

Chroniques génomiques

Remise en cause du « dogme central »

Bertrand Jordan



CoReBio PACA, case 901,
Parc scientifique de Luminy,
13288 Marseille Cedex 9,
France.
brjordan@club-internet.fr

L'émergence des « RDD »

On connaît depuis longtemps le phénomène du *RNA editing*, dans lequel interviennent des enzymes qui modifient après la transcription l'ARN messenger, notamment en transformant l'adénosine (A) en inosine (I), lue par les ribosomes comme la guanine (G), ce qui équivaut donc à une transition A → G. Important pour certains parasites, ce type d'événement était considéré comme marginal chez les mammifères et dans l'espèce humaine. Mais voici qu'un article paru début juillet dans la revue *Science* [1] nous indique que le phénomène serait très fréquent, qu'il provoquerait aussi de nombreuses transversions, et, en somme, que la quasi-identité entre la séquence codante de l'ADN et celle de l'ARN messenger ne tiendrait plus. Il y aurait ainsi dans notre génome plus de 10 000 sites pour lesquels existe une différence entre ADN et ARN, ce qui constitue une sérieuse mise en cause du « dogme central » énoncé en 1958 par Francis Crick [2] et déjà bien écorné par la découverte de la transcription inverse en 1970 !

Une étude approfondie

Il s'agit bien sûr d'une nouvelle retombée des grands travaux de séquençage en cours, notamment du *1000 genomes project*¹ dont les données ont été largement utilisées par les auteurs. Ils ont choisi de focaliser leur étude sur 27 personnes appartenant au panel du CEPH (centre d'étude du polymorphisme humain), dont l'ADN avait été étudié tant par cartographie dans le cadre

du projet HapMap² que, plus récemment, par séquençage intégral. Pour accéder à l'ARN correspondant, ils sont partis des lignées lymphocytaires B établies pour chaque individu, ont séquençé l'ARN messenger par la technique *RNA-seq* à l'aide d'une machine *Illumina*, et ont ensuite comparé ADN et ARN. Les différences observées, désignées dans la suite du texte par l'acronyme RDD (*RNA-DNA difference*) n'ont été prises en compte que si le site dans l'ADN avait été lu avec une haute fiabilité dans le *1000 genomes project*, et si aucun polymorphisme n'y avait été repéré³. De plus, le variant dans l'ARN devait apparaître chez au moins deux personnes et être présent à un niveau supérieur à 10 %. Même dans ces conditions assez draconiennes, on trouve plus de dix mille sites (exoniques) dans lesquels de telles différences apparaissent (*Tableau 1*).

Au total, les auteurs observent près de 30 000 événements de type RDD se produisant dans les 10 000 sites mentionnés, soit une moyenne de 1 000 environ par personne : le phénomène n'est vraiment pas marginal... Comme le montre l'exemple du *Tableau 1*, la même différence est observée chez plusieurs individus : il ne s'agit apparemment pas d'événements aléatoires. En fait, le nombre de différences varie fortement d'une personne à l'autre, allant de 300 à près de 2 000.

¹ <http://www.1000genomes.org/>

² <http://hapmap.ncbi.nlm.nih.gov/>

³ Dans dbSNP, la base de données des Snip, dans le projet HapMap et dans le *1000 genomes project*.



Gène	Chromosome	Différence	N info	N RDD	Fréquence (min/max en %)	dbEST ?
MYL6	12	C vers A	16	16	12/60	oui
RPL23	17	C vers T	12	8	10/22	oui

Tableau I. Deux exemples de RDD (RNA-DNA difference). N info : nombre de cas (individus) où l'on peut conclure à la présence ou non de RDD ; NRDD : nombre de RDD observés (on voit que pour le gène *MYL6* on observe une RDD pour chaque cas informatif) ; fréquence min/max : proportion de la forme variante observée (différente d'un individu à l'autre) ; dbEST ? : retrouve-t-on la forme variante dans les données (bien antérieures) des *expressed sequence tags* ? (données extraites du tableau 1 de [1]).

Gène	Chromosome	Différence	Cellules B	Fibroblastes	Cerveau
EIF2AK2	2	A vers G	8/8	0/8	10/10
DPP7	9	C vers T	2/9	1/8	0/10

Tableau II. Exemples de spécificité tissulaire de RDD. Ces spécificités ont été établies à partir des informations de dbEST. Pour chaque tissu on indique le nombre de RDD observés par rapport au nombre de cas informatifs (données extraites du tableau 2 de [1]).

La forme variante (éditée) de l'ARN messenger est presque toujours minoritaire par rapport à la forme normale et souvent présente à un niveau inférieur à 20 % - qui monte néanmoins jusqu'à 100 % dans quelques cas. Les sites concernés (tous exoniques) sont pour moitié dans les séquences codantes, le reste tombant principalement dans la région 3' non traduite.

Enfin, la majorité des différences repérées dans les régions codantes modifient la séquence protéique, introduisant dans quelques cas un codon stop. Il s'agit donc d'un phénomène majeur, créant chez chaque personne un nombre important de protéines variantes et augmentant de manière jusqu'ici insoupçonnée la diversité de notre protéome.

Est-ce bien vrai ?

C'est un phénomène majeur - s'il est réellement avéré, bien sûr. Quels arguments apportent les auteurs pour affirmer la validité de leurs données ? Les preuves sont de différentes natures. En dehors des critères déjà mentionnés concernant la qualité des séquences et le caractère monomorphique des sites dans l'ADN, et de la vérification ultérieure des séquences par la technique classique de Sanger, une confirmation essentielle est fournie par la consultation de la base de données dbEST⁴.

Le lecteur se souvient peut-être de la grande mode des EST, parfois baptisés chez nous « étiquettes ». Dans les années 1990, et en attendant le séquençage intégral de notre génome, alors à ses débuts, ces courtes séquences obtenues par lecture systématique de clones d'ADNc⁵ devaient donner à peu de frais une idée de l'ensemble des gènes humains [3]. Il en reste une base de données contenant plus de huit millions d'EST humains, aperçu partiel et

biaisé mais néanmoins précieux de notre transcriptome. Eh bien, plus de 60 % des variants RDD découverts par Li *et al.* [1] sont présents dans dbEST, dans ces séquences « historiques » obtenues par de tout autres techniques : ce ne sont donc certainement pas des artefacts liés aux nouvelles méthodes de séquençage. La consultation de dbEST permet aussi d'avoir une idée de la spécificité cellulaire des RDD et montre que certains ne sont présents que dans certains tissus (Tableau II).

Plus fort encore, les auteurs ont effectué une étude protéomique par chromatographie couplée à la spectrométrie de masse (LC-MS/MS, *liquid chromatography coupled to tandem mass spectrometry*) et ont retrouvé des peptides correspondant aux sites RDD, le plus souvent sous la forme « non éditée » (correspondant à la séquence de l'ADN) mais parfois aussi sous la forme éditée, confirmant donc que cet ARN modifié est bel et bien traduit en protéine. Certes, le nombre de peptides édités ainsi repérés est faible, 28, mais on sait que les analyses protéomiques ne sont pas exhaustives, d'autant plus que la forme éditée de chaque ARN est généralement peu abondante (Tableau I). Les conclusions des auteurs, le fait que l'on trouve chez chaque personne un millier d'ARNm édités ayant pour conséquence la présence (à un niveau souvent faible) d'une version modifiée de plusieurs centaines de protéines, semblent donc solidement établies. Évidemment, si ils se confirment, ces résultats introduisent un niveau supplémentaire de complexité dans l'interprétation de notre génome : il est difficile d'imaginer qu'aucune de ces formes modifiées n'ait une importance physiologique.

⁴ <http://www.ncbi.nlm.nih.gov/dbEST/>

⁵ Et dont l'interprétation avait suggéré que nous aurions de 100 à 200 000 gènes...

Décidément...

Une fois de plus, la nature s'avère plus complexe que nous ne l'imaginions, et les nouvelles techniques d'analyse à grande échelle révèlent des phénomènes insoupçonnés. Il reste bien sûr à apprécier l'importance réelle, la signification fonctionnelle de ces protéines modifiées, mais il serait assez étonnant que leur présence n'ait aucune influence sur le phénotype de leurs porteurs. Reste une question : quelle sera la prochaine surprise ? ♦

Chipping away at the « Central Dogma »

CONFLIT D'INTÉRÊTS

L'auteur déclare n'avoir aucun conflit d'intérêts concernant les données publiées dans cet article.

RÉFÉRENCES

1. Li M, Wang IX, Li Y, et al. Widespread RNA and DNA sequence differences in the human transcriptome. *Science* 2011 ; 333 : 53-8
2. Crick FHC. On protein synthesis. *Symp Soc Exp Biol XII* 1958 : 139-63.
3. Jordan BR. Chroniques génomiques. Le festival des ADNc. *Med Sci (Paris)* 1993 ; 9 : 211-6.

TIRÉS À PART

B. Jordan

GenoSplice est une société de biotechnologie spécialisée dans l'**analyse de données transcriptomiques** (gènes, exons, microARN...). Issue du Réseau Européen d'Excellence Eurasnet, GenoSplice s'est imposé comme leader sur son secteur.

Travaillant à l'international (nos produits sont utilisés dans plus de cinquante pays), l'activité de GenoSplice s'est tout d'abord centrée sur la communauté scientifique académique avec la signature de contrats de **partenariat, notamment avec l'Institut Curie**.

La très haute valeur ajoutée du modèle de GenoSplice, unique dans ce domaine, réside dans la convergence de compétences et savoir-faire à la fois en biologie, en bioinformatique et en informatique. Ce modèle permet à la société de proposer des **solutions innovantes et performantes sur-mesure** à ses partenaires.



Par ailleurs, l'explosion des nouvelles techniques dans le domaine de la génomique entraîne une démultiplication pharaonique du nombre et de la complexité des données à analyser. C'est notamment le cas pour les dernières technologies de séquençage à très haut-débit. Dans ce contexte, une part importante de l'activité de R&D de GenoSplice se concentre sur le développement de solutions d'analyses de plus en plus innovantes, notamment dédiées au **séquençage à très haut-débit**.

Forte du savoir-faire de son équipe et de ses technologies de pointe, la société a été **lauréate 2008 et en 2011 du Concours de création d'entreprises innovantes**. Elle bénéficie du soutien financier d'Oséo et a noué de nombreuses collaborations avec des acteurs académiques et industriels à l'international (Howard Hugues Institute, St-Jude Hospital, Universités du Kentucky, de Cambridge...).

www.genosplice.com - contact@genosplice.com - +33 (1) 572 768 39