An atlas of chicken long non-coding RNAs gathering multiple sources : gene models and expression across more than twenty tissues

Frédéric Jehl, Kévin Muret, Maria Bernard, Diane Esquerré, Hervé Acloque, Elisabetta Giuffra, Sarah Djebali, Sylvain Foissac, Thomas Derrien, Tatiana Zerjal, et al.

An atlas of chicken long non-coding RNAs gathering multiple sources: gene models and expression across more than twenty tissues

Frédéric Jehl[a], Kévin Muret[a], Maria Bernard[a], Diane Esquerré, Hervé Acloque, Elisabetta Giuffra, Sarah Djebali, Sylvain Foissac, Thomas Derrien, Tatiana Zerjal, **Christophe Klopp[§] and Sandrine Lagarrigue[§]**

Team « Genetics and Genomics »
UMR INRA – Agrocampus Ouest PEGASE (1348)
Rennes, France

Sunday, January 13th 2019
Non-coding RNA workshop – Plant and Animal Genome XXVII
San Diego, CA

Context: lncRNAs, from cellular processes to complex traits variation

Long non-coding RNAs (lncRNAs) – transcripts longer than 200 nucleotides that are not translated into proteins

- lncRNAs have been showed to act on numerous cellular processes:
  - chromatin compaction (*Xist*)
  - gene expression :
    - → transcription (*Fendrr*, Grote *et al.*, 2013)
    - → translation (*lincRNA-p21*, Yoon *et al.*, 2012)
- Influence on complex traits & diseases ?
  - General agreement that complex traits and diseases are influenced by numerous loci
    ⇒ 90% out of coding regions (Manolio *et al.*, 2009)
  - These loci are located in regulatory regions (Maurano *et al.*, 2012)
    ⇒ effect on genes expression
  - Some of them affect lncRNA expression (Kumar *et al.*, 2013)
- Numerous, poor functional annotation:
  - no function prediction rules (yet)
  - lowly expressed (10-fold less than mRNAs): difficult to validate experimentally

Context: the chicken (*G. gallus*), a specie of great economical and social importance

- **Meat :** currently, one of the most produced and consumed meat in the world
  ⇒ almost 120 millions tonnes produced in 2016
- **Eggs :** good sources of proteins, fatty acids and micronutrients for human consumption
  - Easy to produce and store, no cultural restriction for consumption
  - In 2011, ~200 per year per capita (IEC, 2011)
    ⇒ almost 70 millions tonnes produced in 2016

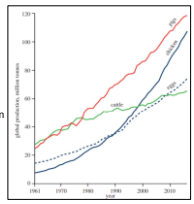


Fig. 1 : Poultry, pig, cattle and egg consumption from 1961 to 2018 (million tonnes). Top left : number of animal slaughtered in 2016.
From Bennett *et al.*, 2018.

Fundamental goal #1: improve our knowledge of the chicken's genes at the genome scale

- **What is the state of the knowledge of the chicken's genes vs. human's ?**

- Chicken :
  → ~ 20 000 protein coding genes
  → ~ 2 000 small non-coding genes
  → ~ 4 600 long non-coding genes

v94, December 2018

- Human :
  → ~ 20 000 protein coding genes
  → ~ 5 000 small non-coding genes
  → ~ 15 000 long non-coding genes

v94, December 2018

⇒ our knowledge of chicken (long)ncRNA is still incomplete

1. in term of number (low expression ⇒ high sequencing depth in numerous tissues)
2. in term of isoforms (similarly to coding genes, requires technologies ⇒ long reads)

⇒ first goal: improve our knowledge of the chicken's long non-coding genes at the genome scale in term of number of loci, using a high number of short reads (364 samples × 90M reads)

Fundamental goal #2: annotate lncRNAs thanks to their expression profile in chicken, focus on tissue-specificity

Problematic: knowing the existence of a gene (coding or non-coding) doesn't inform on its function(s)

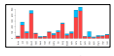
- with coding genes:
  - prediction of the associated amino-acid sequence,
  - prediction of the protein function by motif conservation across species.
- with long non-coding genes:
  - poor sequence conservation,
  - poor / no domains knowledge.

  ⇒ lncRNA are poorly annotated

⇒ second goal: annotate the lncRNAs applying the two following strategies

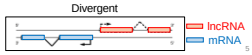
Strategies to annotate lncRNAs function:

- expression profile: lncRNA expressed in tissue *A* and little or not in tissue *B* (= specific to tissue *A*) might have a role related to tissue *A* function



- study their configuration with closest coding gene: *e.g.*, close and divergent ⇒ common regulation ?

Divergent



1. Improvement of our knowledge of chicken's genes at the genome scale

Strategy :

- exploit the volume of expression data from INRA's lab to modelize yet-undiscovered genes, in complement to Ensembl genes
- extension of this catalogue using external sources

Data

- 364 RNA-seq samples from 3 tissues :
  - adipose tissue (56 samples), blood (128 samples) and liver (180 samples)



  - 90 millions reads per sample
  - stranded reads
  - 2 × 150 pb paired-ends

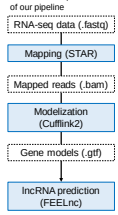
1. Results of gene identification

Methods
- Mapping on *G. gallus* reference genome (GalGal5) with STAR (Dobin *et al.*, 2013), using Ensembl v94 annotation
- Transcriptome reconstruction using Cufflink2 (Trapnell *et al.*, 2013)
- Elimination of modelized genes overlapping on the same strand any Ensembl v94 gene
- lncRNA prediction using FEELnc (Wucher *et al.*, 2017)

Results :
addition of 25 214 genes
to the 24 881 Ensembl genes

⇒ Total of 50 095 genes

Schematic representation of our pipeline



7

1. Selection of the most robust modeled genes based on their expression pattern

Reminder : lncRNA = lowly expressed ($1/10^{th}$ mRNA). Necessity to separate background noise from real expression.

To improve reliability of the 25 214 newly modelized genes, we selected models according to their expression, using 2 expression criteria:
- an normalized expression metric (TPM > 0.1)
- number of reads supporting the model

⇒ **selection of 58.5% of the 25 214 genes (14 760 models)**

INRA genes + Ensembl genes
= **14 760** + 24 881
⇒ **39 641 genes**

Gene type	Number of genes	% of total	Ensembl
lncRNAs	13 009	88.14%	4 641
mRNAs	1 199	8.12%	18 346
Others	552	3.74%	1 894
TOTAL	14 760	100%	24 881

8

1. Selection of the most robust modeled genes based on their expression pattern

Reminder : lncRNA = lowly expressed ($1/10^{th}$ mRNA). Necessity to separate background noise from real expression.

To improve reliability of the 25 214 newly modelized genes, we selected models according to their expression, using 2 expression criteria:
- an normalized expression metric (TPM > 0.1)
- number of reads supporting the model

⇒ **selection of 58.5% of the 25 214 genes (14 760 models)**

Comparison with background noise:
- "No genes" = set of genomic regions out of the Ensembl and our genes models: our models have expression levels superior to the background noise.

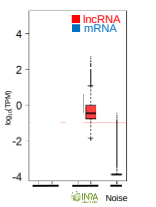


Fig. 4: The 14 760 remaining genes have the same expression characteristics as their Ensembl counterparts, and are not background noise

9

1. Selection of the most robust modeled genes based on their expression pattern

Reminder : lncRNA = lowly expressed ($1/10^{th}$ mRNA). Necessity to separate background noise from real expression.

To improve reliability of the 25 214 newly modelized genes, we selected models according to their expression, using 2 expression criteria:
- an normalized expression metric (TPM > 0.1)
- number of reads supporting the model

⇒ **selection of 58.5% of the 25 214 genes (14 760 models)**

Comparison with background noise:
- "No genes" = set of genomic regions out of the Ensembl and our genes models: our models have expression levels superior to the background noise.

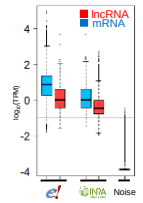


Fig. 4: The 14 760 remaining genes have the same expression characteristics as their Ensembl counterparts, and are not background noise

10

1. Sequential building of an extended annotation using 6 external data sources

From this set of 14 760 INRA genes (88% lncRNA) + 24 881 Ensembl genes,
➡ **extension of the chicken lncRNA annotation using external sources:**
1. NONCODE (Fang *et al.*, 2018) : 9 322 lncRNAs
2. NCBI : 5 738 lncRNAs
3. ALDB (Li *et al.*, 2015) : 5 752 lncRNAs
4. FR-AGENCODE (Foissac *et al.*, submitted) : 6 089 lncRNAs

For each source, we kept only the loci with no overlap with the extended annotation:

- As a result, we increased gene numbers from 14 760 + 24 881 to →→→→ 52 075 genes including
  - 30 084 lncRNAs : **×6.5** compared to Ensembl alone
  - 19 545 mRNAs : **×1.1** compared to Ensembl alone
- This work was done at the locus level since isoforms are poorly known and expression analysis are done at the gene level

11

2. Annotation of the lncRNAs thanks to their expression profile in chicken, focus on tissue-specificity

Thanks to the previous step :
- ☑ extended lncRNA catalogue (from 4 640 to **30 084**)
- ☒ role or function of the lncRNAs ?

⇒ objective of this second part: provide an annotation of the potential lncRNA functions using their expression profiles

12

2. Annotation of the lncRNAs thanks to their expression profile in chicken, focus on tissue-specificity

Hypothesis: if a gene is expressed in one or a few tissues, its function is likely to be related to the function of the tissue(s).

To study tissue-specificity, necessity to have numerous tissues ⇒ public data from other teams.

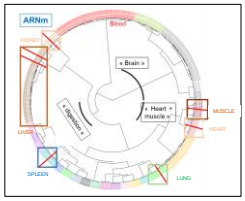
3 datasets were analyzed:
- INRA : 5 tissues (Jehl *et al*, in preparation)
- Sishuan University : 6 tissues (Tang *et al.*, 2017)
- Roslin Institute : 21 tissues (used by Ensembl for the chicken genome annotation)

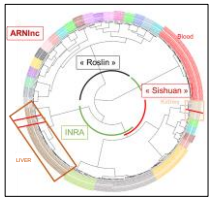
⚠ lncRNAs are lowly expressed ⇒ sensitivity to genotype, physiological status, experimental conditions, …

Preliminary question : can these 3 datasets be studied together after expression normalization, or is there a batch effect ?

2. Datasets comparisons

- The mRNAs expression clusters the sample based on tissues and functions
- The lncRNAs expression clusters the sample mainly by projects

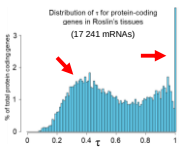


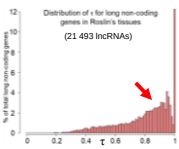


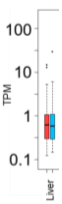
⇒ For the lncRNAs, combining different datasets is not possible
⇒ focus on the 21 tissues from the Roslin Institute data

2. lncRNA and mRNAs have different tissue-specificity patterns

Use of the tissue specificity index τ, which ranges from 0 (same expression in all tissues) to 1 (expression in one tissue only) (Yanai *et al.*, 2005)







There are 2 patterns of tissue-specificity:

⇒ for the mRNAs, a "bump" corresponding to the "housekeeping genes" (τ_{ACTB} = 0.20 ; τ_{GAPDH} = 0.38), then the tissue-specific genes (τ_{TG} = 0.93 ; τ_{MLN} = 1)

⇒ for the lncRNAs, a majority of tissue-specific genes, as expected (Derrien *et al.*, 2012)

665 mRNAs are tissue-specific (τ = 1), with between 6 to 127 mRNA specific to a given tissue

2 634 lncRNAs are tissue-specific (τ = 1), with between 14 to 474 lncRNA specific to a given tissue

2. lncRNA classification using FEELnc

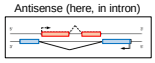
Long non-coding genes are classified by position relatively to the nearest coding genes using FEELnc (Wucher *et al.*, 2017)
⇒ **some configurations suggest potential (co)-regulations**

Examples:

Divergent

In this case, if the genes are close, a co-expression suggests a common regulation, and therefore a common function

Antisense (here, in intron)



In this case, a co-expression suggests that the lncRNAs acts on the mRNA expression

In both case, we infer the function of the lncRNA using the function of the mRNA

	Divergent	Antisense of intron
In the extended catalogue	4 895	2 105
with τ ≥ 0.95	977	418

Conclusion : An atlas of chicken long non-coding RNAs gathering multiple sources : gene models and expression across more than twenty tissues

- We provide a large catalogue of chicken lncRNAs at the gene level
  from **4 640** (Ensembl v94) to **30 084** lncRNAs
- We also provided a rough annotation of all these genes, based on :
  - their expression pattern across 21 chicken tissues
  - their position relative to the nearest coding gene

⇒ this will allow the scientific community to work on different scientific problematics related to lncRNAs in chicken

Perspective:
- we will use these data to study the genetic component of feed efficiency in layer chicken
  ⇒ feed ≃ 60% of production cost in monogastrics
  ⇒ reduction of feed *vs.* food competition
  ⇒ lower environmental impact

Thank you for your attention !

UMR 1313 GABI – team PSGen, Jouy-en-Josas, France
- Tatiana Zerjal

UMR 1313 GABI – team GIS, Jouy-en-Josas, France
- Elisabetta Giuffra

Sigenae platerform – INRA, Toulouse, France
- Maria Bernard
- Christophe Klopp

GeT-PlaGe platerform – GENOTOUL, Toulouse, France
- Diane Esquerré

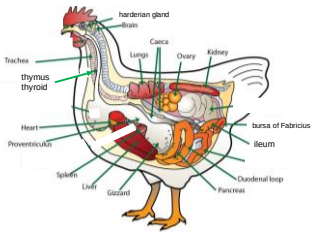
UMR 1348 PEGASE – team GG, Rennes, France
- Frédéric Jehl
- Kévin Muret
- Sandrine Lagarrigue

UMR IGDR, Rennes, France
- Thomas Derrien

UMR 1388 GenPhySE – team GenEpi, Toulouse, France
- Hervé Acloque
- Sarah Djebali
- Sylvain Foissac

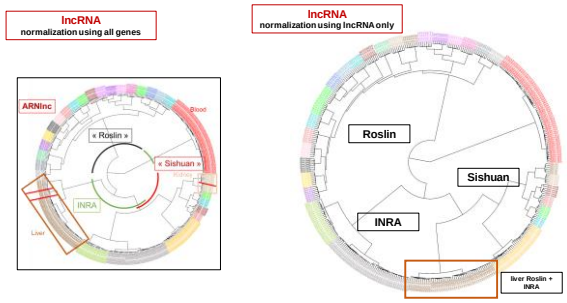
Tissues present in Roslin Institute data

bursa of Fabricius	thyroid	duodenum	proventriculus	trachea	breast muscle	kidney
harderian gland	caecal tonsils	ovary	skin	lung	optical lobe	liver
thymus	ileum	pancreas	fat gizzard	heart	cerebellum	spleen



19

Effect of the biotype of the genes used for TPM normalization on the dendrogram



1. Selection of the most robust modeled genes based on their expression pattern

Reminder : lncRNA = lowly expressed ($1/10^{th}$ mRNA). Necessity to separate background noise from real expression.

To improve reliability of the 25 214 newly modelized genes, we selected models using two criteria:
- an expression metric (TPM)
- number of supporting reads

Additional criterion
- 5 paired-reads or more in 25% of the samples of a tissue: at least 75 supporting reads.

⇒ **leaves 14 760 modelized genes from the 25 214 (= 58.5%).**

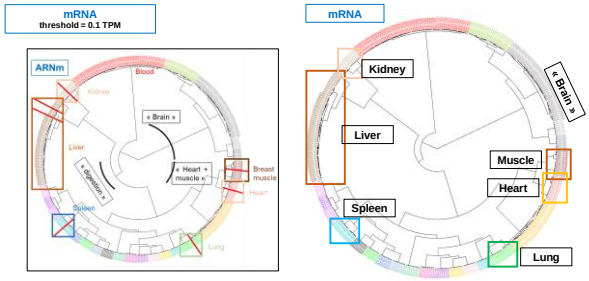
Number of lncRNAs and mRNAs in the remaining catalogue :

Gene type	Number of genes	% of total	Ensembl
lncRNAs	13 009	88.14%	4 641
mRNAs	1 199	8.12%	18 346
Others	552	3.74%	1 894
TOTAL	14 760	100%	24 881

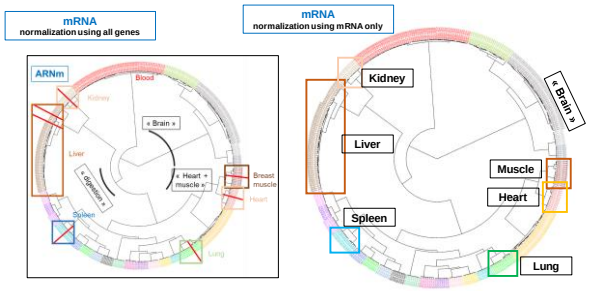
INRA genes + Ensembl genes
= 14 760 + 24 881
= **39 641 genes**

20

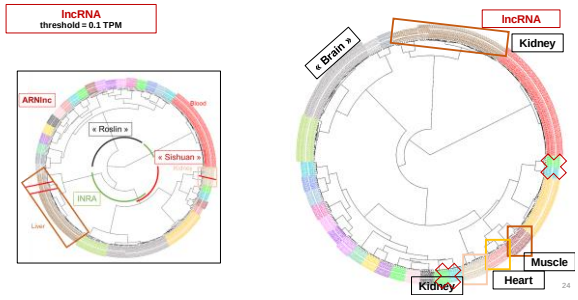
Effects of the expression threshold chosen on the dendrogram : 1 TPM



Effect of the biotype of the genes used for TPM normalization on the dendrogram



Effects of the expression threshold chosen on the dendrogram : 1 TPM



24

A tissue-specificity metrics: the tau (τ)

To study tissue-specificity, we used the tau metrics from Yanai *et al.*, 2005:

- associates a value (τ) to each gene
- for each gene, it accounts for :
  - the expression in each tissue, relatively to the tissue with the highest expression
  - the number of tissues with expression
  - the total number of tissues

- value from 0 to 1
  - τ = 0 ⇒ ubiquist gene: expressed in all tissues at the same level
  - τ = 1 ⇒ tissue-specific gene: expressed in one and only one tissue

$$\hat{x}_i = \frac{x_i}{\max_{1 \le i \le T}(x_i)} \qquad \tau = \frac{\sum_{i=1}^{T}(1 - \hat{x}_i)}{T - 1}$$

x_i : expression in each tissue…

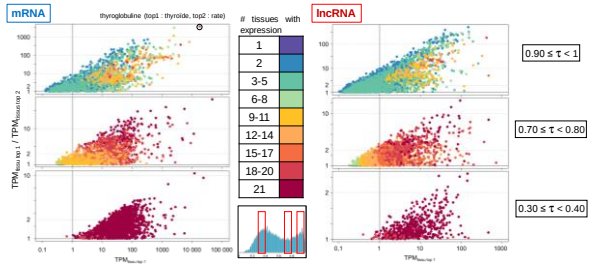
$\max_{1 \le i \le T}(x_i)$: … relatively to the tissue with the highest expression

$\sum_{i=1}^{T}(1 - \hat{x}_i)$: number of tissues with expression

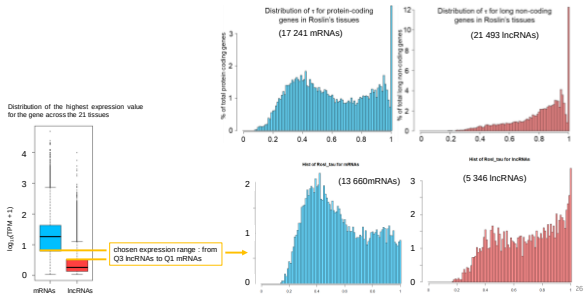
T : total number of tissues

25

Gene expression at different τ values



Tissue-specificity patterns are independant of expression ranges

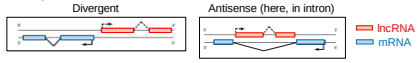


26

lncRNA classification using FEELnc

Long non-coding genes are classified relatively to the nearest coding genes using FEELnc (Wucher *et al.*, 2017)
⇒ some configurations suggest potential (co)-regulations

Examples:

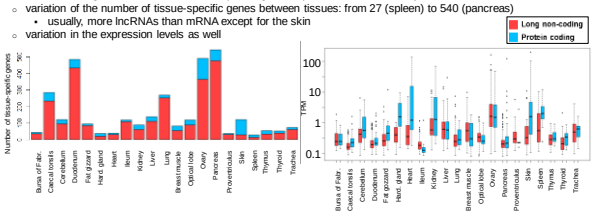


Type					Total	Total classified	Number of genes
Intergenic	Position	distance ≤ 1kb	distance > 1kb	Total			
	same strand	1 403	8 663	10 066	18 204		
	divergent	1 476	3 419	4 895		23 129	30 084
	convergent	584	2 659	3 243			
Genic	Position	direction *sense*	direction *antisense*	Total			
	exonic	2	2 494	2 496	4 925		
	intronic	324	2 105	2 429			
Unclassified					6 955		

29

Number and expression levels of the tissue-specific genes in each tissues

- Focus on the genes with τ = 1 ⇒ tissue-specific gene: expressed in one and only one tissue
- variation of the number of tissue-specific genes between tissues: from 27 (spleen) to 540 (pancreas)
  - usually, more lncRNAs than mRNA except for the skin
- variation in the expression levels as well



1 048 mRNAs are tissue-specific (0.95 ≤ τ ≤ 0.99), with between 5 to 248 mRNA specific to a given tissue

665 mRNAs are tissue-specific (τ = 1), with between 6 to 127 mRNA specific to a given tissue

2 788 lncRNAs are tissue-specific (0.95 ≤ τ ≤ 0.99), with between 17 to 589 lncRNA specific to a given tissue

2 634 lncRNAs are tissue-specific (τ = 1), with between 14 to 474 lncRNA specific to a given tissue

30