

UNE NOUVELLE FACETTE DU GENOME DE POULE : LES GENES “LONGS NON-CODANTS”

Muret Kévin¹, Klopp Christophe², Wucher Valentin³, Esquerré Diane^{4,5}, Legeai Fabrice⁶, Lecerf Frédéric¹, Désert Colette¹, Boutin Morgane¹, Jehl Frédéric¹, Acloque Hervé⁵, Giuffra Elisabetta⁷, Djebali Sarah⁵, Foissac Sylvain⁵, Derrien Thomas³, Lagarrigue Sandrine¹

¹INRA/AGROCAMPUS OUEST UMR PEGASE – 65, rue de St-Brieuc - 35000 RENNES,

²INRA SIGENAE - 24, chemin de Borderouge-Auzeville - 31320 CASTANET-TOLOSAN,

³CNRS/UNIV. RENNES 1 UMR IGDR - 2, avenue du Pr. Léon Bernard - 35000 RENNES,

⁴INRA GENOTOL - 24, chemin de Borderouge-Auzeville - 31320 CASTANET-TOLOSAN,

⁵INRA GENPHYSE/UNIV. DE TOULOUSE - 24, chemin de Borderouge-Auzeville - 31326 CASTANET-TOLOSAN,

⁶INRA/AGROCAMPUS OUEST UMR IGEPP - Domaine de la Motte – 35653 LE RHEU,

⁷INRA GABI/AGROPARISTECH/UNIV. PARIS SACLAY - Domaine de Vilvert - 78352 JOUY-EN-JOSAS

sandrine.lagarrigue@agrocampus-ouest.fr

RÉSUMÉ

De récents travaux effectués chez l'homme et des organismes modèles ont montré l'existence massive d'un nouveau type de gènes : des gènes transcrits en longs ARN (≥ 200 nt) mais ne codant pas de protéine – ces gènes sont appelés gènes à ARN longs non-codants (ARNlnc). Ces ARNlnc sont une composante majeure du transcriptome et seule une petite fraction de ces ARNlnc est décrite aujourd'hui dans les bases de données publiques. Nos travaux sur les ARNlnc de poule ont permis de montrer que le séquençage de transcrits (RNA-Seq) sur plusieurs réplicats biologiques de deux tissus (foie et tissu adipeux) permet de prédire un catalogue d'ARNlnc plus exhaustif pour les tissus considérés que ce qui est proposé dans les bases de données. Nous avons pu confirmer également que les caractéristiques des ARNlnc de poule sont similaires à celles décrites chez la souris ou chez l'Homme (transcrits plus courts, moins d'exons, séquences moins conservées entre les espèces, expression plus tissu spécifique et moins importante que les ARN codant pour des protéines (ARNm)). Nous mettons également en lumière un couple d'ARNlnc et ARNm (DHCR24) adjacent sur le génome et en position divergente dont l'expression est fortement corrélée et dont la position est conservée chez l'Homme suggérant un mécanisme de régulation commun via un promoteur dit « bi-directionnel » qui serait impliqué dans la synthèse du cholestérol.

ABSTRACT

A new side of the chicken genome: Long noncoding RNAs

Recent works in humans and model organisms have shown a massive new type of genes: genes transcribed into long RNAs (≥ 200 nt) but not coding for proteins - these genes are called long noncoding RNA genes (lncRNA). These lncRNAs are a major component of the transcriptome and only a small fraction of them is described in public databases. Our study conducted on chicken shows that transcriptomic sequencing (RNA-Seq) on multiple biological replicates of two tissues (liver and adipose tissue) allows us to predict a lncRNA catalog for these two tissues more exhaustive than lncRNAs catalog available in public databases. We also confirm that chicken lncRNA features are similar to those of human and murine lncRNAs described in the literature (shorter transcripts, less exons, sequences less conserved across species, more tissue-specific expression and lower expression than protein coding RNAs (mRNA)). In conclusion, we provide a comprehensive lncRNA repertoire in the chicken liver and adipose tissue and we highlight a new lncRNA-mRNA pair in divergent position in the chicken genome (as in the human genome), highly co-expressed suggesting a common regulation mechanism through a bidirectional promoter and probably involved in cholesterol metabolism.

INTRODUCTION

Les transcrits de type « longs non-codants » (ARNlnc) font depuis quelques années l'objet de très nombreuses publications au sein de la communauté scientifique. L'avènement des nouvelles technologies de séquençage haut débit (RNA-seq) a permis l'identification de ces nouveaux ARN qui physiquement sont très semblables aux transcrits de type « messenger » (ARNm) [i.e. coiffés, polyadénylés et épissés pour la majorité] mais qui n'ont pas de potentiel codant, c'est-à-dire, qui ne codent pas des protéines. Les régions génomiques portant ces gènes ont longtemps été considérées comme de « l'ADN poubelle », sans aucune entité ou fonction biologique connue, mais à ce jour, le nombre de gènes de type « non-codant » dépassent le nombre de gènes de type « codant une protéine » chez l'Homme (Derrien et al., 2012). Cet engouement nouveau pour ce nouveau type d'ARN est dû au fait qu'ils sont associés, pour la majorité des ARNlnc décrits à ce jour, à des mécanismes de régulation de l'expression d'autres gènes (codant ou non) (Mercer et al., 2009 ; Schmitz et al., 2016). Les premiers catalogues d'ARNlnc humains et murins référencés sur GENCODE (Harrow et al., 2012) datent de 2009 et 2011 (Ørom et al., 2010 ; Derrien et al., 2012). Cette année, avec le nouvel assemblage du génome de la poule (Galgal5), on a vu apparaître les premiers ARNlnc chez la poule dans les bases de données publiques [i.e. en janvier sur le site bio-informatique américain NCBI ; en octobre sur le site européen Ensembl]. Il est important de souligner que ces ARNlnc sont prédits à partir de données expérimentales (séquençage des transcrits d'un tissu) sur la base de règles qui peuvent diverger d'un outil de prédiction à l'autre. Aussi, il peut y avoir de grandes différences entre ces deux catalogues d'ARNlnc prédits qui n'ont fait, à ce jour, l'objet d'aucune publication. Dans ce contexte, nous avons construit un catalogue d'ARNlnc (Muret et al., 2017) à partir de données de séquençage de transcrits de deux tissus (le foie et le tissu adipeux) en utilisant un outil de prédiction FEELnc qui est plus performant que ceux utilisés traditionnellement (Wucher et al., 2016). Nous avons comparé ces ARNlnc avec les bases de données publiques (NCBI et Ensembl) et les avons caractérisés en termes de structure (séquences, localisation génomique) et d'expression. Profitant d'un dispositif unique constitué de 16 répliquats biologiques par tissu, nous avons également étudié la co-expression entre couples de gènes adjacents codant un ARNlnc pour l'un et un ARNm pour l'autre. Nous avons ainsi identifié un couple fort intéressant impliquant un gène codant une enzyme du cholestérol (*DHCR24*) et un ARNlnc susceptible de partager un promoteur bidirectionnel avec le gène *DHCR24*.

1. MATERIELS ET METHODES

1.1. Collecte des échantillons, extraction et séquençage des ARN

16 poulets de chair ont été sacrifiés à l'âge de 9 semaines et leur foie et tissu adipeux abdominal ont été extraits, plongés dans de l'azote liquide et conservés à -80°C. Les ARN totaux ont été extraits selon les instructions du fabricant (RNeasy® Mini Kit, QIAGEN) à partir de 30 et 100 mg de foie et tissu adipeux. Ces ARN ont été quantifiés sur spectrophotomètre NanoDrop® ND-1000. 16 et 8 échantillons de foie et tissu adipeux ont été séquencés sur HiSeq 2000 (RNA-Seq ; lectures de 2x100 bases paires et orientées) et 8 échantillons supplémentaires de tissu adipeux ont été séquencés sur HiSeq 3000 (lectures 2x150 bases paires et orientées). 3,3 Md de lectures ont été obtenus sur les 32 échantillons (111 M et 92 M en moyenne pour les foies et les tissus adipeux).

1.2. Prétraitement informatique

3 Md de lectures ont été alignées sur le génome de référence Galgal4 via l'outil STAR (v2.4.0i) et les duplicats de PCR ont été supprimés pour chaque échantillon avec l'outil rmdup (v0.1.19). L'ensemble des données a été concaténé par l'outil merge (v1.1) puis la modélisation de nouveaux gènes a été effectuée via le logiciel Cufflinks (v2.2.1) en utilisant l'annotation des gènes d'Ensembl (v.82) comme référence. La quantification de l'expression des gènes a été faite à partir du logiciel featureCounts (v1.4.5-p1) sans considération des lectures dites « multi-alignées » sur les fichiers issus de l'alignement. 2,4 Md de lectures ont été considérées sur les 32 échantillons (88 M et 63 M en moyenne pour chaque échantillon de foie et de tissu adipeux).

1.3. Prédiction des ARNlnc

Les ARNlnc ont été prédits par FEELnc (v23.11.2015) (Wucher et al., 2016). Cet outil permet de générer un score de « potentiel codant » pour chaque ARN à partir de caractéristiques qui permettent de discriminer les ARN non-codants des ARN codant (i.e. taille d'ARN, couverture du cadre de lecture maximal, fréquence en « mots » de 1, 2, 3, 6, 9 et 12 lettres) à partir de deux sets d'entraînement : les annotations NONCODE (v.5) pour les ARNlnc et Ensembl (v.78) pour les ARNm. FEELnc permet également de classer les ARNlnc en fonction du gène codant le plus proche. Cet outil, récemment mis au point par T. Derrien, a montré de meilleures performances que les autres outils utilisés massivement depuis quelques années (cf. <http://biorxiv.org/content/early/2016/07/18/064436>).

1.4. Comparaison de nos ARNlnc avec les ARNlnc poulets des bases de données NCBI et Ensembl

Les ARNlnc prédits ont été recherchés dans les nouvelles annotations et génome de la poule de 2016 (i.e. ARNlnc et ARNm d'Ensembl v.86 et du NCBI v.103) via l'outil d'alignement local de séquences Blastn (v.2.4.0+). Il a été choisi de conserver les options par défaut et de ne garder que le meilleur résultat de chaque recherche.

1.5. Conservation des ARNlnc poulet chez l'Homme

La conservation de la séquence entre espèces a été testée contre les ARNlnc humains (GRCh38 ; v.83) avec l'outil Blastn (v.2.4.0+) en adaptant les options pour des séquences éloignées (i.e. taille de mot minimum (7), 50% de couverture de la requête et 70% d'identité).

2. RESULTATS ET DISCUSSION

2.1. Identification et classification des ARNlnc

L'annotation du génome du poulet (v.82) recense 17939 transcrits et 17105 gènes dont 15508 codent pour des protéines. La modélisation de nouveaux gènes à partir des séquences des transcrits des 16 foies et 16 tissus adipeux et l'annotation par l'outil FEELnc de ces nouveaux modèles a permis de définir 2193 nouveaux gènes de type « long non-codant » dont 1670 sont exprimés de façon robuste [i.e. exprimés dans au moins un des deux tissus et dans la majorité des échantillons (10 sur 16)]. Ces ARNlnc sont pour la plupart d'entre eux des ARNlnc de type « intergéniques » (89%), c'est-à-dire localisés entre des gènes connus. Cette classification est évidemment dépendante de l'annotation du génome (ici, v.82). Parmi ces ARNlnc intergéniques, on distingue 3 sous-classes d'ARNlnc définies par rapport au gène le plus proche sur le génome (Fig. 1) : les ARNlnc intergéniques de type « même brin » 48%, de type « convergent » 15% et de type « divergent » 26%. Les 11% d'ARNlnc restant sont des ARNlnc de type « géniques », c'est-à-dire qu'ils sont situés au même locus qu'un autre gène mais en position anti-sens.

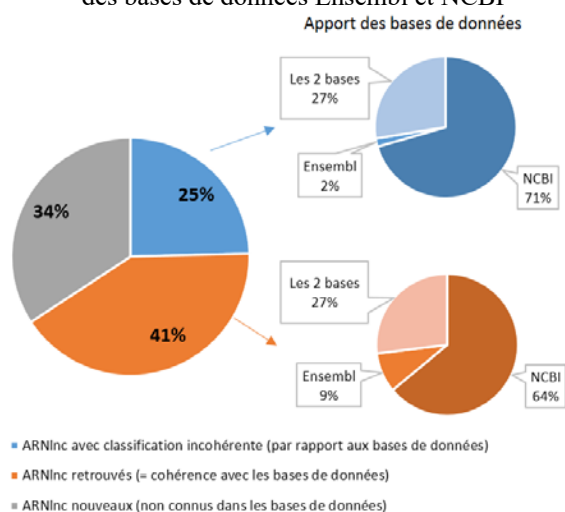
Figure 1. Classification et illustration des ARNlnc (●) en fonction des gènes codant proches (●).

ARNlnc intergéniques [89%]			ARNlnc géniques [11%]
Même brin	Convergent	Divergent	Anti-sens
			
801	255	437	177

2.2. Les ARNlnc en regard des nouveaux ARNlnc des bases de données NCBI et Ensembl.

Nous avons comparé notre catalogue d'ARNlnc avec les deux catalogues disponibles chez le poulet dans les deux bases de données internationales (les 8184 ARNlnc du NCBI - v.103 et les 7841 ARNlnc de Ensembl - v.86). 34% de nos ARNlnc ne sont pas présents dans ces catalogues. Concernant les 66% restant, nous montrons que deux tiers *versus* un tiers de nos ARNlnc correspondent à des ARNlnc *versus* ARNm référencés sur au moins une des bases de données (Fig. 2). La portion différentielle entre ces différents catalogues est due aux différentes stratégies d'annotation des ARN et il est impossible aujourd'hui de définir la part de vérité parmi ce set d'ARN.

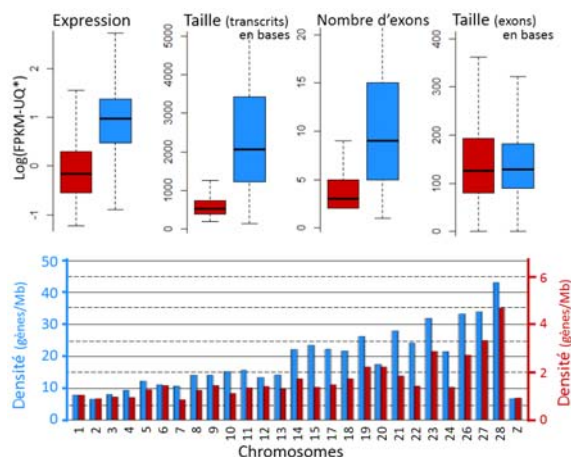
Figure 2. Comparaison de nos ARNlnc avec ceux des bases de données Ensembl et NCBI



2.3. Caractérisation des ARNlnc de la poule (Fig. 3)

Comme déjà observé chez l'Homme, nous montrons également que les gènes de type « long non-codant » ont une expression tissulaire plus faible que celle des gènes codant pour des protéines (en moyenne 10 fois moins) et ont une expression plus tissu spécifique (24% *versus* 3%). Au niveau structural, les ARNlnc sont plus courts et possèdent moins d'exons que les ARNm, alors que la taille des exons n'exerce pas. Les ARNlnc ne sont pas localisés préférentiellement sur certains chromosomes et leur densité chromosomique est corrélée à celle des ARNm ($r=0,91$).

Figure 3. Caractéristiques des ARNlnc (●) par rapport aux ARNm d'Ensembl v.85 (●). *FPKM-UQ est une quantification normalisée reflétant le niveau d'expression des gènes.



2.4. Conservation des ARNlnc

Comme présenté dans plusieurs publications sur la conservation des ARNlnc entre les espèces, nous confirmons que nos ARNlnc sont très peu conservés en séquence entre l'Homme et le poulet (2 ARNlnc d'après nos critères). En revanche, nous avons retrouvé 64% de nos ARNlnc entourés de 2 gènes orthologues et codants qui co-localisent de la même manière chez l'Homme (*i.e.* succession de trois gènes adjacents conservés au niveau positionnel dans les deux génomes). Cette conservation positionnelle des ARNlnc suggère un rôle potentiel de ces gènes « longs non-codants » à cette position donnée.

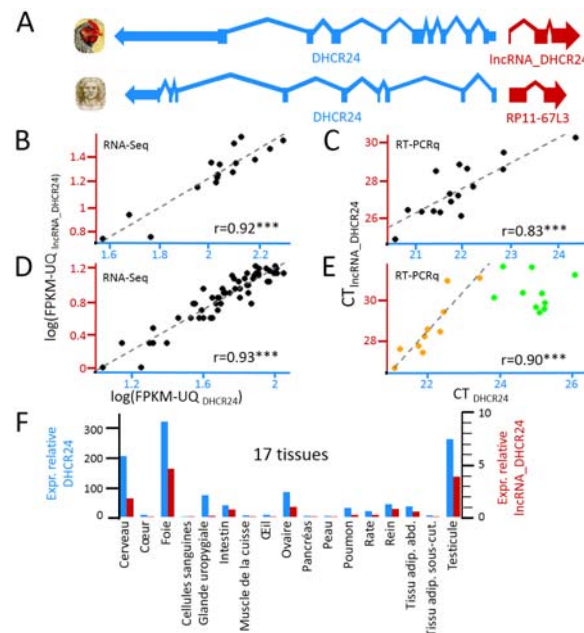
2.5. ARNlnc divergents et coexpression

Parmi les ARNlnc conservés, nous avons pu mettre en évidence que les ARNlnc de type « divergent » et proches sur le génome (<1000 bases) ont une expression plus fortement corrélée avec celle du gène codant le plus proche que les ARNlnc de type « divergent » distants (22% *versus* 13%; test de Fisher, p -valeur=3,7.10⁻²).

L'expression de l'ARNlnc « *lncRNA_DHCR24* », par exemple, est fortement corrélée ($r=0,92$) à celle du gène codant *DHCR24*, enzyme clé du métabolisme du cholestérol (Fig. 4b). Ces deux gènes sont très proches (environ 200 bases), et ils sont retrouvés dans cette même disposition chez la souris, chez l'Homme et d'autres espèces (Fig. 4a). Ce qui laisse penser que cette configuration n'est pas anodine et une étude plus approfondie de ce couple permettrait peut-être de montrer le potentiel rôle régulateur du *lncRNA_DHCR24* dans le métabolisme des lipides. Nous avons validé cette co-expression par une autre technique (la RT-PCRq, Fig. 4c) que le séquençage ($r=0,83$) et ce, dans différents dispositifs de poules ($r=0,93$ pour 50 animaux adultes de 31 semaines (Fig. 4d), $r=0,90$ pour de jeunes animaux de 9 semaines (Fig. 4e)). Cette corrélation est également vraie à travers les tissus (Fig. 4f).

Figure 4. Caractérisation du *lncRNA_DHCR24* (●) avec son gène le plus proche en position divergente et codant l'enzyme *DHCR24* (●). (A) Illustration des modèles de gènes chez la poule et chez l'Homme. (B)

et (C) expression des deux gènes dans les 16 foies de poule (RNA-Seq et validation en RT-PCRq). (D) Idem chez des animaux adultes (RNA-Seq). (E) Idem chez de jeunes animaux (RT-PCRq) avec deux régimes (● : nourri ; ● : à jeun). (F) Expression des deux gènes à travers différents tissus.



CONCLUSION

La modélisation et la classification des ARN à partir du séquençage des transcrits de 16 réplicats biologiques nous a permis de constituer et caractériser un premier répertoire de gènes « longs non-codants » du foie et du tissu adipeux chez le poulet. Leurs caractéristiques sont similaires à ce qui est à ce jour connu dans d'autres espèces. Des cas intéressants de gènes en position divergente dans le génome et avec des expressions corrélées ont été mis en évidence et mériteront de plus amples analyses. Ces premiers travaux sur une espèce d'élevage (Muret et al., 2017) s'inscrivent dans un programme de plus grande envergure FR-AgENCODE, porté par l'INRA, dont l'objectif est d'améliorer l'annotation fonctionnelle des génomes de quatre espèces d'élevage : le porc, le bovin, la chèvre et la poule.

REFERENCES BIBLIOGRAPHIQUES

- Derrien T., Johnson R., Bussotti G., Tanzer A., Djebali S., Tilgner H., et al. 2012. *Genome Res.*, (22), 1775-89
- Harrow J., Frankish A., Gonzalez J.M., Tapanari E., Diekhans M., Kokocinski F., et al. 2012. *Genome Res.* (22), 1760-74
- Mercer T.R., Dinger M.E., Mattick J.S.. 2009. *Nat. Rev. Genet.*, (10), 155-9
- Muret K., Klopp C., Wucher V., Esquerré D., Legeai F., Lecerf F., et al. 2017. *Genet. Sel. Evol.*, (49).
- Ørom U.A., Derrien T., Beringer M., Gumireddy K., Gardini A., Bussotti G., et al. 2010. *Cell*, (143), 46-58
- Schmitz S.U., Grote P., Herrmann B.G.. 2016. *Cell Mol Life Sci.* (73), 2491-509
- Wucher V., Legeai F., Hedan B., Rizk G., Lagoutte L., Leeb T., et al. 2016. *Nucleic Acids Res.*