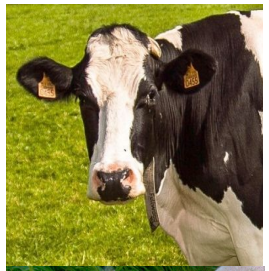
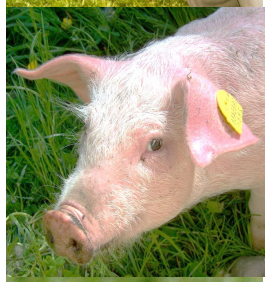
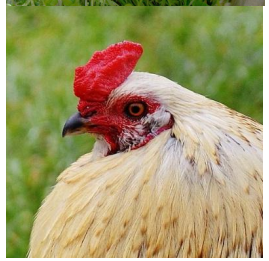
Profiling the landscape of transcription, chromatin accessibility and chromosome conformation of cattle, pig, chicken and goat genomes

[FAANG pilot project "FR-AgENCODE"]

Sarah Djebali

INRA - Toulouse - France

Genetics, Physiology and Breeding Systems Laboratory

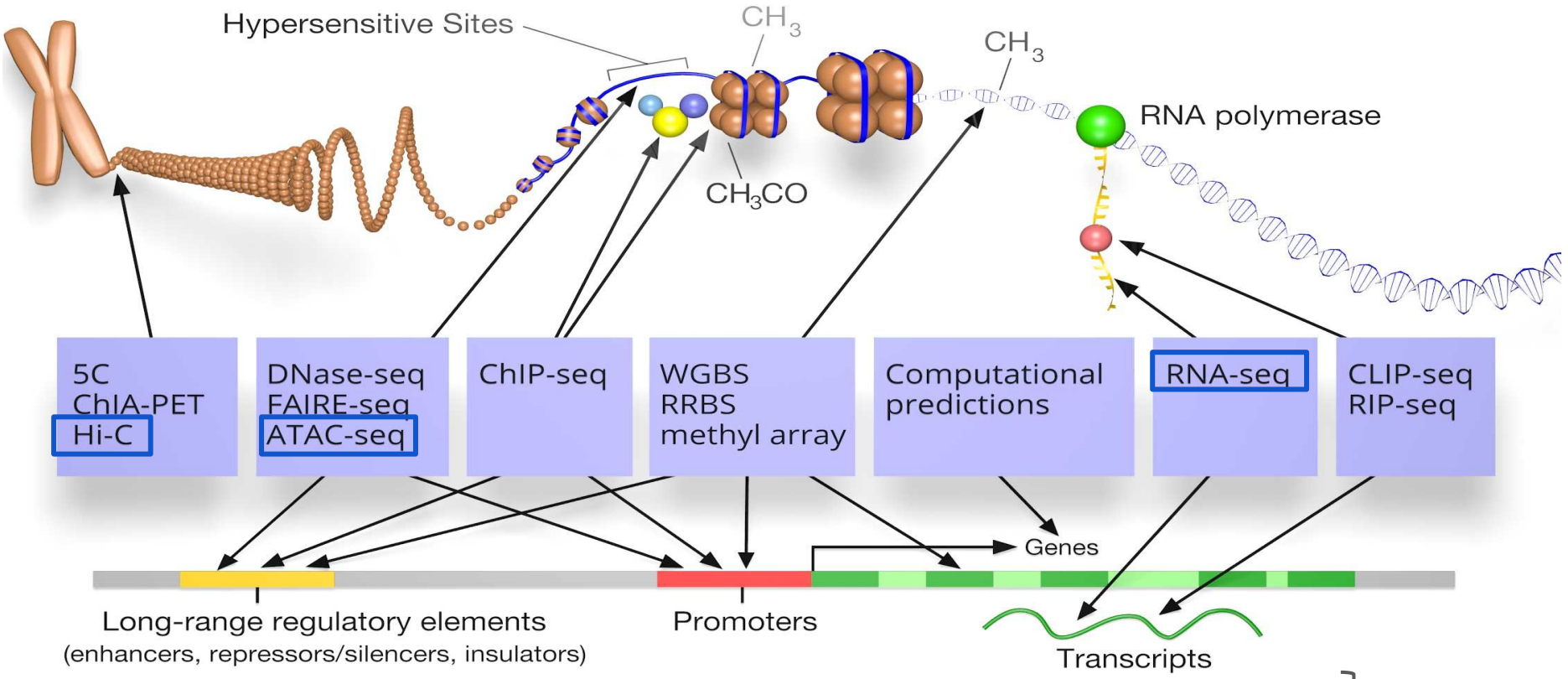
FAANG
Functional Annotation of Animal Genomes

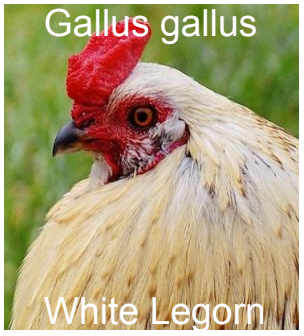
INRA
SCIENCE & IMPACT

AgreenSkills plus
Pathways for inventive researchers

36th International Society for Animal Genetics Conference – July 17th 2017 – UCD – Dublin (Ireland)

FR-AgENCODE data





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

Species	ATAC-seq*			HiC (liver only)			Long RNA-seq			Small RNA-seq		
	Library preparation	Sequencing	Data processing	Library preparation	Sequencing	Data processing	Library preparation	Sequencing	Data processing	Library preparation	Sequencing	Data processing
Bos taurus			**						**			
Capra hircus			**			**			**			
Gallus gallus			**			**			**			
Sus scrofa			**			**			**			

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

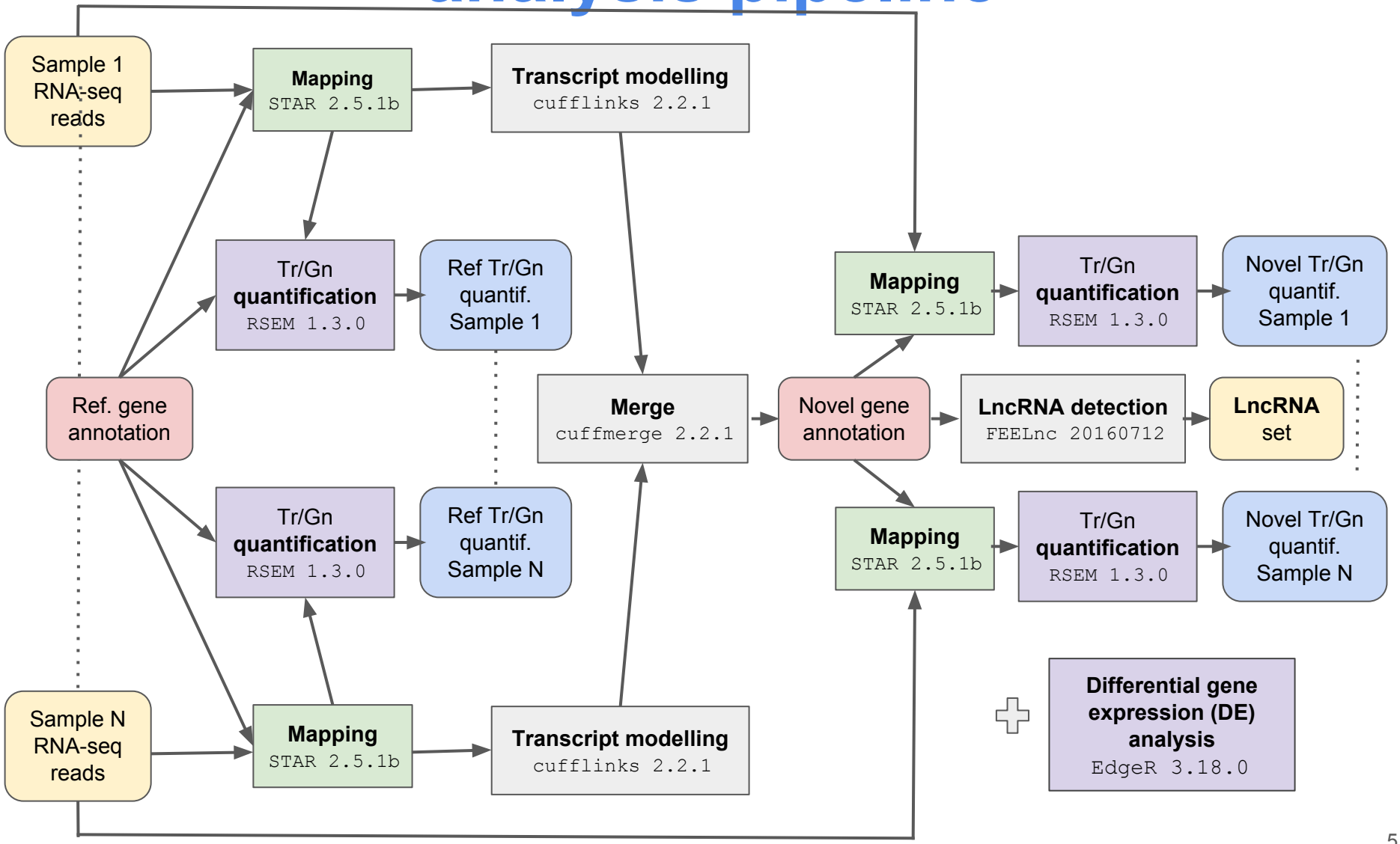
Our 16 animals are called cattle 1,2,3,4, goat 1,2,3,4, chicken 1,2,3,4 and pig 1,2,3,4:

- 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



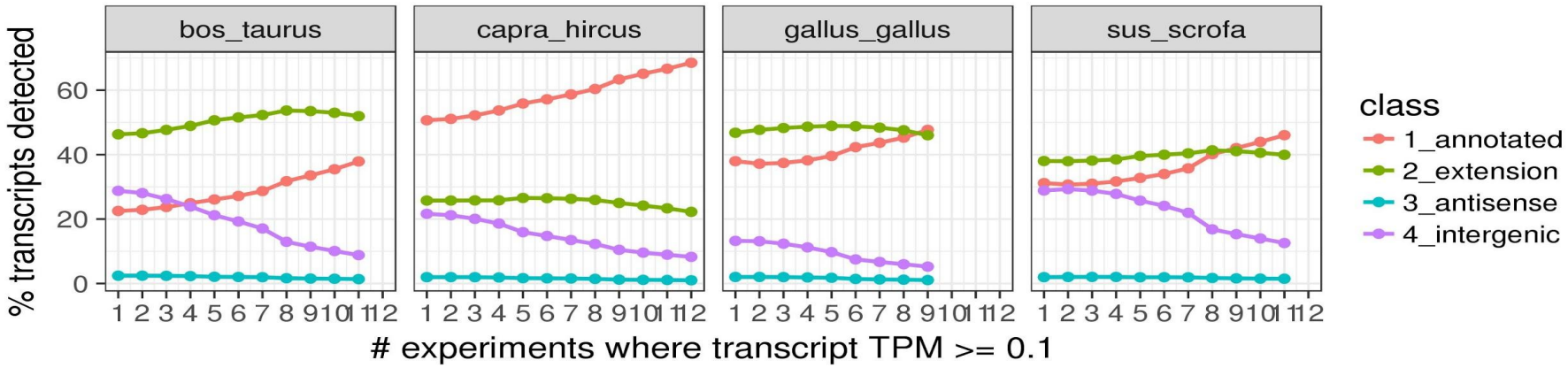
Nextflow implementation @ https://github.com/skptic/lncRNA-Annotation-nf

Many novel transcripts are found, which are globally less expressed than the annotated ones

Species	Genome / Gene annotation	Annotated Transcripts			Number of **novel** transcripts detected*	Number of **novel lncRNAs**
		Total number	Detected*			
			#	% of total		
Bos taurus	UMD 3.1 / Ensembl 84	**26,740**	16,100	**60.2**	**65,538**	**7,929**
Capra hircus	CHIR_ARS 1 / NCBI	**53,266**	34,442	**64.7**	**38,197**	**NA**
Gallus gallus	GalGal 5 / Ensembl 87	**38,118**	22,898	**60.1**	**41,116**	**3,264**
Sus scrofa	SScrofa 10.2 / Ensembl 84	**30,585**	18,746	**61.3**	**57,978**	**6,581**

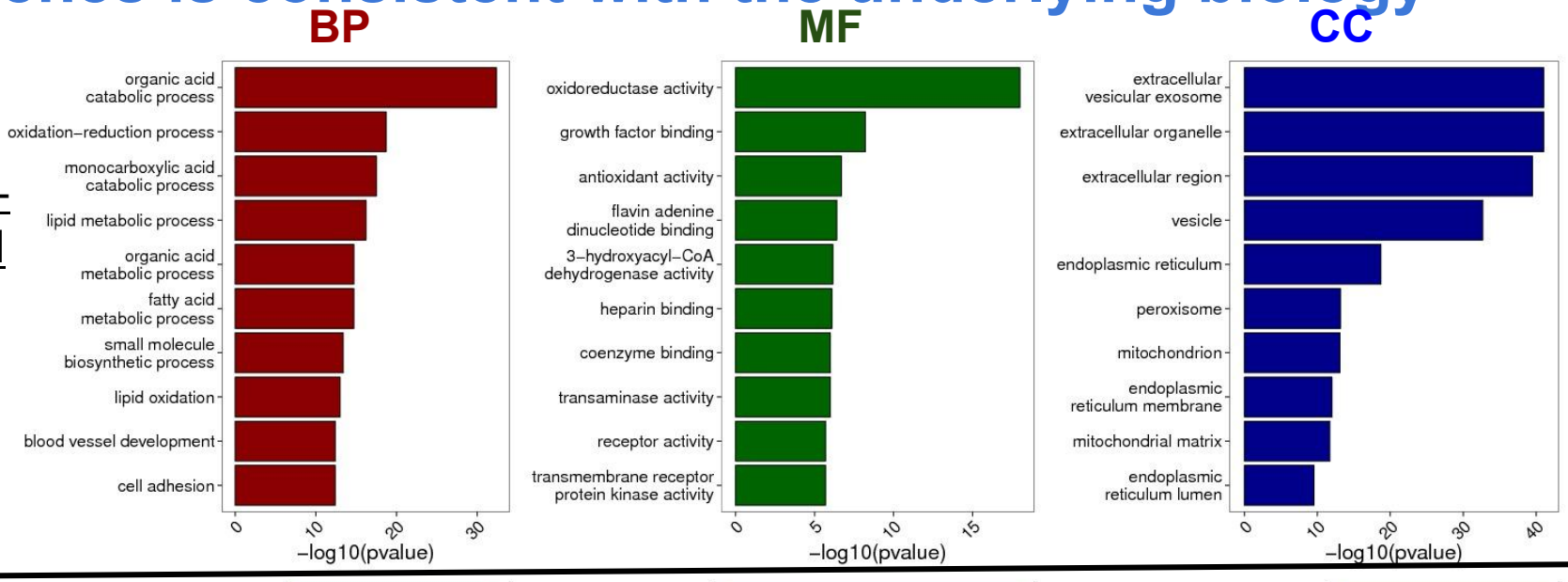
* with TPM ≥ 0.1 in ≥ 2 samples

See Kevin Muret's talk just after this talk

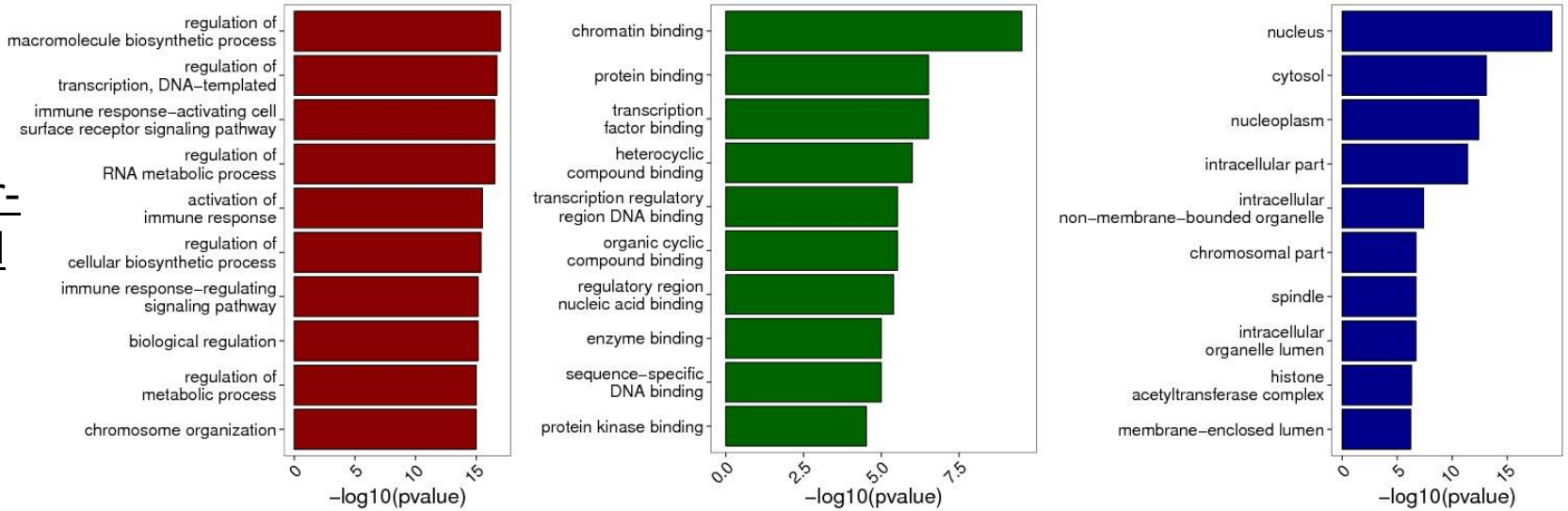


Biological process (BP), molecular function (MF), cell compartment (CC) GO term enrichment on DE genes is consistent with the underlying biology

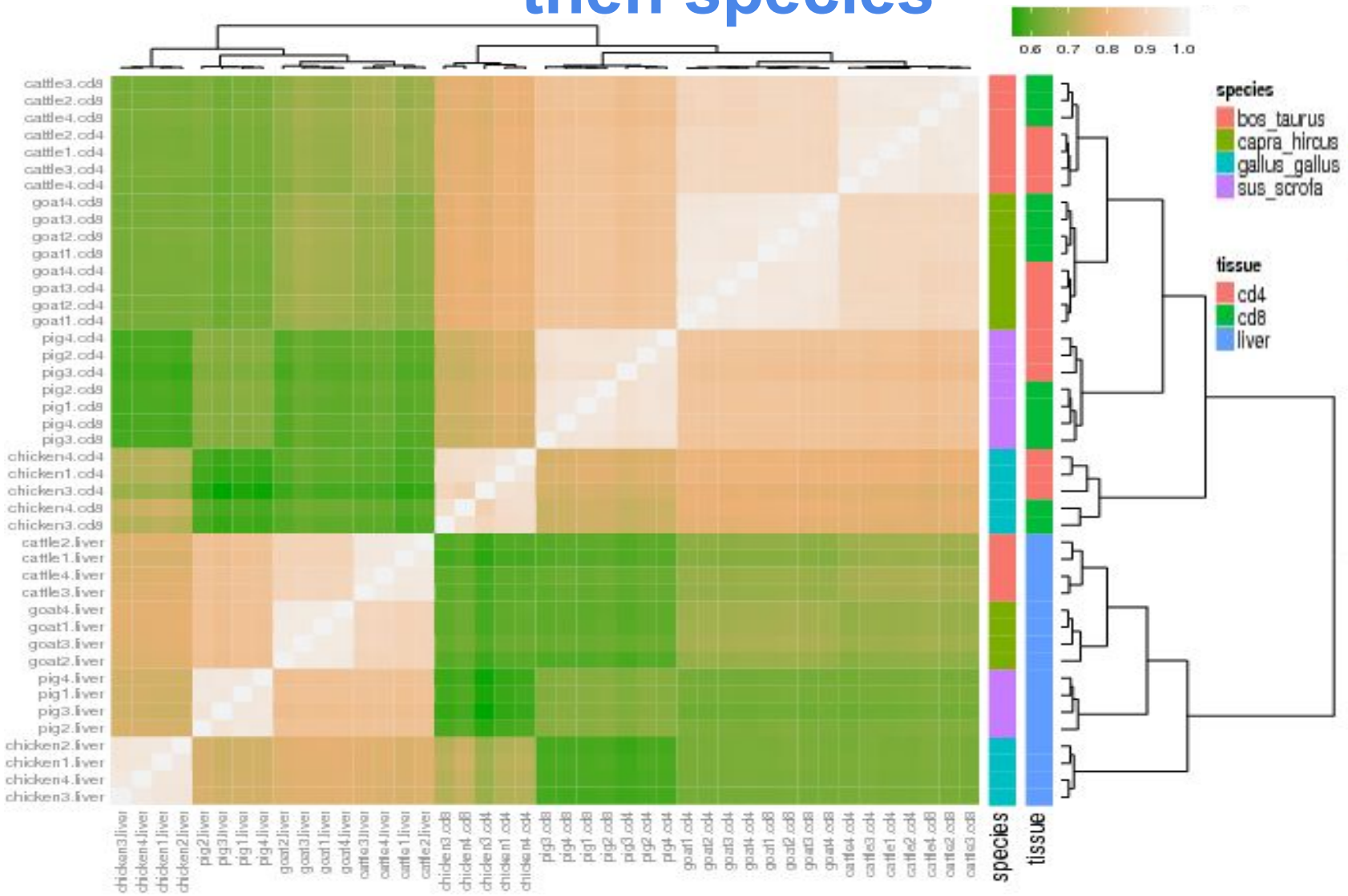
Liver over-expressed genes



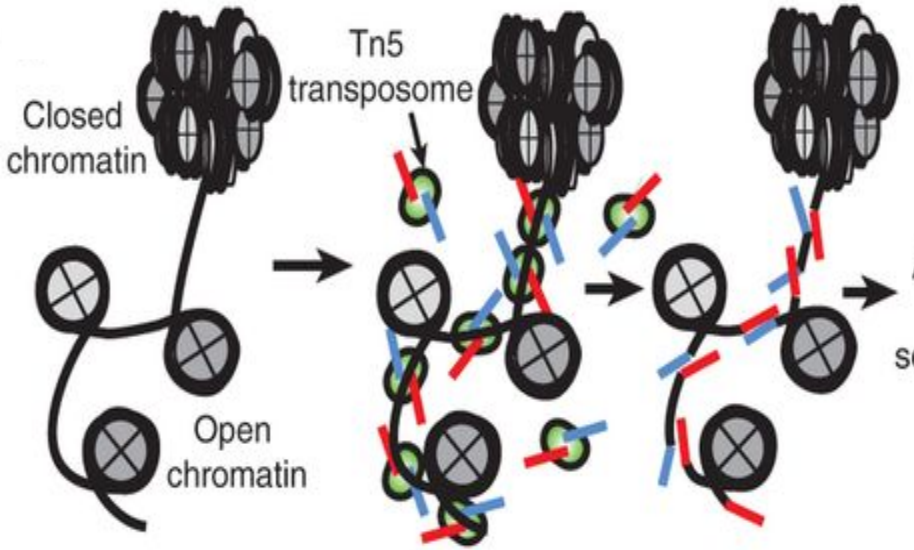
T cell over-expressed genes



All species RNA-seq hierarchical clustering first separates liver from immune cells, and then species

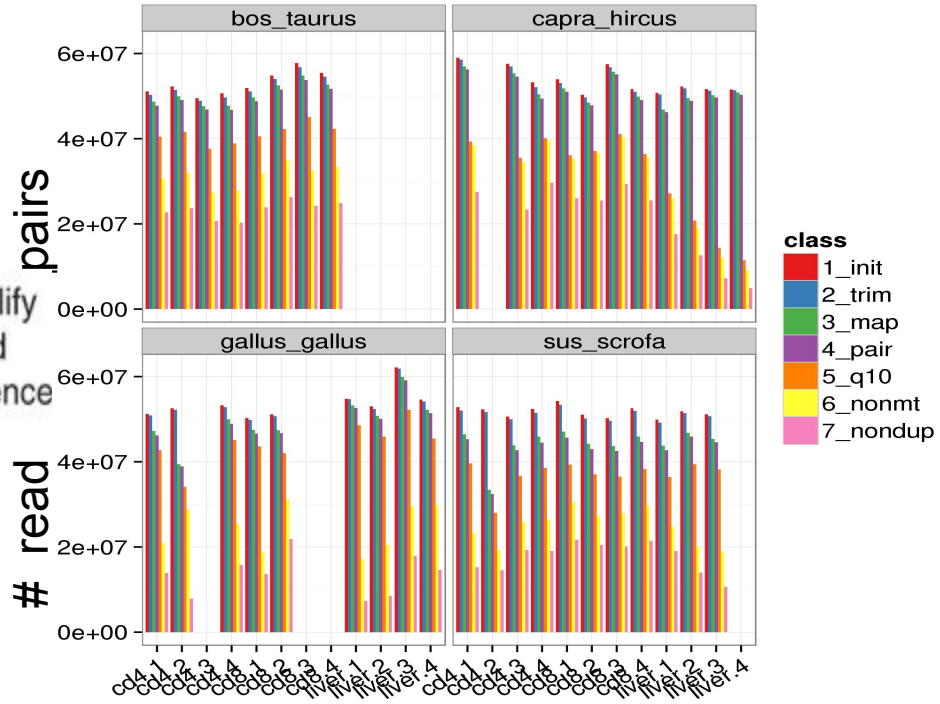


ATAC-seq for open chromatin regions

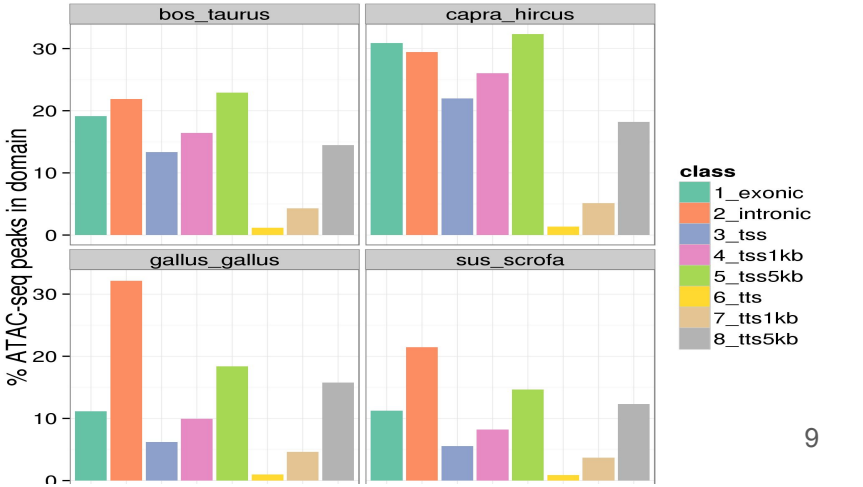


Buenrostro et al., Nature Methods, 2013

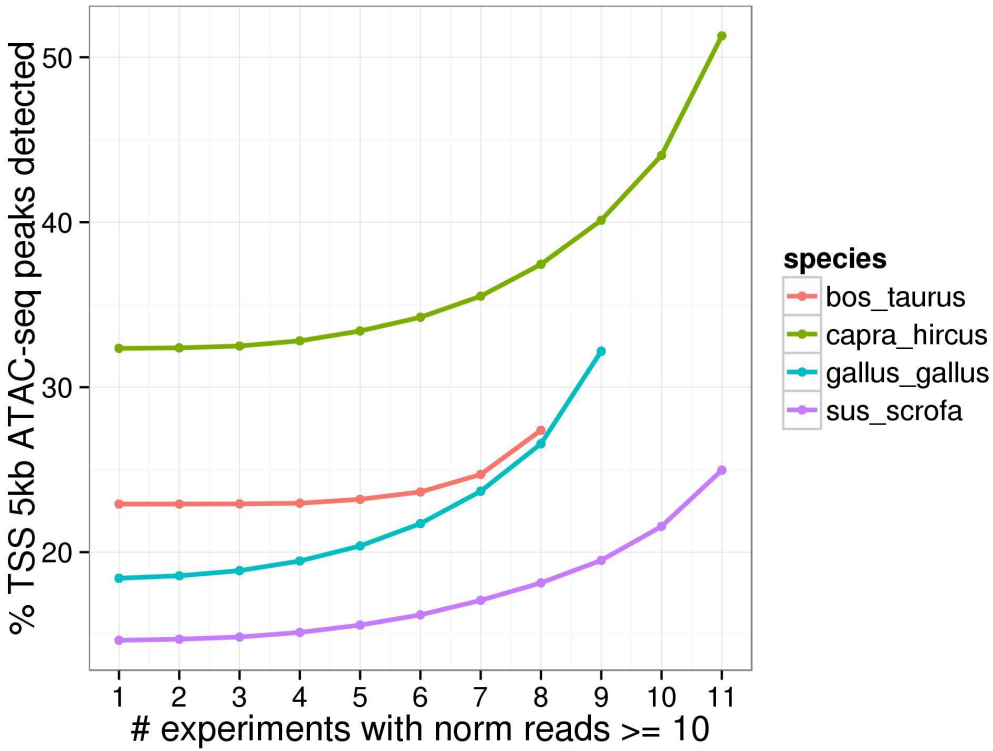
50 million read pairs per sample



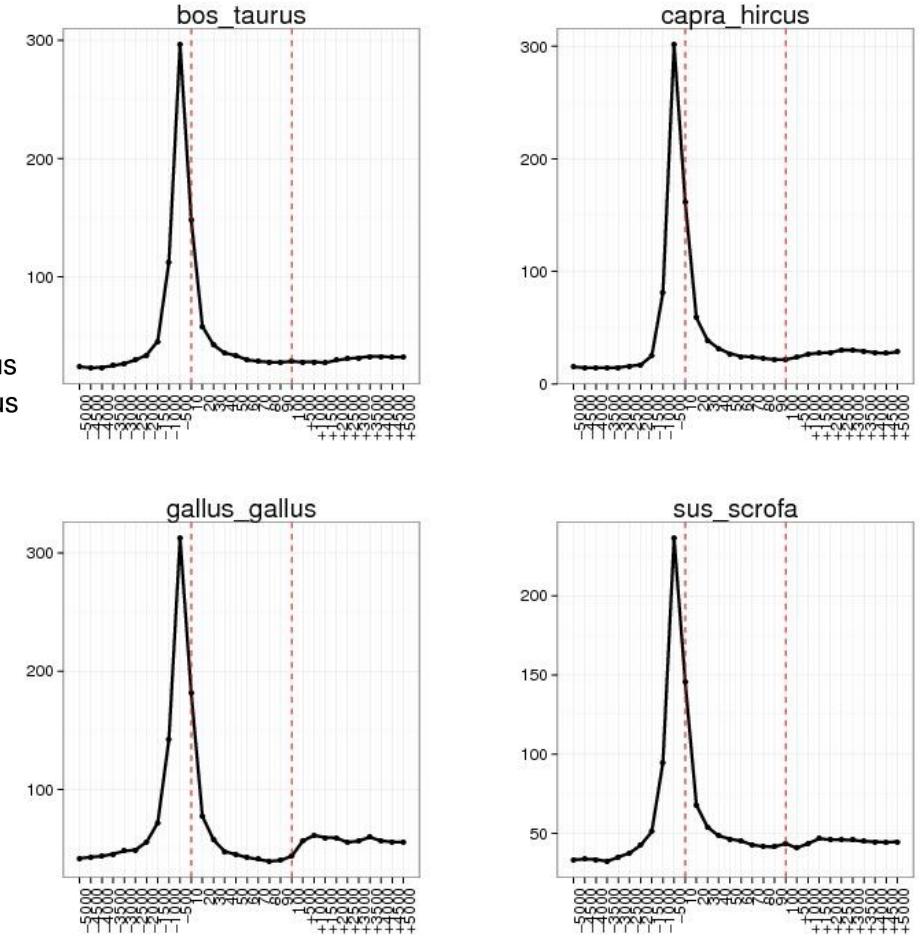
Species	Number of atac-seq peaks	Genome size (bp)	ATAC-seq peak coverage	
			# bp	% of genome
Bos taurus	**58,384**	2,670,422,299	42,288,741	**1.58**
Capra hircus	**46,901**	2,922,813,246	32,951,265	**1.13**
Gallus gallus	**116,893**	1,230,258,557	50,931,713	**4.14**
Sus scrofa	**120,914**	2,808,525,991	72,480,471	**2.58**



ATAC-seq peaks at Transcription Start Sites (TSS); those peaks are more ubiquitous



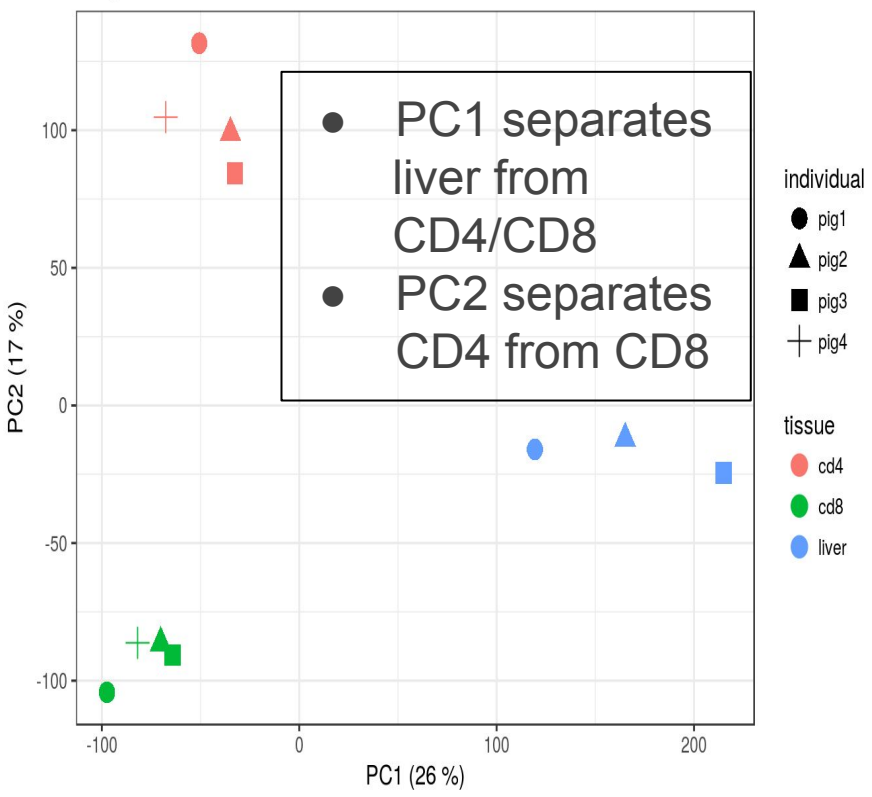
Open chromatin regions located close to TSSs are more ubiquitous than other open chromatin regions



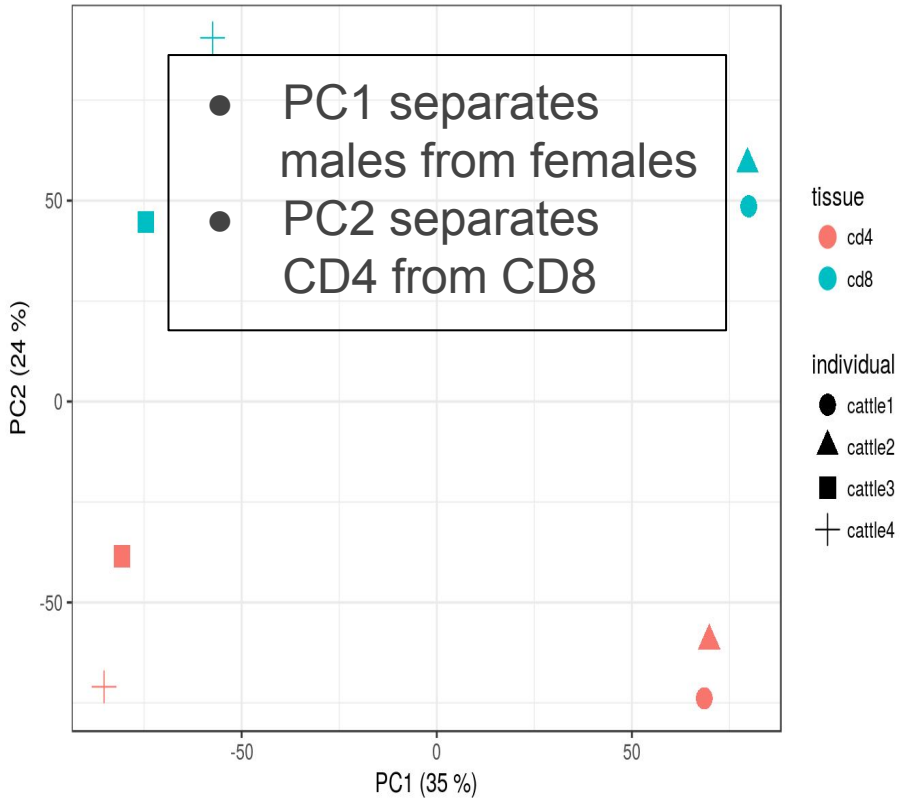
ATAC-seq coverage peaks at annotated TSS in all species

ATAC-seq PCA first separates liver from immune cells; in absence of liver, ATAC-seq PCA first separates males from females

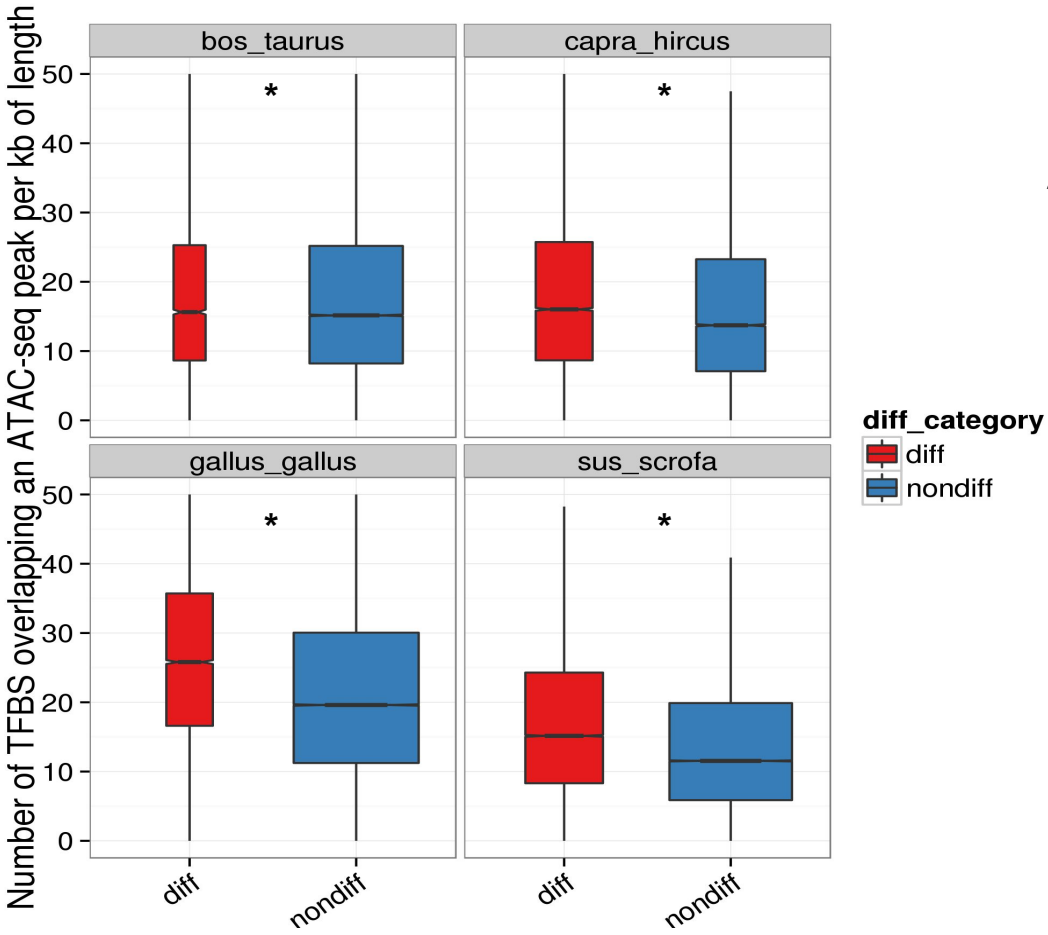
Sus scrofa



Bos taurus



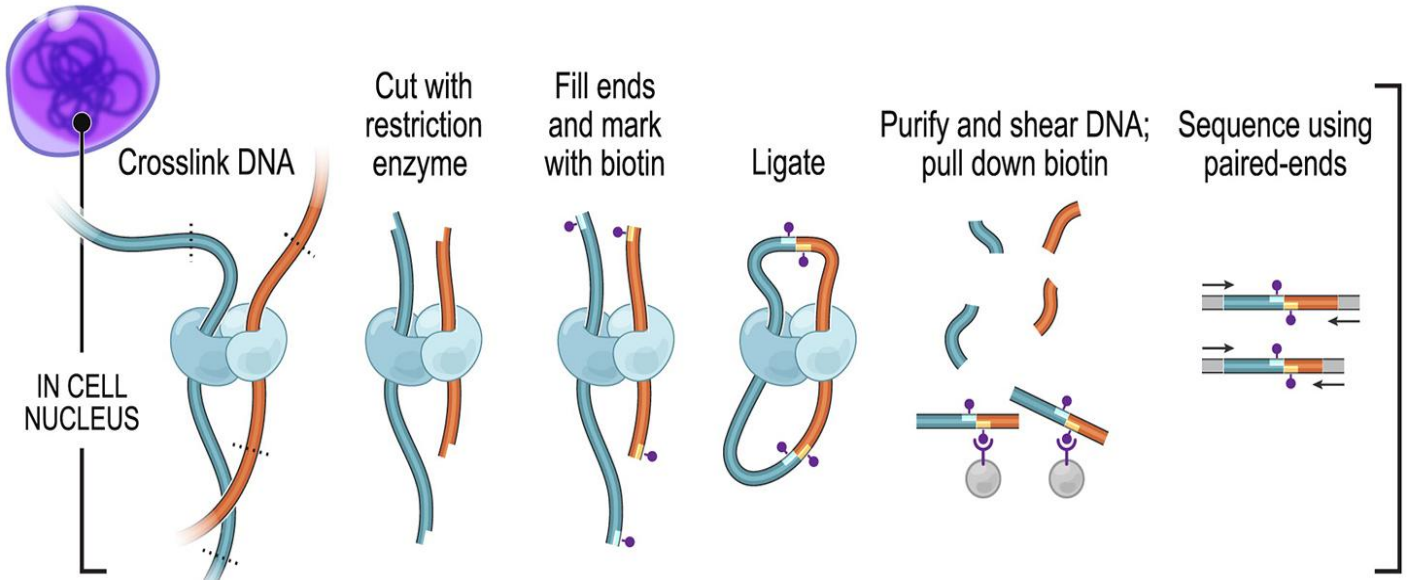
Differential ATAC-seq peaks are more likely to be regulatory



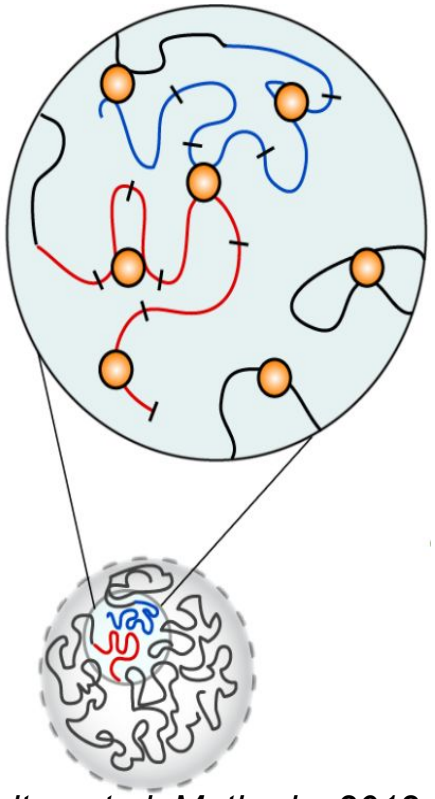
Between tissue **differential** ATAC-seq peaks have a higher **TFBS density** than non differential ATAC-seq peaks (Wilcoxon test, p-value < 10^{-15})

→ Differential ATAC-seq peaks are more likely to have a **regulatory** role

HiC for 3D genomic structure

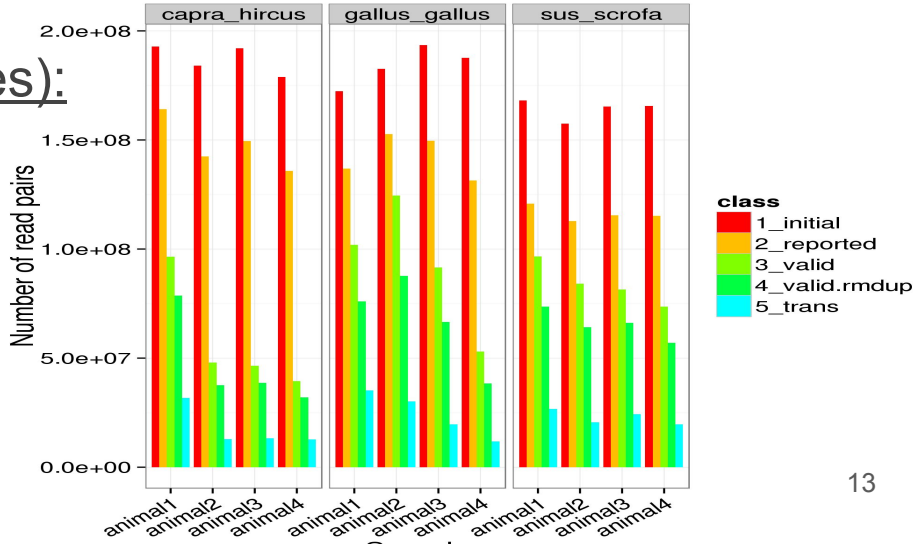


Rao et al, Cell, 2014

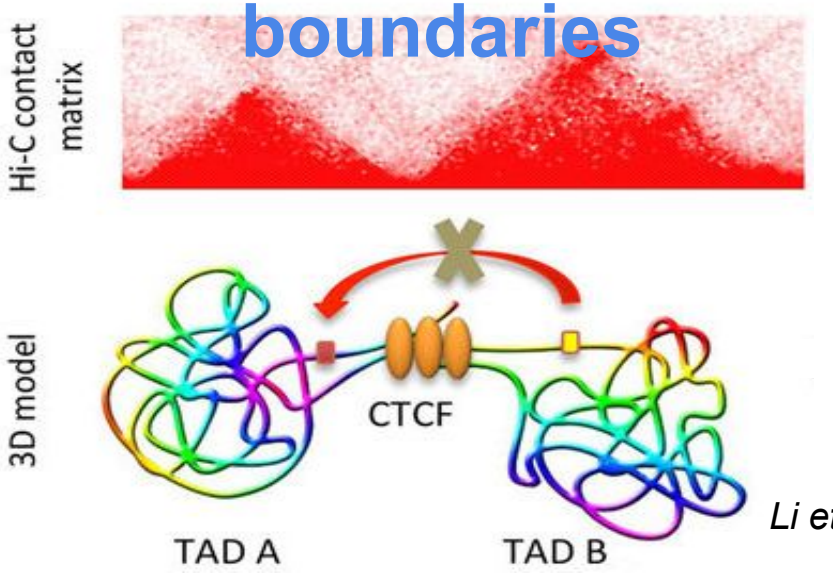


Belton et al, Methods, 2012

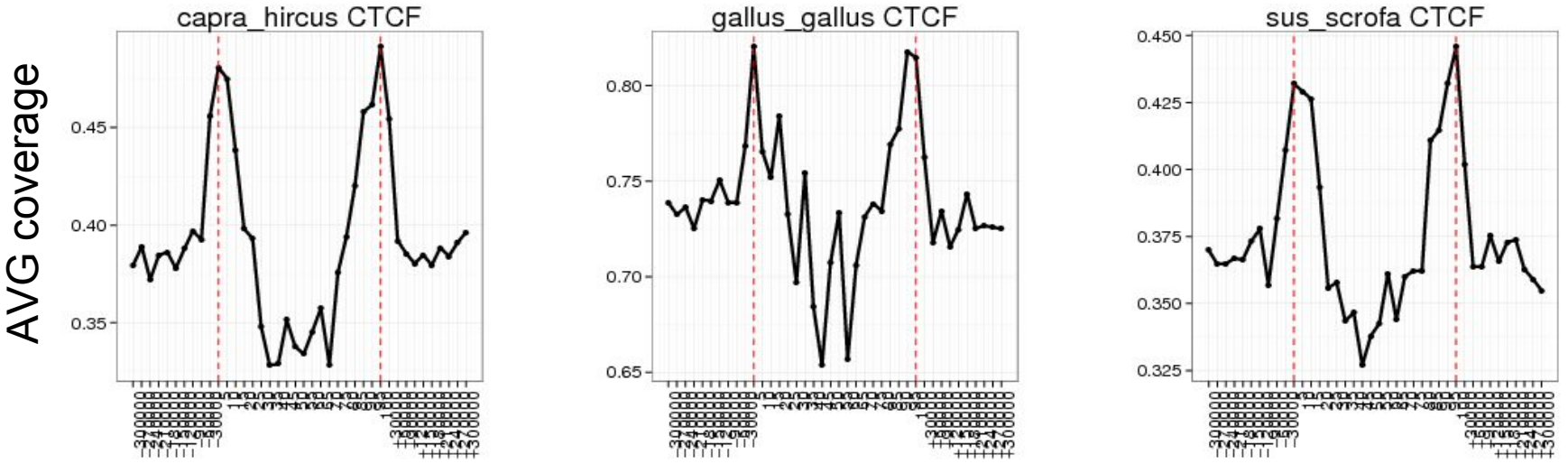
- HiC data (liver of 4 animals of 3 species):
  - 180 million read pairs per sample
- HiC data analysis pipeline:
  - `HiC-Pro`: Read Mapping/Filtering/Normalization
  - `Armatus`: TAD calling
  - `HiTC`: A/B compartment calling



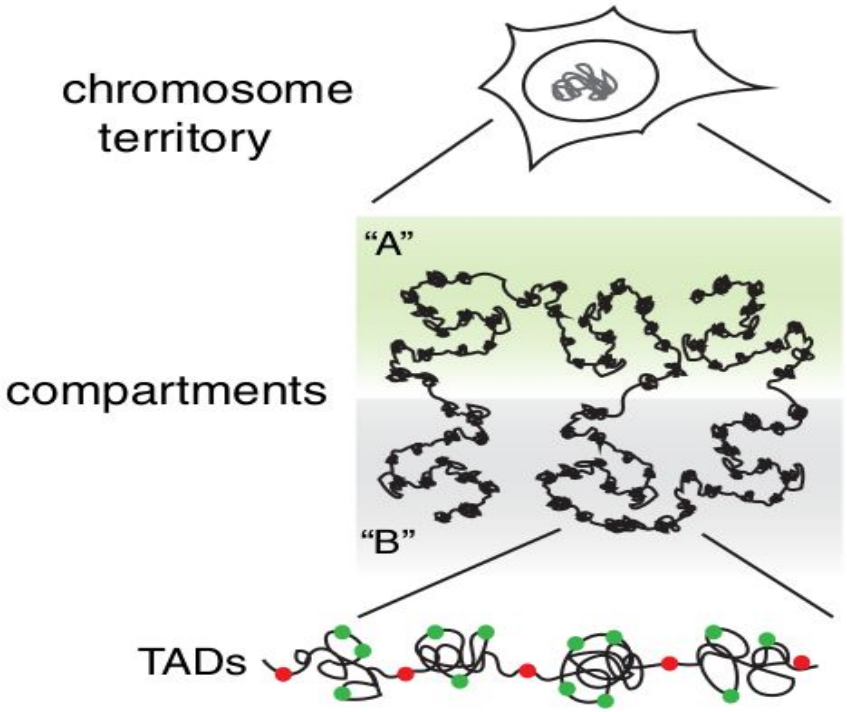
Predicted CTCF binding sites peak at Topologically Associating Domain (TAD) boundaries



Li et al, Scientific Reports, 2016

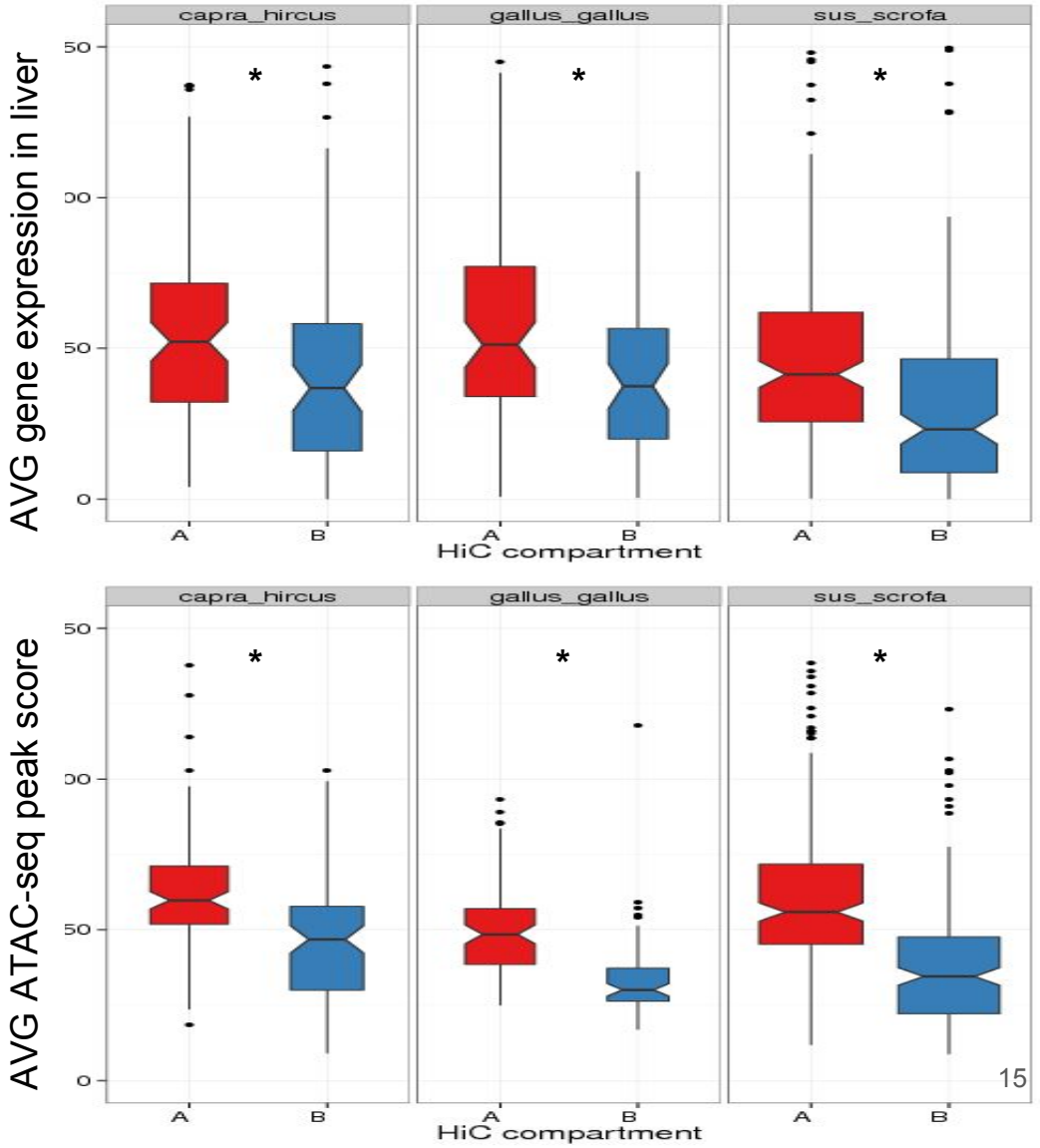


There is global consistency between RNA-seq, ATAC-seq and HiC data



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

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



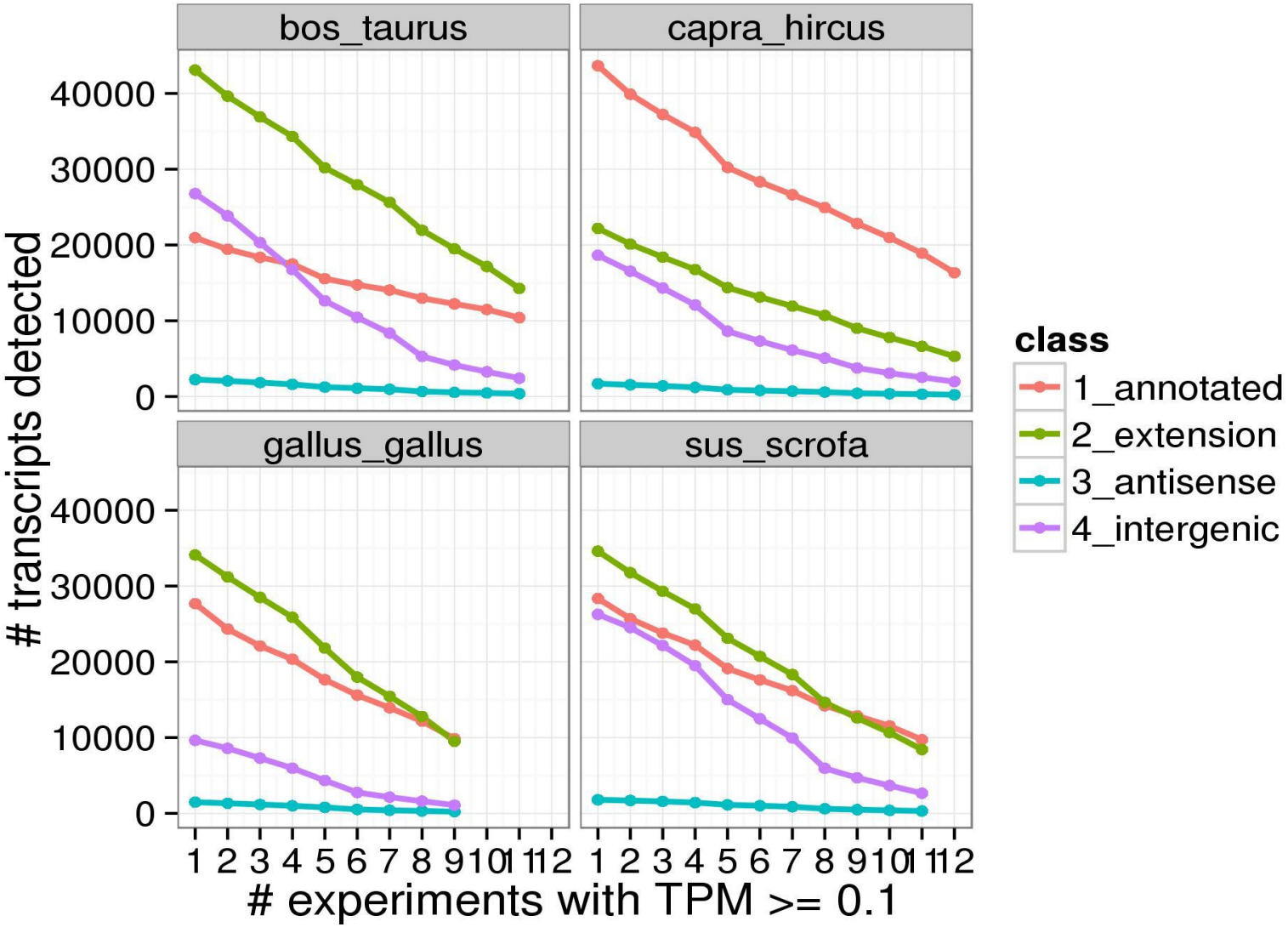
Acknowledgements

- **Management:**
  - Elisabetta Giuffra
  - Sylvain Foissac
  - Sandrine Lagarrigue
  - Marie-Hélène Pinard
- **Sampling:**
  - Michèle Tixier-Boichard
  - Stéphane Fabre
  - Gwenola Tosser-Klopp
  - Pascale Queré
  - 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

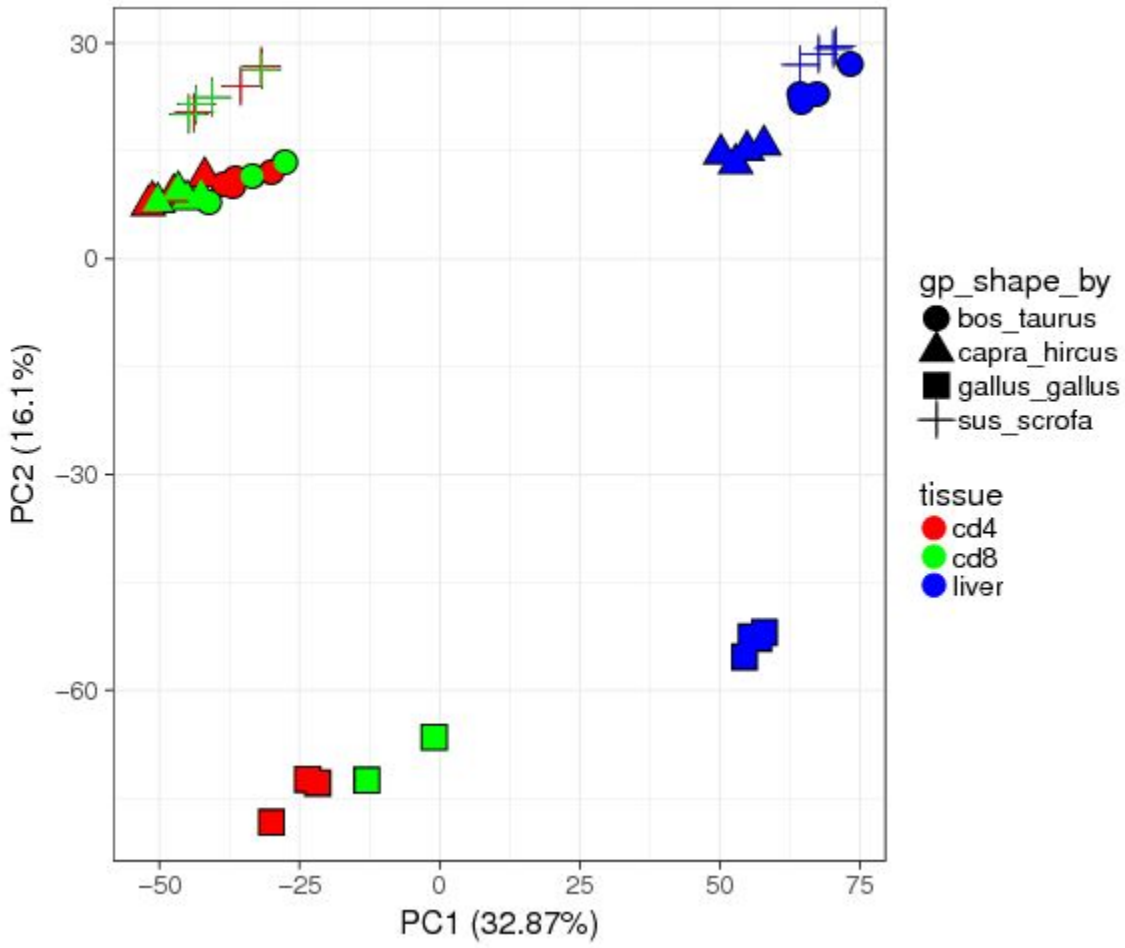
Thanks for your attention!

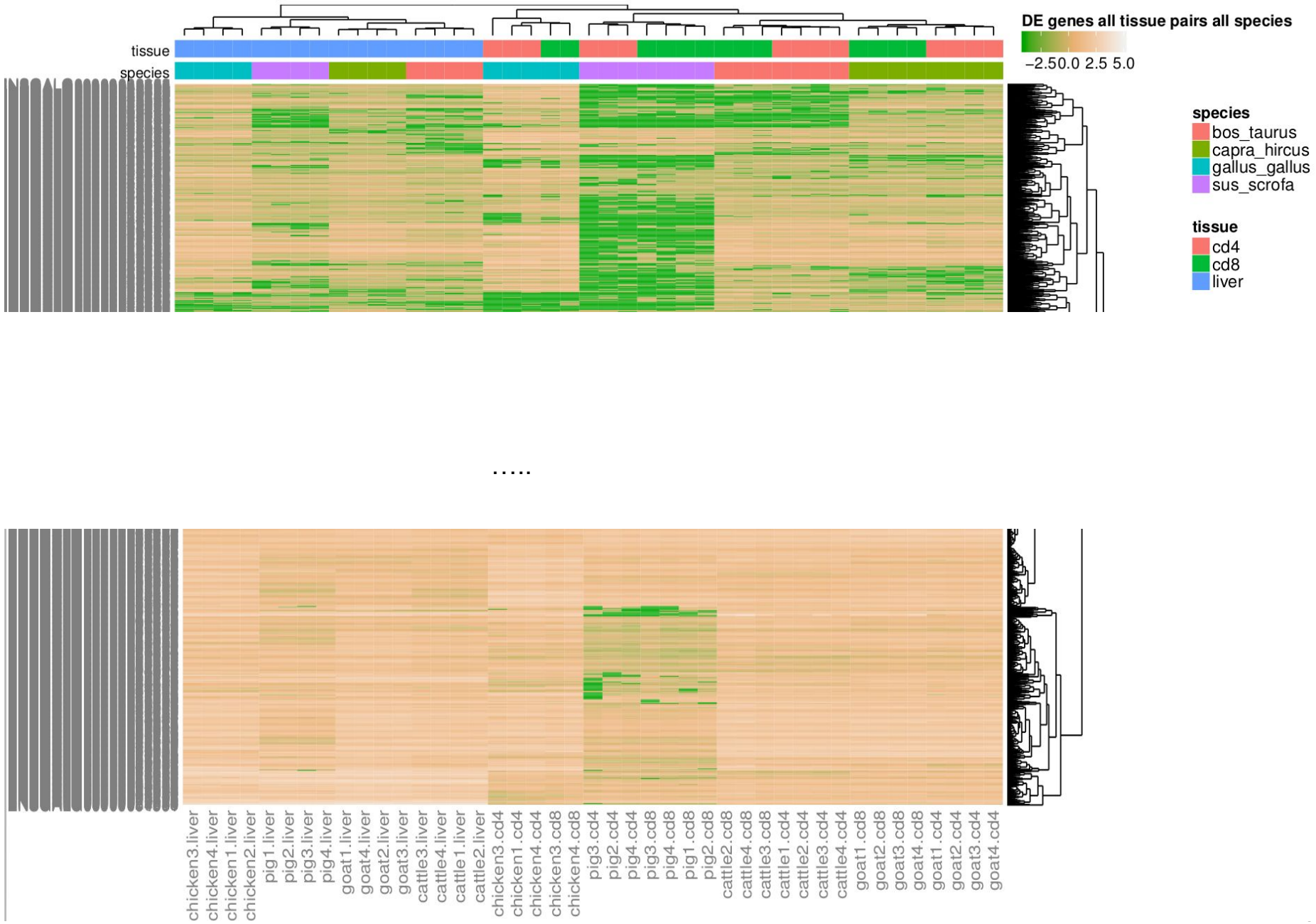
Additional slides

Transcripts detected and their expression

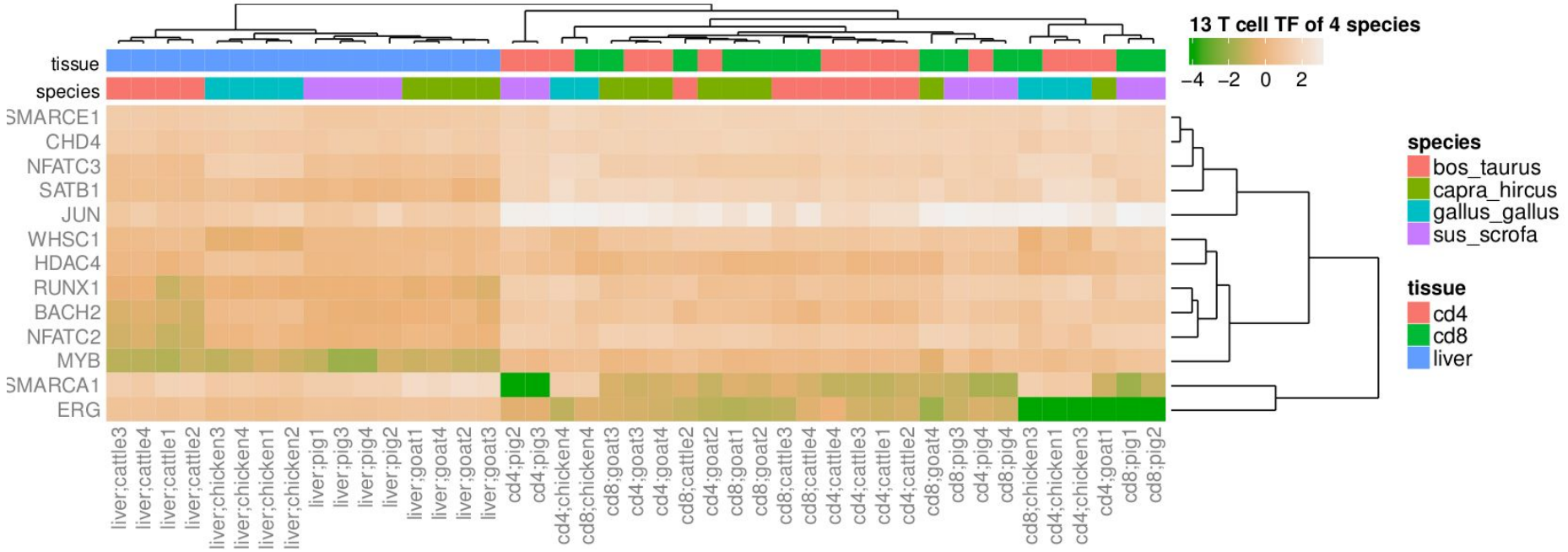


All RNA-seq experiment PCA (~7300 orthologous genes)

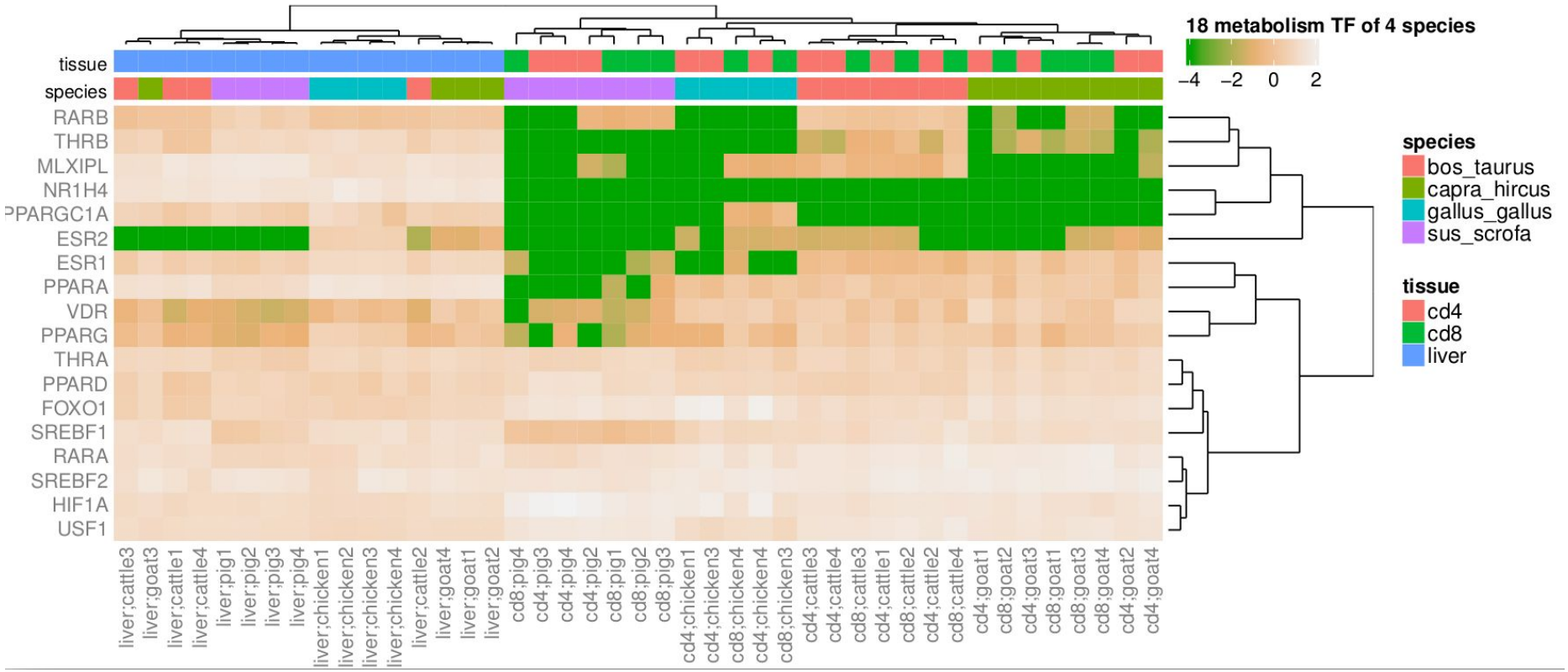



tissue
species
DE genes all tissue pairs all species
−2.5 0.0 2.5 5.0
species
bos_taurus
capra_hircus
gallus_gallus
sus_scrofa
tissue
cd4
cd8
liver
.....
chicken3.liver
chicken4.liver
chicken1.liver
chicken2.liver
pig1.liver
pig2.liver
pig3.liver
pig4.liver
goat1.liver
goat4.liver
goat2.liver
goat3.liver
cattle3.liver
cattle4.liver
cattle1.liver
cattle2.liver
chicken3.cd4
chicken1.cd4
chicken4.cd4
chicken3.cd8
chicken4.cd8
pig3.cd4
pig2.cd4
pig4.cd4
pig3.cd8
pig4.cd8
pig1.cd8
pig2.cd8
cattle2.cd8
cattle4.cd8
cattle3.cd8
cattle1.cd4
cattle2.cd4
cattle3.cd4
cattle4.cd4
goat1.cd8
goat2.cd8
goat3.cd8
goat4.cd8
goat1.cd4
goat2.cd4
goat3.cd4
goat4.cd4

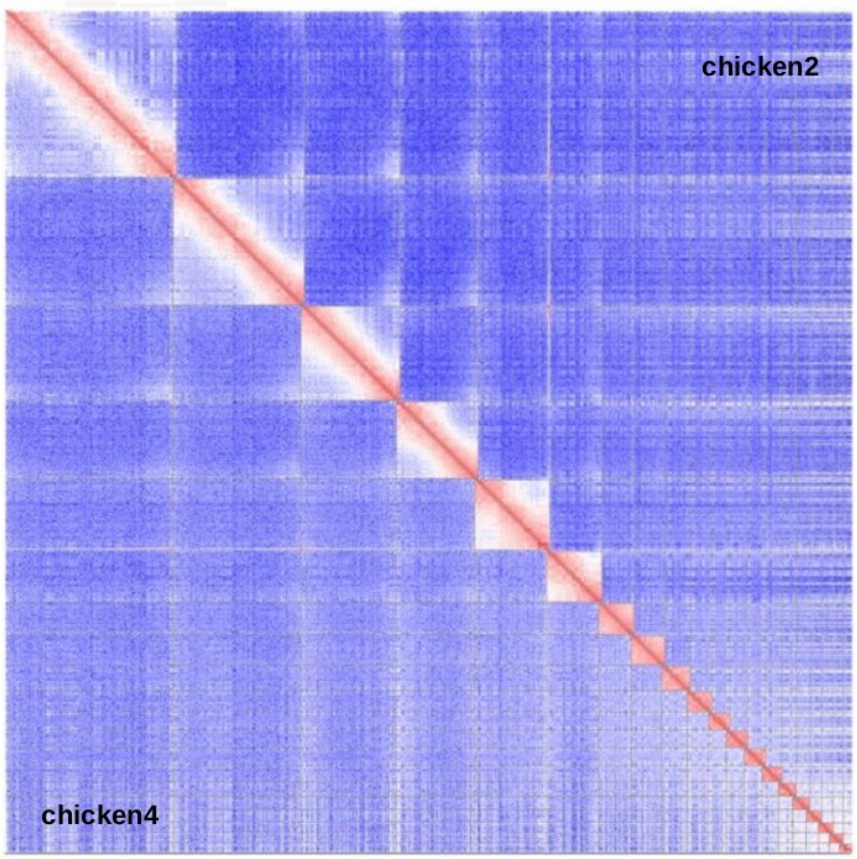
Expression of the 13 T cell TF common to the 4 species (adding goat using gene name)

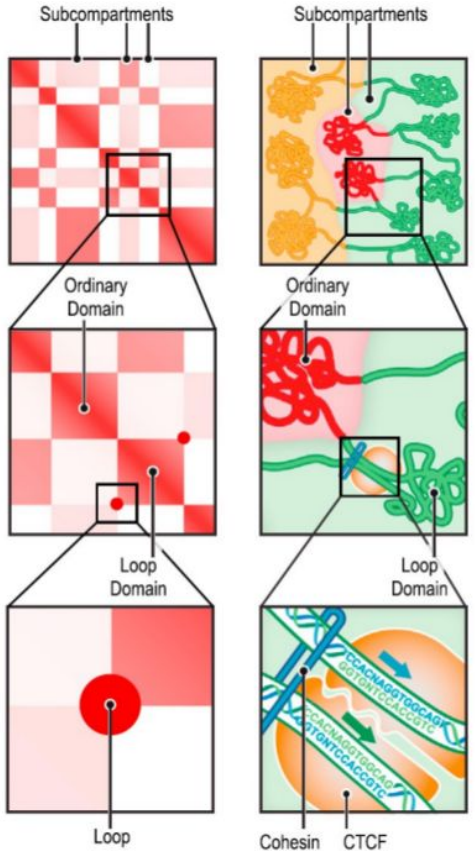


Expression of the 18 metabolism TF common to the 4 species (adding goat using gene name)



HiC results: interaction matrices

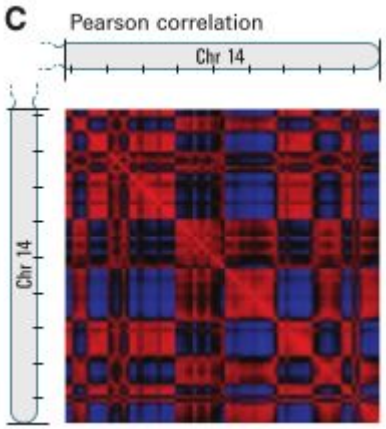


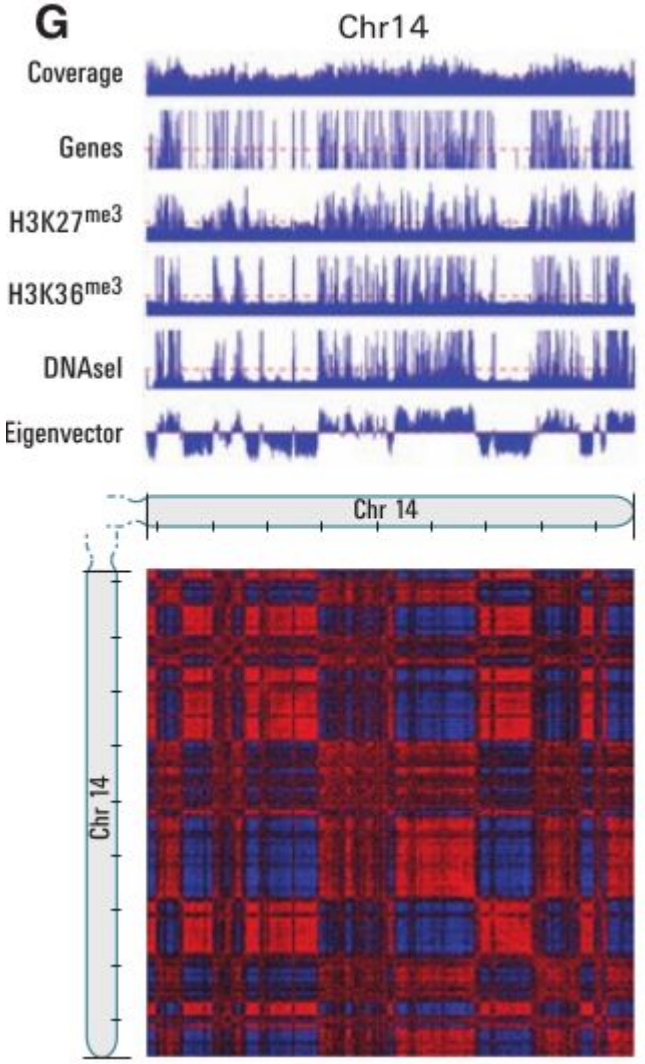


Lieberman-aiden et al, 2009
Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome

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 ith row and jth 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