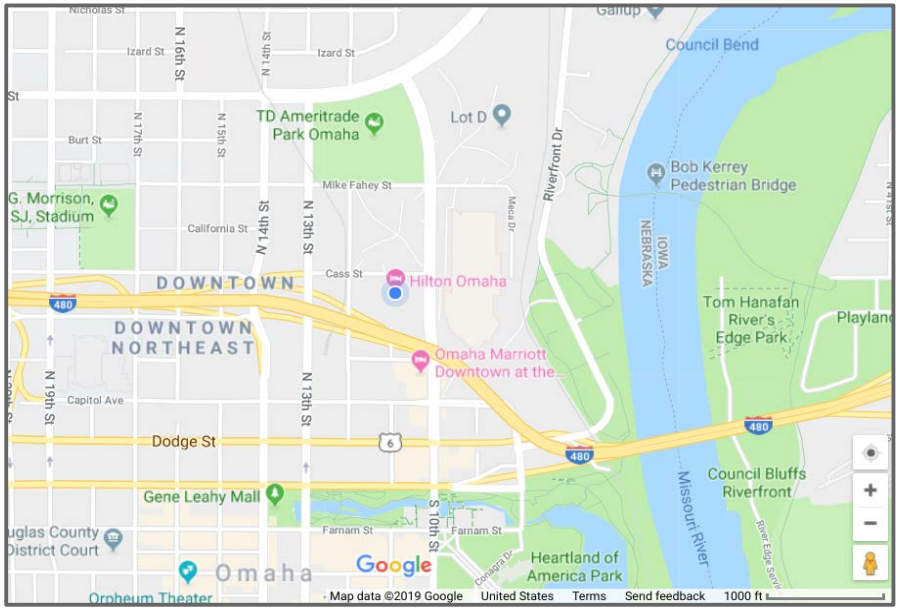


# Functional Annotation of ANimal Genomes

**Genome** — predict / understand → **Phenome**

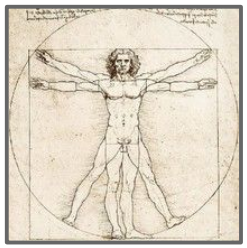
raw DNA sequence

genomic annotation



# Functional Annotation of ANimal Genomes

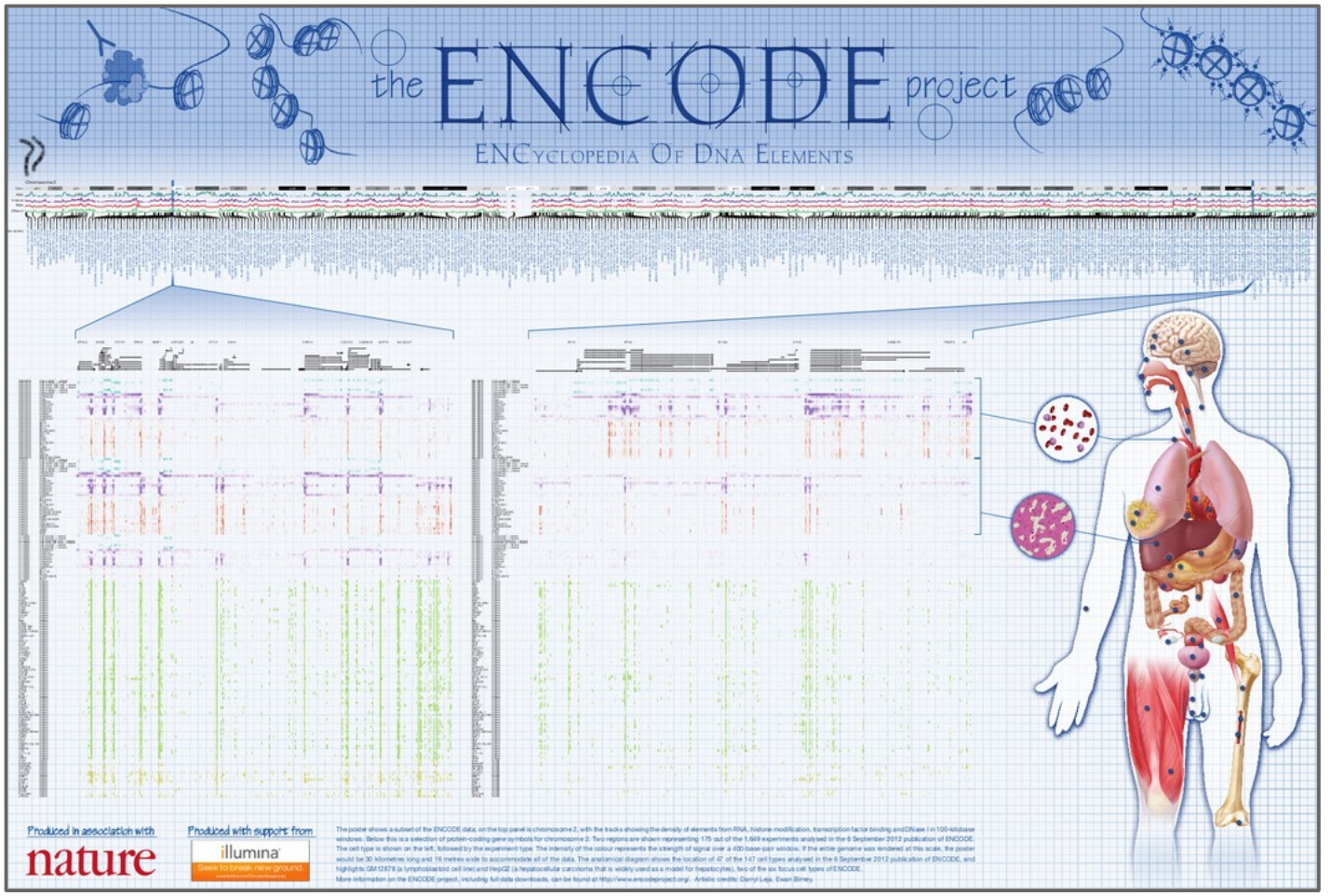
**Genome** — predict / understand → **Phenome**



&

ENCODE

=>



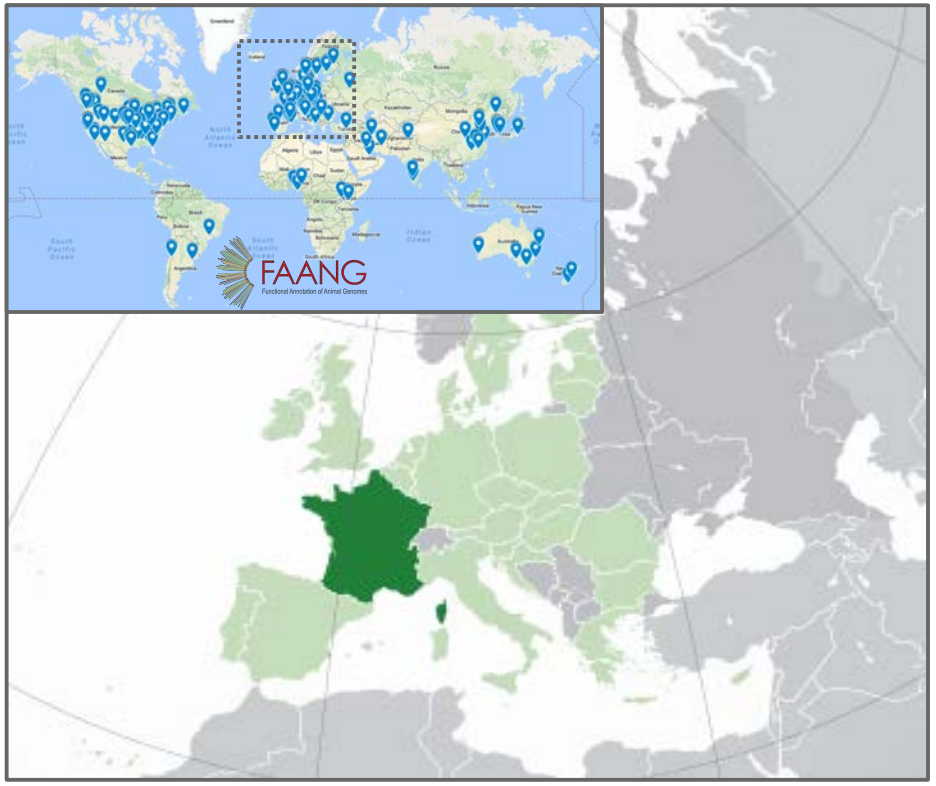
# Functional Annotation of ANimal Genomes

**Genome** — predict / understand → **Phenome**

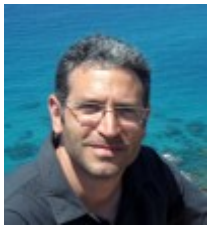
**&** **=> ?**

FAANG
Functional Annotation of Animal Genomes

# The FAANG pilot project FR-AgENCODE



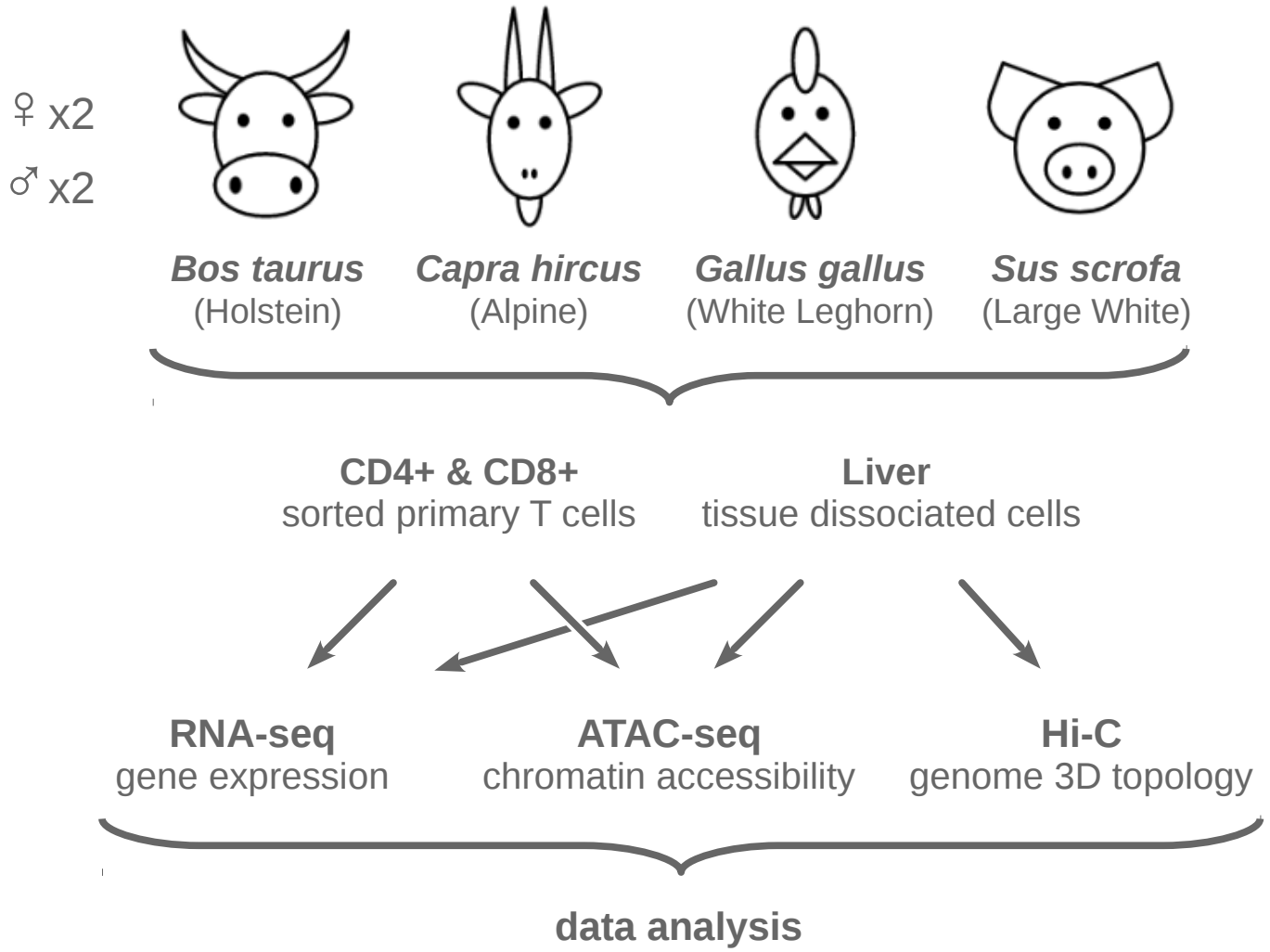

*Elisabetta Giuffra*



*Sylvain Foissac*

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

# The FAANG pilot project FR-AgENCODE

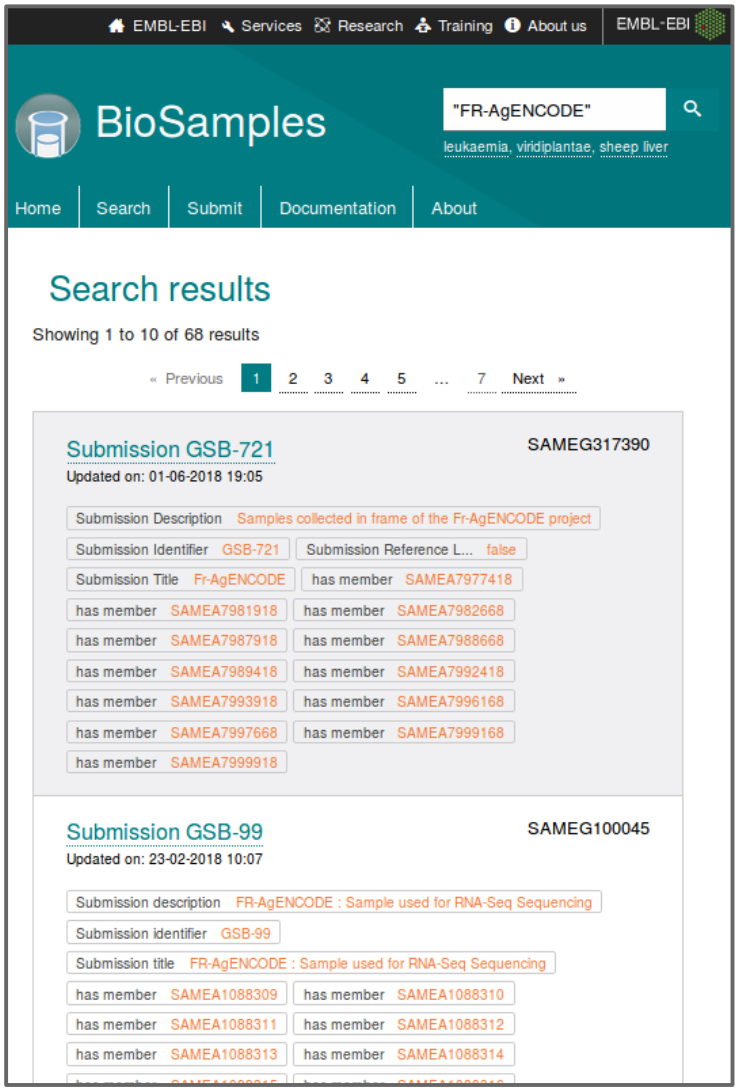


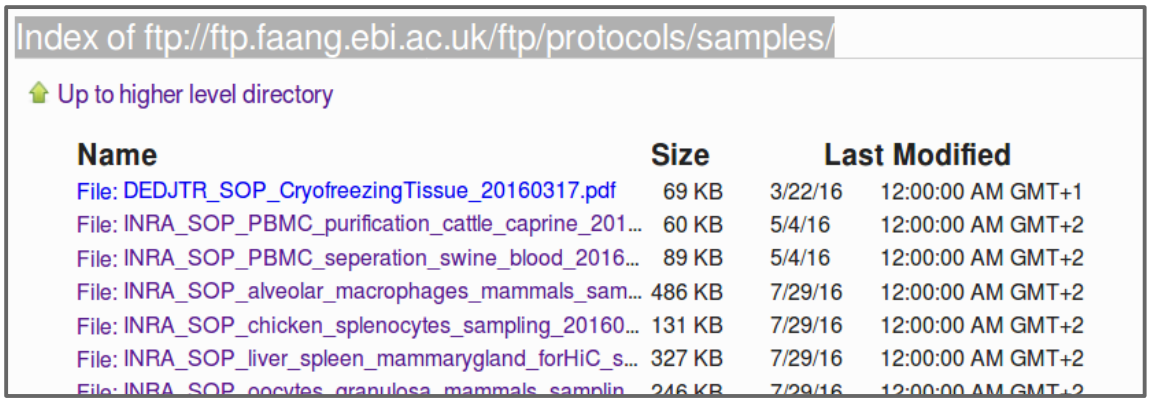
# Step 1: animal and tissue sampling

- 4 animals per species
- 40+ tissues per animal
  Liver, CD4+ T cells, CD8+ T cells, sperm, plasma, heart, lung, skin, fat, duodenum, ileum, jejunum, 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

*Michèle Tixier-Biochard*

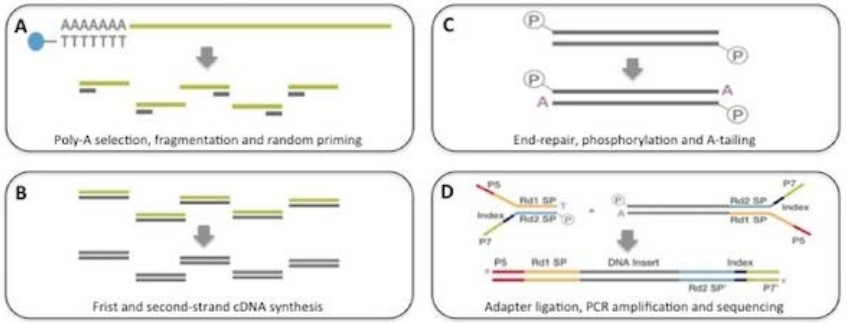
*Stéphane Fabre*



# Step 2: the molecular assays

- RNA-seq: transcriptome profiling

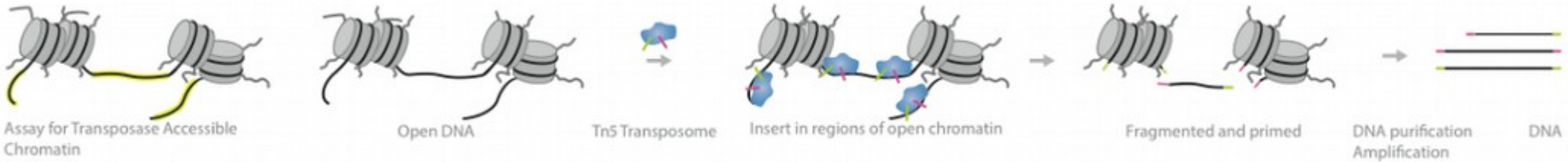



*Diane Esquerré*

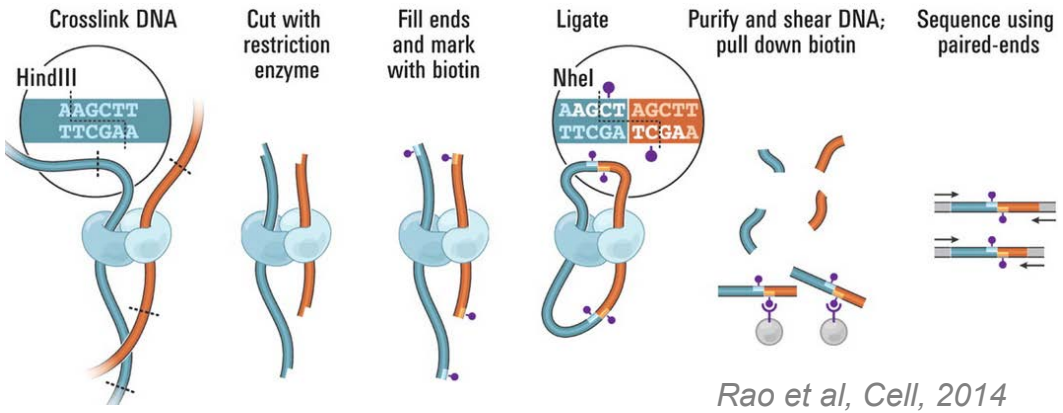
*Hervé Acloque*

- ATAC-seq: chromatin accessibility

*from www.illumina.com*




- Hi-C: chromosome conformation




*Rao et al, Cell, 2014*

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

Up to higher level directory

| Name | Size | Last Modified | |
|---|---|---|---|
| File: DEDJTR_SOP_Sampl... | 134 KB | 3/22/16 | 12:00:00 AM GMT+1 |
| File: INRA_SOP_ATAC-se... | 398 KB | 8/5/16 | 12:00:00 AM GMT+2 |
| File: INRA_SOP_Hi-C_HA... | 900 KB | 12/8/16 | 12:00:00 AM GMT+1 |
| File: INRA_SOP_RNA_extr... | 216 KB | 5/4/16 | 12:00:00 AM GMT+2 |
| File: INRA_SOP_mRNA-se... | 392 KB | 11/24/17 | 12:00:00 AM GMT+1 |
| File: ISU_SOP_ChIP_swin... | 1688 KB | 1/25/19 | 3:34:00 PM GMT+1 |
| File: ISU_SOP_Fetalliver_... | 279 KB | 1/25/19 | 3:34:00 PM GMT+1 |
| File: ISU_SOP_Macropha... | 88 KB | 2/13/19 | 10:14:00 AM GMT+1 |
| File: ROSLIN_SOP_Chrom... | 458 KB | 11/18/16 | 12:00:00 AM GMT+1 |
| File: ROSLIN_SOP_Isolati... | 74 KB | 11/16/16 | 12:00:00 AM GMT+1 |
| File: ROSLIN_SOP_Isolati... | 59 KB | 11/16/16 | 12:00:00 AM GMT+1 |
| File: ROSLIN_SOP_Tissu... | 328 KB | 11/16/16 | 12:00:00 AM GMT+1 |

# Step 2: the molecular assays

**Completed experiments**

**RNA-seq**

| Cattle | cattle1 | cattle2 | cattle3 | cattle4 | Goat | goat1 | goat2 | goat3 | goat4 |
|---|---|---|---|---|---|---|---|---|---|
| cd4 | | | | | cd4 | | | | |
| cd8 | NA | | | | cd8 | | | | |
| liver | | | | | liver | | | | |
| **Chicken** | chicken1 | chicken2 | chicken3 | chicken4 | **Pig** | pig1 | pig2 | pig3 | pig4 |
| cd4 | NA | NA | | | cd4 | NA | | | |
| cd8 | NA | NA | NA | | cd8 | | | | |
| liver | | | | | liver | | | | |

**ATAC-seq**

| Cattle | cattle1 | cattle2 | cattle3 | cattle4 | Goat | goat1 | goat2 | goat3 | goat4 |
|---|---|---|---|---|---|---|---|---|---|
| cd4 | | | | | cd4 | | NA | | |
| cd8 | | | | | cd8 | | | | |
| liver | NA | NA | NA | NA | liver | | | | |
| **Chicken** | chicken1 | chicken2 | chicken3 | chicken4 | **Pig** | pig1 | pig2 | pig3 | pig4 |
| cd4 | | | | NA | cd4 | | | | |
| cd8 | NA | | NA | NA | cd8 | | | | |
| liver | | | | | liver | | | | NA |

**HiC**

| Cattle | cattle1 | cattle2 | cattle3 | cattle4 | Goat | goat1 | goat2 | goat3 | goat4 |
|---|---|---|---|---|---|---|---|---|---|
| liver | NA | NA | NA | NA | liver | | | | |
| **Chicken** | chicken1 | chicken2 | chicken3 | chicken4 | **Pig** | pig1 | pig2 | pig3 | pig4 |
| liver | | | | | liver | | | | |

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

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

- Data integration and comparison
- Statistical analyses

*Sarah Djebali*

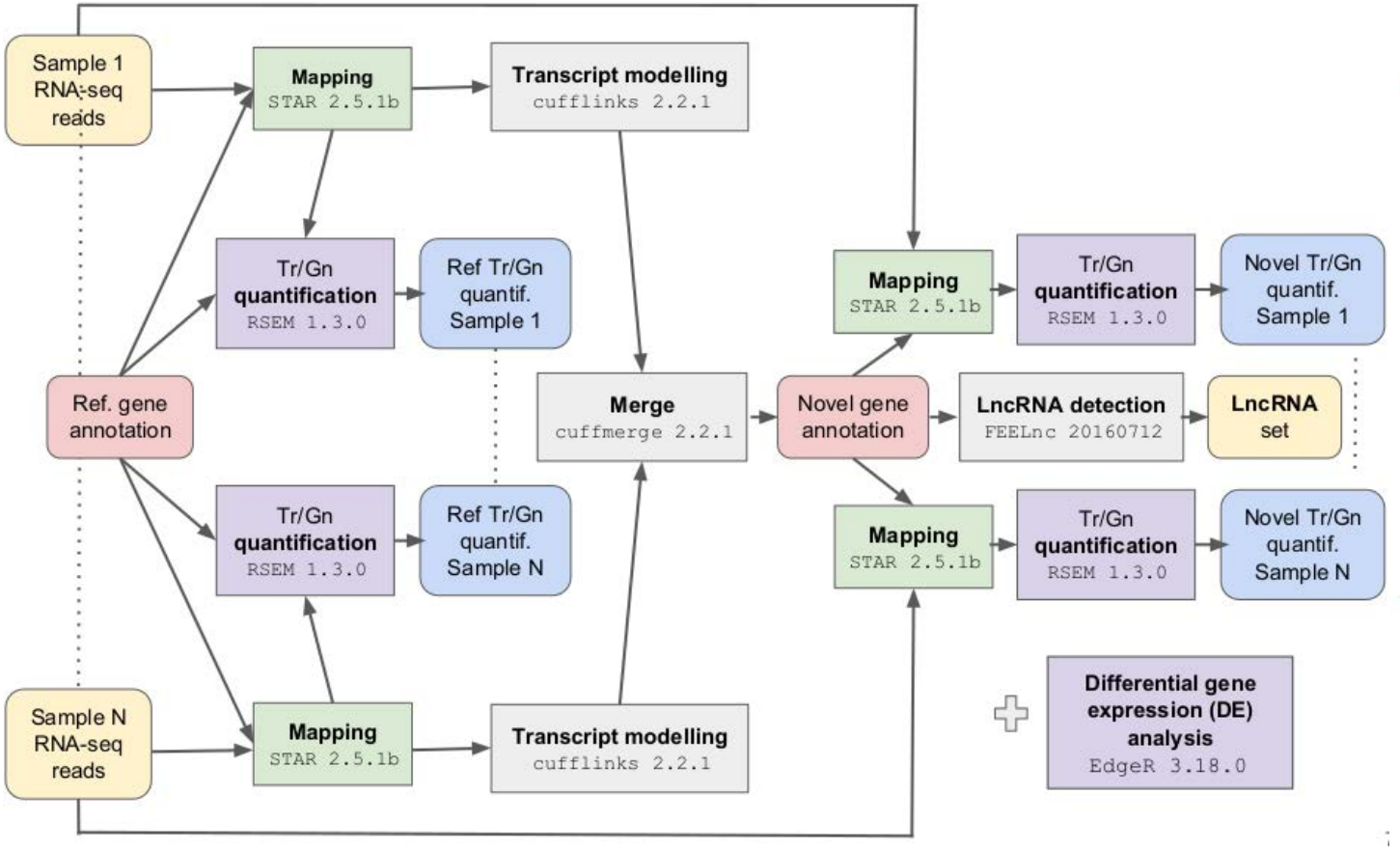
*Nathalie Vialaneix*

# Step 3: data analysis

- Bioinformatics pipelines

RNA-seq

- Read trimming (trimgalore)
- Read mapping (STAR2)
- Transcript modelling (Cufflinks2)
- New gene annotation (Cuffmerge2)
- Transcript/gene expresso quantification (RSEM)
- LncRNA calling (FEELnc)

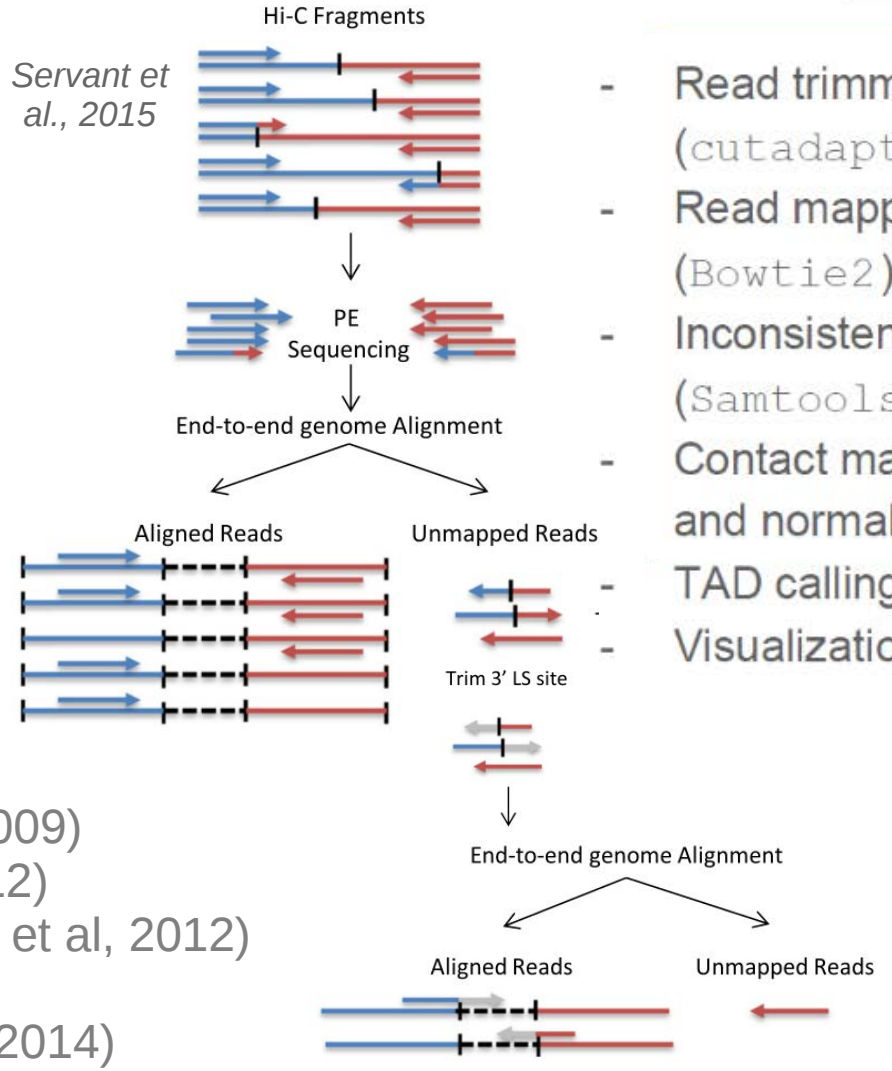


# Step 3: data analysis

- Bioinformatics pipelines

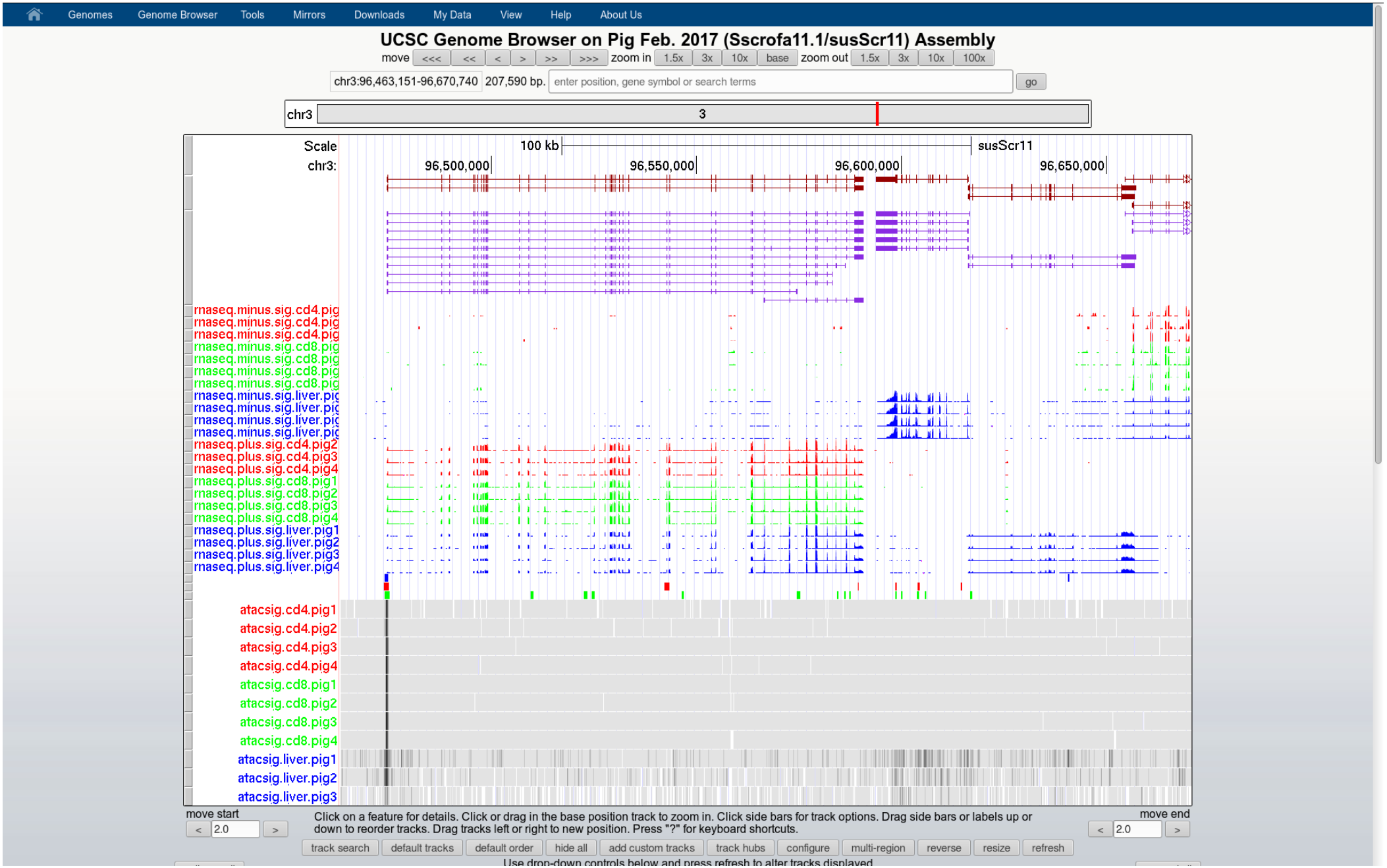
Hi-C detailed workflow
- Trim reads (ligation site)
- **Map on reference genome**
- Discard inconsistent pairs
- **Count reads in pairs of genomic bins & generate contact matrix**
- Normalize contact matrix (non parametric, matrix balancing)
- Generate html report
- **Identify Topological Associated 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)




HiC (HiC-Pro)
- Read trimming (cutadapt)
- Read mapping to genome (Bowtie2)
- Inconsistent pairs filtering (Samtools)
- Contact matrix generation and normalization (ICE)
- TAD calling (Armatus)
- Visualization (Juicebox)

# 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

**Reference and FR-AgENCODE transcripts detected**

| Species | Reference transcripts | | | FR-AgENCODE transcripts | | |
|---|---|---|---|---|---|---|
| | All | Expressed | | | | |
| | | # | % of total | # | mRNAs | lncRNAs |
| Cattle | 26,740 | 16,100 | 60.2 | 84,971 | 59,801 | 22,724 |
| Goat | 53,266 | 34,442 | 64.7 | 78,091 | 64,962 | 13,864 |
| Chicken | 38,118 | 22,180 | 58.2 | 57,817 | 47,567 | 7,502 |
| Pig | 49,448 | 29,786 | 60.2 | 77,540 | 63,721 | 12,587 |

# Results: the transcriptomic landscape

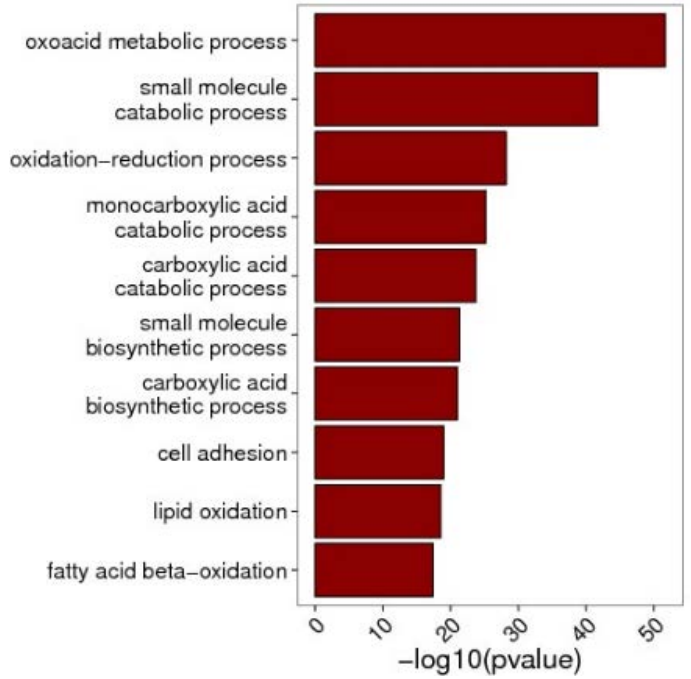
- most of all the known transcripts are detected in liver and T cells
- reference annotation can be extended by a factor 2 to 3
- most of the new transcripts are alternative isoforms of coding genes
- differential expression between liver and T cells:
  - most of the genes are Differentially Expressed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

| | Number of DE reference genes | | | |
|---|---|---|---|---|
| | Cattle | Goat | Chicken | Pig |
| **liver > T cells** | 4,992 | 6,188 | 4,307 | 5,666 |
| **T cells > liver** | 3,943 | 4,384 | 2,640 | 3,772 |

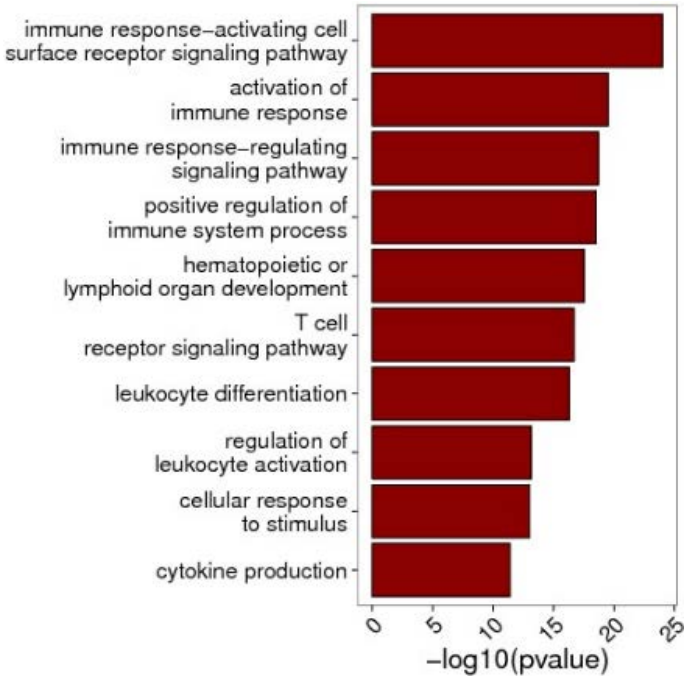
# Results: the transcriptomic landscape

- 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

**liver > T cells**

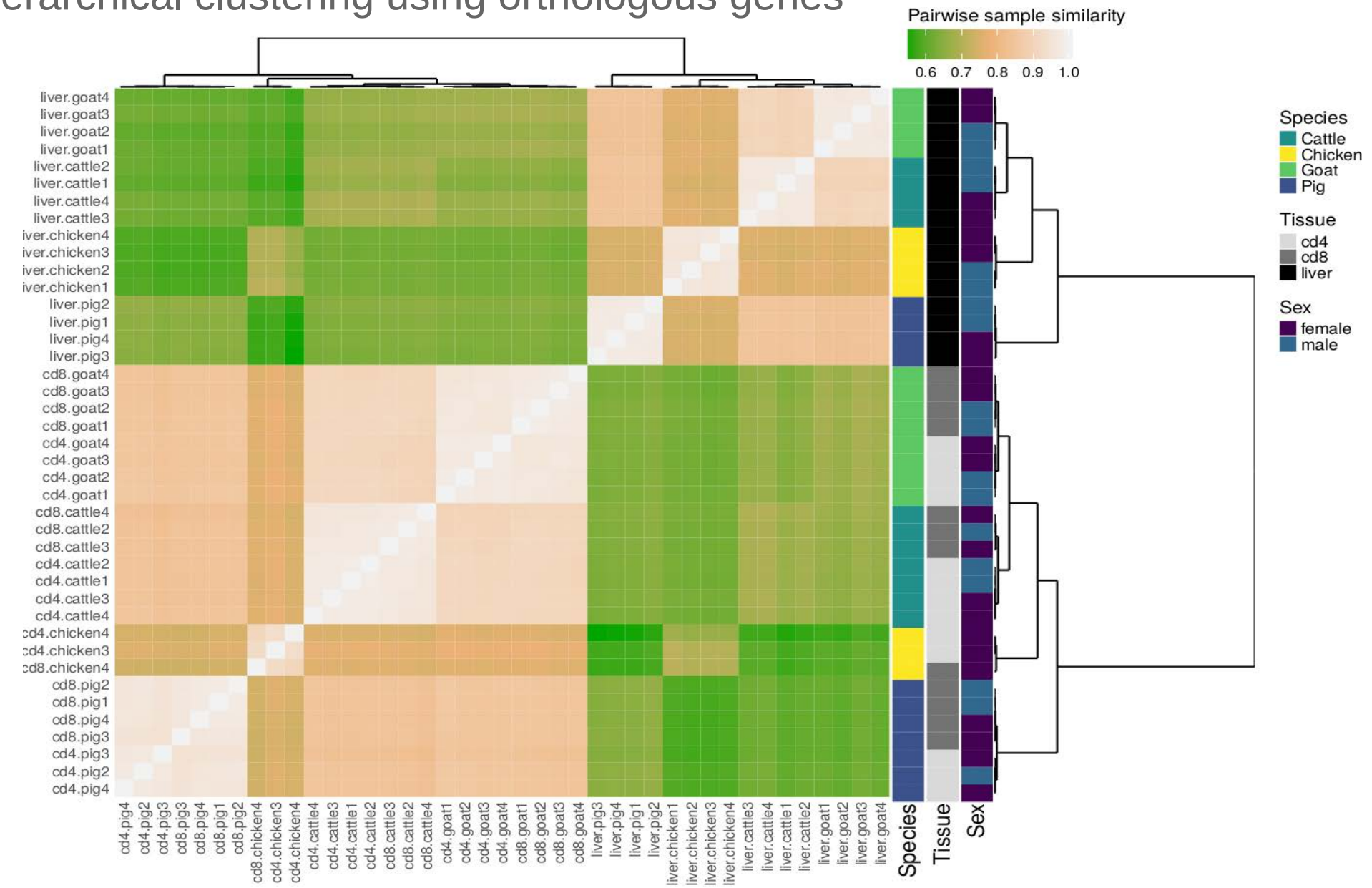



**T cells > liver**



# Results: the transcriptomic landscape

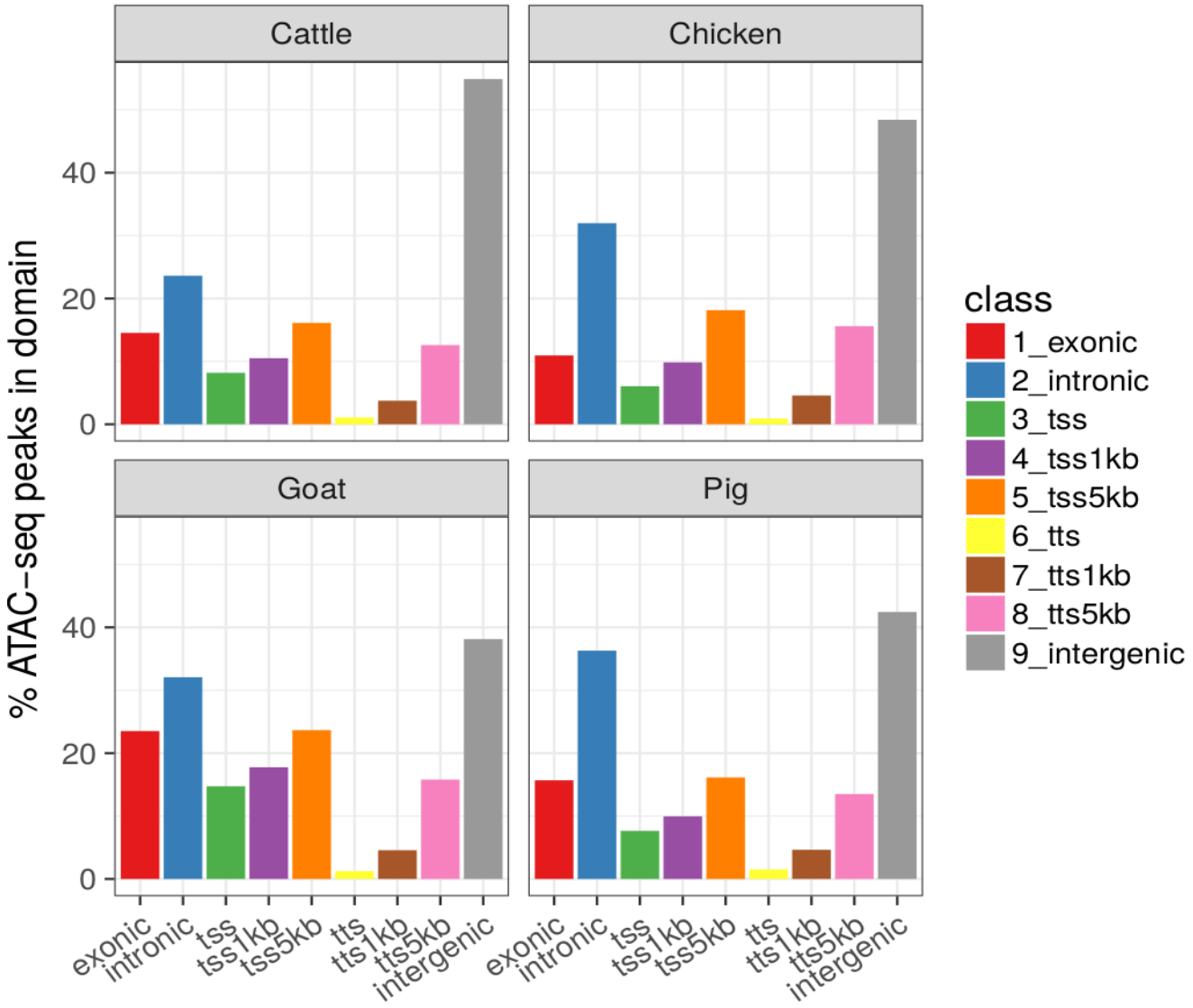
## Hierarchical clustering using orthologous genes




Samples cluster first by tissue, then by species

# Results: the chromatin accessibility landscape

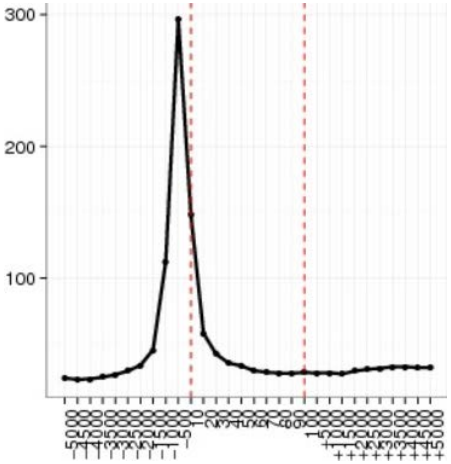
- 75,000-150,000 accessibility sites by species (~2-4% of the genome)
- Most of them are intergenic & intronic

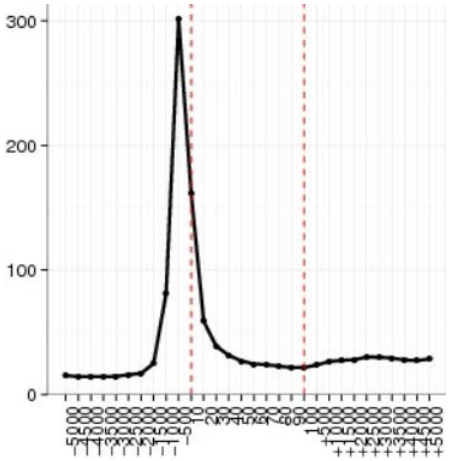


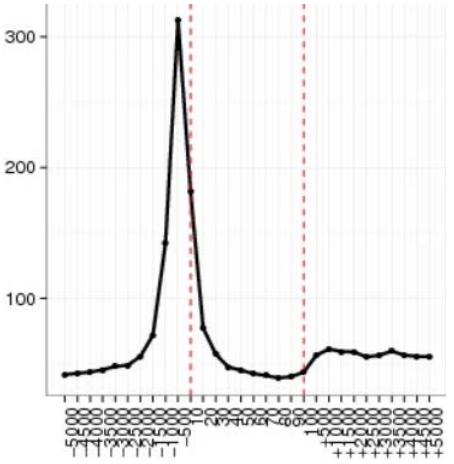
# Results: the chromatin accessibility landscape

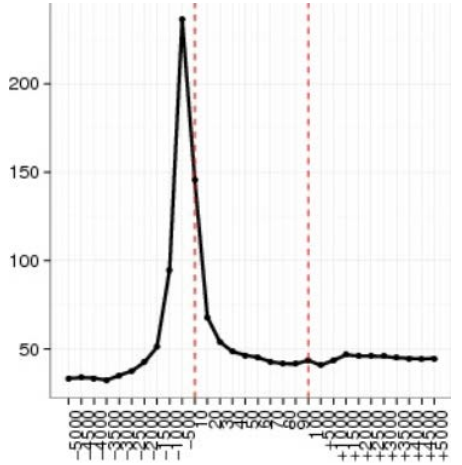
- 75,000-150,000 accessibility sites by species (~2-4% of the genome)
- Most of them are intergenic & intronic
- Promoter accessibility: max within 1Kb upstream of gene starts










Mean ATAC-seq score around and within genes

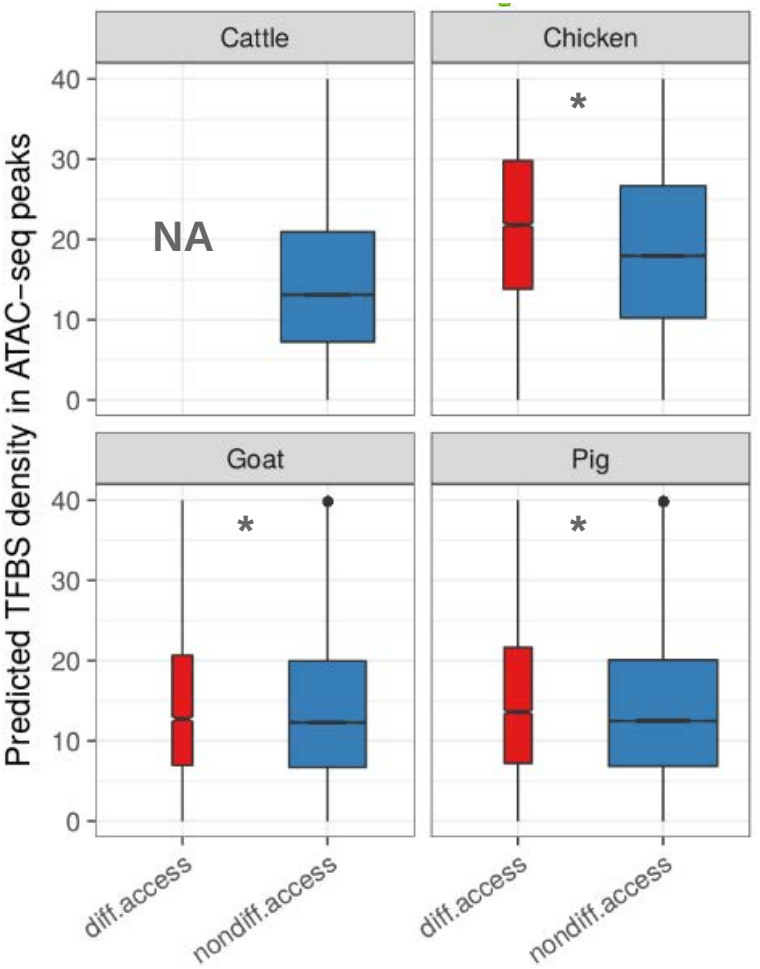
# Results: the chromatin accessibility landscape

- 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

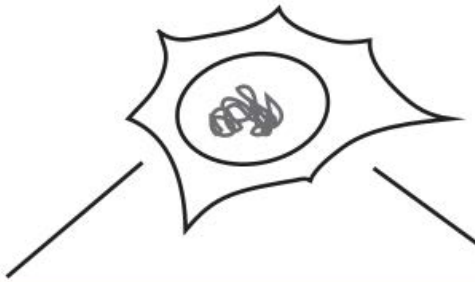
# Results: the chromatin accessibility landscape

- 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



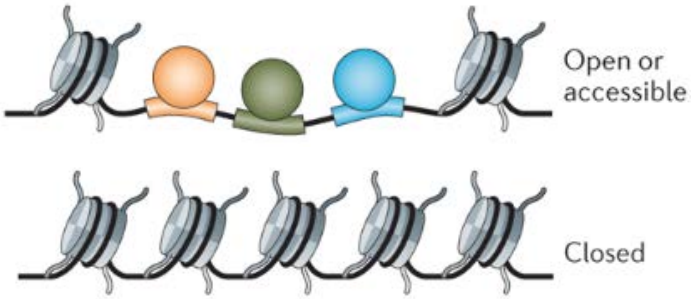
**nuclear genome**
(a few Gb)

*open, active* "A"

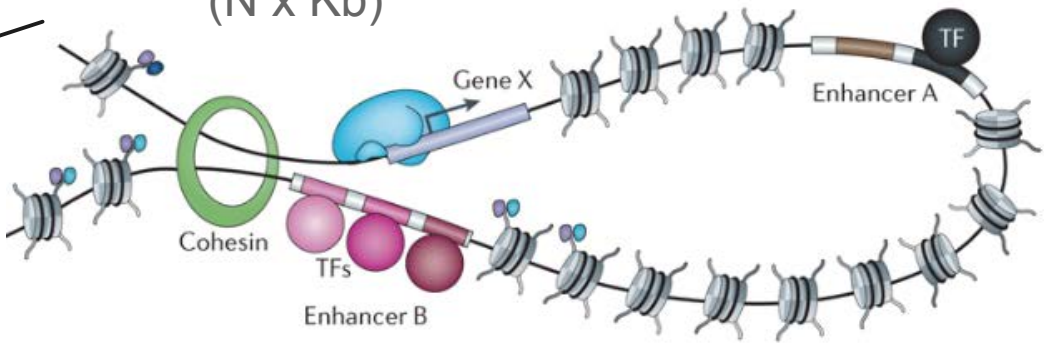
**A/B compartments**
(~n x Mb)

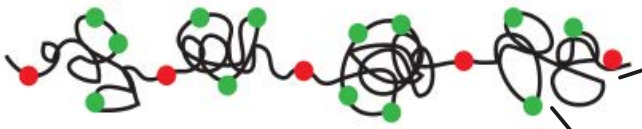
*closed, inactive* "B"

**regulators**
(enhancers, TFBS… n x 100bp)




**regulatory loops**
(N x Kb)



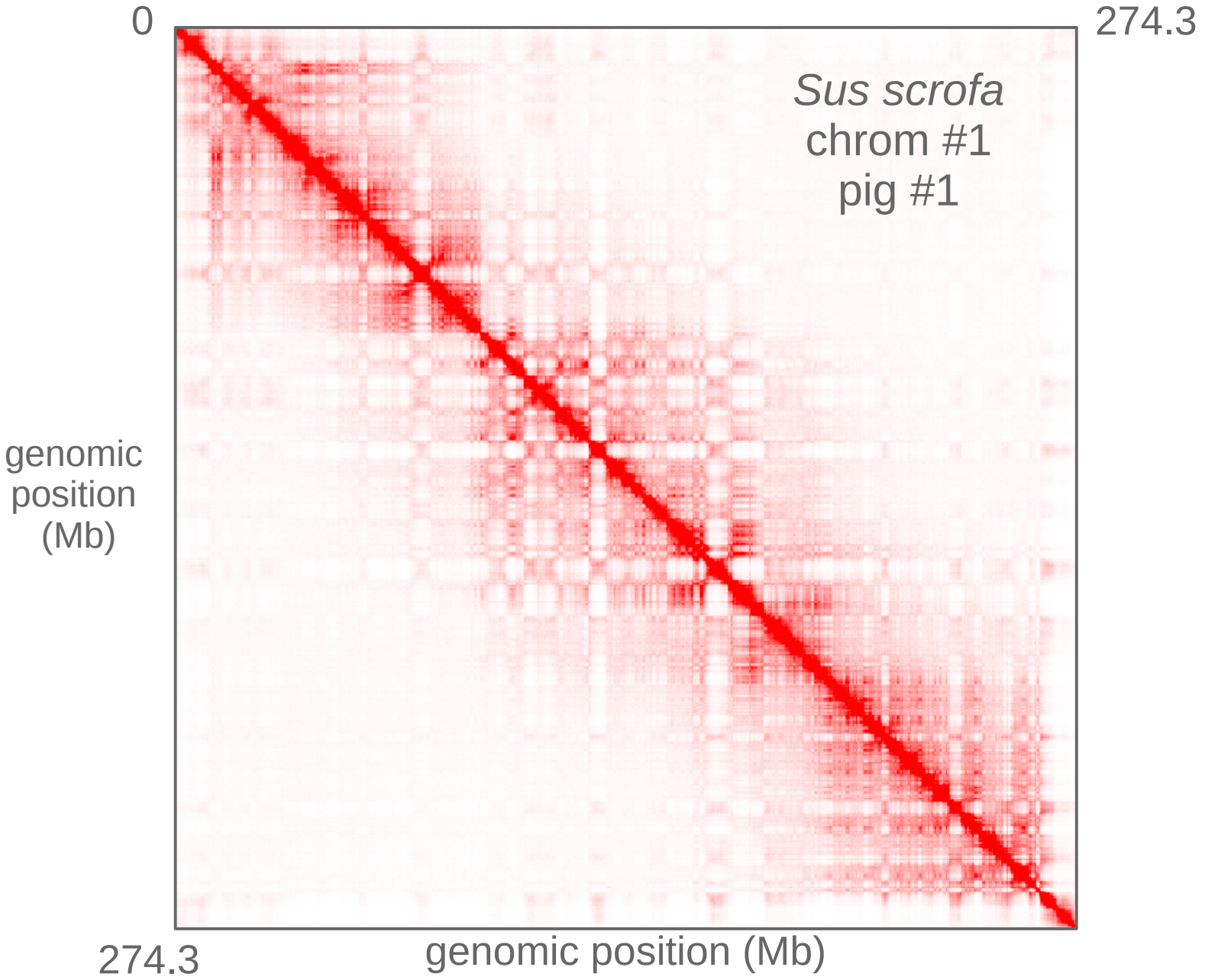



**TADs**
Topologically Associating Domains
(n x 0.1 Mb)

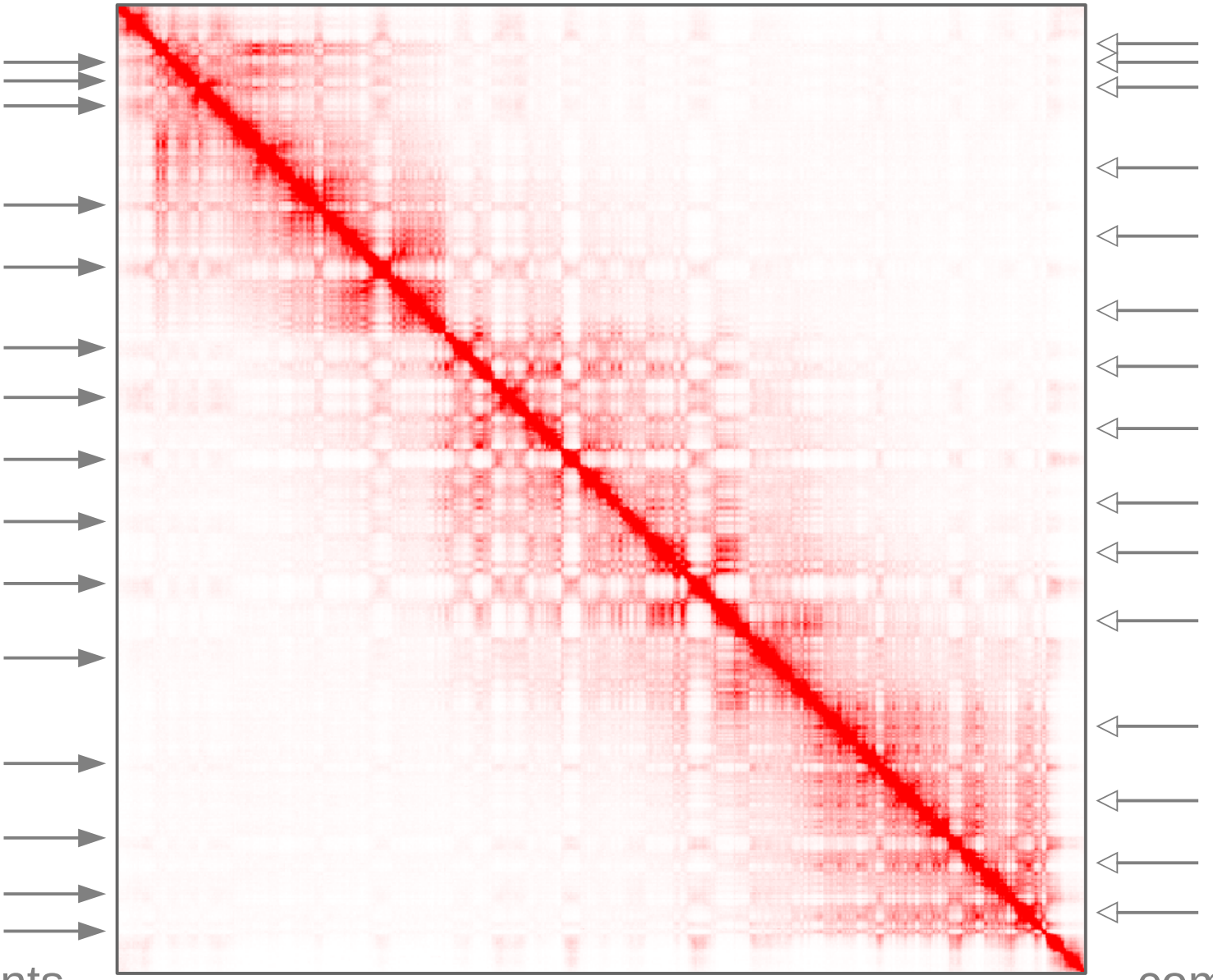
*Giorgetti et al., 2013, Genome Biol.*

*Shlyueva et al, Nat. Rev. Genet., 2014*

# Data overview: the Hi-C interaction matrix



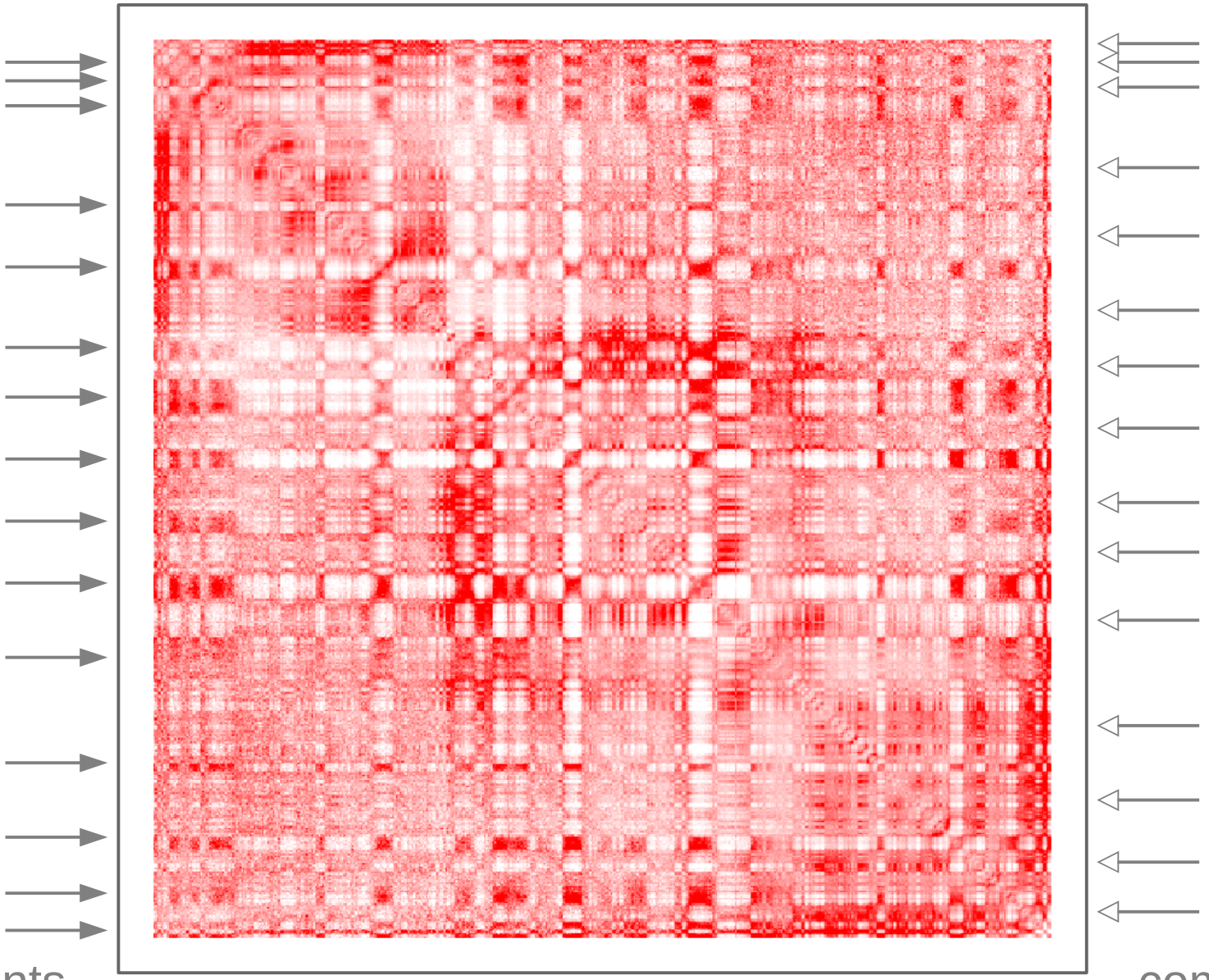
# The A/B (epi)genome compartments



“A” compartments

“B” compartments

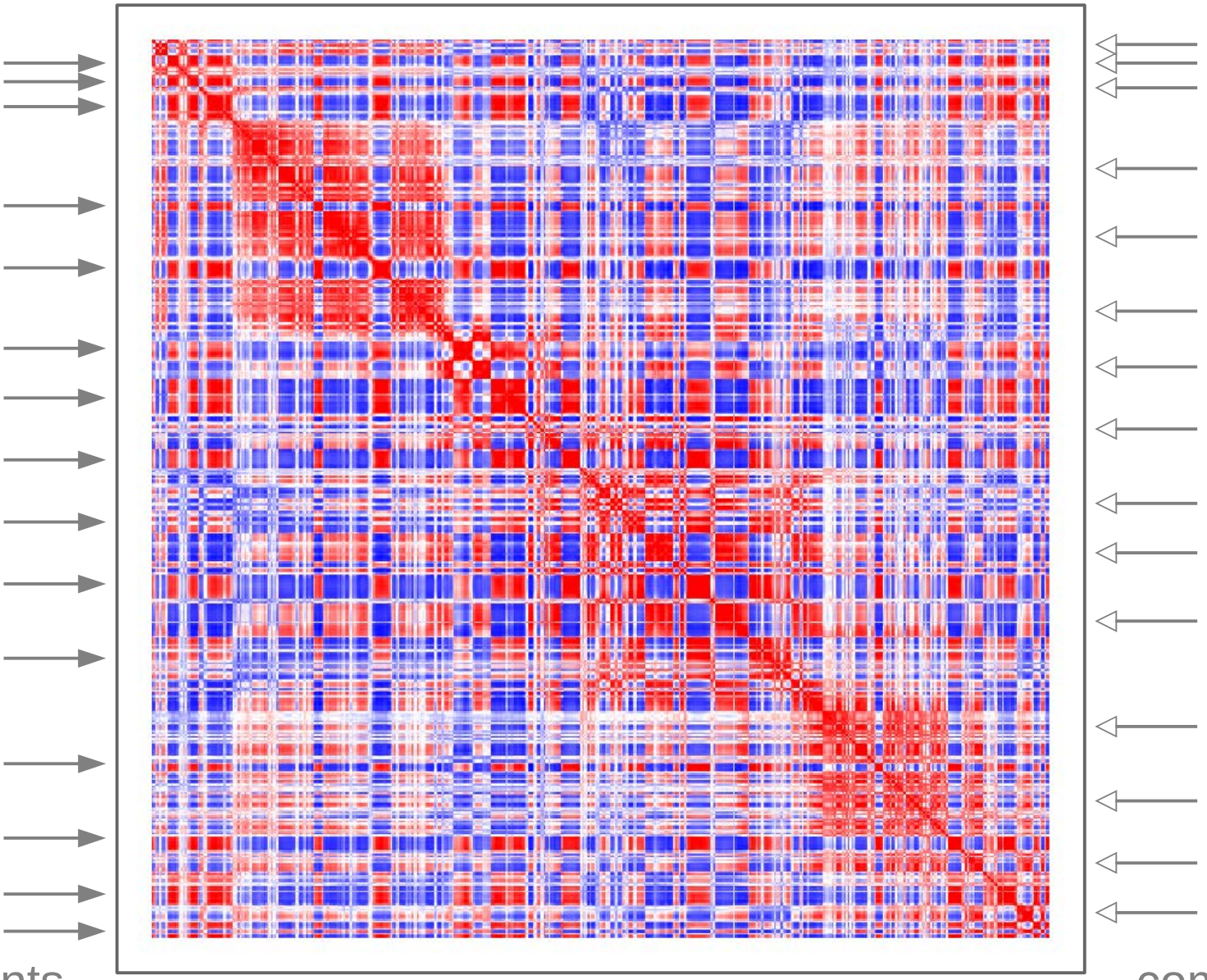
# The A/B (epi)genome compartments



“A” compartments

“B” compartments

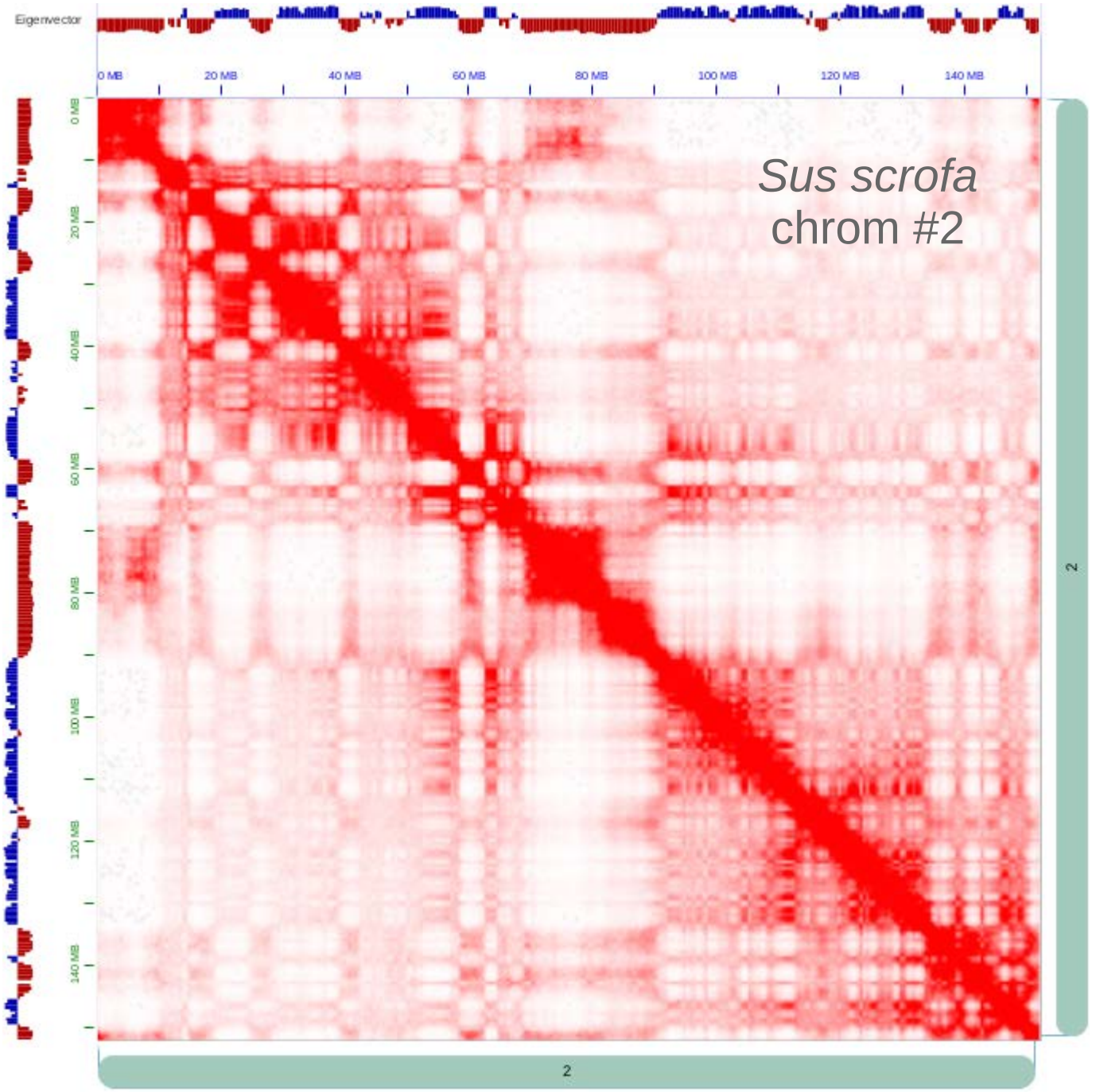
# The A/B (epi)genome compartments



"A" compartments

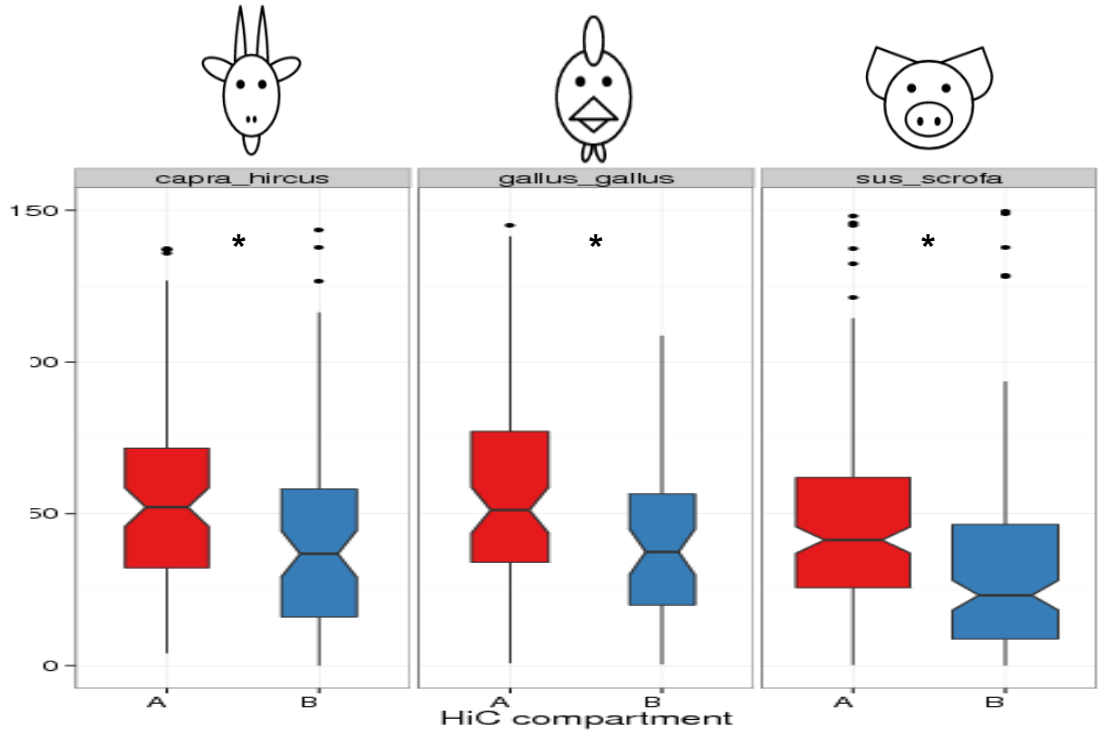
"B" compartments

# The A/B (epi)genome compartments

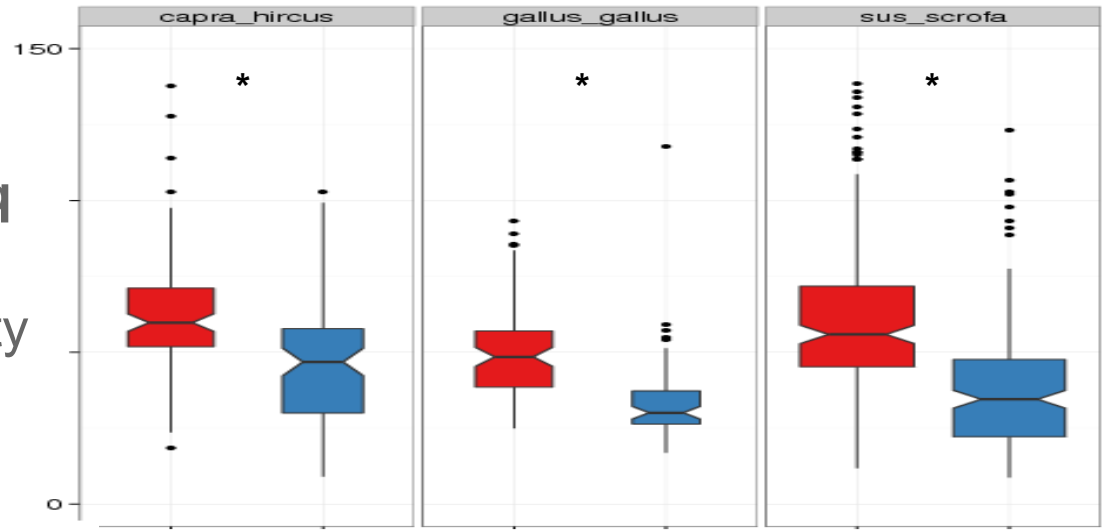


# The A/B (epi)genome compartments

**RNA-seq**
gene
expression




**ATAC-seq**
chromatin
accessibility




**A/B**
compartments

**A** **B** **A** **B** **A** **B**

=>

**Links between genome structure & function**

# 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%**)

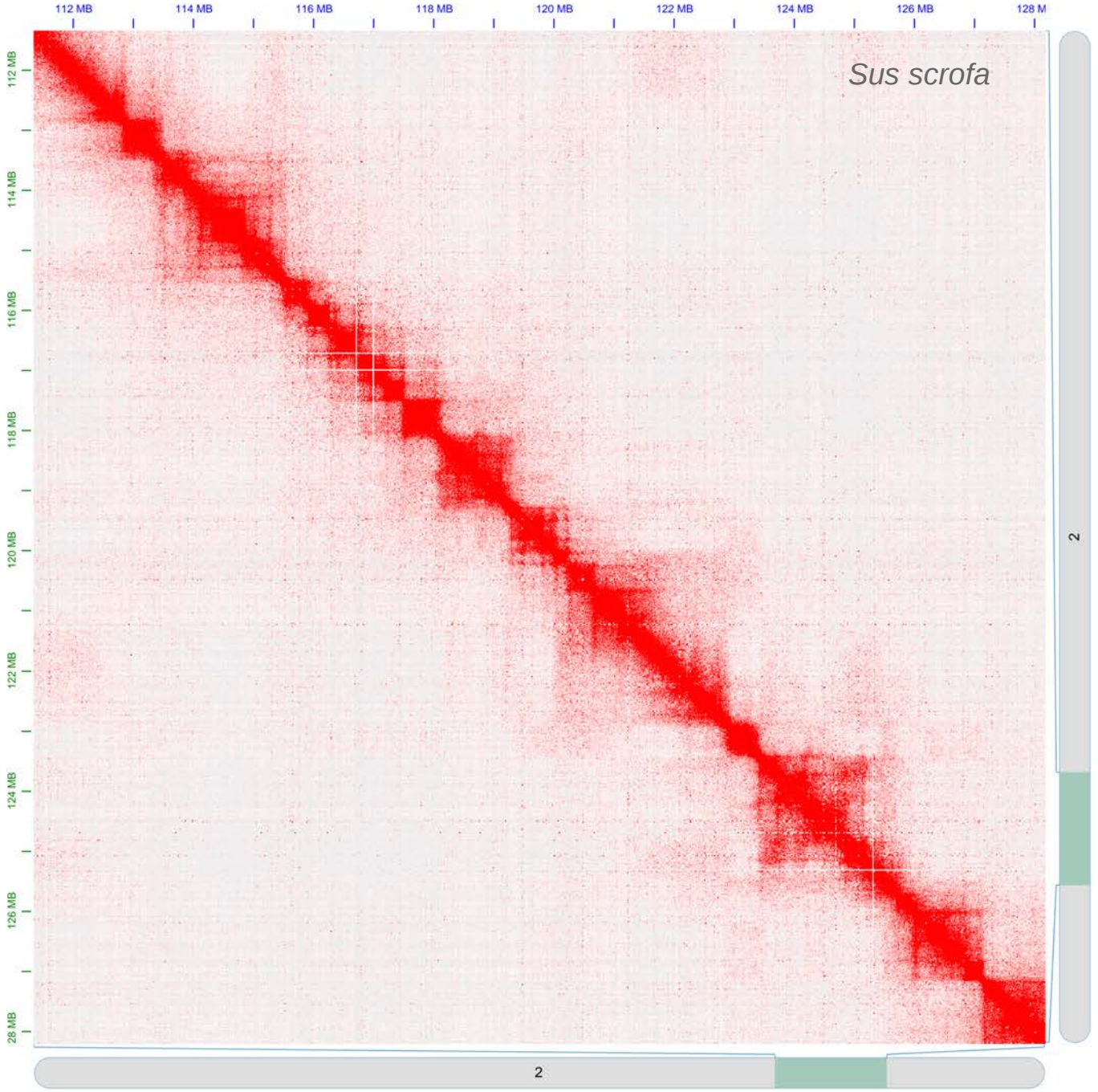
- Focusing on orthologous genes: compartments across species

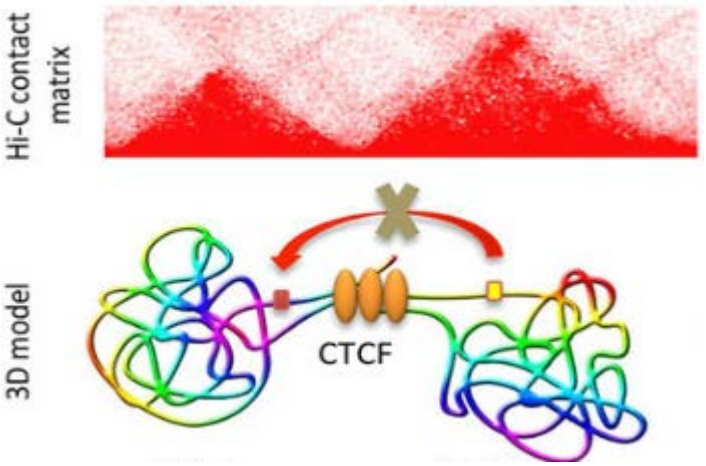
  | "A" in the 3 species | "B" in the 3 species |
  |---|---|
  | expected: 1529 genes (**26.7%**) | expected: 259 genes (**4.5%**) |
  | observed: 2972 genes (**51.9%**) | observed: 611 genes (**10.7%**) |

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

**conservation of genome compartments** => **evidence of functional role**

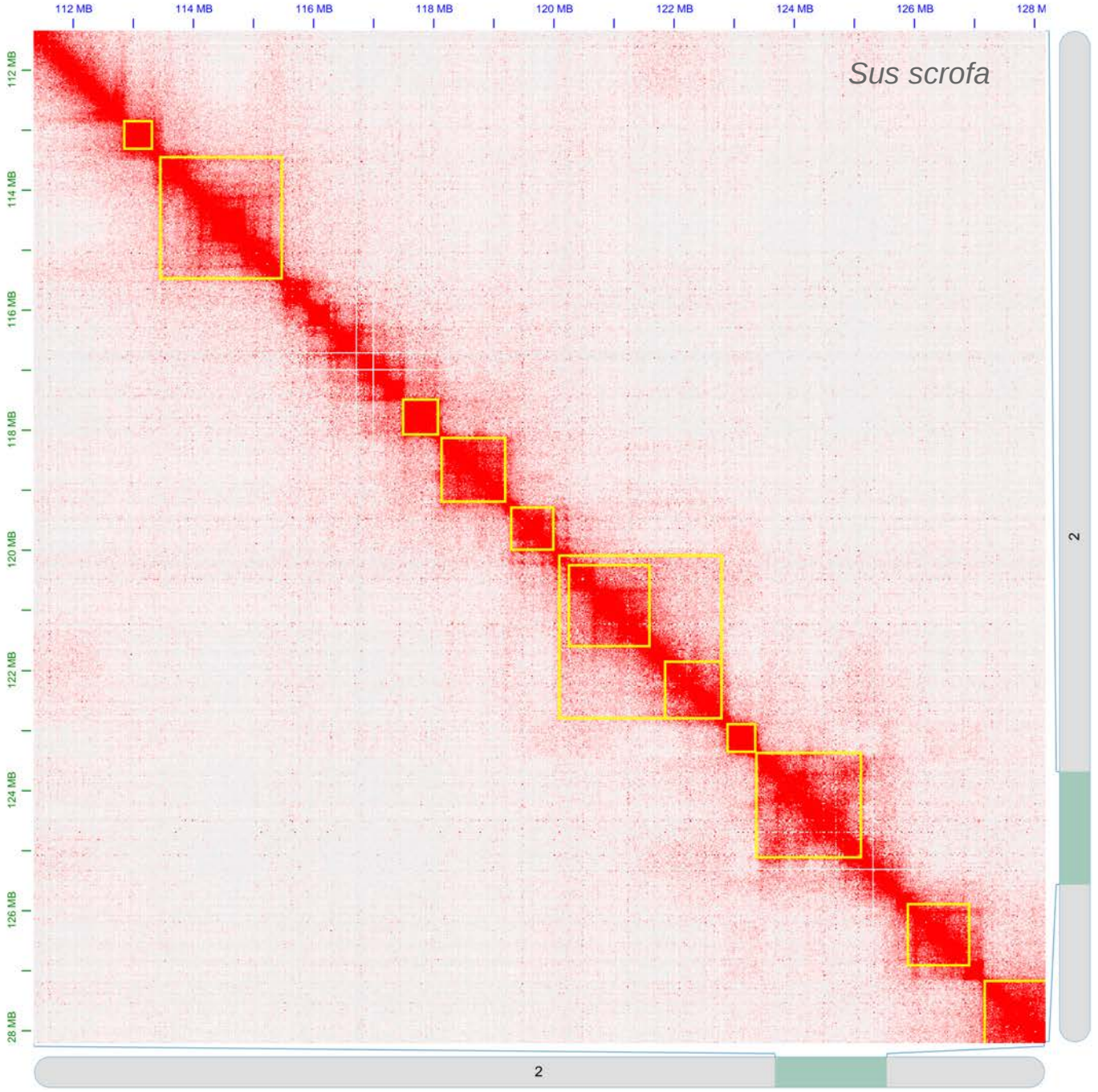
# Topologically Associating Domains (TADs)

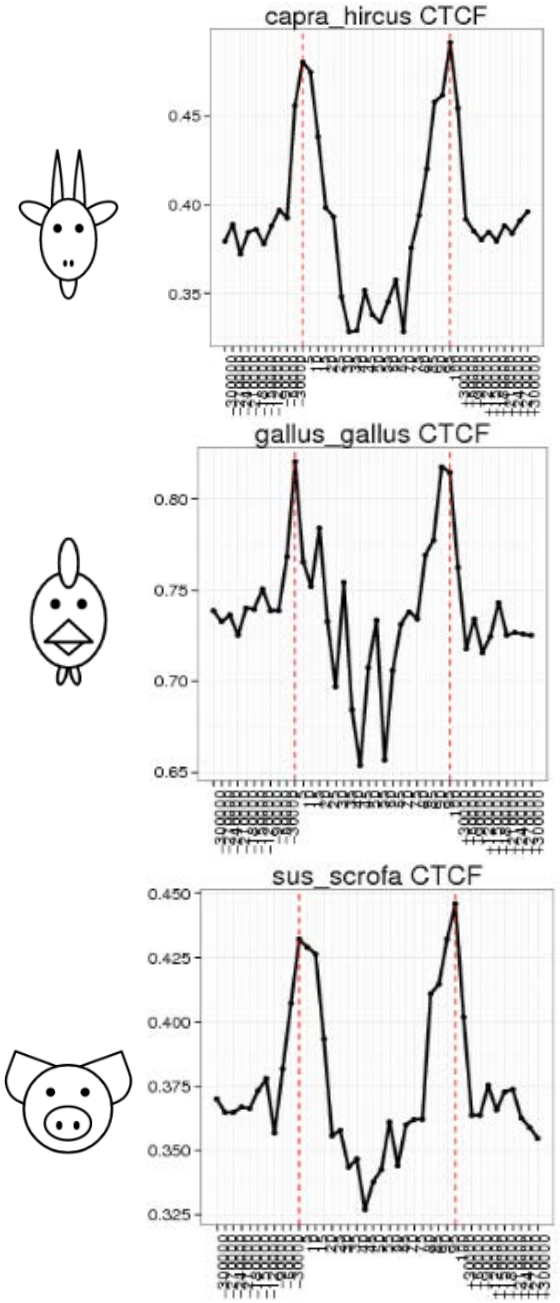



*Li et al, 2016, Scientific Reports*

# Topologically Associating Domains (TADs)



# 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)

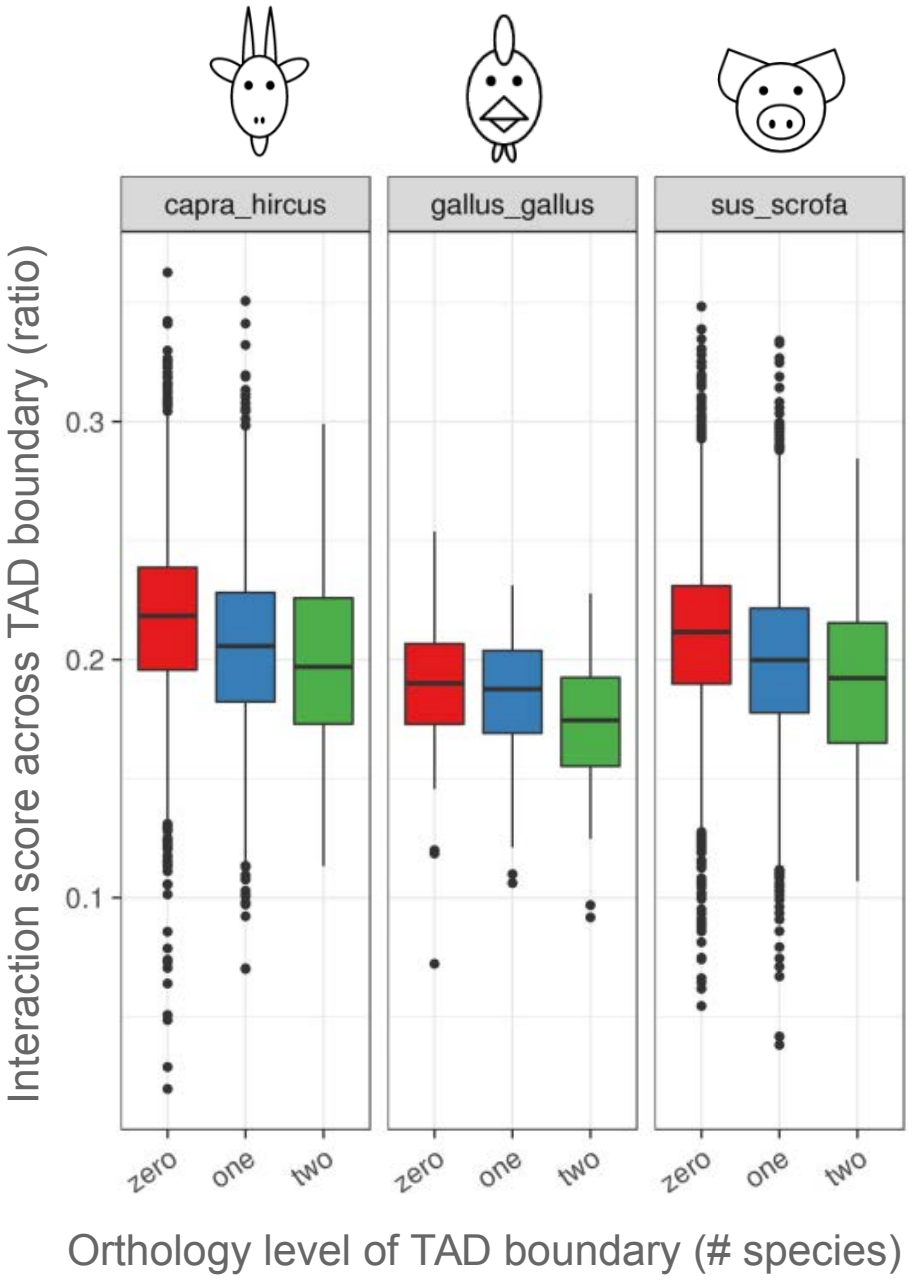
- Identify “orthologous” TAD boundaries (between 2 or 3 species):

  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

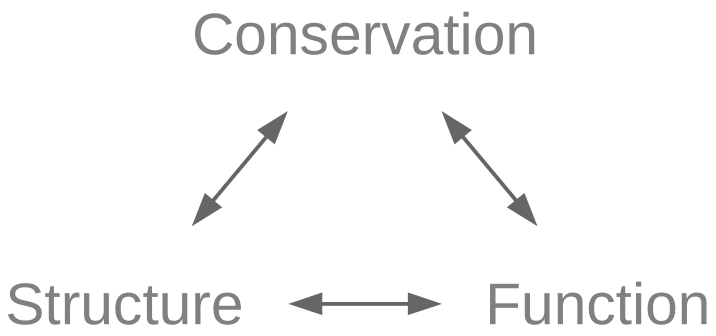
- Compute interaction score across TAD boundaries of each level (the lower the score, the stronger the insulation)

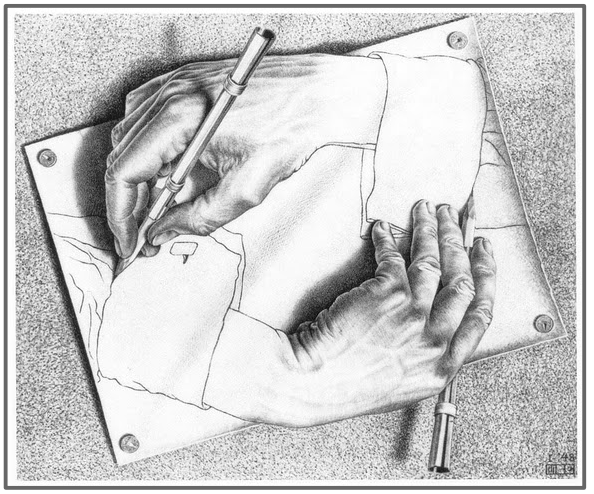
**conserved TAD boundaries have stronger insulations => evidence of selective pressure**



# 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



# Acknowledgments

## Management

Elisabetta Giuffra (coordination)

Sylvain Foissac (coordination)

Sandrine Lagarrigue

Marie-Hélène Pinard-Van der Laan

## Sampling and assays

Hervé Acloque

Cécile Berri

Fany Blanc

Sophie Dhorne-Pollet

Françoise Drouet

Diane Esquerre

Stéphane Fabre

Joël Gautron

Adeline Goubil

Sonia Lacroix-Lamandé

Fabrice Laurent

Florence Mompart

Pascale Queré

Michèle Tixier-Boichard

Gwenola Tosser-Klopp

Silvia Vincent-Naulleau

[ ... ]

## Data analysis

Philippe Bardou

Cédric Cabau

Elisa Crisci

Thomas Derrien

Sarah Djebali-Quelen

Sylvain Foissac

Christine Gaspin

Ignacio Gonzalez

Christophe Klopp

Sandrine Lagarrigue

Sylvain Marthey

Maria Marti-Marimon

Raphaelle Momal-Leisenring

Kylie Munyard

Kévin Muret

Andrea Rau

David Robelin

Magali San Cristobal

Nathalie Vialaneix

Matthias Zytnicki

## + Hi-C team @ Toulouse

**Hervé Acloque**
**Martine Bouissou-Matet**
**Sarah Djebali**
**Yvette Lahbib**
**Laurence Liaubet**
**Maria Marti**
**Florence Mompart**
**Pierre Neuvial**
**David Robelin**
**Nathalie Vialaneix**
**Alain Vignal**
**Matthias Zytnicki**

## + 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**