Caractérisation de l’expression des éléments Alu et du phénomène d’édition de l’ARN chez l’humain et la souris

Pierre Cattenoz

UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE DES SCIENCES DE LA VIE ET DE LA SANTE

UPR 9002 : Architecture et Réactivité de L'ARN

THÈSE présentée par :

Pierre CATTENOZ

Soutenue le : **5 juin 2012**

Pour obtenir le grade de : **Docteur de l'université de Strasbourg**

Discipline/ Spécialité : Biologie moléculaire

Caractérisation de l'expression des éléments Alu et du phénomène d'édition de l'ARN chez l'humain et la souris.

THÈSE dirigée par :

Mr. WESTHOF Eric	Professeur, Université de Strasbourg
Mr. MATTICK John	Professeur, University of Queensland

RAPPORTEURS :

Mme BRANLANT Christianne	Docteur, Université de Nancy
Mr. FILIPOWICZ Witold	Docteur, Friedrich Miesher Institute
Mme FRUGIER Magali	Docteur, Université de Strasbourg

UNIVERSITÉ DE STRASBOURG
ÉCOLE DOCTORALE DES SCIENCES DE LA VIE ET DE LA SANTÉ

THESIS

For the degree of

DOCTEUR DE L'UNIVERSITÉ DE STRASBOURG

Discipline: Life Science
Speciality: Molecular Biology

Presented by

Pierre B. CATTENOZ

The 16th of may 2012

Title:

Characterization of Alu element expression and A-to-I RNA editing in mammals

Directors: **Pr. John MATTICK** (IMB, University of Queensland)
Pr. Eric WESTHOF (IBMC, Université de Strasbourg)

Jury: **Dr. Christiane BRANLANT**
Dr. Witold FILIPOWICZ
Dr. Magali FRUGIER

This thesis was carried out as a co-tutelle agreement between the University of Queensland and the University of Strasbourg

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

UNIVERSITÉ DE STRASBOURG

Acknowledgment

I wouldn't have reached the end of this journey without the assistance of many people.

First, I wish to acknowledge my two supervisors Pr. John Mattick and Pr. Eric Westhof. John and Eric, thank you for initiating this project, rendering it possible and putting me back on track when I followed wrong directions. You allowed me a degree of freedom for my experiments I never had before and the amount I have learnt surpasses by far the frustration. I am grateful to you; it was really inspiring to work with you.

Then, to the people from the Mattick lab, especially Selene Fernandez-Valverde, Kelin Ru, Martin Smith, Paulo Amaral, Dennis Gascoigne, Darren Korbie, Seth Cheetham, Mike Clark, Joanna Crawford, Ryan Taft, Marcel Dinger, Harald Oey, Larry Croft and Martin Hansen, thank you for making the Mattick lab such a stimulating and welcoming place to experiment, I loved working and discussing with you. I will always keep you dearly in my mind. And to the bioinformatician of you, thank you for teaching me how to talk to a computer, the transfer from the bench to the keyboard was tough but now "awk" is my new friend.

To the people from the IBMC, Agnes, Stephanie, Beatrice, Liza, Benoit, Melanie, Christine and Valerie, I was not here as often as I expected, but you rendered each time really enjoyable. Thank you very much.

To other friends, Fabien, Marco, Julie, Wilko, Vikram and Co., you participated to this trip in your own way; it was great to know you and I hope we'll keep in touch.

At last, I wish to thank my soul mate Johana Chicher. Johana, thank you very much for bearing me for that long, particularly these past four years, for encouraging me in this adventure, for organizing my social and cultural life, for kicking my butt when I was out of it and more importantly, for having faith in me. With this thesis, I am achieving eleven years as a student and you accompanied me for most of that time. Now, I cannot tell you "quand on sera grand…" anymore, this time is here and I hope you'll accompany me on this road too.

Thanks a lot mates!

Index

List of Figures

Figure of the French abstract:

Figure of the thesis

List of tables

Abbreviation

5-HT2C	5-HydroxyTryptamine (serotonin) receptor 2C
A	adenosine
A/G	Adenosine to Guanosine
A:C	Adenosine:Cytosine
Ab	antibody
Abcb4	ATP-binding cassette, sub-family B (MDR/TAP), member 4
ADAR	Adenosine DeAminase RNA-specific
ADAT	Adenosine Deaminase Acting on tRNA
ADN	Acide DesoxyriboNucleique
AluJ	Alu family J
AluS	Alu family S
AluY	Alu family Y
APOBEC3G	APOlipoprotein B mRNA Editing enzyme Catalytic polypeptide-like 3G
APS	Ammonium PerSulfate
ARN	Acide RiboNucleique
A-to-C	Adenine to Cytidine
A-to-G	Adenine to Guanine
A-to-I	Adenosine to Inosine
A-to-T	Adenine to Thymine
ATP	Adenosine triphosphate
ATP5E	ATP synthase H+ transporting mitochondrial F1 complex epsilon subunit
AT-rich	Adenine and Thymine rich
BC1	Brain Cytoplasmic 1
BC200	Brain Cytoplasmic RNA of 200nt
BDP1	B Double Prime 1 subunit of RNA polymerase III transcription initiation factor
BRF1	TATA box binding protein (TBP)-associated factor RNA polymerase III GTF3B subunit 2
B-RNA	RNA extracted from the beads in the glyoxal protocol
BSA	Bovine serum albumin
BWA	Burrows-Wheeler Aligner
C	Cytosine
cAMP	cyclic Adenosine MonoPhosphate
CAR	Chromatin Associated RNA
CAT2	Cationic Amino acid Transporter 2
cDNA	coding DNA
CDS	CoDing Sequence
cedar	Caenorhabditis elegans Adenosine Deaminase RNA-specific
CHIPseq	CHromatin ImmunoPrecipitation deep sequencing
co-IP	co-ImmunoPrecipitation
CREB	cAMP response element-binding
CTN-RNA	Cationic amino acid transporter 2 Transcribed Nuclear RNA
C-to-U	Cytosine to Uracil
dADAR	Drosophila Adenosine DeAminase RNA-specific

DEPC	Diethyl pyrocarbonate
Dlgap4	discs, large (Drosophila) homolog-associated protein 4
DMEM	Dulbecco's Modified Eagle Medium
DMSO	DiMethyl SulfOxyde
DNA	DesoxyriboNucleic Acid
dsRNA	double stranded RNA
DTT	DiThioThreitol
ECL	Enhanced ChemiLuminescence
EDTA	Ethylenediaminetetraacetic acid
eIF4A	eukaryotic Initiation Factor 4A
EST	Expressed Sequence Tag
EtOH	Ethanol
FAM	Fossil Alu Monomer
FLAM	Free Left Alu Monomer
FMRP	fragile-X mental retardation protein
FRAM	Free Right Alu Monomer
G	Guanosine
GABRA3	gamma-aminobutyric acid (GABA) A receptor alpha 3
GAPDH	GlycerAldehyde-3-Phosphate DeHydrogenase
GC-rich	Cytosine and Guanine rich
gDNA	genomic DNA
GRIA2	Glutamate Receptor Ionotropic AMPA 2
GTP	Guanosine TriPhosphate
H3	Histone cluster 3
HBSS	Hank's Buffered Salt Solution
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
hg19	human genome version 19
HRP	HorseRadish Peroxydase
I	Inosine
Ig	Immunoglobulin
IgE	Immunoglobulin E
IgG	Immunoglobulin G
IP3	Inositol 1,4,5-triPhosphate
IP6	inositol hexakisphophate
I-RNA	RNA cleaved from the beads in the glyoxal protocol and containing an inosine in 3'end
kb	kilobase
KCNA1	potassium voltage-gated channel shaker-related subfamily member 1
Kd	dissociation constant
LINE	Long Interspersed Element
mGluRA	mouse Glutamate Receptor ionotropic AMPA 1
miRNA	micro RNA
MM	MisMatche
MOPS	3-(N-morpholino)propanesulfonic acid
mRNA	messenger RNA

Mya	Million years ago
NP-40	Nonyl Phenoxypolyethoxylethanol
nt	nucleotide
ORF	Open Reading Frame
PABP	Poly-A Binding Protein
PAGE	Polyacrylamide gel electrophoresis
PBS	Phosphate buffer saline
PCR	Polymerase Chain Reaction
Pin1	Peptidylprolyl cis/trans Isomerase NIMA-interacting 1
PNK	PolyNucleotide Kinase
POLII	RNA polymerase II
POLIII	RNA polymerase III
poly-A	oligomer of Adenine
poly-T	oligomer of thymine
PPIA	PeptidylProlyl Isomerase A (cyclophilin A)
pre-miRNA	precursor miRNA
pre-mRNA	precurssor of the messenger RNA
pri-miRNA	primary miRNA
PTPN6	Protein Tyrosine Phosphatase Non-receptor type 6
Q	glutamine
qPCR	quantitative Polymerase Chain Reaction
R	arginine
RFLP	Restriction Fragment Length Polymorphism
RISC	RNA-induced silencing complex
RNA	RiboNucleic Acid
RNAi	RNA interference
RNAseT1	Ribonuclease T1
RNP	RiboNucleoProtein complex
RPKM	read per kilobase per million of tags
RPLP0	Ribosomal Protein Large P0
rRNA	ribosomal RNA
RT	Room Temperature
RT-PCR	Reverse-Transcription Polymerase Chain Reaction
scAlu	small cytoplasmic Alu element
ScaRNA	Cajal body-specific RNA
SDS	Sodium dodecyl sulfate
SDS-PAGE	Sodium Dodecyl Sulfate PolyAcrylamide Gel Electrophoresis
SINE	Short INterspersed Element
snoRNA	small nucleolar RNA
SNP	Single Nucleotide Polymorphisme
snRNA	small nuclear RNA
Snupn	snurportin 1
SRP	Signal Recognition Particle
ssRNA	single stranded RNA
SUZ12	suppressor of zeste 12 homolog

TAP	Tobacco Acid Pyrophosphatase
Tapbp	TAP binding protein (tapasin)
TBE	Tris Borate EDTA buffer
TBS	Tris Buffer Saline
TEMED	Tetramethylethylenediamine
Tox4	TOX high mobility group box family member 4
Tris	Trishydroxyméthylaminométhane
tRNA	transfert RNA
$tRNA^{Ala}$	Transfert RNA Alanine
Tudor-SN	Tudor Staphylococcal Nuclease
U	Uracil
UCSC	University of California Santa Cruz
UTR	UnTranslated Region
UV	ultraviolet
WWP2	ubiquitin ligase WW domain containing Protein 2

Abstract

The Short INterspersed Element (SINE) Alus are the most prolific retrotransposons in human. With over one million copies, they comprise more than 10% of the human genome and these non-autonomous retrotransposons spread by hijacking the transposition machinery of the autonomous retrotransposon Long Interspersed Element (LINE1). As a response to retrotransposon invasion, organisms developed mechanisms to preserve the integrity of their genome, and one of them is RNA editing. RNA editing is the modification of one or more bases of an RNA molecule. The most abundant type of editing in mammals is A-to-I editing where the ADAR family transforms adenosine into inosine. The main targets of the ADARs in human are the Alu elements. Editing of the Alu elements targets them to the paraspeckles, thus lead to their sequestration in the nucleus which prevents their interaction with the transposition machinery of LINE1. Editing of Alu elements also mutates their internal POLIII promoter and their poly-A tail, thus preventing their subsequent transposition.

The current view on Alu elements is that they are mainly dormant occupants of the genome. In the first part of this study, we challenge this view by characterizing their activity. After demonstrating that Alu element transcripts can be precisely identified on a large scale with current deep-sequencing technology, the primary sequences of Alu elements are screened for active internal RNA polymerase III promoter by screening POLIII-CHIPseq data. The length and identity of the Alu transcripts are then determined in the cytoplasm and nucleus of cell as well as their association with polysomes and chromatin by screening deep-sequencing data performed on each one of these cell compartments. Analysis of a transcriptome Atlas of 16 human tissues reveals that Alu elements transcription is a widespread phenomenon in normal tissues which correlates with functional LINE1 elements expression. This suggests that Alu element retrotransposition may be a natural mechanism in most normal human tissues. Further analyses show that SINE and LINE expression in somatic tissues is not exclusive to human but also occurs in mouse. Finally, attempts are made to identify tissue specific insertions in the human genome resulting from retrotransposition events.

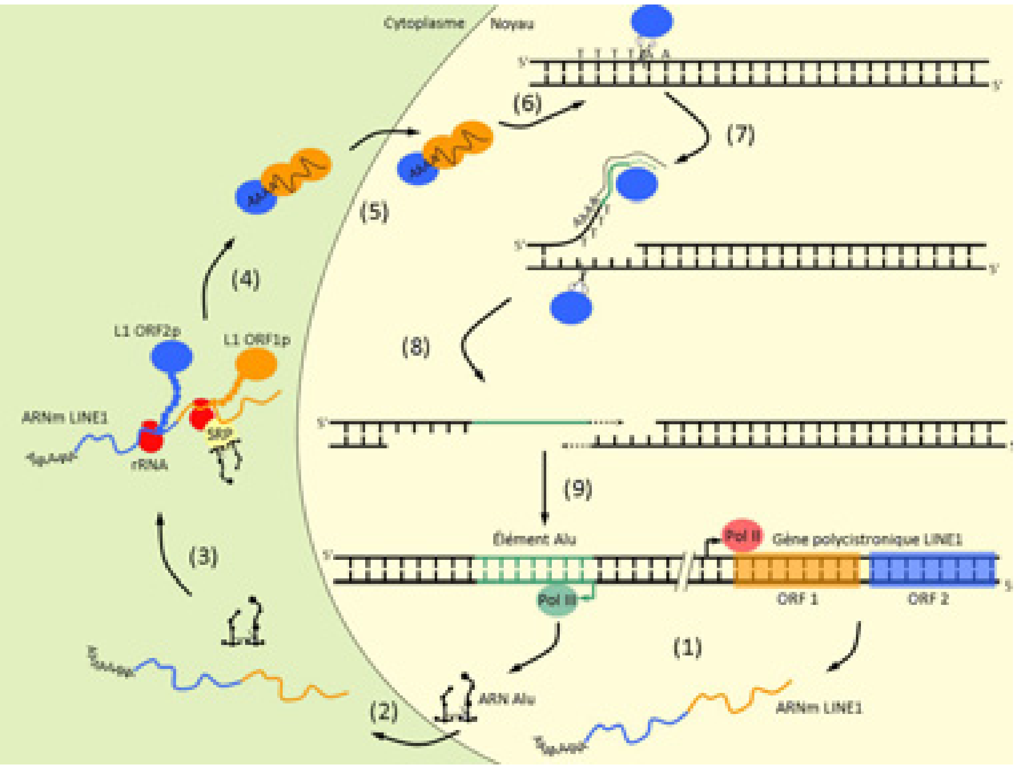
In the second part of this study, a new method is developed to understand the full impact of RNA editing on Alu transcripts and more broadly on whole transcriptomes by characterizing the edited RNA in a high-throughput fashion. In a first unsuccessful attempt, immunoprecipitation was used to pull-down RNA associated with the editing enzymes ADARs. Further fruitless attempts were then made to pre-purify the complex RNA-ADAR by

nuclear fractionation or sucrose gradient before immunoprecipitation. Finally, instead of using an antibody-based approach targeting the ADAR proteins, a protocol targeting directly the inosine in the RNA molecule was developed. First, the RNA is sequestered on magnetic beads. Then an inosine specific cleavage based on RNAseT1 treatment of RNA protected with glyoxal and borate allows the separation of the edited RNA from the total RNA. Finally, deep sequencing is used to identify edited RNA. 1,822 editing sites are found by this method including 28 new editing sites modifying the coding sequences of genes and editing in rRNA, snoRNA and snRNA which were never observed before.

Résumé de thèse (extended summary, in French)

Les SINEs (Short Interspersed Element) Alu représentent la famille de rétrotransposons la plus prolifique chez l'Homme. Avec plus d'un million de copies, ils occupent environ 10% du génome humain [1, 2]. Ces rétrotransposons non-autonomes se sont répandus dans le génome en utilisant la machinerie de rétrotransposition de LINE1 (Long INterspersed Element 1), un rétrotransposon autonome [3, 4] (Figure 1'). Afin de contrecarrer l'expansion des rétro-éléments, les organismes ont développés différents mécanismes pour préserver l'intégrité de leurs génomes. Le plus proéminent, également utilisé pour lutter contre la réinsertion d'ADN viral dans le génome hôte, est l'édition de l'ARN. L'édition de l'ARN est la modification d'une ou plusieurs bases d'une molécule d'ARN. Chez les mammifères, la plus courante est la déamination de l'adénine en inosine catalysée par la famille de protéine ADAR. L'inosine est ensuite lue comme une guanine par les différentes machineries enzymatiques (traduction, reverse transcription) [5]. Si le phénomène d'édition de l'ARN a un impact clair sur l'ARN messager où il modifie les sites d'épissages [6, 7] ou directement la séquence de la future protéine [8], les principales cibles d'ADAR chez l'humain sont les éléments Alu [9]. L'édition des éléments Alu conduit à leur séquestration dans le noyau des cellules [10], empêchant ainsi leur interaction avec la machinerie de transposition de LINE1. L'édition des éléments Alu mute également leurs promoteurs internes, cible de l'ARN polymérase III (POLIII), et leurs queues poly-A, prévenant ainsi leur future rétrotransposition [11].

Figure 1' : Mécanisme de rétrotransposition des éléments Alu par "target-primed reverse transcription". (1) L'élément Alu est transcrit en petit ARN de 300nt depuis son promoteur reconnu par POLIII et LINE1 est transcrit par l'ARN polymérase II. (2) Les transcrits LINE1 et Alu sont exportés vers le cytoplasme. (3) L'ARN polycistronique de LINE1 est traduit en ORF1p qui est une protéine chaperonne et ORF2p qui est une endonuclease et une reverse-transcriptase. Le transcrit Alu interagit avec SRP à proximité des protéines naissantes de LINE1. (4) Le transcrit Alu détourne ORF1p et ORF2p pour former un complexe qui migre vers le noyau (5). (6) ORF2p coupe l'ADN génomique à un locus contenant le motif TTTTAA. (7) la queue poly-A de l'ARN Alu se lie à l'extrémité poly-T de l'ADN génomique et l'ARN Alu est utilisé comme modèle par ORF2p. (8) ORF2p coupe le second brin d'ADN génomique et (9) l'ADN est réparé en incluant la nouvelle insertion de l'élément Alu. Les mécanismes exacts des parties 4, 5, 8 et 9 n'ont pas été validés.



Les éléments Alu sont exprimés dans la plupart des tissues somatiques chez l'humain et la souris.

A cause de leur nature répétitive, les éléments Alu sont systématiquement exclus de la plupart des analyses transcriptomiques et sont difficilement identifiables par les méthodes de biologie moléculaire basées sur l'hybridation de sondes (Northern-blot, PCR…). De ce fait, les propriétés de ces éléments sont très peu connues. La première partie de cette étude est une caractérisation précise de l'expression des éléments Alu réalisée en utilisant le nombre croissant de données de séquençage haut-débit disponibles.

Dans un premier temps, afin de déterminer si les éléments Alu peuvent être caractérisés par séquençage haut-débit, leurs séquences ont été comparées systématiquement. Cette analyse a révélé que 96.6% des éléments Alu sont uniques dans le génome humain et que 80% des éléments contiennent un segment poly-A de plus de 6 nucléotides. Plus de 90% d'entre eux peuvent être précisément identifiés par la plupart des séquenceurs haut-débit de dernière génération et les paramètres utilisés communément lors de l'analyse des données. Le segment poly-A signifie que la plupart des transcrits Alu peuvent être présents dans les banques de données transcriptomiques construites a partir d'ARN poly-adénylé.

Les éléments Alu ont ensuite été annotés en fonction de (i) leur association avec POLIII, (ii) leur localisation dans la cellule et de (iii) leur profil d'expression dans seize tissus humains. (i) Pour déterminer quels éléments Alu ont conservé un promoteur reconnu par POLIII, des données de CHIPseq de POLIII produites à partir de cellules IMR90hTert ont été analysées. 455951 loci Alu ont ainsi été trouvés associés avec POLIII dans cette lignée cellulaire. Contrairement à de précédentes études, la famille d'éléments Alu la plus représentée est la plus ancienne, AluJ, suivit par AluS et enfin AluY. (ii) Pour identifier les éléments Alu localisés dans les différents compartiments cellulaires, des données de séquençage de fractions polysomales, cytoplasmiques et nucléaires de cellules DLD-1 ont été utilisées. 42964 éléments Alu ont été identifiés dans un ou plusieurs compartiments. L'analyse de deux banques de données supplémentaires de petits ARN (<500nt) associés à la chromatine a permis d'identifier 57038 et 27300 éléments Alu associés à la chromatine dans les cellules HeLa et fibroblastes respectivement. (iii) Enfin, l'Atlas de transcriptomes humains généré par Illumina pour seize tissues a été utilisé pour identifier spécifiquement le répertoire d'éléments Alu exprimés dans chaque tissue humain. Apparemment, tous les tissus observés expriment des éléments Alu alors qu'ils sont connus pour être fortement mutagènes et sont supposés être réprimés dans les tissues somatiques.

Suivant l'annotation des éléments Alu et la découverte de leur expression dans la plupart des tissus humains, l'Atlas de transcriptomes a été utilisé pour explorer l'hypothèse que ces éléments rétrotransposent dans les tissues normaux. Etant non-autonome, la retrotransposition des éléments Alu est dépendante de l'expression des protéines chaperonnes et de la reverse-transcriptase/endonuclease du retrotransposons autonome LINE1 (Figure 1'). Seulement 5500 éléments LINE1 complets persistent dans le génome humain et approximativement 100 d'entres eux sont capable d'exprimer leurs ORFs et de rétrotransposer. L'expression de ces 5500 éléments a été suivie dans les 16 tissues de l'Atlas. L'organe exprimant la plus grande quantité de LINE1 est le cerveau, suivi par la glande adrénale, les ovaires, le cœur et les testicules. Il est important de noter que le cerveau, les glandes adrénales, les ovaires et les testicules expriment tous un niveau significatif d'éléments LINE1 et d'éléments Alu, suggérant fortement que ces éléments rétrotransposent activement dans ces tissues.

Afin de voir si l'expression de ces éléments dans les tissus somatiques est limitée à l'humain, neuf transcriptomes de souris ont été investigués. Le génome de la souris contient 564000 éléments B1 qui, comme les éléments Alu, sont dérivés de l'ARN 7SL, 348000 éléments B2 dérivés des ARN de transfert et qui présentent des propriétés similaires aux éléments Alu, et 599000 éléments LINE1. N'ayant pas accès à un atlas de transcriptomes pour la souris, le Northern blot a été utilisé pour caractériser la présence de transcrits de LINE1, B1 et B2 dans les glandes adrénales, le cerveau, le cœur, les reins, le foie, les poumons, les muscles squelettiques, les testicules et le thymus de souris. Comme pour l'humain, LINE1, B1 et B2 sont exprimés à différents niveaux dans tous les tissues observés.

L'ensemble des ces résultats suggère fortement que la rétrotransposition est un mécanisme utilisé dans la plupart des tissus somatiques et dans les gonades, chez l'humain et chez la souris. Etant donné que la rétrotransposition est fortement régulée par le phénomène d'édition de l'ARN, nous avons cherché, dans la deuxième partie de cette étude, à caractériser les ARN ciblés par ADAR.

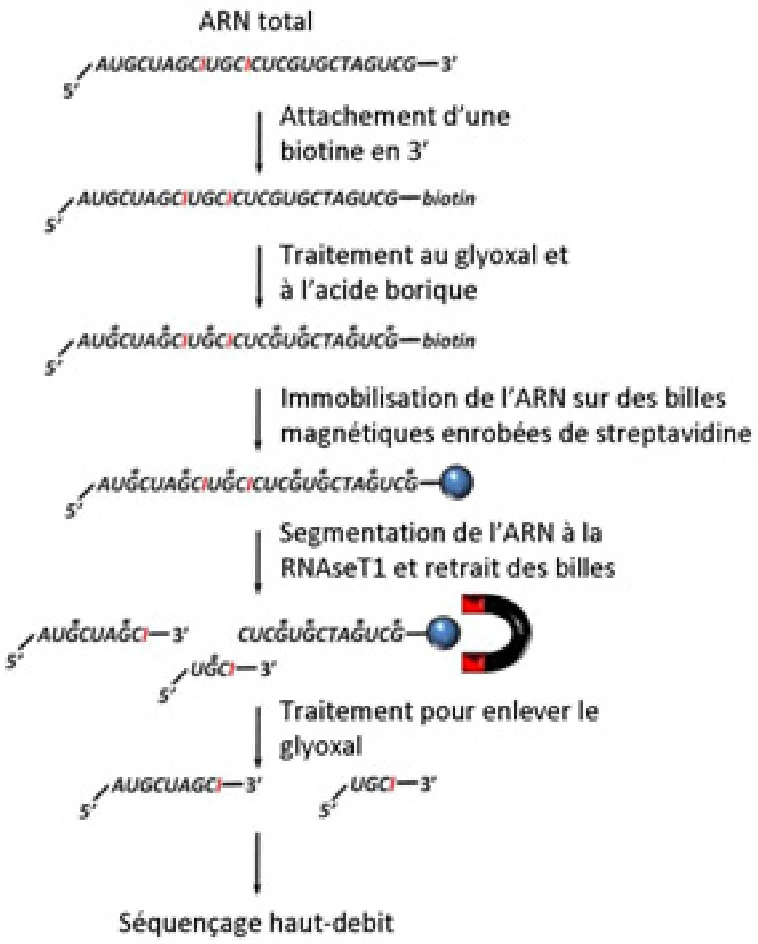
Développement d'une méthode pour identifier les ARN édités par ADAR

Caractériser les sites d'édition dans un transcrit est une étape primordiale pour comprendre la fonction et la régulation du transcrit. La découverte de nouveau sites d'édition s'est faite d'abord aléatoirement quand une mutation A/G était systématiquement observée en comparant la séquence d'un transcrit avec la séquence génomique correspondante. Ceci a ensuite été réalisé à haut-débit en comparant *in silico* des banques d'EST (Expressed Sequence Tags) à la séquence génomique de l'espèce concernée, révélant qu'il est très difficile de discerner les sites d'éditions d'erreurs de séquençage ou de SNP (Single Nucleotide Polymorphisme) et qu'il est nécessaire d'avoir de très larges banques de données, généralement construites a partir de plusieurs individus. Des méthodes expérimentales ont aussi été développées. Ohlson et al. ont utilisé la technique d'immunoprecipitation pour capturer les ARN liés à ADAR et une puce à ADN pour les identifier. Plus récemment, Sakurai et al. ont développé un protocole qui ajoute un groupe cyano-éthyle sur l'inosine pour identifier les sites d'éditions : la cyanoethylation de l'inosine stoppe l'élongation du cDNA au site édité lors de la rétrotranscription de l'ARN. En comparant ensuite la séquence du cDNA traité avec la séquence du cDNA normal, de nouveaux sites d'édition ont pu être observés. Finalement, Morse et Bass ont utilisé un traitement au glyoxal et à l'acide borique suivi d'un traitement à la RNAse T1 pour couper spécifiquement l'ARN au niveau de l'inosine et ont séquencé spécifiquement les ARN coupés. Le problème principal avec ces trois méthodes, c'est qu'elles nécessitent toutes une connaissance de la séquence du transcrit étudié pour synthétiser les sondes de la puces à ADN ou les amorces de la PCR. Le but de la seconde partie de cette étude fut de mettre au point un protocole permettant d'extraire les ARN édités afin de les identifier par séquençage haut-débit sans *a priori* sur leurs identités.

Dans un premier temps, nous avons tenté d'adapter le protocole d'immunoprécipitation d'Ohlson au séquençage à haut débit. Après plusieurs ajustements, un enrichissement en ARN connu pour être édité a été observé, mais la quantité d'ARN non édité ne diminuait pas significativement et la quantité d'ARN collecté n'a jamais été suffisante pour être séquencée. Afin d'augmenter le rendement de l'immunoprécipitation, deux méthodes de pré-enrichissement, par gradient de sucrose et par purification du noyau, ont été testées sans plus de succès.

Ensuite, la méthode développée par Morse et Bass a été expérimentée. Dans leur protocole, le traitement de l'ARN au glyoxal et à l'acide borique protège les guanines contre la RNAse T1. Les inosines, par contre, ne lient pas le glyoxal et restent sensibles au traitement à la RNAse T1. Sur cette base, une nouvelle méthode à été développée : d'abord l'ARN est fixé à des billes magnétiques par leurs extrémités 3', ensuite, les billes sont traitées au glyoxal/acide borique et à la RNAse T1 pour libérer la région 5' des ARN contenant une ou plusieurs inosines, et enfin, les ARN libérés sont séquencés par séquençage haut débit (Figure 2').

Figure 2' : Représentation schématique de la purification de l'ARN contenant des inosines.


ARN total
5' AUGCUAGCIUGCICUCGUGCTAGUCG—3'
Attachement d'une biotine en 3'
5' AUGCUAGCIUGCICUCGUGCTAGUCG—biotin
Traitement au glyoxal et à l'acide borique
5' AUGCUAGCIUGCICUCGUGCTAGUCG—biotin
Immobilisation de l'ARN sur des billes magnétiques enrobées de streptavidine
5' AUGCUAGCIUGCICUCGUGCTAGUCG
Segmentation de l'ARN à la RNAseT1 et retrait des billes
5' AUGCUAGCI—3'
CUCGUGCTAGUCG
5' UGCI—3'
Traitement pour enlever le glyoxal
5' AUGCUAGCI—3'
5' UGCI—3'
Séquençage haut-debit

Une première série d'expériences à déterminée l'efficacité du traitement glyoxal/acide borique et RNAseT1, la meilleure façon de lier l'ARN à des billes magnétiques et les conditions optimales pour décrocher le glyoxal de l'ARN. Ensuite le protocole final a été assemblé et son efficacité a été testée par PCR en estimant le niveau d'enrichissement d'ARN ciblé par ADAR. Finalement, le protocole a été utilisé pour extraire les ARN édités du cerveau de souris afin de compléter la validation du protocole. L'analyse des données de séquençage à haut débit montre que le protocole enrichie efficacement les ARNs contenant un ou plusieurs sites d'édition et a permis d'identifier 1822 sites d'éditions, incluant 28 nouveaux sites présents dans des séquences codantes qui conduisent à des mutations non-synonymes des futur protéines. Des sites d'éditions ont aussi été observés pour la première fois dans les ARN ribosomaux, les snoRNA et les snRNA.

Chapter 1: Introduction

The Alu elements

Prevalence and origin of Alu elements

Alu elements represent the most prolific primate specific Short Interspersed Nuclear Elements (SINE). With over one million copies representing 10.5% of the human genome [1, 2], these elements of roughly 300 nucleotides (nt) are derived from 7SL RNA (Figure 1A and B) [12-15] and retrotranspose by “hijacking” the open reading frame (ORF) 1 and ORF 2 proteins encoded by the autonomous retrotransposon long interspersed element 1 (LINE1) (Figure 2) [3, 4]. Their incorporation into the primate genome occurred in three successive waves over the past 65 million years, and gave rise to three main families [16-18]: the old AluJ which arose 60 million year ago (Mya), the intermediate age AluS infiltrated the genome 44 to 32 Mya, and the youngest AluY 24 to 4 Mya [19]. Due to their ubiquity in the primate’s genomes and their species-specific profile Alu sequences have been extensively used for determining primate phylogeny, and have helped to characterize the evolutionary relationship between the great-apes and humans (for review see: [20]).

Figure 1: A) Secondary structure of 7S/L RNA. The RNA is divided in two functional domains called S and Alu. The S domain of SRP binds nascent chains carrying a signal sequence while they emerge from the ribosome; the Alu domain mediates a transient delay in elongation. Boldface indicates the binding sites of SRP9/14 [21, 22]. Three base pairs are formed between two loops and are indicated by dots. B) Secondary structure of the full length Alu element established on an AluY sequence based on the structure established by [23]. Boldface and dots indicate the binding sites of SRP9/14 and the tertiary base pairing between the two loops, respectively, by analogy to SRP RNA. Open arrow indicates the 3' end of scAlu RNA (116nt) and closed arrows the 5' and 3' ends of sRight RNA (155nt). scAlu and sRight RNAs represent monomeric left and monomeric right arms, respectively. Reprinted by permission from Oxford University Press: Nucleic Acids Research [24] copyright 2006.

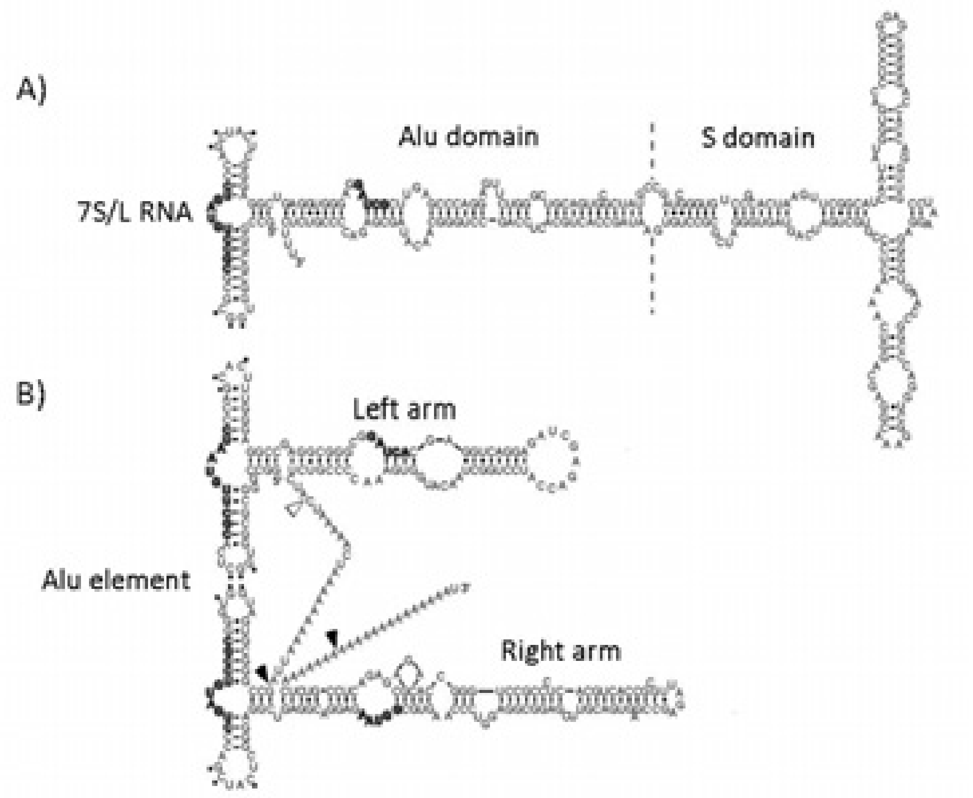
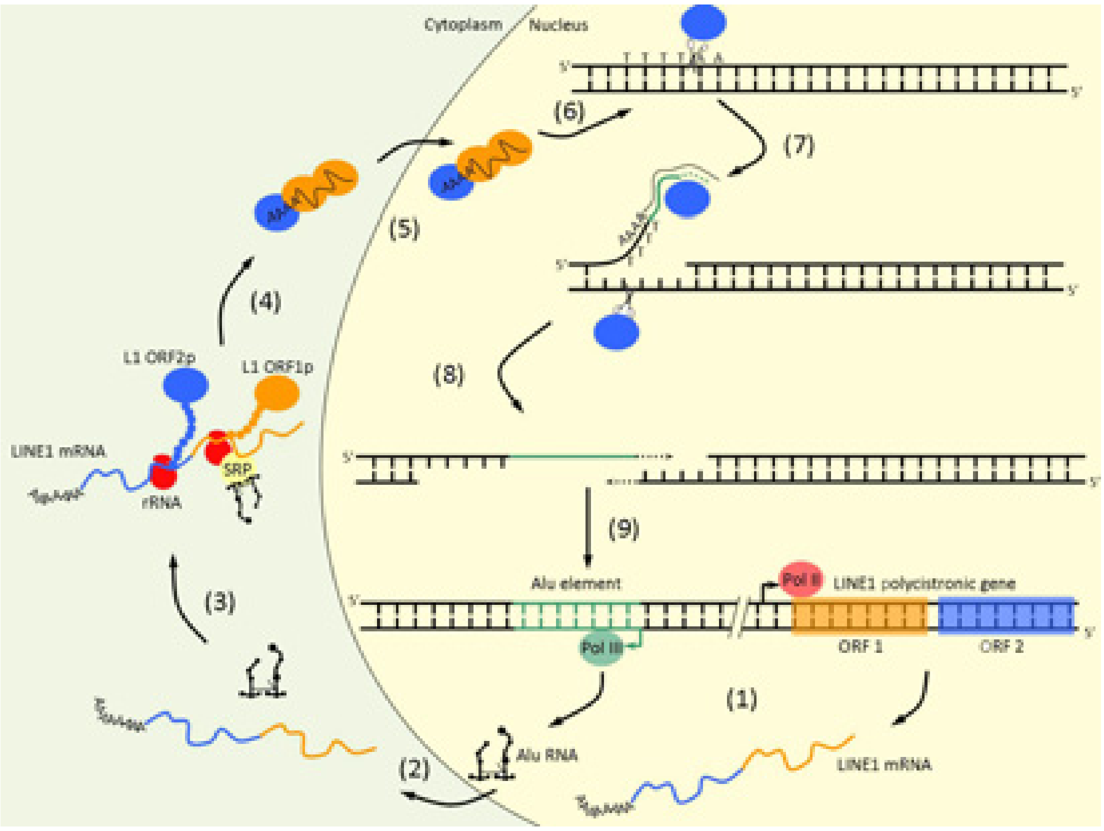


Figure 2: Putative mechanism of retrotransposition of Alu elements by target-primed reverse transcription. (1) Alu is transcribed from its internal POLIII promoter as small RNA of ~300nt and LINE1 is transcribed by POLII (in pink) [25]. (2) LINE1 and Alu transcripts are exported to the cytoplasm. (3) LINE1 polycistronic mRNA is translated in ORF1p (in orange) which is a chaperon protein [26-29] and ORF2p (in blue) which is an endonuclease and a reverse-transcriptase [30, 31]. Alu transcript interacts with SRP (in yellow) at the proximity of the nascent LINE1 proteins [24]. (4) Alu transcript hijacks ORF1p and ORF2p to form a complex which migrates to the nucleus (5). (6) ORF2p cleaves the genomic DNA (gDNA) at the following motif: TTTTAA [32]. (7) The poly-A tail of the Alu transcript binds the poly-T overhang of the gDNA and the Alu RNA serves as template for the ORF2p reverse-transcriptase. (8) ORF2p cleaves the second gDNA strand and (9) the DNA is repaired including the new Alu insertion (review by [33]). The exact mechanisms of parts (4), (5), (8) and (9) have not been validated.



Popular perception of Alu elements

Along with all repetitive elements, the question as to whether Alus are a genomic parasite verses other possible roles as functional integrated genomic elements is still fiercely debated. When Barbara McClintock first discovered transposons in the 1950's [34], she proposed that these mobile elements were responsible for programmed genetic changes during complex development, and therefore called them controlling elements. Although she was awarded a Nobel Prize in 1983 for the discovery of transposons, her original idea that they represented regulatory elements was largely ignored [35] and the idea of transposons as gene regulators was largely dismissed by the 1980's. Moreover, due to their widespread distribution amongst eukaryotes and lack of obvious functions, mobile elements were described as selfish parasitic elements that escaped natural selection by successfully invading and saturating non-critical genomic loci where they had no detrimental effects. Without phenotypic effects, their only function would be their own survival within the genomes they infiltrated, and once inserted they would be under minimal to no positive selection, with only a small minority being exapted into more noble functions [36-38]. This parasitic transposon view was adopted by the majority of the scientific world and is still widespread, despite the accumulation of evidence of the contrary: first, transposons are not passing through selection untouched; second, transposons are not passive occupants of the genome; and third, and most importantly for this study, transposons are not massively switched off transcriptionally by evolution but instead display regulated transcription in different tissues. Each one of these points is discussed below.

Impact of Alu elements on the genome and the genes

Rather than showing a random distribution throughout the genome, retrotransposons are encountered with some degree of sequence and genomic-context bias. Non-exonic regions derived from transposons were shown to be highly conserved and enriched near genes associated with the regulation of transcription and development [39]. More specifically, Alu elements tend to be encountered in GC-rich and gene-dense regions of the genome [40] and are more rapidly eliminated when inserted into AT-rich regions [12], raising the controversial idea of a selective advantage brought on by Alu insertions in gene-dense regions [1, 41, 42].

Retrotransposons have also been shown to impact highly the genome of their host and their expansion played a crucial role in facilitating the rapid evolution of higher eukaryotes [43, 44]. Indeed, even though most retrotransposons are supposed to be transcriptionally inactive components of the genome, they have been shown to increase genomic plasticity and have a strong impact on transcriptomes [45]. At the genomic level they constitute recombination sites that allow exchange of DNA between and within chromosomes [20, 46], and on the transcriptomic level they donate new splice sites for alternative splicing of pre-mRNA [47] or new polyadenylation signals [48], act as targets of A-to-I RNA editing (explained in more detail below) [9-11, 49-51], create functional chimeric proteins by pseudogene retrotransposition into exons [52, 53], and catalyze exon shuffling [54]. As active elements of the genome they also promote gene duplications by transposing part of their flanking sequences along with themselves [55] or by transposing mRNA [56], and by providing new transcription factors binding sites [57, 58] or new promoters [59-61]. As an example of the latter, it has been suggested that as many as 20% of all human micro RNAs (miRNA) are transcribed from Alu repeat POLIII promoters [62].

Transcription of Alu elements

There are several lines of evidence supporting widespread transcription of retrotransposons. The majority of Alu repeats in particular have been shown to contain the necessary POLIII promoter sequences required for transcription initiation [63, 64], and in HeLa cells, the number of Alu transcript was estimated at ~100-1000 per cell under normal growth conditions [65]. Other surveys carried out in a variety of human cells also confirm that Alus are generally transcribed at low levels under normal conditions as either full length Alu repeats or in a truncated form called small cytoplasmic Alus (scAlu). The latter of these are made up with the left Alu monomer alone and arise due to premature POLIII termination sites, RNA degradation or processing of full length Alu transcripts [66, 67]. Not surprisingly, the young AluY which is supposed to be the family having accumulated fewer mutations, were found to be the repeats with the highest expression level in most tissues, although it is also clear that a number of older elements (i.e., AluJ), are still capable of driving transcription [66-69].

The level of transcription is dynamically regulated at individual Alu loci by DNA methylation, chromatin structure, flanking sequences, and promoter sequence variation (reviewed by [70]), and the number of transcripts has been shown to increase dramatically in response to various cellular stress conditions such as heat shock, cycloheximide treatment and viral infection [65, 71]. Alu transcription has also been shown to be up-regulated in some cancerous tissues as compared to normal tissues [72]. However, it is important to note that most investigations into the expression of Alu repeats have been carried out in immortal cell lines rather than in normal tissues [66, 67, 71-74]. This lack of in-depth studies of Alu expression in normal human tissues means that the full complement of Alu repeats that are actually capable of being transcribed is unknown.

Activities of Alu elements

Interaction with SRP

Full length Alus and scAlus are found in both the nucleus and in the cytoplasm where at least some of them are believed to exist as ribonucleoprotein complexes (RNP) comprised of Alu RNA bound to the signal recognition particle 9/14 (SRP) [75], which in turn is involved in the translocation of nascent proteins to the endoplasmic reticulum. In the nucleus, these RNPs are assembled in the nucleolus where there is evidence that they are post-transcriptionally modified at their 3' ends by removal of some terminal nucleotides and by the addition of a single terminal adenylic acid residue prior to export to the cytoplasm [76-78].

Alu RNA bound to SRP 9/14 is believed to affect the synthesis of proteins at the stage of translation initiation upon binding to ribosomes which delays translation [24, 79], notably, it has also been shown that free Alu RNA have the opposite effect on translation and can cause an increase in the synthesis of proteins [24, 80]. In addition, the interaction of Alu repeats with ribosomes is also believed to be essential for their continuing activity as retrotransposons, since the Alu RNA must be in close proximity to an LINE1 transcript that is being actively translated on a ribosome in order to "hijack" the proteins encoded by LINE1 [69].

Interaction with POLII

In addition to the regulatory potential of Alu repeats at the level of protein translation, the murine Alu homologue B1 and the murine SINE B2 have also been found to affect the transcription of protein-coding genes. They carry out this function by binding directly to POLII and inhibiting its transcriptional activity [81-83]. Moreover, in humans the Alu repeats may bind POLII via either of the two Alu monomers, however only the right monomer appears to be capable of repressing transcription [83]. In mice, the related Sine B1, which closely resembles the left Alu monomer, is also able to bind POLII, but it does not affect its transcription. Curiously however, the B2 repeat which is derived from a tRNA and therefore not related to the human Alu repeat or B1 is able to both bind and silence POLII in a manner very similar to that of the right Alu monomer [82, 83]. It is remarkable that such unrelated retrotransposons have acquired identical functional properties in human and mice, and it underlines the regulatory potential of these repeats.

The case of BC200

Another example of Alu repeat exaptation is BC200, an anthropoid primate element that was one of the first Alus to show evidence of transcription [84]. It originated from the retrotransposition of a monomeric Alu element 35-55 Mya, and has given rise to more than 200 pseudogenes in the human genome since then [85, 86]. The 200nt long BC200 transcript can be subdivided into three parts: a monomeric Alu at the 5' end, an A rich region in the central part, and a BC200 specific sequence derived from the locus of integration at the 3' end [87]. The BC200 transcripts are localized almost exclusively in the dendrites of neurons, although some expression in the germ-cells and some cancerous tissues also occur [84, 87, 88].

BC200 is believed to regulate protein synthesis in the post-synaptic regions of dendrites by interacting with at least four proteins involved in translation. First, due to the homology between the Alu-derived part of BC200 and 7SL RNA, it is capable of binding to the SRP [89]. Second, via its A-rich central region BC200/BC1 may also interact with the poly-A binding protein (PABP), a regulator of translation initiation [90]. Third, BC200 and its mouse functional homologue BC1 have also been shown to interact and determine the specificity of the fragile-X mental retardation protein (FMRP), a translational repressor of specific RNAs at synapses. It is believed to fulfill this function by binding to the FMRP protein and linking it to specific mRNAs through complementary strand recognition [91]. Finally, BC200/BC1 has also been shown to repress translation of mRNAs containing 5' secondary structures by blocking the RNA helicase activity of the eukaryotic initiation factor 4A (eIF4A), thus preventing the assembly of the ribosomal complex onto the mRNA [92]. Curiously, BC200 has two functional homologues derived from different transposition events in two unrelated orders: BC1 derived from tRNA^{Ala} and G22 derived from a dimeric Alu element, have been found in rodents [93-95] and in the Lorisoidea branch of prosimians respectively [96, 97].

BC200, the functionally related BC1 and G22 elements, and the similarity between Alus and B1s all represent fascinating examples of retrotransposon-derived noncoding RNAs that have acquired important regulatory functions in mammals. The observation that three evolutionarily unrelated yet functionally similar retrotransposon elements have been independently exapted in three different mammalian orders is remarkable, and highlights the functional significance of SINE elements in the nervous system and in mammalian evolution.

Alu elements and editing proteins

One important feature of Alu repeats that has been characterized in the past few years is their ability to interact with RNA editing enzymes. In the cytoplasm scAlu and full length Alu elements bind to the C-to-U editing enzyme APOBEC3G. Unexpectedly, this interaction does not lead to editing [98] but to the sequestration of Alu elements in high molecular mass ribonucleoprotein complexes like the Staufen-containing granules [99], which are involved in nuclear-cytoplasmic shuttling and in dendritic RNA targeting in neurons (reviewed in [100]). The purpose of this interaction was proposed to be the sequestering of Alu transcripts away from the LINE1 machinery required for their retrotransposition, which would inhibit detrimental Alu retrotransposition [101]; however, this interaction results in the inactivation of the editing function of APOBEC3G [99]. A more elegant explanation would be the transport of Alu transcripts to their "working place" in neurons expressing APOBEC3G [102] but this hypothesis remains untested.

In the nucleus, Alu elements were shown to be the main target of A-to-I editing by the adenosine deaminase acting on RNA (ADAR) family of proteins, with 92% of A-to-I editing events occurring in Alu elements in human [9]. It was shown that as integral elements of RNA, Alus can form a duplex with an adjacent inverted homologue, which promotes hyperediting of the RNA secondary structure and leads to retention of the mRNA in the nucleus [10]. Editing of Alu by ADAR was also proposed to significantly modify the structure and abundance of transcripts by creating alternative splice sites, by changing codons in exons, by interfering with the RNAi pathway and by potentially promoting heterochromatin formation (reviewed in [11]), also this is supported by less evidence. A more detailed description of the interaction between ADAR and repeats can be found in the section entitled "A-to-I editing in the untranslated regions of genes".

A-to-I RNA editing

RNA editing is the modification of nucleotides in an RNA molecule and was first observed in the cytochrome C oxydase subunit II of the *Trypanosoma* where uracil residues were found to be either inserted or deleted in the primary transcript [103]. Shortly after this, a C-to-U editing event was described in the apolipoprotein B transcript in mammals [104, 105], as well as A-to-I editing events in double stranded RNA in *Xenopus* [106, 107]. Since then, a whole diversity of nucleotide modifications have been characterized in all classes of organisms, predominantly in tRNA, rRNA and mRNA [108], and amongst these A-to-I editing is considered the most prominent type of editing in higher eukaryotes [5].

The ADAR family

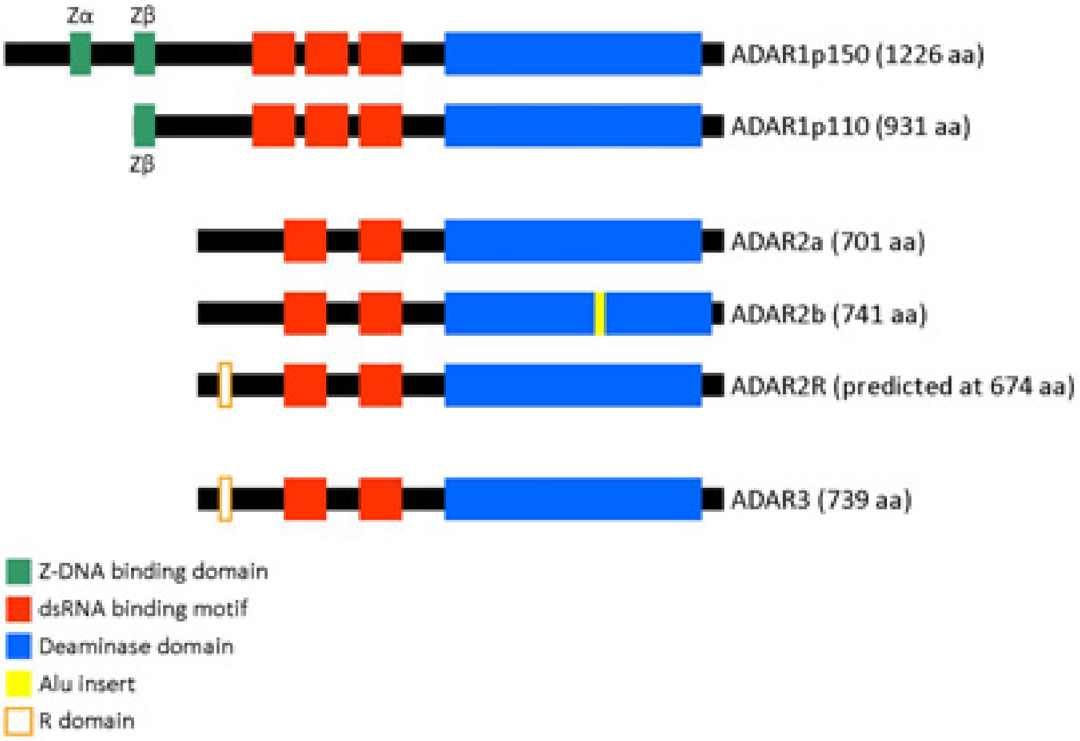
Most A-to-I editing in mammals is catalyzed by the Adenosine Deaminase Acting on RNA (ADAR) family of proteins, which appeared after the split of protozoan and metazoan and are found in all multicellular animals from worms to human, although they are absent in yeast and plants [109]. The ADAR proteins are encoded by three genes in vertebrates, *ADAR 1*, *2* and *3*, and evolved from the Adenosine Deaminase Acting on tRNA (*ADAT*) family, which is present in all eukaryotes including yeast and plants [110], by acquiring a double stranded RNA binding domain.

ADAR1 encodes two main isoforms from three distinct promoters [111, 112] (Figure 3). The long isoform, ADAR1p150, is transcribed from an interferon inducible promoter [113], contains two Z-DNA binding domains Zα and Zβ [114, 115], three dsRNA binding motifs and one deaminase domain and is upregulated in infectious conditions [112, 116]. The short isoform, ADAR1p110, is transcribed from a constitutive promoter into a splice isoform of ADAR1p150 which misses the first Z-DNA binding domain [111, 113]. This isoform contains one Z-DNA binding motifs Zβ unable to bind Z-DNA [115, 117] and like ADAR1p150, three dsRNA binding domains and one deaminase domain. The two isoforms present similar editing activities [118] but their localizations in the cell differ. ADAR1p150 is present mostly in the cytoplasm with minimal localization in the nucleus, whereas ADAR1p110 is localized to the nucleus and accumulates in the nucleolus [113, 119].

Similar to ADAR1p110, ADAR2 also localizes to the nucleus and accumulates in the nucleolus [119], and its expression is induced by the transcription factor cAMP response element-binding (CREB) [120]. The *ADAR2* gene produces many isoforms due to alternative splicing [121-123], and amongst these three have been characterized in depth (Figure 3). The major splice isoform is ADAR2a, which contains two dsRNA binding domains and a deaminase domain, and is the isoform usually used to produce ADAR2 for protein assays [124, 125]. ADAR2b is organized the same way but contains an Alu insert in the deaminase domain, which reduces the deaminase activity of the protein by approximately two fold [126]. Finally, ADAR2R contains an additional exon which adds a single stranded RNA binding domain upstream of the dsRNA binding domains [127]. ADAR2 proteins are expressed ubiquitously and both ADAR2a and ADAR2b are expressed at similar levels and produce the majority of the ADAR2 protein present in the cell [128]. ADAR2R is a minor isoforms whose level of transcription is dependent of the tissue: for example, it represents ~10% of the ADAR transcripts in the hippocampus, 5% in the colon, 2.5% in skeletal muscle and 1% in heart [127].

ADAR3 is localized to the nucleus of cells [129], and only one protein has been characterized to any degree. Similar to ADAR2a, ADAR3 contains two dsRNA binding domains and one deaminase domain but it also possesses an ssRNA binding domain in the N-terminus. ADAR3 is expressed exclusively in the brain, specifically in the amygdala, the thalamus, the cerebral cortex and at a lower level in the cerebellum, the occipital pole, the frontal lobe, the temporal lobe, the caudate nucleus and the hippocampus. Since no editing functions have been shown *in vitro*, it was proposed to function as a competitive inhibitor of ADAR1 and 2 [130, 131]. However, it is important to note that homodimers, the structure necessary for an editing activity for ADAR1 and 2, could not be obtained *in vitro* for ADAR3 even though such structure are present *in vivo* [132].

Figure 3: Structure of the different isoforms of ADAR proteins in human. The Z-DNA binding domains are represented in green, the dsRNA binding domain in red, the deaminase domain in blue, the Alu insert in yellow and the R domain with an orange line.



The A-to-I editing reaction

Conversion of A-to-I by ADAR

ADAR1 and ADAR2 convert adenosine to inosine by C6 hydrolytic deamination of adenosine in dsRNA [106, 133]. In this reaction, the ADAR dsRNA binding domain first recognizes the double stranded structure of a transcript. To be recognized, the double stranded region needs to be at least 20nt long (that is two turns of a dsRNA helix), and can be formed by two individual RNAs or within the same RNA by a stem-loop [134] or a pseudoknot [135]. Next, the ADARs form a homodimer through their dsRNA binding motifs around the dsRNA region [136-138] and deaminate the adenosine. If the dsRNA helix is perfect, the A-to-I conversion can occur in up to 50% of the A present in the dsRNA [139, 140]. However, imperfections in the dsRNA limit A-to-I conversions to specific sites, as observed in the glutamate receptor GRIA2 [141, 142] and in the serotonin receptor 5-HT2C [143, 144]. A:C mismatches in the dsRNA constitute favored editing sites for both ADAR1 and 2 [145], although the edited A are also selected according to their nearest neighbours, ADAR1 and ADAR2 will preferentially edit A with U>A>C>G in 5' position, and in the 3' ADAR1 favors G>C≈A>U, and ADAR2 G>C>U≈A [146, 147]. Finally, the tertiary structure of the dsRNA was also shown to impact the position of the editing site: when adenosines are edited in group, all edited nucleotides appear on the same side of the double helix [148].

Notably, the sites to be edited are not targeted by ADAR1 or ADAR2 with the same efficiency. For example, the serotonin receptor 5-HT2C contains five well characterized editing sites identified by the letters A to E [144, 149, 150]. Amongst those, sites A and B are predominantly edited by ADAR1, C and D are altered by ADAR2, and site E is edited to a lower extent by both enzymes [130, 143, 144, 147, 151]. Similarly, the glutamate receptor GRIA2 possesses two editing sites extensively described: the R/G site is edited by both ADAR1 and 2 whereas the Q/R site is edited exclusively by ADAR2 [152-154]. This disparity in site recognition between the two enzymes is conferred by both their deaminase domain and their dsRNA binding domain [145, 147].

Factors influencing the conversion

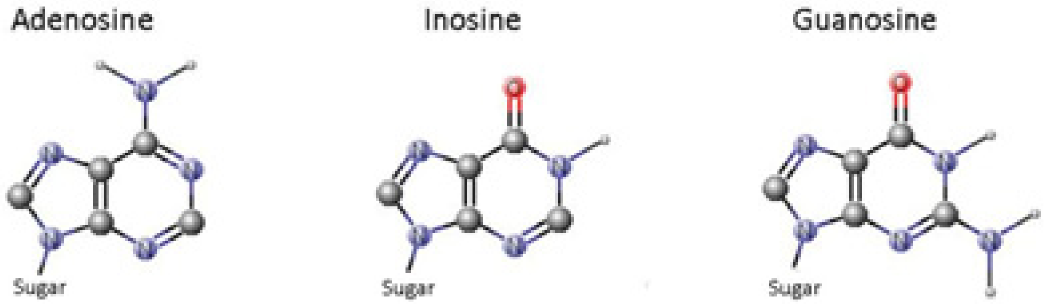
In addition to the structure of its target, ADAR activity is also constrained by its localization in the cell. For example, both ADAR1p110 and ADAR2 localize in the nucleolus, and when a transcript to be edited is produced both enzymes delocalize from the nucleolus and migrate to the location where the substrate transcript accumulates [119]. It was also shown that a transcript expressed directly in the nucleolus could be edited by ADAR2 but not by ADAR1, suggesting that ADAR1 is inactive in the nucleolus [155]. The localization of ADAR2 in the nucleolus was shown to be dependent of the presence of rRNA and PIN1. The interaction of ADAR2 with rRNA is supposed to sequester ADAR2 to the nucleolus on dsRNA fragment where it does not produce any editing [156]. The interaction of ADAR2 with PIN1 is dependent of two phosphorylations in the N-terminal region of ADAR2 and occurs after the binding of ADAR2 to a dsRNA. This protects ADAR2 from degradation promoted by ubiquitin ligase WW domain containing protein 2 (WWP2) and increase ADAR2 editing activity [157].

ADAR activity can be further modulated by several other factors. In Addition to ADAR3 which can acts as a competitive inhibitor [130, 131], ADAR2 activity is impacted by the IP3 pathway: an inositol hexakisphophate (IP6) inclusion into ADAR2 structure is necessary for proper folding of the active protein [158]. Finally, snoRNAs are also able to inhibit editing by adding a 2’-O-methyl to the nucleotide to be edited, which reduces editing efficiency ~200 times [159]. This regulation occurs on 5-HT2C whose editing site C is modified by the C/D box snoRNA HBII-52, thus reducing the efficiency of the editing at this site [155, 160].

Consequences of editing on the transcript

The replacement of the primary amine of the A by an atom of oxygen to make I drastically changes the physicochemical properties of the nucleotide (Figure 4). Indeed, once the A is changed into I, it will be read as a G by the translatory machinery and will pair preferentially with C in the secondary structure of the RNA [161-163]. Such changes in a transcript directly impact its function.

Figure 4: Chemical structures of the adenosine, the inosine and the guanosine



Editing in the coding sequence (CDS)

When the editing occurs in the CDS of a transcript, it changes the sequence of the protein and potentially its property, which is the case for the glutamate receptor and the serotonin receptor that were described previously. In the glutamate receptor subunit GRIA2, A-to-I editing changes systematically a genomic encoded glutamine (Q) into an arginine (R) in the second transmembrane region of GRIA2 [164]. This change alters the permeability of the channel to Ca^{2+} ions [165] which was later shown to be essential for a normal brain development [166]. GRIA2 also contains a second site where arginine is changed to glycine, which modifies the kinetics of the receptor; this editing site is also observed in the subunits GRIA3 and GRIA4 [142]. The coding region of the serotonin receptor 5-HT2C contains five editing sites, A to E, modifying three amino acids (aa) located on its second intracellular loop. The first aa to be modified in position 156 is a genomic encoded isoleucine converted to a valine when A or A and B are edited, or to a methionine when B alone is edited. The second aa in position 158 is an asparagine converted to a serine when C is edited, to an aspartate when E is edited and to a glycine when both C and E are edited. The last aa in position 160 is an isoleucine converted to a valine when D is edited. The conversion of these aa modify the interaction of the receptor with the G protein [144, 150]. Interestingly, a regulatory feedback loop of 5-HT2C was proposed where activation of the receptor by serotonin induces the production of IP6, which in turn activates ADAR2; activated ADAR2 then edits 5-HT2C transcripts to reduce their sensitivity to serotonin [158, 167]. Modifications of the primary sequence of a protein by A-to-I editing has been observed in other genes primarily involved in nervous system development and function [168] such as GABRA3 [169, 170] and KCNA1 [171], and Pullirsch and Jantsch reported an additional 38 proteins modified by A-to-I editing [8].

A-to-I editing in the untranslated regions of genes and ncRNA

The modification of CDS by A-to-I editing represents a small minority of the editing events, as the majority of A-to-I editing occurs in non-coding RNA or non-coding regions of pre-mRNA, mainly in regions containing repetitive elements such as Alu elements or LINE1 [172]. In such cases, two closely related repetitive elements oriented sense and antisense to each other form the double stranded structure required for ADAR editing [49]. In humans, nearly 15,000 editing sites of this kind have been identified in approximately 2,000 genes [9, 50, 51, 173], with 88% to 92% of editing events occurring in Alu elements integrated into longer transcripts [9, 51].

Creation and deletion of splicing sites

Editing in non-coding regions can also modify splicing. For instance, ADAR2 edits its own pre-mRNA to create a new splice junction inside an intron, which leads to the production of a truncated transcript that would produce a non-functional ADAR2. This mechanism of ADAR2 auto-editing was proposed to act as a way to reduce the production of ADAR2 when needed [6]. Removal of splice junction by editing has also been observed in *Drosophila* with 4F-RNP [7] or in mouse with PTPN6 [174]. Notably, it was shown that 5-HT2C splicing pattern was also dependent of editing at specific sites but 5-HT2C does not possess any editing site located on splice junction. This suggests that it is the structural change produced by editing that modifies splicing [175].

Hyperedited RNA in the nucleus and cytoplasm

The dsRNA structure formed by two repetitive elements in a transcript can lead to its hyperediting [176]. Such events were shown to promote the retention of transcripts in nuclear paraspeckles, through their interactions with the nuclear RNA binding protein p54^{nrb} and the non coding RNA Neat 1 [10, 177, 178]. The storage of RNA in paraspeckles was proposed to be a way of releasing transcripts when required, as illustrated by the cationic amino acid transporter 2 (*Cat2)* in mouse. *Cat2* encodes two transcripts, one of which is exported to the cytoplasm to be translated (CAT2) and one which contains one repeat in sense orientation and three antisense repeats in its extended 3'UTR (CAT2 transcribed nuclear RNA: CTN-RNA). As expected, the 3'UTR of CTN-RNA is hyperedited and the transcript is sequestered into paraspeckles in the nucleus. Under stress conditions though, the 3'UTR of CTN-RNA is cleaved, the transcript is freed from the nucleus and migrates to the cytoplasm where it is translated [179].

In the cytoplasm, hyperedited transcripts were shown to be cleaved by the Tudor Staphylococcal Nuclease (TUDOR-SN), a subunit of the RNA-induced silencing complex (RISC) [180], to localize to stress granule along with ADAR1p150 and to down-regulate translation in trans [181, 182]. It was suggested that these interactions are part of a mechanism of cell survival during stress conditions [181]. However, the full pathway of such mechanism is not elucidated yet.

Editing and miRNA

miRNAs are transcribed as long primary miRNAs (pri-miRNA), and are processed into ~60–70-nt precursor miRNAs (pre-miRNA) in the nucleus by the ribonuclease III-like enzyme DROSHA. After nuclear export, the pre-miRNAs undergo further processing by DICER into ~20–22nt mature miRNAs, which are then loaded into RISC. RISC uses the miRNA to scan mRNA and block the translation of partially complementary mRNAs, and can also cleave mRNA showing higher complementarity (reviewed in [183]). Editing on 47 pri-miRNAs were reported so far [184], and A-to-I RNA editing can impact the miRNA pathway in several manners. First, by altering the mRNAs, it can create new regions recognized by miRNAs [185]. ADAR can also directly target the dsRNA structure formed by the pri-miRNA [186, 187], the editing of which can block processing of the pri-miRNA by DROSHA [188] or processing of the pre-miRNA by DICER [184, 189]. If the edited pri-miRNA is successfully processed into a mature miRNA, the editing sites on the miRNA can change the panel of genes to be silenced [190]. It may also change the miRNA strand which is loaded in RISC if the editing occurs at the extremities of the miRNA [191-193].

Consequences of editing on the organism

A-to-I editing has critical effects on coding and non-coding transcripts which can directly translate into an altered phenotype. For example, the ADAR mutants that have been created for several organisms so far highlight the role of ADAR in development and its key involvement in nervous system development. In *Caenorhabditis elegans*, mutants for *ceADR1* and *ceADR2* display altered chemotaxis due to an impaired nervous system [194]. In *Drosophila*, adult mutants with dADAR deletion exhibit altered nervous system functions such as slow uncoordinated locomotion, abnormal body posture, temperature-sensitive paralysis and brain degeneration, effects that can be explained by the lack of editing in ion channels of the nervous system [195]; notably, the dADAR mutant phenotype can be rescued by the human *ADAR2* [196]. In mouse, ADAR2 knock out mutants experience repeated epileptic seizures and die in the three weeks after birth, a phenotype due mainly to lack of editing at the GRIA2 Q/R site [166]. When the GRIA2 Q/R site is rescued, ADAR2 knock out mutants show an increase in IgE level which could potentially modify the response of the mouse to allergens, as well as a change in sensorimotor integration linked to the neurochemical system, hearing deficit and overexpression of genes involved in neuroprotection and synaptic trafficking [197]. Murine ADAR1 knock out mutants manifest severe defects in hematopoiesis and die around embryonic day 12 [198-200], and conditional deletion of ADAR1 in hematopoietic cells revealed that RNA editing by ADAR1 is also primordial for the hematopoiesis in the adult mouse by down-regulating interferon inducible transcripts which can lead to apoptosis [201, 202]. In humans certain alleles of *ADAR2* and *ADAR3* are associated with longevity [203], and conversely, dysfunctions of ADAR proteins are associated with diseases (for a detailed list of human disease in which ADARs are involved refer to Gallo [204, 205]). Mutations in *ADAR1* were associated with dyschromatosis symmetrica hereditaria [206], underediting of GRIA2 at the Q/R sites due to dysfunction of ADAR2 was linked to sporadic amyotrophic lateral sclerosis [207, 208], glyomas [209, 210] and epilepsy [211]; and altered editing of 5-HT2C was linked to depression and suicide [212-214]. Finally, in octopuses, editing was shown to modify a K+ channel in response to cold environment leading to the production of a more effective channel to adapt the organism to the polar conditions [215].

Method to identify the targets

The characterization of A-to-I RNA editing sites is a critical step to comprehend the functions and regulation of a gene, and several approaches have been used to identify these sites.

Bioinformatic approach

Editing sites were first detected serendipitously when A-to-G mismatches were observed by comparing the genomic DNA (gDNA) sequence of a gene to its cDNA sequence. Once the human genome and the mouse genome were successfully sequenced [1, 216], a more thorough comparison was performed by comparing expressed sequence tag (EST) datasets against the respective genomes, and in 2004, four teams published the results of this comparison. Kim et al. found 2674 edited transcripts across 30 human tissues and 91 edited transcripts in mouse [51]. Athanasiadis et al. reported 1445 edited mRNA in 13 human tissues [50]. Levanon et al. identified 1637 edited genes in human [9], and Blow et al. recognized 1727 edited transcripts in the human brain [173]. In their reports, the four teams acknowledged the difficulty of distinguishing editing sites from sequencing errors or single nucleotide polymorphism (SNP) and applied filters to minimize the impact these factors. Another difficulty was that this approach was directly dependent on the tissue and the depth of the EST databases. As a result, neither GRIA2 nor 5-HT2C editing sites were detected.

Experimental approach

Several “wet lab” approaches were also developed to enrich and identify edited RNAs on a moderate scale. Ohlson et al. generated an antibody targeting ADAR2 that would co-immunoprecipitate RNAs bound to the editing protein, followed by identification of the captured RNAs by micro-array (results not available) [217, 218]. Sakurai et al. used inosine cyanoethylation to characterize editing events. The cyanoethylation of inosine prevents the elongation of cDNA at the inosine when reverse transcribing the RNA. By comparing treated cDNA versus untreated cDNA, new editing sites could then be observed. Using specific probes, 642 regions were targeted and 5,072 editing sites were found [219, 220]. Morse and Bass developed a method using glyoxal and RNAseT1 to specifically cleave RNA at inosine sites and followed by sequencing of the cleaved RNAs [221]. Finally, Li et al. used a padlock probe to capture sequences of potentially edited RNAs and their gDNA counter-parts, followed by deep sequencing of both RNA and gDNA to identify A-to-I editing events. Out of the 36,208 sites which were selected for their analysis, they identified 239 edited sites [222]. The main issue with all of these methods is that they require prior knowledge of the sequence of interest to assess its editing level. Also, the use of specific probes to target the chosen sequences is not compatible with high throughput discovery of editing sites.

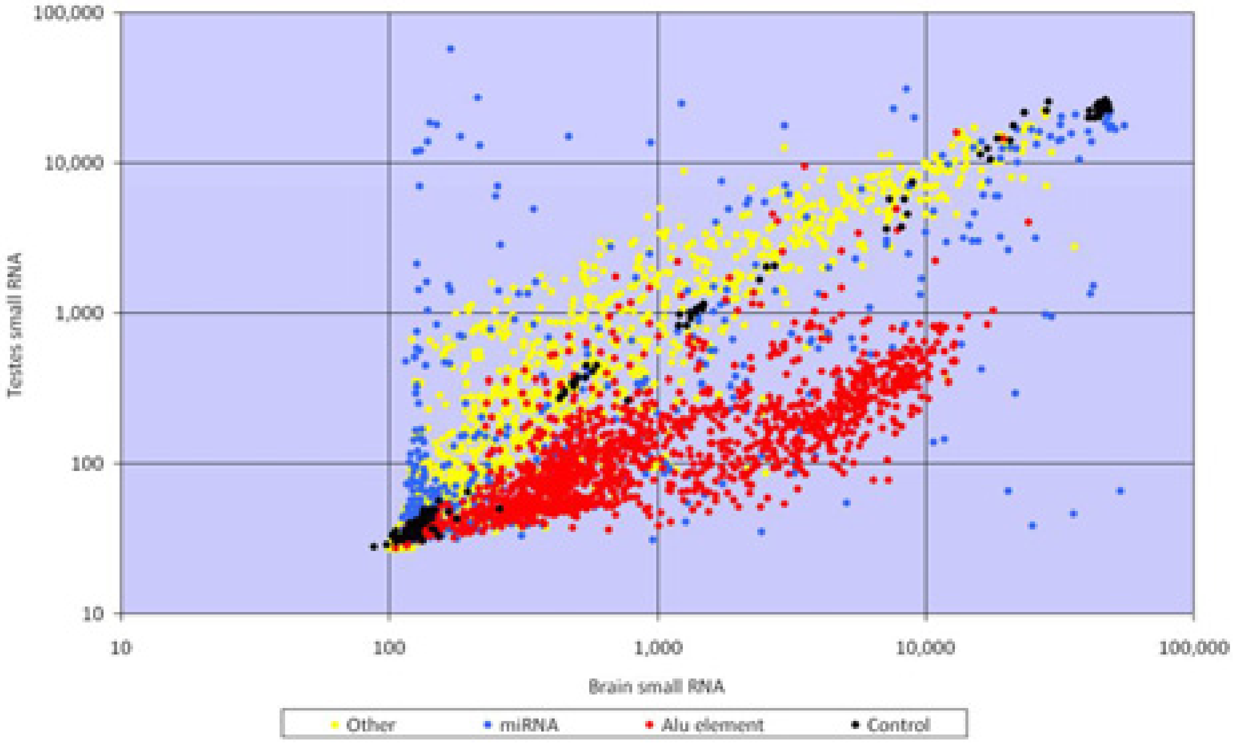
RNA editing and Alu elements encode complexity

The widespread distribution of Alu elements in the human genome and transcriptome, their privileged relationship with editing enzymes, and the preponderance of editing in brain were interpreted by Mattick and Mehler as a way to explain the evolution of cognition in human [223-225]. The development of complex organisms is mainly regulated by non-coding RNA rather than by proteins [223]. Proteins modifying RNA therefore constitute a way to integrate environmental conditions at the molecular level in a system where environmental stimuli could alter encoded genetic information, which in turn modifies gene regulation and function by acting on non-coding RNA. Due to their ability to travel between the dendrites and the nucleus, to be edited, and to induce epigenetic and genetic modifications, Alu elements expressed in the brain might constitute an interface between the environment and the neurons' genome. Acting on neuronal plasticity and brain development, and being one of the major constituents of the human genome, Alu elements were proposed to be the molecular basis of long term memory and higher order of cognition and to have been positively selected for on this basis [224].

In the Mattick and Mehler proposal [224], the mechanism modulating the nervous system is similar to the one observed in the adaptive response of the immune system. In the adaptive immune system, the genomes of the B lymphocytes are modified at the immunoglobulin (Ig) loci by recombination and editing to generate a panel of Ig which provides the organism with the most effective antibody against a specific antigen. The B lymphocytes providing the best response are then used to degrade the antigen, and are preserved as memory of this response in case of a second invasion. By analogy, recoding of DNA in nerve cells was proposed to be the main mechanism by which productive or learned changes induced by RNA editing are rewritten back to the DNA, which would fix the altered genotype once a particular neural circuitry and epigenetic state has been established. In this mechanism, RNA-editing plays a primordial role by establishing a certain epigenetic state as a function of the environment by altering specific RNA elements; this in turn leads to DNA recoding through DNA repair enzymes.

Consistent with the proposal for a central role of Alu sequences in the development and function of the human brain, recent data suggests that a large number of ~300nt long Alu elements is over-expressed in the human brain compared to testes (Croft and Mattick, unpublished data; Figure 5). A recent study also revealed that Alu elements actually retrotranspose in the normal human brain, thus change of the neuronal genome is not exclusive to cancer but also occurs in normal tissue [226].

Figure 5: Scatter plots of small RNAs (<300nt) expression level in human brain and testis measured by microarray. The microarray was designed by Larry Croft to target human miRNA and regions of the human genome predicted to encode RNAs containing stem-loop structures. The signals corresponding to Alu sequences are colored in red, miRNAs are in blue, control probes (spike-in control, GAPDH, Actin and 5S rRNA with match and mismatch probes) are in black and other predictions in yellow. The data were provided by Larry Croft.



Final considerations

There are 1,194,734 Alu elements in the human genome, which represents approximately 10% of its total size. Along with the fact that most Alu elements carry a functional promoter, and that small Alu transcripts are enriched in human brain, a deep analysis of small Alu transcriptions would allow us to uncover a poorly characterized part of the human transcriptome; this in turn may provide insight into still to be discovered regulatory networks in human brain.

Alu elements are also part of the majority of the transcripts and represent the main target of RNA-editing. By extension, RNA-editing is a mechanism able to reach all cellular processes. The prevalence of RNA-editing in human, and specifically in the brain, potentially implicates this mechanism as a powerful driver in what differentiates humans from other animals: the cognition. Uncovering the targets of RNA-editing will reveal the wide repercussion of such a mechanism on human development, and may help to elucidate the molecular basis of the human brain abilities.

In an original model, Mattick and Mehler proposed that the basis of human cognition is the recoding of DNA mediated by RNA-editing and Alu elements. If proved to be true, this hypothesis may revolutionize the understanding of the brain development and challenge the dogma that genomes constitute stable blueprints of the organisms.

Chapter 2: Characterization of small Alu elements in the human transcriptome

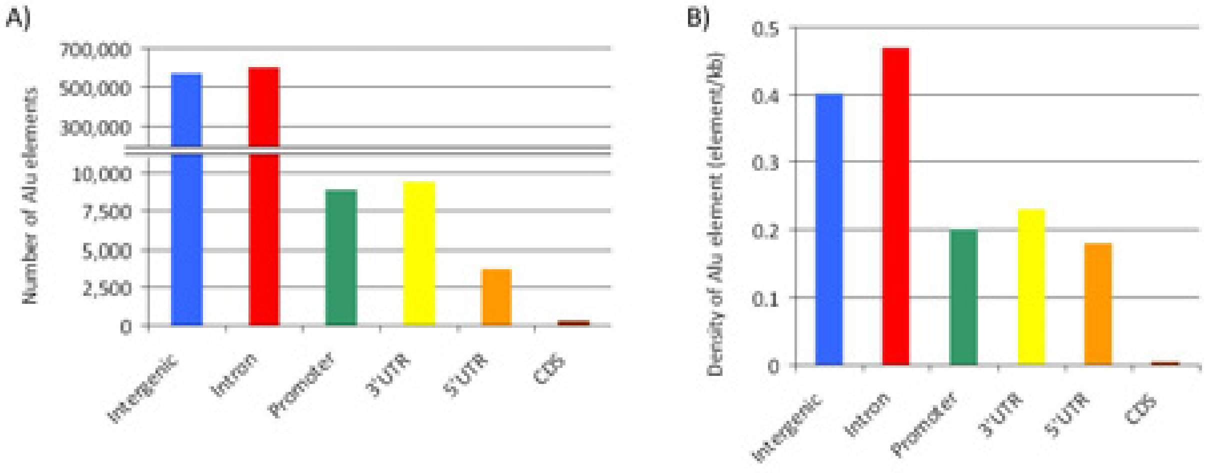
Several studies have attempted to characterize Alu elements, by determining which ones are transcribed [65-67, 227], bind protein complexes [75, 228, 229], control the translation of protein [24, 79, 80] or are able to retrotranspose [3]. However, most of these have performed their analysis on a limited number of Alu candidates, or have used a consensus sequence based on the three main Alu families and pursued subsequent analysis in cell or *in vitro* assays, and thus relate only to a limited subset of Alus amongst the 1,194,734 elements present in the genome. Critically, such studies do not allow accurate inferences to be made about the role of the majority of these elements. To precisely characterize the properties or functions of each Alu element, high-throughput methods are required but since one of the first steps in most of high-throughput analyses is to discard all repetitive elements, only a few published reports can be found with inclusion of repeats in their datasets [45].

In the first part of this thesis, Alu elements were annotated in function of their genomic environment and time of apparition; then the view that repetitive elements cannot be characterized by deep sequencing was challenged, followed by functional annotation based on their potential to bind POLIII and their localization within the cell. Next, sixteen human somatic tissues were screened for Alu and LINE1 elements expression to demonstrate that retrotransposition is possible in normal tissues. Further retrotransposons screening in mouse tissues showed that retrotransposition in somatic tissues may not be limited to human, and pair-end transcriptomic datasets were analyzed for chimeric transcripts to identify new transposon insertion sites in human. Finally, a database that assembles all the information generated in this study was constructed.

Genomic location and conservation of Alu elements

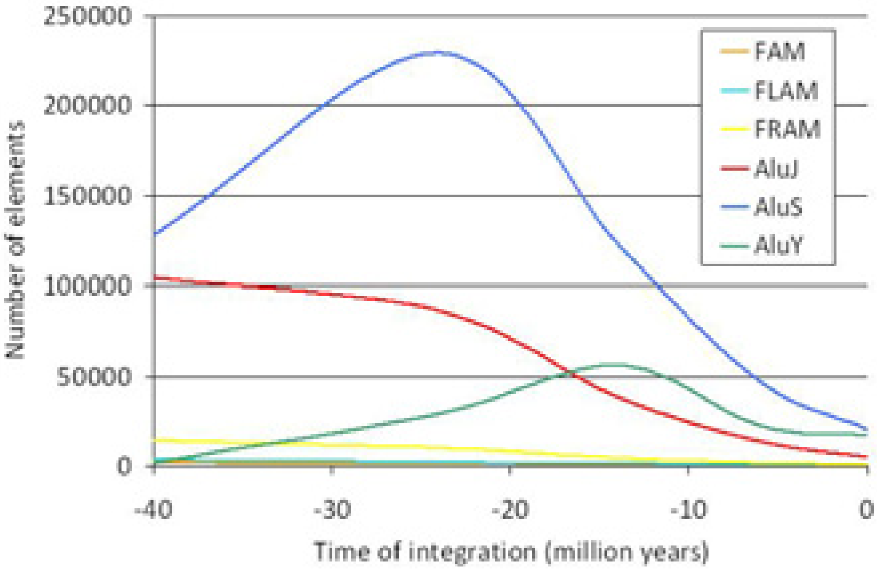
The genomic environment of each Alu elements was determined by intersecting Alu elements coordinates from Repbase [230] with the coordinates of the promoter region of each gene, the 5’UTR, the coding exons, the exons from non-coding RNA, introns, and the 3’UTR of the University of California Santa Cruz (UCSC) genes dataset [231], and all Alu that did not overlap with these regions were considered intergenic. 571,938 Alu elements were located in intergenic regions, 600,628 in introns, 8,855 in promoters, 9,343 in 3’UTR, 3,684 in 5’UTR and 286 in CDS (Figure 6A). It appeared that Alu elements were enriched in intronic regions with a density of 0.47 Alu elements/kb; the intergenic density is slightly lower with 0.40 elements/kb, promoters contained 0.20 elements/kb, 3'UTRs 0.22 elements/kb, 5'UTRs 0.18 elements/kb and the CDS 0.003 elements/kb (Figure 6B).

Figure 6: A) Genomic location of Alu elements in the human genome (hg19). Alu elements inserted in intergenic regions are represented in blue, introns in red, promoters in green, 3'UTRs in yellow, 5'UTRs in orange and CDS in brown. The y-axis is the number of Alu. B) Relative coverage of each genomic location by Alu elements. The y-axis is the number of Alu elements per kilobase of genomic location.



To estimate the time of apparition of each Alu element, homologues were searched for in the genomes of Marmoset, Rhesus, Orangutan and Chimpanzee that split from Human 35-40Mya, 23-25Mya, 14Mya and 6Mya respectively [232]. 253,425 Alu elements were present in the genome before the separation of the new world monkey from other primates at 35-40Mya, mostly AluJ and AluS but also the less prolific ancestral families FAM, FLAM and FRAM (Figure 7). At the separation of the old world monkey from the other primates at 23-25Mya, 359,202 additional Alu elements were integrated in the primate genome, the majority of them being from the AluS family. At the split between Orangutan and other primates, around 14Mya, 223,786 other Alu elements have appeared which was also the point when the majority of the AluY population appeared. At 6Mya, the separation period of chimpanzee and human, the rate of expansion was reduced for all the families, and only 84,986 Alu elements integrated into the genome. Finally, since the separation of chimpanzee and human, 43,209 Alu elements have appeared in the human genome (Figure 7).

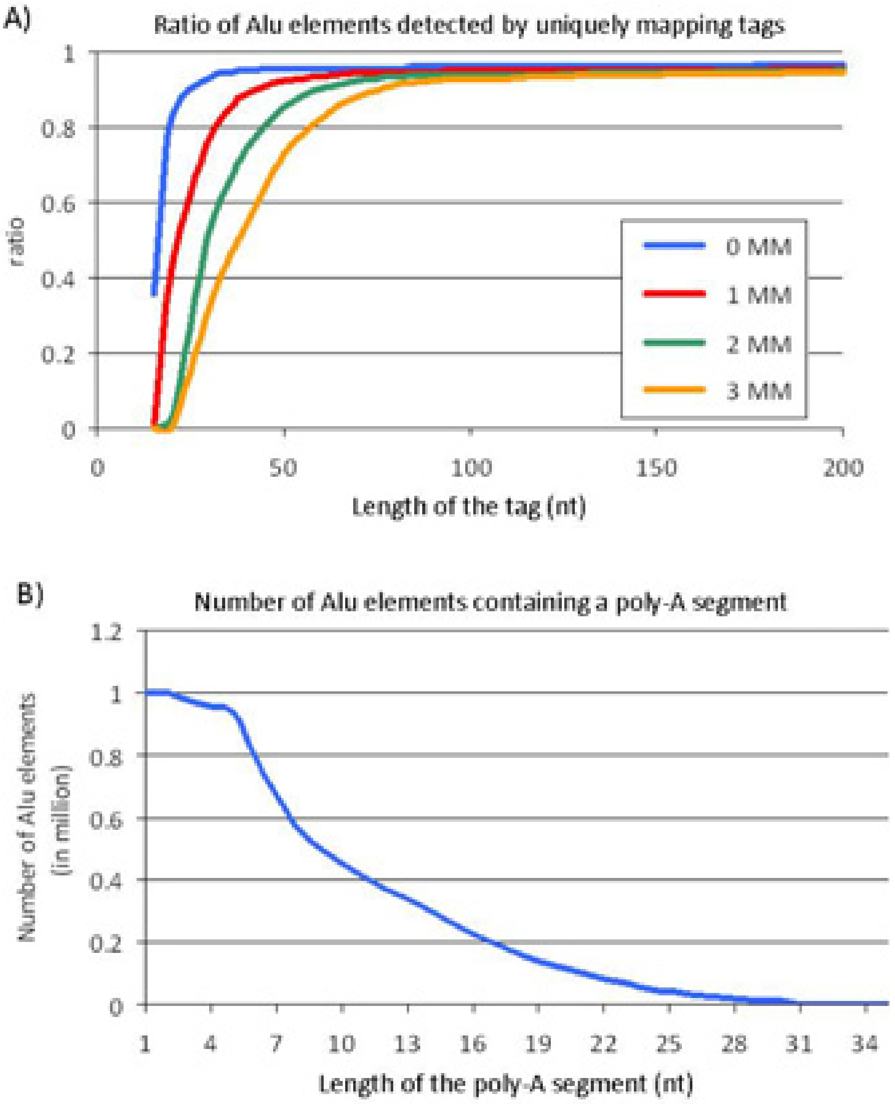
Figure 7: Time of integration (in million years) of the human Alu elements in the genome per family: FAM in orange, FLAM in light blue, FRAM in yellow, AluJ in red, AluS in blue and AluY in green. The time of integration is based on the conservation of the human Alu elements in Marmoset, Rhesus, Orangutan and Chimpanzee and the phylogeny of primates calculated by Marques-Bonet et al. [232].



Alu elements are mostly unique and can be identified by deep sequencing

Although Alu repeats are quite similar in their primary sequences, systematic comparison of all human Alu sequences showed that 96.6% of Alu elements present a unique sequence in the human genome (personal data and [233]). In this thesis, transcriptomic data generated by deep sequencing was used to identify transcribed Alu elements. To assess the extent to which Alu elements can be accurately identified with this technology, the number of Alu elements that could be identified by a uniquely mapping tag was calculated, as a function of the length of the tag and the number of mismatches allowed (Figure 8A). Over 90% of Alu elements could be uniquely identified by sequence reads of at least 60nt when allowing for two mismatches, which corresponds to the length given by most current deep-sequencers. The first generation of deep-sequencers returned sequences of 36nt long, and at this length up to 94% of Alu elements can still be uniquely identified when the mapping is performed without mismatches; if two mismatches are permitted for the mapping, this number drops to 66%. Moreover, since most of the transcriptomic data available was from poly-adenylated RNA, the number of Alu elements containing a poly-A segment was also estimated as this would render them prone to poly-A RNA purification techniques (Figure 8B). 80% of Alu elements (957,596 elements) contained a poly-A segment of at least six adenosines and 50% of Alu elements contained nine or more adenosines (Figure 8B). According to this analysis the transcriptional state of most Alu elements can be determined by screening deep-sequencing datasets even if those datasets are generated with first generation deep-sequencers or from poly-A purified RNA.

Figure 8: A) Proportion of Alu elements that can be identified by a unique tag in function of the length of the tag assuming zero to three mismatches. B) Number of Alu elements containing a poly-A segment of the length indicated on the x-axis.



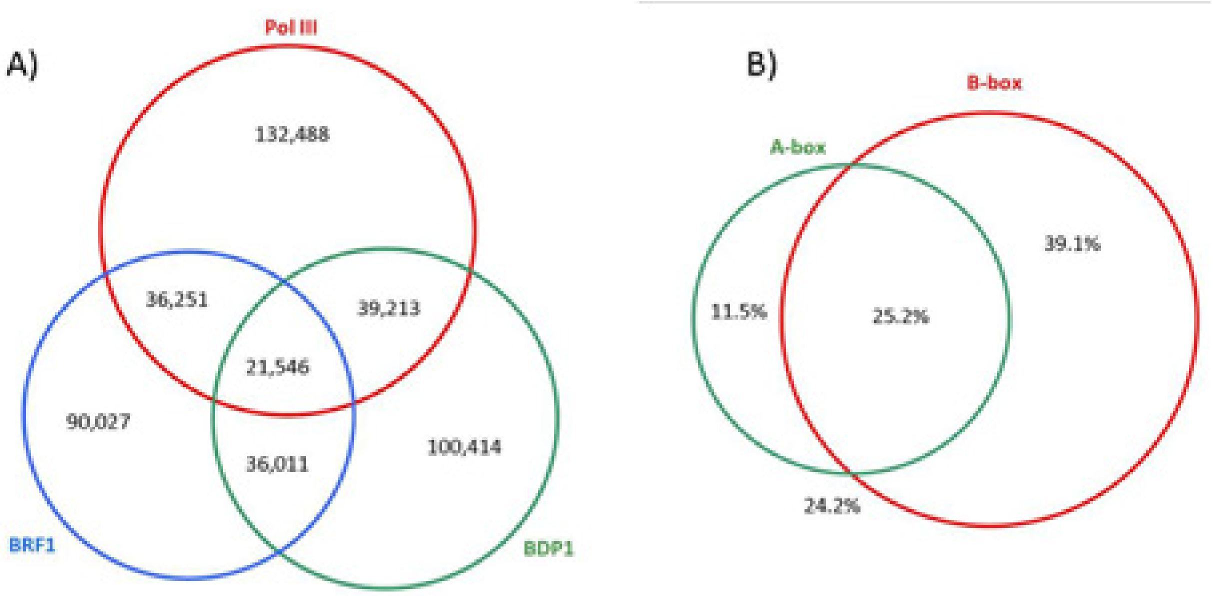
Transcription of Alu elements by POLIII

Alu elements detected in transcriptomic data can be derived from POLII transcription of genes containing Alu insert or by POLIII from transcription of the Alu element itself due to its internal promoter. To determine which elements were more likely to be transcribed by POLIII, publically available POLIII CHIP-seq data from IMR90hTert cells (human lung fibroblasts) [234] was analyzed for Alu elements binding. The pull-down assays were performed using antibodies targeting either POLIII directly or two sub-units necessary for the building of POLIII complex on the promoter, BDP1 and BRF1. The DNA associated with the proteins was sequenced with Illumina with a protocol returning 35nt long sequences (Table 1) [234]. The dataset was generated initially to identify genes transcribed by POLIII and Canella et al. focused mainly on rRNA, tRNA and others genes but no deep analysis of Alu elements bond to POLIII was performed. The dataset was reanalyzed targeting mainly Alu elements, detecting 455,950 elements that associated with at least one of the above POLIII proteins in IMR90hTert cells (Figure 9A). It was estimated that 64% of these elements contained the B-box necessary for transcription [63], and 25.2% contained both an A-box and B-box (Figure 9B). It is also important to note that 24.2% of the Alu elements that associated with POLIII did not contain POLIII internal promoter. This may reflect experimental bias or an unknown POLIII binding site, or may be the consequence of Alu loci pull-down along an adjacent POLIII transcribed gene.

Table 1: POLIII Chip-seq data summary (analyzed from [234]).

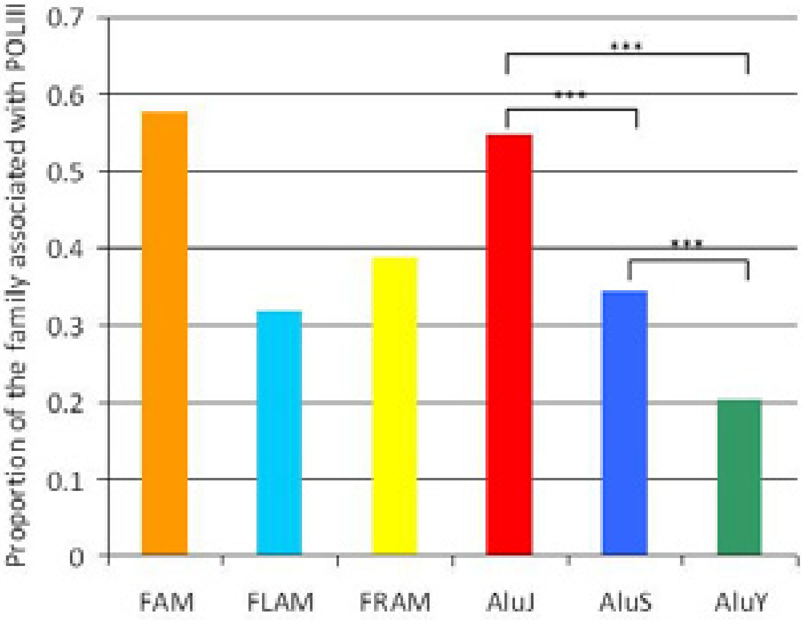
Dataset	Number of tags	Mappable tags	Tags mapping to Alu elements	Number of Alu elements
POLIII	37,655,093	15,799,677	306,777	229,498
BRF1	25,293,242	16,801,904	320,925	183,835
BDP1	20,335,637	13,490,225	292,081	197,184

Figure 9: A) Venn diagram representing the number of Alu elements present in the POLIII (red), the BRF1 (blue) and the BDP1 (green) Chip-seq dataset. B) Venn diagram representing the percentage of Alu elements detected by Chip-seq containing an A-box (green) or a B-box (red).



Previous studies reported that the young AluY family is the family predominantly transcribed in NTera2D1 pluripotent cells [67, 68]. However, the analysis on IMR90hTert cells reveals that the old AluS and AluJ families are more frequently associated with POLIII than the young AluY family, with 55%, 34% and 21% of the members of each family found associated with the POLIII complex respectively (Figure 10), which in turn suggests that the old AluS and AluJ families are more frequently transcribed than the young family AluY. To determine if this observation was due to removing all repetitive elements from the analysis, multi-mapping tags were included in a secondary mapping analysis that returned highly similar results (Pearson coefficient between uniquely mapping tags and all tags is 0.933).

*Figure 10: Proportion of each Alu family associated with the POLIII complex in the CHIP-seq experiment conducted on IMR90hTert cells (POLIII, BRF1 and BDP1 combined) [234]. The proportion is the ratio between the number of elements associated with POLIII, BRF1 or BDP1 and the number of member of each family (FAM in orange, FLAM in light blue, FRAM in yellow AluJ in red, AluS in dark blue and AluY in green). The p-values were measured using a one-tailed two proportion z-test: *** for p-value < 0.0001.*



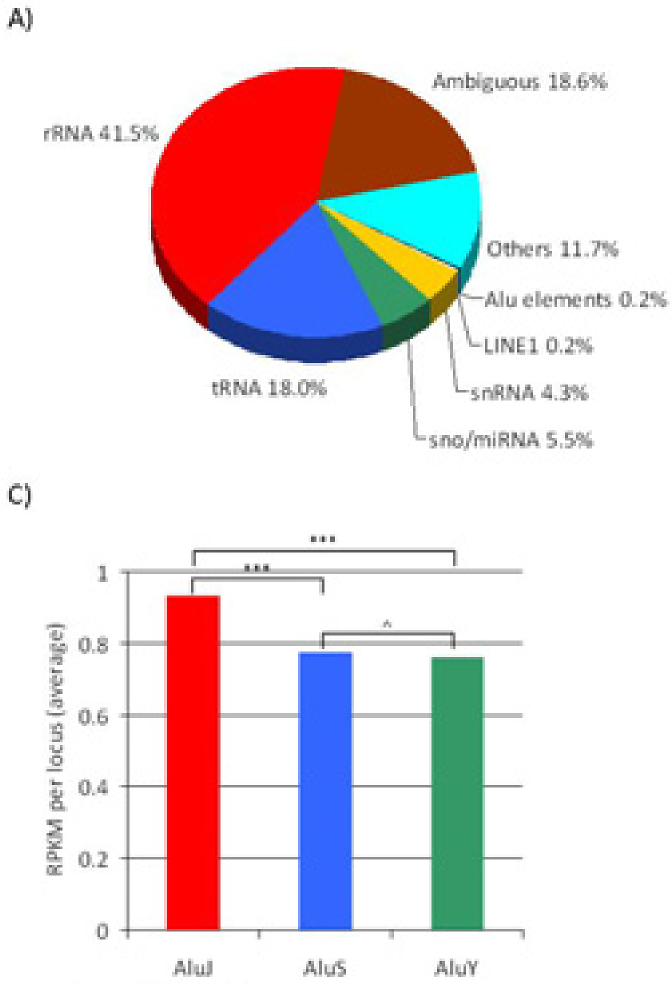
Alu elements are transcribed as small transcripts

Alu elements transcripts from 100nt to 300nt long have been previously reported [65-67], but their identification was rendered tedious because of their repetitive nature, as well as the hybridization based technology used for their detection. Current deep-sequencing technology allows this problem to be bypassed but the 50 to 350nt fraction of RNA is usually avoided in transcriptomic analyses due to the presence of tRNAs, 5S and 5.8S rRNAs, snoRNAs and snRNAs which reduces sequencing depth. Hence, this fraction is usually removed from RNA preparations by gel or column purification, to enrich for miRNA and other such small RNA fractions, or longer RNAs like mRNA and long non-coding RNAs. A direct consequence of these enrichment steps is that very little information is available about the content of the 100-300nt fraction of RNA, and no deep-sequencing dataset is available to interrogate - in an unbiased manner - the extent of Alu element expression; similarly, no deep-sequencing protocol is currently available to prepare a deep-sequencing library of this type.

To address this need a protocol to extract and deep-sequence specifically RNA from 50 to 350nt long was developed. First, the RNA is pre-fractionated on a sucrose gradient, followed by a more precise size-fractionation by polyacrylamide gel electrophoresis (PAGE). The extracted RNA is then treated with tobacco acid pyrophosphatase (TAP) and T4 polynucleotide kinase (PNK) to remove the 5' cap which prepares the extremity of the RNA for deep sequencing. Finally the RNA is deep-sequenced using a strand specific protocol that sequences the 5'end of RNA. This protocol was used with RNA from THP1 cells, a human acute monocytic leukemia cell line. This deep sequencing method returned 4,789,283 tags of 36nt representing 967,207 different sequences.

As expected, the majority of the reads were derived from rRNA and tRNA, which represented 55% and 9% of the dataset respectively (Figure 11A). 17,026 tags mapped to Alu elements, and 76% of these were uniquely mapping. The most abundant family detected was the AluS (59% of the Alu) which also represents the most numerous family in the human genome (58% of the Alu elements in hg19 are from the AluS family), followed by the old AluJ family with 22% of the Alu elements detected, and finally the young AluY representing 15% of the elements expressed in THP1 small RNA (Figure 11B). In terms of expression level, the AluJ family was significantly expressed at the highest level with an average of 0.93 read per kilobase per million of tags (RPKM), followed by AluS and AluY families which were expressed at the same level of 0.77 RPKM. These results are concordant with the POLIII chip data previously analyzed, in the sense that the family with the highest level of expression (i.e. AluJ) is also the family the most commonly found associated with POLIII. However, these results challenge previous observations that concluded the AluY family is the only one which remains transcribed [67-69]. Notably, the fact that the old AluJ and AluS families remain transcribed at a significant levels, even though the majority of them have lost their abilities to retrotranspose [69], suggests another function for these elements.

*Figure 11: A) Distribution of reads from 5'end deep sequencing of the 50 to 350nt fraction of THP1 RNA. Eight categories were distinguished in the dataset: rRNA in red, tRNA in blue, snoRNA and miRNA in green, snRNA in yellow, LINE1 in white, Alu elements in black, others in light blue correspond to all tags mapping to other regions that the ones mentioned, and ambiguous in brown correspond to all tags mapping to more than one category. B) Distribution of the number of different Alu elements detected per Alu family (AluJ in red, AluS in blue, AluY in green, FAM in brown, FLAM in yellow and FRAM in black). C) Average expression (in RPKM) of each Alu family; AluJ in red, AluS in blue and AluY in green. The p-values were measured using a one-tailed z-test: *** for p-value < 0.0001 and ^ for p-value > 0.05 (not significant).*

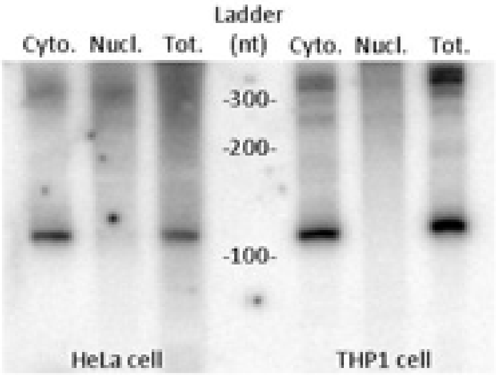


Localization of Alu transcripts

Alu elements can be detected in both the nucleus and the cytoplasm of NTera2Dl pluripotent cells, with a higher abundance of Alus in the cytoplasm [68]. These results were first confirmed by Northern blot analysis of nuclear and cytoplasmic fractions extracted from Hela and THP1 cells (Figure 12). scAlu elements of ~100nt long were detected only in the cytoplasm of Hela and THP1 cells, as expected [75]; larger bands of 280 and 300nt, corresponding to full length Alu elements, were detected in both the cytoplasm and nucleus of

Hela and THP1 cells, although expression seems more abundant in cytoplasm compared to nucleus for THP1.

Figure 12: Northern blot probing total RNA (Tot.), the cytoplasmic (Cyto.) fraction and the nuclear (Nucl.) fraction of Hela and THP1 cells for Alu elements using the probe Alu4. The RNA profiles are in Appendix 1.



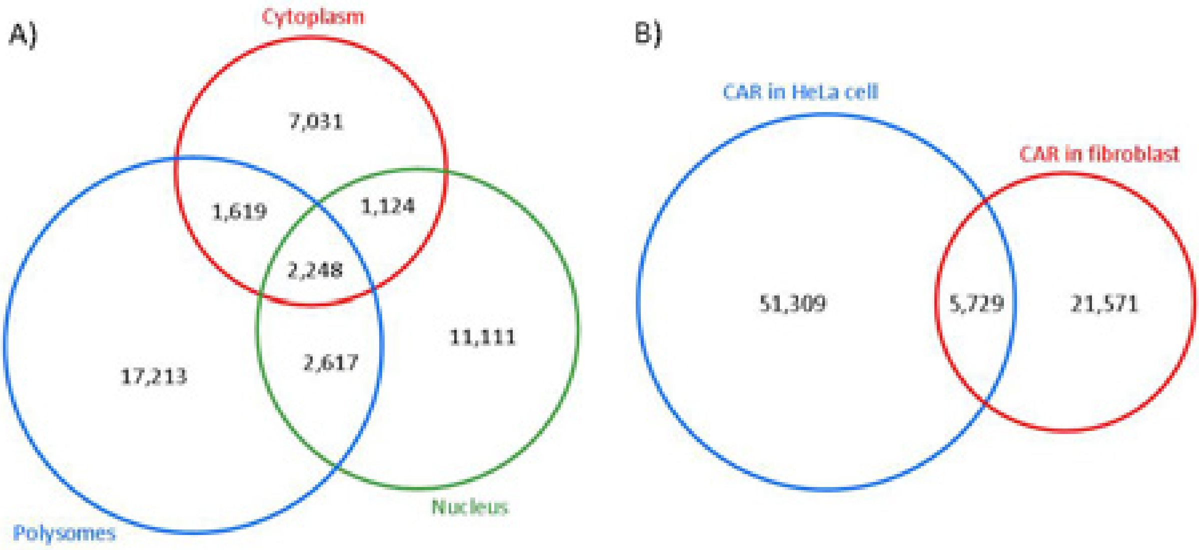
Alu elements seem to be predominant in the cytoplasm of both THP1 and Hela cells. To identify which Alus are localized in the cytoplasm or the nucleus, publically available deep-sequencing datasets from the nuclear, cytoplasmic and polysomale fractions of DLD-1 (colorectal adenocarcinoma cell line) were analyzed [235]. These datasets were generated with Illumina that returned 36nt long reads, and since they were not detecting small transcripts specifically, a pipeline was elaborated to be able to remove Alu expressed as part of a bigger transcript and select only Alu elements expressed as independent element (materiel and methods). 42,964 Alu elements were identified in either one or more of these datasets (Table 2): 11,111, 7,031 and 17,213 were observed exclusively in the nucleus, the cytoplasm and the polysomes respectively; 2,248 Alu elements were observed in the three fractions (Figure 13A). The relative abundance of Alu elements in each fraction was estimated by summing the RPKM. The cytoplasmic fraction returned an RPKM sum of 2,577, the polysomes 4,777 and the nucleus 3,581 (Table 2). In accordance with the previous Northern blot, Alu elements seem to be more abundant in the cytoplasm (polysomes + cytoplasmic fractions) with a sum of RPKM of 7354 to be compared to 3581 for the nucleus.

Table 2: Summary of the deep-sequencing data of the different cell compartments of DLD-1 cells (analyzed from [235]).

Dataset	Number of tags	Mappable tags	Mapping Alu	Number of Alu detected	Sum RPKM
Nucleus	47,120,831	31,774,535	27,399	17,100	3,581
Cytoplasm	46,354,139	31,585,388	19,579	12,022	2,577
Polysome	54,901,628	32,519,458	37,564	23,697	4,778

Two other datasets aiming at identifying chromatin associated RNA (CAR) were also screened for Alu elements. The first dataset was generated by extracting total chromatin from Hela cells, isolating the small RNA (<500nt) and deep-sequencing this fraction with GAII (Illumina). This dataset contained 13,218,503 reads 65nt long and identified 57,038 Alu elements. Using a similar protocol, Mondal et al. also sequenced chromatin associated RNA from fibroblasts [236], and their dataset contained 7,924,072 reads of 36nt long and identified 27,300 Alu elements; 5,729 Alu elements were found associated with chromatin in both datasets (Figure 13B). It is important to note that the localization of each Alu element might be dependent of the tissue in which it is expressed; this was well reflected over the datasets analyzed, where some Alus were detected in all compartment explored or also in cytoplasm of DLD-1 and associated with chromatin in HeLa and fibroblast.

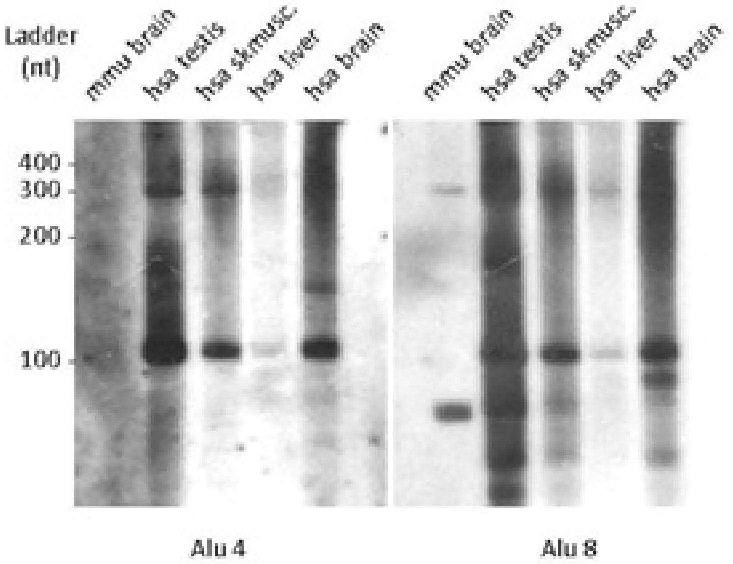
Figure 13: A) Venn diagram representing the number of Alu elements present in the cytoplasm (red), the polysomes (blue) or the nucleus (green) of DLD1 cells. B) Venn diagram representing the number of Alu elements overlapping between the two CAR datasets.



Transcription of Alu elements in normal tissues

Alu elements are discarded from most transcriptomic studies due to their repetitive nature. As such, there are only few reports of direct detection or sequencing of Alu element transcripts in cell lines [66-69, 75] and even fewer in normal tissues [72]. One paper investigated the expression level of Alu elements between normal liver and hepatocellular carcinoma and observed a higher expression of Alu elements in the tumor, but also detected POLIII transcribed Alu elements in normal liver tissue [72]. In a first step, four human tissues (brain, liver, skeletal muscle and testis) were assayed by Northern blot analysis with two Alu specific probes to screen for Alu element transcripts of ~300nt. Mouse brain RNA which doesn't contain full length Alu elements, was used to assess the specificity of the probes. Both probes gave strong bands around 100nt and 300nt in the four human tissues (Figure 14). The human brain and testis seem to contain a lot of small Alu elements since the signal is saturated. However, two different types of Alu elements were observed in these two tissues: Alu elements between 100nt and 200nt were primarily detected in human testis whereas human brain tissue expressed mostly Alu elements between 200nt and 400nt. Human liver and skeletal muscle appeared to express Alu elements in a more moderate manner and sharp bands of 100nt and 300nt long are detected in both tissues. Mouse brain RNA that does not contain Alu elements, does not show any band for Alu4 probe clearly indicating a strong specificity. One band around 300nt is, however, observed for Alu8 indicating that this probe may not be as specific as expected and we should rely more on the result given by probe Alu4.

Figure 14: Northern blot probing the RNA from four human tissues (brain, liver, skeletal muscle and testis) and one mouse tissue (brain) for Alu elements with the probe Alu 4 and Alu 8. mmu – mouse; hsa – human. The RNA profiles are in Appendix 1.



We have shown that normal tissues also express Alu elements and produce transcripts between 100nt and 300nt as observed in human cell lines, suggesting a transcription directed by POLIII. To identify which Alu elements are expressed in specific tissues, we screened transcriptomic data on sixteen normal human tissues (adipose, adrenal, brain, breast, colon, heart, kidney, liver, lung, lymph, ovary, prostate, skeletal muscle, testis, thyroid and white blood cells) produced by Illumina (GEO accession number in Appendix 2). The Atlas comprises sixteen unstranded deep-sequencing datasets of 75nt long reads from poly-adenylated RNA and one stranded dataset of 100nt long reads from a mix of the RNAs from which the sixteen tissues were depleted for rRNA (Table 3 and Figure 15). 465,160 Alu elements were detected across the sixteen tissues, the vast majority of which were present in one tissue only (Figure 16A). The tissue with the highest diversity of Alu elements was the adrenal tissue, with 193,573 elements detected. Adrenal tissue was also the tissue that presented the highest number of tissue-specific Alu elements (Figure 16B) and the highest level of expression with an RPKM sum of 53,042 (Table 3). Since transcription of Alu elements is stimulated by stress conditions [65, 71], it is notable that the organ expressing the largest panel of Alu elements is also the organ in charge of producing stress hormones when stimulated by stress conditions.

Table 3: Summary of the Illumina tissue atlas.

Library	Number of tags	Mappable tags	Mapping Alu	Nb of Alu detected	Sum RPKM
100nt	430,504,182	300,584,754	832,960	160,088	10,718
Adipose	76,269,225	69,161,694	291,503	77,780	16,250
Adrenal	76,171,569	67,379,978	937,215	193,573	53,042
Brain	64,313,204	55,956,508	333,018	121,280	22,770
Breast	77,195,260	69,356,954	461,749	116,412	25,287
Colon	80,257,757	72,722,629	252,998	55,299	13,256
Heart	76,766,862	69,401,226	241,973	90,514	13,228
Kidney	79,772,393	70,463,602	444,656	84,469	24,070
Liver	77,453,877	69,089,131	173,143	43,148	9,380
Lung	81,255,438	74,092,257	371,504	50,128	18,905
Lymph	81,916,460	71,388,339	525,414	54,900	28,021
Ovary	81,003,052	73,255,949	515,432	143,579	27,043
Prostate	83,319,902	77,176,702	339,977	51,751	16,695
Skelmusc	82,864,636	76,294,703	244,096	15,297	12,827
Testes	82,044,319	74,983,397	405,462	118,688	20,573
Thyroid	80,246,657	72,265,772	461,993	108,235	24,662
Whiteblood	82,785,673	76,047,837	256,148	26,094	12,687

Figure 15: Boxplot representing the distribution of the RPKM for all Alu elements detected in each tissue in the Illumina datasets. The width of each box is proportional to the number of Alu detected in each dataset.

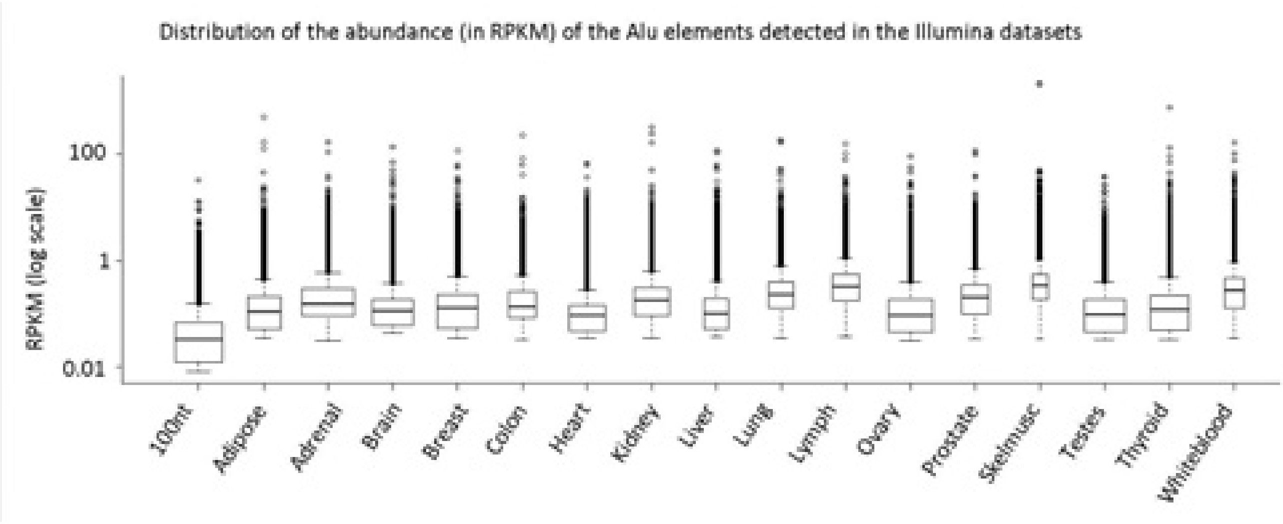
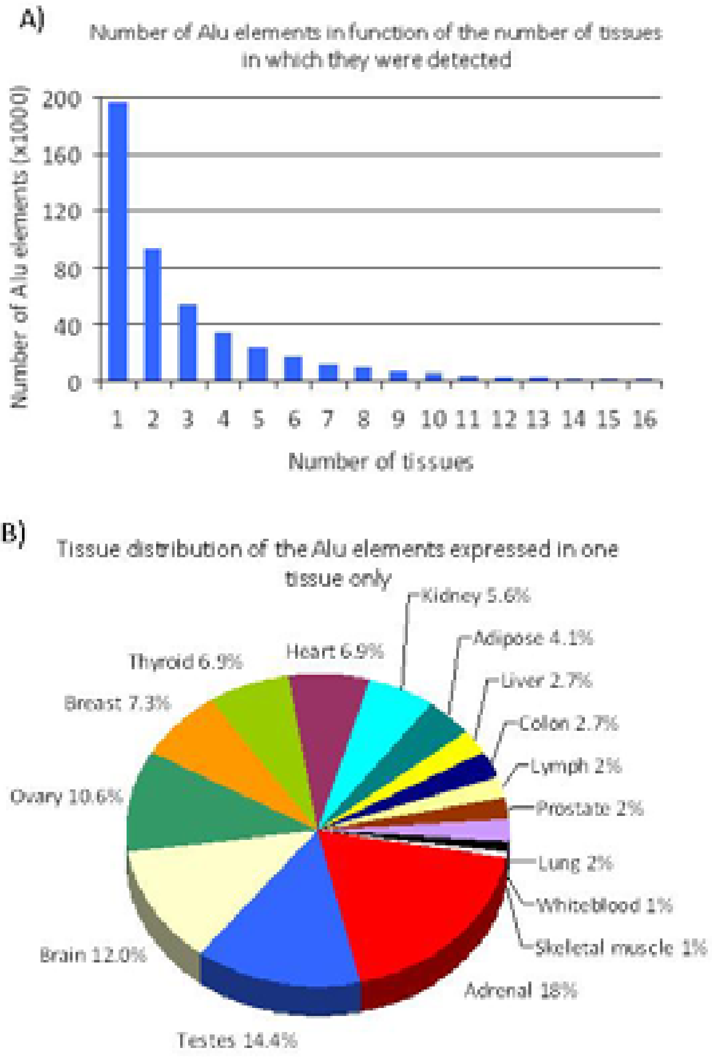


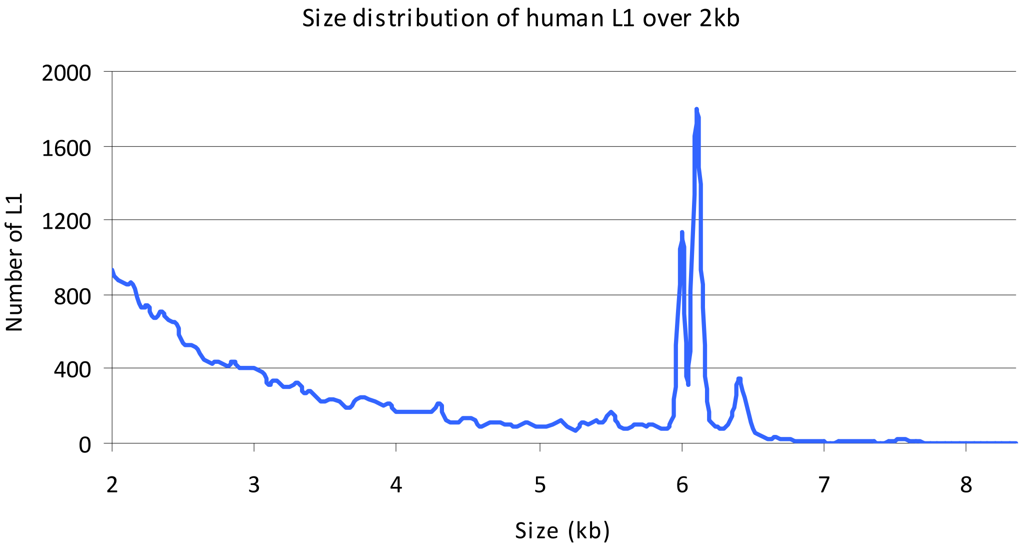
Figure 16: A) Distribution of the Alu elements as a function of the number of tissues in which they are expressed. B) Distribution of the Alu elements expressed in one tissue only, in function of the tissue in which they are expressed.



Expression of LINE1 elements in human tissues

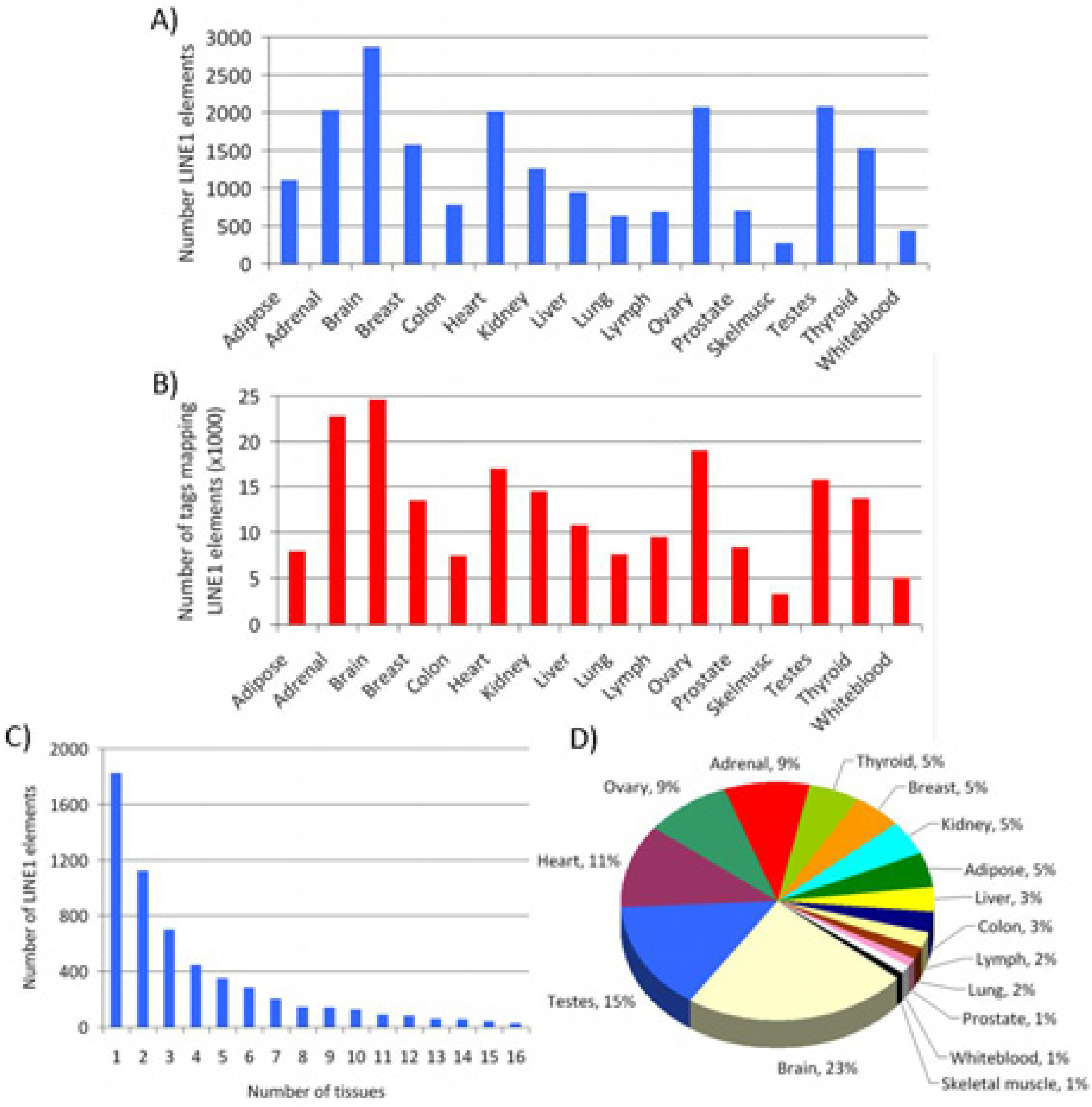
As previously mentioned, the role of Alu transcripts is normal human tissue is largely unclear. To explore the hypothesis that Alu elements retrotranspose in normal human tissues LINE1 expression was investigated, since Alu elements need the two proteins expressed by the full length LINE1 to be reverse transcribed back into the genome. Although the vast majority of LINE1 are 5'end truncated, contain internal rearrangements or harbor debilitating mutations in their ORFs [1, 237], about 5,500 of them remain full-length (i.e. ~6kb - Figure 17) and amongst these, 80 to 100 were shown to express these two key proteins and to be retrotransposable [238].

Figure 17: Size distribution of human LINE1 over 2kb. The length of LINE1s were calculated using the genomic coordinate of all human LINE1s (repeat masker hg19 [230, 239]).



The expression of the full length LINE1 in the 16 normal human tissues datasets from Illumina was determined. The organs with the highest abundance of full-length LINE1 transcripts were brain, adrenal, ovary, heart and testes with more than 15,000 tags for each tissue (Figure 18A). In terms of diversity of full length LINE1s expressed brain comes first, with more than 2,800 different elements expressed; followed by adrenal, ovary, heart and testes which expressed around 2,000 LINE1s each (Figure 18A). Like Alu elements most LINE1s are tissue specific (Figure 18B), but unlike Alu elements the brain expressed the most tissue specific LINE1s (Figure 18C). Interestingly, adrenal, brain, ovary and testis express both a significant level of LINE1 and Alu elements, strongly suggesting that those elements retrotranspose actively in normal tissues.

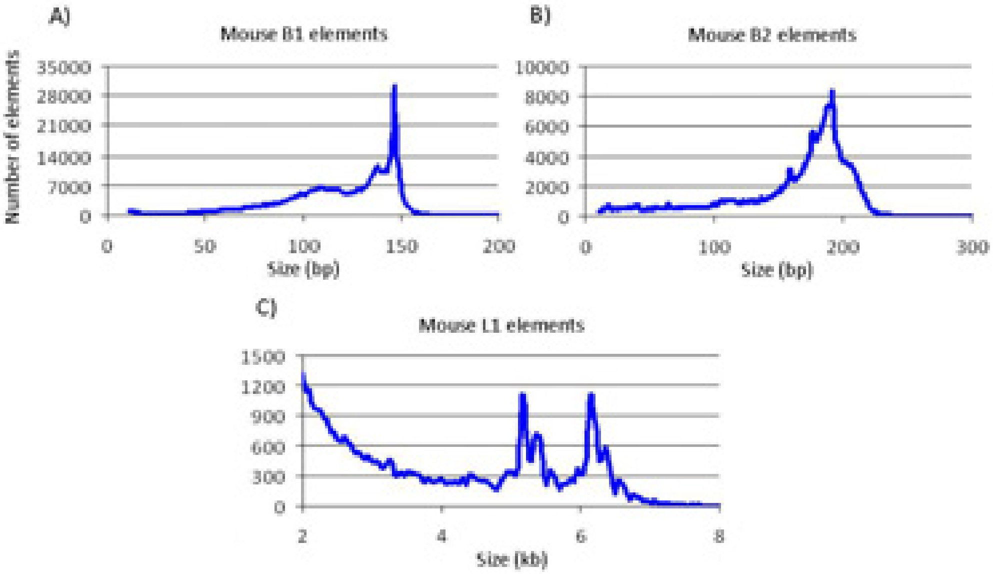
Figure 18: A) Diversity in number of different LINE1 expressed (in blue) and B) level of expression in number of tags mapping to LINE1 (in red) of full length LINE1 (>5.9kb) across 16 human tissues. C) Distribution of the LINE1 elements as a function of the number of tissues in which they are expressed. D) Distribution of the LINE1 elements expressed in one tissue only, as a function of the tissue in which they are expressed.



SINE and LINE1 expression in mouse tissues

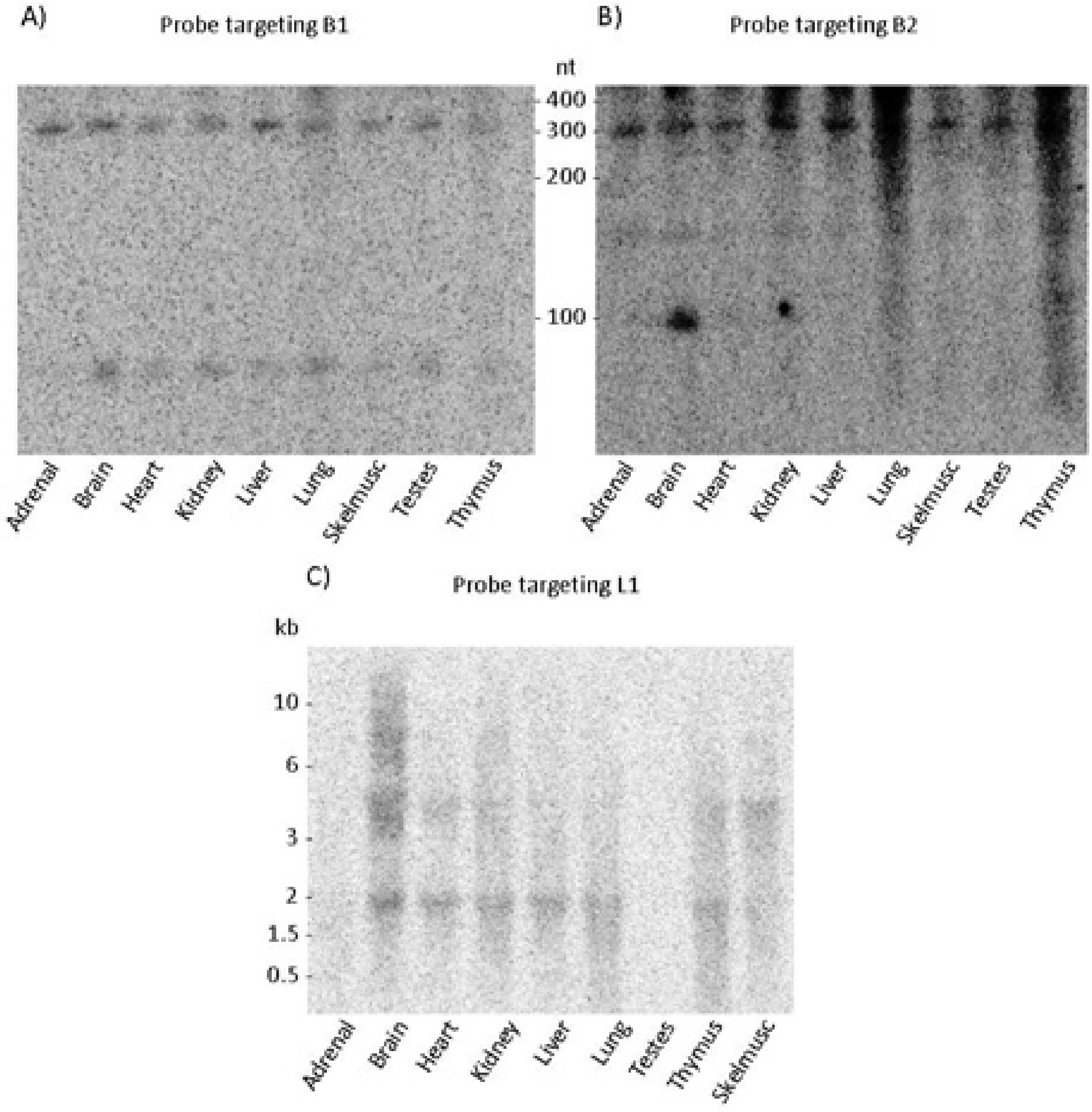
The mouse genome contains ~564,000 B1 elements that are derived from 7SL RNA. B1 elements are similar to Alu elements but are only ~140nt long, and correspond to the left arm of Alu elements [240] (Figure 19A); there are ~348,000 B2 elements which are retrotransposons derived from tRNA, each around 180nt long (Figure 19B) and which present similar properties to Alus [81, 241, 242]; and ~599,000 LINE1 elements are present in the mouse genome [216], with 15,399 of them being longer than 5kb (Figure 19C).

Figure 19: Size distribution of three transposon families in mouse. A) Number of SINE B1 in function of their size in base. B) Number of SINE B2 in function of their size in base. C) Number of LINE1 (L1) in function of their size in kilobase. The length of each repeat was calculated from its bed coordinate of all the mouse B1, B2 and LINE1 (repeat masker mm9 [230, 239]).



Given the lack of deep-sequencing data on the transcriptome of a large panel of mouse tissues, Northern blot analysis was performed using RNA from mouse adrenal, brain, heart, kidney, liver, lung, skeletal muscle, testes and thymus tissues to observe the expression of mouse retrotransposons. For B1, bands around 70nt were observed in all tissues except for adrenal tissue, and 300nt bands were observed ubiquitously (Figure 20A). The majority of B1 were around 150nt in size, the 70nt bands correspond to truncated B1, and the 300nt band may be due to cross-hybridization with 7S/L RNA or another transcript containing B1 sequences. The Northern blot probe targeting B2 shows a band around 100nt that is specific to brain, bands around 150nt in all tissues that may correspond to the full length B2 and bands around 300nt that may correspond to transcript containing B2 (Figure 20B). Finally, for LINE1, the bands around 2kb correspond to the transcripts from truncated LINE1 loci. The bands around 4-5kb expressed in all tissues and bands over 6kb expressed mainly in brain, correspond to the full length LINE1 (Figure 20C). Similar to human, the Northern blot analysis indicates that the SINE and LINE1 are co-expressed in a wide panel of tissues, suggesting that retrotransposition has the potential to occur in somatic tissues in mouse.

Figure 20: Northern-blot analysis of mouse RNA extracted from adrenal, brain, heart, kidney, liver, lung, skeletal muscle, testes and thymus with a probe targeting B1 (A), B2 (B) and LINE1 elements (C). The loading controls are in Appendix 1.



New insertion of retroelements in specific tissues

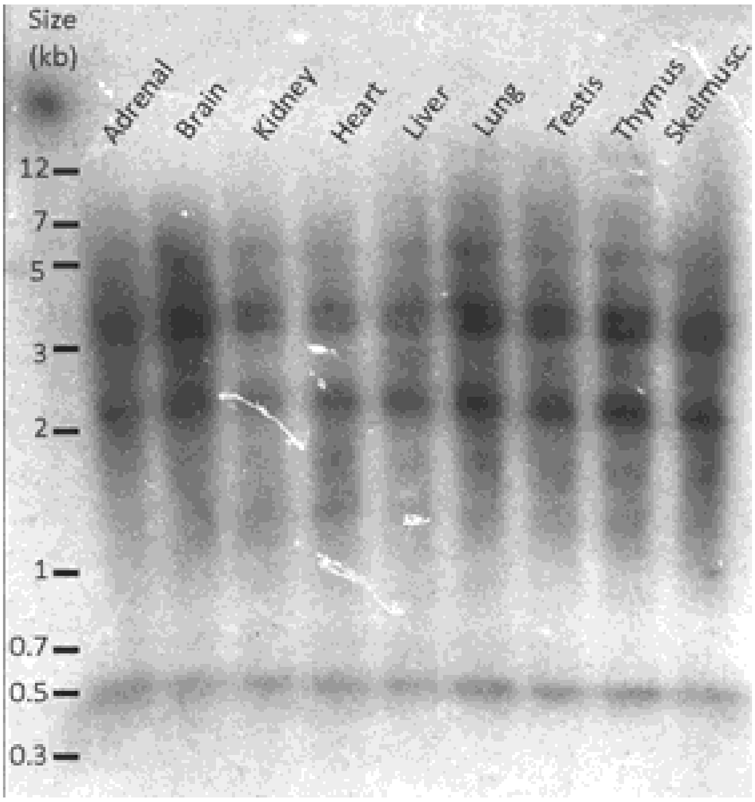
Given the above, evidence demonstrated that SINEs and full length LINE1s are co-expressed in many tissues in human and in mouse, evidence for tissue specific retrotransposition was looked for by screening each genome for new insertion events.

Analysis of retroelements insertion by Southern blot

It was rationalized that new tissue specific B2 insertions could be visualized as specific bands using Southern-blot analysis and restriction fragment length polymorphism (RFLP). Genomic

DNA isolated from 9 mouse tissues (adrenal, brain, kidney, heart, liver, lung, testis, thymus and skeletal muscle) was digested with Msp1 and probed with a B2 probe. However, no tissue specific bands were detected, most likely because the method is not sensitive enough to detect a B2 insertion event which may only occur in a restricted number of cells within each tissue (Figure 21).

Figure 21: RFLP of nine mouse tissues (adrenal, brain, kidney, heart, liver, lung, testis, thymus and skeletal muscle). The DNA was digested by MspI and probed with an oligonucleotide targeting B2.



Analysis of retroelements insertion by deep-sequencing data analysis

The Illumina tissue atlas provided extensive pair-end transcriptomic data on 16 human tissues (adipose, adrenal, brain, breast, colon, heart, kidney, liver, lung, lymph, ovary, prostate, skeletal muscle, testis, thyroid and whiteblood). To look for new insertions, we screened for pairs mapping uniquely to the genome but unrelated (that is, they would map to different chromosomes or on the same strand), these instances were referred as chimeric pairs, and were used to build a library of chimeric transcripts. It is important to note that this library is based on transcriptomic data. It is not possible to deduce from this dataset if the chimeric transcripts are due to some sort of RNA recombination or expression of a recombined gene. Another important consideration is that all tissues used in the Illumina tissue atlas come from different individuals, meaning that a tissue specific chimeric transcript may be either the result of a tissue specific recombination event or the product of a new allele specific to the individual.

282,389 loci were identified in the chimeric pairs dataset across the sixteen tissues. On average, 6.3% of the mapped pairs have unrelated mates across the 16 datasets. However, a refined analysis to confirm the distinct location of each mate of the pairs revealed that most of these chimeric transcripts were from the result of mapping errors. In these cases, the proper pair was not found because one of the mates contained mismatches which caused it to map perfectly to another incorrect location (data not shown).

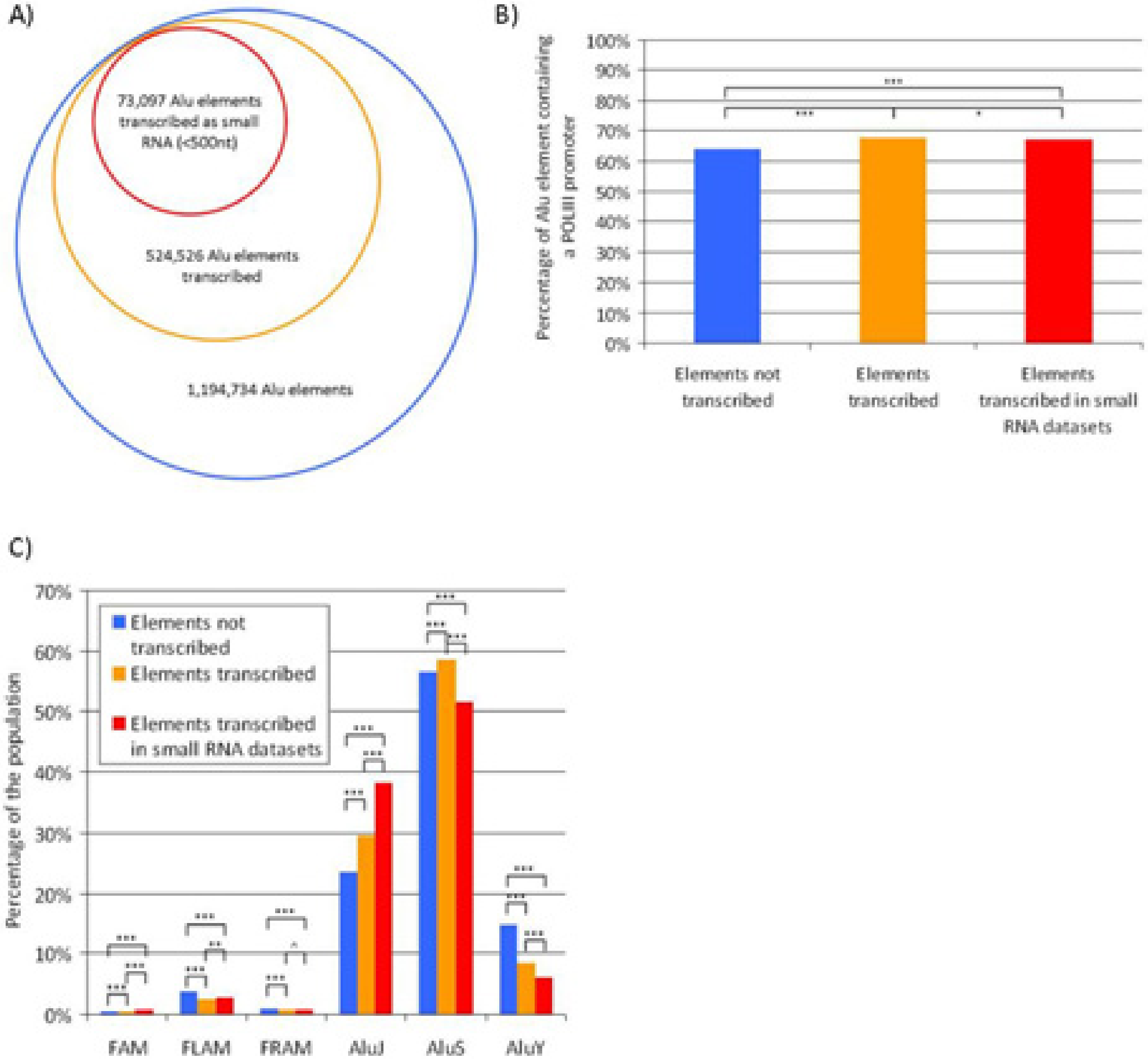
This means that direct sequencing is not efficient to detect new insertions in the genome, and an enrichment step is necessary. Recently, new insertions were discovered in human brain by coupling capture-arrays with deep sequencing. In this method, Baillie et al. [226] designed an array with probes targeting Alu and LINE1 elements, captured fragmented gDNA from five regions of the brains of three donors, and performed pair-end deep sequencing on the captured gDNA. Similar to our analysis the unpaired mates were selected to identify new insertions, and 7,743 LINE1 and 13,692 Alu elements insertion events were discovered in a tissue specific fashion. The newly inserted Alu elements were derived 83% from the AluY family and 17% from the AluS family, but it is important to note that all the probes present on the array were designed on AluY elements.

Generation of a database of Alu elements

All the information generated in this chapter was assembled in a table that contained each of the 1,194,734 Alu elements and for each described: (i) a unique identifier, (ii) the family, (iii) the bed coordinates, (iv) the presence of a POLIII promoter A-box and B-box, (v) the genomic context, (vi) the conservation, (vii) the expression in small RNA from THP1, (viii) the association with POLIII complex in IMR90hTert cells, (ix) the localization in DLD-1 cells, (x) the association with chromatin in HeLa and fibroblast cells and (xi) the level of expression in sixteen human tissues (table description in Appendix 2). Overall, 670,209 Alu elements were not detected in any of the 23 transcriptomic datasets analyzed (untranscribed group), suggesting that they are switched off, and leaving 524,526 Alu elements that are transcribed in at least one of the tissues investigated (transcribed group). Amongst the 26 datasets, only two analyzed directly the fraction containing Alu elements transcribed as small RNA (<500nt): the THP1_smallRNA and CAR_HeLa libraries; 73,097 Alu elements were detected in these two datasets (small RNA group) (Figure 22A).

The comparison of the content in POLIII promoters, for the transcribed verses untranscribed Alu element groups, revealed that transcribed elements have a slightly higher propensity to contain an effective promoter than the elements that are not transcribed (67% for transcribed Alu and 64% for untranscribed Alu: Figure 22B). Also, comparison of the family content of the transcribed and untranscribed groups showed that the old AluJ are significantly more abundant in the transcribed group than in the untranscribed group (30% versus 24% respectively). This trend is even more obvious when looking at Alu elements from the small RNA group in which AluJ represents 38% of the dataset. The young AluY, however, is more abundant in the untranscribed group compared to the transcribed group and the small RNA group with 15%, 8% and 6% respectively.

*Figure 22: A) Out of all the Alu elements (blue circle), the number of Alu elements that are detected in the database when considering only transcriptomic datasets (orange circle) and transcriptomic datasets on small RNA <500nt (red circle). B) Percentage of elements containing an effective POLIII promoter (an A-box and a B-Box or a B-Box only) for the elements detected as not transcribed (blue), transcribed (orange) and transcribed in the small RNA datasets (red). C) Distribution of the elements per family (in %) for the elements detected as not transcribed (blue), transcribed (orange) and transcribed in the small RNA datasets (red). The p-values were measured using a one-tailed two proportion z-test: *** for p-value < 0.0001, ** for p-value < 0.001, * for p-value < 0.05 and ^ for p-value > 0.05 (not significant).*



Materials and methods

Alu elements sequences, coordinates and annotations

The Alu elements library was built from the human genome (hg19/GRCh37 February 2009) using repeat masker [230]. The library contained 1,194,734 elements and a unique identifier was allocated to each element. The genomic context of each element was determined by intersecting the Alu elements positions with the positions of promoters (1000nt before transcription start site), 5'UTRs, exons, introns or 3'UTRs of the UCSC genes track containing 77,614 gene entries for hg19. All loci failing to map any of these regions were considered as intergenic. Alu elements were allocated to a region if at least 80% of their lengths overlapped with the genomic region. The conservation in primates was determined by intersecting the Alu elements coordinates with the coordinates of the blocks of conservation between human and chimpanzee, orangutan, marmoset and rhesus uploaded from the UCSC genome browser Primate Chain/Net track for hg19.

Estimation of the probability for an Alu mapping tag to be uniquely mapping

Bins of sequences from 15nt long to full length were generated from the Alu library. These bins were collapsed and tags presenting a unique occurrence were mapped with 0, 1, 2 or 3 mismatches with Bowtie [243] against the Alu masked human genome (hg19) to remove all tags mapping elsewhere in the genome. The non-mapping tags represented the uniquely mapping tags. The number of Alu elements identified by these tags was counted and the ratio [number of Alu identified by unique tags / number of Alu identified] was calculated for each bin.

Identification of A-box, B-box and poly-A segment in Alu elements sequences

Each Alu sequence retrieved from repeat masker in hg19 was screened first for containing this tag "GWTYRANNC" for the B-box and then for the following tag "KGGCNNRGTNS" for the A-box [63]. Finally each Alu sequence was screened for poly-A elements 2 to 40nt long.

Deep sequencing of the fraction 50 to 350nt from THP1 RNA

Pre-fractionation of total RNA in a sucrose gradient

A 5% to 20% w/v sucrose gradient was set up in 8.9mL ultracentrifuge tubes (Beckman Coulter #361623). Four solutions of RNAse free sucrose at 5%, 10%, 15% and 20% were prepared in 1x TBE with RNAseOUT 100U/mL and 1mM DTT. 2mL of each solution beginning by the 20% sucrose solution up to the 5% sucrose solution were successively poured in the ultracentrifuge tube and deep frozen in dry ice. The frozen gradient was then equilibrated at 4˚C for 30hours. 100µg of THP1 total RNA in 1x TBE was load on the top of the gradient [244]. The gradient was then sealed and centrifuged for 15hrs at 25000rpm at 4˚C. 500µL fractions were extracted carefully from top to bottom. The quality of the fractionation was then assessed by running 5µL of each fraction on a 1.5% agarose gel containing ethidium bromide. The RNA from the five low sucrose density fractions (determined on the gel) was then extracted by phenol/chloroform pH4. After centrifuging 15min at 12000g at 4˚C, the supernatants were collected, 2.5vol. of ethanol 100%, 1/10 vol. of sodium acetate and 1µL of glycoblue 20mg/mL were added and samples were incubated overnight at -20˚C and then centrifuged at 12000g at 4˚C for 25min. The pellets were washed two times with 70% cold EtOH, dried and finally all pellets were pulled in 32uL of DEPC water.

Precise RNA fractionation by PAGE

The RNA was mixed with 1vol. of loading buffer (95% formamide, 18mM EDTA, 0.025% SDS, 0.00025% of Xylene cyanol and 0.00025% of bromophenol blue), denatured at 65˚C for 5min and run on a polyacrylamide gel (8% acrylamide/bis-acrylamide (19/1), 1x TBE and 480g/L urea) in 1xTBE at 150V for 40min. The gel was then stained 10min in 1x TBE containing ethidium bromide (5µg/L) and the band corresponding to the fraction 70 to 400nt length RNA was visualized and cut under UV light. The gel band was then shredded in small pieces with an RNAse free blade and the RNA was eluted from the gel fragments overnight at room temperature in 150µL of elution buffer (20mM Tris pH 7.7, 150mM NaCl, 0.2mM EDTA pH 8.0, 0.5% v/w SDS). The eluted RNA was purified and precipitated by phenol/chloroform and ethanol precipitation as described previously and resuspended in 15µl DEPC-treated water.

Treatment of the extremities of the RNA for deep sequencing

The RNA was then treated with 20U of tobacco acid pyrophosphatase (Epicentre #T19050) for 1hr at 37˚C in 20µL following manufacturer's instructions. 6µL of 5x ligation buffer (250mM Tris-HCl pH 7.6, 50mM $MgCl_2$, 5mM ATP, 5mM DTT, 25% w/v PEG8000) and 30U of T4 polynucleotide kinase (NEB #M0201S) were added to the TAP treated RNA, the mix was incubated 1hr at 37˚C and the RNA was cleaned and precipitated by phenol/chloroform and ethanol precipitation as described previously and re-suspended in 15µl DEPC-treated water. The RNA was deep-sequenced with Illumina using a protocol returning 36nt long read corresponding to the 5'end of the transcripts.

Identification of Alu elements in deep-sequencing datasets

For each dataset screened, the following procedure was employed: first the dataset was uploaded in fasta format for the first generation of deep-sequencing data or in fastq format for the more recent data. The 3' adapters were trimmed and the tags were collapsed using the fastx toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html). The dataset was then mapped with Bowtie [243] against the human genome (hg19) from which Alu elements have been masked using an in house algorithm (maskOutFa). The remaining unmappable tags were then mapped with Bowtie against a splice junction library build from Gencode, UCSC, Refseq and Emsembl gene prediction to get rid of all tags mapping spliced mRNA. Finally, the remaining unmappable tags were mapped against the Alu library; uniquely mapping tags and multi-mapping tags were treated separately. RPKM (Read per Million tags per kilobase) were calculated for each Alu elements using the total number of mapping tags (against masked hg19, splice library and Alu library) and the length of each Alu elements to normalize the score. Finally, to exclude Alu elements included in longer transcripts, all Alu elements were mapped to a transcriptome index build with Cufflinks [245], integrating the data from the 16 tissues. Each analysis considered only the transcripts <500nt long.

Design of probes targeting Alu elements for Northern blot

One consensus sequence per Alu family (AluY, AluS, AluJ, FRAM, FLAM and FAM) was generated with all the family members using Clustalw [246]. 30 oligo-nucleotides of 25nt were designed per Alu consensus sequences. Each probe was then rated in function of the number of Alu elements it targets (which needs to be high), its similarity with 7SL RNA (which needs to be null), then the number of times it maps to non-Alu related regions of human region (which needs to be low) and finally the number of ESTs it maps to and the secondary structure. 2 probes out of 180 generated were selected for Northern blot (Table 4)

Table 4: Properties of the probes selected for Northern blot. Columns 2 to 7 represent the fraction of each family detected by each probe, column 8 the number total of Alu elements detected, column 9 the number of unrelated loci in the genome, column 10 the number of unrelated transcripts and column 11 the secondary structure of the probe.

Name	Alu S	Alu J	Alu Y	FAM	FRAM	FLAM	Nb Alu	Unrelated targets	ESTs	2nd struct
Alu4	8.1%	0.5%	86.1%	0.1%	0.0%	0.2%	183,891	106	11	weak
Alu8	9.8%	0.0%	82.8%	0.0%	0.1%	0.1%	189,169	95	7	moderate

Northern blot on nuclear and cytoplasmic fraction of Hela cells and THP1 cells

The fractionation of the THP1 RNA and Hela RNA was done by Anupma Choudhary (University of Queensland, Institute for Molecular Bioscience, lab-book 4940, pages 54, 73 and 88), the efficiency of the fractionation was assessed by observing the RNA profile (tRNA, 5S and 5.8S rRNA were observed only in the cytoplasm whereas bands corresponding to snRNA remained in the nucleus) and by PCR using snoRNA primers and tRNA primers. Respectively, 10ug, 1.04ug, 8.96ug of total, nuclear and cytoplasmic RNA from THP1 and 26.67ug, 6.67ug and 20ug of total, nuclear and cytoplasmic RNA from HeLa cell were loaded in a 4M urea 8% polyacrylamide gel. The amount of RNA loaded on the gel was adjusted for each sample to correspond to the amount of RNA present in cytoplasm and nucleus in the two kinds of cells: the ratio nucleus RNA/cytoplasmic RNA is 0.116 for THP1 cells and 0.33 for Hela cells. After migration of the RNA at 200V for 3hrs, the gel was transferred on a nitrocellulose membrane (GE Healthcare Hybond N-+) with a semi-dry transfer unit and incubated overnight with Alu4 probe radioactive labeled using terminal transferase (NEB#M0315) and alpha 32P-dCTP (Perkin Elmer #NEG513H-500uCi). The picture of the membrane was taken using a phosphoimager.

Northern blot on RNA from four human tissues and one mouse tissue

5ug of human brain, testis, skeletal muscle and liver (Ambion tissue panel #AM6000) and 5ug of mouse brain RNA (negative control) were run in a gel, transferred and probed with Alu 4 and Alu 8 as described previously.

Northern blot on mouse RNA

Total RNA from adrenal, brain, heart, kidney, liver, lung, skeletal muscle, testes and thymus were extracted from freshly dissected organs from a mouse BL6C57 male using Trizol (Invitrogen #15596-026) according to the manufacturer's instructions. For small RNA Northern-blot, 5µg of RNA of each tissue were run in a urea gel and transferred as previously described and probed with a B2 probe (TCCTGCTTGGACCAGCCTCA) or a B1 probe (TGTAGCTTTAGCTGGCCCAGAACT). For Northern-blot on large RNA, 5µg total RNA from the 9 tissues were denatured in 1X MOPS (0.1M MOPS, 50mM EDTA pH8.0, 200mM Na-acetate pH7), 50% formamide and 5% formaldehyde for 15min at 65˚C and run on a MOPS-agarose gel (1X MOPS, 1% formaldehyde) in 1X MOPS at 70V for 2 hours. The gel was then transferred on a nitrocellulose membrane and probed with a LINE1 probe (CCCCTACGCACCCTCTCCCA) as previously described.

Extraction of chromatin associated RNA from HeLa cell

Extraction of the nuclei of HeLa cells

HeLa cells were grown in 4 flasks of 400mL at 37˚C in 5% CO2 in DMEM containing 4mM L-glutamate, 4.5mg/ml glucose and 0.11mg/ml sodium pyruvate, supplemented with 100U/ml Penicillin and 100µg/ml Streptomycin until >90% confluence (approx. 10^7 cells per dish). The cells were collected by trypsinization, rinsed twice with PBS and resuspended in 5mL of buffer A (10mM Hepes, pH7.9, 10mM KCl, 1.5mM $MgCl_2$, 0.5mM DTT). After 5min of incubation on ice, the suspension was transferred in a pre-cooled 7mL Dounce tissue homogenizer (Wheaton Scientific Product Cat no: 357542) and homogenized ~30 times using a tight pestle (Kontes pestle B: 0.0010" - 0.0030" clearance), while keeping the homogenizer on ice. The homogenized cells were then collected at 218g (1000 rpm, Beckman GS-6 centrifuge, GH-3.8 rotor) for 5 min at 4°C. The pellet contained enriched, but not highly pure, nuclei. The nuclear pellet was resuspended in 3ml of S1 (0.25M Sucrose, 10mM $MgCl_2$), layered over 3ml of S3 (0.88M Sucrose, 0.5mM $MgCl_2$) and centrifuged at 3000g (3500 rpm, Beckman GS-6 centrifuge, GH-3.8 rotor) for 15min at 4°C. This pellet contained pure nuclei.

Extraction of the chromatin

The sucrose gradient was prepared according to the protocol developed by Luthe [244] in 8.9mL ultracentrifuge tubes (part number 361623, tube polyallomer, optiseal 8.9mL, 16x60mm). Two solutions of RNAse free sucrose at 5% and 40% were prepared in a solution of 100mM NaCl, 10mM Tris-HCl (pH 8.0), 0.2 mM EDTA, 10mM vanadyl ribonucleoside complex, 50U RNaseOUT/mL and 1mM DTT. 4mL of 40% and 5% sucrose were successively poured and deep frozen in dry ice in the ultracentrifuge tube. The frozen gradient was then equilibrated at 4°C for 30hours.

The pure nuclei were resuspended in 2mL of PXL lysis buffer (1X PBS, 0.1% SDS, 0.5% deoxycholate, 0.5% NP-40, one tablet of protease inhibitor cocktail (Roche), 100U RNaseOUT/mL and 1mM DTT per 10mL solution) and treated 5min at 37°C with 4µL of micrococcal nuclease 0.2U/uL (Sigma N3755) and 20µL of 0.1M $CaCl_2$. The reaction was stopped by adding 200µL of 0.5M EDTA and incubating on ice for 2min. The nuclei were then disrupted by pipetting and centrifuged at 8000g for 2min at 4°C to pellet insoluble nucleus components. The soluble chromatin was loaded on a sucrose gradient (5% to 40% sucrose) and centrifuged for 15hrs at 22000rpm at 4°C. The gradient was then fractionated in 500uL fractions and dot-plot was used to determine which fractions contained histones. The RNA from the selected fractions was extracted with Trizol LS (Invitrogen #10296-028) according to the manufacturer's instructions.

Extraction of the small RNA associated with the chromatin

The RNA from the selected fractions was run in a urea-PAGE (see the section deep sequencing of the fraction 50 to 350nt from THP1 RNA described previously) and small RNA <500nt from the early fractions of the gradient containing high percentage of sucrose and heavy complexes were cut out from the gel, eluted in elution buffer (20mM Tris pH7.7, 150mM NaCl, 0.2mM EDTA pH8.0, 0.5% v/W SDS) overnight at room temperature. The eluted RNA was deep-sequenced with Illumina using a protocol returning 65nt long stranded reads (v1.5 Illumina).

New genomic insertion of retroelements

Restriction Fragment Length Polymorphism

Genomic DNA was extracted from freshly dissected tissues (adrenal, brain, kidney, heart, liver, lung, testis, thymus and skeletal muscle) from one mouse BL6C57 male according to Strauss protocol [247]. The Southern-blot was performed according to Brown protocol [248]. For each tissue, 10µg of gDNA was treated with 40U of MspI (NEB # R0106S) overnight at 37˚C, the enzyme was then deactivated for 20min at 80˚C and the samples were run in a 0.7% agarose 1XTBE gel. The gel was depurinated 10min in 0.25M HCl, rinsed in water and treated by 0.5M NaOH and 1M NaCl to denature the DNA. The gel was then transferred by capillarity on a nitrocellulose membrane (GE Healthcare Hybond N-+) using 0.025M $NaPO_4$ pH6.5. The membrane was cross-linked with 120mJ of UV and prehybridized at 55˚C for 2hrs in 1% BSA, 1mM EDTA, [0.5M Na]HPO_4 pH7.2 and 7% SDS. The B2 probe used for the mouse northern was treated with terminal transferase as previously described and the membrane incubated with the radiolabelled probe overnight at 55˚C. The membrane was washed 4x10min at 55˚C with 1mM EDTA, [40mM Na]HPO_4 pH7.2 and 1% SDS. The picture of the membrane was taken using a phosphoimager.

Deep sequencing based approach

To look for new genomic insertion, the 50nt long pair-end dataset on 16 human tissues from Illumina (http://www.ncbi.nlm.nih.gov/projects/geo/query/acc.cgi?acc=GSE30611) was screened for pairs for which the two mates mapped uniquely to the genome but are unrelated (on different chromosomes, on the same strand). First, the original fastq file were mapped using Tophat [249] selecting for uniquely mapping tags. Then, the tags for which both mates were mapped but labeled as unpaired were selected using SAMtools [250]. Each tag of the unmated pairs of tags was classified as intronic, exonic, 3'UTR, 5'UTR, Alu element, LINE1, other repeat (repeat masker without LINE1 and Alu) or intergenic and the chimeric transcripts built by these unrelated pairs were built using Cufflinks [245]. The 50 transcripts presenting the highest number of chimeric pairs were looked at in the UCSC genome browser [251] and each component of the chimera was remapped using Blat [252].

Chapter 3: Identification of RNA edited by ADAR proteins

A-to-I RNA editing was first discovered when A-to-G substitutions were systematically observed in the comparison between transcripts encoding certain ion channels and ligand-gated receptors and their genomic counter-parts. This phenomenon was later discovered as being essential for normal life in metazoans and, more specifically, was critical for the development of the nervous system in mammals. Extensive genome-wide comparisons between cDNA and gDNA emphasized the fact that A-to-I editing is not a mechanism limited to a few genes but occurs throughout the transcriptome.

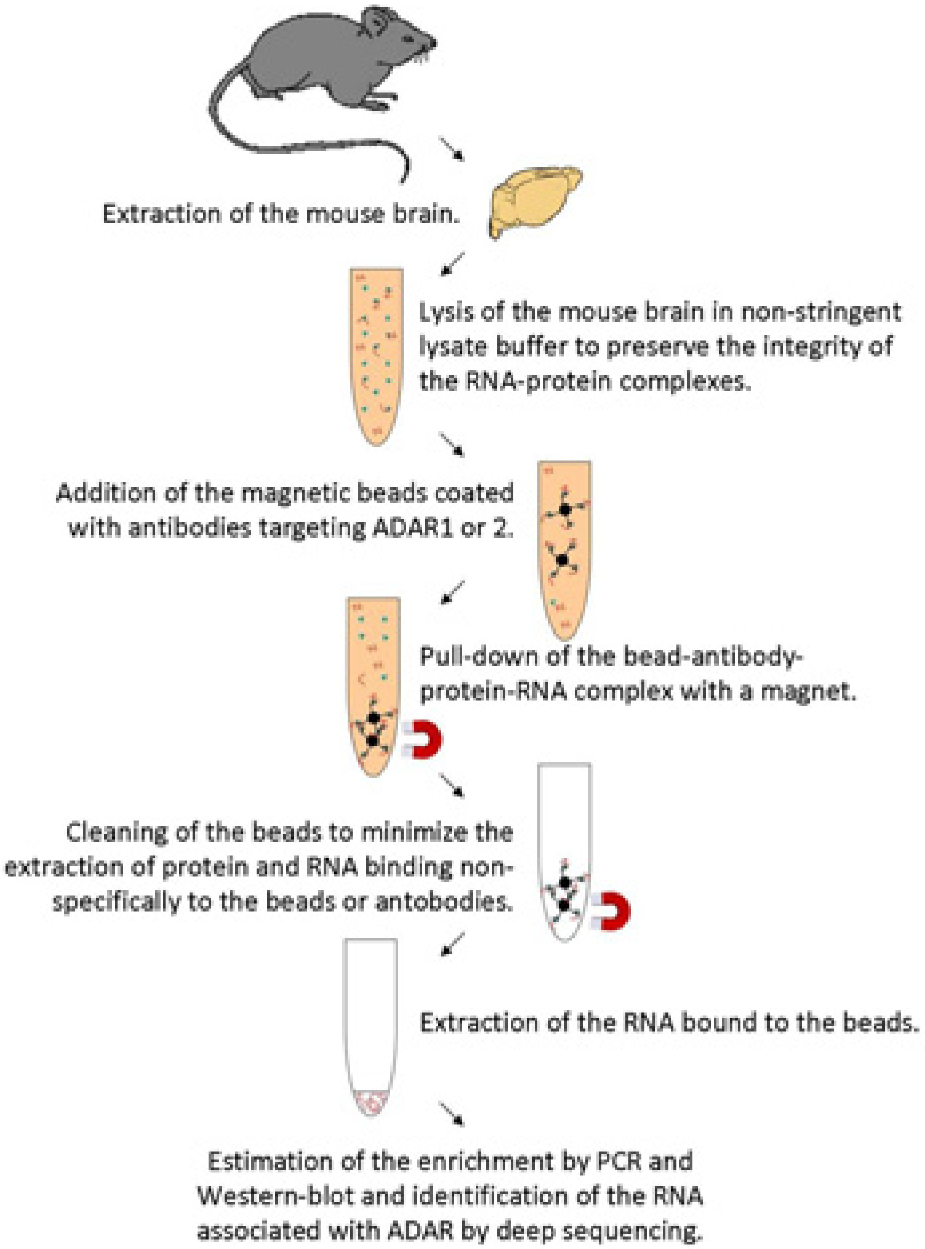
To more precisely characterize A-to-I editing in the transcriptome, a protocol to identify the targets of A-to-I editing enzymes on a large scale and in an unbiased way was developed. Initial analyses attempted to isolate RNAs associated with ADAR proteins using an immunoprecipitation protocol, as described by Ohlson et al [217, 218] followed by deep sequencing of the isolated transcripts. However, despite successive refinements and optimizations of this protocol, the technique was not robust enough to reach the level of enrichment and quantity of RNA required. An alternate strategy was then developed based on an adaptation of Morse and Bass' protocol for deep sequencing. Proof of concept experiments performed on mouse brain RNA successfully assessed the enrichment of edited targets by PCR, were followed by deep sequencing analysis, comparison to known editing sites and screened for new editing sites.

Isolation of RNA bound to the ADAR proteins

Standard immunoprecipitation

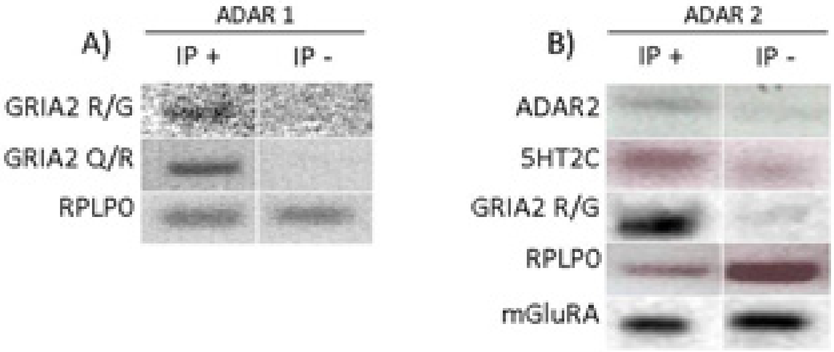
The first approach attempted to identify RNA associated with ADAR proteins using co-immunoprecipitation (co-IP), based on the protocol developed by Ohlson et al. [217, 218]. Briefly, fresh mouse brain was dissected and lysed and an ADAR1 or ADAR2 specific antibody bound to magnetic beads was added to the lysate. RNA-ADAR-antibody-magnetic beads complexes were collected with a magnet and the RNA extracted from the complex (Figure 23). The specificity of the immunoprecipitation was then estimated by PCR using primers that target transcripts which are known to be edited, such as GRIA2 R/G, GRIA2 Q/R, ADAR2 and 5HT2C compared to unedited transcripts like ARPP0 and mGluRA; Western-blot analysis was also performed.

Figure 23: Outline of the immunoprecipitation protocol. RNA is represented by red lines and ADAR protein by green triangles.


Extraction of the mouse brain.
Lysis of the mouse brain in non-stringent lysate buffer to preserve the integrity of the RNA-protein complexes.
Addition of the magnetic beads coated with antibodies targeting ADAR1 or 2.
Pull-down of the bead-antibody-protein-RNA complex with a magnet.
Cleaning of the beads to minimize the extraction of protein and RNA binding non-specifically to the beads or antobodies.
Extraction of the RNA bound to the beads.
Estimation of the enrichment by PCR and Western-blot and identification of the RNA associated with ADAR by deep sequencing.

After successive refinements of the protocol including modification of the brain lysate preparation (lysate buffer from Boyer et al. [253]) and final RNA extraction (using Trizol as described by Niranjanakumari [254]), the RT-PCR showed clear enrichment of GRIA2 in the ADAR1 co-IP, and similarly ADAR2, 5HT2C and GRIA2 were present in the ADAR2 co-IP as compared to the negative control IP- performed with pre-immune serum (Figure 24). However, Western-blot analysis failed to show any ADAR1 or ADAR2 protein in the co-IP (data not shown), perhaps because the amount of protein present in the IP was too low to be detected. Another concern was the strength of the signal of the nonspecific targets RPLP0 and mGluRA in the IP, for which the bands detected were stronger than the specific targets GRIA2, ADAR2, and 5HT2C; this suggested that subsequent sequencing would return many nonspecific transcripts.

Figure 24: RT-PCR followed by PCR amplification of co-IP purified RNAs associated with ADAR1 and ADAR2 in mouse brain. A) PCR signal of cDNA enriched from ADAR1 co-IP with GRIA2 R/G, GRIA2 Q/R and RPLP0 primers. B) PCR signal of cDNA enriched from ADAR2 co-IP with ADAR2, 5HT2C, GRIA2 R/G, ARPP0 and mGluRA primers.

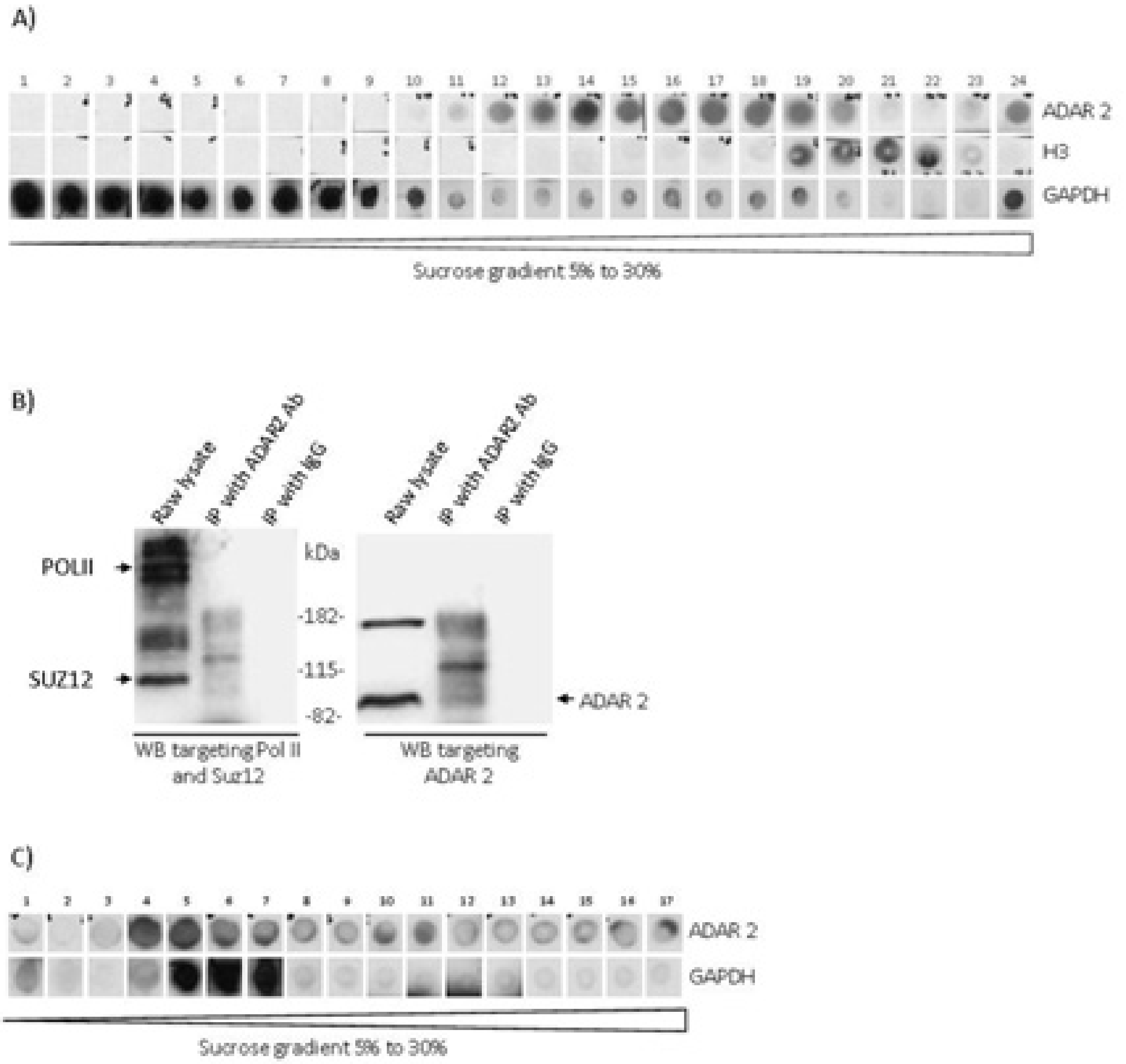


Immunoprecipitation using sucrose gradient purified fractions

One way to reduce the amount of nonspecific transcript in the co-IP and increase the yield is to pre-purify the complex of interest. This reduces the complexity of the lysate and increases the concentration of the complex compared to nonspecific elements prior to co-IP. The first method of pre-purification attempted was by sucrose gradient, which can separate ribonucleoprotein complexes (RNP) according to their density. First, mouse brain lysate was fractionated on a sucrose gradient ranging from 5% to 30%. Detection of fractions that contained ADAR2 complexes was carried out by blotting an aliquot of each fraction on nitrocellulose, followed by detection using specific antibodies targeting ADAR2. Fractions containing ADAR2 were pooled and IP using ADAR2 antibodies was carried out. The dot plot on each fraction of the sucrose gradient revealed that ADAR2 RNPs migrated in fractions 12 through 19; the gradient fractions were also probed with antibodies targeting H3 and GAPDH and showed that GAPDH migrated to fractions 1 through 10, and H3 to fractions 19 through 23 (Figure 25A). This clearly indicated that the sucrose gradient was reducing the complexity of the lysate by removing all the complexes of lower or higher density than the ADAR2. Fractions 12 through 19 (those containing ADAR2) were pooled to perform the ADAR2 IP; a negative control IP using IgG was also performed, and Western-blots were used to monitor the enrichment of ADAR2 in the IP. The Western-blot analysis using antibodies targeting POLII and SUZ12 showed that the ADAR2 IP was efficient at removing unrelated protein from the mix: the strong signal observed in the pool of fractions for SUZ12 and POLII completely disappeared in the IP product. The Western-blot analysis with an antibody targeting ADAR2 showed bands corresponding to ADAR2 in the fraction 12-19 pool, a weak band in the ADAR2-IP product, and no band in the IP-IgG (Figure 25B). The reduction in nonspecific target and detection of ADAR2 in the ADAR2-IP indicated that ADAR2 was specifically pulled down, although the weak ADAR2 signal in the Western-blot analysis indicated that IP enrichment was generally low, and subsequent extraction and analysis of RNA from this sample did not return any detectable signal by PCR (data not shown). This may be due to the integrity of the ADAR2-RNA being lost due to the buffer used for the fractionation, which could explain the detection of the ADAR2 protein but lack of associated RNA.

To control for this the original buffer used for the sucrose gradient was depleted in EDTA and supplemented with 100mM NaCl and 2mM $MgCl_2$ (i.e. physiological conditions) which could help to preserve the structure of the RNP. The results from this experiment demonstrated that the migration profile of the different proteins detected was completely different from that observed without NaCl and $MgCl_2$. Instead of being spread throughout the low-density fractions, GAPDH was concentrated in fractions 5 to 7, whereas ADAR2 was spread from fractions 3 to 17 (Figure 25C). This suggests that under physiological conditions, GAPDH assembles into homogeneous complexes of the same density, whereas ADAR2 assembles in heterogeneous complexes across all densities. This is consistent with a specific function and a limited number of partners for GAPDH, and a wide range of targets of different sizes for ADAR2. Unfortunately, this also implies that sucrose gradient fractionation is not a robust approach to concentrate ADAR2 RNP.

Figure 25: A) Dot-plot screening for ADAR2, H3 and GAPDH in the 24 fractions of a sucrose gradient from 5% to 30% sucrose in which mouse brain lysate was separated. B) Western-blots targeting POLII and Suz12 or ADAR2 in the pool of fractions 12 to 19 from the sucrose gradient, in ADAR2 IP performed on this pool and in IP- with IgG). C) Dot-plot screening for ADAR2 and GAPDH in the 17 fractions of a sucrose gradient from 5% to 30% sucrose containing 100mM NaCl and 2mM MgCl$_2$ in which mouse brain lysate was separated.



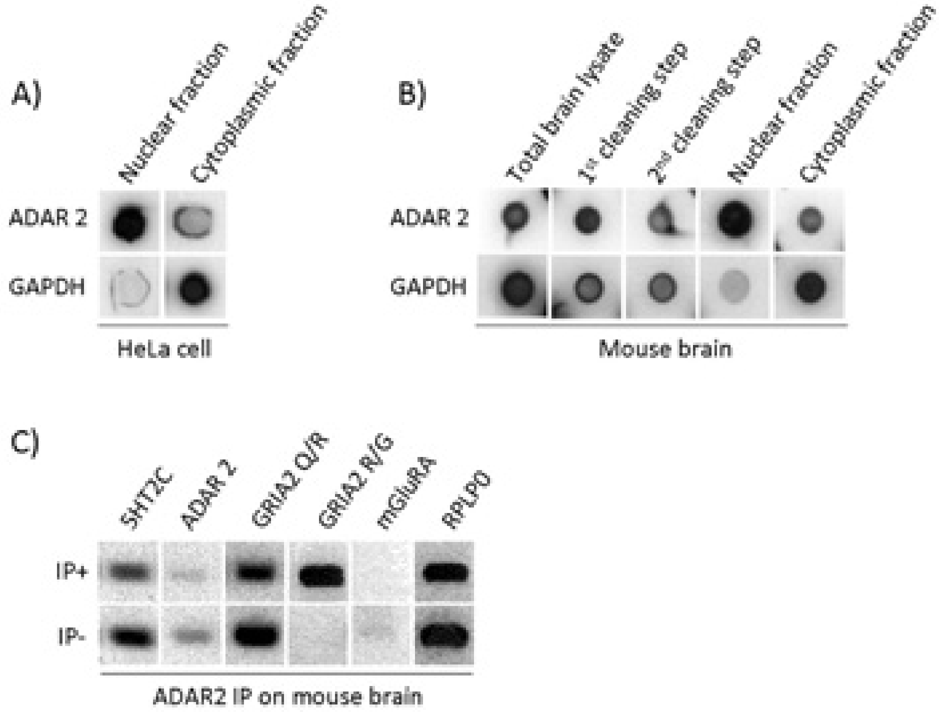
Immunoprecipitation using isolated nuclei

ADAR proteins are present in both the nucleus and the cytoplasm of cells. However, A-to-I editing of mRNA is believed to occur mainly in the nucleus prior to splicing. Since the sucrose gradient demonstrated limited capacity to enrich for ADAR2 RNP, IP was attempted on nuclei extracted from mouse brain. In a preliminary experiment, ADAR2 location was investigated by dot-plot screening of the nuclear and cytoplasmic fraction of HeLa (Figure 26A). The two cellular fractions were separated as follows [255]: cells were first ruptured in a mild lysis buffer using a dounce homogenizer with a tight interstice that breaks the cytoplasmic membrane but leaves the nuclei intact. A centrifugation step then separated the cytoplasmic fraction (supernatant) from the nuclei (pellet), and the cytoplasmic fraction was collected. The nuclei were then rinsed twice, "filtered" on a sucrose cushion and lysed with sonication using a strong lysis buffer. The soluble fraction of the lysed nuclei was collected. Dot-plot analysis (Figure 26A) showed a clear enrichment of GAPDH in the cytoplasmic fraction of HeLa cells, with the majority of ADAR2 remaining in the nucleus.

This protocol was then adapted to total tissue, where the cells from the tissue were separated using a dounce homogenizer with a large interstice before following the procedure described above. Each step of the protocol showed an enrichment of ADAR2 and a reduction of GAPDH that was used as a cytoplasmic marker (Figure 26B). This indicated that nuclei fractionation was an efficient method to enrich for ADAR RNP before the IP.

The IP was then carried out as described previously, with ADAR2 antibodies on lysates prepared from the purified nuclei and using pre-immune serum as a negative control for the IP; Western-blot analysis targeting ADAR2 did not show any bands in the co-IP products (data not shown). The PCR on 5HT2C, ADAR2 and GRIA2 Q/R did not show enrichment in the co-IP compared to negative control, although GRIA2 R/G was successfully enriched in the co-IP. Finally, the levels of mGluRA and RPLP0 were observed to decrease in the co-IP, although by marginal amount. The nuclei fractionation step did not notably improve the yield and quality of the enrichment by co-IP, meaning that the successful characterization of edited transcripts on a transcriptome-wide scale required a fundamentally different approach.

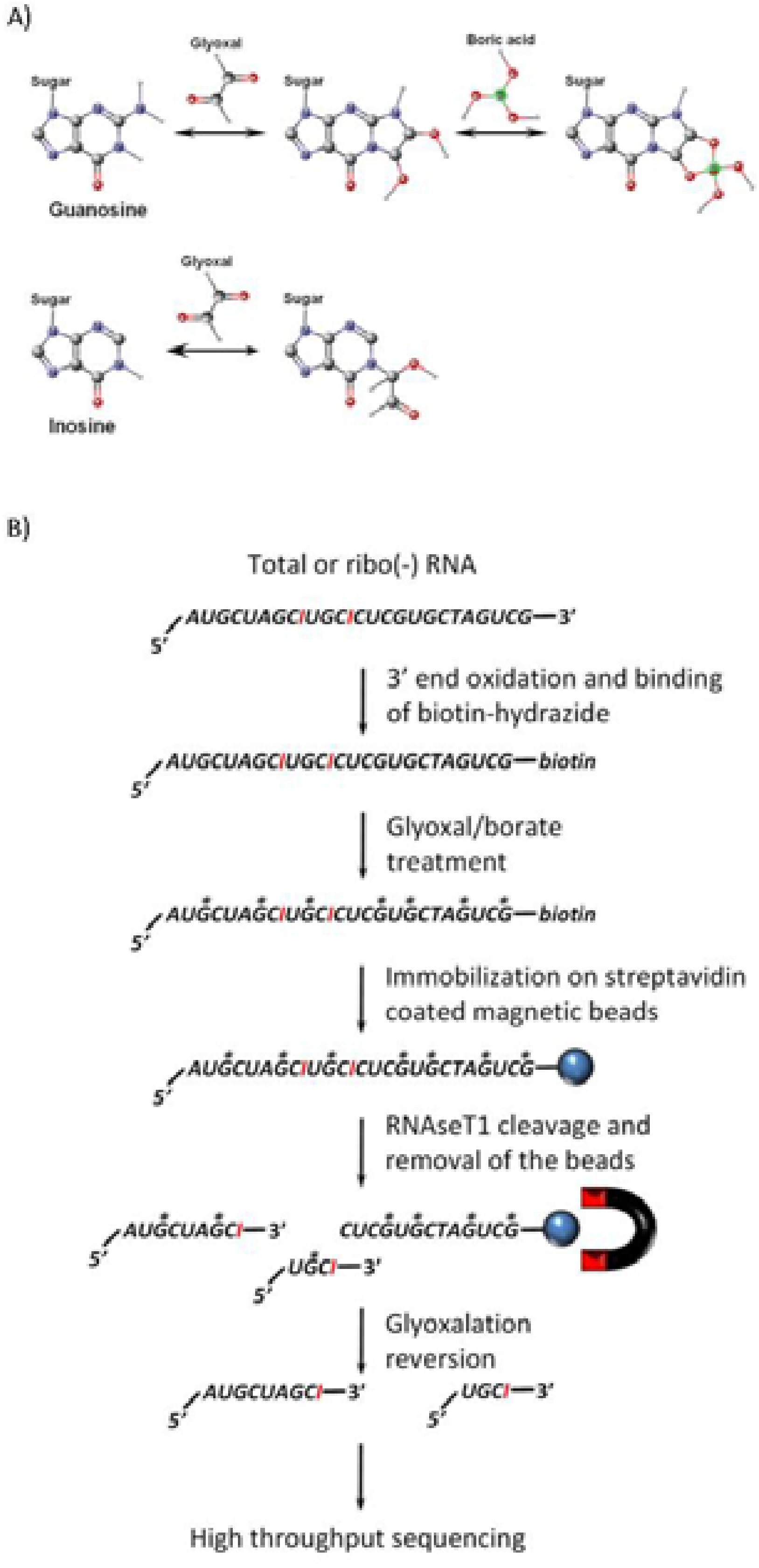
Figure 26: A) Dotplot screening for ADAR2 and GAPDH in HeLa cell nuclear and cytoplasmic fractions. B) Dotplot screening for ADAR2 and GAPDH in the different fractions produced during the nucleus purification from mouse brain. C) RT-PCR targeting 5HT2C, ADAR2, Gria2 Q/R and R/G, mGluRA and RPLP0 transcripts in ADAR2 IP (IP+) and negative control made with nonspecific IgG (IP-).



Development of a new method to purify edited RNA using Glyoxal/RNAseT1 cleavage

Due to limited success with the immunoprecipitation another approach was required. Critically, instead of targeting ADAR proteins to isolate the A-to-I edited RNA, the inosine in the modified RNA was directly targeted. The protocol developed aimed to extract RNA containing inosine, followed by sequencing of the purified RNA to identify editing sites and quantify the level of editing. First, the RNA was biotinylated and bound to magnetic beads, followed by treatment with glyoxal and borate to protect guanosine against RNAseT1 [221] (Figure 27A), and then by treatment with RNAseT1. The RNAseT1 cleaved the RNA off from the beads at the 3'end of inosine nucleotides. The RNAs freed from the beads were collected (I-RNA), which represented the 5' side of the editing site and contained the inosine at their 3'ends. The RNAs still bound to the beads were eluted (B-RNA), and this sample contained the RNAs that remained intact because they do not contain inosine and the 3' side of the RNAs that were cleaved. The edited sites were then identified by deep-sequencing the I-RNA and B-RNA (Figure 27B).

Figure 27: A) Guanosine forms a stable complex with glyoxal and borate that protects guanosine against RNAseT1 cleavage. Inosine fail to bind glyoxal in a stable fashion and remains cleavable by RNAseT1 [221, 256]. B) The strategy for separating RNA containing inosine from normal RNA before identification.

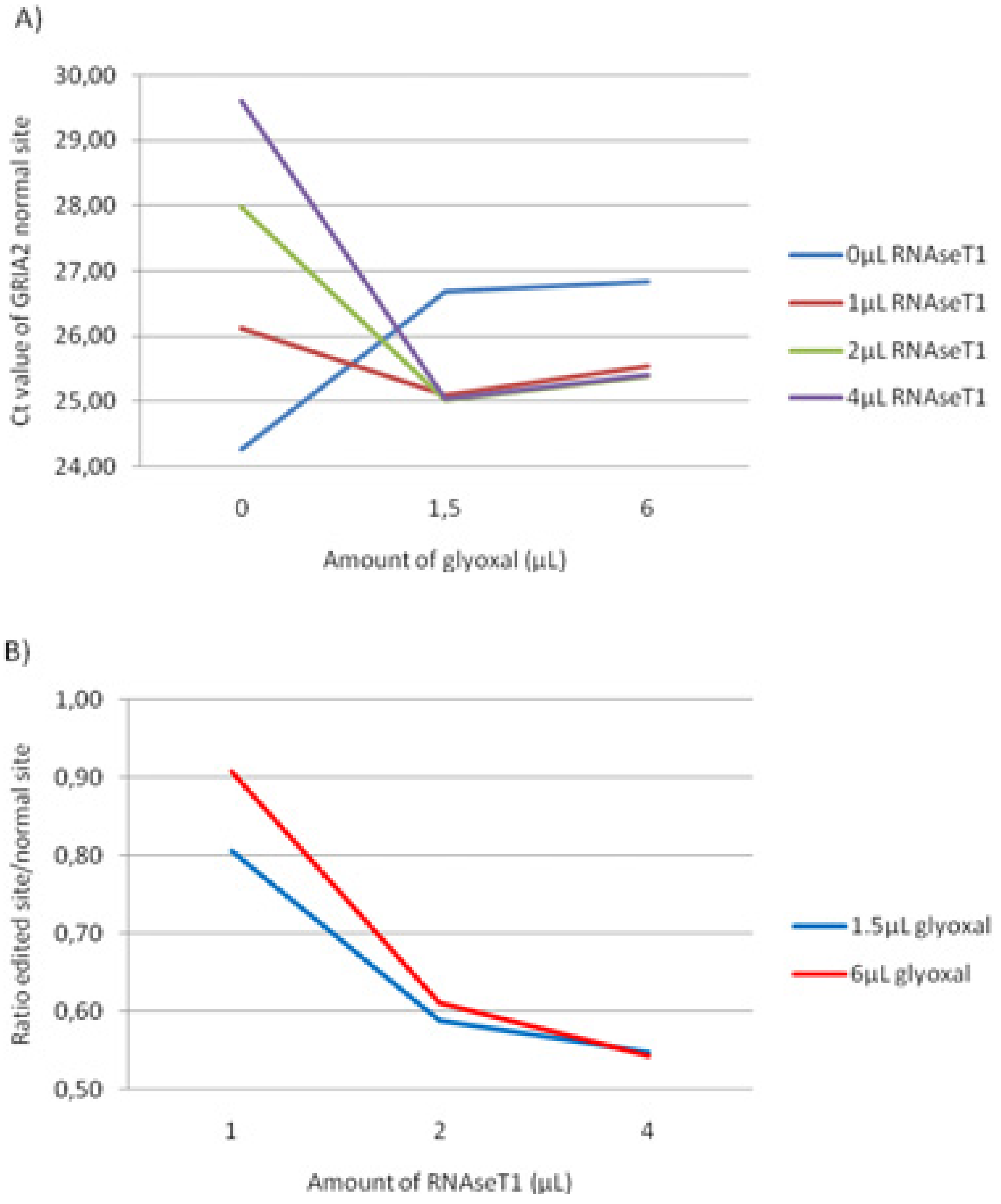

A)
Glyoxal
Boric acid
Sugar
Sugar
Sugar
Guanosine
Glyoxal
Sugar
Sugar
Inosine
B)
Total or ribo(-) RNA
5′ AUGCUAGCIUGCICUCGUGCTAGUCG—3′
3′ end oxidation and binding of biotin-hydrazide
5′ AUGCUAGCIUGCICUCGUGCTAGUCG—biotin
Glyoxal/borate treatment
5′ AUGCUAGCIUGCICUCGUGCTAGUCG—biotin
Immobilization on streptavidin coated magnetic beads
5′ AUGCUAGCIUGCICUCGUGCTAGUCG
RNAseT1 cleavage and removal of the beads
5′ AUGCUAGCI—3′
CUCGUGCTAGUCG
5′ UGCI—3′
Glyoxalation reversion
5′ AUGCUAGCI—3′
5′ UGCI—3′
High throughput sequencing

Protocol optimization

Specificity of the Glyoxal/RNAseT1 cleavage

To assess the efficiency and specificity of the cleavage using the glyoxal/RNAseT1 approach, primers were designed to amplify GRIA2 at both the Q/R site which is edited at 100% in brain [141, 164], and upstream of this site as well. qPCR was then used to query the level of the two targets to optimize the glyoxal/RNAseT1 treatment. The optimal amount of RNAseT1 and glyoxal necessary for a specific cleavage of edited sites was determined empirically by performing a series of RNA cleavages with varying amounts of both glyoxal and RNAseT1. The results from this indicated that treating RNA with glyoxal altered the stability of the RNA: a reduction of the level of GRIA2 non-edited site was observed when treating RNA with 1.5µL or 6µL of glyoxal without RNAseT1 treatment (Figure 28A, 0µL RNAseT1); however, the glyoxal efficiently protected the RNA against RNAseT1 treatment. Without glyoxal protection, the level of GRIA2 decreased 4 times, 13 times and 41 times (calculated with the following formula: $2^{\wedge(\text{Ct X}\mu\text{L RNAseT1 - 0}\mu\text{L RNAseT1})}$) when treated by 1µL, 2µL and 4µL of RNAseT1 respectively, whereas the level of GRIA2 remained stable at a Ct value of ~25.1 and 25.4 upon protection by 1.5µL and 6µL of glyoxal respectively for any amount of RNAseT1 (Figure 28A). Comparison of the level of the GRIA2 edited site with the level of normal site across the series of RNA cleavage experiments revealed that the cleavage was indeed specific (Figure 28B): a reduction of the level of the edited site was observed when treating the RNA with glyoxal and RNAseT1, the optimal conditions being 4µL of RNAseT1 and 1.5µL or 6µL of glyoxal, which cleaved up to 45% of the edited site [calculated with the following formula: 1- (edited site / normal site)].

Figure 28: A) Variation of the level (in Ct value) of GRIA2 normal site upon glyoxal and RNAseT1 treatments of mouse brain RNA. B) Variation of the level of GRIA2 site Q/R relative to the variation of the level of GRIA2 upstream Q/R as a function of the amount of RNAseT1 and glyoxal used to treat the mouse brain RNA.



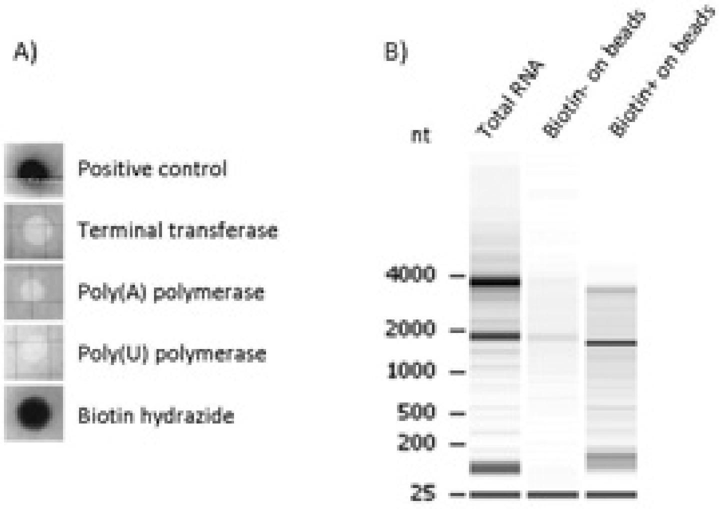
Binding of total RNA to magnetic beads

Before cleavage with RNAseT1, the RNA needed to be attached to magnetic beads to separate the edited RNA from the unedited pool. A common way to reversibly bind two molecular entities together is with streptavidin and biotin, where one entity is bound to biotin and the second to streptavidin. When mixed together the biotin quickly forms a strong non-covalent bond with streptavidin (Kd ~ $4*10^{-14}$M), which can be reversed using deionized water at temperature greater than 60˚C [257].

Although streptavidin coated magnetic beads are available commercially, the RNA still needs to be conjugated to biotin. Since the glyoxal/RNAseT1 treatment cleaves RNA at the 3'end of the inosine [256], the biotin must be incorporated at the 3'end of the RNA so that the inosine on the RNA strand which is freed from the bead can be isolated. Fortunately, several methods are available to bind biotin to RNA, such as enzymatic treatment with terminal transferase [258], poly(A)polymerase [259], and poly(U)polymerase [260], all of which can add biotinylated nucleotides to the 3'end of RNA [261, 262]. Chemically, the 3'end of RNA can also be oxidized, transforming the 3'end cis-diol into two aldehydes able to form a covalent bond with biotin-hydrazide [263]. To determine which treatment was the most efficient way to incorporate biotin into RNA each method was tried on total mouse brain RNA and the level of incorporation was estimated by dot-plot with streptavidin-HRP to detect the biotinylated RNA. The chemical binding of biotin to RNA was by far the most efficient being the only one detected by streptavidin-HRP (Figure 29A). This method also generally preserved the integrity of the RNA, as demonstrated by comparing the RIN scores [264] of the biotinylated RNA to total RNA (7.80 verses 8.10 respectively; data not shown).

Next, the impact of the linkage between the biotinylated RNA and beads was tested by analyzing the profile of the RNA before and after magnetic beads binding, a negative control was done with streptavidin coated magnetic beads and non-biotinylated RNA (Figure 29B). The bioanalyzer profile of the biotinylated RNA eluted from magnetic beads revealed an intensity reduction for the RNAs over 2000nt compared to the total biotinylated RNA, which indicated that the binding of RNA to the beads was slightly biased toward short RNA. In addition, the RNA bands remained sharp on the profile, indicating that the binding of RNA on the beads did not impact the integrity of the RNA.

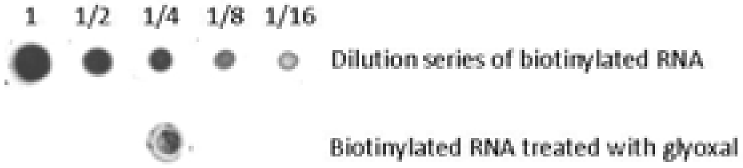
Figure 29: A) Dot-plot of mouse brain RNA biotinylated with four different methods and detected with streptavidin-HRP. B) Bioanalyzer profile of total RNA, total RNA incubated with streptavidin coated magnetic beads then eluted, and biotinylated total RNA incubated with streptavidin coated magnetic beads and then eluted.



Interaction between glyoxal and the biotin-streptavidin bond

To establish if the RNA needed to be attached to the beads before or after the glyoxal treatment, the impact of the glyoxal treatment on the biotin-streptavidin bond was observed by treating biotinylated RNA with glyoxal and detecting it with streptavidin-HRP. The dot-plot indicated that glyoxal reduced binding of streptavidin to biotin by a factor of 4, since the intensity of the 2.5µg glyoxal treated RNA showed an intensity equivalent to ~25% of the intensity of non-treated RNA (Figure 30). This result was reproduced also in the binding efficiency of the magnetic beads: 1mg of beads typically binds 2.6µg of biotinylated RNA, but this decreased to 1.38µg when the biotinylated RNA was treated with glyoxal beforehand. In addition, the binding efficiency measured for untreated RNA was 4 times lower than the efficiency announced by the manufacturer (10µg of nucleic acid/mg of beads), although this could be explained by the fact that the binding capacity was inversely related to the molecule size, and that total RNA is predominantly composed of large RNA species (manufacturer instructions for Invitrogen #653-05).

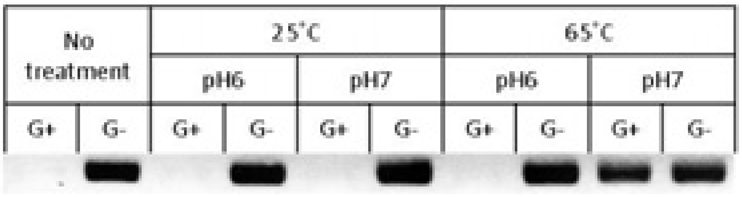
Figure 30: Dot-plot of biotinylated RNA detected by streptavidin-HRP. Series of untreated biotinylated RNA (from 1 corresponding to 2.5µg of RNA to 1/16 corresponding to 156ng of RNA) and 2.5µg of biotinylated RNA treated with glyoxal were spotted on the blot. The intensity of the biotinylated RNA treated with glyoxal was compared to the dilution series to estimate the impact of the glyoxal treatment on the biotin-streptavidin interaction.


1
1/2
1/4
1/8
1/16
Dilution series of biotinylated RNA
Biotinylated RNA treated with glyoxal

Removal of the glyoxal

One key requirement of the glyoxal/RNAseT1 protocol was that it was necessary to remove glyoxal after isolating the RNA to reverse transcribe it for sequencing [221]. To find the optimal pH and temperature for glyoxal removal, glyoxalated RNA was cleaned at pH6 or pH7 at either 25˚C or 65˚C for 3hrs. If glyoxal was not removed from the RNA with the cleaning step, it was expected that no signal would be detected in the RT-PCR since glyoxal blocks reverse-transcriptase. Based on this experiment it appeared that the best condition to remove glyoxal from RNA was 65˚C at pH7 (Figure 31).

Figure 31: RT-PCR targeting β-actin of mouse brain RNA which was successively treated (G+) or not treated (G-) with glyoxal and cleaned at 25˚C or 65˚C at pH6 or pH7, or not cleaned (No treatment) as indicated. The detection of a band in the glyoxal treated sample means that the conditions allow a successful removal of the glyoxal.



Evaluation of the extraction protocol

The experimental results established above were synthesized into a complete protocol, taking into account the optimal conditions determined previously and the potential biases introduced by each step (refer to the Supplemental Materials for a more detailed protocol description). The performances in terms of amount of RNA recovered and specificity of the RNA are presented below.

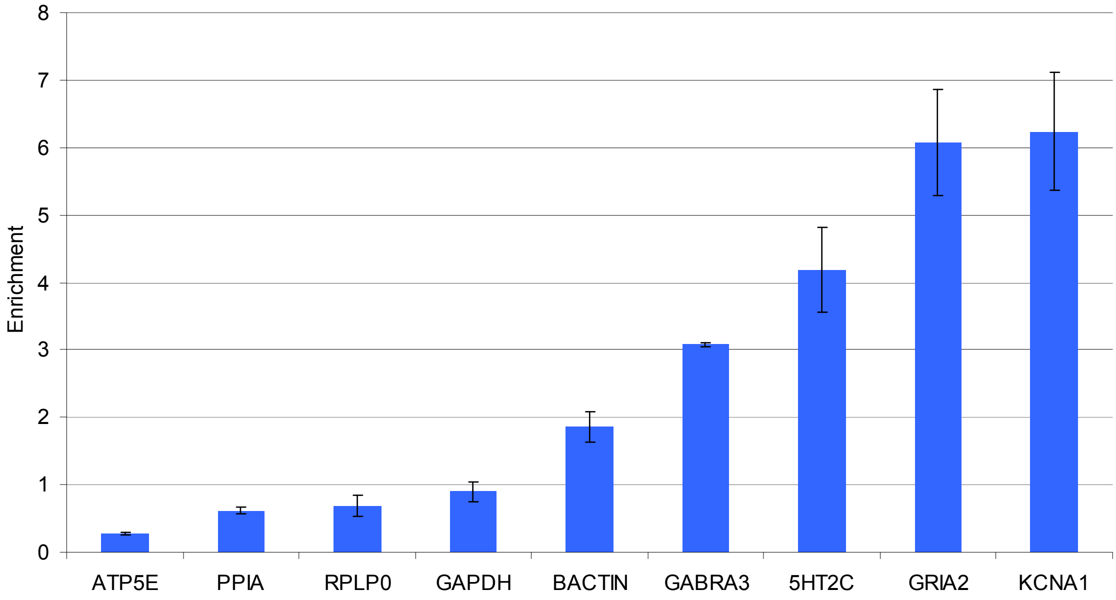
Extraction of edited RNA

Starting with 30µg of mouse brain RNA, the glyoxal/RNAseT1 protocol produced 400ng to 600ng of B-RNA and 50ng to 80ng of I-RNA. However, current deep-sequencing technologies require at least 100ng of RNA for library preparation, meaning that the protocol needs to be scaled to reach that amount. It is important to note that the amount of I-RNA collected is directly dependent on the prevalence of editing in the tissue; for example, the same experiment done on mouse liver RNA, which contains less editing sites [265], provided ~8ng of I-RNA for the same amount of B-RNA.

Specific enrichment of known targets of the editing enzymes

To assess the efficiency of inosilated RNA extraction using this protocol, the amounts of known edited transcripts in the I-RNA fraction were compared to their amounts in B-RNA. GRIA2, HTR2C, the potassium voltage-gated channel KCNA1 and the γ-amino-butyric acid receptor GABRA3 are extensively reported as targets of the ADAR proteins and are all expressed in the mouse brain. GRIA2 is edited at 100% at the Q/R site in brain [141, 164]. 5HT2C contains at least four different editing sites, each one being edited at various levels from 40% to 80% [144]. KCNA1 is edited at the site I/V at ~47% by ADAR2 [168, 171]; and GABRA3 is edited at the site I/M by both ADAR1 and ADAR2 at 50% in new born mice, and up to ~100% in adult mice [169, 170]. Their levels of enrichment were compared to the enrichment of β-actin, GAPDH, PPIA and ATP5E for which no editing was reported in the DARNED database [266] and RPLP0 which was specifically used as a negative control in previous experiments on A-to-I editing [217, 218]. GABRA3, 5HT2C, GRIA2 and KCNA1 showed from 3 fold to 6 fold enrichment whereas the levels of ATP5E, PPIA, RPLP0 and GAPDH decreased in I-RNA (ratio I-RNA/B-RNA <1) indicating that the protocol allowed an efficient enrichment of edited RNA (Figure 32). Out of the five negative controls, β-actin was the only one being enriched in the I-RNA with almost 2-fold enrichment. This may indicate that β-actin is edited at a low level in mouse brain.

Figure 32: Enrichment of GRIA2, 5HT2C, KCNA1, GABRA3, RPLP0, β-actin, GAPDH, PPIA and ATP5E in I-RNA compared to B-RNA in mouse brain by qPCR.



Deep sequencing of the I-RNA and B-RNA

Illumina deep sequencing was performed on 200ng of I-RNA and 500ng of B-RNA to produce 65nt paired-end sequences. The I-RNA was sequenced without fragmentation or size selection to obtain the 3'end of each edited transcript, which contained the inosine; the B-RNA library was prepared with the fragmentation step to remove the 3'end biotin. The I-RNA sequencing produced 34,623,034 pairs of sequences and the B-RNA 83,049,769 pairs. The discrepancy between the numbers of sequences of the two libraries could be explained by the difference in the preparation. The B-RNA library preparation included fragmentation and size selection, which meant that the library was homogeneous and the sequencing optimal. This is in contrast to the I-RNA library which did not include either fragmentation or size selection, producing a heterogeneous library that contained both long and short sequences which could reduce the sequencing efficiency.

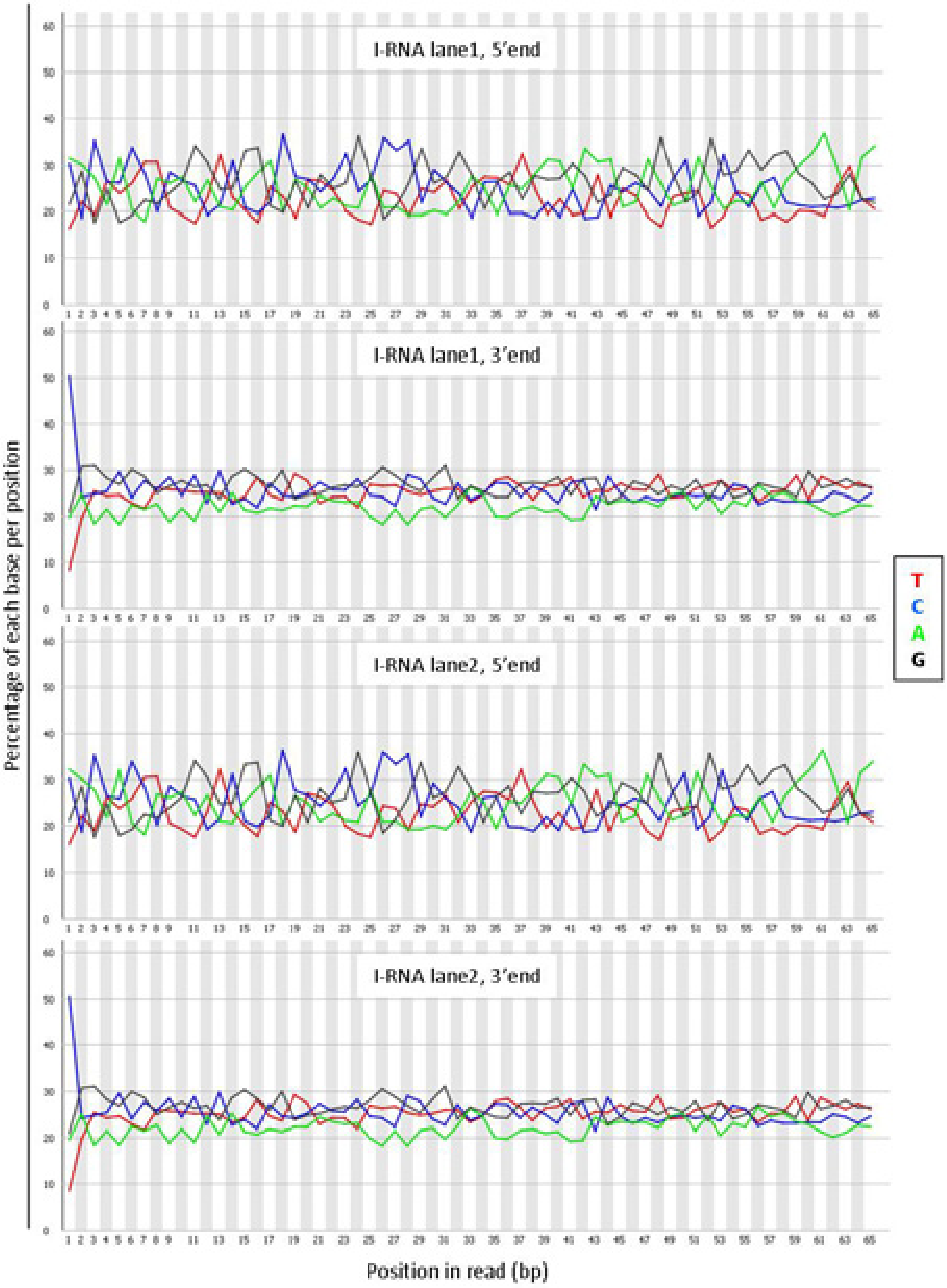
Raw sequence analysis

The deep-sequencing analysis returned 8 sets of sequences corresponding to the two lanes for each end of each library (Table 5). The nucleotide content of the reads from the I-RNA library showed a strong bias toward cytosine at the first base of the 3'end, a bias that was not observed in the 5'end (Figure 33). The 3'end reads corresponded to the reverse sequence of the cleaved RNA 3'end, which meant that the bias toward cytosine in the sequence corresponded to a bias towards guanine in the library. This bias confirmed that the RNAseT1 cleavage was specific to guanine and inosine.

Table 5: Files details generated by the deep sequencing of I-RNA and B-RNA libraries

Library	End sequenced	Sense of the reads	File name	Number of tags
I-RNA lane1	5’end	Sense	mmu_Brain_I_RNA_a_read1_sequence.txt	22,024,909
I-RNA lane1	3’end	Antisense	mmu_Brain_I_RNA_a_read2_sequence.txt	22,024,909
I-RNA lane2	5’end	Sense	mmu_Brain_I_RNA_b_read1_sequence.txt	12,598,125
I-RNA lane2	3’end	Antisense	mmu_Brain_I_RNA_b_read2_sequence.txt	12,598,125
B-RNA lane1	5’end	Sense	mmu_Brain_B_RNA_a_read1_sequence.txt	45,201,504
B-RNA lane1	3’end	Antisense	mmu_Brain_B_RNA_a_read2_sequence.txt	45,201,504
B-RNA lane2	5’end	Sense	mmu_Brain_B_RNA_b_read1_sequence.txt	37,848,265
B-RNA lane2	3’end	Antisense	mmu_Brain_B_RNA_b_read2_sequence.txt	37,848,265

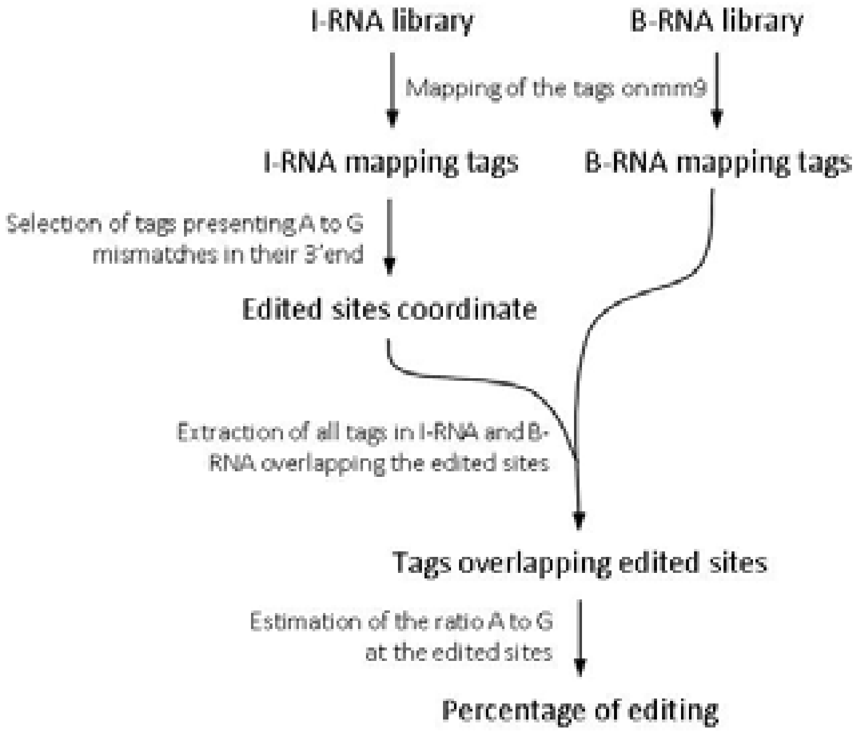
Figure 33: Proportion of each base, thymine in red, cytosine in blue, adenine in green and guanine in black in function of the position in the reads of the I-RNA libraries (from top to bottom: lane1 5'end, lane1 3'end, lane2 5'end and lane2 3'end).



Identification of editing sites

To identify the exact position of the edited sites and the edited transcripts the libraries were mapped to the mouse genome, and the coordinates of the edited sites were extracted from the I-RNA tags by identifying those reads that mapped to the genome with an A to G mismatch at their 3' extremities. All tags from B-RNA and I-RNA libraries overlapping these edited sites were selected and aligned to each other, and the number of A and G present at the edited site were tallied to calculate the percentage of editing at each edited position (Figure 34).

Figure 34: Outline of the algorithm used to identify A-to-I edited sites


I-RNA library
B-RNA library
Mapping of the tags on mm9
I-RNA mapping tags
B-RNA mapping tags
Selection of tags presenting A to G mismatches in their 3'end
Edited sites coordinate
Extraction of all tags in I-RNA and B-RNA overlapping the edited sites
Tags overlapping edited sites
Estimation of the ratio A to G at the edited sites
Percentage of editing

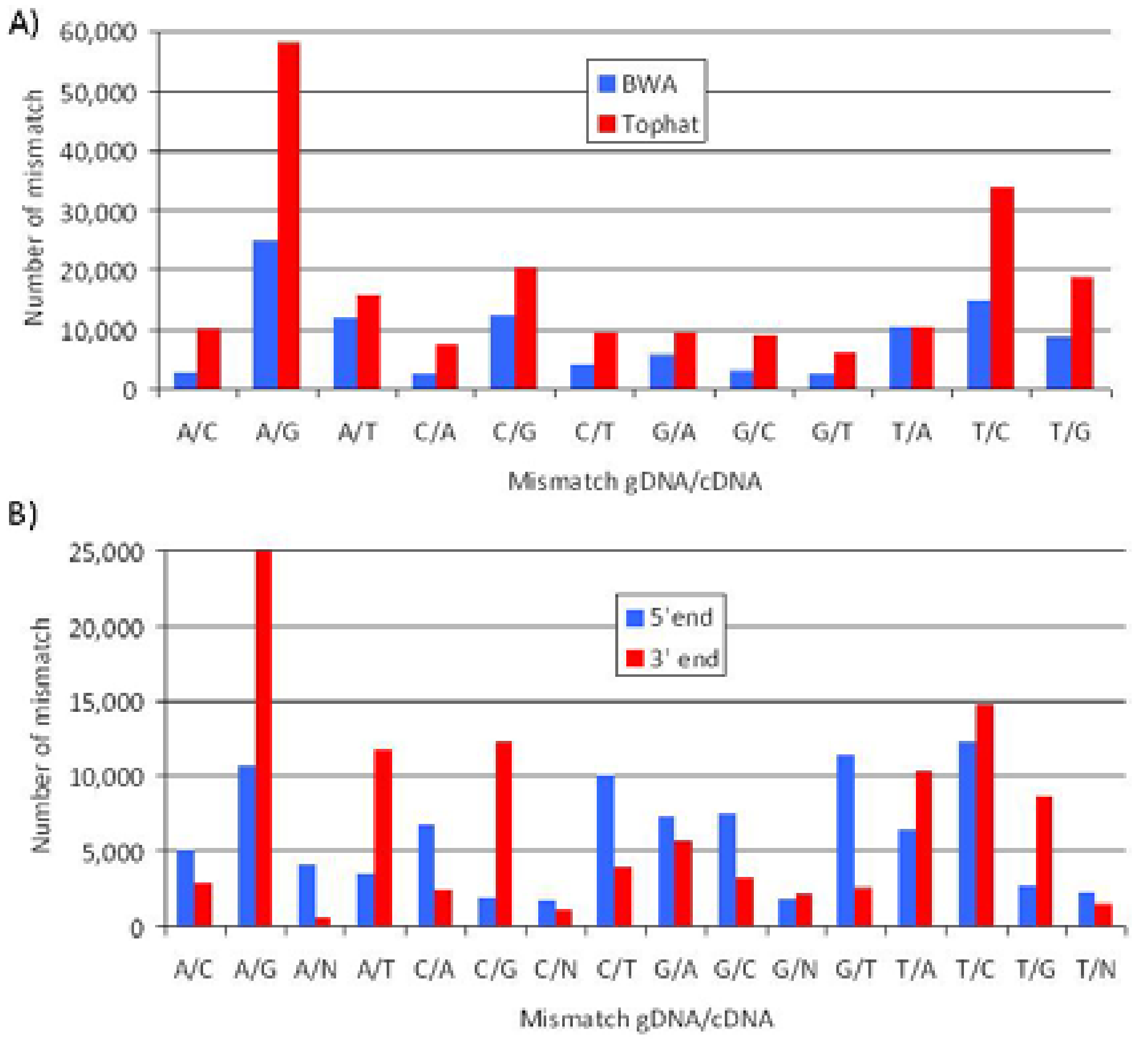
The mapping strategy

Several algorithms are available to map deep-sequencing data (reviewed in [267]). To maximize the detection of editing sites two algorithms were compared: BWA [268] and Tophat [249]. The results given by the two mapping strategy are given in Table 6. It is important to note that Tophat map tags more thoroughly than BWA since Tophat looks for splice junctions. Tophat detected 2.3 times more tags with an A-to-G mismatch and 1.4 times more editing sites than BWA (Figure 35A). The analysis of the mismatches present at the 5' and 3'ends of the mapping tags of the I-RNA library showed enrichment for A-to-G mismatches over all other mismatches in 3'end, a result which confirmed that our method enriched for RNAs containing inosine (Figure 35B).

Table 6: Result of the mapping of the I-RNA and B-RNA libraries with BWA and Tophat.

	BWA	Tophat
I-RNA mapping tags (from 34,623,034 pairs of tags)	13,661,778 (39.4%)	15,093,027 (43.6%)
B-RNA mapping tags (from 83,049,769 pairs of tags)	76,985,712 (92.7%)	75,213,148 (90.6%)
Number of 3'end extremities from I-RNA (number of loci)	1,526,719	1,398,042
Number of A to G mismatches at the 3'end extremities of I-RNA (number of tags)	24,965	58,206
Number of loci identified by a uniquely mapping tag containing an A-to-G mismatch in 3'end	9,231	12,644

Figure 35: A) Distribution of the mismatches at the 3'end of the tags from the I-RNA library mapped with BWA (in blue) and Tophat (in red). B) Distribution of the mismatches at the 5'end (in blue) and at the 3'end (in red) of the tags from the I-RNA library mapped with BWA.



Filtering of the edited loci

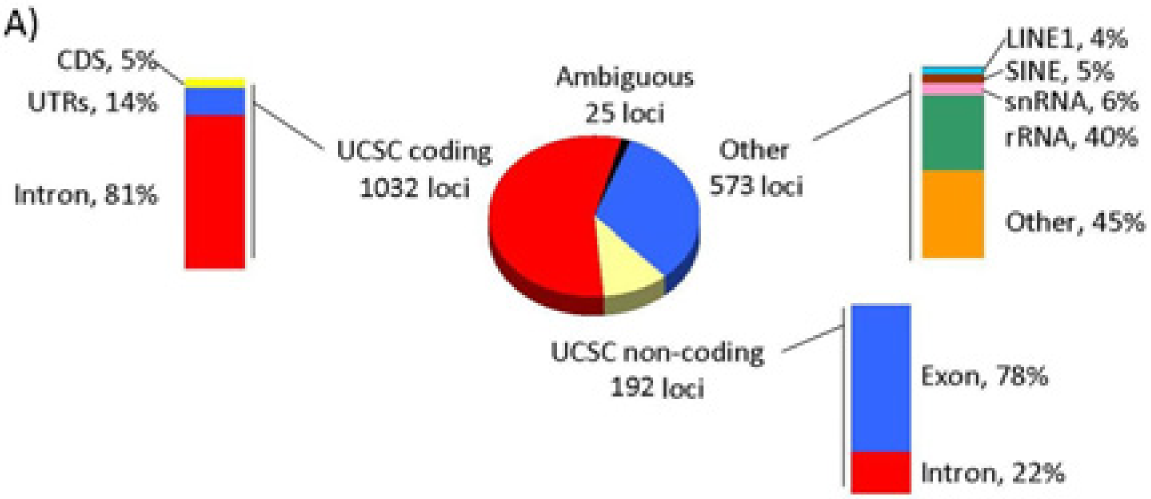
The following analyses were carried out using the tags mapped using Tophat. The Tophat alignment algorithm provided a file in bam format [250] from which it was possible to extract the exact location of the mismatch, and to measure the number of tags covering this location and the number of each kind of mismatches. This information was collected for each one of the 12,644 loci from the I-RNA library presenting an A-to-G mismatch in their 3' extremity, and assembled in a table (table available in supplemental material, description of the table in Appendix 6). The following filtering was then applied to the data: all the loci that only contained one A-to-G mismatch or were covered by less than five tags were discarded. On the 2,891 remaining loci, all loci presenting less A-to-G mismatches than A-to-T and A-to-C together were also discarded which returned 1,971 loci. Amongst those, 1,069 possessed the condition required (detailed in material and method) to perform a z-test to determine if the number of A-to-G mismatches is significantly higher than the sum of the number of the two other mismatches. 983 loci returned a p-value < 0.0001, 20 loci < 0.001 and 20 loci < 0.05, and 46 loci had a non-significant enrichment in A-to-G mismatch and were discarded. Of the 843 loci that did not have the conditions required for statistical analysis, all loci that presented at least five times more A-to-G mismatches than any other were kept, thus rescuing 799 loci. All together, this analysis returned a total of 1,822 edited loci (data available in supplemental material, description of the table in Appendix 6).

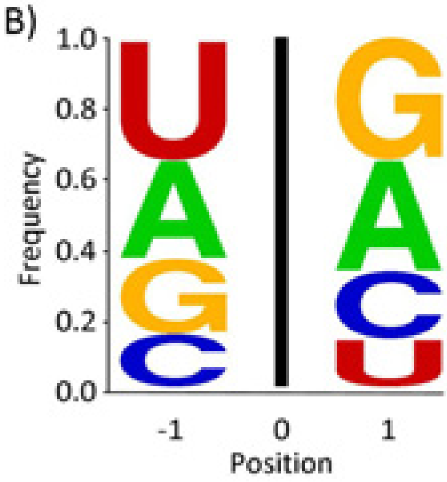
Characterization of the editing sites

To characterize the 1,822 edited loci, their locations were intersected with the locations of the SINEs, the LINE1s, the rRNAs, the snoRNAs, the snRNAs, the miRNAs and the UCSC genes with a distinction between coding genes and non-coding genes (Figure 36A). 56% of the loci covered coding genes and 10% non-coding genes. In the coding genes, most of the edited sites were intronic (81%) and only 5% were present in coding exons. In the non-coding genes, 78% of the sites were present in the exons and only 22% in the introns. This suggested that the primary role of editing in coding genes is to modify the splicing pattern of the transcript whereas in non-coding gene it is to modify the sequence of the mature RNA which may have an impact on its structure and target recognition. In the 895 loci seating in the introns of the UCSC gene database, 103 were localized in LINE1, 27 in SINE, 63 in rRNA, 33 in snoRNA and 26 in snRNA. Finally, 573 loci came from transcripts that were not referenced by the UCSC gene database. 40% of them came from rRNA, 6% from snRNA, 5% from SINE and 4% from LINE1 (Figure 36A).

The frequency of nucleotides surrounding the 1,822 edited sites was then estimated; it appeared that editing shows a preference for sites with the following neighbors: U>A>G>C in 5’ and G>A>C>U in 3’ (Figure 36B) which is similar to previous findings [146, 147].

Figure 36: A) Identification of the transcripts containing the editing sites. The pie chart represent the number of editing sites located in UCSC coding genes in red, in UCSC non-coding genes in yellow, ambiguous loci in black (unclear if it is coding or not) and loci which are not covered by any genes (other) in blue. For each category, except for ambiguous, the location was characterized further in the bar graphs: for the coding genes the sites were distributed in three classes, intron in red, CDS in yellow and UTR in blue; for the non-coding genes in two category, exon in blue and intron in red; and for the category other, loci were distributed in LINE1 in light blue, SINE in brown, snRNA in pink, rRNA in green and unclassified loci in other in orange. B) Nucleotide density map in 5’ (position -1) and 3’ (position 1) of the editing site, the graph was constructed using WebLogo [269].





Editing sites in the CDS of protein coding genes

48 editing sites were found in the coding regions of 40 genes, 8 of them produced synonymous mutation, leaving 35 genes where the editing produced a change in amino acid (Table 7). Among these, several targets were already known, like Gria2 site R/G [152-154], Kcna1 [171], Gabra3 [169, 170], 3 of the 5 sites present on 5Ht2c [144, 149, 150], Grik2 [270], Cyfip2 [271], Blcap [271] and Cadps [222]. The site Gria2 Q/R described previously was not present in this dataset even though it is edited at 100% [141, 164]. Further analysis at this specific site revealed that it was excluded because of its low coverage.

In term of frequency of editing, Gabra3 was measured to be edited at 100% with coverage by 16 tags, which was concordant with previous reports for adult mouse brain [169, 170]. For 5Ht2c, the editing frequency decreased from site A to site D, 75% for A, 50% for B and 33% for D, which was mainly explained by experimental biases on editing sites that were nearby to each other (12nt between site A and site D): the cleavage in site A reduced the coverage in B and even more in D. Further examination of our datasets at the 5Ht2c sites revealed that the C site was also detected but was discarded because of low coverage. For Kcna1, we measured an editing frequency of 53%, which was close to the 47% previously reported [168, 171]. Overall, 28 new edited sites that changed the protein sequence of 26 genes were discovered.

*Table 7: List of the coding genes for which a non-synonymous mutation occurs when they are edited. Freq. : frequency of A-to-I editing; Gene symbol: the ones which are underlined have been previously described and the ones in red are reported as edited in the DARNED database for human [266]; gCodon: genomic encoded codon; eCodon: edited codon; Mutation: amino acid substitution created by the editing (aa from genome/aa from edited codon); * this codon possess two edited site which can produce two different aa.*

Location	Freq.	Strand	Gene symbol	gCodon	eCodon	Mutation
chrX:143604228-143604229	0.75	+	5Ht2c site A	ATA	GT[A/G]*	I/V
chrX:143604230-143604231	0.50	+	5Ht2c site B	ATA	[A/G]TG*	I/[M/V]
chrX:143604240-143604241	0.33	+	5Ht2c site D	ATT	GTT	I/V
chr4:42290071-42290072	0.17	+	AK048672	TAT	TGT	Y/C
chr11:30119575-30119576	0.81	-	AK137356	AAG	GAG	K/E
chr4:116310362-116310363	0.96	-	Akr1a4	AGG	GGG	R/G
chr3:107990695-107990696	0.09	+	Amigo1	AAG	GAG	K/E
chr18:34459717-34459718	1.00	+	Apc	AAG	GAG	K/E
chrX:8470505-8470506	0.40	+	B630019K06Rik	TAC	TGC	Y/C
chr2:157383884-157383885	0.37	-	Blcap	TAT	TGT	Y/C
chr2:157383875-157383876	0.32	-	Blcap	CAG	CGG	Q/R

Location	Freq.	STD	Gene symbol	gCodon	eCodon	Mutation
chr14:13244095-13244096	0.38	-	Cadps	GAG	GGG	E/G
chr2:71872179-71872180	1.00	+	cAMP-GeFII	ATC	GTC	I/V
chr9:110158808-110158809	0.95	+	Cspg5	ACT	GCT	T/A
chr11:46086144-46086145	0.85	-	Cyfip2	AAG	GAG	K/E
chr7:17470695-17470696	1.00	+	Dact3	AGG	GGG	R/G
chr4:68459718-68459719	0.57	-	Dbc1	ACA	GCA	T/A
chr15:54701754-54701755	0.95	-	Enpp2	AAG	GAG	K/E
chr14:46102178-46102179	0.06	-	Fermt2	AAG	GAG	K/E
chr10:13232535-13232536	1.00	+	Fuca2	AGA	GGA	R/G
chrX:69690630-69690631	1.00	-	Gabra3	ATA	ATG	I/M
chr3:80496207-80496208	0.77	-	Gria2	AGA	GGA	R/G
chr6:64225893-64225894	1.00	+	Grid2	AGC	GGC	S/G
chr10:48992581-48992582	0.92	-	Grik2	TAC	TGC	T/C
chr10:48992594-48992595	0.73	-	Grik2	ATT	GTT	I/V
chr17:27639739-27639740	0.20	-	Grm4	CAG	CGG	Q/R
chr3:96024192-96024193	0.06	+	Hist2h2ab	AAG	GAG	K/E
chr3:96024435-96024436	0.09	-	Hist2h2ac	AAC	AGC	N/S
chr9:40612739-40612740	0.01	+	Hspa8	AAG	GAG	K/E
chr6:126592175-126592176	0.53	-	Kcna1	ATT	GTT	I/V
chr15:74939114-74939115	0.95	-	Ly6c2	AAG	AGG	K/R
chr8:96703791-96703792	1.00	+	Mt1	GAG	GGG	E/G
chr14:32001840-32001841	0.07	-	Nisch	GAG	GGG	E/G
chr11:79316468-79316469	0.01	-	Omg	ACG	GCG	T/A
chr9:25109124-25109125	0.98	+	Sept7	TAC	TGC	Y/C
chr16:91656333-91656334	0.50	+	Son	AGG	GGG	R/G
chr16:91655859-91655860	0.25	+	Son	ACC	GCC	T/A
chr8:26128628-26128629	1.00	+	Tm2d2	AGG	GGG	R/G
chr9:64083672-64083673	1.00	+	Uchl4	AAA	GAA	K/E
chr18:24119047-24119048	0.25	+	Zfp397	AAA	AGA	K/R

Editing sites in the non-coding sequence of protein coding genes

Hyperediting of RNA was shown to lead to the sequestration of transcripts in paraspeckles which can later be released by cleavage under certain conditions (e.g. Cat2 [179]). Following this, an *in silico* analysis predicted 107 coding genes in which specific cleavage of hyperedited region may occur [272]. The editing of five of these genes were confirmed in our analysis as containing editing sites in 3'UTR and introns: Dlgap4 (NM_146128.5) contained 2 editing sites in introns, Abcb4 (NM_008830.2) and Snupn (NM_178374.3) one each in intron, Tox4 (NM_023434.3) two in 3'UTR overlapping a B1 element, and Tapbp (NM_009318.2) four in 3'UTR overlapping two B1 elements.

Editing sites in rRNA

343 editing sites were located in regions annotated as rRNA, among them, 30 presented an editing frequency over 0.5. Of those sites, 24 were located in 28S rRNA and 6 in the 18S rRNA. To precise the site of editing in the rRNA sequence, the tags defining editing sites were mapped to the mouse 45S rRNA precursor and then to human 28S and 18S rRNAs for which extensive information was available about the modifications occurring [273]. Intriguingly, it turned out that the six A-to-G mismatches found in the 18S rRNA regions corresponded to G in the 45S rRNA precursor and in the human 18S sequences; the same was observed for 21 editing sites found in the 28S rRNA regions and in the human 28S rRNA sequence, only three A-to-G mismatches corresponded to A in the mouse precursor and in the human 28S sequences (the sequences of the human 28S and 18S rRNA annotated with the editing sites discovered here, the methylation and pseudouridylation sites and the small RNAs initiating the modifications are described in Appendix 7). It is important to note that, at this stage, only uniquely mapping tags were used, and that they were paired-end tags, which mean that the loci identified were highly reliable. These observations suggested that most of the editing observed in rRNA corrected nucleotides which mutated to adenosine in the new rRNA locus compared to the 45S precursor rRNA. 17 of the 30 edited sites were located nearby modified sites (methylation or pseudouridylation; i.e. Appendix 7), modifications which are dependent on the base-pairing of a complementary snoRNA [274, 275]; thus, editing at these sites may alter the modification process when modifying a conserved nucleotide and disrupting the base-pairing, or restore the modification process when the editing change a mutated nucleotide back to the conserved nucleotide and restore the base-pairing.

Editing sites in snoRNA

The mouse snoRNAs are not well annotated and the coordinates of only 170 snoRNAs are available for mouse [276] whereas 381 are known for human [277]. To maximize the detection of editing sites in mouse snoRNAs, the editing sites were intersected with the mouse snoRNAs and also the human snoRNAs for which orthologous sequences were found in mouse. 29 A-to-I editing sites were located in 13 snoRNAs and, among these, 11 sites were on snoRNAs that were not annotated in mouse (Table 8).

The editing sites present on Snord45c and Snora22 were located in the regions involved in the target recognition. For Snora22, the nucleotides 87 and 88 located in the stem loop above the pseudouridylation site were edited (Figure 37A). The editing site in Snord45c was located in the region pairing with the 18S rRNA, on the A pairing with the U adjacent to the methylated nucleotide A159 (Figure 37B). The consequence of such editing may be critical for the functions of the snoRNA; however the level of editing at these sites remained really low with editing frequencies below 1% (Table 8). Coincidentally, the nucleotide located in position 4974 on the 28S rRNA in 5' of the uridine modified by Snora22, was shown previously to be edited in mouse in an rRNA gene containing an A at this position (Figure 37A). Similarly, the snoRNA Snord5 is edited and catalyze the methylation of the C2409 of 28S rRNA, a site located 15 nucleotides upstream of a conserved A-to-I editing site (i.e. Appendix 7). This can be due to blind chance, or it might also suggest some concomitance between the A-to-I editing and other modifications of rRNA. Co-localization of ADAR protein with rRNA and snoRNA in the nucleolus was previously demonstrated [119] but its signification remains largely unclear. The sequences of all the edited snoRNA with their edited loci are in Appendix 8.

Figure 37: A) Region from the human snoRNA Snora22 pairing with the 28S rRNA to modify the uridine U4975 to pseudouridine. B) Region from the human snoRNA Snord45c pairing with the 18S rRNA to modify the adenine A159 to methylated A159. The red arrows indicate the location of the editing site discovered in this thesis, the green arrow represents a position which was previously found edited in mouse 28S rRNA. The drawings are modified from the snoRNABase (http://www-snorna.biotoul.fr).

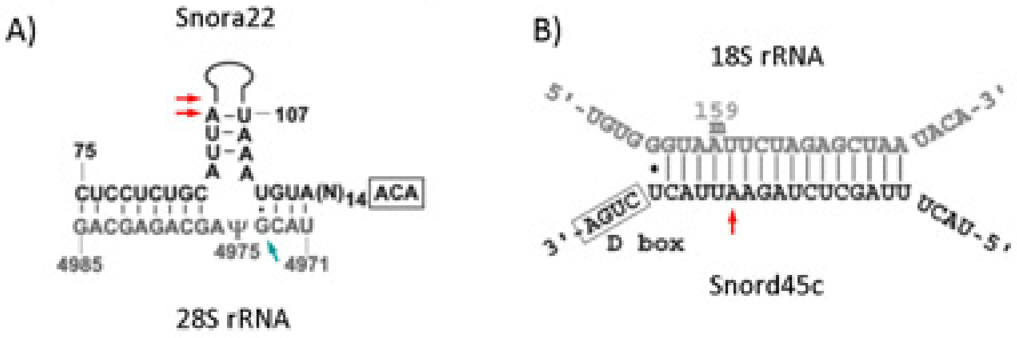
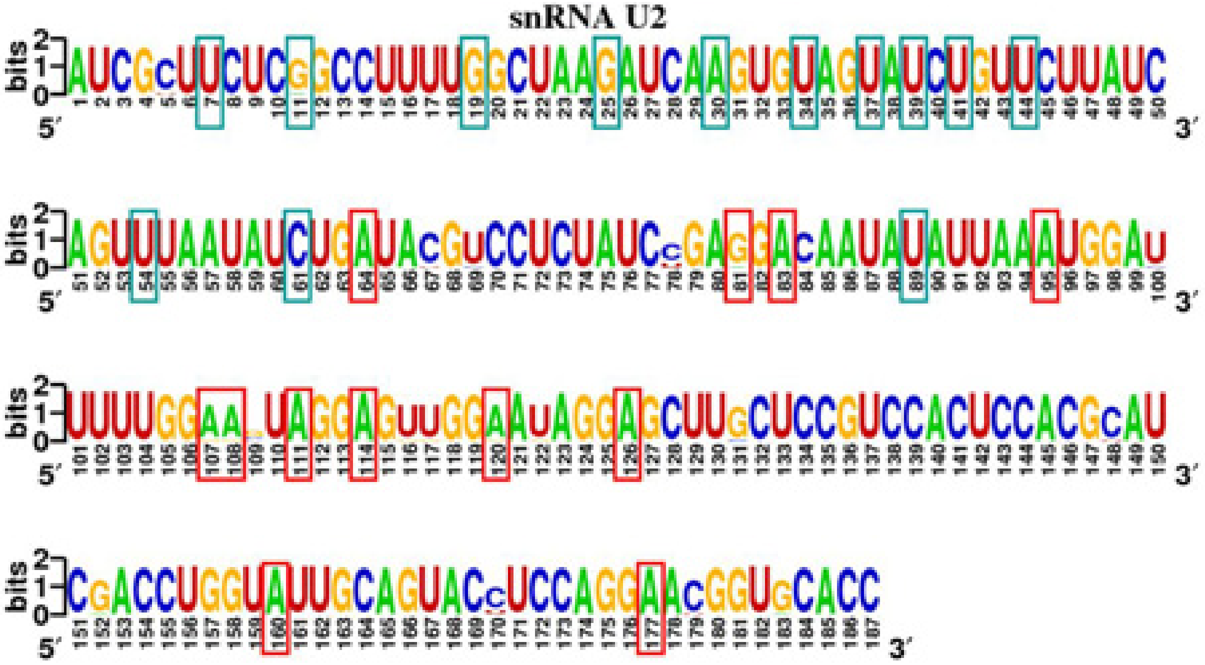


Table 8: List of the editing sites located in snoRNA. Freq. : frequency of editing; Predicted targets from the snoRNABase [277].

Location	Strand	Freq.	Mouse snoRNA	Human orthologue	Type	Predicted target
chr5:130295500-130295501	+	0.007	N/A	snora22	H/ACA	U4966 and U4975 of 28S rRNA
chr5:130295501-130295502	+	0.002	N/A	snora22	H/ACA	U4966 and U4975 of 28S rRNA
chr1:87983952-87983953	+	0.024	N/A	snora26	H/ACA	28S rRNA U4522
chr12:60235135-60235136	+	0.011	N/A	snora26	H/ACA	28S rRNA U4522
chr10:90581200-90581201	-	0.482	N/A	snora53	H/ACA	Unknown
chr11:69482390-69482391	-	0.004	N/A	snord10	C/D	U6 snRNA C77 and 28S rRNA C3787
chr7:67024349-67024350	-	0.009	N/A	snord116	C/D	Unknown
chr7:67029393-67029394	-	0.014	N/A	snord116	C/D	Unknown
chr9:15118582-15118583	+	0.007	N/A	snord5	C/D	28S rRNA C2409
chr7:51505898-51505899	+	0.005	N/A	snord88b	C/D	28S rRNA C3680
chr6:71832598-71832599	-	0.006	N/A	snord94	C/D	U6 snRNA C62
chr12:112779253-112779254	+	0.010	Snora28	snora28	H/ACA	U815 and U866 of 18S rRNA
chr12:112779265-112779266	+	0.040	Snora28	snora28	H/ACA	U815 and U866 of 18S rRNA
chr7:106631332-106631333	-	0.010	snord15a	snord15a	C/D	28S rRNA A3764
chr7:106631336-106631337	-	0.005	snord15a	snord15a	C/D	28S rRNA A3764
chr7:106631351-106631352	-	0.031	snord15a	snord15a	C/D	28S rRNA A3764
chr7:106631352-106631353	-	0.005	snord15a	snord15a	C/D	28S rRNA A3764
chr7:106631377-106631378	-	0.009	snord15a	snord15a	C/D	28S rRNA A3764
chr7:106628117-106628118	-	0.001	snord15b	snord15b	C/D	28S rRNA A3764
chr2:144091786-144091787	-	0.013	Snord17	snord17	C/D	28S rRNA U3797
chr2:144091787-144091788	-	0.081	Snord17	snord17	C/D	28S rRNA U3797
chr2:144091789-144091790	-	0.052	Snord17	snord17	C/D	28S rRNA U3797
chr2:144091790-144091791	-	0.006	Snord17	snord17	C/D	28S rRNA U3797
chr2:144091799-144091800	-	0.068	Snord17	snord17	C/D	28S rRNA U3797
chr2:144091858-144091859	-	0.108	Snord17	snord17	C/D	28S rRNA U3797
chr2:144091873-144091874	-	0.008	Snord17	snord17	C/D	28S rRNA U3797
chr19:8800422-8800423	+	0.003	Snord22	snord22	C/D	Unknown
chr3:153575103-153575104	-	0.006	snord45c	snord45c	C/D	18S rRNA A159

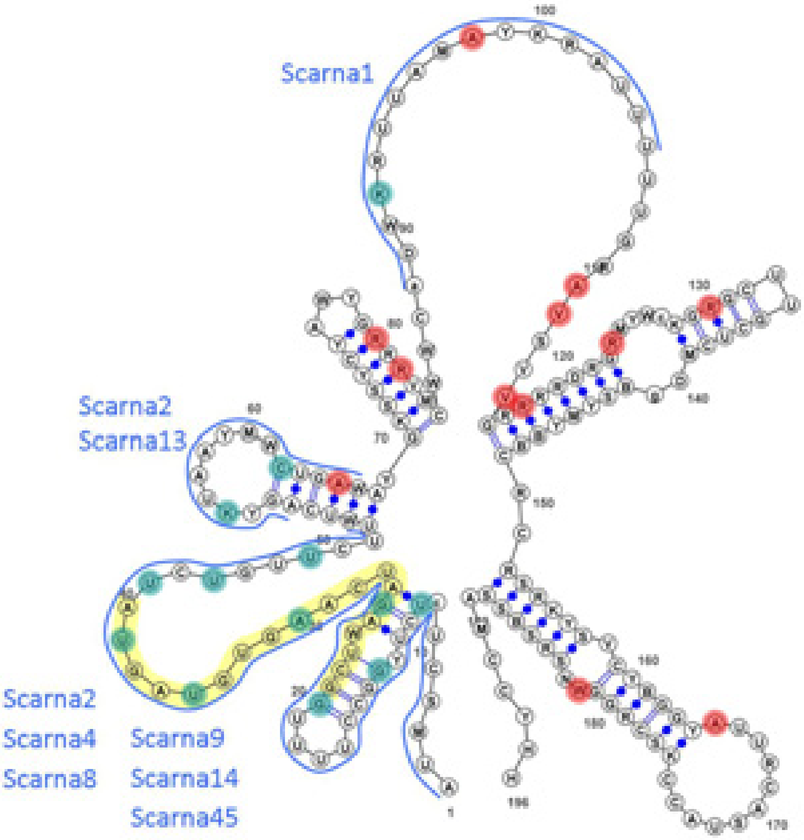
60 editing sites were located in 25 distinct snRNAs: eleven from the U2 family, five from the U3, two from the U4, four from the U5, two from the U14 and one from the U17; 59 of these sites were edited at less than 3% and the last site, on a U2, was edited at 99.6%. Alignment of the eleven U2 homologues that presented editing sites revealed that the highly edited site corresponded to a guanosine in most U2 sequences (position 81 on Figure 38). This may indicate that editing corrects a deleterious mutation at this precise site as observed previously in the rRNA. U2 snRNA contained 11 additional A-to-I editing sites and 13 sites that were modified by snoRNAs (Figure 38). The methylation site C61, modified by Scarna2 [277], is three nucleotides downstream the editing site A64 which is located in the site recognized by Scarna2 (Figure 39). Similarly, the pseudouridylation site U89, modified by Scarna1 [277], is six nucleotides downstream the editing sites A95 which is located in the site recognized by Scarna1 (Figure 38 and Figure 39). This means that one of the modifications might modulate the efficiency of the other one. The modifications occurring on the six snRNAs are presented in Appendix 9.

Figure 38: Sequence composition of the snRNA U2 built with the eleven sequences containing edited sites, aligned with ClustalW [246] and generated with WebLogo [269]. The nucleotides which are pseudouridylated or methylated by snoRNA (data from [277]) are boxed in green and the nucleotides which are edited in red.



The analysis of the secondary structure of snRNA U2 revealed that none of the editing sites were present in the region of U2 which interact with snRNA U6 or the mRNA during splicing [278, 279] which suggested that editing of U2 does not impact its splicing function (Figure 39). It also revealed that none of the editing sites were present in a stem-loop of at least 20nt long and seven of the twelve editing sites were located in loops (Figure 39), which indicated that editing in snRNA U2 was dependent of its binding to a partner to form the double stranded structure necessary for its editing [134].

Figure 39: Secondary structure of the family snRNA U2 from Rfam [280, 281] annotated with pseudouridylation and methylation sites (outlined in green; data from [277]), A-to-I editing sites (outlined in red), the nucleotides interacting with ScaRNAs are indicated with a blue line, the names of the ScaRNAs responsible of the modifications are in blue, and the nucleotides interacting with the mRNA and snRNA U6 during splicing are outlined in yellow [278, 279].



Materials and methods

Purification of edited RNA by co-immunoprecipitation

Preparation of the beads

For each IP, 100µL of Dynabeads M-280 sheep anti-rabbit IgG 10mg/mL (Invitrogen catalogue #112-03D) were prepared. First the beads were washed 3 times in block solution (1X PBS, 0.5% BSA) and incubated with 10µg of target specific antibody (IP+) or pre-immune serum (IP-) for at least 2hrs. The mix was then washed 3 times with the block solution and resuspended in 50µL of block solution. For the ADAR IP the following antibodies were used: ADAR1 (Santa Cruz biotechnology #sc-33179) and ADAR2 (Santa Cruz Biotechnology #sc-33180).

Brain lysate preparation

For standard IP

One brain from a BL6C57 male was homogenized in 3mL 1X HBSS and 0.01M HEPES (pH 7.3) with a glass grinder. The suspension was aliquoted in 2*1.5mL and centrifuged 10 min at 600g at 4˚C. The supernatants were discarded and the pellets were washed in 1mL ice cold 1X HBSS and resuspended in 0.5mL PXL (1X PBS, 0.1% SDS, 0.5% (v/v) sodium deoxycholate, 0.5% (v/v) NP-40, 1X protease inhibitor cocktail (Roche #11836170001), RNAseOut 100U/mL (Ambion #10777-019) and 0.5µL DTT 100mM). The two suspensions were sonicated by applying 2*5min 30s pulse 30s pause at high power in ice with a Bioruptor 200 (Diagenode #UCD-200 TO) and centrifuged at 10,000g for 20min at 4˚C, the supernatants were cleared for 30min at 4˚C on a rotating platform with 100µL Dynabeads M-280 coupled with sheep anti-rabbit IgG (Invitrogen #112-03D) washed 3 times with block solution as described above. The cleared lysates were collected after placing the tube on a magnetic stand for 2min.

Including the sucrose gradient purification step

The first sucrose gradient containing EDTA was prepared as follow: first the tube containing the sucrose gradient (Beckmann Coulter #361623) was treated RNAse free 1h with 2N NaOH solution and then washed thoroughly with sterile water. 2mL of RNAse free sucrose solution containing 30%, 22%, 14% and 5% sucrose in [5mM HEPES-NaOH pH 7.4, 0.2mM EDTA, 1X of protease inhibitor (Roche #11836170001), 50U/mL RNaseOUT (Invitrogen #10777-019) and 1mM DTT] were successively poured in the tube and frozen in dry ice. The sucrose gradient was then stabilized at 4˚C for 24hrs. 1mL of brain lysate prepared as described in the section "Brain lysate preparation for standard IP" was delicately poured over the sucrose gradient and the sucrose gradient was centrifuged for 15hrs at 22000rpm at 4C. The sucrose gradient was then fractionated in ~350µL fractions from bottom to top by piercing a hole at the bottom of the tube using a needle.

A dot-plot was done to assess which fractions contain ADAR2, H3 or GAPDH complexes as follow: a nitrocellulose membrane Hybond-P (Amersham #RPN2020F) large enough to contain 30 dots (10*3cm) was conditioned for 10min in methanol and then put in transfer buffer for at least 1min. The membrane was then put on a glass slide and 5µL of each fraction was deposited on the membrane. Once the membrane was dry, it was blocked in [10% skim milk, 1X TBS, 0.1% Tween 20] at RT for 2hrs, incubated with primary antibody (1/1,000) targeting ADAR2 (Santa Cruz Biotechnology #sc-33180), H3 (Cell Signaling Technology #9715) or GAPDH (Trevigen #2275-PC-1) in [2% skim milk, 1X TBS, 0.1% Tween 20] for at least 2hrs at RT or overnight at 4˚C, washed 3 times 5min in [1X TBS, 0.1% Tween], incubated 30min at RT in [10% skim milk, 1X TBS, 0.1% Tween 20], incubated with the secondary antibody [1/2,000 for anti-rabbit HRP-antibody (Invitrogen #62-6120) in [2% skim milk, 1X TBS, 0.1% Tween 20] for 1hr at RT, washed 3 times 10min in [1X TBS, 0.1% Tween], soaked for 1min in ECL solution (VWR #GEHRPN2132) and exposed to a photosensitive film (Fuji Medical X-RAY Film RX 18X24CM 100SHTS, Fujifilm #497690). The fraction containing ADAR2 complexes were then pooled to perform the IP as described in the pull-down part described below.

The second sucrose gradient containing NaCl and $MgCl_2$ was prepared as described for the sucrose gradient containing EDTA by replacing the sucrose solution by the following solution: 100mM NaCl, 10mM Tris pH 8.0, 2mM $MgCl_2$, 2mM vanadyl ribonucleoside, 50U/mL RNaseOUT (Invitrogen #10777-019) and 1mM DTT.

Including the nuclei fractionation step

To extract the nuclei from total brain, first 2 mouse brain were homogenized in 6mL of ice cold HBSS with a Dounce homogenizer (Wheaton Scientific #357542) using the loose pestle (0.114 ±0.025mm clearance). The cells from the homogenate were pelleted at 2500rpm for 3 min at 4°C, rinsed with 6mL of ice cold HBSS, resuspended in 5ml of Buffer A [10mM

Hepes pH 7.9, 10mM KCl, 1.5mM $MgCl_2$, 0.5mM DTT] and incubated on ice for 5 min. The cell suspension was then transferred to a pre-cooled 7 ml Dounce tissue homogenizer (Wheaton Scientific #357542) and homogenized with 30 strokes using the tight pestle (0.05 ±0.025mm clearance), while keeping the homogenizer on ice. The homogenized cells were centrifuged at 218g (1000 rpm, Beckman GS-6 centrifuge, GH-3.8 rotor) for 5 min at 4°C. The pellet contained enriched, but not highly pure, nuclei. The supernatant contained the cytoplasmic fraction. The nuclear pellet was resuspended in 3 ml of [0.25M Sucrose, 10mM $MgCl_2$], layered over 3 ml of [0.88 M Sucrose, 0.5 mM $MgCl_2$] and centrifuged at 3000g (3500 rpm, Beckman GS-6 centrifuge, GH-3.8 rotor) for 15 min at 4°C. The supernatant was discarded. The pellet contained pure nuclei, and was resuspended in 1.5mL of PXL and prepared as described in the protocol for standard IP.

Pull down

Each aliquot of cleared lysate was mixed either with the 50µL of beads incubated with the specific antibody (IP+) or with the 50µL of beads incubated with the pre-immune serum (IP-) and incubated overnight at 4°C on a rotating platform. The beads were then rinsed 3 times in 500µL of cold wash buffer [1× PBS, 2 mM $MgCl_2$, 15mM EDTA, 1% NP-40, 0.5% Tween-20, 1X protease Inhibitor Cocktail (Roche #11836170-001)], and once with 1mL of cold 1× PBS. The complexes antibody/specific targets were then eluted from the beads in 100µL [1× PBS, 1% SDS] at 65°C for 10min and the supernatants were collected for protein and RNA extraction.

RNA and protein purification

The RNA and proteins from the IP+ and IP- eluates were extracted using Trizol LS (Invitrogen #10296-028) according to the manufacturer instruction. Briefly, each sample volume was adjusted to 250µL with DEPC water, mixed with 750µL of Trizol LS, incubated at room temperature (RT) for 5 min, mixed with 200µL of chloroform, incubated 2-3min at RT and centrifuged at 12000g for 15min at 4°C. The suspension separated into two phases, the colorless aqueous phase contained the RNA and the pink organic phase and the interphase contained the proteins.

The RNA was precipitated by adding 500µL of isopropanol to the aqueous phase, incubating 10min at RT and centrifuging at 12000g for 10min at 4°C. The pellet was rinsed once with 70% ethanol, air-dried and resuspended in 10µL of DEPC-water.

The proteins were extracted from the organic phase by adding first 300µL of 100% ethanol to the organic phase, incubating 2-3min at RT, centrifuging at 2,000g for 5min at 4°C. The proteins were then precipitated by mixing the supernatant with 1.5mL isopropanol, incubating 10min at RT and centrifuging at 12,000g for 10min at 4°C. The pellet was washed 3 times in [95% ethanol, 0.3M guanidine hydrochloride], once in 100% ethanol, air-dried and resuspended in 100µL 1% SDS. After resuspension, the insoluble material was sedimented by centrifuging at 12,000g for 10min at 4°C. The proteins were ready for the Western-blot.

Reverse-transcription and PCR

The RNA from IP+ and IP- were treated with 2U of DNaseI (Invitrogen #18068-015) according to the manufacturer instructions and cleaned with MinElute cleanup-kit (Qiagen #74204) according to the following instruction (the standard protocol from Qiagen excludes the RNAs < 200nt, this protocol retains them): each DNaseI I treated sample was adjusted to 100µL with DEPC water and mixed with 350µL buffer RLT and 675µL 100%EtOH. The mix was passed through a column inserted in a collection tube by centrifuging 15s at 8,000g, the flow through was discarded. The column was then washed with 500µL buffer RPE and with 500µL of 80% EtOH. The column was dried by centrifuging at full speed for 1 min with the lid of the column open, and then placed in a new collection tube. The RNA was eluted with 14µL of DEPC-water placed directly on the column and centrifugation for 1min at full speed. The RNA was then reverse transcribed with the superscript III kit (Invitrogen #18080-051) according to manufacturer instruction using random hexamers.

As indicated for each IP, the abundance of GRIA2 site R/G, GRIA2 site Q/R, 5HT2C, mGluRA, RPLP0 or ADAR2 were queried by PCR. Each reaction was performed in 20µL containing 1µL of cDNA, 0.25µM of primers forward and reverse (sequences in Appendix 4), 1X PCR buffer (fisher biotec #TAQ-1), 1.25mM MgCl2, 1 unit of Taq polymerase (fisher biotec #TAQ-1) with the following program: 95˚C for 2min, [95˚C for 30s, 60˚C for 30s, 72˚C for 30s] for 35 cycles, 72˚C for 7min. PCR products were then run in a 1.5% agarose gel containing ethidium bromide.

SDS-PAGE and Western-blot

The two phases gel used in the SDS-PAGE and Western-blot was prepared the following way: first, the separating gel mix (10% acrylamide/bis 30:0.8, 0.4M Tris-HCl pH8.8, 0.1% SDS, 0.05% APS, 0.05% TEMED) was poured in a vertical gel cast. Then, the stacking gel mix (4% acrylamide/bis 30:0.8, 0.4M Tris-HCl pH6.8, 0.1% SDS, 0.05% APS, 0.05% TEMED) was poured on top of the polymerized separating gel and the comb was placed in the gel cast. Once polymerized, the gel was incubated overnight at 4˚C to complete the polymerization. 20µL of the purified proteins (IP+ and IP-) were denatured at 95˚C for 5 min in 1X Laemmli dye (80mM Tris-Cl pH 6.8, 2% SDS, 10% glycerol, 0.2g/L Bromophenol blue, 0.5% B-mercaptoethanol) and loaded on the gel. Each sample was loaded in 2 wells, one for the SDS-PAGE and one for the Western-blot. The gel was then run at 200V in Laemmli running buffer (3.03g/L Tris base, 14.4g/L glycine, 1g/L SDS) until the dye goes out of the gel. The gel was then divided in two parts, one performed Western-blot and the other for Coomassie blue coloration.

The gel part to be stained was incubated 20min at RT in Coomassie blue solution (1g/L Coomassie brilliant blue powder, 50% methanol, 10% acetic acid) and rinsed in 10% acetic acid until the band reach the desired definition.

The gel part to be blotted was incubated in transfer buffer for 10min (15% methanol, 3.03% Tris w/v, 14.4 % glycine w/v). A nitrocellulose membrane Hybond-P (Amersham #RPN2020F) the same size than the gel was conditioned for 10min in methanol and then put in transfer buffer for at least 1min. 6 pieces of Whatmann paper the same size than the gel were soaked in transfer buffer. The following mounting [3*Whatmann paper, membrane, gel, 3xWhatmann paper] was assembled in a semi-dry transfer unit (Hoefer SemiPhor TE-77), the membrane facing the anode grid and the gel facing the cathode grid. The transfer was carried out at 0.8mA/cm^2 for 2hrs. The membrane was then blocked in [10% skim milk, 1X TBS, 0.1% Tween 20] at RT for 2hrs, incubated with primary antibody (1/1,000) in [2% skim milk, 1X TBS, 0.1% Tween 20] for at least 2hrs at RT or overnight at 4˚C, washed 3 times 5min in [1X TBS, 0.1% Tween], incubated 30min at RT in [10% skim milk, 1X TBS, 0.1% Tween 20], incubated with the secondary antibody [1/2,000 for anti-rabbit HRP-antibody (Invitrogen #62-6120) and 1/10,000 for protein A-HRP (Invitrogen #10-1023)] in [2% skim milk, 1X TBS, 0.1% Tween 20] for 1hr at RT, washed 3 times 10min in [1X TBS, 0.1% Tween], soaked for 1min in ECL solution (VWR #GEHRPN2132) and exposed to a photosensitive film (Fuji Medical X-RAY Film RX 18X24CM 100SHTS, Fujifilm #497690).

Purification of edited RNA by Glyoxal/RNAseT1 cleavage

Protocol optimization

Specificity of the Glyoxal/RNAseT1 cleavage

Mouse brain RNA was combined with glyoxal according to the following conditions:

Tube	1 to 4	5 to 8	9 to 12
Mouse brain RNA 1ug/µL	2.5µL	2.5µL	2.5µL
$NaPO_4$ 100mM pH7	10µL	10µL	10µL
DMSO	50µL	50µL	50µL
Deionized Glyoxal 40%	-	1.5µL	6µL
Water	37.5µL	37.5µL	37.5µL

The samples were incubated for 45min at 37˚C, mixed with 100µL of 1M sodium borate pH7.5 and precipitated with 1mL of 100% ethanol by centrifuging at 17,000g for 15min at 4˚C. The pellets were dried thoroughly under vacuum and resuspended in 21µL of Tris-borate buffer (10mM Tris-HCl pH7.8, 1M sodium borate pH7.5).

The RNAseT1 (Ambion #AM2283) was added to each sample as follow:

Tubes	1, 5, 9	2, 6, 10	3, 7, 11	4, 8, 12
RNAseT1 1U/µL	0µL	1µL	2µL	4µL
DEPC water	4µL	3µL	2µL	0µL

All tubes were incubated 30min at 37˚C and the treatment was stopped by adding 500µL of phenol/chloroform, vortexing, and completing the volume up to 1mL with DEPC water. The emulsion was centrifuged 10min at 13,000g at 4˚C. The phenol extraction was repeated on the aqueous phase, and then the RNA in the aqueous phase was precipitated with 1µL of glycoblue 15mg/mL (Ambion #AM9516), 1/10 vol. of sodium acetate 3M and 0.7vol. of isopropanol. Samples were incubated 5min at RT and centrifuged at 17,000g for 15min at 4˚C. The pellets were washed three times with 70% ethanol, air-dried and resuspended in 500µL of DEPC water. The phenol extraction was repeated on the 500µL RNA to get rid of the excess borate before removing the glyoxal. The pellets were carefully resuspended in 40uL of water (the pellet may be hard to resuspend), 10µL of 100mM sodium phosphate pH7.0 and 50µL DMSO. The samples were incubated 3hrs at 65˚C to remove the glyoxal. The volume of each sample was adjusted to 500µL with DEPC water and precipitated with sodium acetate and isopropanol as described previously. The pellets were resuspended in 10µL of DEPC water, 8µL of RNA was reverse-transcribed with random hexamers and qPCR were performed using the primers GRIA2 and GRIA2 Q/R (sequences in supplemental information) as described below in the PCR paragraph of this section.

Binding of total RNA to streptavidin coated magnetic beads

The different Biotinylation methods were performed on mouse brain total RNA. For the terminal transferase treatment, 10µg of RNA was mixed with 50U of terminal transferase (NEB # M0315S) in the buffer provided by the manufacturer with 0.2mM biotin-ATP (Perkin Elmer #NEL544001EA), incubated for 1hr at 37˚C, cleaned with phenol chloroform as described previously and resuspended in 10µL of DEPC water. For the poly(A)polymerase treatment, 10µg of RNA was mixed with 5U of poly(A)polymerase (NEB # M0276S) in the buffer provided by the manufacturer with 0.05mM biotin-ATP (Perkin Elmer #NEL544001EA) and 0.05mM ATP, incubated for 30min on a temperature gradient from 30˚C to 60˚C [261], cleaned with phenol chloroform and resuspended in 10µL of DEPC water. For the poly(U)polymerase, 10µg of RNA was mixed with 10U of poly(U)polymerase (NEB # M0337S) in the buffer provided by the manufacturer with 0.1mM biotin-ATP (Perkin Elmer #NEL544001EA) and 0.1mM ATP, incubated 1hr at 37˚C, cleaned with phenol chloroform as described previously and resuspended in 10µL of DEPC water. Finally, for the labeling using biotin-hydrazide, 10µg of RNA was oxidized with 10mM sodium periodate freshly prepared in 100mM sodium acetate for 90min et RT in the dark, the reaction was stopped with 250mM KCl (final concentration) for 10min on ice. The RNA was cleaned with phenol chloroform as described previously, resuspended in 90µL of 100mM sodium acetate pH5, mixed with 10µL of 40mM biotin-hydrazide (Sigma B7639) dissolved in DMSO and incubated 4hrs at RT in the dark. The RNA was cleaned with phenol chloroform and resuspended in 10µL of DEPC water. The concentration of each sample is measured with a NanoDrop 1000 Spectrophotometer (ThermoFisher Scientific) and adjusted to 0.5µg/µL. The efficiency of the biotinylation was determined by dot-plot as follow: 2.5µg of biotinylated RNA and 20pmol of biotinylated probe for the positive control (the sequence does not matter, the following one was used: atccaagtactaaccaggcccgaccctgc-biotin) were plotted on a membrane Hybond–N+ (Amersham #RPN119B). The membrane was blocked for 2hrs at RT in [1x PBS, 0.1% Tween 20, 10% skim milk], incubated 1hr with 1/2500 HRP-avidin (Invitrogen #43-4323) in [1x PBS, 0.1% Tween 20, 2% skim milk], washed 3 times in [1x PBS, 0.1% Tween 20], soaked for 1min in ECL solution (VWR #GEHRPN2132) and exposed to a photosensitive film (Fuji Medical X-RAY Film RX 18X24CM 100SHTS, Fujifilm #497690).

To assess the efficacy of the interaction magnetic beads – biotinylated RNA, 5µg of RNA biotinylated with biotin-hydrazide were incubated for 30min at RT on a rotating platform with 0.5mg of streptavidin coated magnetic beads (Invitrogen #653-05) treated to be RNAse free following manufacturer instruction in 1xB&W buffer (5mM Tris-HCl pH7.5, 0.5mM EDTA, 1M NaCl). The beads were washed twice with 1xB&W buffer. The RNA was eluted from the beads 5min at 65˚C in [95% formamide, 10mM EDTA], cleaned with phenol chloroform, resuspended in DEPC water and the quantity and quality of the RNA was determined by running a bioanalyzer nanochip (Agilent #5067-1511).

Interaction between glyoxal and the biotin-streptavidin bond

10µg of biotinylated RNA prepared with biotin-hydrazide was treated with 6µL of 40% glyoxal and 1M sodium borate and resuspended in 10µL Tris-borate buffer as described previously. 2.5µg of treated RNA and series of untreated biotinylated RNA (2.5µg, 1.25µg, 0.625µg, 0.313µg and 0.156µg) were blotted on a membrane Hybond–N+ and revealed as described previously.

Removal of the glyoxal

5 tubes (tubes 1 to 5) containing 2.5µg of mouse brain total RNA, 6µL of Deionized Glyoxal 40%, 10mM Na-PO$_4$ pH7 and 50% DMSO in 100µL and 5 tubes (tubes 6 to 10) containing 2.5µg of mouse brain total RNA, 10mM Na-PO$_4$ pH7 and 50% DMSO in 100µL were incubated for 45min at 37˚C, mixed with 100µL of 1M sodium borate pH7.5 and precipitated with 1mL of 100% ethanol by centrifuging at 17,000g for 15min at 4˚C. The pellets were dried thoroughly under vacuum and resuspended in 21µL of Tris-borate buffer (10mM Tris-HCl pH7.8, 1M sodium borate pH7.5). Each sample was then cleaned by adding 500µL of phenol/chloroform, vortexing, and completing the volume up to 1mL with DEPC water. The emulsion was centrifuged 10min at 13,000g at 4˚C. The phenol extraction was repeated on the aqueous phase, and then the RNA in the aqueous phase was precipitated with 1µL of glycoblue 15mg/mL (Ambion #AM9516), 1/10 vol. of sodium acetate 3M and 0.7vol. of isopropanol. Samples were incubated 5min at RT and centrifuged at 17,000g for 15min at 4˚C. The pellets were washed three times with 70% ethanol, air-dried and resuspended in 500µL of DEPC water. The phenol extraction was repeated on the 500µL RNA to get rid of the excess borate before removing the glyoxal.

The pellets were carefully resuspended according to the following table:

Tube	1 and 6	2, 3 and 7, 8	4, 5 and 9, 10
$NaPO_4$ 100mM pH6		10µL	
$NaPO_4$ 100mM pH7			10µL
DMSO		50µL	50µL
Water	8µL	40µL	40µL

Samples 1 and 6 were kept at -20˚C. The samples 2, 4, 7 and 9 were incubated 3hrs at 25˚C and the samples 3, 8, 5 and 10 were incubated 3hrs at 65˚C. The volume of each sample, except for samples 1 and 6, was adjusted to 500µL with DEPC water and precipitated with sodium acetate and isopropanol as described previously. The pellets were resuspended in 8µL of DEPC water and all samples (1 to 10) were reverse-transcribed with random hexamers. PCRs were performed in 20µL containing 1µL of cDNA, 0.25µM of β-actin primers forward and reverse (sequences in Appendix 3), 1X PCR buffer (fisher biotec #TAQ-1), 1.25mM MgCl2, 1 unit of Taq polymerase (fisher biotec #TAQ-1) with the following program: 95˚C for 2min, [95˚C for 30s, 60˚C for 30s, 72˚C for 30s] for 35 cycles, 72˚C for 7min. PCR products were then run in a 1.5% agarose gel containing ethidium bromide.

Preparation of the I-RNA and B-RNA

A detailed protocol is described in Appendix 5. Briefly, 30µg total RNA was biotinylated, coated with glyoxal and boric acid and bond to streptavidin coated magnetic beads. The RNA-magnetic beads complexes were then treated with RNAseT1 to cleave the inosinylated RNA (I-RNA) off the beads, and the I-RNA was collected. The RNA bound to the beads (B-RNA) was eluted using formamide. The glyoxal was removed by heating from both I-RNA and B-RNA and both samples were cleaned. In order to perform 3'end sequencing on the I-RNA, the I-RNA need 3'end OH and 5'end phosphate, therefore, the I-RNA was treated by TAP to remove the potential 5'cap and by PNK to repair the extremities.

PCR

To assess the efficiency of the enrichment in inosilated RNA, the quantity of the following transcripts were queried by PCR in B-RNA and I-RNA: GRIA2, 5HT2C, KCNA1, GABRA3, RPLP0, β-actin, GAPDH, PPIA and ATP5E. The primers (sequence in Appendix 4) were designed upstream of the editing sites for GRIA2, 5HT2C, KCNA1 and GABRA3. 30ng of I-RNA and 30ng of B-RNA were reverse transcribed with the superscript III kit (Invitrogen #18080-051) according to manufacturer instruction using random hexamer. Real time PCR was performed in triplicate using SYBR Green PCR Master Mix (Applied Biosystems #4309155), 0.5uL of cDNA and 0.25µM primers in 10µL on the ViiA7 Real-Time PCR System (Applied Biosystems). The enrichment of a given target in I-RNA compared to B-RNA was calculated according to the following formula: enrichment = $2^{(Ct_{\text{B-RNA}} - Ct_{\text{I-RNA}})}$ in an experiment performed in triplicate on RNA from the same mouse brain.

Library preparation

The library preparation and deep sequencing were done by Genework (http://www.geneworks.com.au). The I-RNA library was prepared on 200ng I-RNA using the Illumina TruSeq Small RNA kit (Illumina #RS-200-0012) and the library size selection was achieved using AMPure beads (Beckman Coulter #A63880) to select for insert >100nt. The I-RNA library was sequenced on two lanes, generating pair-end reads of 2*65nt. The B-RNA library was prepared on 500ng B-RNA using the Illumina TruSeq RNA kit (Illumina #FC-122-1001) and sequenced on two lanes generating pair-end reads of 65nt.

Deep-sequencing data analysis

The nucleotide density of the raw sequences was measured using the Fastqc software developed by Babraham Bioinformatics (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/). Both libraries were then mapped as paired-end libraries to the mouse genome assembly NCBI37/mm9 (ftp://ftp.ensembl.org/pub/release-64/fasta/mus_musculus/dna/) using BWA [268] and Tophat [249] allowing three mismatches for the mapping of the I-RNA library and two mismatches for the mapping of the B-RNA library. The sequences were then further analyzed using bash scripts and the samtools package [250] (scripts available upon request) to select uniquely mapping tags, to count the number of 3'end extremities, to estimate the ratio of each MM in 3'end and 5'end, to count the number of loci containing an A-to-G MM in 3'end in the I-RNA library. The location of each A-to-I MM was then extracted, and the coverage (number of tags covering this locus) and number of each MM at this exact position were determined.

The loci were then filtered according to the following criteria: (i) contain more than two A-to-G MM, (ii) are covered by more than five tags from both the I-RNA library and the B-RNA library, (iii) contain more A-to-G MM than any other MM, (iv) if the z-test can be done (detail below), the p-value measured by a z-test for the amount of A-to-I MM compared to the other must be < 0.05, or if the z-test is not relevant, the locus must contain at least five times more A-to-G MM than any other one.

The one-tailed two proportion z-test was performed as follow: the proportion of A-to-G mismatch (Pa = number of A-to-G MM / number of tags) and the proportion of other mismatches (Pb = number of A-to-C and A-to-T MM / number of tags) were calculated. The Z-score was then determined applying the following formula with N being the number of tags covering the locus.

$$Zexp = \frac{Pa-Pb}{\sqrt{2 \cdot P \cdot Q / N}}$$

$$P = \frac{Pa+Pb}{2}$$

$$Q = 1-P$$

This test was applied only to loci for which P*N ≥ 5 and Q*N ≥ 5. The p-values were then compiled by comparing the Z-score to the Standard normal table.

Characterization of editing in rRNA

For each site edited at more than 50% and overlapping with region annotated as rRNA (repeat masker: [230]), the full rRNA sequence of the locus was collected and aligned to the mouse 45S rRNA precursor (NR_046233.1), the mouse 28S rRNA (NR_003279), the mouse 18S rRNA (NR_003278.3), the human 18S rRNA (X03205) and the human ribosomal DNA complete repeating unit (U13369.1) using ClustalW [246]. The BioEdit alignment editor [282] was used to assign the exact edited location in mouse to human 18S and 28S rRNAs and these locations were reported on the rRNA modification map realized by Piekna-Przybylska et al. [273] (for 28S rRNA: http://people.biochem.umass.edu/fournierlab/3dmodmap/hum28sseq.php and for 18S rRNA: http://people.biochem.umass.edu/fournierlab/3dmodmap/hum18sseq.php).

Chapter 4: Discussion

Alu elements and ADAR proteins are two highly connected entities, whose expansions correlated strongly with an increase in complexity, and particularly in the human brain. The first one, Alu element, can regulate protein translation, RNA transcription and has the potential to modify the genome. It constitutes the main target of the second one, the ADAR proteins that modify the sequence of Alu element, thereby restraining Alu activities. The roles that each of them play in the organism remain far from being elucidated, for instance no information is available about the transcriptional fate of Alu element, and the targets of ADARs are far from being all identified. The aim of this thesis was to extend the knowledge about Alu elements and the targets of ADAR proteins to have a better understanding of their involvement in the regulation of complexity.

Characterization of Alu elements

The first chapter of this thesis focused on identifying and characterizing the active Alu elements. Two factors needed to be considered before starting the characterization: first, Alu elements are repetitive elements, which mean that all the methods relying on hybridization could not be used to identify them precisely. Thus, sequencing appeared to be the only reliable method. Second, Alu elements are the most abundant repeat in the human genome with more than one million copies, which means that high-throughput methods were required to reach anywhere near the characterization of all the elements. Conveniently, deep-sequencing technology, a method allying sequencing and high-throughput, became available approximately at the beginning of this project [283] and alongside more and more deep-sequencing datasets were published.

In a first part, a preliminary analysis of each Alu element sequence, coordinates and conservation permitted the annotation of Alus in function of their genomic environment, their time of apparition and number of identical sequences in the human genome, revealing that 48% of them are located in intergenic regions and 50% in introns. A similar study reported previously that 66% of Alu elements were located in intron [284], but the analysis was performed on a previous version of the human genome (hg17 NCBI35) which lacked approximately 50Mb and 100,000 Alu elements compared to hg19 on which this analysis was performed.

Following this, an *in silico* analysis of the sequences of Alu elements revealed an important point: the majority of Alu elements preserved enough unique features to be identified precisely by deep sequencing according to the performance of the latter. Confident that deep sequencing could identify Alu elements, the first dataset analysis was conducted to evaluate the potential of Alu elements to bind POLIII and, thus, be transcribed from their internal promoters. It appeared that POLIII recognized 455,950 Alu elements, 64% of them having the B-box necessary for initiating their transcriptions. However, binding of POLIII does not prove that transcription occurs. To identify the panel of Alu elements that are transcribed as independent elements, a protocol was developed to sequence specifically the fraction of RNA containing Alu transcripts of 100 to 300nt, and applied to THP1 RNA. This allowed the identification of approximately 17,000 Alu elements actively transcribed as independent elements in this cell line.

In a next step, the question of the cell localization of these transcripts was addressed. If present in the nucleus, Alu elements may regulate POLII or retrotranspose into the genome; if present in the cytoplasm, they may regulate translation. Three sets of data were used: the first one on DLD-1 cells, sequenced the RNA localized in the nucleus, the cytoplasm or the polysomes [235], the second one in fibroblasts, sequenced the RNA associated to chromatin and the third one, generated in house, sequenced the small RNA <500nt associated with chromatin in HeLa cells; 111,844 Alu elements transcripts could be allocated to cell compartments.

The last screening of Alu elements in deep-sequencing data aimed at profiling Alu elements expression in sixteen normal human tissues. 465,160 elements were detected in a least one of the tissue and the tissue expressing the highest diversity of Alu elements was the adrenal, followed by the ovary, the brain and the testis. The same data were also screened for full length LINE1 expression to see if the material necessary for Alu retrotransposition is also present in normal tissue, and, surprisingly, full length LINE1 elements were found at their highest level in adrenal, brain, ovary and testis, suggesting that Alu retrotransposition in normal tissue was happening. To observe if this phenomenon was limited to organism presenting higher order of cognition, mouse tissues were tested for SINEs and LINE1 expression revealing that mouse also co-expressed autonomous and non-autonomous retrotransposons in most of its tissues.

Finally, two approaches were attempted to show active retrotransposition in tissues: for the first one the gDNA from nine tissues from one mouse were compared by RFLP to look for disparities, and for the second one, pair-end transcriptomic data were screened to find transcripts issued from loci containing a new insertion. None of the two methods worked, the first one was not sensitive enough to detect insertions which are probably happening at a low level and for the second one, the pipeline used to extract transcripts issued from new insertion enriched actually for mapping artifacts. Shortly after this analysis, Baillie et al. [226] showed that retrotransposition was occurring actively in normal human brain by using a method probing the genomes of matched tissues from the same individual like we did in our first approach, with deep sequencing to look for chimeric pair of tags similarly to what we did in our second approach. This study clearly confirmed that retrotransposition was happening in normal somatic tissues in human.

Characterization of the target of editing enzymes

The second chapter of this thesis aimed at identifying by an experimental approach and in a high throughput fashion, the transcripts that are A-to-I edited by the ADARs. First, antibodies were used to pull-down ADARs with the associated RNAs. In a first attempt, a standard immunoprecipitation protocol was used [217, 218] but did not produce enough RNA and the level of enrichment was not satisfying. Then, several pre-enrichment steps were tested to reduce the complexity of the tissue lysate and increase the binding efficiency of the antibody: sucrose gradients containing EDTA were efficient for concentrating ADAR proteins in distinct sucrose fractions and separating ADARs from other proteins such as H3 and GAPDH, however no RNA could be extracted from the ADAR fractions, probably because the conditions did not preserve the structure of the complexes ADAR-RNA; sucrose gradients without EDTA and containing NaCl and $MgCl_2$, to be closer to physiological conditions, preserved the structure of the complexes containing ADARs, however were inefficient at concentrating the complexes; and finally, nucleus fractionation was efficient at enriching for ADAR protein, yet it did not improve the efficiency of the IP.

Another strategy was then considered to target directly the inosine in the edited RNA in a method where RNAs were first immobilized on magnetic beads, then a glyoxal/RNAseT1 treatment cleaved specifically the edited RNAs at their inosine, thus separating edited RNAs from the normal RNA, and at last, the RNAs were identified using deep sequencing. Each components of the method were optimized, the complete protocol compiled and used to characterize the RNAs that are edited in mouse brain. Analysis of the data from the deep sequencing revealed that the protocol was enriching for A-to-G mismatches and identified precisely editing sites. 1,822 editing sites in mouse brain RNA were identified by this method:

1,249 sites were present on known genes, among them, 48 editing sites were in coding region of genes including 28 sites that changed the sequence of the protein and were not reported in previous studies, 895 sites were seating in intronic regions of genes and 306 sites were in UTRs or in non-coding genes; 573 sites were in intergenic regions. A-to-I editing sites were also characterized in RNA families where they have never been observed before, 343 sites were located in 18S and 28S rRNAs, 29 sites in snoRNAs and 60 sites in snRNAs.

Limitation of the study

The first part of this thesis aimed at characterizing the expression of Alu elements as independent element was based mainly on observations of data produced by deep sequencing that were initially generated to interrogate long poly-adenylated transcripts (i. e. DLD-1 dataset, fibroblast dataset and human transcriptomic Atlas), Only two datasets truly identified independent Alu transcripts by excluding physically during the library preparation all long RNAs that could contain Alu in their sequences (i.e. THP1 dataset and CAR dataset on HeLa). To distinguish independent Alu transcripts from Alu included in longer transcripts in the poly-A datasets several filters were applied during the analyses (chapter 2, material and method, Identification of Alu elements in deep-sequencing datasets) to remove all tags mapping to Alu elements but overlapping with known transcripts (e.g. UCSC genes) and uncharacterized transcripts longer than 500nt that correspond to a database of transcripts built with the Illumina dataset. These steps introduced a bias in the analyses of these datasets by excluding systematically all Alu elements overlapping with exons. For instance, 20,680 Alu elements were found in the HeLa and THP1 datasets and were completely absent from all the other transcriptomic datasets, mainly because they were overlapping with larger transcripts. A second bias was introduced by the fact that all these poly-A datasets were generated with protocols that did not give any information about the strand of the transcript, meaning that all Alu elements located in antisense of exons were also excluded and that it was not possible to distinguish between transcripts coming from Alu elements and transcripts coming from the antisense of Alu elements. Again, this bias was not present in the two in-house datasets that were generated with protocols preserving the strand information. Finally, a last bias needs to be considered, it is the propensity of Alu elements to be edited. When analyzing the chimeric transcripts from pair-end data (chapter 2; New insertion of retroelements in specific tissues; Analysis of retroelements insertion by deep-sequencing data analysis), false mapping were discovered for 6.3% of the tags for which sequence polymorphism lead to their mis-mapping. Since Alu elements are repetitive, editing of their sequences can easily lead to false identification of the Alus expressed, and since nothing is known about the editing of the independent Alu transcripts, this bias is hard to rectify, however all the tools necessary to address this problematic were developed during this thesis (see future directions).

In the second part of this thesis, the protocol developed to characterize A-to-I editing sites showed good concordance with sites previously reported in term of precision of the site location and level of editing. However, the deep sequencing of the edited RNA revealed that the protocol could be improved. For instance, in theory, each pair of tag from the I-RNA library is issued from an editing event; 1,398,042 loci were identified in the I-RNA library, approximately 50% of them have a G in 3'end (Figure 33) and only 12,644 loci corresponded to an A-to-G mismatch when mapped to the genome (Table 6). First, the fact that only 50% of the I-RNA 3'end possesses a G, whereas the protocol is meant to identify cleaved sites which should all be read as guanosine, could be explained by degradation occurring after the cleavage, probably during the removal of the glyoxal for 3hrs at 65°C, leading to the cleavage of the 3'end nucleotides. This effect could be reduced by adding EDTA in the solution to avoid fragmentation due to heat and the presence of dications. To remove glyoxal without heating the sample, other possibilities that need to be tested would be the use of high concentration of competitors such as GTP, or the use of a chemical reaction transforming the glyoxal into a compound that does not interact with RNA. Second, the fact that only 12,644 loci with an A-to-G mismatch were identified could be explained by an incomplete protection of the guanosine by the glyoxal: if the RNA was not treated with a sufficient amount of glyoxal, unprotected guanosines would be cleaved like inosine thus increasing the number of tags that did not identify an editing site. The amount of glyoxal required for an optimal cleavage of inosine was determined experimentally (Figure 28), however only two concentrations were tested and repeating the experiment including more conditions may allow a more precise estimation of the optimal concentration of glyoxal. A last concern with this protocol is that so far it has been tried only on mouse tissues. Initially, it was decided to develop the method on mouse tissue for several evident facts such as the cost and availability of the tissue compared to human tissue and the fact that the mouse genome which is available is from the strain of mouse used in our experiment (C57BL/6J), which means that heterogeneity observed between transcripts and genome are more likely to be due to editing than allelic differences between the gDNA donor and RNA donor. However, realizing this experiment on mouse also rendered hard the comparison between our data and previous high throughput analyses which were all realized in human. For instance, the DARNED database on editing sites in human lists 42,042 A-to-I sites in the human genome hg19 [266], of which only 1,794 have orthologous positions in the mouse genome mm9 and this does not mean that the orthologous site is edited in mouse: the intersection of our data with these 1,794 sites return 13 loci edited in mouse which are all located in exons of protein coding genes. Also, the mouse brain transcriptome is edited at a lower extent than the human brain transcriptome, which means that the protocol may behave differently with the human brain RNA which contains at least 17 times more inosine than mouse brain RNA [285].

Future directions

This thesis described precisely the expression pattern and cell localization of 524,526 Alu elements in human and an efficient method to characterize A-to-I editing sites. However, several experiments are required to complete this study.

First, considering the biases described previously for the Alu transcription analysis, the investigation of the expression of Alu transcripts in normal human tissue would benefit from using the protocol developed in this thesis on RNA from THP1. This would allow a more comprehensive identification of transcriptionally active Alu elements. Following this, another direction that need to be addressed, is the function of Alu elements transcription. Attempts were made to characterize the elements able to retrotranspose, and Baillie et al. developed a capture-seq method to identify them [226], however our attempts failed and Baillie et al. limited their study to brain. We proved here that Alu elements are expressed in a broad range of human tissues and in a tissue specific fashion, it would be interesting to use our results to design a capture-array targeting Alu expressed in specific tissues to probe the genome of these tissues for new insertions of Alu element. We have also shown that retroelements were expressed in mouse normal tissues; it is of high importance to confirm this finding. If retrotransposition can be observed in mouse normal tissues, knock-out experiments can be designed to reveal the importance of such mechanism in tissue development and function.

Second, the characterization of editing sites was performed in mouse brain RNA only. Efforts should be made to apply this protocol to human brain where editing was shown to be crucial for its normal function. In addition, the editing of snRNA, snoRNA and rRNA should be confirmed and may help elucidating the regulation and selectivity of these elements.

Finally, the question of the editing of independent Alu transcripts should be addressed. If demonstrated, this would show that independent Alu transcripts can regulate editing in trans, giving a new role to Alu elements. Presently, it is known that Alu elements promote editing when two inverted repeats are integrated in a longer transcript: they form a double stranded structure which is targeted by ADAR. If a transcript contain only one repeat, it is unlikely that it will form the double stranded structure required by ADAR, excepting if an independent Alu transcript is coupled with the integrated Alu element. Using the protocol developed on THP1 to enrich for RNA < 500nt followed by the protocol using glyoxal to identify editing sites would be a good approach to investigate the editing of independent Alu transcripts.

Conclusion

This thesis aimed at characterizing Alu elements transcription and A-to-I RNA editing in human to clarify the role they play in human biology. In the first part, the first comprehensive database on Alu elements was generated giving detailed information for each Alu elements such as location, conservation, genomic environment, cell localization and level of expression in normal human tissues. The generation of this database revealed that the old Alu families, AluJ and AluS, were expressed at a larger extent than the young AluY; this observation was supported by POLIII CHIP-seq data and deep sequencing of small RNA, although it is in contradiction with previous reports [67-69]. It also revealed that Alu elements were expressed in a broad range of normal tissues, actually at least 15,000 different Alu elements were observed in each tissue investigated (i.e. the tissue expressing the least Alu elements was skeletal muscle with 15,297 elements; Table 3); and extension of the analysis to LINE1 indicated that full length LINE1 were co-expressed with Alu elements in all tissue observed, rendering retrotransposition possible. On this basis, I propose that retrotransposition is a tightly regulated mechanism in normal tissue that aims at modifying the core of the genome. This proposition is supported by previous reports describing genomic mosaicism promoted by retrotransposons in several part of the human brain [226, 286-289]. In addition, I propose that this mechanism is occurring in most tissues and is not exclusive to human but also occurs in mouse.

In the second part, a new protocol was developed to identify the RNA edited by ADAR at the transcriptome scale and was used to characterize the mouse brain RNA editome. 1,822 editing sites were characterized, the majority of which were located in intronic regions similarly to what is observed in human [172]. A-to-I editing was detected for the first time in rRNA, snoRNA and snRNA, an observation rendered possible by applying the protocol on total RNA instead of polyadenylated RNA or known genes as it is usually done in the targeted approach previously described [9, 50, 51, 173, 217-220, 222]. In order to fully understand the impact of A-to-I editing on Alu elements, I suggest continuing this study by applying this protocol to RNA from human brain or adrenal, two somatic tissues expressing high level of Alu elements.

This will characterize not only the editing of genes expressed in these tissues but also the editing occurring on independent Alu transcripts which may have an impact on their propensity to retrotranspose.

This project was initiated based on the hypothesis that editing and Alu elements expression play a role in the regulation of the cognition, a human brain specific trait distinctive by its complexity. Evidence that both phenomena occur at a lesser extent in mouse and in other tissues suggests that their synergy fulfill also other roles than cognition, but it is undeniable that they lead to a more complex system by increasing its plasticity. To conclude, this thesis only started to clarify the role that Alu elements and A-to-I editing could play in complex mechanisms by identifying the different performers. To complete this study, several questions remain to be answered. What are the roles and partners of all those Alu transcripts? Can they retrotranspose in somatic tissues? Are they edited? Can they regulate editing? How Alu elements and editing are controlled? If this thesis did not address any of these questions, it gave the tools necessary to answer several of them.

References

1. Lander ES et al. (2001) Initial sequencing and analysis of the human genome. Nature, **409**:860-921.

2. Grover D et al. (2004) Alu repeat analysis in the complete human genome: trends and variations with respect to genomic composition. Bioinformatics, **20**:813-7.

3. Dewannieux M et al. (2003) LINE-mediated retrotransposition of marked Alu sequences. Nat Genet, **35**:41-8.

4. Wallace N et al. (2008) LINE-1 ORF1 protein enhances Alu SINE retrotransposition. Gene, **419**:1-6.

5. Bass BL (2002) RNA editing by adenosine deaminases that act on RNA. Annu Rev Biochem, **71**:817-46.

6. Rueter SM et al. (1999) Regulation of alternative splicing by RNA editing. Nature, **399**:75-80.

7. Petschek JP et al. (1997) RNA editing and alternative splicing generate mRNA transcript diversity from the Drosophila 4f-rnp locus. Gene, **204**:267-76.

8. Pullirsch D et al. (2010) Proteome diversification by adenosine to inosine RNA editing. RNA Biol, **7**:205-12.

9. Levanon EY et al. (2004) Systematic identification of abundant A-to-I editing sites in the human transcriptome. Nat Biotechnol, **22**:1001-5.

10. Chen LL et al. (2008) Alu element-mediated gene silencing. EMBO J, **27**:1694-705.

11. Chen LL et al. (2008) Gene regulation by SINES and inosines: biological consequences of A-to-I editing of Alu element inverted repeats. Cell Cycle, **7**:3294-301.

12. Jurka J et al. (2004) Duplication, coclustering, and selection of human Alu retrotransposons. Proc Natl Acad Sci U S A, **101**:1268-72.

13. Labuda D et al. (1994) Evolution of secondary structure in the family of 7SL-like RNAs. J Mol Evol, **39**:506-18.

14. Jurka J (2004) Evolutionary impact of human Alu repetitive elements. Curr Opin Genet Dev, **14**:603-8.

15. Ullu E et al. (1984) Alu sequences are processed 7SL RNA genes. Nature, **312**:171-2.

16. Britten RJ et al. (1988) Sources and evolution of human Alu repeated sequences. Proc Natl Acad Sci U S A, **85**:4770-4.

17. Jurka J et al. (1991) Reconstruction and analysis of human Alu genes. J Mol Evol, **32**:105-21.

18. Batzer MA et al. (1996) Standardized nomenclature for Alu repeats. J Mol Evol, **42**:3-6.

19. Price AL et al. (2004) Whole-genome analysis of Alu repeat elements reveals complex evolutionary history. Genome Res, **14**:2245-52.

20. Xing J et al. (2007) Mobile DNA elements in primate and human evolution. Am J Phys Anthropol, **Suppl 45**:2-19.

21. Strub K et al. (1991) Binding sites of the 9- and 14-kilodalton heterodimeric protein subunit of the signal recognition particle (SRP) are contained exclusively in the Alu domain of SRP RNA and contain a sequence motif that is conserved in evolution. Mol Cell Biol, **11**:3949-59.

22. Weichenrieder O et al. (2000) Structure and assembly of the Alu domain of the mammalian signal recognition particle. Nature, **408**:167-73.

23. Sinnett D et al. (1991) Alu RNA secondary structure consists of two independent 7 SL RNA-like folding units. J Biol Chem, **266**:8675-8.

24. Hasler J et al. (2006) Alu RNP and Alu RNA regulate translation initiation in vitro. Nucleic Acids Res, **34**:2374-85.

25. Scott AF et al. (1987) Origin of the human L1 elements: proposed progenitor genes deduced from a consensus DNA sequence. Genomics, **1**:113-25.

26. Martin SL et al. (2008) A single amino acid substitution in ORF1 dramatically decreases L1 retrotransposition and provides insight into nucleic acid chaperone activity. Nucleic Acids Res, **36**:5845-54.

27. Martin SL et al. (2001) Nucleic acid chaperone activity of the ORF1 protein from the mouse LINE-1 retrotransposon. Mol Cell Biol, **21**:467-75.

28. Martin SL et al. (2005) LINE-1 retrotransposition requires the nucleic acid chaperone activity of the ORF1 protein. J Mol Biol, **348**:549-61.

29. Martin SL (2006) The ORF1 Protein Encoded by LINE-1: Structure and Function During L1 Retrotransposition. J Biomed Biotechnol, **2006**:45621.

30. Feng Q et al. (1996) Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. Cell, **87**:905-16.

31. Mathias SL et al. (1991) Reverse transcriptase encoded by a human transposable element. Science, **254**:1808-10.

32. Jurka J (1997) Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. Proc Natl Acad Sci U S A, **94**:1872-7.

33. Ostertag EM et al. (2001) Biology of mammalian L1 retrotransposons. Annu Rev Genet, **35**:501-38.

34. McClintock B (1956) Controlling elements and the gene. Cold Spring Harb Symp Quant Biol, **21**:197-216.

35. Comfort NC (2001) From controlling elements to transposons: Barbara McClintock and the Nobel Prize. Trends Genet, **17**:475-8.

36. Orgel LE et al. (1980) Selfish DNA: the ultimate parasite. Nature, **284**:604-7.

37. Doolittle WF et al. (1980) Selfish genes, the phenotype paradigm and genome evolution. Nature, **284**:601-3.

38. Orgel LE et al. (1980) Selfish DNA. Nature, **288**:645-6.

39. Lowe CB et al. (2007) Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. Proc Natl Acad Sci U S A, **104**:8005-10.

40. Medstrand P et al. (2002) Retroelement distributions in the human genome: variations associated with age and proximity to genes. Genome Res, **12**:1483-95.

41. Brookfield JF (2001) Selection on Alu sequences? Curr Biol, **11**:R900-1.

42. Cordaux R et al. (2006) Recently integrated Alu retrotransposons are essentially neutral residents of the human genome. Gene, **373**:138-44.

43. Kidwell MG et al. (2000) Transposable elements and host genome evolution. Trends Ecol Evol, **15**:95-99.

44. Muotri AR et al. (2007) The necessary junk: new functions for transposable elements. Hum Mol Genet, **16 Spec No. 2**:R159-67.

45. Faulkner GJ et al. (2009) The regulated retrotransposon transcriptome of mammalian cells. Nat Genet, **41**:563-71.

46. Lee J et al. (2008) Chromosomal inversions between human and chimpanzee lineages caused by retrotransposons. PLoS ONE, **3**:e4047.

47. Lev-Maor G et al. (2008) Intronic Alus influence alternative splicing. PLoS Genet, **4**:e1000204.

48. Chen C et al. (2009) Using Alu elements as polyadenylation sites: A case of retroposon exaptation. Mol Biol Evol, **26**:327-34.

49. Kawahara Y et al. (2006) Extensive adenosine-to-inosine editing detected in Alu repeats of antisense RNAs reveals scarcity of sense-antisense duplex formation. FEBS Lett, **580**:2301-5.

50. Athanasiadis A et al. (2004) Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. PLoS Biol, **2**:e391.

51. Kim DD et al. (2004) Widespread RNA editing of embedded alu elements in the human transcriptome. Genome Res, **14**:1719-25.

52. Courseaux A et al. (2001) Birth of two chimeric genes in the Hominidae lineage. Science, **291**:1293-7.

53. Lee Y et al. (2006) Evolution and expression of chimeric POTE-actin genes in the human genome. Proc Natl Acad Sci U S A, **103**:17885-90.

54. Sayah DM et al. (2004) Cyclophilin A retrotransposition into TRIM5 explains owl monkey resistance to HIV-1. Nature, **430**:569-73.

55. Goodier JL et al. (2000) Transduction of 3'-flanking sequences is common in L1 retrotransposition. Hum Mol Genet, **9**:653-7.

56. Kaessmann H et al. (2008) RNA-based gene duplication: mechanistic and evolutionary insights. Nat Rev Genet.

57. Bourque G et al. (2008) Evolution of the mammalian transcription factor binding repertoire via transposable elements. Genome Res, **18**:1752-62.

58. Antonaki A et al. (2011) Genomic analysis reveals a novel nuclear factor-kappaB (NF-kappaB)-binding site in Alu-repetitive elements. J Biol Chem, **286**:38768-82.

59. Speek M (2001) Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. Mol Cell Biol, **21**:1973-85.

60. Buzdin A et al. (2006) At least 50% of human-specific HERV-K (HML-2) long terminal repeats serve in vivo as active promoters for host nonrepetitive DNA transcription. J Virol, **80**:10752-62.

61. Okamura K et al. (2008) Retrotransposition as a source of new promoters. Mol Biol Evol, **25**:1231-8.

62. Borchert GM et al. (2006) RNA polymerase III transcribes human microRNAs. Nat Struct Mol Biol, **13**:1097-101.

63. Umylny B et al. (2007) Evidence of Alu and B1 expression in dbEST. Arch Androl, **53**:207-18.

64. Shankar R et al. (2004) Evolution and distribution of RNA polymerase II regulatory sites from RNA polymerase III dependant mobile Alu elements. BMC Evol Biol, **4**:37.

65. Liu WM et al. (1995) Cell stress and translational inhibitors transiently increase the abundance of mammalian SINE transcripts. Nucleic Acids Res, **23**:1758-65.

66. Maraia RJ et al. (1993) Multiple dispersed loci produce small cytoplasmic Alu RNA. Mol Cell Biol, **13**:4233-41.

67. Shaikh TH et al. (1997) cDNAs derived from primary and small cytoplasmic Alu (scAlu) transcripts. J Mol Biol, **271**:222-34.

68. Sinnett D et al. (1992) Alu RNA transcripts in human embryonal carcinoma cells. Model of post-transcriptional selection of master sequences. J Mol Biol, **226**:689-706.

69. Bennett EA et al. (2008) Active Alu retrotransposons in the human genome. Genome Res, **18**:1875-83.

70. Schmid CW (1998) Does SINE evolution preclude Alu function? Nucleic Acids Res, **26**:4541-50.

71. Li TH et al. (2001) Differential stress induction of individual Alu loci: implications for transcription and retrotransposition. Gene, **276**:135-41.

72. Tang RB et al. (2005) Increased level of polymerase III transcribed Alu RNA in hepatocellular carcinoma tissue. Mol Carcinog, **42**:93-6.

73. Chu WM et al. (1998) Potential Alu function: regulation of the activity of double-stranded RNA-activated kinase PKR. Mol Cell Biol, **18**:58-68.

74. Johanning K et al. (2003) Potential for retroposition by old Alu subfamilies. J Mol Evol, **56**:658-64.

75. Chang DY et al. (1996) Monomeric scAlu and nascent dimeric Alu RNAs induced by adenovirus are assembled into SRP9/14-containing RNPs in HeLa cells. Nucleic Acids Res, **24**:4165-70.

76. Chen Y et al. (1998) Accurate 3' end processing and adenylation of human signal recognition particle RNA and alu RNA in vitro. J Biol Chem, **273**:35023-31.

77. Jacobson MR et al. (1998) Localization of signal recognition particle RNA in the nucleolus of mammalian cells. Proc Natl Acad Sci U S A, **95**:7981-6.

78. Perumal K et al. (2001) Purification, characterization, and cloning of the cDNA of human signal recognition particle RNA 3'-adenylating enzyme. J Biol Chem, **276**:21791-6.

79. Hasler J et al. (2006) Alu elements as regulators of gene expression. Nucleic Acids Res, **34**:5491-7.

80. Rubin CM et al. (2002) Selective stimulation of translational expression by Alu RNA. Nucleic Acids Res, **30**:3253-61.

81. Espinoza CA et al. (2004) B2 RNA binds directly to RNA polymerase II to repress transcript synthesis. Nat Struct Mol Biol, **11**:822-9.

82. Espinoza CA et al. (2007) Characterization of the structure, function, and mechanism of B2 RNA, an ncRNA repressor of RNA polymerase II transcription. RNA, **13**:583-96.

83. Mariner PD et al. (2008) Human Alu RNA is a modular transacting repressor of mRNA transcription during heat shock. Mol Cell, **29**:499-509.

84. Watson JB et al. (1987) Primate brain-specific cytoplasmic transcript of the Alu repeat family. Mol Cell Biol, **7**:3324-7.

85. Kuryshev VY et al. (2001) Birth of a gene: locus of neuronal BC200 snmRNA in three prosimians and human BC200 pseudogenes as archives of change in the Anthropoidea lineage. J Mol Biol, **309**:1049-66.

86. Martignetti JA et al. (1993) BC200 RNA: a neural RNA polymerase III product encoded by a monomeric Alu element. Proc Natl Acad Sci U S A, **90**:11563-7.

87. Tiedge H et al. (1993) Primary structure, neural-specific expression, and dendritic location of human BC200 RNA. J Neurosci, **13**:2382-90.

88. Chen W et al. (1997) Expression of neural BC200 RNA in human tumours. J Pathol, **183**:345-51.

89. Kremerskothen J et al. (1998) Heterodimer SRP9/14 is an integral part of the neural BC200 RNP in primate brain. Neurosci Lett, **245**:123-6.

90. Muddashetty R et al. (2002) Poly(A)-binding protein is associated with neuronal BC1 and BC200 ribonucleoprotein particles. J Mol Biol, **321**:433-45.

91. Zalfa F et al. (2003) The fragile X syndrome protein FMRP associates with BC1 RNA and regulates the translation of specific mRNAs at synapses. Cell, **112**:317-27.

92. Lin D et al. (2008) Translational control by a small RNA: dendritic BC1 RNA targets the eukaryotic initiation factor 4A helicase mechanism. Mol Cell Biol, **28**:3008-19.

93. Muslimov IA et al. (2002) A small RNA in testis and brain: implications for male germ cell development. J Cell Sci, **115**:1243-50.

94. Tiedge H et al. (1991) Dendritic location of neural BC1 RNA. Proc Natl Acad Sci U S A, **88**:2093-7.

95. Sutcliffe JG et al. (1984) Control of neuronal gene expression. Science, **225**:1308-15.

96. Khanam T et al. (2007) Two primate-specific small non-protein-coding RNAs in transgenic mice: neuronal expression, subcellular localization and binding partners. Nucleic Acids Res, **35**:529-39.

97. Ludwig A et al. (2005) An unusual primate locus that attracted two independent Alu insertions and facilitates their transcription. J Mol Biol, **350**:200-14.

98. Hulme AE et al. (2007) Selective inhibition of Alu retrotransposition by APOBEC3G. Gene, **390**:199-205.

99. Chiu YL et al. (2006) High-molecular-mass APOBEC3G complexes restrict Alu retrotransposition. Proc Natl Acad Sci U S A, **103**:15588-93.

100. Miki T et al. (2005) The role of mammalian Staufen on mRNA traffic: a view from its nucleocytoplasmic shuttling function. Cell Struct Funct, **30**:51-6.

101. Bulliard Y et al. (2009) Functional analysis and structural modeling of human APOBEC3G reveal the role of evolutionarily conserved elements in the inhibition of human immunodeficiency virus type 1 infection and Alu transposition. J Virol, **83**:12611-21.

102. Hill MS et al. (2006) APOBEC3G expression is restricted to neurons in the brains of pigtailed macaques. AIDS Res Hum Retroviruses, **22**:541-50.

103. Benne R et al. (1986) Major transcript of the frameshifted coxII gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. Cell, **46**:819-26.

104. Chen SH et al. (1987) Apolipoprotein B-48 is the product of a messenger RNA with an organ-specific in-frame stop codon. Science, **238**:363-6.

105. Driscoll DM et al. (1989) An in vitro system for the editing of apolipoprotein B mRNA. Cell, **58**:519-25.

106. Bass BL et al. (1988) An unwinding activity that covalently modifies its double-stranded RNA substrate. Cell, **55**:1089-98.

107. Bass BL et al. (1987) A developmentally regulated activity that unwinds RNA duplexes. Cell, **48**:607-13.

108. Brennicke A et al. (1999) RNA editing. FEMS Microbiol Rev, **23**:297-316.

109. Jin Y et al. (2009) Origins and evolution of ADAR-mediated RNA editing. IUBMB Life, **61**:572-8.

110. Keegan LP et al. (2004) Adenosine deaminases acting on RNA (ADARs): RNA-editing enzymes. Genome Biol, **5**:209.

111. George CX et al. (1999) Human RNA-specific adenosine deaminase ADAR1 transcripts possess alternative exon 1 structures that initiate from different promoters, one constitutively active and the other interferon inducible. Proc Natl Acad Sci U S A, **96**:4621-6.

112. George CX et al. (2005) Expression of interferon-inducible RNA adenosine deaminase ADAR1 during pathogen infection and mouse embryo development involves tissue-selective promoter utilization and alternative splicing. J Biol Chem, **280**:15020-8.

113. Patterson JB et al. (1995) Expression and regulation by interferon of a double-stranded-RNA-specific adenosine deaminase from human cells: evidence for two forms of the deaminase. Mol Cell Biol, **15**:5376-88.

114. Herbert A et al. (1997) A Z-DNA binding domain present in the human editing enzyme, double-stranded RNA adenosine deaminase. Proc Natl Acad Sci U S A, **94**:8421-6.

115. Schwartz T et al. (1999) Proteolytic dissection of Zab, the Z-DNA-binding domain of human ADAR1. J Biol Chem, **274**:2899-906.

116. Shtrichman R et al. (2002) Tissue selectivity of interferon-stimulated gene expression in mice infected with Dam(+) versus Dam(-) Salmonella enterica serovar Typhimurium strains. Infect Immun, **70**:5579-88.

117. Athanasiadis A et al. (2005) The crystal structure of the Zbeta domain of the RNA-editing enzyme ADAR1 reveals distinct conserved surfaces among Z-domains. J Mol Biol, **351**:496-507.

118. Liu Y et al. (1997) Functionally distinct double-stranded RNA-binding domains associated with alternative splice site variants of the interferon-inducible double-stranded RNA-specific adenosine deaminase. J Biol Chem, **272**:4419-28.

119. Desterro JM et al. (2003) Dynamic association of RNA-editing enzymes with the nucleolus. J Cell Sci, **116**:1805-18.

120. Peng PL et al. (2006) ADAR2-dependent RNA editing of AMPA receptor subunit GluR2 determines vulnerability of neurons in forebrain ischemia. Neuron, **49**:719-33.

121. Kawahara Y et al. (2005) Novel splice variants of human ADAR2 mRNA: skipping of the exon encoding the dsRNA-binding domains, and multiple C-terminal splice sites. Gene, **363**:193-201.

122. Agranat L et al. (2010) A novel tissue-specific alternatively spliced form of the A-to-I RNA editing enzyme ADAR2. RNA Biol, **7**:253-62.

123. Slavov D et al. (2002) Phylogenetic comparison of the pre-mRNA adenosine deaminase ADAR2 genes and transcripts: conservation and diversity in editing site sequence and alternative splicing patterns. Gene, **299**:83-94.

124. Ohman M et al. (2000) In vitro analysis of the binding of ADAR2 to the pre-mRNA encoding the GluR-B R/G site. RNA, **6**:687-97.

125. Kallman AM et al. (2003) ADAR2 A-->I editing: site selectivity and editing efficiency are separate events. Nucleic Acids Res, **31**:4874-81.

126. Gerber A et al. (1997) Two forms of human double-stranded RNA-specific editase 1 (hRED1) generated by the insertion of an Alu cassette. RNA, **3**:453-63.

127. Maas S et al. (2009) Novel exon of mammalian ADAR2 extends open reading frame. PLoS One, **4**:e4225.

128. Mittaz L et al. (1997) Cloning of a human RNA editing deaminase (ADARB1) of glutamate receptors that maps to chromosome 21q22.3. Genomics, **41**:210-7.

129. Maas S et al. (2009) Identification of a selective nuclear import signal in adenosine deaminases acting on RNA. Nucleic Acids Res, **37**:5822-9.

130. Chen CX et al. (2000) A third member of the RNA-specific adenosine deaminase gene family, ADAR3, contains both single- and double-stranded RNA binding domains. Rna, **6**:755-67.

131. Melcher T et al. (1996) RED2, a brain-specific member of the RNA-specific adenosine deaminase family. J Biol Chem, **271**:31795-8.

132. Cho DS et al. (2003) Requirement of dimerization for RNA editing activity of adenosine deaminases acting on RNA. J Biol Chem, **278**:17093-102.

133. Polson AG et al. (1991) The mechanism of adenosine to inosine conversion by the double-stranded RNA unwinding/modifying activity: a high-performance liquid chromatography-mass spectrometry analysis. Biochemistry, **30**:11507-14.

134. Nishikura K et al. (1991) Substrate specificity of the dsRNA unwinding/modifying activity. EMBO J, **10**:3523-32.

135. Reenan RA (2005) Molecular determinants and guided evolution of species-specific RNA editing. Nature, **434**:409-13.

136. Gallo A et al. (2003) An ADAR that edits transcripts encoding ion channel subunits functions as a dimer. EMBO J, **22**:3421-30.

137. Poulsen H et al. (2006) Dimerization of ADAR2 is mediated by the double-stranded RNA binding domain. RNA, **12**:1350-60.

138. Jaikaran DC et al. (2002) Adenosine to inosine editing by ADAR2 requires formation of a ternary complex on the GluR-B R/G site. J Biol Chem, **277**:37624-9.

139. Polson AG et al. (1994) Preferential selection of adenosines for modification by double-stranded RNA adenosine deaminase. EMBO J, **13**:5701-11.

140. Kimelman D et al. (1989) An antisense mRNA directs the covalent modification of the transcript encoding fibroblast growth factor in Xenopus oocytes. Cell, **59**:687-96.

141. Higuchi M et al. (1993) RNA editing of AMPA receptor subunit GluR-B: a base-paired intron-exon structure determines position and efficiency. Cell, **75**:1361-70.

142. Lomeli H et al. (1994) Control of kinetic properties of AMPA receptor channels by nuclear RNA editing. Science, **266**:1709-13.

143. Wang Q et al. (2000) Altered G protein-coupling functions of RNA editing isoform and splicing variant serotonin2C receptors. J Neurochem, **74**:1290-300.

144. Burns CM et al. (1997) Regulation of serotonin-2C receptor G-protein coupling by RNA editing. Nature, **387**:303-8.

145. Wong SK et al. (2001) Substrate recognition by ADAR1 and ADAR2. RNA, **7**:846-58.

146. Lehmann KA et al. (2000) Double-stranded RNA adenosine deaminases ADAR1 and ADAR2 have overlapping specificities. Biochemistry, **39**:12875-84.

147. Eggington JM et al. (2011) Predicting sites of ADAR editing in double-stranded RNA. Nat Commun, **2**:319.

148. Enstero M et al. (2009) Recognition and coupling of A-to-I edited sites are determined by the tertiary structure of the RNA. Nucleic Acids Res, **37**:6916-26.

149. Fitzgerald LW et al. (1999) Messenger RNA editing of the human serotonin 5-HT2C receptor. Neuropsychopharmacology, **21**:82S-90S.

150. Niswender CM et al. (1999) RNA editing of the human serotonin 5-hydroxytryptamine 2C receptor silences constitutive activity. J Biol Chem, **274**:9472-8.

151. Schmauss C et al. (2010) The roles of phospholipase C activation and alternative ADAR1 and ADAR2 pre-mRNA splicing in modulating serotonin 2C-receptor editing in vivo. RNA, **16**:1779-85.

152. Maas S et al. (1996) Structural requirements for RNA editing in glutamate receptor pre-mRNAs by recombinant double-stranded RNA adenosine deaminase. J Biol Chem, **271**:12221-6.

153. Melcher T et al. (1996) A mammalian RNA editing enzyme. Nature, **379**:460-4.

154. Yang JH et al. (1997) Purification and characterization of a human RNA adenosine deaminase for glutamate receptor B pre-mRNA editing. Proc Natl Acad Sci U S A, **94**:4354-9.

155. Vitali P et al. (2005) ADAR2-mediated editing of RNA substrates in the nucleolus is inhibited by C/D small nucleolar RNAs. J Cell Biol, **169**:745-53.

156. Sansam CL et al. (2003) Modulation of RNA editing by functional nucleolar sequestration of ADAR2. Proc Natl Acad Sci U S A, **100**:14018-23.

157. Marcucci R et al. (2011) Pin1 and WWP2 regulate GluR2 Q/R site RNA editing by ADAR2 with opposing effects. EMBO J, **30**:4211-22.

158. Macbeth MR et al. (2005) Inositol hexakisphosphate is bound in the ADAR2 core and required for RNA editing. Science, **309**:1534-9.

159. Yi-Brunozzi HY et al. (1999) Synthetic substrate analogs for the RNA-editing adenosine deaminase ADAR-2. Nucleic Acids Res, **27**:2912-7.

160. Cavaille J et al. (2000) Identification of brain-specific and imprinted small nucleolar RNA genes exhibiting an unusual genomic organization. Proc Natl Acad Sci U S A, **97**:14311-6.

161. Basilio C et al. (1962) Synthetic polynucleotides and the amino acid code. V. Proc Natl Acad Sci U S A, **48**:613-6.

162. Nishikura K (1992) Modulation of double-stranded RNAs in vivo by RNA duplex unwindase. Ann N Y Acad Sci, **660**:240-50.

163. Kim U et al. (1993) Double-stranded RNA adenosine deaminase as a potential mammalian RNA editing factor. Semin Cell Biol, **4**:285-93.

164. Sommer B et al. (1991) RNA editing in brain controls a determinant of ion flow in glutamate-gated channels. Cell, **67**:11-9.

165. Hume RI et al. (1991) Identification of a site in glutamate receptor subunits that controls calcium permeability. Science, **253**:1028-31.

166. Higuchi M et al. (2000) Point mutation in an AMPA receptor gene rescues lethality in mice deficient in the RNA-editing enzyme ADAR2. Nature, **406**:78-81.

167. Schmauss C (2005) Regulation of serotonin 2C receptor pre-mRNA editing by serotonin. Int Rev Neurobiol, **63**:83-100.

168. Hoopengardner B et al. (2003) Nervous system targets of RNA editing identified by comparative genomics. Science, **301**:832-6.

169. Ohlson J et al. (2007) Editing modifies the GABA(A) receptor subunit alpha3. RNA, **13**:698-703.

170. Daniel C et al. (2010) Adenosine-to-inosine RNA editing affects trafficking of the gamma-aminobutyric acid type A (GABA(A)) receptor. J Biol Chem, **286**:2031-40.

171. Bhalla T et al. (2004) Control of human potassium channel inactivation by editing of a small mRNA hairpin. Nat Struct Mol Biol, **11**:950-6.

172. Morse DP et al. (2002) RNA hairpins in noncoding regions of human brain and Caenorhabditis elegans mRNA are edited by adenosine deaminases that act on RNA. Proc Natl Acad Sci U S A, **99**:7906-11.

173. Blow M et al. (2004) A survey of RNA editing in human brain. Genome Res, **14**:2379-87.

174. Beghini A et al. (2000) RNA hyperediting and alternative splicing of hematopoietic cell phosphatase (PTPN6) gene in acute myeloid leukemia. Hum Mol Genet, **9**:2297-304.

175. Flomen R et al. (2004) Evidence that RNA editing modulates splice site selection in the 5-HT2C receptor gene. Nucleic Acids Res, **32**:2113-22.

176. Carmi S et al. (2011) Identification of widespread ultra-edited human RNAs. PLoS Genet, **7**:e1002317.

177. Chen LL et al. (2009) Altered nuclear retention of mRNAs containing inverted repeats in human embryonic stem cells: functional role of a nuclear noncoding RNA. Mol Cell, **35**:467-78.

178. Zhang Z et al. (2001) The fate of dsRNA in the nucleus: a p54(nrb)-containing complex mediates the nuclear retention of promiscuously A-to-I edited RNAs. Cell, **106**:465-75.

179. Prasanth KV et al. (2005) Regulating gene expression through RNA nuclear retention. Cell, **123**:249-63.

180. Scadden AD (2005) The RISC subunit Tudor-SN binds to hyper-edited double-stranded RNA and promotes its cleavage. Nat Struct Mol Biol, **12**:489-96.

181. Scadden AD (2007) Inosine-containing dsRNA binds a stress-granule-like complex and downregulates gene expression in trans. Mol Cell, **28**:491-500.

182. Weissbach R et al. (2012) Tudor-SN and ADAR1 are components of cytoplasmic stress granules. RNA.

183. Bartel DP (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. Cell, **116**:281-97.

184. Kawahara Y et al. (2008) Frequency and fate of microRNA editing in human brain. Nucleic Acids Res, **36**:5270-80.

185. Borchert GM et al. (2009) Adenosine deamination in human transcripts generates novel microRNA binding sites. Hum Mol Genet, **18**:4801-7.

186. Luciano DJ et al. (2004) RNA editing of a miRNA precursor. Rna, **10**:1174-7.

187. Blow MJ et al. (2006) RNA editing of human microRNAs. Genome Biol, **7**:R27.

188. Yang W et al. (2006) Modulation of microRNA processing and expression through RNA editing by ADAR deaminases. Nat Struct Mol Biol, **13**:13-21.

189. Kawahara Y et al. (2007) RNA editing of the microRNA-151 precursor blocks cleavage by the Dicer-TRBP complex. EMBO Rep, **8**:763-9.

190. Kawahara Y et al. (2007) Redirection of silencing targets by adenosine-to-inosine editing of miRNAs. Science, **315**:1137-40.

191. Khvorova A et al. (2003) Functional siRNAs and miRNAs exhibit strand bias. Cell, **115**:209-16.

192. Schwarz DS et al. (2003) Asymmetry in the assembly of the RNAi enzyme complex. Cell, **115**:199-208.

193. Iizasa H et al. (2010) Editing of Epstein-Barr virus-encoded BART6 microRNAs controls their dicer targeting and consequently affects viral latency. J Biol Chem, **285**:33358-70.

194. Tonkin LA et al. (2002) RNA editing by ADARs is important for normal behavior in Caenorhabditis elegans. EMBO J, **21**:6025-35.

195. Palladino MJ et al. (2000) A-to-I pre-mRNA editing in Drosophila is primarily involved in adult nervous system function and integrity. Cell, **102**:437-49.

196. Keegan LP et al. (2011) Functional conservation in human and Drosophila of Metazoan ADAR2 involved in RNA editing: loss of ADAR1 in insects. Nucleic Acids Res, **39**:7249-62.

197. Horsch M et al. (2011) Requirement of the RNA-editing enzyme ADAR2 for normal physiology in mice. J Biol Chem, **286**:18614-22.

198. Hartner JC et al. (2004) Liver disintegration in the mouse embryo caused by deficiency in the RNA-editing enzyme ADAR1. J Biol Chem, **279**:4894-902.

199. Wang Q et al. (2000) Requirement of the RNA editing deaminase ADAR1 gene for embryonic erythropoiesis. Science, **290**:1765-8.

200. Wang Q et al. (2004) Stress-induced apoptosis associated with null mutation of ADAR1 RNA editing deaminase gene. J Biol Chem, **279**:4952-61.

201. XuFeng R et al. (2009) ADAR1 is required for hematopoietic progenitor cell survival via RNA editing. Proc Natl Acad Sci U S A, **106**:17763-8.

202. Hartner JC et al. (2009) ADAR1 is essential for the maintenance of hematopoiesis and suppression of interferon signaling. Nat Immunol, **10**:109-15.

203. Sebastiani P et al. (2009) RNA editing genes associated with extreme old age in humans and with lifespan in C. elegans. PLoS One, **4**:e8210.

204. Gallo A et al. (2012) ADARs: allies or enemies? The importance of A-to-I RNA editing in human disease: from cancer to HIV-1. Biol Rev Camb Philos Soc, **87**:95-110.

205. Galeano F et al. (2011) A-to-I RNA editing: The "ADAR" side of human cancer. Semin Cell Dev Biol.

206. Miyamura Y et al. (2003) Mutations of the RNA-specific adenosine deaminase gene (DSRAD) are involved in dyschromatosis symmetrica hereditaria. Am J Hum Genet, **73**:693-9.

207. Kawahara Y et al. (2004) Glutamate receptors: RNA editing and death of motor neurons. Nature, **427**:801.

208. Aizawa H et al. (2010) TDP-43 pathology in sporadic ALS occurs in motor neurons lacking the RNA editing enzyme ADAR2. Acta Neuropathol, **120**:75-84.

209. Maas S et al. (2001) Underediting of glutamate receptor GluR-B mRNA in malignant gliomas. Proc Natl Acad Sci U S A, **98**:14687-92.

210. Ishiuchi S et al. (2002) Blockage of Ca(2+)-permeable AMPA receptors suppresses migration and induces apoptosis in human glioblastoma cells. Nat Med, **8**:971-8.

211. Maas S et al. (2006) A-to-I RNA editing and human disease. RNA Biol, **3**:1-9.

212. Niswender CM et al. (2001) RNA editing of the human serotonin 5-HT2C receptor. alterations in suicide and implications for serotonergic pharmacotherapy. Neuropsychopharmacology, **24**:478-91.

213. Gurevich I et al. (2002) Altered editing of serotonin 2C receptor pre-mRNA in the prefrontal cortex of depressed suicide victims. Neuron, **34**:349-56.

214. Iwamoto K et al. (2003) RNA editing of serotonin 2C receptor in human postmortem brains of major mental disorders. Neurosci Lett, **346**:169-72.

215. Garrett S et al. (2012) RNA Editing Underlies Temperature Adaptation in K+ Channels from Polar Octopuses. Science.

216. Waterston RH et al. (2002) Initial sequencing and comparative analysis of the mouse genome. Nature, **420**:520-62.

217. Ohlson J et al. (2005) A method to find tissue-specific novel sites of selective adenosine deamination. Nucleic Acids Res, **33**:e167.

218. Ohlson J et al. (2007) A method for finding sites of selective adenosine deamination. Methods Enzymol, **424**:289-300.

219. Sakurai M et al. (2010) Inosine cyanoethylation identifies A-to-I RNA editing sites in the human transcriptome. Nat Chem Biol, **6**:733-40.

220. Sakurai M et al. (2011) Biochemical identification of A-to-I RNA editing sites by the inosine chemical erasing (ICE) method. Methods Mol Biol, **718**:89-99.

221. Morse DP (2004) Identification of substrates for adenosine deaminases that act on RNA. Methods Mol Biol, **265**:199-218.

222. Li JB et al. (2009) Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. Science, **324**:1210-3.

223. Mattick JS (2007) A new paradigm for developmental biology. J Exp Biol, **210**:1526-47.

224. Mattick JS et al. (2008) RNA editing, DNA recoding and the evolution of human cognition. Trends Neurosci, **31**:227-33.

225. Mehler MF et al. (2007) Noncoding RNAs and RNA editing in brain development, functional diversification, and neurological disease. Physiol Rev, **87**:799-823.

226. Baillie JK et al. (2011) Somatic retrotransposition alters the genetic landscape of the human brain. Nature, **479**:534-7.

227. Liu WM et al. (1994) Alu transcripts: cytoplasmic localisation and regulation by DNA methylation. Nucleic Acids Res, **22**:1087-95.

228. Chang DY et al. (1993) A cellular protein binds B1 and Alu small cytoplasmic RNAs in vitro. J Biol Chem, **268**:6423-8.

229. Bach D et al. (2008) Characterization of APOBEC3G binding to 7SL RNA. Retrovirology, **5**:54.

230. Jurka J (2000) Repbase update: a database and an electronic journal of repetitive elements. Trends Genet, **16**:418-20.

231. Hsu F et al. (2006) The UCSC Known Genes. Bioinformatics, **22**:1036-46.

232. Marques-Bonet T et al. (2009) Sequencing primate genomes: what have we learned? Annu Rev Genomics Hum Genet, **10**:355-86.

233. Umylny B et al. (2007) Most human Alu and murine B1 repeats are unique. J Cell Biochem, **102**:110-21.

234. Canella D et al. (2010) Defining the RNA polymerase III transcriptome: Genome-wide localization of the RNA polymerase III transcription machinery in human cells. Genome Res, **20**:710-21.

235. Takeda J et al. (2009) H-DBAS: human-transcriptome database for alternative splicing: update 2010. Nucleic Acids Res, **38**:D86-90.

236. Mondal T et al. (2010) Characterization of the RNA content of chromatin. Genome Res, **20**:899-907.

237. Grimaldi G et al. (1984) Defining the beginning and end of KpnI family segments. EMBO J, **3**:1753-9.

238. Brouha B et al. (2003) Hot L1s account for the bulk of retrotransposition in the human population. Proc Natl Acad Sci U S A, **100**:5280-5.

239. Smit AFA et al. (1996-2010) RepeatMasker Open-3.0 http://www.repeatmasker.org.

240. Labuda D et al. (1991) Evolution of mouse B1 repeats: 7SL RNA folding pattern conserved. J Mol Evol, **32**:405-14.

241. Allen TA et al. (2004) The SINE-encoded mouse B2 RNA represses mRNA transcription in response to heat shock. Nat Struct Mol Biol, **11**:816-21.

242. Fornace AJ, Jr. et al. (1986) Induction of B2 RNA polymerase III transcription by heat shock: enrichment for heat shock induced sequences in rodent cells by hybridization subtraction. Nucleic Acids Res, **14**:5793-811.

243. Langmead B et al. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol, **10**:R25.

244. Luthe DS (1983) A simple technique for the preparation and storage of sucrose gradients. Anal Biochem, **135**:230-2.

245. Trapnell C et al. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol, **28**:511-5.

246. Larkin MA et al. (2007) Clustal W and Clustal X version 2.0. Bioinformatics, **23**:2947-8.

247. Strauss WM (2001) Preparation of genomic DNA from mammalian tissue. Curr Protoc Mol Biol, **Chapter 2**:Unit2 2.

248. Brown T (2001) Southern blotting. Curr Protoc Mol Biol, **Chapter 2**:Unit2 9A.

249. Trapnell C et al. (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics, **25**:1105-11.

250. Li H et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics, **25**:2078-9.

251. Kent WJ et al. (2002) The human genome browser at UCSC. Genome Res, **12**:996-1006.

252. Kent WJ (2002) BLAT--the BLAST-like alignment tool. Genome Res, **12**:656-64.

253. Boyer LA et al. (2006) Polycomb complexes repress developmental regulators in murine embryonic stem cells. Nature, **441**:349-53.

254. Niranjanakumari S et al. (2002) Reversible cross-linking combined with immunoprecipitation to study RNA-protein interactions in vivo. Methods, **26**:182-90.

255. Greenberg ME et al. (2007) Identification of newly transcribed RNA. Curr Protoc Mol Biol, **Chapter 4**:Unit 4 10.

256. Morse DP et al. (1997) Detection of inosine in messenger RNA by inosine-specific cleavage. Biochemistry, **36**:8429-34.

257. Holmberg A et al. (2005) The biotin-streptavidin interaction can be reversibly broken using water at elevated temperatures. Electrophoresis, **26**:501-10.

258. Deng G et al. (1983) Terminal transferase: use of the tailing of DNA and for in vitro mutagenesis. Methods Enzymol, **100**:96-116.

259. Cao GJ et al. (1992) Identification of the gene for an Escherichia coli poly(A) polymerase. Proc Natl Acad Sci U S A, **89**:10380-4.

260. Rissland OS et al. (2007) Efficient RNA polyuridylation by noncanonical poly(A) polymerases. Mol Cell Biol, **27**:3612-24.

261. Martin G et al. (1998) Tailing and 3'-end labeling of RNA with yeast poly(A) polymerase and various nucleotides. RNA, **4**:226-30.

262. Yue D et al. (2008) Template-independent DNA polymerases. Curr Protoc Mol Biol, **Chapter 3**:Unit3 6.

263. Bayer EA et al. (1988) Biocytin hydrazide--a selective label for sialic acids, galactose, and other sugars in glycoconjugates using avidin-biotin technology. Anal Biochem, **170**:271-81.

264. Schroeder A et al. (2006) The RIN: an RNA integrity number for assigning integrity values to RNA measurements. BMC Mol Biol, **7**:3.

265. Paul MS et al. (1998) Inosine exists in mRNA at tissue-specific levels and is most abundant in brain mRNA. EMBO J, **17**:1120-7.

266. Kiran A et al. (2010) DARNED: a DAtabase of RNa EDiting in humans. Bioinformatics, **26**:1772-6.

267. Ruffalo M et al. (2011) Comparative analysis of algorithms for next-generation sequencing read alignment. Bioinformatics, **27**:2790-6.

268. Li H et al. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, **25**:1754-60.

269. Crooks GE et al. (2004) WebLogo: a sequence logo generator. Genome Res, **14**:1188-90.

270. Kohler M et al. (1993) Determinants of Ca2+ permeability in both TM1 and TM2 of high affinity kainate receptor channels: diversity by RNA editing. Neuron, **10**:491-500.

271. Levanon EY et al. (2005) Evolutionarily conserved human targets of adenosine to inosine RNA editing. Nucleic Acids Res, **33**:1162-8.

272. Osenberg S et al. (2009) Widespread cleavage of A-to-I hyperediting substrates. RNA, **15**:1632-9.

273. Piekna-Przybylska D et al. (2008) The 3D rRNA modification maps database: with interactive tools for ribosome analysis. Nucleic Acids Res, **36**:D178-83.

274. Ni J et al. (1997) Small nucleolar RNAs direct site-specific synthesis of pseudouridine in ribosomal RNA. Cell, **89**:565-73.

275. Omer AD et al. (2002) In vitro reconstitution and activity of a C/D box methylation guide ribonucleoprotein complex. Proc Natl Acad Sci U S A, **99**:5289-94.

276. Ellis JC et al. (2010) The small nucleolar ribonucleoprotein (snoRNP) database. RNA, **16**:664-6.

277. Lestrade L et al. (2006) snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. Nucleic Acids Res, **34**:D158-62.

278. Madhani HD et al. (1994) Randomization-selection analysis of snRNAs in vivo: evidence for a tertiary interaction in the spliceosome. Genes Dev, **8**:1071-86.

279. Newby MI et al. (2001) A conserved pseudouridine modification in eukaryotic U2 snRNA induces a change in branch-site architecture. RNA, **7**:833-45.

280. Gardner PP et al. (2011) Rfam: Wikipedia, clans and the "decimal" release. Nucleic Acids Res, **39**:D141-5.

281. Zwieb C (1997) The uRNA database. Nucleic Acids Res, **25**:102-3.

282. Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symposium Series, **41**:95-98.

283. Mardis ER (2008) The impact of next-generation sequencing technology on genetics. Trends Genet, **24**:133-41.

284. Sela N et al. (2007) Comparative analysis of transposed element insertion within human and mouse genomes reveals Alu's unique role in shaping the human transcriptome. Genome Biol, **8**:R127.

285. Eisenberg E et al. (2005) Is abundant A-to-I RNA editing primate-specific? Trends Genet, **21**:77-81.

286. Yurov YB et al. (2007) Aneuploidy and confined chromosomal mosaicism in the developing human brain. PLoS One, **2**:e558.

287. Westra JW et al. (2008) Aneuploid mosaicism in the developing and adult cerebellar cortex. J Comp Neurol, **507**:1944-51.

288. Coufal NG et al. (2009) L1 retrotransposition in human neural progenitor cells. Nature, **460**:1127-31.

289. Muotri AR et al. (2005) Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. Nature, **435**:903-10.

Appendixes

Appendix 1: Loading controls of the Northern blots

Appendix 2: Alu database

Appendix 3: Poster presented at the GRC

Appendix 4: List of PCR primers targeting mouse genes used in the glyoxal protocol

Appendix 5: Detailed protocol for the extraction of RNA containing inosine

Appendix 6: Editing sites in mouse brain

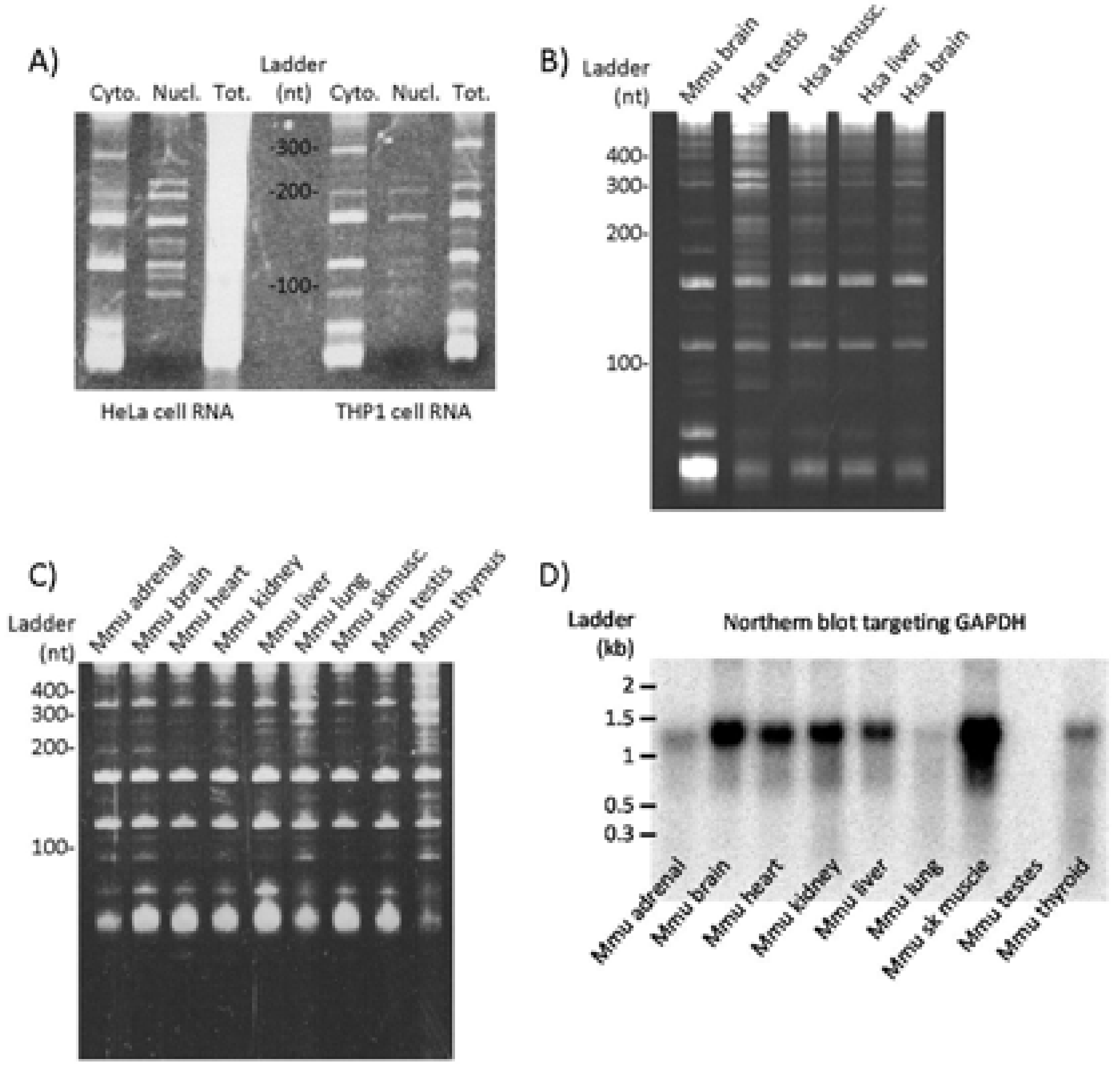
Appendix 7: Editing sites and other modifications on rRNA

Appendix 8: Editing sites on snoRNA

Appendix 9: Editing sites on snRNA

Appendix 1: Loading controls of the Northern blots

To assess the integrity and amount of RNA used for the Northern-blots, the profile of small RNA were observed for Northern blots on the human and mouse SINEs (A, B and C) and the profile of GAPDH was observed for the mouse LINE1 Northern-blot (D). For the small RNAs (A, B and C)The bands remain really sharp for the small RNAs and the intensity of the bands are of similar intensity, meaning that the RNAs were loaded in similar quantity and were not degraded during the migration in denaturating condition. One exception is observed in (A) where the total HeLa cell RNA signal seems saturated, which may be due to RNA degradation or overloading of the gel. For the mouse long RNAs in (D), the difference observed in the intensity of the bands means either that the RNAs were not loaded in equal quantity or that the transfer on membrane was not homogeneous. The bands remain however sharp, so no RNA degradation occurred before or during the migration.



Appendix 2: Alu database

The database of Alu elements generated in this study is a table summarizing the information in 42 columns for each of the 1,194,735 Alu elements of the human genome (hg19). In each row, the data are separated by a single tabulation and the first row of the table is the label of each column. It is advised to read the table in a command line interface. The table compressed with 7-zip is on the attached DVD under the following name: Alu_database.7z.

Description of the table:

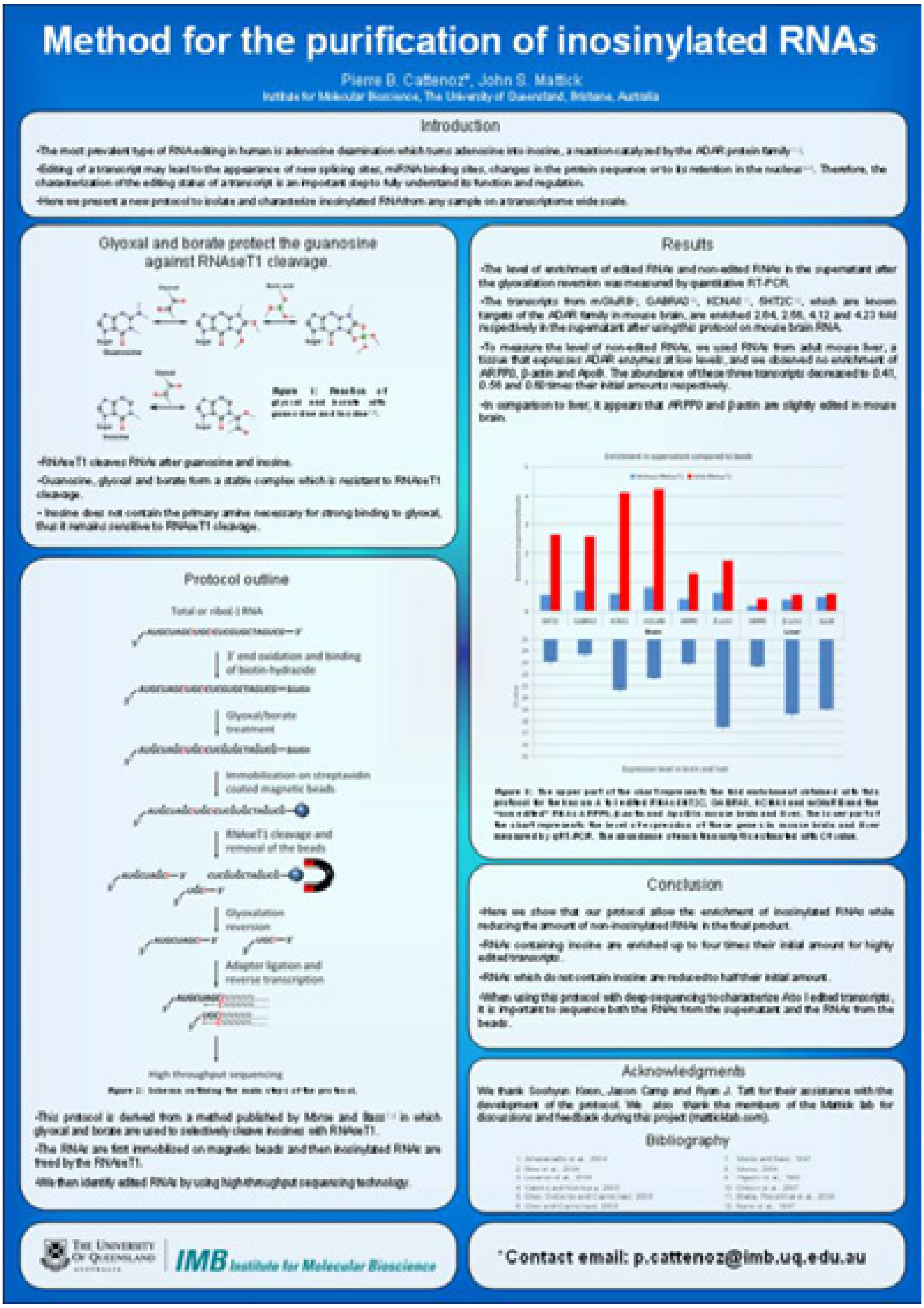
Col. #	Col. name	Description	GEO accession number or link to the deep-sequencing data
1	Alu_ID	Unique identifier.	
2	Family	Family of the Alu elements.	
3	CHR	Chromosome name.	
4	CHR_BEG	Position of the left most nucleotide on the chromosome.	
5	CHR_END	Position of the right most nucleotide on the chromosome.	
6	Strand	Strand on which is located the Alu element.	
7	Length_(nt)	Length of the element in nucleotide.	
8	Nb_of_occurence	Number of time the element's sequence exist in the genome.	
9	POLIII_A-box	Presence of the A-box region of the POLIII promoter.	
10	POLIII_B-box	Presence of the B-box region of the POLIII promoter.	
11	Genomic_context	Location of the Alu element in promoter, 5'UTR, exon, intron, 3'UTR or intergenic.	
12	Age	Time of apparition of the element in Million years ago (Mya).	
13	Alu_ID	Unique identifier.	
14	THP1_smallRNA	Coverage of the element by tags from deep sequencing of RNA from THP1 of 50nt to 350nt long (in read per kilobase per million of reads: RPKM)	In house data
15	CHIP_BDP1	Coverage of the element by tags from CHIPseq done on the POLIII complex subunit BDP1 in IMR90hTert cells (in RPKM).	GSM454596:
16	CHIP_BRF1	Coverage of the element by tags from CHIPseq done on the POLIII complex subunit BRF1 in IMR90hTert cells (in RPKM).	GSM454597
17	CHIP_POLIII	Coverage of the element by tags from CHIPseq done on the POLIII complex subunit POLIII (In read per kilobase per million of reads RPKM) in IMR90hTert cells (in RPKM).	GSM454595

Col. #	Col. name	Description	GEO accession number or link to the deep-sequencing data
18	Cytoplasmic	Coverage of the element by tags from cytoplasmic RNA sequencing in DLD1 cells (in RPKM).	http://www.h-invitational.jp/download/h-dbas/cyt.fa.tar.gz
19	Nuclear	Coverage of the element by tags from nuclear RNA sequencing in DLD1 cells (in RPKM).	http://www.h-invitational.jp/download/h-dbas/nuc.fa.tar.gz
20	Polysomal	Coverage of the element by tags from polysomal RNA sequencing in DLD1 cells (in RPKM).	http://www.h-invitational.jp/download/h-dbas/pol.fa.tar.gz
21	CAR_HeLa	Coverage of the element by tags from the sequencing of RNA associated with chromatin in HeLa cells (in RPKM).	In house data
22	CAR_fibroblast	Coverage of the element by tags from the sequencing of RNA associated with chromatin in fibroblasts (in RPKM).	GSM530678
23	Alu_ID	Unique identifier.	
24	100nt	Coverage of the element by tags from the sequencing of a pool of 16 human tissues with a protocol returning 100nt long read (in RPKM).	GSM759522-37
25	Adipose	Coverage of the element by tags from the sequencing of RNA from human adipose with a protocol returning 75nt long read (in RPKM).	GSM759490
26	Adrenal	Coverage of the element by tags from the sequencing of RNA from human adrenal with a protocol returning 75nt long read (in RPKM).	GSM759492
27	Brain	Coverage of the element by tags from the sequencing of RNA from human brain with a protocol returning 75nt long read (in RPKM).	GSM759494
28	Breast	Coverage of the element by tags from the sequencing of RNA from human breast with a protocol returning 75nt long read (in RPKM).	GSM759496
29	Colon	Coverage of the element by tags from the sequencing of RNA from human colon with a protocol returning 75nt long read (in RPKM).	GSM759498
30	Heart	Coverage of the element by tags from the sequencing of RNA from human heart with a protocol returning 75nt long read (in RPKM).	GSM759500
31	Kidney	Coverage of the element by tags from the sequencing of RNA from human kidney with a protocol returning 75nt long read (in RPKM).	GSM759502
32	Liver	Coverage of the element by tags from the sequencing of RNA from human liver with a protocol returning 75nt long read (in RPKM).	GSM759504

Col. #	Col. name	Description	GEO accession number or link to the deep-sequencing data
33	Lung	Coverage of the element by tags from the sequencing of RNA from human lung with a protocol returning 75nt long read (in RPKM).	GSM759506
34	Lymph	Coverage of the element by tags from the sequencing of RNA from human lymph node with a protocol returning 75nt long read (in RPKM).	GSM759508
35	Ovary	Coverage of the element by tags from the sequencing of RNA from human ovary with a protocol returning 75nt long read (in RPKM).	GSM759510
36	Prostate	Coverage of the element by tags from the sequencing of RNA from human prostate with a protocol returning 75nt long read (in RPKM).	GSM759512
37	Skelmusc	Coverage of the element by tags from the sequencing of RNA from human skeletal muscle with a protocol returning 75nt long read (in RPKM).	GSM759514
38	Testes	Coverage of the element by tags from the sequencing of RNA from human testes with a protocol returning 75nt long read (in RPKM).	GSM759516
39	Thyroid	Coverage of the element by tags from the sequencing of RNA from human thyroid with a protocol returning 75nt long read (in RPKM).	GSM759518
40	Whiteblood	Coverage of the element by tags from the sequencing of RNA from human whiteblood cells with a protocol returning 75nt long read (in RPKM).	GSM759520
41	Alu_ID	Unique identifier.	
42	Sequence	Sequence of the element.	

Appendix 3: Poster presented at the GRC

The work on the development of the protocol to identify edited RNA with glyoxal and RNAseT1 was presented with the following poster at the Gordon Research Conference and at the Gordon Research Seminar on RNA editing in Galveston (Texas) in January 2011.



Method for the purification of inosinylated RNAs

Pierre B. Cattenoz*, John S. Mattick
Institute for Molecular Bioscience, The University of Queensland, Brisbane, Australia

Introduction

-The most prevalent type of RNA editing in human is adenosine deamination which turns adenosine into inosine, a reaction catalyzed by the ADAR protein family[1-3].

-Editing of a transcript may lead to the appearance of new splicing sites, miRNA binding sites, changes in the protein sequence or to its retention in the nucleus[4-6]. Therefore, the characterization of the editing status of a transcript is an important step to fully understand its function and regulation.

-Here we present a new protocol to isolate and characterize inosinylated RNA from any sample on a transcriptome wide scale.

Glyoxal and borate protect the guanosine against RNAseT1 cleavage.

Figure 1: Reaction of glyoxal and borate with guanosine and inosine[7].

-RNAseT1 cleaves RNAs after guanosine and inosine.

-Guanosine, glyoxal and borate form a stable complex which is resistant to RNAseT1 cleavage.

- Inosine does not contain the primary amine necessary for strong binding to glyoxal, thus it remains sensitive to RNAseT1 cleavage.

Protocol outline

Total or ribo(-) RNA

3' end oxidation and binding of biotin-hydrazide

Glyoxal/borate treatment

Immobilization on streptavidin coated magnetic beads

RNAseT1 cleavage and removal of the beads

Glyoxalation reversion

Adapter ligation and reverse transcription

High throughput sequencing

Figure 2: Scheme outlining the main steps of the protocol.

-This protocol is derived from a method published by Morse and Bass[7,8] in which glyoxal and borate are used to selectively cleave inosines with RNAseT1.

-The RNAs are first immobilized on magnetic beads and then inosinylated RNAs are freed by the RNAseT1.

-We then identify edited RNAs by using high throughput sequencing technology.

Results

-The level of enrichment of edited RNAs and non-edited RNAs in the supernatant after the glyoxalation reversion was measured by quantitative RT-PCR.

-The transcripts from mGluR-B, GABRA3[9], KCNA1[10], 5HT2C[11], which are known targets of the ADAR family in mouse brain, are enriched 2.04, 2.56, 4.12 and 4.23 fold respectively in the supernatant after using this protocol on mouse brain RNA.

-To measure the level of non-edited RNAs, we used RNAs from adult mouse liver, a tissue that expresses ADAR enzymes at low levels, and we observed no enrichment of ARPP0, β-actin and ApoB. The abundance of these three transcripts decreased to 0.41, 0.56 and 0.50 times their initial amounts respectively.

-In comparison to liver, it appears that ARPP0 and β-actin are slightly edited in mouse brain.

Enrichment in supernatant compared to beads
Without RNAseT1
With RNAseT1
Brain
Liver
Expression level in brain and liver

Figure 3: The upper part of the chart represents the fold enrichment obtained with this protocol for the known A to I edited RNAs (5HT2C, GABRA3, KCNA1 and mGluR-B) and the "non edited" RNAs (ARPP0, β-actin and ApoB) in mouse brain and liver. The lower part of the chart represents the level of expression of these genes in mouse brain and liver measured by qRT-PCR. The abundance of each transcript is reflected with Ct value.

Conclusion

-Here we show that our protocol allow the enrichment of inosinylated RNAs while reducing the amount of non-inosinylated RNAs in the final product.

-RNAs containing inosine are enriched up to four times their initial amount for highly edited transcripts.

-RNAs which do not contain inosine are reduced to half their initial amount.

-When using this protocol with deep-sequencing to characterize A to I edited transcripts, it is important to sequence both the RNAs from the supernatant and the RNAs from the beads.

Acknowledgments

We thank Soohyun Kwon, Jason Carey and Ryan J. Taft for their assistance with the development of the protocol. We also thank the members of the Mattick lab for discussions and feedback during this project (matticklab.com).

Bibliography

1. Athanasiadis et al., 2004
2. [illegible] et al., 2004
3. Levanon et al., 2004
4. Valente and Nishikura, 2005
5. Chen, DeCerbo and Carmichael, 2008
6. Chen and Carmichael, 2009
7. Morse and Bass, 1997
8. Morse, 2004
9. Higuchi et al., 1993
10. Ohlson et al., 2007
11. [illegible] et al., 2000
12. Burns et al., 1997

The University of Queensland Australia | IMB Institute for Molecular Bioscience

*Contact email: p.cattenoz@imb.uq.edu.au

Appendix 4: List of PCR primers targeting mouse genes used in the glyoxal protocol

Target	Forward primer	Reverse primer	Length	Used in (*)
5HT2C	TGCCCCTGTCTCTGCTTGCAA	ACCGGTCCAGCGATATGGCG	138	Glyo
ATP5E	CCGGTTTGAGGCTACTCTGA	AGATCTGGGAAAACCGGATG	117	Glyo
β-actin	CCACAGCTGAGAGGGAAATC	TCTCCAGGGAGGAAGAGGAT	108	Glyo
GABRA3	TGTCCCTGCCCGCACAGTCT	TCCATGGCCGTCGCGTATGC	108	Glyo
GAPDH	AACTTTGGCATTGTGGAAGG	GGATGCAGGGATGATGTTCT	132	Glyo
GRIA2	TCGAAGCCATTCATGAGCCTTGGA	AGTGTGCCACTCGTAGGGGCTA	186	Glyo
KCNA1	AGAAGGGCGAGCAGGCCACT	TCAGCTTCTTCCGCCTCCGC	220	Glyo
PPIA	GCTGGACCAAACACAAACG	CGCAAGTCAAAAGAAATTAGAGC	205	Glyo
5HT2C	AGATATTTGTGCCCCGTCTG	CTTAGTCCGCGAATTGAACC	141	IP
ADAR2	TGTGGCTAAAGGAAGCTCGT	ATGTTGTCCAGATTGCGGTT	253	IP
GRIA2 Q/R	CAGCAGATTTAGCCCCTACG	AGCCGTGTAGGAGGAGATGA	226	IP & Glyo
GRIA2 R/G	ATTCCAAAGGCTACGGCATC	TTTTCAATTTGTCCAACAGGC	108	IP
mGluRA	CCAGAGCTGGTGCTGGTCAGCTCTCG	GAAGTATATACGACCACTGTCATC	203	IP
RPLP0	GCAGTGGAAGTCCAACTACTTC	TGAGGTCCTCCTTGGTGAACAC	266	IP & Glyo

*The primers used in the IP were designed to cover editing sites and are indicated by IP. The primers used in the glyoxal method were designed upstream of the editing sites and are indicated by Glyo.

Appendix 5: Detailed protocol for the extraction of RNA containing inosine

Oxidation of the 3'end of RNA

1. Mix 30ug of RNA (in 225µL) with 75µL of 0.5M Na-acetate and 75µL of 50mM sodium periodate (freshly prepared) in 375µL. (sodium periodate MM=213.89g/mol, solution of 10.695g/L for 50mM).

2. Incubate 90min at RT in the dark.

3. Stop the reaction with 113µL of 1M KCl and place on ice for 10min.

4. Adjust the volume to 750µL with 262µL of DEPC water.

5. Add 1vol. of phenol/chloroform/isoamyl-alcohol (25/24/1) and mix well.

6. Centrifuge 10min at 13000g at 4˚C.

7. Precipitate the RNA with 1/10 vol. of sodium acetate 3M and 0.7vol. of isopropanol. 1µL of glycoblue 10mg/mL may be added to the mix to be able to see the pellet more easily.

8. Incubate at RT for 5min and centrifuge at max speed (17krpm) for 15min at 4°C.

9. Remove the supernatant.

10. Wash the pellet 3 times in 70% EtOH. If the pellet is dispersed during the wash centrifuge at max speed for 2min.

11. Air dry the pellet for 5-10min.

Biotinylation of the RNA

12. Resuspend the pellet in 675µL of 100mM NaOAc pH5.

NB: the pellet may be really hard to resuspend, I have found easier to resuspend it first in a smaller volume of DEPC water, then adjust the concentration of sodium acetate to 100mM and then to adjust the volume to 675µL with 100mM NaOAc pH5.

13. Add 75µL of 40mM (~10mg/mL) biotin-hydrazide (Sigma B7639) dissolved in DMSO.

14. Incubate 4hrs at RT in the dark.

15. Phenol extract the RNA as described previously from steps 5 to 11 of this protocol.

16. Resuspend the pellet in 20µL of DEPC water.

17. Measure the concentration of RNA with the nanodrop.

NB: you can assess the Biotinylation of the RNA by dot-plot: plot 0.5 to 1µg of RNA on a Northern membrane, let the drop dry, and incubate the membrane in a solution of 1XPBS, 0.1%Tween20 and 10% milk at room temperature for 1hr on a shaker. Incubate the membrane in a solution of 2% milk in PBS 1X / Tween20 0.1% containing 1/2000 dilution of avidin-peroxidase. Wash the membrane 3 times 10min with PBS 1X / Tween20 0.1% and reveal the membrane with ECL reagent following the manufacturer instruction.

Magnetic beads treatment

First the beads (Dynabeads M-270 Invitrogen #653-05) need to be treated to be RNAse free, and then they are incubated with the RNAs. 100µL of beads can bind up to 2.5ug of total RNA. 50µL of beads are used for 5ug of biotinylated RNA.

18. Wash 50µL of beads (per sample) for 2min with 800µL of DEPC treated NaOH 0.1M and NaCl 0.05M.

19. Place the bead on the magnetic stand for 2min and remove the supernatant.

20. Repeat the NaOH treatment once.

21. Wash the beads twice with 0.1M DEPC treated NaCl in the same volume than in step 1.

22. Resuspend the beads in 25µL of Tris-borate buffer (10mM Tris-HCl pH 7.8, 1M sodium borate pH 7.5).

Inosine specific RNAseT1 treatment

23. Add the following reagents to a tube (can be scaled in separate tubes according to the amount of RNA used):

Tube	1
2.5ug biotin-RNA in water	34µL
Na-PO4 100mM pH7	10µL
DMSO	50µL
Glyoxal 40%	6µL

24. Incubate the tube for 45min at 37°C.

25. Add 100µL of 1M sodium borate pH7.5 to each sample.

26. Precipitate the sample with 1mL of EtOH 100%.

27. Centrifuge the tubes at maximum speed for 15min at 4°C, and dry the pellet very well (speedvac for 20min).

28. Resuspend the sample (2*2.5ug) in 25µL of Tris-borate buffer (10mM Tris-HCl pH 7.8, 1M sodium borate pH 7.5).

29. Pool the 25µL samples in the 25µL of beads and incubate the tube for 30min at RT with gentle mixing

NB: you can collect the supernatant to assess the diminution of biotinylated RNA in the sample.

30. Rinse the beads gently, twice with 1mL of Tris-borate buffer (10mM Tris-HCl pH 7.8, 1M sodium borate pH 7.5) and resuspend them in 47µL of Tris-borate buffer.

31. Add 3µL of RNAseT1 (Ambion #AM2280) to your sample.

32. Incubate the tubes at 37°C for 30min, resuspend the beads gently every 5min.

33. Collect the supernatant (this is the I-RNA), adjust the volume to 100uL with DEPC water.

34. Clean-up the I-RNA with MinElute cleanup-kit (Qiagen #74204): add 350µL of buffer RLT and 675µL 100%EtOH to 100µL sample. Add 700µL of the mix on a column

inserted in a collection tube, centrifuge 15s at >8000g, discard flow through, pipet the remaining sample on the column and centrifuge 15s at >8000g, discard flow through. Wash the column with 500µL buffer RPE and centrifuge 15s at >8000g, discard flow through, repeat the wash step with 500µL of 80% EtOH and centrifuge 15s at >8000g, discard flow through and collection tube, place column in a new collection tube and centrifuge at full speed for 1 min with the lid of the column open. Place the column in a new collection tube, add at least 14µL of DEPC-water directly on the column and centrifuge for 1min at full speed to collect the RNA. (The column has a retention volume of 2µL).

35. Resuspend the beads in 100µL of a solution of 95% formamide and 10mM EDTA (2µL EDTA 0.5M), and heat the sample 5min at 65C to dissociate the RNA from the beads (to be done under the fume hood, heating formamide make cyanide). Collect the beads eluate (this is the B-RNA) and extract as described step 34.

36. Adjust the volume of I-RNA and B-RNA to 100µL with DEPC water and repeat the column cleaning as described step 34 of this protocol (the columns can be reused given that they are cleaned with 500uL DEPC water right after first use) to get rid of the excess borate before removing the glyoxal.

37. Elute the I-RNA and B-RNA in 42µL of water (40 µL remain after elution), add 10µL of 100mM sodium phosphate pH7.0 and 50µL DMSO per tubes.

38. Incubate all samples for 3hrs at 65C.

39. Clean up the sample as described step 34 of this protocol, elute in 17µL DEPC water.

40. Measure the concentration of the samples with a nanodrop or bioanalyzer. If well resuspended, you'd expect around 400 to 600ng of B-RNA and around 50 to 80ng of I-RNA when you start with 25ug of biotinylated RNA.

Important if deep-sequencing these RNAs: the cleavage of the inosine by RNAse T1 leave a phosphorylated 3'end. It is important to treat the RNA from the supernatant with PNK (NEB #M0236S) in order to be able to ligate adapters for generating the deep-sequencing library. For deep-sequencing the RNA bond to the beads, those ones contain a biotin and are oxidized on their 3'end which completely prevent a consecutive ligation of adapters. To deep-sequence these beads RNA, I suggest to reverse transcribe them with random hexamer and deep-sequence the cDNA generated.

TAP and PNK treatment of the I-RNA

41. Mix 15uL of I-RNA with 5µL of 10X TAP buffer, 4µL of TAP (Epicentre #T19050) in a 50µL reaction.

42. Incubate 1hr at 37ºC and then place the sample at -20ºC to stop the reaction.

43. Clean up the sample as described step 34 of this protocol, elute in 17µL DEPC water.

44. Mix 15µL of TAP treated supernatant RNA or 15µL of beads RNA with 6µL of 5X T4 DNA ligase buffer (check the composition, it needs to contain ATP) (250mM Tris-HCl pH 7.6, 50mM $MgCl_2$, 5mM ATP, 5mM DTT, 25% w/v PEG8000) and 3µL of T4 polynucleotide kinase (NEB #M0201S), adjust the volume to 30µL with water.

45. Incubate 1hr at 37ºC.

46. Clean up the sample as described step 34 of this protocol, elute in 17µL DEPC water. The I-RNA is ready for deep sequencing.

Appendix 6: Editing sites in mouse brain

All the editing sites characterized by the glyoxal protocol were compiled in two excel tables (sheets Raw data and Significant data in the excel file Edited_position_in_mouse_brain.xls in the enclosed DVD) which present the same criteria for all the 12,644 A-to-G mismatches found in the 3'end of the I-RNA library in table and for the subset of 1,882 editing sites which remain after the filtering described in the material and method, a part from the last column which is present only in the significant data sheet.

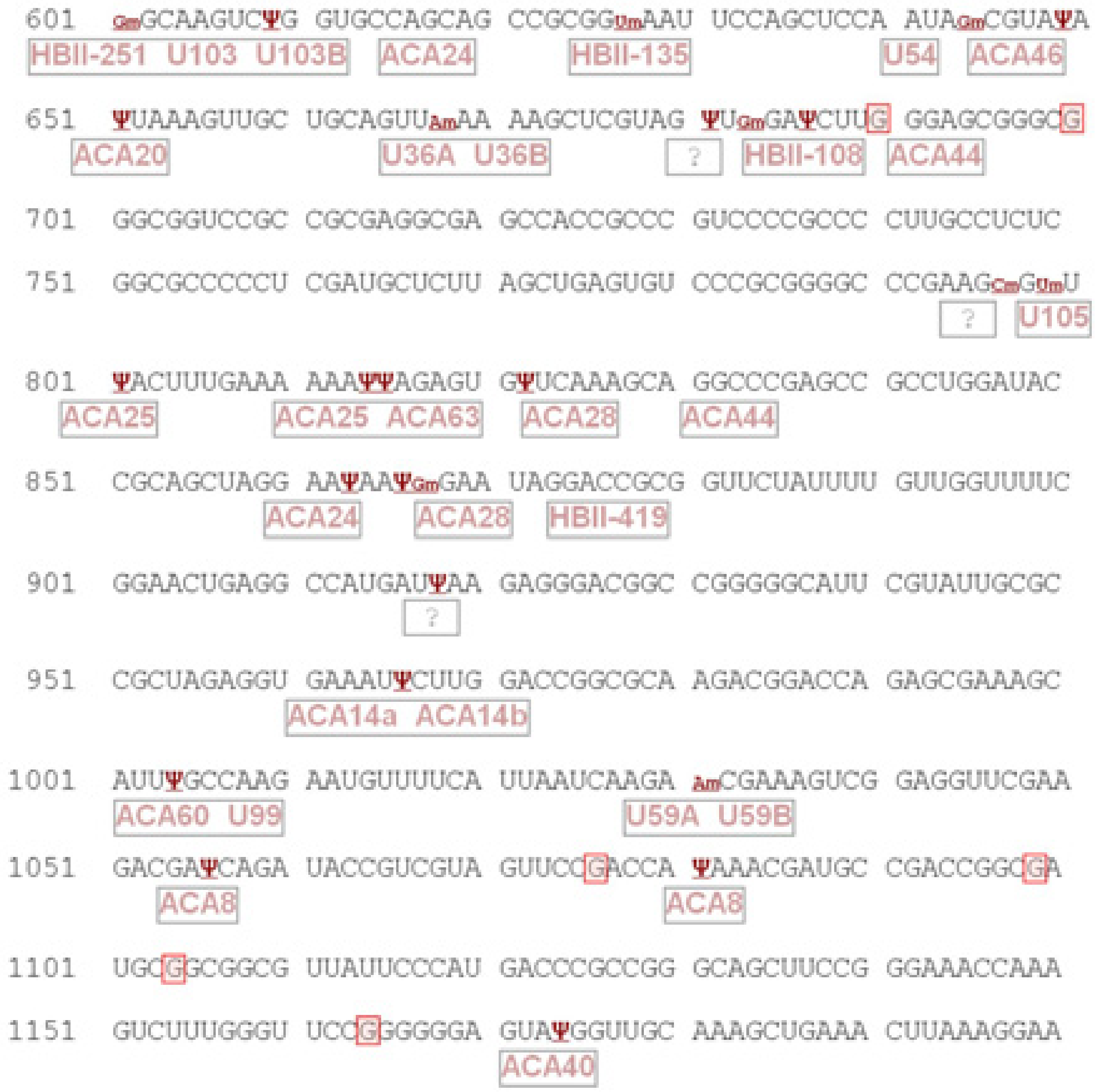
Description of the tables:

Col. #	Col. name	Description
1	Site	Coordinate of the locus in the format {chromosome:chromosome right coordinate – chromosome left coordinate}.
2	STD	Strand on which is present the site.
3	G in I-RNA	Number of G counted at this locus in the I-RNA library.
4	G in all	Number of G counted at this locus in both I-RNA and B-RNA libraries.
5	A in all	Number of A counted at this locus in both I-RNA and B-RNA libraries.
6	C and T in all	Number of C and T counted at this locus in both I-RNA and B-RNA libraries.
7	Tag #	Number of tags covering this locus.
8	A freq.	Frequency of A at this locus.
9	T freq.	Frequency of T at this locus.
10	G freq.	Frequency of G at this locus.
11	C freq.	Frequency of C at this locus.
12	p-value	P-value for the comparison of A frequency against C and T frequency: "^" if not significant, "*" if below 0.05, "**" if below 0.001, "***" if below 0.0001 and N/A if non measurable.

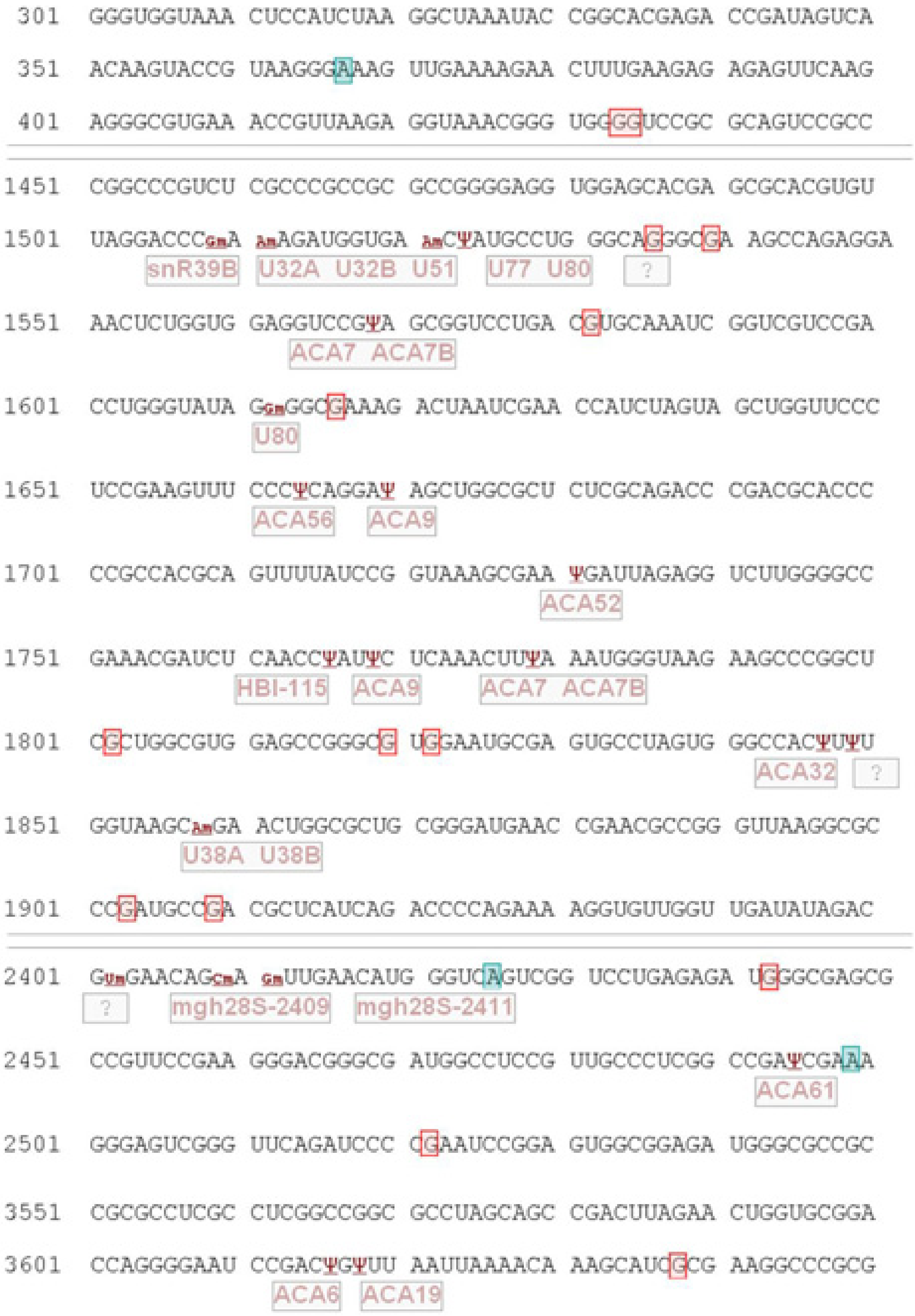
Appendix 7: Editing sites and other modifications on rRNA

The location of the A-to-I editing sites discovered in this thesis, presenting a editing frequency >0.5, and present in mouse rRNA were reported on the sequences of the human 28S and 18S rRNAs for which a detailed map of modification is available [273]. The sequences with methylation (**m**) and pseudouridine (ψ) mark and the snoRNA responsible for the modification were taken from [273], A-to-I editing sites newly discovered in orthologue sequences were implemented to the sequences in green when the modified adenosine was conserved from the identified locus to the human 28S rRNA and in red when the adenosine was replaced by a guanosine in the human 28S rRNA.

***Human* 18S rRNA (X03205)**


601 GmGCAAGUCΨG GUGCCAGCAG CCGCGGUmAAU UCCAGCUCCA AUAGmCGUAΨA
HBII-251 U103 U103B
ACA24
HBII-135
U54
ACA46
651 ΨUAAAGUUGC UGCAGUUAmAA AAGCUCGUAG ΨUGmGAΨCUUG GGAGCGGGCG
ACA20
U36A U36B
?
HBII-108
ACA44
701 GGCGGUCCGC CGCGAGGCGA GCCACCGCCC GUCCCCGCCC CUUGCCUCUC
751 GGCGCCCCCU CGAUGCUCUU AGCUGAGUGU CCCGCGGGGC CCGAAGCmGUmU
?
U105
801 ΨACUUUGAAA AAAΨΨAGAGU GΨUCAAAGCA GGCCCGAGCC GCCUGGAUAC
ACA25
ACA25 ACA63
ACA28
ACA44
851 CGCAGCUAGG AAΨAAΨGmGAA UAGGACCGCG GUUCUAUUUU GUUGGUUUUC
ACA24
ACA28
HBII-419
901 GGAACUGAGG CCAUGAUΨAA GAGGGACGGC CGGGGGCAUU CGUAUUGCGC
?
951 CGCUAGAGGU GAAAUΨCUUG GACCGGCGCA AGACGGACCA GAGCGAAAGC
ACA14a ACA14b
1001 AUUΨGCCAAG AAUGUUUUCA UUAAUCAAGA AmCGAAAGUCG GAGGUUCGAA
ACA60 U99
U59A U59B
1051 GACGAΨCAGA UACCGUCGUA GUUCCGACCA ΨAAACGAUGC CGACCGGCGA
ACA8
ACA8
1101 UGCGGCGGCG UUAUUCCCAU GACCCGCCGG GCAGCUUCCG GGAAACCAAA
1151 GUCUUUGGGU UCCGGGGGGA GUAΨGGUUGC AAAGCUGAAA CUUAAAGGAA
ACA40

Human 28S rRNA (U13369)

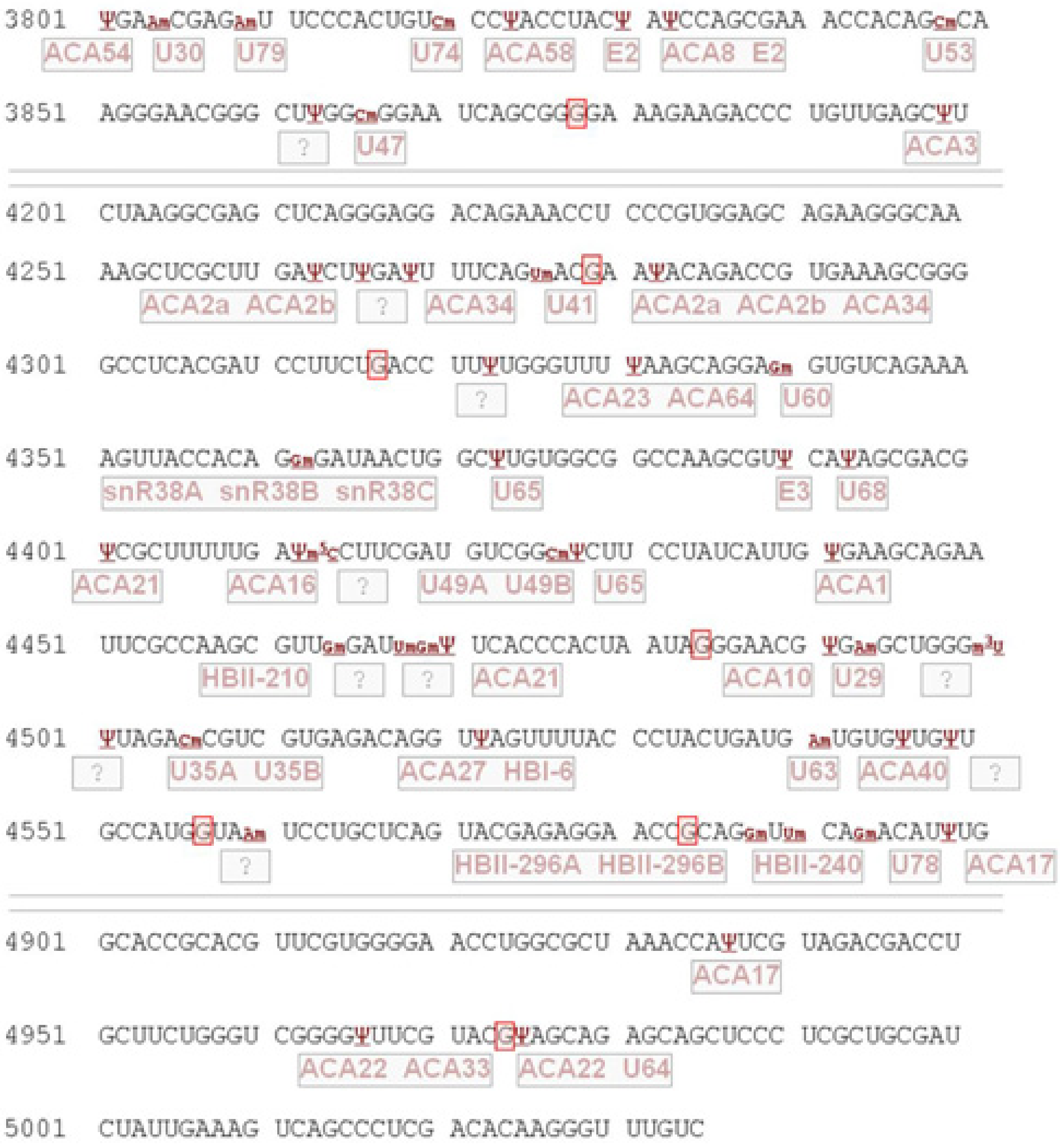


```
 301  GGGUGGUAAA CUCCAUCUAA GGCUAAAUAC CGGCACGAGA CCGAUAGUCA
 351  ACAAGUACCG UAAGGGAAAG UUGAAAAGAA CUUUGAAGAG AGAGUUCAAG
 401  AGGGCGUGAA ACCGUUAAGA GGUAAACGGG UGGGGUCCGC GCAGUCCGCC

1451  CGGCCCGUCU CGCCCGCCGC GCCGGGGAGG UGGAGCACGA GCGCACGUGU
1501  UAGGACCCGmA AmAGAUGGUGA AmCΨAUGCCUG GGCAGGGCGA AGCCAGAGGA
          snR39B  U32A U32B U51  U77 U80  ?
1551  AACUCUGGUG GAGGUCCGΨA GCGGUCCUGA CGUGCAAAUC GGUCGUCCGA
                ACA7 ACA7B
1601  CCUGGGUAUA GGmGGCGAAAG ACUAAUCGAA CCAUCUAGUA GCUGGUUCCC
              U80
1651  UCCGAAGUUU CCCΨCAGGAΨ AGCUGGCGCU CUCGCAGACC CGACGCACCC
              ACA56   ACA9
1701  CCGCCACGCA GUUUUAUCCG GUAAAGCGAA ΨGAUUAGAGG UCUUGGGGCC
                                     ACA52
1751  GAAACGAUCU CAACCΨAUΨC UCAAACUUΨA AAUGGGUAAG AAGCCCGGCU
             HBI-115  ACA9   ACA7 ACA7B
1801  CGCUGGCGUG GAGCCGGGCG UGGAAUGCGA GUGCCUAGUG GGCCACΨUΨU
                                                ACA32  ?
1851  GGUAAGCAmGA ACUGGCGCUG CGGGAUGAAC CGAACGCCGG GUUAAGGCGC
              U38A U38B
1901  CCGAUGCCGA CGCUCAUCAG ACCCCAGAAA AGGUGUUGGU UGAUAUAGAC

2401  GUmGAACAGCmA GmUUGAACAUG GGUCAGUCGG UCCUGAGAGA UGGGCGAGCG
      ?    mgh28S-2409  mgh28S-2411
2451  CCGUUCCGAA GGGACGGGCG AUGGCCUCCG UUGCCCUCGG CCGAΨCGAAA
                                                ACA61
2501  GGGAGUCGGG UUCAGAUCCC CGAAUCCGGA GUGGCGGAGA UGGGCGCCGC

3551  CGCGCCUCGC CUCGGCCGGC GCCUAGCAGC CGACUUAGAA CUGGUGCGGA
3601  CCAGGGGAAU CCGACΨGΨUU AAUUAAAACA AAGCAUCGCG AAGGCCCGCG
                ACA6 ACA19
```


3801 ΨGAAmCGAGAmU UCCCACUGUCm CCΨACCUACΨ AΨCCAGCGAA ACCACAGCmCA
ACA54 U30 U79 U74 ACA58 E2 ACA8 E2 U53
3851 AGGGAACGGG CUΨGGCmGGAA UCAGCGGGGA AAGAAGACCC UGUUGAGCΨU
? U47 ACA3
4201 CUAAGGCGAG CUCAGGGAGG ACAGAAACCU CCCGUGGAGC AGAAGGGCAA
4251 AAGCUCGCUU GAΨCUΨGAΨU UUCAGUmACGA AΨACAGACCG UGAAAGCGGG
ACA2a ACA2b ? ACA34 U41 ACA2a ACA2b ACA34
4301 GCCUCACGAU CCUUCUGACC UUΨUGGGUUU ΨAAGCAGGAGm GUGUCAGAAA
? ACA23 ACA64 U60
4351 AGUUACCACA GGmGAUAACUG GCΨUGUGGCG GCCAAGCGUΨ CAΨAGCGACG
snR38A snR38B snR38C U65 E3 U68
4401 ΨCGCUUUUUG AΨmm^{5}CCUUCGAU GUCGGCmΨCUU CCUAUCAUUG ΨGAAGCAGAA
ACA21 ACA16 ? U49A U49B U65 ACA1
4451 UUCGCCAAGC GUUGmGAUUmGmΨ UCACCCACUA AUAGGGAACG ΨGAmGCUGGGm^{3}U
HBII-210 ? ? ACA21 ACA10 U29 ?
4501 ΨUAGACmCGUC GUGAGACAGG UΨAGUUUUAC CCUACUGAUG AmUGUGΨUGΨU
? U35A U35B ACA27 HBI-6 U63 ACA40 ?
4551 GCCAUGGUAAm UCCUGCUCAG UACGAGAGGA ACCGCAGGmUUm CAGmACAUΨUG
? HBII-296A HBII-296B HBII-240 U78 ACA17
4901 GCACCGCACG UUCGUGGGGA ACCUGGCGCU AAACCAΨUCG UAGACGACCU
ACA17
4951 GCUUCUGGGU CGGGGΨUUCG UACGΨAGCAG AGCAGCUCCC UCGCUGCGAU
ACA22 ACA33 ACA22 U64
5001 CUAUUGAAAG UCAGCCCUCG ACACAAGGGU UUGUC

Appendix 8: Editing sites on snoRNA

Below are the sequences of the snoRNAs containing editing sites (**A**), the C/D box and ACA box are indicated in red and the site recognizing the targets are in blue. The sequences and target site are taken from the snoRNABase [277].

>snora28

acactctgtggcagatgaacaaaaccgtctgacacaatttgagcttgctatagcaagaaagtctaacctattccggtgttctctttccccatgagac**A**agccgttatat**A**ggctctaacaaa

>snord15a

cttcgatgaagaggtgatgacgagtctgagtaggaagtgttgtctttgtccaagatgcctcact**A**tgctgcgttctgtggcacagctga**AA**gcactgtggtcaaa**A**gaa**A**cttcctaaagatgaccaagaggcatttgtctgagaagg

>snord15b

cttcagtgatgacacgatgacgagtcagaatggccacgtcttgctcttggtccctgtcagtgccatgttctgtggtgctgtacatggttcccttggcaa**A**agtgtcctgcgcactgattgatttagaggcatttgtctgagaagg

>snord17

ctgtttatccattcgctgagtacgctgctctgaccttcttcccagtctcggttcctgttctggg**A**gcttggggctgagt**A**gccaccagccctgctctctgcagtgttctattgtggattgcttgtgtgctggcaggct**A**ctactggt**AA**g**AA**tggctagtgtcagcagggatggctcctctctgggttccatctcaccaagatgagtggtgcaaatctgat

>snord22

tccccatgaagaaatgttcacacgtcctacttcctgtcctagctccagagcctgaaaaggtgaaccc**A**ctggggctggctgggggaaaagaggaaactttgttccagaaggaactgtctgagggat

>snord45c

ggcatgtattctgaatctaaagttgattataaaccactttagctctaga**A**ttactctgaga

>snora22

ttgcacagtgaacacccaagtgtgctttatagttcccttggctttgaccctgtgctagagcattgcctgctcttctcctctgcatt**AA**aaggaatatttatcctttaaatgtattcagaaagccagcacatta

>snora26

gtgcccttttaaggttgacccagtgctttaagaggctaacacagaagggtaaagtaagtctccataaaacccagaga**A**gagactggaa
agctcctctttggatcctgtctggagtcacaact

>snora53

aacatgcttccttagatccacctttgtggatgaatcttgaactgagttccacttgtaaacttcttgtttcttgtggttcc**A**gtagtcaaagaaac
atccagcaacttttttggttgtatagtcaaaggtgcttgagtcattggcatgtaagagaaatatacctgcatgttagtctaacgttctgataga
aatgacatgcatttatgctgccatttgttactatcaggactcgactcgtgtgcggacattt

>snord10

gctctgtgatggagcccatgcgtgtcatctgagcctctggcttccctgccagtgcagccctggcagtgtcctacttcccagggctgttgtc
tgcctggcggga**A**ggtcctgggcaaaggatcagtctttgtactctgagagcagacta

>snord116

taggttgatgatgacttacatatatacgttttttttttttttttggaaaggtgaacaaaatgagtgaaaactcagtaccatcatcctcatctaactg
Aggtcca

>snord5

gttcagatgatgaatttaactgttcaactgctgaatgataacgggcatgaactaaaacttaattctg**A**cagag

>snord88b

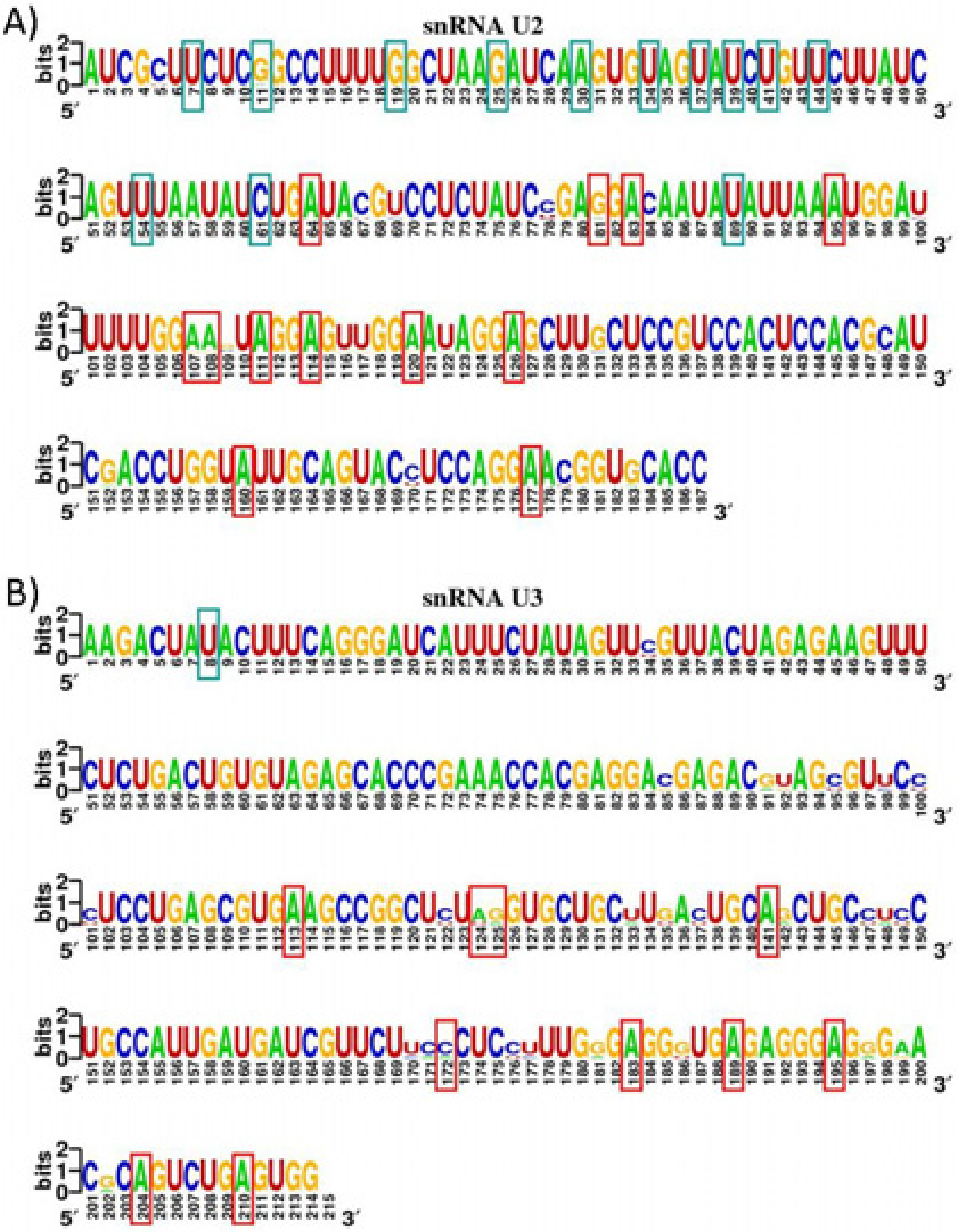
ttggggaccccgtgatgtccagcactgggctctgactgcccctgaggacacggtgcaccccgggacctttgacatccgg**A**gttctga
ggggccccac

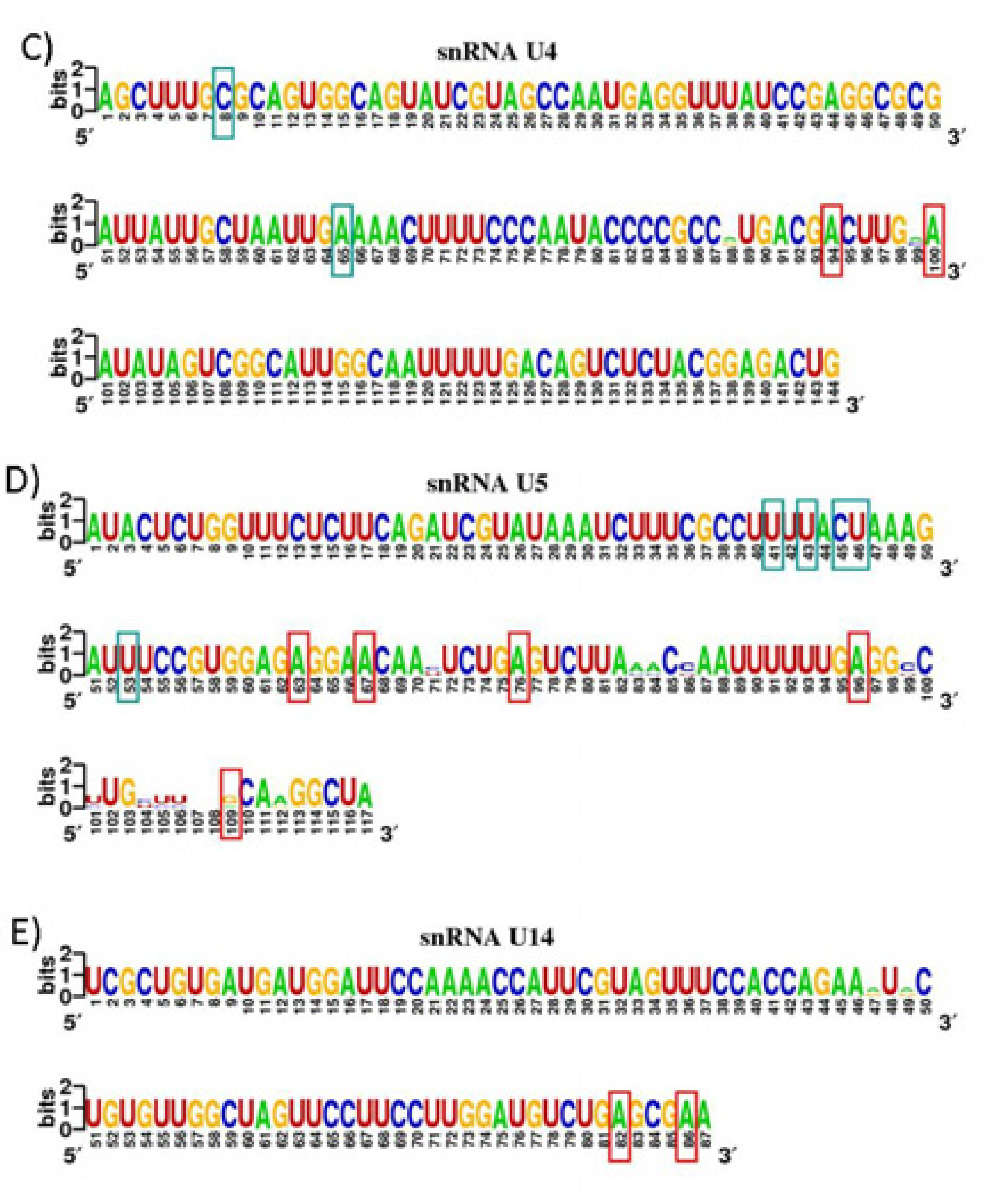
>snord94

caggctgtgatgattggcgcaggggtacggacctcagctgagtcatgggagctgaatgtatgtgtttctcctttgtcctgcatgtggc**A**g
gctgatggggagcacttacatgagactgttgcctcaatctgagcctg

Appendix 9: Editing sites on snRNA

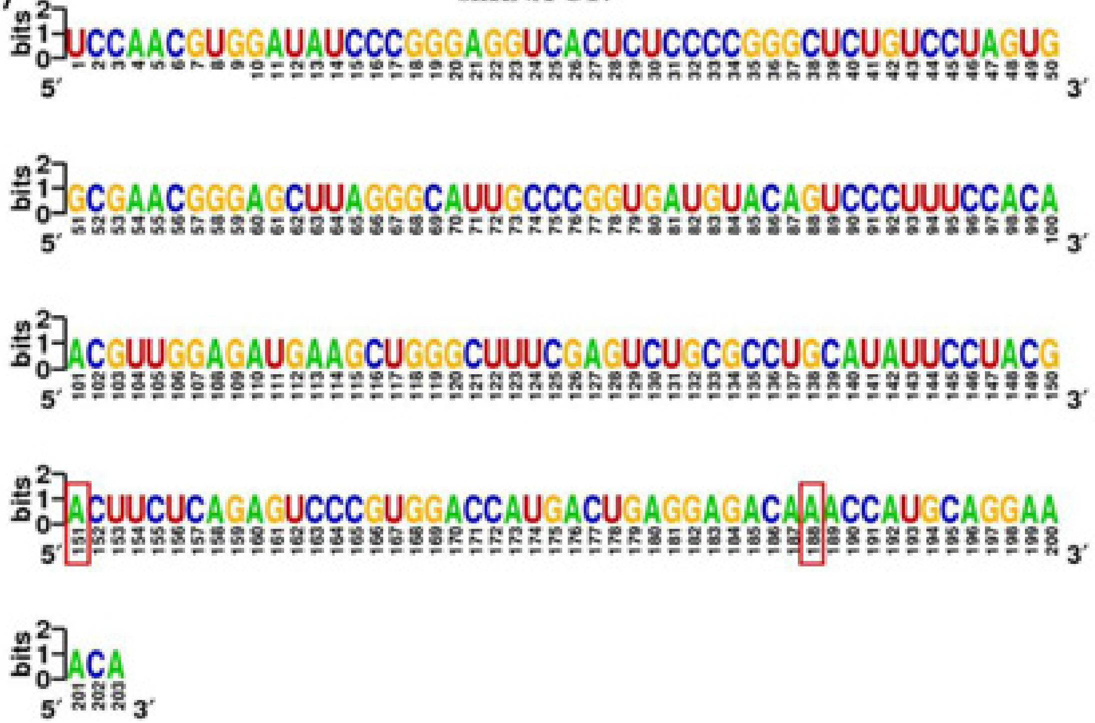
The sequence composition of the snRNAs U2 (A), U3 (B), U4 (C), U5 (D), U14 (E) and U17 (F) was built with the sequences containing edited sites, aligned with ClustalW [246] and generated with WebLogo [269]. The nucleotides which are pseudouridylated or methylated by snoRNA (data from [277]) are indicated by a green contour and the nucleotides which are edited by a red contour.




C)
snRNA U4
bits
AGCUUUGCGCAGUGGCAGUAUCGUAGCCAAUGAGGUUUAUCCGAGGCGCG
AUUAUUGCUAAUUGAAAACUUUUCCCAAUACCCCGCCUGACGACUUGA
AUAUAGUCGGCAUUGGCAAUUUUGACAGUCUCUACGGAGACUG
5′
3′
D)
snRNA U5
bits
AUACUCUGGUUUCUCUUCAGAUCGUAUAAAUCUUUCGCCUUUUACUAAAG
AUUUCCGUGGAGAGGAACAAUCUGAGUCUUACAAUUUUUUGAGGC
UGCAGGCUA
5′
3′
E)
snRNA U14
bits
UCGCUGUGAUGAUGGAUUCCAAAACCAUUCGUAGUUUCCACCAGAAUC
UGUGUUGGCUAGUUCCUUCCUUGGAUGUCUGAGCGAA
5′
3′

F)

snRNA U17


bits
2
1
0
5′
UCCAACGUGGAUAUCCCGGGAGGUCACUCUCCCCGGGCUCUGUCCUAGUG
3′
GCGAACGGGAGCUUAGGGCAUUGCCCGGUGAUGUACAGUCCCUUUCCACA
ACGUUGGAGAUGAAGCUGGGCUUUCGAGUCUGCGCCUGCAUAUUCCUACG
ACUUCUCAGAGUCCCGUGGACCAUGACUGAGGAGACAAACCAUGCAGGAA
ACA

Pierre Cattenoz

Caractérisation de l'expression des éléments Alu et du phénomène d'édition de l'ARN chez l'humain et la souris

Résumé

Dans la 1ère partie de cette étude, l'analyse de données de séquençage haut-débit révèle que ~40% des éléments Alu sont reconnus par POLIII, qu'ils sont présents en tant que petits ARN dans le cytoplasme et le noyau des cellules, que certain d'entre eux sont associés à la chromatine, et que la transcription des éléments Alu est un phénomène courant dans les tissus somatiques qui concorde avec l'expression d'éléments LINE1 fonctionnels. Ceci suggère que la rétrotransposition peut être un mécanisme normal dans la plupart des tissus humains. Enfin, l'analyse de l'expression des éléments Alu et LINE1 chez la souris montre que la transcription de rétrotransposons n'est pas spécifique de l'humain.

Dans la seconde partie de cette étude, une nouvelle méthode a été développée pour explorer l'impact de l'édition de l'ARN sur le transcriptome en identifiant les ARN édités par séquençage haut-débit. Dans un premier temps, un anticorps ciblant ADAR a été utilisé pour extraire les ARN associés aux protéines de l'édition. Cette méthode n'étant pas suffisamment efficace, une autre stratégie, qui extrait directement les ARN contenant de l'inosine, a été développée : dans un premier temps, l'ARN est fixé à des billes magnétiques par leurs extrémités 3', ensuite, les billes sont traitées au glyoxal/acide borique et à la RNAse T1 pour libérer la région 5' des ARN contenant une ou plusieurs inosines, et enfin, les ARN libérés sont séquencés par séquençage haut débit. En utilisant cette méthode, 1822 sites d'éditions ont été identifiés dans l'ARN de cerveau de souris, incluant 28 nouveaux sites présents dans des séquences codantes qui conduisent à des mutations non-synonymes des futures protéines. Des sites d'éditions ont aussi été observés pour la première fois dans les ARN ribosomaux, les snoRNA et les snRNA.

Résumé en anglais

In the first part of this study, we challenged the view that Alu elements are dormant occupant of the genome by characterizing their activity. Deep-sequencing data analyses revealed that ~40% of Alu elements can bind POLIII, they present a definite localization in the cell and associate with chromatin and polysomes, and that Alu elements transcription is a widespread phenomenon in normal tissues which correlates with functional LINE1 elements expression. This suggested that Alu element retrotransposition may be a natural mechanism in most normal human tissues. Further analyses showed that SINE and LINE expression in somatic tissues was not exclusive to human but also occurs in mouse. Finally, attempts were made to identify tissue specific insertions in the human genome resulting from retrotransposition events.

In the second part of this study, a new method was developed to understand the full impact of RNA editing on transcriptomes by characterizing the edited RNA in a high-throughput fashion. First, immunoprecipitation was attempted to pull-down RNA associated with the editing enzymes ADARs. Since this method was inefficient, another approach purifying directly the edited RNA was developed. First, the RNA was sequestered on magnetic beads. Then an inosine specific cleavage based on RNAseT1 treatment of RNA protected with glyoxal and borate allowed the separation of the edited RNA from the total RNA. Finally, deep sequencing was used to identify edited RNA. 1,822 editing sites were found in mouse brain RNA by this method, including 28 new editing sites modifying the coding sequences of genes and editing in rRNA, snoRNA and snRNA which were never observed before.

Key words: RNA, retrotransposons, deep-sequencing, ADAR, SINE, Alu elements, inosine, editing.