Livestock genome annotation: transcriptome and chromatin structure profiling in cattle, goat, chicken and pig

Sylvain Foissac, Sarah Djebali, Kylie Munyard, Nathalie Vialaneix, Andrea Rau, Kévin Muret, Diane Esquerré, Matthias Zytnicki, Thomas Derrien, Philippe Bardou, et al.

RESEARCH

Livestock genome annotation: transcriptome and chromatin structure profiling in cattle, goat, chicken and pig

Sylvain Foissac[1†], Sarah Djebali[1†], Kylie Munyard[2], Nathalie Villa-Vialaneix[3], Andrea Rau[4], Kevin Muret[5], Diane Esquerré[1,6], Matthias Zytnicki[3], Thomas Derrien[7], Philippe Bardou[1], Fany Blanc[4], Cédric Cabau[1], Elisa Crisci[4], Sophie Dhorne-Pollet[4], Françoise Drouet[8], Ignacio Gonzales[3], Adeline Goubil[4], Sonia Lacroix-Lamandé[8], Fabrice Laurent[8], Sylvain Marthey[4], Maria Marti-Marimon[1], Raphaelle Momal-Leisenring[4], Florence Mompart[1], Pascale Quéré[8], David Robelin[1], Magali San Cristobal[1], Gwenola Tosser-Klopp[1], Silvia Vincent-Naulleau[9], Stéphane Fabre[1], Marie-Hélène Pinard-Van der Laan[4], Christophe Klopp[3], Michèle Tixier-Boichard[4], Hervé Acloque[1,4], Sandrine Lagarrigue[5] and Elisabetta Giuffra[4*^]

[*]Correspondence:
elisabetta.giuffra@inra.fr
[4]GABI, AgroParisTech, INRA, Université Paris Saclay, F-78350 Jouy-en-Josas, France
Full list of author information is available at the end of the article
[†]Contributed equally. [^]To whom correspondence should be addressed: sylvain.foissac@inra.fr and elisabetta.giuffra@inra.fr.

Abstract

Background: Functional annotation of livestock genomes is a critical step to decipher the genotype-to-phenotype relationship underlying complex traits. As part of the Functional Annotation of Animal Genomes (FAANG) action, the FR-AgENCODE project aims at profiling the landscape of transcription (RNA-seq) and chromatin accessibility and conformation (ATAC-seq and Hi-C) in four livestock species representing ruminants (cattle, goat), monogastrics (pig) and birds (chicken), using three target samples related to metabolism (liver) and immunity (CD4+ and CD8+ T cells).

Results: Standardized protocols were applied to produce transcriptome and chromatin datasets for the four species. RNA-seq assays considerably extended the available catalog of protein-coding and non-coding transcripts. Gene expression profiles were consistent with known metabolic/immune functions and revealed differentially expressed transcripts with unknown function, including new lncRNAs in syntenic regions. The majority of ATAC-seq peaks of chromatin accessibility mapped to putative regulatory regions, with an enrichment of predicted transcription factor binding sites in differentially accessible peaks. Hi-C provided the first set of genome-wide maps of three-dimensional interactions across livestock and showed consistency with results from gene expression and chromatin accessibility in topological compartments of the genomes.

Conclusions: We report the first multi-species and multi-assay genome annotation results obtained by a FAANG pilot project. The global consistency between gene expression and chromatin structure data in these four livestock species confirms previous findings in model animals. Overall, these results emphasize the value of FAANG for research on domesticated animals and strengthen the importance of future meta-analyses of the reference datasets being generated by this community on different species.

Keywords: functional annotation; livestock; RNA-seq; ATAC-seq; Hi-C

Background

The majority of complex trait-associated loci lie outside protein-coding regions, and comparative genomics studies have shown that the majority of mammalian-conserved and recently adapted regions consist of non-coding elements [1–3]. This evidence prompted the first large-scale efforts into genome annotation for human and model organisms [4–6]. The genome-wide annotation maps generated by these projects helped to shed light on the main features of genome activity. For example, chromatin conformation or transcription factor occupancy at regulatory elements can often be directly tied to the biology of the specific cell or tissue under study [3, 7, 8]. Moreover, although a subset of core regulatory systems are largely conserved across humans and mice, the underlying regulatory systems often diverge substantially [9–11], implying that understanding the phenotypes of interest requires organism-specific information for any specific physiological phase, tissue and cell.

The Functional Annotation of Animal Genomes (FAANG) initiative (`www.faang.org`) aims to support and coordinate the community in the endeavor of creating reference functional maps of the genomes of domesticated animals across different species, tissues and developmental stages, with an initial focus on farm and companion animals [12, 13]. To achieve high quality functional annotation of these genomes, FAANG carries out activities to standardize core biochemical assay protocols, coordinate and facilitate data sharing. It also establishes an infrastructure for the analysis of genotype-to-phenotype data (`www.faang.org`, `www.faang-europe.org`). Substantial efforts are being dedicated to farm animal species, as deciphering the genotype-to-phenotype relationships underlying complex traits such as production efficiency and disease resistance is a prerequisite for exploiting the full potential of livestock ([12]).

Here we report the main results of a pilot project (FR-AgENCODE) launched at the beginning of the FAANG initiative. The broad aim was to generate standardized FAANG reference datasets from four livestock species (cattle, goat, chicken and pig) through the adaptation and optimization of molecular assays and analysis pipelines. We first collected a panel of samples from more than 40 tissues from two males and two females of four species: *Bos taurus* (cattle, Holstein breed), *Capra hircus* (goat, Alpine breed), *Gallus gallus* (chicken, White Leghorn breed) and *Sus scrofa* (pig, Large White breed), generating a total of 4,115 corresponding entries registered at the EMBL-EBI BioSamples database (see Methods). For molecular characterization, three tissues were chosen to represent a "hub" organ (liver) and two broad immune cell populations (CD4+ and CD8+ T cells). This allowed obtaining a partial representation of energy metabolisms and immunity functions, as well as the optimization of the protocols for experimental assays for both tissue-dissociated and primary sorted cells. In addition to the transcriptome, we analyzed chromatin accessibility by the assay for transposase-accessible chromatin using sequencing (ATAC-seq, [14]), and we characterized the three-dimensional (3D) genome architecture by coupling proximity-based ligation with massively parallel sequencing (Hi-C, [15]) (Figure 1). Using this combination of tissues/assays, we assessed the expression of a large set of coding and non-coding transcripts in the four species, evaluated their patterns of differential expression in light of chromatin accessibility at promoter regions, and characterized active and inactive topological domains of these genomes.

The integrative analysis showed a global consistency of all data, emphasizing the value of a coordinated action to improve the genomic annotation of livestock species.

Results and discussion

High-depth RNA-seq assays provide gene expression profiles in liver and immune cells from cattle, goat, chicken and pig

For each animal (two males, two females) of the four species, we used RNA-seq to profile the transcriptome of liver, CD4+ and CD8+ cells (see Methods, Figure 1 and Table S1). Briefly, stranded libraries were prepared from polyA+ selected RNAs longer than 200bp, and sequencing was performed on an Illumina HiSeq3000 (see Methods). Between 250M (chicken) and 515M (goat) read pairs were obtained per tissue, of which 94% (chicken) to 98% (pig) mapped on their respective genome using STAR (see Methods and Tables S2-S4). As an initial quality control step, we processed the mapped reads with RSEM to estimate the expression levels of all genes and transcripts from the Ensembl reference annotation (hereafter called "reference" genes/transcripts/annotation) (Figure S1, Supplementary data file 1). As expected, a large majority of the reads (from 62% in cattle to 72% in goat) fell in annotated exons of the reference genes (Figure S2). In spite of the specialized scope of this initial study limited to liver and immune cells, a large share of all reference genes were detected (from 58% in chicken to 65% in goat) with an expression level higher than 0.1 TPM for the same transcript isoform in at least two samples (see Methods and Table 1).

For each species, we explored the similarity among samples using the expression profiles of the reference genes. Principal component analysis (PCA) revealed quite consistent patterns across species, where the first principal component (explaining 84 to 91% of the variability among samples) clearly separated samples according to their tissue of origin (liver vs. T cells). A more moderate yet systematic separation was observed between CD4+ and CD8+ T cells on the second principal component (Figure S3). The consistency of these patterns across species supports the reliability of our RNA-seq data.

As expected, the gene expression ratio in males versus females was globally uniform genome-wide, with the exception of the sex chromosomes (Figure 2). Dosage compensation (leading for instance to X chromosome inactivation in mammals) can indeed be observed in all species but chicken, which is consistent with previous reports on dosage compensation in mammals but not in avian species [16] (Figure 2).

To get insight into the pattern of gene expression across species, we hierarchically clustered our samples using the expression of 9,461 genes found to be orthologous across the four species (see Methods and Supplementary data file 2). We observed that, regardless of the species, all liver samples clustered together while immune cell samples clustered first by species and then by type (*i.e.* CD4+ separately from CD8+) (Figure 3). For most species and tissues, samples also clustered by sex (Figure 3). This observation confirms the higher conservation of liver compared to blood gene expression across vertebrates, and highlights the fact that when RNA-seq is performed on multiple tissues and species, samples could primarily cluster by either [17].

Most of the known reference genes are differentially expressed between liver and T cells

To provide functional evidence supporting our RNA-seq data, we performed a differential gene expression analysis across tissues per species. The number of exonic reads was first counted for each gene of the reference annotation and was then TMM-normalized ([18], see Methods). Taking into account the specificities of our experimental design, in which samples from different tissues come from the same animal, we fitted generalized linear models (GLM) to identify genes with differential expression in:

1 All pairwise comparisons between liver, CD4+ and CD8+
2 Liver vs. T cells globally
3 Males vs. females in liver and in T cells.

As expected, liver and T cells were characterized by the largest differences in their respective RNA composition. In all species, the comparison between both tissue types yielded the majority of differentially expressed genes identified (from 7,000 genes in chicken to 10,500 genes in goat), while fewer differentially expressed genes were detected between CD4+ and CD8+ samples (Table S5 and Supplementary data file 3). In addition, Gene Ontology (GO) analysis provided results in line with the role of liver in metabolism and of T cells in immunity for all species (Figures S4-7 and Methods).

Due in part to the limited number of male and female replicates, fewer genes could be revealed by the sex-specific differential analysis (Table S5). Among them, we identified 12 orthologous genes that showed consistent overexpression either in males or females across all the mammalian species, with various degrees of agreement considering the literature. We could indeed find previous reports of sex-specific effects or functions for 6 of these 12 orthologous genes (*PLA2G7*, *L1CAM*, *PGLYRP2*, *THBS2*, *MYBL1* and *IGF2BP3*) but not for five others (*PATJ*, *CRYAB*, *NREP*, *PSAT1* and *MCM2*), [19–24]. For instance, the *MYBL1* transcription factor, which is consistently overexpressed in female T-cells in our data, is described in mammals as a master regulator of meiosis in males [25]. Interestingly, the last of the 12 genes, *CUX2*, which was overexpressed in liver vs. T cells, was consistently overexpressed in males, as seen in human but not in mouse, where overexpression was observed in females [26, 27].

As an additional control, we performed hierarchical clustering of RNA-seq samples by focusing on a manually curated list of T cell- or liver-related transcription factors (TF) (Supplementary data file 4). As expected, this also led to a clear separation between liver and T cells in all cases (Figures S8-S9).

Few immune genes differentiate CD4+ and CD8+ cell populations

In line with the results of the hierarchical clustering (Figure 3), most orthologous genes found to be differentially expressed in CD4+ versus CD8+ within species showed high variability of expression levels across species (not shown). This variability is likely caused by the natural heterogeneity in the relative proportions of cell subtypes among the different species, as already reported between mammals [28, 29]. Nevertheless, 39 orthologous genes could consistently differentiate CD4+ and CD8+ in the four species, including both mammals and chicken. Among those,

10 and 29 genes showed significant overexpression in CD4+ and in CD8+ cells respectively (Table S6). We searched for these genes in the baseline expression dataset of human CD4+ and CD8+ $\alpha\beta$ T cell subsets generated by the Blueprint Epigenome Project [30, 31] and considered their relative enrichment in either cell subset (Table S6). With one exception (*ACVRL1*), all genes were found to be expressed in human CD4+ and/or CD8+ $\alpha\beta$ T cells and 25 of them show relative enrichment in CD4+ (or CD8+) human cells, which is consistent with our data across the four species. Out of these 25 genes, six and eight genes could be respectively associated with CD4+ and CD8+ T cell differentiation, activation and function (Table S6).

The limited number of genes composing the common signature of CD4+ and CD8+ T cell functions is primarily caused by the limited resolution of the sorting strategies (see Methods), leading to heterogeneity in the relative proportions of cell subtypes among the different species. As an example, the sorting of porcine CD8α+ cells included TCR-$\gamma\delta$ T cells [28, 32], an important subpopulation which is known to be present at higher frequency in pigs and ruminants (15-60%) compared to mice and humans (<5%) [29]. Despite these limits, and considering that only partial knowledge is available across the farm species used in this study, our results are in agreement with studies in human and model species: despite global transcription profiles that are conserved between corresponding immune cell lineages between human and mouse, several hundred genes show divergent expression [33].

Analysis of new transcripts improves and extends gene structure annotation

In order to test if our data could improve the reference gene annotation for each species, we used STAR and Cufflinks to identify all transcripts present in our samples and predict their exon-intron structures. We then quantified their expression in each sample using STAR and RSEM (Methods, Figure S1) and only retained the transcripts and corresponding genes expressed in at least two samples with TPM $\geq$ 0.1. We identified between 58,000 and 85,000 transcripts depending on the species (Table 1, Supplementary data file 5), hereafter called "FR-AgENCODE transcripts".

To characterize these FR-AgENCODE transcripts with respect to the reference ones, we grouped them into four positional classes (see Methods): (1) *known*: a FR-AgENCODE transcript whose exon-intron structure is strictly identical to a reference transcript (*i.e.* exact same exons and introns); (2) *extension*: same as (1), but the FR-AgENCODE transcript extends a reference transcript by at least 1 bp on at least one side; (3) *alternative*: a FR-AgENCODE transcript that shares at least one intron with a reference transcript but does not belong to the previous categories (only spliced transcripts can be in this class); and (4) *novel*: a FR-AgENCODE transcript that does not belong to any of the above categories. We found that the majority of the FR-AgENCODE transcripts were of the alternative class, therefore enriching the reference annotation with new splice variants of known genes. We note that the proportion of completely novel transcripts is relatively high for cattle, which may be due to the lack of RNA-seq data in the Ensembl gene annotation pipeline of the UMD3.1 version, leading to a less complete reference annotation (Figure S10A-C, Table 1 and Table S7).

In order to further validate the FR-AgENCODE gene annotation and identify interesting new transcripts involved in immunity and metabolism, we performed a

gene expression analysis similar to the one performed on reference genes (see above and Methods). PCAs were highly similar to those previously found, with a clear separation between liver and T cell samples on the first component and a weaker separation on the second component between the two types of T cells (Figure S11). Using the same approach as that used for the reference genes, we performed a differential expression (DE) analysis on the FR-AgENCODE genes (see above and Methods). As expected, the number of DE genes between T cells and liver was again higher than between CD4+ and CD8+ cells (Table S8 and Supplementary data file 6). The proportion of DE genes was smaller than for the reference genes (Table S5), which is expected given their globally lower gene expression levels. We also found a high level of consistency between the sets of DE genes found for the reference versus the FR-AgENCODE genes (more than 88% of the DE reference genes correspond to a DE FR-AgENCODE gene). Among the DE genes of the FR-AgENCODE set, between 202 (chicken) and 1,032 (goat) of them have at least one transcript that is classified as coding mRNA and does not overlap any reference gene on the same strand. This highlights the potential to identify novel interesting candidates for further functional characterization.

Identification, classification and comparative analysis of lncRNAs

Using the FEELnc method [34] trained with mRNAs from the reference annotation as the coding set and shuffled mRNA sequences as the non-coding training set (Methods), we identified 22,724, 13,864, 7,502 and 12,587 putative lncRNA transcripts for cattle, goat, chicken and pig, respectively (Tables S9-10 and Supplementary data file 7), many more than compared to the reference (Ensembl and NCBI) annotations (0, 4,483, 4,643 and 875 lncRNAs respectively). The smaller number of lncRNAs found in chicken was due to the larger number of non-assembled small scaffolds in that genome, and can be extended to include the 2,718 additional unclassified lncRNAs located on them (217, 707 and 83 for the other species). Consistent with previous reports in several species including human [35], dog [34], and chicken [36], the predicted lncRNA genes were much less expressed than the reference protein-coding genes (Figure S12). The structural features of these predicted lncRNA transcripts were consistent between the four species: lncRNAs are spliced but with fewer exons (1.5 vs. 10) and higher median exon length (660 vs. 130bp) compared to mRNAs (Figure S12). lncRNAs are also smaller than mRNAs (1800 vs. 3600bp). Notably, the lower number of exons and consequent smaller size of lncRNAs compared to mRNAs could also be due to the weaker expression of lncRNAs, which makes it more difficult for computational methods to properly identify their structure [37].

In addition to the coding/non-coding classification, FEELnc can also categorize lncRNAs as intergenic or intragenic based on their genomic positions with respect to a provided set of reference genes (usually protein coding genes), and considering their transcription orientation with respect to these reference genes. Table S10 shows this distribution for all species. Similarities were observed between species with, on average, 6% to 12% intragenic lncRNA genes and more than 88% intergenic lncRNA genes, as observed in human [35] and chicken [36], respectively.

We and others previously showed a sharp decrease in sequence conservation of lncRNAs when the phylogenetic distance increases [35, 36, 38], in particular between chicken and human that diverged 300M years ago. We therefore analyzed lncRNA conservation between the four livestock species using a synteny approach based on the orthology of protein-coding genes surrounding the lncRNA and not on the lncRNA sequence conservation itself ([36], see Methods). We found 73 such conserved lncRNAs across cattle, goat and pig, 19 across cattle, chicken and pig, and 6 across all four species (Supplementary data file 7). All were expressed in these species and located in the same orientation with respect to the flanking orthologous genes. An example of such a conserved lncRNA, hitherto unknown in our four species, is provided in Figure 4. In human, this lncRNA was named *CREMos* for "*CREM* opposite sense" since it is in divergent position with respect to the neighboring *CREM* protein-coding gene. The synteny is conserved across species from fishes to mammals. Interestingly, the *CREMos* lncRNA is overexpressed in T cells while the *CREM* protein-coding gene is overexpressed in liver in goat, cattle and chicken (Figure 4).

Finally, we performed a comparative transcriptomic approach, using the 73 and 19 lncRNAs conserved between three species "cattle-goat-pig" and "cattle-chicken-pig" respectively. The clustering obtained with these two gene sets shows a clear tissue separation between liver and T cells but does not separate species according to their evolutionary distance, as was observed with the 9,461 orthologous protein-coding genes from the four species (Figure S13). Such results, in line with studies using multiple tissues from 11 tetrapod species [39], probably reflects a faster evolution of lncRNA sequences with less constraints than protein coding sequences and a specific pattern of expression.

RNA-seq mapped reads can confirm transcript 3' end positions

In order to confirm the 3' end positions of the annotated transcripts, we looked for non genomic stretches of As in RNA-seq mapped reads that could originate from unmapped polyA tails (see Methods). After filtering for internal priming and other artefacts, we obtained from 150,000 (chicken) to 250,000 (pig) distinct positions of potential polyA sites (Table S11 and Supplementary data file 8). Neighboring sites in the same orientation were further merged into 58,000 (chicken) to 108,000 (pig) polyA clusters, among which 81% contain a single polyA position. Despite their weak coverage that involved less than 0.02% of the genomic space in all species, these clusters tended to fall close to transcription termination sites (TTS). Approximately 11-14% of the reference and FR-AgENCODE transcripts had at least one polyA cluster near their TTS on the same strand (Table S11). The distribution of polyA clusters around reference and novel genes showed a clear peak at the 3' end of genes in all species (Figure 5 and Figure S14), indicating that predicted polyAs might help support TTS positions. A quantitative analysis correlating the number of polyA-supporting reads 1Kb around the TTS of reference genes and the total number of RNA-seq reads in their exons showed only a moderate correlation (around 0.5, data not shown). Notably, most of the largest polyA clusters were found in the mitochondrial chromosome, where the cluster density was several orders of magnitude higher than genome-wide (from 11 to 30/Kb vs. 0.04 to

0.06/Kb). This suggests a high abundance of internally polyadenylated truncated mRNAs that may represent prematurely terminated transcripts or stalled degradation intermediates, as previously suggested from observations in human [40, 41]. Mitochondrial polyadenylation thus appears to be prevalent across a wide range of vertebrates, from mammals to birds.

ATAC-seq peaks allow the characterization of accessible regions of chromatin in the four species

We used ATAC-seq to profile the accessible chromatin of liver and of CD4+ and CD8+ immune cells in animals from the four species. Between 480M (chicken) and 950M (pig) ATAC-seq fragments were sequenced per species, and were processed by a standard pipeline: reads were trimmed, mapped to the genome (requiring a MAPQ higher than 10) and filtered for mitochondrial and PCR duplicate reads (see Methods). Most reads were filtered out by the MAPQ 10 and PCR duplicate filters (Figure S15). Peaks were then called in each tissue separately (see Methods), resulting in between 26,000 (pig, liver) and 111,000 (pig, cd8) peaks per tissue (Table S12). Those peaks were further merged into a unique set of peaks per species, resulting in between 75,000 (goat) and 149,000 (pig) peaks (Table S12 and Supplementary data file 9), and covering between 1 and 5% of the genome. The average peak size was around 600 bp for all species, except for chicken where it was less than 500 bp. Merging tissue peaks did not result in much wider peaks (Figure S16).

Similarly to GWAS variants associated to human diseases [3], the vast majority of our ATAC-seq peaks were either intronic or intergenic (Figure S17, Methods and Supplementary data file 9). In particular, from 38% (goat) to 55% (cattle) of the peaks lie at least 5Kb away from any reference gene (Figure S17). These peaks potentially represent many new candidate regulatory regions that could be further correlated with genetic variants known from previous studies to be associated with phenotypes. Focusing on the promoter regions of the genes, we noted that around 15% of the ATAC-seq peaks lie at 5Kb or less from a reference TSS (Transcription Start Site). More precisely, the ATAC-seq peak distribution within and around reference genes showed a clear signal at the TSS for all species, thereby supporting the quality of both the reference gene annotation, and of our ATAC-seq data and processing (Figure 5). Repeating the distribution analysis of the ATAC-seq peaks within and around the FR-AgENCODE transcripts that were not in the reference annotation (i.e. not from the known class; Figure S14E-H) resulted in a similar enhanced 5' signal (Figure S14A-D). This supports the quality of the structural annotations we propose, and the fact that ATAC-seq peaks can be used to support TSS prediction.

The ATAC-seq peaks of each species were quantified in each sample and resulting read counts were normalized using a loess correction (see Methods). Differential analyses similar to those used for RNA-seq were then performed on normalized counts (see Methods). The number of Differentially Accessible (DA) peaks between T cells and liver per species was between 4,800 (goat) and 13,600 (chicken), and, with the exception of pig, there were roughly as many more accessible peaks in T cells than in liver (Table S13 and Supplementary data file 10).

As accessible regions of the chromatin can contain regulatory sites for DNA binding proteins, we computed the density of transcription factor binding sites (TFBS)

in ATAC-seq peaks genome-wide. Interestingly, the TFBS density was significantly higher in ATAC-seq peaks with a differential accessibility (p-value $<$ 7.10e-4 for goat and $<$ 2.2e-16 for chicken and pig, Wilcoxon test, see Methods). This was also observed for distal ATAC-seq peaks, at least 5kb away from promoters (not shown). This suggests that differentially accessible peaks are more likely to have a regulatory role than the others.

Integrative analysis of RNA-seq and ATAC-seq data suggests complex regulatory mechanisms of gene expression

We then investigated the correlation between chromatin accessibility and gene expression by comparing RNA-seq and ATAC-seq data. Given the specific distribution of the ATAC-seq signal, we initially focused on the gene promoter regions and considered proximal chromatin peaks (at a maximum distance of 1Kb from the TSS) to assign a promoter accessibility value to each gene.

Within each sample, genes with highly accessible promoters showed higher expression values globally (Figure S18), as already reported in mouse and human for instance [42]. In pig and goat, the number of available samples allowed us to compute for each gene the correlation between promoter accessibility and gene expression across all samples (Figure S19 and Methods). Interestingly, while the distribution of the correlation values appeared to be unimodal for non-differentially expressed genes, a bimodal distribution was obtained for differentially expressed genes, with an accumulation of both positive and negative correlation values (Figure 6, Figures S19-20). This pattern suggests the existence of different types of molecular mechanisms involved in gene expression regulation.

Hi-C reveals topological aspects of the genome in the nucleus

In order to profile the structural organization of the genome in the nucleus, we performed *in situ* Hi-C on liver cells from the two male and the two female samples of pig, goat and chicken. The *in situ* Hi-C protocol was applied as previously described [43] with slight modifications (see FAANG protocols online and Methods). Reads were processed using a bioinformatics pipeline based on HiC-Pro [44] (Methods). From 83 to 91% of the reads could be genomically mapped depending on the sample, and after filtering out all inconsistent mapping configurations we obtained a total of 185, 266 and 260M valid read pairs in goat, chicken and pig respectively (Table S14 and Figure S21). These sequencing depths allowed us to build interaction matrices (or Hi-C contact heatmaps) at the resolution of 40 and 500Kb in order to detect Topologically Associating Domains (TADs) and A/B compartments (Figure S23).

We identified from $\approx$ 5,400 (chicken) to 11,000 (pig) TADs of variable sizes (150 to 220Kb on average, Table S15 and Supplementary data file 11), with a 79-86% genome-wide coverage. In order to validate the relevance of the predicted domains, obtained by maximizing the interaction densities across different resolutions [45] (see Methods), we used two approaches. First, we used the original TAD-finding metric called the Directionality Index (DI) to quantify the degree of upstream or downstream interaction bias for any genomic region [46]. As expected, the average DI along TADs showed a bias for downstream interactions at the beginning of the domains and for upstream interactions at the ends of the domains for all species

(Figure 7). Secondly, since TADs are known to be flanked by the CTCF insulating factor, at least in model organisms [46], we investigated if this could be detected by a prevalence of CTCF binding sites at the boundaries of our TADs. We observed that the density of *in silico* predicted CTCF sites (see Methods) consistently peaked at TAD boundaries (Figure 7), in agreement with previous reports on experimental data in human [43, 46]. Altogether, these findings show a global consistency of our topological annotation with previous results and support the generality of the 3D structural organization across animals, ranging from mammals to birds.

Considering a higher-order topological organization of the chromosomes in the nucleus, we then investigated the partition of the genome into "A" and "B" compartments [15]. By applying the original method described in [15] as described in Figure S23 (see also Methods) we divided the genomes into "active" (A) and "inactive" (B) compartments and obtained from $\approx$ 580 to 700 compartments per genome with a mean size between 1.6Mb (chicken) and 3.4Mb (goat) (see Table S15 and Supplementary data file 11). To assess the robustness of the A/B compartment calling, we performed the same analysis on the interaction matrix of each replicate separately and compared the four resulting partitions within each species. About 80% of the informative loci (genomic bins of the matrix) were assigned to the same compartment in each animal of the three species (Figure S23), emphasizing the stability of the detected signal in liver cells.

Considering functional aspects of genome topology, it has been proposed that A compartments represent genomic regions enriched for open chromatin and transcription compared to B compartments [46]. A higher gene expression and chromatin accessibility should therefore be expected in A versus B compartments. In all species for which we have Hi-C data, such a pattern was indeed observed, as both the average RNA-seq gene expression values and ATAC-seq chromatin accessibility values in liver samples were significantly higher in A than in B compartments (Figure 8, p-value $<$ 2.2e-16 for each comparison, Wilcoxon test). This result highlights the high consistency and coherence among the three types of data we have generated and validates the general relevance of the annotation we propose.

Conclusion

We report the first multi-species and multi-assay genome annotation results obtained by a pilot FAANG project. The main outcomes were consistent with our expectations:

- Despite that only three tissue samples were used, a majority of the reference transcripts could be detected. Moreover, the new identified "FR-AgENCODE transcripts" considerably enrich the reference annotations: the transcript repertoire is augmented by about 50% in all species but cattle, where it is more than tripled.
- Differential analyses of gene expression in liver and T cells yielded results consistent with known metabolism and immunity functions.
- As expected, ATAC-seq results revealed an abundance of potential regulatory regions. When integrated with RNA-seq data, these results suggest complex regulatory mechanisms of gene expression.

- Hi-C experiments provided the first set of genome-wide 3D interaction maps of the same tissue from three livestock species. Beyond the chromosome topology annotation, the analysis shows high consistency with the gene expression and chromatin accessibility results.

As such, the FR-AgENCODE group has delivered a strong proof of concept of a successful collaborative approach at a national scale to apply FAANG guidelines to various experimental procedures and animal models. This notably includes the set up of a combination of sequencing assays on primary cells and tissue-dissociated cells, as well as a large collection of documented tissue samples available for further characterization. It also confirmed, in line with several studies in model species [4–6, 8] the value of combining molecular characterizations on the same samples to simultaneously identify the transcriptional output of the cell and investigate the underlying regulatory mechanisms.

In the context of the global domesticated animal genome annotation effort, lessons learned from this pilot project confirmed conclusions drawn by the FAANG community regarding the challenges to be addressed in the future [12]. As all possible combinations of experimental assays, tissues, developmental stages and phenotypes cannot be undertaken by a single project on all species, standardized procedures for both data production and analysis need to be widely adopted [47]. Furthermore, the mosaic nature of a global annotation effort that gathers contributions from various partners worldwide emphasizes the challenge of translating recent advances from the field of data science into efficient methods for the integrative analysis of 'omics data and the importance of future metanalysis of several datasets. Altogether, these annotation results will be useful for future studies aiming to determine which subsets of putative regulatory elements are conserved, or diverge, across animal genomes representing different phylogenetic taxa. This will be beneficial for devising efficient annotation strategies for the genomes of emerging domesticated species.

Methods

Animals, sampling and tissue collections

Animals and breeds

Well-characterized breeds were chosen in order to obtain well-documented samples. Holstein is the most widely used breed for dairy cattle. For goat, the Alpine breed is one of the two most commonly used dairy breeds, and for pigs, the Large white breed is widely used as a dam line. For chickens, the White Leghorn breed was chosen as it provides the genetic basis for numerous experimental lines and is widely used for egg production.

Four animals were sampled for each species, two males and two females. They all had a known pedigree. Animals were sampled at an adult stage, so that they were sexually mature and had performance records, obtained in known environmental conditions. Females were either lactating or laying eggs.

All animals were fasted at least 12 hours before slaughter. No chemicals were injected before slaughtering, animals were stunned and bled in a licensed slaughter facility at the INRA research center in Nouzilly.

Samples

Liver samples of 0.5 cm^3 were taken from the edge of the organ, avoiding proximity with the gallbladder and avoiding blood vessels. Time from slaughter to sampling varied from 5 minutes for chickens to 30 minutes for goats and pigs and 45 minutes for cattle. For the purpose of RNA-seq, samples were immediately snapfrozen in liquid nitrogen, stored in 2ml cryotubes and temporarily kept in dry ice until final storage at -80°C.

For mammals, whole blood was sampled into EDTA tubes before slaughter; at least one sampling took place well before slaughter (at least one month) and another just before slaughter, in order to obtain at least 50 ml of whole blood for separation of lymphocytes (PBMC). PBMC were re-suspended in a medium containing 10% FCS, counted, conditioned with 10% DMSO and stored in liquid nitrogen prior to the sorting of specific cell types: CD3+CD4+ ("CD4") and CD3+CD8+ ("CD8").

For chicken, a specific procedure was implemented to separate spleen cells from red blood cells, in order to get a population of immune cells without any contamination with platelets, prior to CD4+ and CD8+ sorting.

All protocols for liver sampling, PBMC separation, splenocyte purification and T cell sorting can be found at ftp://ftp.faang.ebi.ac.uk/ftp/protocols/samples/

Experimental assays and protocols

All assays were performed according to FAANG guidelines and recommendations, available at http://www.faang.org. All detailed protocols used for RNA extraction and libraries production for RNA-seq, ATAC-seq and Hi-C are available at http://ftp.faang.ebi.ac.uk/ftp/protocols/assays/.

RNA extraction

Cells and tissues were homogenized in TRIzol reagent (Thermo) using an ULTRA-TURRAX (IKA-Werke) and total RNAs were extracted from the aqueous phase. They were then treated with TURBO DNase (Ambion) to remove remaining genomic DNA and then process to separate long and small RNAs using the mirVana miRNA Isolation kit. Small and long RNAs quality was assessed using an Agilent 2100 Bioanalyzer and RNA 6000 nano kits (Agilent) and quantified on a Nanodrop spectrophotometer.

RNA-seq

Stranded mRNA libraries were prepared using the TruSeq Stranded mRNA Sample Prep Kit -V2 (Illumina) on 200 ng to 1μg of total long RNA with a RNA Integrity Number (RIN) over 8 following the manufacturer's instructions. Libraries were PCR amplified for 11 cycles and libraries quality was assessed using the High Sensitivity NGS Fragment Analysis Kit DNF-474 and the Fragment Analyser system (AATI). Libraries were loaded onto a High-seq 3000 (Illumina) to reach a minimum read numbers of 100M paired reads for each library.

Hi-C

In situ Hi-C libraries were performed according to [43] with a few modifications. For all species, fresh liver biopsies were dissociated using Accutase, and each resulting cell suspension was filtered using a 70 μm cell strainer. Cells were then fixed with 1% formaldehyde for 10 minutes at 37°C and fixation was stopped by adding Glycine to a final concentration of 0.125M. After two washes with PBS, cells were pelleted and kept at -80°C for long term storage. Subsequently, cells were thawed on ice and 5 million cells were processed for each Hi-C library. Cell membranes were disrupted using a potter-Elvehjem PTFE pestle and nuclei were then permeabilized using 0.5% SDS with digestion overnight with HindIII endonuclease. HindIII-cut restriction sites were then end-filled in the presence of biotin-dCTP using the Klenow large fragment and were religated overnight at 4°C. Nuclei integrity was checked using DAPI labelling and fluorescence microscopy. Nuclei were then lysed and DNA was precipitated and purified using Agencourt AMPure XP beads (Beckman Coulter) and quantified using the Qubit fluorimetric quantification system (Thermo). Hi-C efficiency was controlled by PCR using specific primers for each species and, if this step was successful, DNA was used for library production. DNA was first treated with T4 DNA polymerase to remove unligated biotinylated ends and shared by sonication using a M220 Covaris ultra-sonicator with the DNA 550pb SnapCap microtube program (Program length: 45s; Picpower 50; DutyF 20; Cycle 200; Temperature 20°C).

Sheared DNA was then sized using magnetic beads, and biotinylated fragments were purified using M280 Streptavidin Dynabeads (Thermo) and reagents from the Nextera_Mate_Pair Sample preparation kit (Illumina). Purified biotinylated DNA was then processed using the TrueSeq nano DNA kit (Illumina) following the manufacturer's instructions. Libraries were amplified for 10 cycles and then purified using Agencourt AMPure XP beads. Library quality was assessed on a Fragment Analyser (AATI) and by endonuclease digestion using NheI endonuclease. Once validated, each library was sequenced on an Illumina Hi-Seq 3000 to reach a minimum number of 150M paired reads per library.

ATAC-seq

ATAC-seq libraries were prepared according to Buenrostro et al. (2013) with a few modifications. For liver, cells were dissociated from the fresh tissue to get a single cell suspension. Cells were counted and 50,000 cells were processed for each assay. Transposition reaction was performed using the Tn5 Transposase and TD reaction buffer from the Nextera DNA library preparation kit (Illumina) for 30 minutes at 37°C. DNA was then purified using the Qiagen MinElute PCR purification kit. Libraries were first amplified for 5 cycles using custom-synthesized index primers (see supplementary methods) and then a second amplification was performed. The appropriate number of additional PCR cycles was determined using real-time PCR, permitting the cessation of amplification prior to saturation. The additional number of cycles needed was determined by plotting the Rn versus Cycle and then selecting the cycle number corresponding to one-third of the maximum fluorescent intensity. After PCR amplification, libraries were purified using a Qiagen MinElute PCR purification kit followed by an additional clean-up and sizing step using AMPure

XP beads (160 μl of bead stock solution was added to 100 μl of DNA in EB buffer) following the manufacturer's instructions. Library quality was assessed on a BioAnalyser (Agilent) using Agilent High Sensitivity DNA kit (Agilent), and libraries were quantified using a Qubit Fluorometer (Thermo).

Bioinformatics and Data Analysis

All software used in this project along with the corresponding versions are listed in Table S3. The reference gene annotation was obtained from the Ensembl v90 release. Since *Capra hircus* was not part of the Ensembl release, we used the NCBI CHIR_1.0.102 annotation (see Table S2).

RNA-seq pipeline

Prior to any processing, all RNA-seq reads were trimmed using cutadapt version 1.8.3. Reads were then mapped twice using STAR v2.5.1.b [48, 49]: first on the genome indexed with the reference gene annotation to quantify expression of reference transcripts, and secondly on the same genome indexed with the newly generated gene annotation (FR-AgENCODE transcripts) (see below and Figure S1) [50]. The STAR `--quantMode TranscriptomeSAM` option was used in both cases in order to additionally generate a transcriptome alignment (bam) file. After read mapping and CIGAR-based softclip removal, each sample alignment file (bam file) was processed with Cufflinks 2.2.1 [51, 52] with the `max-intron-length` (`100000`) and `overlap-radius` (`5`) options, guided by the reference gene annotation (`--GTF-guide` option) ([50], Figure S1). All cufflinks models were then merged into a single gene annotation using Cuffmerge 2.2.1 [51, 52] with the `--ref-gtf` option. The transcript and gene expressions on both reference and newly generated gene annotation were quantified as TPM (transcripts per million) using RSEM 1.3.0 [53] on the corresponding transcriptome alignment files ([50], Figure S1). The newly generated transcripts were then processed with FEELnc version 0.1.0 [34] in order to classify them into "lncRNA", "mRNA" and "otherRNA" (Figure S1, Tables S9-10, Figure S10). The newly generated transcripts with a TPM value of at least 0.1 in at least 2 samples were called FR-AgENCODE transcripts and kept as part of the new annotation. The 0.1 threshold was chosen knowing that the expression values of polyadenylated transcripts usually go from 0.01 to 10,000 [54] and that we wanted to simultaneously capture long non coding RNAs that are generally lowly expressed and reduce the risk of calling artefactual transcripts.

PCA based on gene expression

Principal Component Analysis (PCA) was performed using the **mixOmics** R package [55] on the RNA-seq sample quantifications of each species. This was done using the expression (TPM) of two different sets of genes: reference genes with TPM 0.1 in at least two samples (Figure S3) and FR-AgENCODE genes with TPM 0.1 in at least two samples (Figure S15).

Annotated gene orthologs

We used Ensembl Biomart [56] to define the set of orthologous genes across cattle, chicken and pig. Only "1 to 1" orthology relationships were kept (11,001 genes,

Supplementary data file 1). Since goat was not part of the Ensembl annotation, goat gene IDs were added to this list using gene name as a correspondence term. The resulting 4-species orthologous set contained 9,461 genes (Supplementary data file 1).

RNA-seq sample hierarchical clustering

Based on the expression of the 9,461 orthologous genes in the 39 RNA-seq samples from the four species, the sample-by-sample correlation matrix was computed using the Pearson correlation of the $\log_{10}$ gene TPM values. We then represented this sample by sample correlation matrix as a heatmap where the samples were also clustered using a complete linkage hierarchical clustering (Figure 3).

RNA-seq normalization and differential analysis

To perform the differential analysis of gene expression, raw read counts were first measured by counting the number of RNA-seq reads overlapping the exonic regions of each gene in each library using bedtools [57]. RNA-seq library size normalization factors were calculated using the weighted Trimmed Mean of M-values (TMM) approach of [58] as implemented in the R/Bioconductor package **edgeR** [18]. The same package was used to fit three different per-gene negative binomial (NB) generalized log-linear models [59].

- In **Model 1**, the expression of each gene was explained by a tissue effect; because all three tissues (liver, CD4, CD8) were collected from each animal, an animal effect was also included to account for these repeated measures:

  $$\frac{\log \mu_{gi}}{s_i} = \beta_{g,\text{tissue}(i)} + \gamma_{g,\text{animal}(i)},$$

  where μ_{gi} represents the mean expression of gene g in sample i, s_i the TMM normalization factor for sample i, $\text{tissue}(i) \in \{\text{liver, CD4, CD8}\}$ and $\text{animal}(i) \in \{1, 2, 3, 4\}$ the tissue and animal corresponding to sample i, and $\beta_{g,\text{tissue}(i)}$ and $\gamma_{g,\text{animal}(i)}$ the fixed tissue and animal effects, respectively, of gene g in sample i. Hypothesis tests were performed to identify significantly differentially expressed genes among each pair of tissues, *e.g.*

  $$\mathcal{H}_{0g} : \beta_{g,\text{liver}} = \beta_{g,\text{CD4}}.$$

- **Model 2** is identical to the previous model, where gene expression was modeled using both a tissue and an animal effect, with the exception that the CD4 and CD8 tissues were collapsed into a single group. In this model, the only hypothesis of interest is thus between the liver and global CD cell group:

  $$\mathcal{H}_{0g} : \beta_{g,\text{liver}} = \beta_{g,\text{CD}}.$$

- In **Model 3**, the expression of each gene was explained using a nested factorial model, which included a sex effect as well as tissue and animal effects nested within sex:

  $$\frac{\log \mu_{gi}}{s_i} = \beta_{g,\text{sex}(i)} + \gamma_{g,\text{sex}(i):\text{tissue}(i)} + \delta_{g,\text{sex}(i):\text{animal}(i)},$$

where μ_{gi} and s_i are as in Model 1, $\text{sex}(i) \in \{\text{M, F}\}$, $\text{tissue}(i) \in \{\text{liver, CD4, CD8}\}$, and $\text{animal}(i) \in \{1, 2\}$, respectively represent the sex, tissue, and animal nested within sex corresponding to sample i, $\beta_{g,\text{sex}(i)}$ is the sex effect for gene g in sample i, and $\gamma_{g,\text{sex}(i):\text{tissue}(i)}$ and $\delta_{g,\text{sex}(i):\text{animal}(i)}$ are the nested interaction effects of tissue and animal with sex. Note that this model enables comparisons between sexes to be made between animals, while comparisons among tissues are performed within animals. Specifically, we focused on two sets of hypothesis tests: 1) pairwise comparisons of tissues, after correction of the sex effect; and 2) comparisons of differences between males and females within each tissue. For the former, in any of the pairwise tissue comparisons or any of the four species, 50% to 85% of reference genes (56% to 90% of the FR-AgENCODE genes) identified as DE in Model 1 or 3, were commonly identified as DE by both Models 1 and 3.

All hypothesis tests were performed using likelihood-ratio tests and were corrected for multiple testing with the Benjamini-Hochberg (FDR, [60]) procedure. Genes with an FDR smaller than 5% and an absolute log-fold change larger than 1 were declared differentially expressed. Due to the small number of available samples and the fact that Model 3 was only possible for a subset of species due to incomplete designs, we primarily focused on the simpler but more robust Models 1 and 2 by default, and only used Model 3 to investigate the sex effect.

GO analysis of differentially expressed genes

For each species, GO term enrichment analysis was performed on the genes found to be over-expressed in liver versus T cells, and reciprocally. This analysis was done separately for each species (Figures S4-S6) and then on genes present in all species (Figure S7), using the three following ontologies: biological process (BP), molecular function (MF) and cell compartment (CC), and using the **GOstat** R/Bioconductor package [61]). This analysis tests the differential genes against all the genes, and here we always restricted analyses to genes with a human ortholog.

Sample and gene hierarchical clustering based on T cell and metabolism Transcription Factors

Manually curated lists of T cell and metabolism related Transcription Factors (TF) were established (Supplementary data file 3). As a quality control we performed in each species and for both T cell and metabolism related TFs, a double hierarchical clustering of RNA-seq samples and TF genes, using one minus the squared Pearson correlation on the base 10 logarithm of the TPM as a distance and the complete linkage aggregation (Figures S8-9).

FR-AgENCODE transcript positional classification

The FR-AgENCODE transcript models were first classified according to their position with respect to reference transcripts:

known the FR-AgENCODE transcript is strictly identical to a reference transcript (same intron chain and same initial and terminal exons)

extension the FR-AgENCODE transcript extends a reference transcript (same intron chain but one of its two most extreme exons extends the reference transcript by at least one base pair)

alternative the FR-AgENCODE transcript corresponds to a new isoform (or variant) of a reference gene, i.e. the FR-AgENCODE transcript shares at least one intron with a reference transcript but does not belong to the above categories

novel the FR-AgENCODE transcript is in none of the above classes

FR-AgENCODE transcript coding classification

The FR-AgENCODE transcript models were also classified according to their coding potential. For this, the FEELnc program (release v0.1.0) was used to discriminate long non-coding RNAs from protein-coding RNAs. FEELnc includes three consecutive modules: $\text{FEELnc}_{\text{filter}}$, $\text{FEELnc}_{\text{codpot}}$ and $\text{FEELnc}_{\text{classifier}}$. The first module, $\text{FEELnc}_{\text{filter}}$, filters out non-lncRNA transcripts from the assembled models, such as transcripts smaller than 200 nucleotides or those with exons strandedly overlapping exons from the reference annotation. This module was used with default parameters except `-b transcript_biotype=protein_coding, pseudogene` to remove novel transcripts overlapping protein-coding and pseudogene exons from the reference. The $\text{FEELnc}_{\text{codpot}}$ module then calculates a coding potential score (CPS) for the remaining transcripts based on several predictors (such as multi k-mer frequencies and ORF coverage) incorporated into a random forest algorithm [62]. In order to increase the robustness of the final set of novel lncRNAs and mRNAs, the options `--mode=shuffle` and `--spethres=0.98,0.98` were set. Finally, the $\text{FEELnc}_{\text{classifier}}$ classifies the resulting lncRNAs according to their positions and transcriptional orientations with respect to the closest annotated reference transcripts (sense or antisense, genic or intergenic) in a 1Mb window (`--maxwindow=1000000`).

It is worth noting that between 83 and 2,718 lncRNA transcripts were not classified because of their localization on the numerous unassembled contigs in livestock species with no annotated genes.

Syntenic conservation of lncRNAs

Briefly, a lncRNA was considered as "syntenically" conserved between two species if (1) the lncRNA was located between two orthologous protein-coding genes, (2) the lncRNA was the only one in each species between the two protein-coding genes and (3) the relative gene order and orientation of the resulting triplet was identical between species. Using these criteria, we found six triplets shared between the four species, 73 triplets shared between cattle, goat and pig, and 19 triplets shared between cattle, chicken and pig.

Annotation of 3' polyA sites

The identification of potential polyA sites was performed from the already mapped RNA-seq data in three steps: prediction, cleaning and clustering.

Prediction The samToPolyA utility (Table S3) was used on RNA-seq mapped reads to find traces of sequenced yet unmapped A stretches in reads that overlap polyA sites. Taking a bam file as input, samToPolyA finds stretches of As in the terminal part of the reads that had to be trimmed to allow the read to be mapped (soft-clip in the CIGAR string). The bam file of each sample was

processed with the options `--minClipped=20 --minAcontent=0.9 --discardInternallyPrimed`, which outputs one bed record per potential site. For each site the number of supporting reads was computed per tissue and in total.

Cleaning Extensive filtering was done to discard false positive sites that could have originated from internal priming during the library construction or from excessive read trimming during the mapping. The first can be detected by the presence of an A-rich region upstream of the candidate polyA site (12 As or more within the upstream 20bp), the second by a similar region downstream. Any site with such a region upstream or downstream was discarded. Unknown bases "N" were considered as "A"s for this filtering step.

Clustering Consecutive positions on the same strand that were not separated by more than 10bp were then grouped into clusters of polyA sites. PolyA clusters are provided for each species as additional files in BED format, where the positions of the distinct polyA sites within the clusters and the corresponding numbers of supporting reads at each of them are indicated in fields #7 and #8 respectively. The total number of supporting reads for each cluster is reported in the score field #5.

ATAC-seq data analysis pipeline

ATAC-seq reads were trimmed with trimgalore 0.4.0 using the `--stringency 3`, `-q 20`, `--paired` and `--nextera` options (Table S3). The trimmed reads were then mapped to the genome using bowtie 2 2.3.3.1 with the `-X 2000` and `-S` options [63]. The resulting sam file was then converted to a bam file with samtools 1.3.1, and this bam file was sorted and indexed with samtools 1.3.1 [64]. The reads for which the mate was also mapped and with a MAPQ $\geq$ 10 were retained using samtools 1.3.1 (`-F 12` and `-q 10` options, [64]), and finally only the fragments where both reads had a MAPQ $\geq$ 10 and which were on the same chromosome were retained.

Mitochondrial reads were then filtered out, as well as duplicate reads (with picard tools, MarkDuplicates subtool, see Table S14). The peaks were called using MACS2 2.1.1.20160309 [65] in each tissue separately using all the mapped reads from the given tissue (`-t` option) and with the `--nomodel`, `-f BAMPE` and `--keep-dup all` options. To get a single set of peaks per species, the tissue peaks were merged using mergeBed version 2.26.0 [57]. These peaks were also quantified in each sample by simply counting the number of mapped reads overlapping the peak.

ATAC-seq peaks were also classified with respect to the reference gene annotation using these eight genomic domains and allowing a peak to be in several genomic domains:

exonic the ATAC-seq peak overlaps an annotated exon by at least one bp

intronic the ATAC-seq peak is totally included in an annotated intron

tss the ATAC-seq peak includes an annotated TSS

tss1Kb the ATAC-seq peak overlaps an annotated TSS extended 1Kb both 5' and 3'

tss5Kb the ATAC-seq peak overlaps an annotated TSS extended 5Kb both 5' and 3'

tts	the ATAC-seq peak includes an annotated TTS
tts1Kb	the ATAC-seq peak overlaps an annotated TTS extended 1Kb both 5' and 3'
tts5Kb	the ATAC-seq peak overlaps an annotated TTS extended 5Kb both 5' and 3'
intergenic	the ATAC-seq peak does not overlap any gene extended by 5KB on each side

Differential analysis: normalization and model. Contrary to RNA-seq counts, ATAC-seq counts exhibited trended biases visible in log ratio-mean average (MA) plots between pairwise samples after normalization using the TMM approach, suggesting that an alternative normalization strategy was needed. In particular, trended biases are problematic as they can potentially inflate variance estimates or log fold-changes for some peaks. To address this issue, a fast loess approach [66] implemented in the normOffsets function of the R/Bioconductor package **csaw** [67] was used to correct differences in log-counts vs log-average counts observed between pairs of samples.

As for for RNA-seq, we used three different differential models: Model 1 for tissue pair comparisons, Model 2 for T cell versus liver comparisons and Model 3 for tissue pair and sex comparisons taking the sex effect into account (see corresponding RNA-seq section for more details). In any of the pairwise tissue comparisons or any of the four species, 26% to 56% of ATAC-seq peaks identified as differentially accessible (DA) in Model 1 or 3 were commonly identified as DA by both Models 1 and 3.

Although sex-specific effects appeared to be stronger in the ATAC-seq data than in the RNA-seq data, for simplicity and consistency, we have focused our discussions on results from Models 1 and 2 in the manuscript, unless specifically discussing sex effects.

ATAC-seq peak TFBS density. In order to identify Transcription Factor Binding Sites (TFBS) genome-wide, we used the FIMO [68] software (Table S3) to look for genomic occurrences of the 519 TFs catalogued and defined in the Vertebrate 2016 JASPAR database [69] . We then intersected these occurrences with the ATAC-seq peaks of each species and computed the TFBS density in differential vs non differential ATAC-seq peaks (Figure S18). Among the predicted TFBSs, those obtained from the CTCF motif were used to profile the resulting density with respect to Topological Associating Domains from Hi-C data (Figure 7).

Hi-C data analysis pipeline

Our Hi-C analysis pipeline includes HiC-Pro v2.9.0 [70] (Table S3) for the read cleaning, trimming, mapping (this part is internally delegated to Bowtie 2 v2.3.3.1), matrix construction, and matrix balancing ICE normalization [71]. TAD finding was done using Armatus v2.1 [45] (Table S3). Graphical visualization of the matrices was produced with the **HiTC** R/Bioconductor package v1.18.1 [44] (Table S3). Export to JuiceBox [72] was done through the Juicer Tools v0.7.5 (Table S3). These tools were called through a pipeline implemented in Python. Because of the high number of unassembled scaffolds (e.g. for goat) and/or micro-chromosomes (e.g. for chicken)

in our reference genomes, only the longest 25 chromosomes were considered for TAD and A/B compartment calling. For these processes, each chromosome was considered separately.

The Directionality Index (DI) was computed using the original definition introduced by [46] to indicate the upstream vs. downstream interaction bias of each genomic region. Interaction matrices of each chromosome were merged across replicates and the score was computed for each bin of 40Kb. CTCF sites were predicted along the genomes by running FIMO with the JASPAR TFBS catalogue (see section "ATAC-seq peak TFBS density").

A/B compartments were obtained using the method described in [15] as illustrated in Figure S23: first, ICE-normalized counts, K_{ij}, were corrected for a distance effect with:

$$\widehat{K}_{ij} = \frac{K_{ij} - \overline{K}^d}{\sigma^d},$$

in which $\widehat{K}_{ij}$ is the distance-corrected count for the bins i and j, $\overline{K}^d$ is the average count over all pairs of bins at distance $d = d(i,j)$ and σ^d is the standard deviation of the counts over all pairs of bins at distance d. Within-chromosome Pearson correlation matrices were then computed for all pairs of bins based on their distance-corrected counts and a PCA was performed on this matrix. The overall process was performed similarly to the method implemented in the R/Bioconductor package **HiTC** [44]. Boundaries between A and B compartments were identified according to the sign of the first PC (eigenvector). Since PCAs had to be performed on each chromosome separately, the average counts on the diagonal of the normalized matrix were used to identify which PC sign (+/-) should be assigned to A and B compartments for each chromosome. This allowed to automatically obtain an homogeneous assignment across chromosomes. In line with what was originally observed in human, where the first PC was the best criterion for separating A from B compartments (apart from few exceptions like the chromosome 14 for instance [15]), we also observed a good agreement between the plaid patterns of the normalized correlation matrices and the sign of the first PC (Figure S23).

To estimate the robustness of A/B compartment calling, the method was tested on each replicate separately (four animals). Since the HiTC filtering method can discard a few bins in some matrices, resulting in missing A/B labels, the proportion of bins with no conflicting labels across replicates was computed among the bins that had at least two informative replicates (Figure S24).

Integrative analysis

ATAC-seq vs. RNA-seq correlation: intra- and inter-sample analysis. For each ATAC-seq peak that overlapped a promoter region (1Kb upstream of the TSS, as suggested in Figure 5) its loess-normalized read count value (see differential analysis) was associated with the TMM-normalized expression of the corresponding gene from the reference annotation. Intra- and inter-sample correlations were then investigated: within each sample, genes were ranked according to their expressions and the distribution of the corresponding ATAC-seq values was computed for each

quartile (Figure S19). Across samples, the Pearson correlation coefficient was computed for each gene using the ATAC-seq and RNA-seq normalized values from all the available samples with both values (10 for pig, for instance), as illustrated in Figure S20.

Chromatin accessibility and gene expression in A/B compartments. To compute the general chromatin accessibility in A and B compartments, we first computed the average of the normalized read count values across all liver samples for each ATAC-seq peak. For each compartment, the mean value of all contained peaks was then reported and the resulting distributions for all A and B compartments were reported (Figure 8).

The same approach was used to assess the general expression of genes in A and B compartments, using the average of the normalized expression values from the liver samples.

Difference between A and B distributions was tested for statistical significance using a Wilcoxon test.

List of abbreviations

DE	Differentially Expressed
FAANG	Functional Annotation of Animal Genomes
lncRNA	long non-coding RNA
mRNA	messenger RNA
PE	Paired-End
polyA	polyAdenylation
RT-PCR	Reverse Transcriptase Polymerase Chain Reaction
TAD	Topological Associating Domain
TF	Transcription Factor
TFBS	Transcription Factor Biding Site
TPM	Transcript Per Million
TSS	Transcription Start Site
TTS	Transcription Termination Site

Declarations

Ethics approval and consent to participate

All animal handling and sampling was realized in conformity with the French legislation on animal experimentation. Blood samples were collected under the authorizations APAFIS#334-2015031615255004_v4, APAFIS#333-2015031613482601_v4 and APAFIS#3066-201511301610897_v2. Chicken immune cells were obtained from spleen sampled after slaughter (no need for animal experiment authorization).

Consent for publication

Not applicable.

Availability of data and material

Experimental protocols for tissue sampling and molecular assays are available at the FAANG portal of the EMBL-EBI ftp website: `ftp://ftp.faang.ebi.ac.uk/ftp/`

protocols/samples/. Sample records are available at the BioSamples database using the keyword "FR-AgENCODE", under the submission code GSB-99. Biological material from the 16 animals (tissue samples and aliquots) are available at the INRA CRB-Anim BioBanking facility on request. Fastq sequences are available at the NCBI SRA under the accession project ERP023985 or using the keyword "FR-AgENCODE". Additional data and materials, including annotation files with RNA-seq, ATAC-seq and Hi-C results are available at the FR-AgENCODE portal: www.fragencode.org.

Competing interests

The authors declare that they have no competing interests.

Funding

This study has been supported by the INRA "SelGen metaprogramme", grant "FR-AgENCODE: A French pilot project to enrich the annotation of livestock genomes" (2015-2017). S. Djebali and A. Rau are supported by the Agreenskills fellowship program with funding from the EU's Seventh Framework Program under grant agreement FP7-609398. Additional financial support for tissue biobanking was provided by the CRB-Anim infrastructure project, ANR-11-INBS-0003, funded by the French National Research Agency in the context of the "Investing for the Future" program.

Authors' Contributions

Animal and sampling: MTB and SFa (coordination), AG, EG, FB, FD, FL, GTK, HA, MTB, PQ, SFa, SLL, SVN (sampling and cell sorting). Molecular Assays: DE and HA (coordination), SDP (RNA-seq libraries), AG, EG, KMun (ATAC-seq libraries), FM, HA, MM (Hi-C libraries). Bioinformatics and Data Analysis: CK, SD, SFo (coordination), CC, SD (RNA-seq), KMun, SD, SFo (ATAC-seq), KMur, SL, TD (lncRNAs), AR, NVV, RML (differential analyses), DR, IG, MM, MSC, MZ, NVV, SFo (Hi-C pipeline), EC, EG, SD, SL, TD (functional analyses), SD, SFo (metadata), AR, NVV, SD, SFo (integrative analyses), HA, SFo, SM, PB (data submission). Manuscript writing: AR, CK, EG, HA, KMun, KMur, MTB, MZ, NVV, SD, SFo, SL, TD. Management committee: EG, MHP, SFo, SL. Project coordination: EG, SFo. All authors have read and approved the manuscript.

Acknowledgements

We would like to thank the FAANG community for the general support and in particular A. Archibald (Roslin Institute, UK), L. Clarke, P. Harrison (EMBL-EBI, UK) and M. Groenen (WUR, Netherlands) for their collaboration in the organization of the project, and B. Rosen (ARS, USDA) for providing information about the goat sexual chromosomes. We wish to thank all field operators at INRA experimental animal facilities, units and platforms in France for the access and the assistance in animal handling and sampling, including: Yves Gallard, UE Le Pin (Gouffern en Auge), Frédéric Bouvier and Thierry Fassier, UE Bourges (Osmoy), Stéphane Ferchaud, UE GenESI (Magneraud/Rouillé), Yannick Baumard, UE PEAT (Nouzilly), Cécile Berri and Joel Gautron, UR BOA (Nouzilly), Gilles Gomot, Jean-Philippe

Dubois and Albert Arnould, UR PRC CIRE (Nouzilly), Elodie Guettier, Damien Capo and Jonathan Savoie, UE PAO (Nouzilly).

We are grateful to Alessandra Breschi (Stanford, USA) for sharing R visualization scripts, Julien Lagarde (CRG, Spain) for the samToPolyA.pl tool, and Nicolas Servant (Institut Curie, France) for assistance on HiC-PRO. Additional acknowledgements go to Cécile Donnadieu and Olivier Bouchez from the Get-Plage sequencing platform (INRA Toulouse) and to Christine Gaspin and her staff at the GenoToul bioinformatics platform (INRA Toulouse), especially Didier Laborie and Marie-Stéphane Trotard for IT support.

Author details

[1]GenPhySE, Université de Toulouse, INRA, INPT, ENVT, Chemin de Borde Rouge, F-31326 Castanet-Tolosan Cedex, France. [2]Curtin University, School of Biomedical Sciences, CHIRI Biosciences, 24105 Perth, Australia. [3]MIAT, Université de Toulouse, INRA, Chemin de Borde Rouge, F-31326 Castanet-Tolosan Cedex, France. [4]GABI, AgroParisTech, INRA, Université Paris Saclay, F-78350 Jouy-en-Josas, France. [5]PEGASE, Agrocampus-Ouest, INRA, F-35590 Saint-Gilles Cedex, France. [6]GetPlaGe, Université de Toulouse, INRA, Chemin de Borde Rouge, F-31326 Castanet-Tolosan Cedex, France. [7]UMR6290 IGDR, CNRS, Université Rennes 1, Rennes, France. [8]UMR1282 ISP, INRA, F-37380 Nouzilly, France. [9]IRCM, CEA, Université Paris Saclay, F-78350 Jouy-en-Josas, France.

References

1. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Research . 2005; 15:1034–1050.
2. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences . 2009; 106:9362–9367.
3. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science . 2012; 337:1190–1195.
4. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature . 2012; 489:57–74.
5. Yue F, Cheng Y, Breschi A, Vierstra J, Wu W, Ryba T, et al. A comparative encyclopedia of DNA elements in the mouse genome. Nature . 2014; 515:355–364.
6. Gerstein MB, Rozowsky J, Yan KK, Wang D, Cheng C, Brown JB, et al. Comparative analysis of the transcriptome across distant species. Nature . 2014; 512:445–448.
7. Vernot B, Stergachis AB, Maurano MT, Vierstra J, Neph S, Thurman RE, et al. Personal and population genomics of human regulatory variation. Genome research . 2012; 22:1689–1697.
8. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, et al. Integrative analysis of 111 reference human epigenomes. Nature . 2015; 518:317–330.
9. Stergachis AB, Neph S, Sandstrom R, Haugen E, Reynolds AP, Zhang M, et al. Conservation of trans-acting circuitry during mammalian regulatory evolution. Nature . 2014; 515:365–370.
10. Cheng Y, Ma Z, Kim BH, Wu W, Cayting P, Boyle AP, et al. Principles of regulatory information conservation between mouse and human. Nature . 2014; 515:371–375.
11. Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, et al. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science . 2010; 328:1036–1040.
12. Andersson L, Archibald AL, Bottema CD, Brauning R, Burgess SC, Burt DW, et al. Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. Genome Biology . 2015; 16:57.
13. Tuggle CK, Giuffra E, White SN, Clarke L, Zhou H, Ross PJ, et al. GO-FAANG meeting: a gathering on functional annotation of animal genomes. Animal Genetics . 2016; 47:528–533.
14. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nature Methods . 2013; 10:1213–1218.
15. Lieberman-Aiden E, Van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science . 2009; 326:289–293.
16. Mank JE. Sex chromosome dosage compensation: definitely not for everyone. Trends in Genetics . 2013; 29:677–683.
17. Lin S, Lin Y, Nery JR, Urich MA, Breschi A, Davis CA, et al. Comparison of the transcriptional landscapes between human and mouse tissues. Proceedings of the National Academy of Sciences . 2014; 111:17224–17229.
18. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics . 2010; 26:139–140.
19. Jiang D, Zheng D, Wang L, Huang Y, Liu H, Xu L, et al. Elevated PLA2G7 gene promoter methylation as a gender-specific marker of aging increases the risk of coronary heart disease in females. PLoS ONE . 2013; 8:e59752.
20. Stumpel C, Vos YJ. L1 syndrome. In M Adam, H Ardinger, R Pagon, et al., editors, GeneReviews, Seattle, USA: University of Washington. 2015; Available from: https://www.ncbi.nlm.nih.gov/books/NBK1484/.

21. Arentsen T, Khalid R, Qian Y, Heijtz RD. Sex-dependent alterations in motor and anxiety-like behavior of aged bacterial peptidoglycan sensing molecule 2 knockout mice. Brain, Behavior, and Immunity . 2018; 67:345–354.
22. Yamaguchi S, Yamada Y, Matsuo H, Segawa T, Watanabe S, Kato K, et al. Gender differences in the association of gene polymorphisms with type 2 diabetes mellitus. International Journal of Molecular Medicine . 2007; 19:631–637.
23. Goodman S, Zhang L, Cheng L, Jiang Z. Differential expression of IMP3 between male and female mature teratomas – immunohistochemical evidence of malignant nature. Histopathology . 2014; 65:483–489.
24. Clark EL, Bush SJ, McCulloch ME, Farquhar IL, Young R, Lefevre L, et al. A high resolution atlas of gene expression in the domestic sheep (*Ovis aries*). PLoS Genetics . 2017; 13:e1006997.
25. Bolcun-Filas E, Bannister LA, Barash A, Schimenti KJ, Hartford SA, Eppig JJ, et al. A-MYB (MYBL1) transcription factor is a master regulator of male meiosis. Development . 2011; 138:3319–3330.
26. Zhang Y, Klein K, Sugathan A, Nassery N, Dombkowski A, Zanger UM, et al. Transcriptional profiling of human liver identifies sex-biased genes associated with polygenic dyslipidemia and coronary artery disease. PLoS ONE . 2011; 6:e23506.
27. Conforto TL, Zhang Y, Sherman J, Waxman DJ. Impact of CUX2 on the female mouse liver transcriptome: activation of female-biased genes and repression of male-biased genes. Molecular and Cellular Biology . 2012; 32:4611–4627.
28. Gerner W, Käser T, Saalmüller A. Porcine T lymphocytes and NK cells – an update. Developmental & Comparative Immunology . 2009; 33:310–320.
29. Guzman E, Hope J, Taylor G, Smith AL, Cubillos-Zapata C, Charleston B. Bovine $\gamma\delta$ T cells are a major regulatory T cell subset. The Journal of Immunology . 2014; 193:208–222.
30. Kapushesky M, Adamusiak T, Burdett T, Culhane A, Farne A, Filippov A, et al. Gene Expression Atlas update – a value-added database of microarray and sequencing-based functional genomics experiments. Nucleic Acids Research . 2011; 40:D1077–D1081.
31. Petryszak R, Burdett T, Fiorelli B, Fonseca NA, Gonzalez-Porta M, Hastings E, et al. Expression Atlas update – a database of gene and transcript expression from microarray-and sequencing-based functional genomics experiments. Nucleic Acids Research . 2014; 42:D926–D932.
32. Takamatsu H, Denyer MS, Stirling CMA, Cox S, Aggarwal N, Dash P, et al. Porcine $\gamma\delta$ T cells: possible roles on the innate and adaptive immune responses following virus infection. Veterinary Immunology and Immunopathology . 2006; 112:49–61.
33. Shay T, Jojic V, Zuk O, Rothamel K, Puyraimond-Zemmour D, Feng T, et al. Conservation and divergence in the transcriptional programs of the human and mouse immune systems. Proceedings of the National Academy of Sciences . 2013; 110:2946–2951.
34. Wucher V, Legeai F, Hédan B, Rizk G, Lagoutte L, Leeb T, et al. FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome. Nucleic Acids Research . 2017; 45:e57.
35. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, et al. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Research . 2012; 22:1775–1789.
36. Muret K, Klopp C, Wucher V, Esquerré D, Legeai F, Lecerf F, et al. Long noncoding RNA repertoire in chicken liver and adipose tissue. Genetics Selection Evolution . 2017; 49:6.
37. Lagarde J, Uszczynska-Ratajczak B, Santoyo-Lopez J, Gonzalez JM, Tapanari E, Mudge JM, et al. Extension of human lncRNA transcripts by RACE coupled with long-read high-throughput sequencing (RACE-Seq). Nature Communications . 2016; 7:12339.
38. Hezroni H, Koppstein D, Schwartz MG, Avrutin A, Bartel DP, Ulitsky I. Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. Cell Reports . 2015; 11:1110–1122.
39. Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, et al. The evolution of lncRNA repertoires and expression patterns in tetrapods. Nature . 2014; 505:635–640.
40. Shepard PJ, Choi EA, Lu J, Flanagan LA, Hertel KJ, Shi Y. Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. Rna . 2011; 17:761–772.
41. Chang JH, Tong L. Mitochondrial poly (A) polymerase and polyadenylation. Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms . 2012; 1819:992–997.
42. Scott-Browne JP, López-Moyado IF, Trifari S, Wong V, Chavez L, Rao A, et al. Dynamic changes in chromatin accessibility occur in CD8+ T cells responding to viral infection. Immunity . 2016; 45:1327–1340.
43. Rao S, Huntley M, Durand N, Stamenova E, Bochkov I, Robinson J, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell . 2014; 159:1665–1680.
44. Servant N, Lajoie BR, Nora EP, Giorgetti L, Chen CJ, Heard E, et al. HiTC: exploration of high-throughput 'C' experiments. Bioinformatics . 2012; 28:2843–2844.
45. Filippova D, Patro R, Duggal G, Kingsford C. Identification of alternative topological domains in chromatin. Algorithms for Molecular Biology . 2014; 9:14.
46. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature . 2012; 485:376–380.
47. Tagu D, Colbourne JK, Nègre N. Genomic data integration for ecological and evolutionary traits in non-model organisms. BMC genomics . 2014; 15:490.
48. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics . 2013; 29:15–21.
49. Dobin A, Gingeras TR. Mapping RNA-seq reads with STAR. Current Protocols in Bioinformatics . 2015; 51:11–14.
50. Djebali S, Wucher V, Foissac S, Hitte C, Corre E, Derrien T. Bioinformatics pipeline for transcriptome sequencing analysis. In U Ørom, editor, Enhancer RNAs, New York, NY, USA: Humana Press, volume 1468, pp. 201–219. 2017; .
51. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.

Nature Biotechnology . 2010; 28:511–515.
52. Roberts A, Pimentel H, Trapnell C, Pachter L. Identification of novel transcripts in annotated genomes using RNA-Seq. Bioinformatics . 2011; 27:2325–2329.
53. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics . 2011; 12:323.
54. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, et al. Landscape of transcription in human cells. Nature . 2012; 489:101.
55. Rohart F, Gautier B, Singh A, Lê Cao KA. mixOmics: an R package for omics feature selection and multiple data integration. PLoS Computational Biology . 2017; 13:e1005752.
56. Kinsella RJ, Kähäri A, Haider S, Zamora J, Proctor G, Spudich G, et al. Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database . 2011; 2011:bar030.
57. Quinlan AR. BEDTools: the Swiss-army tool for genome feature analysis. Current Protocols in Bioinformatics . 2014; 47:11–12.
58. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology . 2010; 11:R25.
59. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Research . 2012; 40:4288–4297.
60. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, Series B (Methodological) . 1995; 57:289–300.
61. Falcon S, Gentleman R. Using GOstats to test gene lists for GO term association. Bioinformatics . 2006; 23:257–258.
62. Breiman L. Random forests. Machine Learning . 2001; 45:5–32.
63. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature Methods . 2012; 9:357–359.
64. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics . 2009; 25:2078–2079.
65. Feng J, Liu T, Qin B, Zhang Y, Liu XS. Identifying ChIP-seq enrichment using MACS. Nature Protocols . 2012; 7:1728–1740.
66. Ballman KV, Grill DE, Oberg AL, Therneau TM. Faster cyclic loess: normalizing RNA arrays via linear models. Bioinformatics . 2004; 20:2778–2786.
67. Lun AT, Smyth GK. csaw: a Bioconductor package for differential binding analysis of ChIP-seq data using sliding windows. Nucleic Acids Research . 2015; 44:e45.
68. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics . 2011; 27:1017–1018.
69. Mathelier A, Fornes O, Arenillas DJ, Chen Cy, Denay G, Lee J, et al. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. Nucleic Acids Research . 2015; 44:D110–D115.
70. Servant N, Varoquaux N, Lajoie BR, Viara E, Chen CJ, Vert JP, et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biology . 2015; 16:259.
71. Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, et al. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nature Methods . 2012; 9:999–1003.
72. Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, et al. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. Cell Systems . 2016; 3:99–101.
73. Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Research . 2016; 44:W242–W245.

Figures

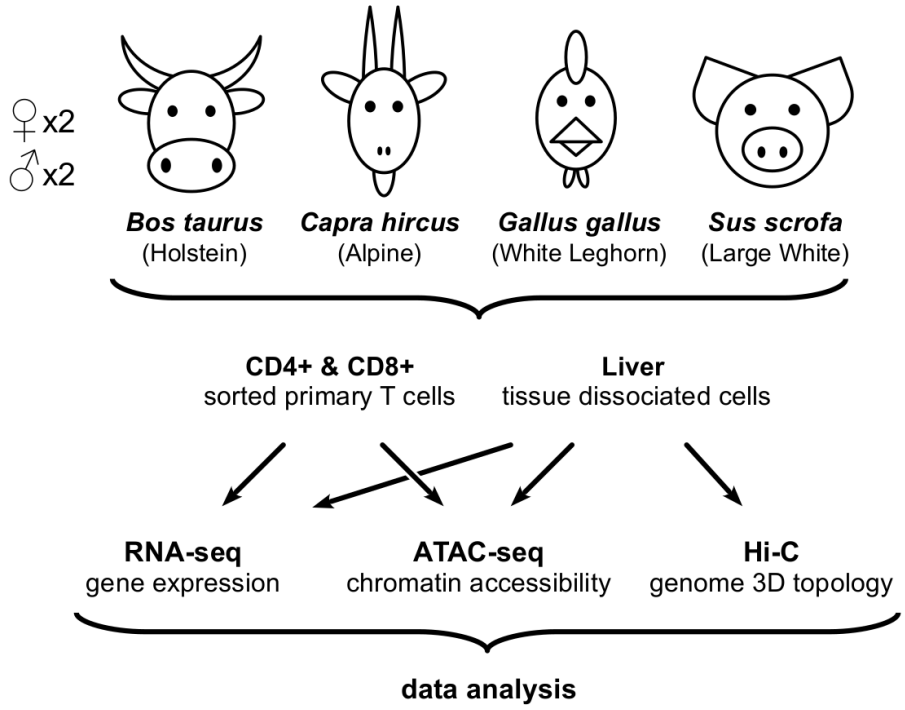


Figure 1 Experimental design overview. For each species, samples from liver and T cells of two males and two females were processed by RNA-seq, ATAC-seq and Hi-C assays. See Table S1 for a complete list of performed experiments and available datasets.

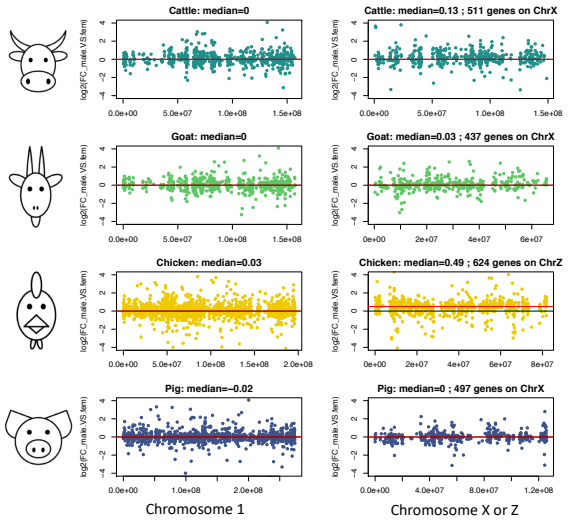


Figure 2 Gene expression ratio in males versus females on chromosomes X or Z. Gene expression was compared in males vs. females using the base 10 logarithm of the TPM values. y-axis: the $\log_2$ of the male/female expression ratio for the genes located on the chromosome of interest. For each plot, the median of these expression ratios is indicated in the title. Left: genes located on chromosome 1 used as a control. Right: genes located on chromosome X for mammals and Z for chicken. x-axis: the position of the gene on the chromosome of interest.

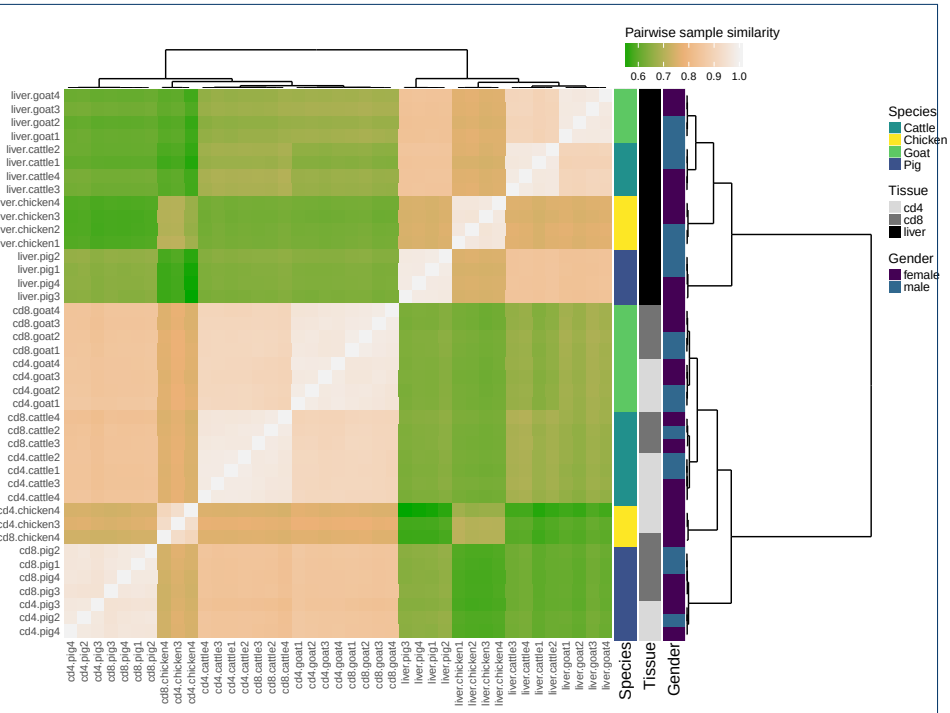


Figure 3 RNA-seq sample heatmap and hierarchical clustering based on the expression of the 9,461 orthologous genes across the four species. Pairwise similarity between samples is computed as the Pearson correlation between the base 10 logarithm of the expression (TPM) of the 9,461 orthologous genes. These similarities are plotted as a heatmap, where samples appear both as rows and columns and are labelled by their species, tissue and the sex of the animal. The color of each heatmap cell also reflects the similarity (Pearson correlation) between each sample pair (the lighter, the higher). Hierarchical clustering is also performed using one minus the squared Pearson correlation as a distance and the complete linkage aggregation method.

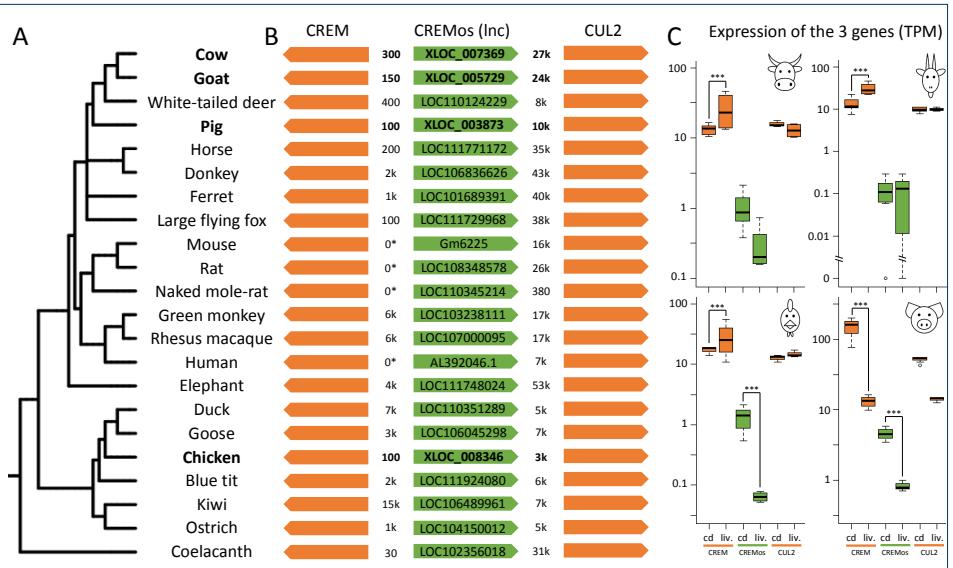


Figure 4 A novel lncRNA conserved across multiple species. Phylogenetic representation based on the NCBI taxonomy of the 22 annotated species from fishes to mammals using iTOL [73] **(A)** Three gene syntenic region centered on *CREMos* lncRNA with NCBI IDs for lncRNAs already annotated in reference databases and distance between entities in nucleotides ; The cases where *CREM* and *CREMos* genes are overlapping are indicated by the "0*" distance. **(B)** Expression of the 3 genes in cattle, goat, chicken and pig: *CREMos* is generally less expressed in liver than in CD cells (in cattle, chicken and pig) whereas *CREM* is generally more expressed in liver than in CD cells (in cattle, chicken and goat).

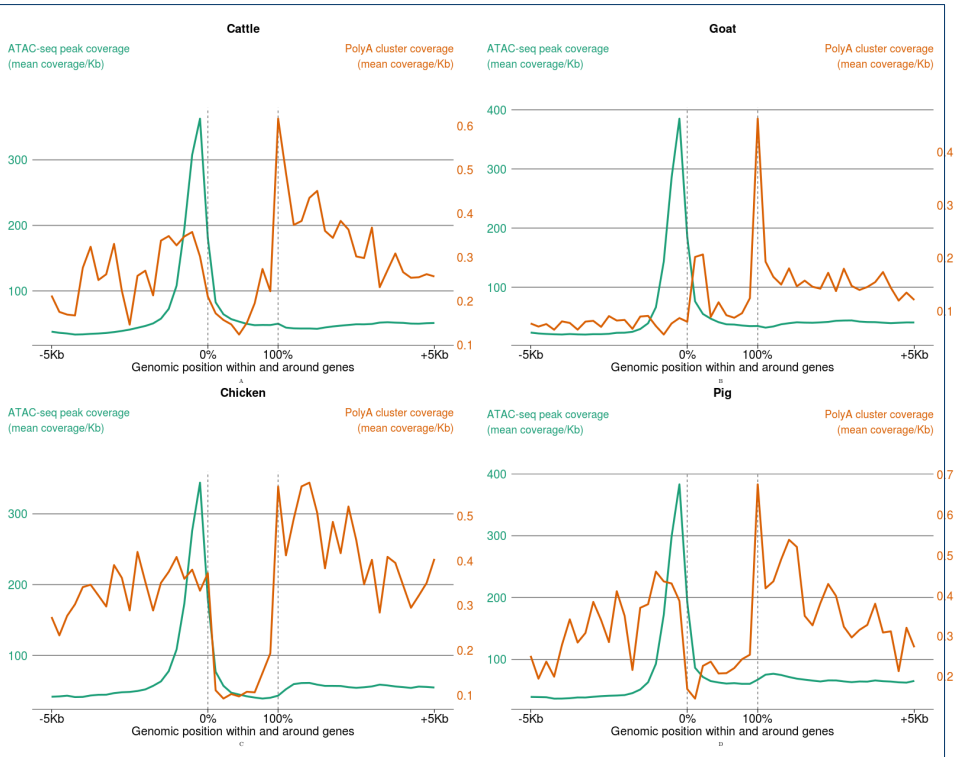
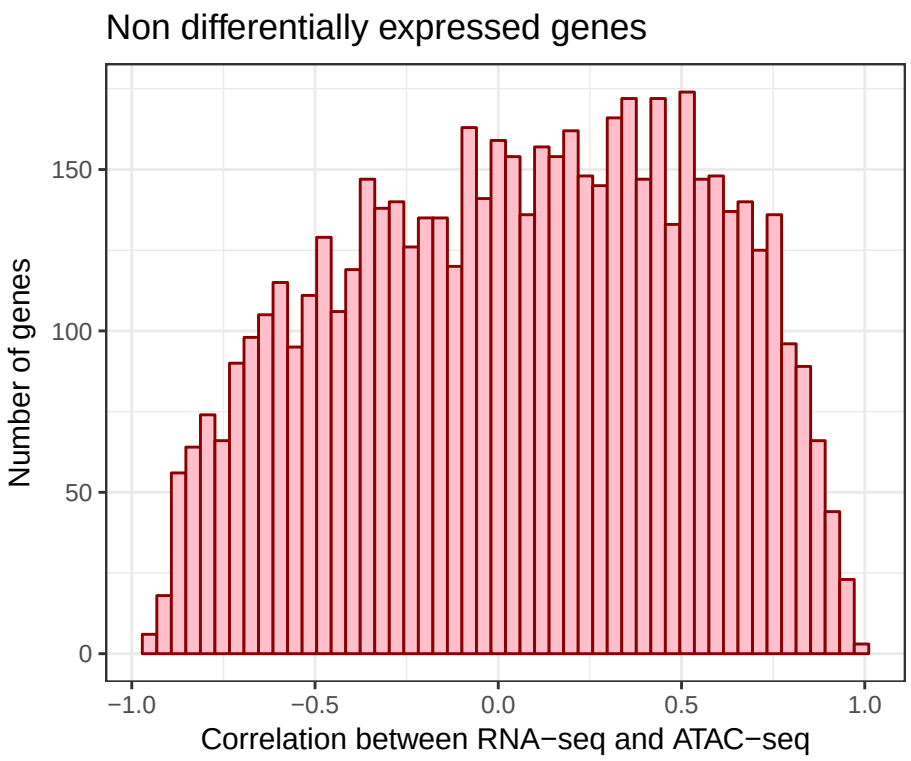
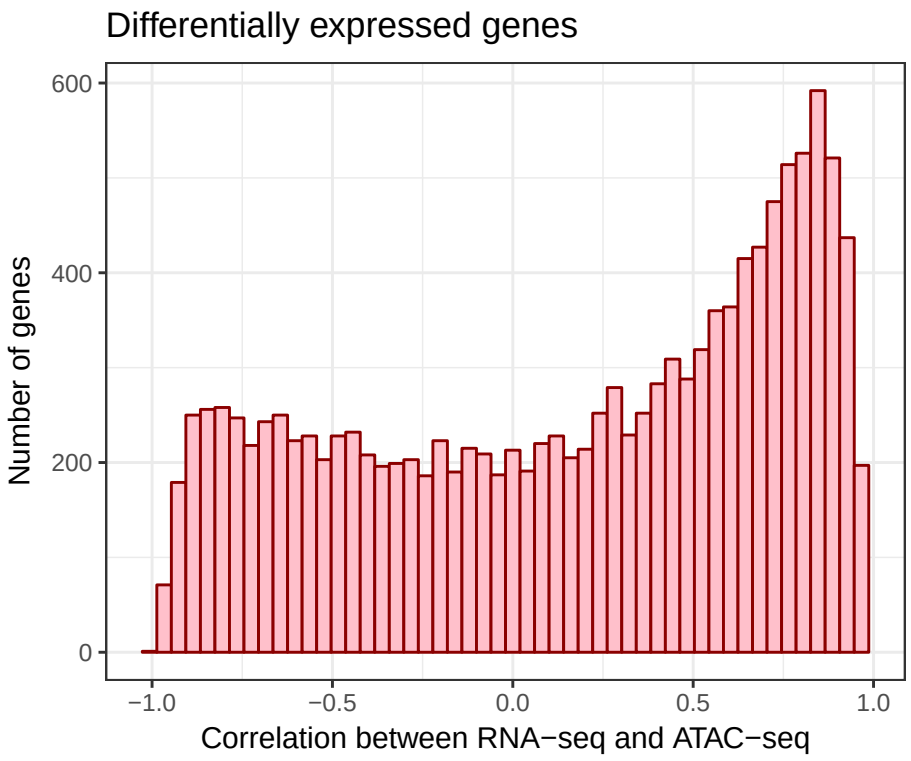


Figure 5 Density of ATAC-seq peaks and polyA clusters within and around reference genes for cattle (A), goat (B), chicken (C) and pig (D). Mean coverage values of ATAC-seq peaks (y-axis, left) and polyA clusters (y-axis, right) were computed genome-wide relatively to the position of the genes (x-axis). Dotted lines indicate gene boundaries. Since genes have variable sizes, an absolute scale is used for the upstream and downstream flanking regions (from -5Kb before to +5Kb after the genes) while a relative scale is used within the gene regions (from 0 to 100% of the gene length). As expected, the ATAC-seq and the polyA signals tend to accumulate at the 5' and 3' ends of the genes respectively.

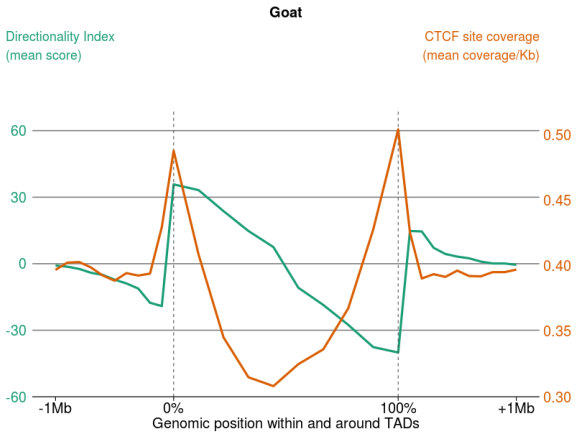


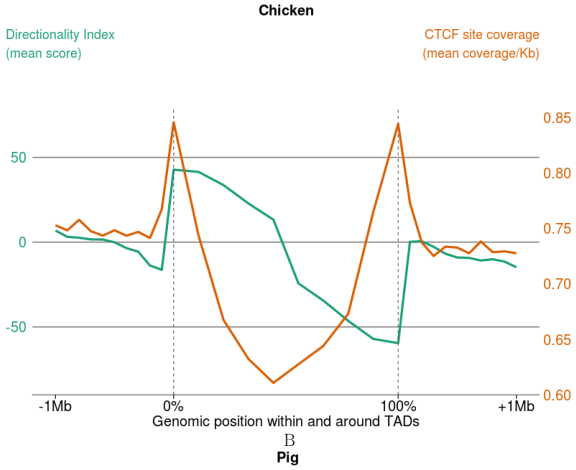
A



B

Figure 6 Correlation between gene expression and promoter accessibility in pig. For each expressed FR-AgENCODE gene with an ATAC-seq peak in the promoter region, the Pearson correlation is computed between the base 10 logarithm of the RNA-seq gene expression and the base 10 logarithm of the ATAC-seq chromatin accessibility. The distribution is represented for genes with no significant differential expression between liver and T cells **(A, top)** and for differentially expressed genes **(B, bottom)**. The distribution obtained for differentially expressed genes showed an accumulation of both positive and negative correlation values, suggesting a mixture of regulatory mechanisms.





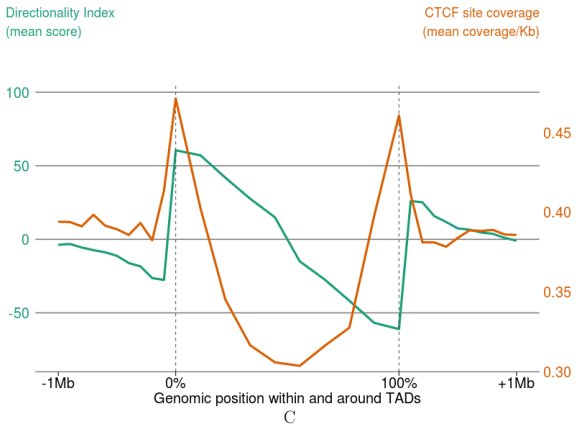
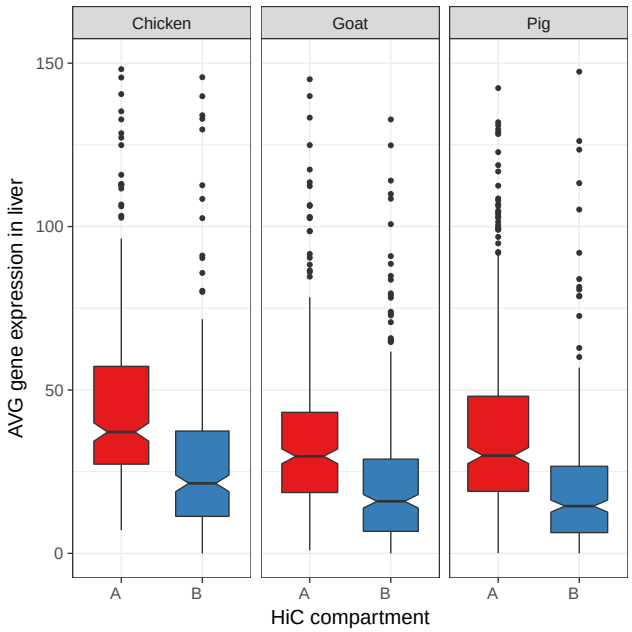
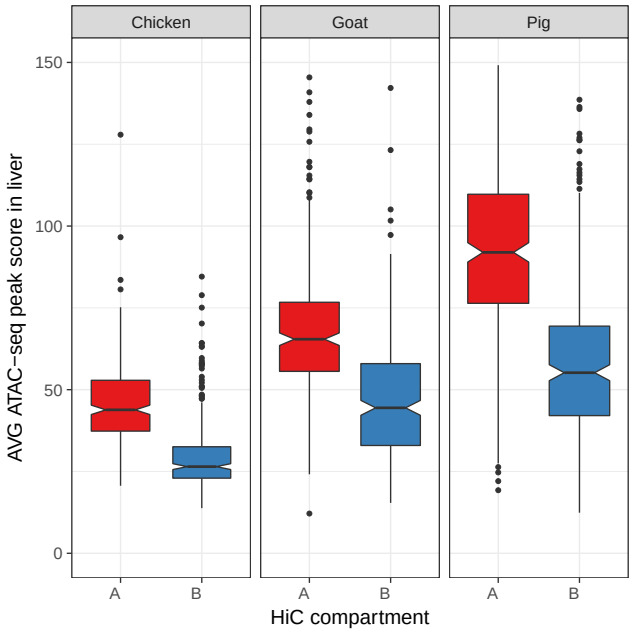


Figure 7 CTCF motif density and Directionality Index within and around TADs. Spatial proximity bias toward upstream vs. downstream positions is measured from Hi-C interaction matrices using the Directionality Index and represented on the y-axis (left). The mean density of predicted CTCF binding sites is also shown on the y-axis (right). Mean DI and CTCF density are plotted relative to the positions of Hi-C-derived Topologically Associating Domains. Dotted lines indicate TAD boundaries. Absolute scale is used on the x-axis up to 1Mb on each side of the TADs while relative positions are used inside the domains (from 0 to 100% of the TAD length).



A



B

Figure 8 Gene expression and chromatin accessibility in A and B topological compartments. For the three species with Hi-C-derived A and B compartments, the distribution of the RNA-seq gene expression values (normalized read counts, top panel) and ATAC-seq chromatin accessibility values (normalized read counts, bottom panel) is shown per compartment type. A: "active" compartments. B: "repressed" compartments. As Hi-C data was only available from liver, only RNA-seq and ATAC-seq values from the same samples were considered. The significant and systematic difference of gene expression and chromatin accessibility values between A and B compartments (p-values $<$ 2.2e-16, Wilcoxon test) confirm a general consistency between RNA-seq, ATAC-seq and Hi-C data across species.

Tables

Table 1 Reference and FR-AgENCODE detected transcripts. This table provides the total number of reference transcripts for each species, the number and percent of those that were detected by RNA-seq (TPM $\geq$ 0.1 in at least 2 samples), the total number of FR-AgENCODE transcripts, and the subsets of them that were classified as mRNAs or lncRNAs by FEELnc. Overall, the transcript repertoire is augmented by about 50% in all species but cattle, where it is more than tripled.

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

Additional files

Additional files are available in the supplementary data file section and on the FR-AgENCODE website `www.fragencode.org`.

Additional file 1 — `SF1.pdf`

Supplementary figures (S1-S23) and tables (S1-S15).

Addisional file 2 – `refgn.tar.gz`

Reference genes and transcripts (structure, expression) of the 4 species. Archive content:

- `bos_taurus.gtf`
- `bos_taurus.refgn.tpm.tsv`
- `capra_hircus.gtf`
- `capra_hircus.refgn.tpm.tsv`
- `gallus_gallus.gtf`
- `gallus_gallus.refgn.tpm.tsv`
- `sus_scrofa.gtf`
- `sus_scrofa.refgn.tpm.tsv`

Additional file 3 – `ensembl_orth_1to1_chicken_pig_cow_goat_hsid.tsv`

Orthologs between the 4 livestock species. We used Biomart to retrieve the 1 to 1 orthology relationships between chicken, pig and cow and added goat via gene name. The human gene id is given for reference.

Additional file 4 – `de.refgn.tar.gz`

Reference DE genes (all combinations): the archive contains four folders, one for each species (`bos_taurus`, `capra_hircus`, `gallus_gallus`, `sus_scrofa`). Each folder contains itself three subfolders, one for each model: `diffcounts.nominsum` (Model 1), `diffcounts.cdvsliver` (Model 2) and `diffcounts.withsex` (Model 3).

Results of Model 1 are given in:

- `refgenes.counts.min2tpm0.1.normcounts.diff.readme.idx`
- `refgenes.counts.min2tpm0.1.normcounts.diff.cd4.cd8.bed`
- `refgenes.counts.min2tpm0.1.normcounts.diff.cd4.liver.bed`

- `refgenes.counts.min2tpm0.1.normcounts.diff.cd8.liver.bed`

Results of Model 2 are given in:

- `refgenes.counts.min2tpm0.1.normcounts.diff.readme.idx`
- `refgenes.counts.min2tpm0.1.normcounts.diff.cd.liver.bed`

Results of Model 3 are given:

- for all species but chicken in:
  - `refgenes.counts.min2tpm0.1.normcounts.diff.readme.idx`
  - `refgenes.counts.min2tpm0.1.normcounts.diff.cd4.cd8.bed`
  - `refgenes.counts.min2tpm0.1.normcounts.diff.cd4.liver.bed`
  - `refgenes.counts.min2tpm0.1.normcounts.diff.cd8.liver.bed`
  - `refgenes.counts.min2tpm0.1.normcounts.diff.sex.cd4.bed`
  - `refgenes.counts.min2tpm0.1.normcounts.diff.sex.cd8.bed`
  - `refgenes.counts.min2tpm0.1.normcounts.diff.sex.liver.bed`
- for chicken in:
  - `refgenes.counts.min2tpm0.1.normcounts.diff.readme.idx`
  - `refgenes.counts.min2tpm0.1.normcounts.diff.F.M.bed`

All bed files contain the coordinates and id of the genes found to be differentially expressed between the two conditions. The file also contains the normalized read counts of those genes in the different samples as well as the adjusted pvalue, logFC and normLogFC (see readme.idx file for more details)

Additional file 5 – `TF.curated.tar.gz`

Manually curated lists of immunity and metabolism TFs.:

- `liver_TFnames.txt`
- `Tcell_TFnames.txt`

Additional file 6 – `fraggn.tar.gz`

FR-AgENCODE genes and transcripts (structure, expression, positional and coding classes).

- `bos_taurus_cuff_tpm0.1_2sample_complete.gff`
- `bos_taurus_cuff_tpm0.1_2sample_trid_4posclasses_3codingclasses_booleans.tsv`
- `bos_taurus.frag.gnid.posclasslist.codclasslist.tsv`
- `bos_taurus.fraggn.tpm.tsv`
- `capra_hircus_cuff_tpm0.1_2sample_complete.gff`
- `capra_hircus_cuff_tpm0.1_2sample_trid_4posclasses_3codingclasses_booleans.tsv`
- `capra_hircus.frag.gnid.posclasslist.codclasslist.tsv`
- `capra_hircus.fraggn.tpm.tsv`
- `gallus_gallus_cuff_tpm0.1_2sample_complete.gff`
- `gallus_gallus_cuff_tpm0.1_2sample_trid_4posclasses_3codingclasses_booleans.tsv`
- `gallus_gallus.frag.gnid.posclasslist.codclasslist.tsv`
- `gallus_gallus.fraggn.tpm.tsv`
- `sus_scrofa_cuff_tpm0.1_2sample_complete.gff`
- `sus_scrofa_cuff_tpm0.1_2sample_trid_4posclasses_3codingclasses_booleans.tsv`
- `sus_scrofa.frag.gnid.posclasslist.codclasslist.tsv`
- `sus_scrofa.fraggn.tpm.tsv`

Additional file 7 – `de.fraggn.gz`

FR-AgENCODE DE genes (all combinations). The archive has the same structure than `de.refgn.tar.gz` with names starting with `cuffgenes` instead of `refgenes`.

Additional file 8 – `lncRNAs.tar.gz`

lncRNAs (information from FEELnc, orthology, structure ...). Archive content:

- `bos_taurus.lncrna.TPM0.1in2samples.classif.tsv`
- `capra_hircus.lncrna.TPM0.1in2samples.classif.tsv`
- `ConservedLncRNABySynteny_73_19_6.xlsx`
- `gallus_gallus.lncrna.TPM0.1in2samples.classif.tsv`
- `sus_scrofa.lncrna.TPM0.1in2samples.classif.tsv`

Additional file 9 – `polyAsite.clusters.tar.gz`

polyA site clusters.

- `bos_taurus.polyAsites.minclip10.merged.maxdist10.minreads2.bed`
- `capra_hircus.polyAsites.minclip10.merged.maxdist10.minreads2.bed`
- `gallus_gallus.polyAsites.minclip10.merged.maxdist10.minreads2.bed`
- `sus_scrofa.polyAsites.minclip10.merged.maxdist10.minreads2.bed`

Additional file 10 – `atac.peaks.tar.gz`

ATAC-seq peaks (coordinates, quantification, positional classification): the archive contains four folders, one for each species (`bos_taurus`, `capra_hircus`, `gallus_gallus`, `sus_scrofa`). Each folder contains the following six files:

- `mergedpeaks_allinfo2.tsv`
- `mergedpeaks_allinfo.tr2.tsv`
- `mergedpeaks_allinfo.tr.tsv`
- `mergedpeaks_allinfo.tsv`
- `mergedpeaks.peaknb.allexp.readnb.bed.readme.idx`
- `mergedpeaks.peaknb.allexp.readnb.bed`

Additional file 11 – `da.atacpeaks.tar.gz`

DA ATAC-seq peaks (all combinations). The archive has the same structure than `de.refgn.tar.gz` with names starting with `mergedpeaks.peaknb.allexp.readnb` instead of `refgenes.counts.min2tpm0.1`.

Additional file 12 – `hic.tad.ab.tar.gz`

Hi-C TADs and A/B compartments: the archive contains three folders, one for each species (`capra_hircus`, `gallus_gallus`, `sus_scrofa`). Each folder contains the following two files:

- `compartments.bed`
- `mat.40000.longest25chr.tad.consensus.bed`