Des longs ARN non codants humains activateurs de la transcription des gènes. [Long non-coding RNAs with enhancer-like function in human cells].

Thomas Derrien, Roderic Guigó

Des longs ARNs non codants humains activateurs de la transcription des gènes

(Long noncoding RNAs with enhancer-like function in human cells)

Thomas Derrien* et Roderic Guigó**

*Institut de génétique et Développement de Rennes 1, CNRS UMR6061, 2 av du Pr. Léon Bernard, Faculté de Médecine, Université de Rennes1, Rennes, France.
** Bioinformatics and genomics group, Center for Genomic Regulation, Barcelona, Catalonia, Spain.

Email: toma.derrien@gmail.com

Le transcriptome d'une cellule représente l'ensemble des molécules d'ARN codants et non codants pour des protéines d'une cellule ou d'un tissu. L'émergence des nouvelles techniques de séquençage haut débit permet à présent d'appréhender la complexité d'un transcriptome entier. Chez l'homme, ces études ont montré qu'une majeure partie de son génome était transcrite [1] et que la proportion et le rôle des ARN non codants (ncARN) pour des protéines était largement sous estimée [2].

La cellule, un monde ARN?

Il est admis que des mécanismes essentiels (voire "primitifs") de la machinerie cellulaire sont assurés par de multiples classes d'ARN non codants. Ceux-ci interviennent soit lors de la traduction des ARN messagers (ARNm) en protéines grâce aux ARN ribosomaux et aux ARN de transfert, pour le contrôle de l'épissage impliquant, entre autres, les small nuclear ARN ou encore, plus récemment mis en évidence, lors de l'inactivation spécifique de l'expression de certains gènes par les microARN (miARN) [3].
Les connaissances sur le nombre, la localisation et le rôle des long ARNs non codants (lncARN) restent partielles, contrairement à celles sur les ncARN de petites tailles. On définit arbitrairement les lncARN comme ayant une taille supérieure à 200 nucléotides et n'ayant pas de capacité à coder des protéines [10]. Cette définition est généralement affinée en fonction de la localisation du lncARN, c'est à dire intergénique ou bien chevauchant (sens ou anti-sens) un (ou des) gène(s) codant pour des protéines. De plus, la quantité de grands transcrits non codant annotés dans le génome humain est, certes en constante expansion, mais reste assez variable selon les bases de données (de 4 000 à plus de 10 000 [4]). Grâce à la mise à disposition croissante de séquences caractérisant des transcriptomes complets (*RNA-Seq*)[1], on peut anticiper que ces estimations sont bien loin du nombre réel chez l'homme. Par exemple, une étude exhaustive visant à séquencer plusieurs transcriptomes de souris a démontré que le nombre de lncARN murins se situait autour de 35 000 [5].
À l'instar des ~ 21 000 gènes humains codant pour des protéines pour lesquels les fonctions

[1] *RNA-Seq* : technique de séquençage haut-débit qui permet d'identifier et d'analyser quantitativement une population entière d'ARNs dans un échantillon.

biologiques ne sont pas encore toutes connues, le rôle spécifique du répertoire des lncARN reste à élucider. Cependant, des travaux pionniers sur quelques lncARN permettent de voir qu'ils sont impliqués dans une régulation de la transcription du génome [10], souvent par modifications de la structure de la chromatine. Ainsi, le lncARN *Xist,* d'une taille d'environ 19 kb, contrôle l'inactivation d'un des deux chromosomes X durant la différentiation précoce des cellules souches embryonnaire chez les placentaires femelles. Plus récemment, une classe de lncARN intergéniques a été mise en évidence chez la souris par une approche de caractérisation des patrons de méthylation de la chromatine renforçant le lien entre lncARN, épigénétique et niveau d'expression [6]. Un autre exemple de lncARN concerne *HotAIR* qui peut réprimer la transcription du locus HOXD en modifiant l'état de méthylation de la chromatine et donc l'expression de ce locus.

Identifier et caractériser les longs ARN non codants

Dans le but d'analyser à grande échelle la fonction des lncARN, nous avons utilisé l'annotation Gencode [4] du génome humain, produite dans le cadre du projet ENCODE (*Encyclopedia of DNA elements*) [1], qui vise à identifier tous les éléments fonctionnels du génome. Pour cela, une combinaison d'outils bioinformatiques, de données expérimentales et de vérifications réalisées par des experts en annotation a permis d'identifier les régions transcrites du génome humain par l'alignement de séquences exprimées totales (ADNc) ou partielles (EST, *expressed sequence tags*) ou de transcriptomes complets. Ensuite, leur définition est affinée en les classant par catégories fonctionnelles telles que les gènes codant pour des protéines, les gènes non codants, les pseudogènes ou encore les gènes polymorphiques (dont la séquence codante varie en fonction des polymorphismes entre individus).

En utilisant cette annotation, nous avons identifié 3 019 nouveaux transcrits non codants ayant une taille moyenne de 800 nucléotides et exclusivement localisés entre les régions de gènes codants pour des protéines [7]. Ces lncARN ont certaines caractéristiques similaires aux gènes codant des protéines puisqu'ils sont épissés et la moitié d'entre eux possèdent au moins un intron. Dans cette étude, seulement le tiers du génome couvert par l'annotation Gencode a pu être analysé, ce qui suppose, par projection, que le génome humain possèderait près de 10 000 longs ncARN intergéniques.

Contrairement aux petits ncARN (tels que les miARN), la conservation des séquences des lncARN mesurée grâce à des alignements multiples de 44 génomes de vertébrés [8] apparaît relativement faible ce qui rend leur identification par les approches de similarité de séquences (identification d'homologie) très délicate. Le faible niveau de conservation observé peut refléter une plus grande adaptabilité de ces séquences au cours de l'évolution par acquisition de mutations à l'opposé des fortes contraintes sélectives qui s'exercent sur les phases codantes des séquences codant pour des protéines. Néanmoins, la conservation de séquences des lncARN, et plus particulièrement celle de leurs promoteurs, est statistiquement plus forte que des régions du génome prises aléatoirement, ceci renforçant la probabilité d'un rôle fonctionnel de ces lncARN.

En utilisant des données de séquençage de 15 transcriptomes issus de 10 tissus et 5 lignées cellulaires [9], il s'avère que plus de 75 % des 3 019 lncARN sont retrouvés exprimés dans au moins un tissu et cette expression semble tissu-spécifique. Une caractéristique importante des lncARN concerne leur niveau d'expression qui est 10 à 20 fois plus faible que les transcrits

produits à partir des gènes codants pour des protéines. Ceci conforte la nécessité de séquençage en profondeur pour l'identification de transcrits faiblement exprimés.

Mise en évidence du rôle activateur de la transcription des longs ncARN.

La plupart des fonctions des lncARN décrites dans l'inactivation du chromosome X ou dans l'empreinte parentale (*imprinting*) ont mis en évidence leur rôle sur la répression de l'expression des gènes voisins codants pour des protéines. Afin de tester cette hypothèse, nous avons analysé la corrélation d'expression entre les lncARN et tous les gènes codants dans les 15 échantillons (provenant de tissus et cellules) mentionnés ci-dessus. De façon surprenante, nous avons trouvé que l'expression des lncARN est essentiellement corrélée positivement avec l'expression des gènes codants pour des protéines et que cette corrélation est plus forte avec des gènes voisins plutôt qu'avec des gènes distants sur le génome. Pour tester expérimentalement ces résultats, des petits ARN interférants (siARN) ciblant exclusivement les lncARN ont été utilisés. Un premier lncARN candidat (ncARN-a7) a été spécifiquement réprimé par l'utilisation de siARN. Nous avons ensuite évalué les conséquences de cette inactivation en mesurant le niveau d'expression des gènes situés dans une fenêtre de 1 Mb autour du lncARN *(Figure 1)*. Comme prédit au niveau bioinformatique, l'inactivation du lncARN entraîne une répression concomitante de l'expression d'un gène voisin *Snai1*, ce qui implique un rôle activateur du ncARN-a7 sur la transcription du gène *Snai1*, dont l'expression a un rôle dans l'adhésion et la migration cellulaire. Nous avons répété ces expériences d'interférence pour plusieurs lncARN candidats et suggérons que ce mécanisme d'activation de la transcription serait généralisable. Les gènes activés par les lncARN ont des fonctions variées. Parmi ces gènes, *ECM1* est impliqué dans la constitution de matrice extra-cellulaire, *KLHL12* est un régulateur négatif de la voie Wnt-beta caténine et *SCL* (ou *TAL1*) est un facteur important de la régulation de l'hématopoïèse.

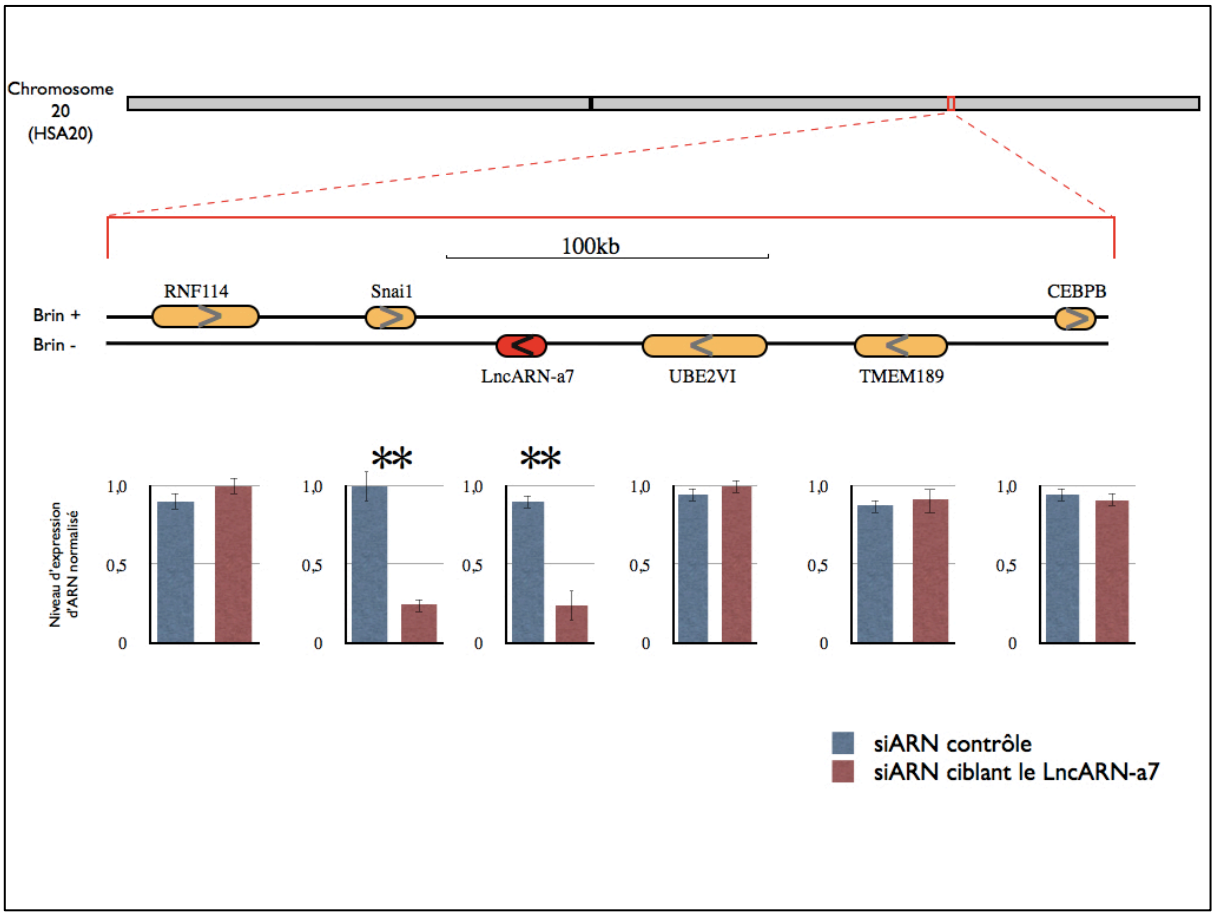

Figure1 : La région chromosomique du chromosome 20 humain (HSA20) aux environs de la position 48Mb est encadrée en rouge. Ce locus contient le long ARN non codant (LncARN activateur 7, ici en rouge) ainsi que cinq gènes codant pour des protéines (en orange). L'inactivation (*knock-out*) du lncARN activateur7 par siARN provoque la répression spécifique du gène *Snai1* dont la séquence codante est située en amont sur le brin opposé. L'expression des autres gènes n'est pas altérée. Six expériences indépendantes ont été réalisées ici ; ** $p < 0.01$ (test de Student).

Le génome humain possède donc une forte proportion de longs ARNs non codants pour des protéines. Une nouvelle classe de lncARN, certes modestement conservés et faiblement exprimés, régule l'activation de la transcription de gènes voisins au même titre que les régions activatrices de la transcription (*enhancer*). Cependant, à l'inverse des séquences *enhancer* qui activent la transcription par fixation de facteurs de transcription au niveau de l'ADN, c'est bien l'ARN qui est ici l'élément central de l'activation de la transcription même si le mécanisme d'activation reste à préciser. Une hypothèse intéressante serait que le lncARN servirait de plate-forme par similarité de séquence avec son gène cible pour favoriser le recrutement de protéines activatrices de la transcription. L'achèvement de l'annotation Gencode à la totalité du génome humain permettra d'identifier l'ensemble du répertoire des lncARN et ainsi préciser leurs rôles fonctionnels.

CONFLIT D'INTÉRÊTS

Les auteurs déclarent n'avoir aucun conflit d'intérêts concernant les données publiées dans cet article.

RÉFÉRENCES

1. ENCODE Projet Consortium, Birney E, Stamatoyannopoulos JA*, et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 2007 ; 447 : 799-816.
2. Amaral PP, Dinger ME, Mercer TR, Mattick JS. The eukaryotic genome as an RNA machine. *Science* 2008 ; 319 : 1787-9.
3. Lee RC, Feinbaum RL, Ambros V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* 1993 ; 75 : 843-54.
4. Harrow J, Denoeud F, Frankish A*, et al.* GENCODE: producing a reference annotation for ENCODE. *Genome Biol* 2006 ; 7 Suppl 1 : S4.1-9.
5. Carninci P, Kasukawa T, Katayama S*, et al.* The transcriptional landscape of the mammalian genome. *Science* 2005 ; 309 : 1559-63.
6. Guttman M, Amit I, Garber M*, et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, 2009 ; 458 : 223-7.
7. Ørom UA, Derrien T, Beringer M*, et al.* Long noncoding RNAs with enhancer-like function in human cells. *Cell*, 2010 ; 143 : 46-58.
8. Siepel A, Bejerano G, Pedersen JS*, et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 2005 ; 15 : 1034-50.
9. Wang ET, Sandberg R, Luo S*, et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* 2008 ; 456 : 470-6.
10. Pasmant E, Laurendeau I, Sabbagh A, *et al*. ANRIL ou l'étrange histoire d'un grand ARN non codant. *Med Sci (Paris)* 2010 ; 26 : 564-6.