CORE-SINE : Une nouvelle classe de rétroposons des génomes eucaryotes.

par

Nicolas Gilbert

Thèse de doctorat effectuée en cotutelle

au

Programme de biologie moléculaire
Faculté des études supérieures
Université de Montréal

et

Ecole Doctorale des Sciences de la Vie et de la Santé
Université Blaise Pascal, Univesité d'Auvergne

Thèse présentée à la Faculté des études supérieures
en vue de l'obtention du grade de
Philosophiæ Doctor (Ph.D.)
en biologie moléculaire
et à
l'Université Blaise Pascal
en vue de l'obtention du grade de
Docteur d'Université

Mars 1999

Université de Montréal
Faculté des études supérieires

et

Université Blaise Pascal, Universuté d'Auvergne
Ecole Doctorale des Sciences de la Vie et de la Santé

Cette thèse intitulée :

CORE-SINE : une nouvelle classe de rétroposons des génomes eucaryotes.

Présentée et soutenue à l'Université de Montréal par :

Nicolas Gilbert

à été évaluée par un jury composé des personnes suivantes :

Thèse acceptée le : 99.05.21

UNIVERSITE BLAISE PASCAL UNIVERSITE D'AUVERGNE
ANNEE 1999

ECOLE DOCTORALE
DES SCIENCES DE LA VIE ET DE LA SANTE
N° d'ordre

Thèse

Présentée à l'Université Blaise Pascal
Pour l'obtention du grade de

DOCTEUR D'UNIVERSITE

SPECIALITÉ : BIOLOGIE MOLECULAIRE

Soutenue le :

Nicolas Gilbert

CORE-SINE : une nouvelle classe de retroposons des génomes eucaryotes.

Président :
Membres :
Rapporteurs :

SOMMAIRE

Chez l'humain, près de 30% de la masse génomique est constituée de séquences répétées dispersées qui se sont amplifiées par le mécanisme de rétroposition. Ce processus, présent dans tous les génomes eucaryotes, implique la transcription inverse de l'ARN d'un élément répété et l'intégration de l'ADNc qui en résulte dans une nouvelle localisation génomique. Les "Long Interspersed Elements" (LINE) codent pour les activités spécifiques de la rétroposition, telles que la transcriptase inverse et l'endonucléase. A l'inverse les "Short Interspersed Elements" (SINE) ne codent pour aucune activité enzymatique et sont considérés comme des "satellites" des éléments LINE.

Nous avons caractérisé 5 nouvelles familles de rétroposon SINE chez les mammifères. Celles-ci font partie des SINE dérivés d'ARNt et ont, comme caractéristique commune, un domaine central nommé "core". Les régions 3' sont distinctes pour chacune des familles, mais fortement identiques aux extrémités 3' de différents LINE. D'autres séquences SINE possédant ces mêmes critères sont présentes dans les génomes d'oiseaux, de reptiles, de poissons et de céphalopodes. Nous avons ainsi identifié une nouvelle "super-famille" de rétroposon appelée CORE-SINE présente chez tous les vertébrés. L'étude du rôle de chaque segment des CORE-SINE ; région dérivée d'ARNt, "core" et région dérivée de LINE, nous a permis de donner de nouveaux éléments de réponse sur l'évolution des rétroposons dans les génomes eucaryotes.

Enfin, nous avons décrit la présence d'un nouvel élément LINE dans les génomes de marsupiaux. Celui-ci est fortement identique au rétroposon Bov-B des génomes bovins et reptiles. Sa présence dans ces différents génomes soulève la possibilité d'un transfert horizontal de cet élément.

ABSTRACT

Almost 30% of the human genome consists of copies of interspersed repeats that amplified by retroposition, a process widely spread among eukaryotic taxa. Retroposition involves reverse transcription of the transcribed copies and reintegration of the resulting cDNAs into the host genome. Retroposition requires specific activities in addition to the enzymatic machinery commonly found in the host cells. The reverse transcriptase as well as the endonuclease involved in the cDNA synthesis and integration, are coded by the actively retroposing long elements such as LINEs. In contrast, short elements (SINEs) do not encode any protein facilitating their proliferation. However, these elements must have used both host-specific and retroposition-specific activities provided in *trans* to secure their efficient amplification.

We have characterised 5 new SINE retroposon families from mammalian genomes. They belong to tRNA-derived SINEs and have also a common central domain called "core". The 3' end regions of all families are distinct but they display high identity with the 3'extremities of different LINEs. Several SINEs with the same characteristics have been found in bird, reptile, fish, and cephalopod genomes. These data point to the existence of a new "super-family" of SINE retroposons, named CORE-SINE, present in all vertebrate genomes. The study of each CORE-SINE segments, i.e. tRNA-derived region, "core" and LINE-derived region, gave new insight into the evolution of retroposon in eukaryotic genomes.

Finally, we also described a new LINE element from marsupial genomes. It presents high identity with the Bov-B element from bovine and reptile genomes, which raises the possibility of a horizontal transfer of this element between genomes.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES.

LISTE DES ABRÉVIATIONS et NOMENCLATURE.

ADN:	Acide désoxyribonucléique
ARN:	Acide ribonucléique
ARNm:	ARN messager
ARN Pol II:	ARN polymérase II
ARN PolIII:	ARN polymérase III
ARNt:	ARN de transfert
ATP:	Adénosine triphosphate
°C:	degré Celsius
C/AI:	Chloroforme/alcool Isoamylique (24:1)
Ci:	Curie (unité de radioactivité)
CTP:	Cytosine triphosphate
cm:	centimêtre
cpm:	coup par minute (rayonnement radioactif)
dNTP:	désoxiribonucléotide triphosphate (ATP, CTP, GTP ou TTP)
GTP:	Guanosine triphosphate
g:	force centrifuge
Kb:	mille paires de bases
KDa:	kilodalton (unité de masse moléculaire des protéines)
LINE:	"Long Interspersed Element"
M:	concentration molaire
MA:	Millions d'années

mM:	millimolaire
ml:	millilitre
mmol:	millimole
µg:	microgramme
µJ/cm^2:	microjoule/centimètre carré
µl:	microlitre
µM:	micromolaire
ng:	nanogramme
nM:	nanomolaire
nm:	nanomètre
ORF:	cadre de lecture (Open Reading Frame)
p:	poids
pBS:	pBlueScript KS+
pb:	paire de base
Phénol/C/AI:	Phénol/Chlorophorme/Alcool Isoamylique (50:48:2)
pmole:	picomole
poly-A:	répétition simple d'Adénine ou aussi "queue" poly-Adénylée
rpm:	tour par minute (agitation ou centrifugation)
RTD:	Répétition Terminale Directe
SINE:	"Short Interspersed Element"
snRNA:	petit ARN nucléaire ("small nuclear RNA")
TE:	Tris EDTA
TTP:	Thymidine triphosphate

UV:	Ultra Violet
V:	Volt
v:	volume

Nomenclature internationale des nucléotides.

A : Adénine

C : Cytosine

G : Guanine

T : Thymine

U : Uridine

R : Purine (A ou G)

Y : Pyrinidine (C ou T)

M : A ou C

N : A ou C ou G ou T

REMERCIEMENTS

J'aimerais, pour commencer, remercier tous les membres du jury qui ont accepté d'évaluer ce travail.

Je voudrais ensuite exprimer toute ma reconnaissance au Dr. Damian Labuda pour m'avoir guidé avec confiance pendant ces années pour la réalisation de ce projet. Je remercie aussi le Dr. Jean-Marc Deragon pour sa confiance et sa grande patience.

De plus, j'ai une reconnaissance particulière pour les Drs. Ewa Zietkiewicz et Daniel Sinnett pour les nombreuses conversations qui ont aidé à l'avancée de mes travaux.

Merci à Chantal et Vania pour avoir dépassé le Maniatis dans toutes les situations ainsi que pour leurs ondes positives. Merci à nos informaticiens, David et Jean-François, pour la création "d'Analyse" et les solutions à mes "ratés" informatiques. Merci à tous les membres du groupe d'hémato-onco, Damian, Daniel, Wagner, Yves, Ewa, Chantal, Vania, Nathalie, Geneviève, Maya, Debi, Andrzej, Hugues, Stéphanie, Géraldine, Annie, Zeina, Sébastien, Brahim, Caroline, Gino, Alexandre, Sylvie, Hugo, Anne-Julie, et Raffaela pour les échanges scientifiques et culturels et pour les ha-ha-ha quotidiens.

Grand merci aux travailleurs de l'ombre, Anne, Gilles, Guillaume, Jean-Christophe et Stéphane qui ont bien voulu lire cette thèse avant les autres pour la rendre plus lisible.

Un immense merci à mes parents et à ma famille pour leur soutien.

Pour finir je remercie les organismes qui m'ont apporté un soutient financier : Le Ministère de l'Education Nationale, de la Recherche et de la Technologie (France), le fond du Téléthon des étoiles de l'Hôpital Sainte-Justine et le Programme de biologie moléculaire de l'Université de Montréal.

A Anne

INTRODUCTION

Tout génome eucaryote peut être divisé en trois catégories de séquences désoxyribonucléiques. La première comprenant les séquences hautement répétées, la deuxième les séquences moyennement répétées et la troisième les séquences uniques. Chacune de ces catégories est définie par une cinétique de renaturation différente de l'ADN en solution après dénaturation à haute température (Britten et Kohne, 1968). Les séquences hautement répétées se renaturant le plus rapidement. Chez les vertébrés, la proportion de séquences répétées est très variable, oscillant de près de 80% pour certains amphibiens à moins de 20% pour certains poissons (Britten et Davidson, 1971). Au sein des génomes mammifères, la variation de la quantité des séquences répétitives est moins forte, 45% chez les bovins (Britten et Kohne, 1968), de 20 à 40% chez les primates (Britten et Davidson, 1971) et de 30 à 60 % chez l'Homme (Houck *et al.*, 1979). Des études sur ces séquences répétées par hybridation croisée entre différents génomes, notamment entre la souris et le rat, montrent que des mécanismes d'amplification "rapide", d'un point de vue évolutif, doivent intervenir. En effet, deux génomes d'espèces proches peuvent ne pas partager les mêmes séquences répétées (Britten et Davidson, 1971).

Avec le développement de la biologie moléculaire, dans les années 80, et notamment les techniques de séquençage, il a été possible d'obtenir de nouvelles connaissances sur la structure, la nature et l'origine des séquences répétées. La première distinction a été de diviser les séquences répétitives en deux catégories majeures selon le type de distribution dans les génomes. La première catégorie représente l'ADN répété en tandem, ou encore ADN dit satellite (John et Miklos, 1979). Ce sont en général des motifs simples regroupés en bloc de répétition. La seconde représente l'ADN répété dispersé (Rogers, 1985; Schmid

et Shen, 1985). À l'intérieur de cette catégorie existe une grande hétérogénéité structurale des séquences et celles-ci sont généralement distribuées de façon aléatoire dans le génome.

La diversité structurale de toutes ces séquences a permis de définir différents mécanismes responsables de l'amplification de l'ADN répété dispersé, tels que la transposition, la duplication, le réarrangement chromosomique et l'intégration de génomes viraux (Finnegan, 1989). Cependant, les mécanismes biologiques dominants responsables de la formation des familles majeures des séquences répétées dispersées sont la rétrotransposition (Boeke *et al.*, 1985; Varmus et Brown, 1989) et la rétroposition (Rogers, 1985). Ces deux mécanismes font appel à la transcription, la transcription inverse et l'intégration, qui permettent l'amplification des éléments appelés rétrotransposons (Boeke *et al.*, 1985) et rétroposons (Rogers, 1983). La rétrotransposition est un mécanisme bien défini qui présente de fortes homologies avec l'intégration des rétrovirus (Varmus et Brown, 1989). En revanche, la rétroposition reste un mécanisme dont les étapes n'ont pas été décrites de façon directe. Même si les étapes importantes du mécanisme commencent à être comprises, de nombreuses questions restent posées :

- Quelle est la régulation spatiale et temporelle de l'amplification ?
- Comment les retroposons, au vu de leur large distribution chez les eucaryotes, sont-ils maintenus dans les génomes au cours de l'évolution ?
- Comment les rétroposons ont-ils évolué dans les génomes eucaryotes ?
- Quel est l'impact de la rétroposition et des rétroposons sur la dynamique et la variabilité des génomes ?

C'est grâce à l'étude d'un ancien rétroposon que nous allons essayer d'amener des éléments de réponses à ces questions.

1-1- Les Séquences Répétées.

1-1.1- L'ADN répété en tandem.

L'ADN répété en tandem a été largement étudié, et plusieurs types de ces séquences ont été décrits. Les ADN ribosomiques (ADNr) et les ADN satellites en sont des exemples. Tous les groupes de répétitions en tandem sont retrouvés dans tous les génomes eucaryotes. Ces séquences participent en général à des fonctions primordiales des cellules et peuvent être indispensables à leur survie.

1-1.1.1- Les ADNr.

Les ARN ribosomiques sont codés par deux gènes. Les ARN 18S, 5,8S et 28S sont transcrits à partir d'une unité polycistronique par l'ARN polymérase I (13.7 Kb chez l'Homme). Cette unité se présente sous la forme de plusieurs blocs de répétition en tandem "tête à queue". Le gène de la sous-unité 5S est séparé de l'autre gène ribosomique, mais est retrouvé aussi dans l'organisation d'un ou plusieurs blocs de répétition en tandem. Leur localisation est en général retrouvé à proximité du bloc de répétition des autres gènes ribosomiques. Les séquences d'ADN qui codent pour les ARN ribosomiques sont rassemblées dans des formations visibles au microscope optique, appelées nucléoles. Ces ARN sont les constituants principaux de la machinerie traductionnelle que constitue le ribosome.

Les gènes des histones sont un autre exemple de gènes organisés en répétition en tandem (Hentschel et Birnstiel, 1981). Leurs protéines participent à la structure

condensée de l'ADN.

L'organisation en tandem de ces gènes a pour effet d'augmenter le pouvoir transcriptionnel.

1-1.1.2- Les ADN satellites.

Les ADN satellites composent la majorité de l'ADN répété des génomes eucaryotes. Ils peuvent représenter jusqu'à 30% de la masse génomique, ne sont en général ni transcrits ni traduits, sont constitués de motifs peu complexes, et sont retrouvés généralement dans les régions centromériques et télomériques des chromosomes. Ils forment ainsi la majeure partie de l'hétérochromatine constitutive. De façon générale, les ADN satellites sont regroupés en tandem (John et Miklos, 1979). La répétition d'une même séquence peut être trouvée à différents locus d'un même génome.

Dans les génomes, les ADN satellites sont représentés par les télomères, les "grands satellites", les minisatellites et les microsatellites.

Les télomères sont constitués, chez les vertébrés, d'un motif de 6 nucléotides (TTAGGG) hautement répété (taille pouvant atteindre 30 Kb) (Morin, 1989). Ils sont situés à chacune des extrémités des chromosomes. La synthèse des télomères est effectuée par une transcriptase inverse spécifique, la télomérase (Greider et Blackburn, 1987). Plusieurs fonctions sont attribuées aux télomères : protection vis-à-vis de la dégradation par les nucléases, maintien de la longueur des chromosomes lors de la réplication, rôle dans l'organisation structurale via un attachement à la membrane nucléaire (Blackburn, 1991).

Les "grands" satellites sont des séquences regroupées en un ou plusieurs blocs généralement situés dans les régions centromériques (Willard, 1991). Les ADN α satellites

peuvent constituer de 3 à 5% de chacun des chromosomes (Willard, 1991). Ils peuvent subir des amplifications très rapides au cours de l'évolution. En effet, une étude a montré que 25% du génome des singes vert d'Afrique est constitué d'un ADN α satellite particulier, celui-ci ne représentant que 1 ou 2% de l'ADN chez l'humain (Kurnit et Maio, 1973), (Manuelidis, 1976). La fonction possible de ces familles de séquences satellites pourrait être en relation avec le processus de ségrégation des chromosomes (Wevrick *et al.*, 1990). L'origine des satellites alphoïdes est inconnue. Ils pourraient avoir évolué à partir d'unités heptanucléotidiques, provenant de satellites déjà présents dans le génome, et dont l'expansion aurait été induite par des recombinaisons inégales (Smith, 1976). Cependant un rapport récent démontre que l'unité d'un satellite α des cétacés, dont la taille est de 1600 nucléotides environ, est composée en partie d'un fragment fortement similaire à la région 3'terminale de l'élément LINE L1 des mammifères (Kapitonov *et al.*, 1998). Le premier modèle d'évolution est donc en partie remis en question, car il apparaît que des séquences qui n'ont pas, *a priori*, de prédisposition à participer à l'élaboration des satellites, peuvent en être partie intégrante. Chez l'humain il existe d'autres ADN satellites : les satellites I, dont l'unité de répétition est de 42 nucléotides, et les satellites II et III qui sont à l'origine issus de la répétition basale ATTCC.

Les minisatellites, ou VNTR (Variable Number Tandem Repeat) (Nakamura *et al.*, 1987), sont des répétitions définies par un motif central dont la taille peut varier de 10 à 60 nucléotides. La structure générale de ces minisatellites est fortement conservée chez les eucaryotes, de l'humain (Jeffreys *et al.*, 1985) aux végétaux (Tourmente *et al.*, 1994). Ils ont la particularité d'être répétés, dispersés et très polymorphes. Leur haute fréquence de recombinaison (10 fois supérieure à la moyenne) est due non seulement à l'organisation en

tandem, mais aussi à la nature intrinsèque du motif central, peu différent de la séquence chi de l'ADN d'*E. coli* (Jeffreys *et al.*, 1985). Ce polymorphisme très fort dans la population permet leur utilisation comme marqueur pour la cartographie (Nakamura *et al.*, 1987). L'utilisation d'un ensemble de ces marqueurs donne une empreinte génétique, ou *fingerprint*, particulière à chaque individu (Wong *et al.*, 1986).

Les microsatellites, ou SSR (Simple Sequence Repeats), sont des répétitions en tandem de un à cinq nucléotides. Ces motifs sont en général dispersés par petits blocs, d'un maximum de 100 paires de bases, à différents loci (Tautz, 1989). Le motif principal chez l'humain est la répétition $(CA)_n$ - $(GT)_n$ pour les dinucléotides et $(A)_n$ - $(T)_n$ pour les mononucléotides, séquences provenant essentiellement des terminaisons poly-adénylée (poly-A) des séquences répétées des familles de rétroposons *Alu* et L1 (Arcot *et al.*, 1995; Nadir *et al.*, 1996). Comme les minisatellites, ces répétitions sont polymorphes et peuvent servir d'outils pour l'étude des populations (Tautz, 1989), de marqueur génétique pour la cartographie (Weissenbach *et al.*, 1992) et les empreintes génétiques (Economou *et al.*, 1990). La variabilité de répétition microsatellite trinucléotidique est un facteur responsable de l'induction d'un certain nombre de maladies génétiques récessives, comme l'ataxie spinocérébrale et l'ataxie de Friedreich (Orr *et al.*, 1993; Campuzano *et al.*, 1996) ou dominantes, comme la dystrophie myotonique (Hunter *et al.*, 1992).

1-1.2- L'ADN répété dispersé.

Les séquences répétées dispersées sont très variées et de structures généralement complexes. Elles représentent plus de 35% de la masse génomique nucléaire chez l'humain. Elles appartiennent à l'ensemble des séquences moyennement répétées, mais une seule

famille peut représenter dans certains cas plus de 10% du génome, comme les éléments *Alu* du génome humain (Smit, 1996). Ces motifs sont distribués tout le long du génome de façon aléatoire entre des séquences uniques de l'euchromatine, ou entre des séquences de l'hétérochromatine (Rogers, 1985). Plusieurs de ces séquences ont été étudiées dans différents organismes allant de la bactérie aux eucaryotes supérieurs. Ces études ont permis d'établir que la majorité de ces séquences sont ou ont été mobiles. Elles ont donc été appelées éléments transposables. La diversité de leurs arrangements et structures a conduit à les classer en deux groupes différents selon leur mode de propagation. Le premier est constitué d'éléments mobilisés par un mécanisme de transposition ADN-ADN, les transposons. Le second groupe se propage à l'aide d'intermédiaires ARN et ADNc, les rétroéléments.

1-1.2.1- Les transposons.

Les transposons sont les éléments mobiles qui ont été trouvés dans tous les phyla procaryotes ou eucaryotes. Malgré leur grande variété, il est possible de donner des caractéristiques structurales générales. Leur taille varie entre 500 et 5000 nucléotides. Chaque extrémité de la séquence possède des répétitions de taille variable, de 13 à 1250 pb, en orientation inverse, appelées répétitions terminales inverses. Lors de l'insertion de l'élément dans l'ADN cible, une duplication du site d'insertion de chaque côté du transposon est créée, appelée répétition directe. La taille de ces répétitions, est aussi variable, de quelques nucléotides à une vingtaine.

Les transposons possèdent, en général, dans leur structure la protéine nécessaire à leur déplacement, la transposase ou l'intégrase. En effet, une seule enzyme est nécessaire

et suffisante, *in vitro*, pour induire le mécanisme de transposition (Kaufman et Rio, 1992). Les transposases et les intégrases possèdent un site catalytique fortement conservé qui est responsable de la transphosphorylation nécessaire pour la coupure de l'ADN et le transfert pendant la transposition. Le site catalytique a pour particularité de rassembler deux résidus aspartate et un glutamate (DDE) (Labrador et Corces, 1997). La transposition peut être soit réplicative, auquel cas une nouvelle copie est donnée à l'ADN, soit conservative, le transposon est déplacé d'un site de l'ADN à un autre. Dans tous les cas de figure, le transposon ne se trouve jamais à l'état libre dans la cellule. De façon générale, la transposition est un événement rare. Chez l'humain les transposons représentent moins de 2% du génome (Smit, 1996) et semblent avoir perdu toute activité de mobilité (Oosumi *et al.*, 1995; Smit et Riggs, 1996).

1-1.2.2- Les rétroéléments.

Les rétroéléments ont des structures très variées. Ils utilisent divers mécanismes de propagation, cependant tous nécessitent l'activité d'une transcriptase inverse, impliquant donc un intermédiaire ARN (Jagadeeswaran *et al.*, 1981; Van Arsdell *et al.*, 1981; Hollis *et al.*, 1982; Sharp, 1983; Boeke *et al.*, 1985). Suivant les structures des éléments et leurs identités avec les rétrovirus, les rétroéléments ont été séparés en deux groupes majeurs : la famille des rétroéléments de type viral et la famille des rétroéléments de type non viral.

1-1.2.2.1- La famille des rétroéléments de type viral.

Les rétroéléments de type viral possèdent les mêmes caractéristiques que les rétrovirus. Il existe deux catégories de ces types d'éléments dans les génomes : les

rétrovirus endogènes et les rétrotransposons (Temin, 1985). La première catégorie diffère des rétrovirus du fait que les éléments ont perdu leur caractère infectieux. La majorité d'entre eux, chez l'humain, se réduit à la présence d'un LTR solitaire (Long Terminal Repeat) (Smit, 1996). La seconde catégorie diffère des rétrovirus par l'absence du troisième cadre de lecture codant pour les gènes de l'enveloppe virale (Figure A_1 et A_2). Ces gènes de l'enveloppe sont responsables du pouvoir infectieux des virus. Les rétrotransposons sont très faiblement représentés, et sans doute même absents, dans le génome humain (Smit, 1996), mais peuvent dans certains génomes, comme celui du maïs, représenter une très large portion de la masse d'ADN (>50%) (SanMiguel *et al.*, 1996).

Les séquences des rétrotransposons sont construites en général de la manière suivante (Figure A_2) :

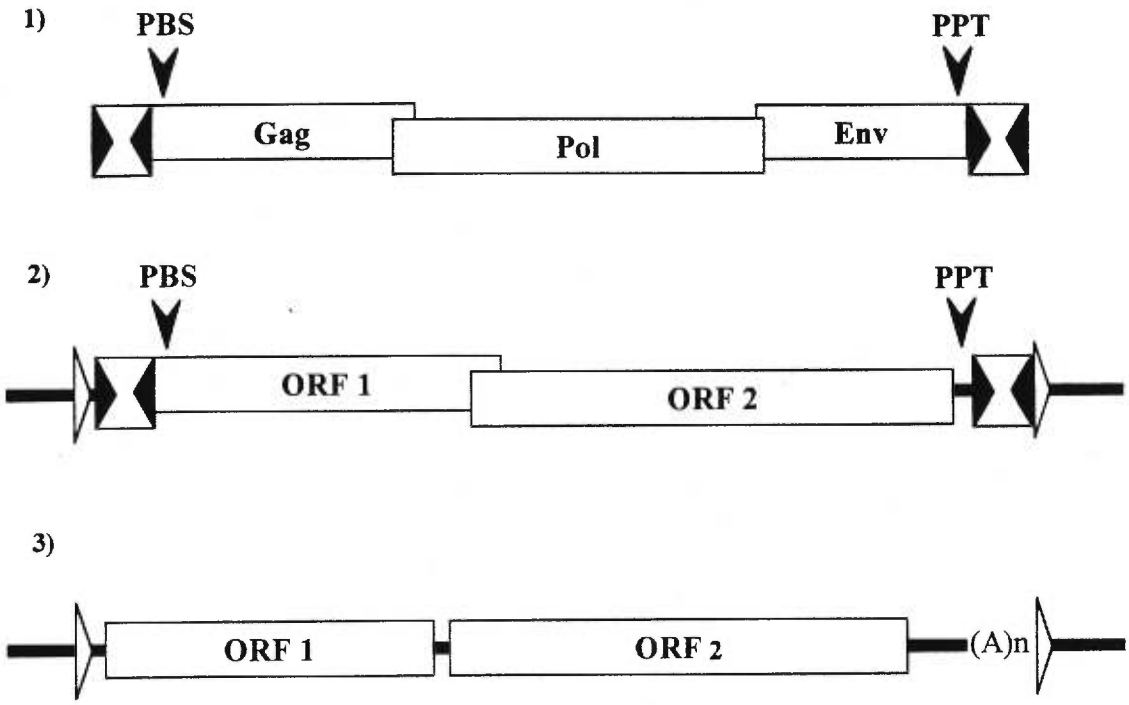
-Deux LTR identiques à chaque extrémité, indispensables à la régulation de l'expression, à la transcription inverse et à l'intégration de l'élément.

-Un site PBS (Primer Binding Site), adjacent à l'extrémité 3' du LTR gauche, indispensable pour l'amorçage de la transcription inverse et la synthèse du premier brin d'ADNc. Cet amorçage se fait avec un ARNt, spécifique pour chacun des rétrotransposons, qui s'apparie avec la partie PBS du transcrit (un modèle du mécanisme de rétrotransposition est décrit dans la Figure B).

-Deux cadres de lecture (ORF), dont un, ORF2, codant pour les activités transcriptase inverse (TR), endonucléase (EN), RNAse H (RH) et protéase (PR).

-Un site PPT (PolyPurine Tract) indispensable à la synthèse du second brin de l'ADNc de l'élément (Varmus, 1982; Temin, 1985) (figure B).

FIGURE A : Représentation schématique des rétroéléments actifs.



1) les rétrovirus

2) les rétrotransposons

3) les LINE (L1Hs)

PBS (Primer Binding Site)
PPT (Poly-Purine Tract)

: RTD, répétition directe flanquante dont la taille varie de quelques nucléotides à une vingtaine.

: LTR (Long Terminal Repeat), régulent la transcription, la transcription inverse et l'intégration.

(A)n : Répétition simple de type poly-A à l'extrémité 3' de la séquence.

Figure établie à partir des références Varmus et Brown (1989), Finnegan (1989, 1992), Hutchison *et al.* (1989).

FIGURE B : Mécanisme de transcription inverse rétrovirale.

La première étape (B1) consiste en la fixation de l'amorce, un ARNt, sur son site homologue PBS. Cette région d'homologie est de 16 à 19 nucléotides et permet la synthèse de 100 à 200 nucléotides par la transcriptase inverse en copiant le brin d'ARN du rétrotransposon. Le petit fragment d'ADN, appelé "strong stop", saute de l'extrémité 5' de l'ARN vers l'extrémité 3' du même ARN, où il se fixe sur une région homologue, la région R. Le fragment "strong stop" va ainsi servir d'amorce pour la synthèse complète du premier brin d'ADNc et du LTR droit.

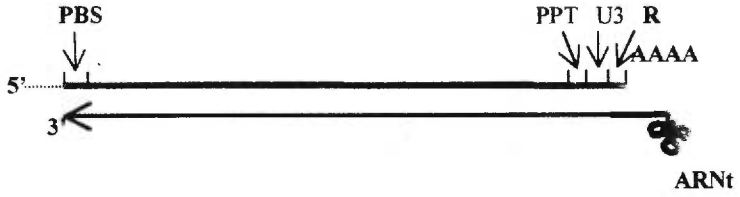
La seconde étape (B2) commence par la dégradation du brin d'ARN en complexe avec le premier brin de l'ADNc, par l'activité RNAse H. Seul un petit fragment de 20 nucléotides plus résistant à la RNAse est préservé, c'est la région PPT. Elle sert d'amorce pour la synthèse du second brin de l'ADNc. La synthèse complète nécessite un autre saut qui est réalisé lorsque la portion 3' du premier brin se referme en cercle par association de la portion 3' du second brin à la région homologue PBS. Le second brin est ainsi complété et le LTR gauche est formé.

Figure établie à partir des références Varmus (1982) et Temin (1985).

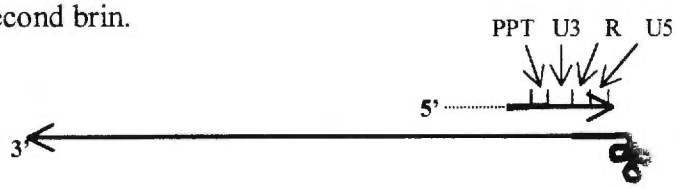
FIGURE B

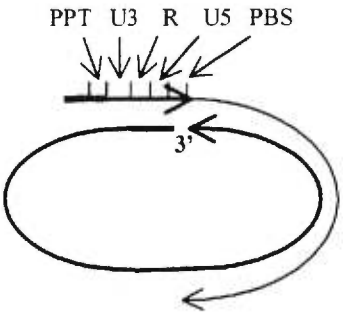
Initiation de la transcription inverse.

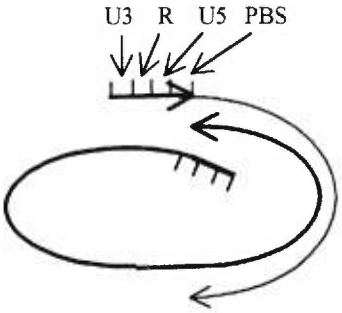
Synthèse du premier brin.



Synthèse du second brin.







ADNc du rétrotransposon.

Deux répétitions terminales directes (RTD) identiques se trouvent à chaque extrémité de l'élément et représentent la duplication du site d'intégration, signature d'une insertion d'un élément mobile.

Deux familles de rétrotransposons ont été définies à l'origine chez les animaux et les levures. Celles-ci se retrouvent également chez les végétaux supérieurs (Grandbastien, 1992). Les membres de la première famille sont caractérisés par leur identité avec les rétrotransposons Ty3 de la levure et gypsy de la drosophile. Ceux de la seconde famille, présentent des identités avec Ty1 de la levure et Copia de la drosophile. Ces deux familles ne diffèrent entre elles que par l'enchaînement des gènes du deuxième cadre de lecture, domaine fonctionnel des gènes *pol*, PR-TR-RH-EN pour la première famille et PR-EN-TR-RH pour la seconde (Xiong et Eickbush, 1990; Grandbastien, 1992).

1-1.2.2.2- La famille des rétroéléments de type non viral.

Au début des années 80 se sont accumulées les découvertes de nouveaux éléments mobiles qui, à la grande différence des rétroéléments de type viral, ne possèdent pas de LTR. De ce fait, certains auteurs ont appelé ces éléments des rétrotransposons non-LTR (Weiner *et al.*, 1986; Xiong et Eickbush, 1988b). D'autres auteurs, et notamment Rogers en 1983, leur ont donné le nom de rétroposons. C'est ce terme que j'utiliserai préférentiellement dans ce mémoire.

Le site d'intégration des rétroposons révèle toujours les RTD, constituant la spécificité des éléments transposables et la caractéristique de leur mobilité. Les RTD peuvent avoir des tailles variables pour les éléments d'une même famille de séquence. L'absence des LTR implique bien entendu un ou plusieurs mécanismes différents pour la

transcription inverse des ARN issus des séquences et pour la transposition. Trois groupes de rétroposons ont pu être constitués : les rétropseudogènes, les LINE et les SINE (Singer, 1982; Rogers, 1985). Ces derniers vont être décrits de façon plus détaillée dans les paragraphes qui suivent.

1-2- Les rétroposons.

1-2.1- Les rétropseudogènes.

L'existence de transcriptases inverses dans les virions de rétrovirus avait été établie en 1970 par deux laboratoires (Baltimore, 1970; Temin et Mizutani, 1970). Mais c'est essentiellement grâce à la découverte de rétropseudogènes dans le génome humain que l'activité d'une transcriptase inverse endogène fut établie de façon indirecte (Van Arsdell *et al.*, 1981; Denison et Weiner, 1982; Hollis *et al.*, 1982). Les rétropseudogènes sont divisés en deux catégories. La première comprend les pseudogènes homologues aux ARN des gènes de classe II (transcrits par l'ARN polymérase II). Ces pseudogènes diffèrent du gène originel par l'absence des régions introniques mais aussi par l'existence d'une terminaison poly-A du coté 3' (Hollis *et al.*, 1982; Moos et Gallwitz, 1983). Ces caractéristiques dues à l'épissage et à la maturation post-transcriptionnelle démontrent bien le passage par un ARN messager, et induisent aussi la perte de l'activité du pseudogène. Ces pseudogènes sont d'autant plus inactifs qu'ils ont perdu leur promoteur et qu'ils accumulent de nombreuses mutations dans leur séquence. De façon générale les rétropseudogènes sont rares et doivent apparaître par "accident" dans le génome.

La seconde catégorie est constituée essentiellement de pseudogènes des petits ARN

nucléaires (snRNA pour "small nuclear RNA"), telles que les familles U1 à U6, ou encore de petits ARN cytoplasmiques tel que BC200 chez l'humain ou BC1 chez les rongeurs. À l'inverse de la première catégorie, ces pseudogènes sont plus fréquents et peuvent même être représentés en plusieurs milliers d'exemplaires dans le génome, comme c'est le cas pour les familles U1 à U6 (Zieve, 1981).

1-2.2- Les LINE.

Le terme LINE, pour "Long INterspersed Element", a été introduit en 1982 pour décrire de nouvelles séquences répétitives de grande taille (>4 Kb) des génomes mammifères (Singer, 1982). Il a représenté une nouvelle classe de séquence après la caractérisation du premier élément, le LINE L1Hs (L1 Homo sapiens) du génome humain (aussi appelé KpnI) (Adams *et al.*, 1980; Skowronski et Singer, 1985). Dès lors, une définition plus précise a été attribuée aux LINE : ce sont des rétroposons actifs de grandes tailles. Le terme actif est associé directement au fait que les LINE possèdent au moins un cadre de lecture qui code pour les activités nécessaires à leur amplification (Hutchison III *et al.*, 1989). Rapidement, d'autres éléments ont été découverts et à l'heure actuelle, ils ont été décrits dans de nombreux génomes eucaryotes : chez les vertébrés, les invertébrés, les plantes, les champignons et les protozoaires (Tableau A, voir page 62 à la fin du chapitre). De façon générale, les LINE possèdent deux ORF. Cependant, un des premiers LINE identifié chez les insectes, l'élément R2, ne dispose que d'un seul cadre de lecture (Burke *et al.*, 1987). Dans le génome de *Trypanosoma cruzi*, un élément LINE, L1Tc, possède, lui, trois cadres de lecture (Martin *et al.*, 1995). Il semble que ces éléments soient des exceptions à la règle en ce qui concerne le nombre de cadres de lecture qui caractérise les

LINE. Une représentation générale des éléments LINE est montrée sur la figure A3.

Environ 95% des éléments LINE sont tronqués dans leur région 5' (Hutchison III *et al.*, 1989). Dans certains cas, les éléments "pleine longueur" peuvent représenter moins de 1% du nombre de copies présentes dans le génome, comme c'est le cas pour l'élément PsCR1 chez la tortue (Kajikawa *et al.*, 1997). La taille des éléments LINE d'une même famille dans un même génome peut ainsi varier de 60 nucléotides jusqu'à la taille "pleine longueur" (plusieurs Kb).

Il a été proposé à la fin des années 80, que les éléments LINE "pleine longueur" possèdent un promoteur interne reconnu par l'ARN polymérase II, notamment par le fait qu'ils sont constitués de séquences riches en G+C (Nur *et al.*, 1988; Di Nocera et Sakaki, 1990). Celui-ci a été par la suite décrit pour les éléments *Jockey* et F de la drosophile (Mizrokhi *et al.*, 1988; Minchiotti et Di Nocera, 1991). Pour l'élément *Jockey*, la transcription est fortement diminuée par l'action de l'α-amanitine qui est un inhibiteur spécifique de l'ARN Polymérase II. Dans la cas de l'élément F, il existe deux promoteurs (F-in et F-out) en sens opposés dans les 270 premiers nucléotides de la séquence. Ces deux promoteurs sont indépendants et F-in est responsable de la transcription de l'ARN du LINE. Un heptamère (GACGTGY) retrouvé dans les régions 5' non traduite (5'UTR) des éléments LINE F, G, *Jockey*, I, et Doc de la drosophile pourrait être le constituant fonctionnel du promoteur interne (Minchiotti et Di Nocera, 1991). Les études faites sur l'élément L1Hs décrivent qu'une région de 600 nucléotides dans la partie 5'UTR est impliquée dans la transcription par l'ARN polymérase II (Swergold, 1990). Cependant d'autres chercheurs ont déterminé que le promoteur des LINE L1 est reconnu par l'ARN polymérase III (Kurose *et al.*, 1995). En effet, la transcription des éléments L1Hs est

sensible à la tagetitoxine qui est un inhibiteur de l'ARN polymérase III. Ces auteurs ont démontré qu'un facteur de l'ARN polymérase II, YY1, se lie sur les 40 premiers nucléotides de L1Hs où se trouve une séquence très similaire à la boîte A des promoteurs Pol III. Ce facteur participerait à la transcription en étant associé à l'ARN polymérase III. Ce type de coopération entre les facteurs de transcription de l'ARN polymérase II avec l'ARN polymérase III a déjà été observé, notamment pour le facteur TFIID (Gabrielsen et Sentenac, 1991). La difficulté majeure pour déterminer la structure du promoteur interne est qu'il n'est pas conservé entre tous les éléments LINE. Il apparaît même que celui-ci n'est pas conservé à l'intérieur de certaines familles, notamment chez le LINE L1 des rongeurs. Il est proposé que le promoteur 5' des jeunes éléments L1 chez la souris ne dérive pas du promoteur des éléments plus vieux de cette même famille. Au cours de l'évolution, L1 aurait acquis un nouveau promoteur (Adey *et al.*, 1994).

Toutefois, malgré l'existence d'un promoteur interne, la majorité des éléments "pleine longueur" sont inactifs d'un point de vue traductionnel, du fait des nombreuses mutations observées dans les séquences des cadres de lecture (Skowronski *et al.*, 1988).

Les éléments LINE ont la particularité de posséder une terminaison 3' constituée d'une répétition simple de type microsatellite. Cette répétition simple est le plus souvent une "queue" poly-A, comme pour L1, mais elle peut être constituée de plusieurs nucléotides, comme par exemple un octamère (TATTCTAT) pour le LINE PsCR1 de tortue (Kajikawa *et al.*, 1997).

Chaque groupe de LINE est caractérisé par un consensus qui représente la séquence moyenne de tous les éléments de ce groupe présents dans un génome. Ce consensus, qui définit une famille LINE, est construit en utilisant pour chacune des positions nucléiques de

la séquence, le nucléotide le plus fréquent à cette position pour l'ensemble des membres de la famille. Le consensus représente ainsi la séquence ancestrale de la famille. En général, pour chaque famille d'élément, il est possible de diviser la population de séquences en sous-familles qui diffèrent du consensus général par la présence d'une ou plusieurs mutations communes. Ces mutations, appelées positions diagnostiques, suggèrent que l'amplification des rétroposons n'est générée que par un nombre restreint d'éléments fondateurs. L'origine des sous-familles sera développée ultérieurement dans le paragraphe décrivant les modèles d'amplification par rétroposition (paragraphe 1-3-).

1-2.3- Les SINE.

A l'inverse des éléments LINE, les SINE ("Short Interspersed Element") sont des éléments de petite taille (entre 100 et 400 nucléotides), définis comme étant des rétroposons passifs. Ils ne codent pour aucune protéine, et sont donc dépendants de l'activité rétropositionnelle d'éléments actifs comme par exemple les LINE ou les rétrovirus (Hutchison III *et al.*, 1989; Sinnett *et al.*, 1992). Contrairement aux LINE, tous les membres des familles SINE sont constitués de leur séquence "pleine longueur", avec des extrémités 5' et 3' bien définies (Weiner *et al.*, 1986).

De la même façon que pour les LINE, un consensus général peut être reconstruit pour chaque famille d'éléments SINE. Il représente la structure originelle de la famille considérée. Il est aussi possible de diviser la famille en sous-familles (paragraphe 1-3-).

Les SINE ont été découverts dans de nombreux génomes eucaryotes (Tableau B, page 64). À partir de ceux-ci, il a été établi une structure générale des éléments, identifiée par deux caractères dominants : le promoteur dans le segment 5' de la séquence et la

terminaison de répétition simple en 3'. Un schéma général de la structure des SINE est présenté sur la figure C.

Le promoteur est un promoteur interne de l'ARN polymérase III, identifié par les boîtes A et B (Figure C et D). Ces mêmes boîtes sont retrouvées dans toutes les séquences des gènes ARNt et de nombreux snRNA. De ce fait, les SINE possèdent en règle générale une forte identité avec les ARNt et on dit de ces SINE qu'ils dérivent de ces ARN de transfert (Lawrence *et al.*, 1985; Sakamoto et Okada, 1985; McNamara *et al.*, 1990; Okada, 1991a). Cependant, il existe des exceptions à cette règle. En effet, l'élément SINE le plus étudié chez les vertébrés, le rétroposon *Alu* des génomes des primates, est dérivé de l'ARN du gène 7SL (Ullu et Tschudi, 1984). Le gène 7SL possède lui aussi un promoteur interne de l'ARN polymérase III mais est légèrement différent de celui des ARNt (Figure D). *Alu* n'est pas la seule exception à la règle, mais nous reviendrons sur les origines possibles des rétroposons de type SINE de façon plus exhaustive dans ce chapitre au paragraphe 4.

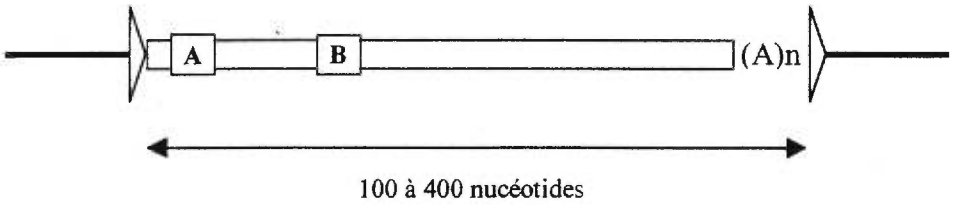
Les éléments SINE, comme les rétroposons LINE, sont caractérisés par une répétition simple de type poly-A ou plus complexe, comme par exemple des trinucléotides, à leur extrémité 3'.

1-3- La rétropositon.

1-3.1- Le modèle général.

Les premières hypothèses sur le mécanisme de rétroposition ont été fondées sur l'analyse de séquences rétroposons des génomes mammifères. Celles-ci proposent que la rétroposition implique trois étapes fondamentales qui sont la transcription spécifique de

FIGURE C : Structure générale des éléments SINE.



▷ : RTD dont la taille varie de quelques nucléotides à une vingtaine.

A B : Ces boîtes représentent le promoteur interne de l'ARN polymérase III présent chez tous les SINE.

(A)n : Répétition simple de type poly-A à l'extrémité 3' de la séquence.

FIGURE D : Structure des promoteurs de l'ARN polymérase III de différents SINE et de séquences reliées.

Séquences	5'	Boîte A		Boîte B
Consensus Pol III *	7	...TGGCNNAGTGG	....25-35....	GGTTCGANNCC
ARNt met (humain)	7	...TGGCGCAGcGG	.....31.....	GGaTCtAAACC
Alu type II (galago)	14	...TaGCACAGTGG	.....32.....	GGTTCGAACCC
B2 (rongeur)	13	...TGGCTCAGTGG	.....35.....	aGTTCaAATCC
MIR (mammifère)	7	...yaGCATAGTGG	.....33.....	GGTTCGAATCC
ARN 7SL (humain)	3	...**g**GGCGC**g**GTGG	.....**59**.....	**a**GTTCtgGGCt
Alu type I -gauche (primate)	4	...**g**GGCGC**g**GTGG	.....**60**.....	**a**GTTCGAGACC
Alu type I -droite (primate)		.....**g**GGCGT**g**GTGG	.....**58**.....	GaggCGgAGgt
B1 (rongeur)	3	...**g**GGCGT**g**GTGG	.....**50**.....	**a**GTTCGAGGCC

* Galli *et al.* (1981)

Les chiffres indiquent le nombre de nucléotides avant la boîte A et entre les deux boîtes. Les lettres en petit caractère indiquent la divergence avec le consensus général qui se trouve à la première ligne. Les caractères en gras sont communs au promoteur du gène de 7SL.

l'élément, la transcription inverse et enfin l'intégration dans un nouveau locus chromosomique. Ces trois étapes ont été tout d'abord considérées indépendantes (Jagadeeswaran *et al.*, 1981; Van Arsdell *et al.*, 1981). Cependant, de nombreuses études sur ces trois étapes ont permis d'affiner les premières hypothèses et de proposer le modèle actuel du mécanisme de rétroposition. Ce dernier suggère que la transcription inverse s'effectue au site d'intégration.

1-3.2- La transcription.

1-3.2.1- La transcription *in vitro*.

La transcription représente la première étape déterminante pour la rétroposition. Même si les LINE et les SINE ne sont pas transcrits par la même ARN polymérase, tous les rétroposons possèdent un promoteur interne qui assure ainsi l'intégrité et la conservation de l'activité pour les éléments qui sont issus de l'amplification. Parce que les promoteurs des différentes familles LINE ne semblent pas posséder une structure commune conservée, les études sur la transcription ont été essentiellement effectuées sur des éléments SINE de mammifères tel que *Alu*, B1, B2 et ID.

La transcription *in vitro* des séquences *Alu*, pour l'humain, et ID, pour les rongeurs, a déterminé que les éléments SINE sont effectivement transcrits par l'ARN Pol III et que l'initiation de la transcription se fait à la position +1 de la séquence. La transcription s'achève lors de la rencontre d'un terminateur externe, qui se caractérise par une succession d'au moins 4 thymines, en aval de la séquence (Jagadeeswaran *et al.*, 1981; Van Arsdell *et al.*, 1981; Schmid et Maraia, 1992). L'intégrité des boîtes A et B du promoteur interne est nécessaire et suffisante pour initier la transcription.

Ainsi, toutes les séquences ayant un promoteur interne sont considérées comme potentiellement actives (Duncan *et al.*, 1979; Elder *et al.*, 1981; Fuhrman *et al.*, 1981; Jagadeeswaran *et al.*, 1981; Gutierrez-Hartmann *et al.*, 1984). Ces premiers résultats supportent le modèle d'amplification en cascade, qui propose que tout élément possédant un promoteur interne intact peut générer de nouveaux éléments. Il en découle une amplification exponentielle des éléments dans le génome hôte (Jagadeeswaran *et al.*, 1981; Van Arsdell *et al.*, 1981). Or, ce modèle, basé sur l'hypothèse que l'efficacité de la rétroposition est étroitement liée à la transcription, n'explique pas comment d'autres séquences possédant aussi un promoteur interne pour l'ARN Pol III (par exemple les gènes ARNt, 5S, 7SL, 4,5S) ne sont pas amplifiées avec la même efficacité. De plus, si le modèle en cascade était effectivement le mécanisme d'amplification, il impliquerait alors que la distribution des mutations sur les séquences des éléments SINE soit aléatoire. Or les analyses approfondies des séquences montrent que ce n'est pas le cas. Pour retrouver une distribution aléatoire des mutations, il faut diviser les éléments d'une même famille en groupe de séquences qui forment les sous-familles (Labuda et Striker, 1989; Deragon *et al.*, 1994). Ces dernières sont caractérisées par la présence de positions diagnostiques (Willard *et al.*, 1987; Jurka et Smith, 1988; Quentin, 1988; Quentin, 1989). De nombreux travaux ont permis d'identifier les sous-familles pour la majorité des SINE provenant aussi bien des génomes mammifères (Willard *et al.*, 1987; Jurka et Smith, 1988; Quentin, 1989; Batzer *et al.*, 1990; Matera *et al.*, 1990; Jurka et Milosavljevic, 1991; Shen *et al.*, 1991; Leeflang *et al.*, 1992; Zietkiewicz *et al.*, 1994; Zietkiewicz et Labuda, 1996; Zietkiewicz *et al.*, 1998) que des génomes d'autres vertébrés, comme les poissons (Kido *et al.*, 1994; Kido *et al.*, 1995) ou encore des génomes de plantes (Yoshioka *et al.*, 1993; Deragon *et al.*, 1994). La

classification en sous-familles a aussi été possible pour les éléments LINE, et notamment pour le rétroposon L1 des mammifères (Smit *et al.*, 1995). Pour l'élément *Alu* du génome des primates, la nomenclature des sous-familles a été standardisée, car une même sous-famille pouvait posséder plusieurs noms suivant l'origine du laboratoire (Batzer *et al.*, 1996).

Les études de la transcription *in vivo* ont permis d'éliminer complètement l'hypothèse d'une amplification en cascade.

1-3.2.2- La transcription *in vivo*.

Les études de la transcription *in vivo*, à partir des mêmes éléments SINE, ont permis de donner des éléments de réponse aux questions soulevées par les résultats des expériences de transcription *in vitro*. Tout d'abord, il a été démontré que les transcrits observés sont essentiellement générés par cotranscription d'éléments SINE présents dans les unités de transcription de l'ARN Pol II. En effet, les séquences *Alu* constituent près de 10% du contenu des ARN pré-messagers et quelques-unes persistent dans les ARNm matures (Weiner *et al.*, 1986; Barnett *et al.*, 1993). Ces transcrits ainsi formés ne peuvent pas participer à la rétroposition. Il a été ensuite observé que les transcrits spécifiques de l'ARN Pol III sont en fait très peu représentés dans les cellules (Paulson et Schmid, 1986; Sapienza et St-Jacques, 1986; DeChiara et Brosius, 1987). Ces résultats vont à l'encontre du modèle en cascade qui voudrait, au vu du nombre très élevé de copies dans les génomes, que l'expression des éléments SINE soit forte. En revanche ces observations ont conduit à proposer un second modèle dit, modèle des séquences maîtresses (Shen *et al.*, 1991). Il suggère que parmi les éléments d'un génome donné, seule une petite fraction est active,

c'est-à-dire apte à être amplifiée. Cette hypothèse est soutenue par des études *in vivo*, démontrant que seule est exprimée une fraction des éléments SINE appartenant en majorité aux sous-familles les plus jeunes (Sinnett *et al.*, 1992; Liu *et al.*, 1994; Deragon *et al.*, 1996). Cette aptitude à être rétroposée est liée, chez les éléments SINE dérivés de l'ARN 7SL, à la conservation de la structure secondaire de l'ARN semblable à celle de leur géniteur. Il s'agit donc d'une sélection au niveau ARN (Sinnett *et al.*, 1991; Sinnett *et al.*, 1992; Labuda et Zietkiewicz, 1994).

Une autre version de ce modèle postule qu'une seule séquence maîtresse ou gène maître, est à l'origine de chaque sous-famille. Par opposition au modèle précédent, la sélection se ferait au niveau ADN. De plus ces gènes maîtres ne sont jamais actifs en même temps mais successivement. C'est l'hypothèse "Master genes amplification" proposée par Deininger (Deininger et Slagel, 1988; Deininger *et al.*, 1992). Cette hypothèse est soutenue par les faits que les sous-familles peuvent être datées et qu'elles ont été amplifiées dans un ordre séquentiel. L'analyse des positions diagnostiques observées sur les consensus des sous-familles montre qu'elles s'ajoutent aux positions diagnostiques de la sous-famille plus vieille (Shen *et al.*, 1991). De plus, l'analyse de trois insertions *Alu* récentes presque identiques suggère qu'elles proviennent d'un unique gène maître (Deininger et Slagel, 1988). Les premières classifications en sous-familles du LINE L1 chez la souris laissaient aussi supposer un seul gène maître (Hardies *et al.*, 1986; Deininger *et al.*, 1992; Schichman *et al.*, 1993).

Cependant d'autres groupes proposent que les gènes maîtres puissent s'amplifier pendant des périodes qui se chevauchent (Britten *et al.*, 1988; Jurka et Milosavljevic, 1991; Schmid et Maraia, 1992). Cette nouvelle proposition a été élaborée et

développée grâce, non seulement à l'identification de nouvelles insertions d'éléments SINE (*Alu* ou autre) issues d'anciennes sous-familles ou de plusieurs séquences d'une même sous-famille (Matera *et al.*, 1990; Leeflang *et al.*, 1992; Sinnett *et al.*, 1992; Hutchinson *et al.*, 1993; Jurka, 1993; Deragon *et al.*, 1996; Tachida, 1996), mais aussi grâce à l'analyse informatique des séquences présentes dans les banques de données (Jurka et Milosavljevic, 1991). L'analyse exhaustive des éléments L1 des mammifères conclut aussi que plusieurs séquences ont généré, dans la même période, de nouvelles copies génomiques de rétroposons (Smit *et al.*, 1995). L'accumulation de ces données a conduit à proposer que certaines séquences ont été générées par des éléments qui étaient inactifs et dont l'activité aurait été restaurée (Deininger et Batzer, 1995). C'est le modèle dit, formation parallèle des sous familles (parallel subfamily formation).

De toutes ces analyses le modèle d'amplification retenu est le suivant. L'amplification des rétroposons se fait de manière continue dans les génomes. L'ensemble des copies est issu d'une succession de séquences actives provenant de plusieurs loci génomiques, mais peu nombreux. Chaque sous-famille peut être le résultat de l'amplification de plusieurs séquences génératrices puisque plusieurs séquences peuvent être actives en même temps.

Il faut noter que le terme "amplification continue" ne signifie pas que le taux d'amplification est constant au cours du temps. En effet si on observe l'amplification des séquences *Alu* au cours du temps, on s'aperçoit qu'elle n'est pas constante mais a été provoquée par une succession de vagues. De plus l'amplification semble s'être beaucoup ralentie depuis les 30 derniers millions d'années (Britten, 1994). Il est en effet estimé que le taux de rétroposition est aujourd'hui 100 fois plus lent qu'il y a 40-50 millions d'années

(Deininger et Batzer, 1993). Cette diminution d'efficacité de rétroposition serait liée à la perte d'affinité de liaison de l'ARN du SINE *Alu* avec un polypeptide multimérique ribosomique, SRP9/14 (SRP9/14 fait partie de la ribonucléoprotéine cytoplasmique 11S, appelée aussi SRP, "signal recognition particule") (Sarrowa *et al.*, 1997). Il est aussi intéressant de noter que le taux de rétroposition d'une même famille SINE diffère selon les espèces. En effet, des études effectuées chez le rat, la souris, le hamster et le cobaye (cochon d'Inde), ont montré que le nombre de copies de l'élément ID de ces rongeurs varie de 130 000 (pour le rat) à 200 (pour le cobaye) (Sapienza et St-Jacques, 1986; Anzai *et al.*, 1987; Deininger et Batzer, 1993). Ces observations démontrent que le taux d'amplification des SINE est très variable et peut dépendre du génome hôte (Kass *et al.*, 1996). Par observation des représentations du rétroposon L1 chez les mammifères, nous pouvons en déduire qu'il en est de même pour les LINE.

La transcription est donc primordiale pour la rétroposition. Cependant, une meilleure compréhension du contrôle de la transcription devrait permettre d'expliquer la sélectivité des séquences maîtresses.

1-3.2.3- Le contrôle de la transcription.

Pourquoi les éléments SINE sont si peu transcrits alors qu'ils sont nombreux dans le génome et qu'ils possèdent un promoteur interne?

La première hypothèse est que des régions chromosomiques peuvent être éteintes d'un point de vue transcriptionnel empêchant ainsi l'accès de la machinerie transcriptionnelle aux séquences promotrices des éléments SINE. Ce mécanisme de répression transcriptionnelle peut être induit par la méthylation. La méthylation se fait

essentiellement sur les Cytosines des dinucléotides CpG. L'hyper-méthylation de ces sites a pour effet d'inhiber complètement la transcription dépendante de l'ARN Pol III (Juttermann *et al.*, 1991; Tachida, 1996). De plus la méthylation induit la mutation rapide des dinucléotides CpG (Bird, 1980). En effet, la base 5-méthylcytosine peut subir une trans-amination sur le carbone 4 de la base, et se transformer en thymine. Le dinucléotide devient ainsi TpG, ou encore CpA pour son complémentaire. De nombreux dinucléotides CpG se trouvent dans les boîtes A et B du promoteur interne de l'ARN Pol III, ou à proximité ; ainsi les mutations ont pour effet d'induire une perte d'affinité avec la machinerie transcriptionnelle. Ce mécanisme mutationnel explique le fait que la majorité des éléments SINE perdent très vite leur potentiel transcriptionnel et deviennent définitivement inactifs (Schmid, 1991). L'étude comparative de transcription *in vitro* de différents éléments SINE chez le galago (monomère, *Alu* type I et *Alu* type II) tend à démontrer que l'affinité du promoteur avec l'ARN polymérase III décroît si un dinucléotide CpG de la boîte B des promoteurs devient CpA ou TpG (Daniels et Deininger, 1991; Schmid et Maraia, 1992). Des résultats similaires sont obtenus lors de la mutation de trois CpG en CpA situés dans ou près de la boîte A du promoteur du rétroposon *Alu* (Liu et Schmid, 1993). Finalement l'analyse des transcrits Alu dans les cellules Ntera2D1 montrent que ceux-ci possèdent de nombreux dinucléotides CpG intacts (Sinnett *et al.*, 1992).

L'étude des rétroposons *Alu* dans différents tissus a montré que la majorité des sites sont méthylés (Schmid, 1991). Il apparaît donc que la méthylation est le principal agent répresseur de la transcription des rétroposons via les sites CpG. La répression pourrait être médiée par une protéine de liaison à l'ADN, MeCP1 (Liu et Schmid, 1993). Cette protéine est responsable de la répression de la transcription par l'ARN Pol II induite par la

reconnaissance des sites méthylés (Boyes et Bird, 1991). Ceci est d'autant plus remarquable que les éléments jeunes de la famille *Alu* sont très méthylés *in vivo* (Schmid, 1991).

D'autres expériences, en accord avec les précédentes, ont montré que les rétroposons SINE sont hypométhylés dans les cellules germinales mâles et que ce phénomène est associé à une hausse de la transcription *in vitro* (Kochanek *et al.*, 1993; Rubin *et al.*, 1994). Cette levée d'inhibition dans les cellules germinales est utile à la rétroposition : en effet, si la nouvelle insertion doit être fixée dans le génome, alors elle doit se faire dans les cellules qui seront à l'origine de la génération suivante.

La seconde hypothèse, en complément de la première, est que seul sont actifs les éléments SINE se trouvant dans des régions de chromatines actives, c'est-à-dire près d'unités de transcription de l'ARN Pol II. Ces régions ouvertes permettent l'accessibilité des promoteurs au complexe de transcription (Slagel et Deininger, 1989). Dans de tels environnements, la proximité d'un promoteur à ARN Pol II peut contribuer à l'expression des rétroposons (Schmid et Maraia, 1992). Il a été déterminé que certains facteurs couplés à l'ARN Pol II pouvaient interagir avec l'ARN Pol III, notamment le facteur TFIID. Par exemple, celui-ci interagit avec l'ARN Pol III par l'intermédiaire du facteur TFIIIB pour la transcription du gène U6 (Das *et al.*, 1988; Gabrielsen et Sentenac, 1991). Il a aussi été décrit que l'élément Bm-1 de *Bombyx mori* utilise la boîte TATA des promoteurs à ARN Pol II pour sa transcription (Wilson *et al.*, 1988). L'importance des séquences flanquantes a été aussi démontrée pour l'expression de certains éléments *Alu*. Dans deux cas, la région en amont d'un élément, d'environ 40 nucléotides, possède soit de courtes répétitions en tandem de 9 nucléotides (Ullu et Weiner, 1985) soit un site de fixation du facteur de transcription Ap1 (Chesnokov et Schmid, 1996). Dans ces deux cas, la délétion de la région

de 40 nucléotides fait disparaître l'expression de l'élément *Alu*.

En complément de ces résultats, il a été observé que l'expression des rétroposons *in vivo* possède dans certains cas une spécificité tissulaire. Dans les cellules germinales, les cellules indifférenciées et les cellules cancéreuses, l'expression des éléments SINE est généralement plus forte que dans les tissus somatiques. Dans le cas particulier du rétroposon ID des rongeurs, les éléments, même s'ils sont transcrits dans tous les tissus (Sapienza et St-Jacques, 1986; Mellon *et al.*, 1988), sont fortement transcrits dans les tissus nerveux. Cette expression différentielle est sans doute liée au fait que le SINE ID prend son origine du gène BC1 (Kim *et al.*, 1995), qui est exprimé spécifiquement dans les tissus nerveux (Sutcliffe *et al.*, 1984), mais aussi à la présence de régions flanquantes favorables dans le contexte tissulaire du système nerveux.

De façon plus précise, il a été observé chez les plantes que l'expression de certains éléments est liée aux tissus. Pour une même plante, *Brassica napus* (Colza), ce ne sont pas les mêmes éléments SINE S1 qui sont exprimés dans les racines, dans le mélange tiges et feuilles ou dans les cellules en culture (Deragon *et al.*, 1996). Ces résultats suggèrent aussi que l'environnement génétique des rétroposons influence la transcription.

Des expériences de transformation de cellules suggèrent que des facteurs en *trans* ont une activité qui stimule la transcription par un signal se trouvant en *cis* dans la région flanquante de l'élément SINE. En effet, une transformation de cellules par le virus SV40 stimule fortement la transcription des éléments B1 et B2 chez les rongeurs (Carey *et al.*, 1986). Cet effet activateur de transcription se fait par l'intermédiaire de la formation d'un complexe de transcription sur des éléments décrits au départ comme étant inactifs (Carey et Singh, 1988). Ce *trans*-activateur serait fortement inhibé dans des conditions

normales et ainsi limiterait l'expression des rétroposons. Des résultats similaires ont été obtenus en étudiant la réponse à des stress cellulaires. Des stress physiques (chaleur) ou chimiques (inhibiteur de traduction par le cycloheximide) ont pour effet d'augmenter la transcription des éléments SINE de façon spécifique (Liu *et al.*, 1995).

1-3.2.4- La transcription et la traduction des éléments LINE.

Les LINE possèdent des promoteurs internes et leurs séquences sont riches en dinucléotides CpG. Comme pour les SINE, la transcription est sensible à la méthylation. Il a été démontré que le traitement de fibroblastes embryonnaires à la 5-azadeoxycytidine (empêchant la méthylation) provoque une augmentation d'expression des éléments L1 de près de 4 fois par rapport à une expression basale (Woodcock *et al.*, 1997). De même, la méthylation de 5 des 22 sites CpG présents sur le promoteur de L1Hs induit une inhibition de 75 % de la transcription (Nur *et al.*, 1988). De façon plus précise, il a été déterminé que les quatre premiers CpG de la séquence des éléments L1 sont essentiels à l'inhibition (Hata et Sakaki, 1997). Il semble donc que le phénomène de méthylation soit un des principaux contrôles de la transcription.

La transcription est indispensable à la rétroposition des éléments LINE pour deux raisons : la première est que son mécanisme implique une étape ARN; la seconde est que les éléments codent pour les activités enzymatiques, comme par exemple la transcriptase inverse, nécessaires à l'amplification. Il existe donc une étape de traduction des ARN LINE qui n'existe pas chez les SINE. Cette étape peut induire des contrôles supplémentaires pour la rétroposition. Les transcrits LINE sont généralement bi-cistroniques. La première initiation de la traduction se fait à la position 1035 pour les

éléments L1Hs, dans un environnement favorable présentant une séquence de Kozak (GCCRCCAUGG) (Leibold *et al.*, 1990). À la fin du premier cadre de lecture, le ribosome se dissocie et une réinitiation est nécessaire pour la traduction du deuxième cadre de lecture (McMillan et Singer, 1993). Des modèles bi-cistroniques ont été étudiés chez les primates et il a été constaté que la traduction du second gène s'effectue à un taux variant de 10 à 20 % par rapport au premier gène (Peabody et Berg, 1986). Le taux de réinitiation dépend de la distance qui sépare le codon stop de l'AUG suivant. Plus la distance est grande, plus la réinitiation est efficace (Kozak, 1987). Dans le cas du rétroposon L1Hs il a été déterminé que la traduction de l'ORF2 est près de 100 fois plus faible que celle de l'ORF1 (McMillan et Singer, 1993). La rétroposition dépend donc aussi de l'efficacité de traduction.

Des études récentes sur l'activité rétropositionnelle, par transfection transitoire dans des cellules humaines ou murines, des éléments L1 chez l'humain et la souris ont démontré qu'une mutation sur un site actif de l'ORF2 inhibe complètement la rétroposition de l'élément transfecté. Cette observation permet de conclure que les protéines fonctionnelles produites par un élément actif agissent préférentiellement sur l'ARN qui les génère : les protéines agissent préférentiellement en *cis* (Moran *et al.*, 1996; Boeke, 1997; Naas *et al.*, 1998). Ce phénomène restreint considérablement le nombre d'éléments LINE qui peuvent être amplifiés, et ces derniers ont été dénombrés à environ 40 copies chez l'humain (tous membres de la sous-famille L1Ta) (Sassaman *et al.*, 1997) et près de 300 chez la souris (Naas *et al.*, 1998). Les expériences faites sur ces éléments actifs de la souris ont permis aussi de vérifier non seulement que le promoteur interne est suffisant pour la transcription mais aussi que la présence d'un promoteur Pol II en amont de l'élément contribue à une forte augmentation de la transcription (Naas *et al.*, 1998). Ainsi

l'environnement chromosomique d'un rétroposon peut agir sur son activité.

1-3.3- La transcription inverse et l'intégration des rétroposons.

Le premier modèle de mécanisme de rétroposition pour les éléments SINE, proposé au début des années 80, sépare les étapes de transcription inverse et d'intégration. Dans ce modèle, la transcription inverse se fait dans le cytoplasme par auto-amorçage (Jagadeeswaran *et al.*, 1981). L'extrémité 3' de l'ARN est composée d'un segment poly-Uridine, créé par le site de terminaison de transcription de l'ARN Polymérase III. Ce segment vient s'apparier avec la terminaison poly-adénine de la séquence du rétroposon pour initier la transcription inverse (Figure E). Ce mécanisme permet aux ARN de perdre leurs séquences 3' adjacentes ne faisant pas partie du SINE. L'ADNc simple brin est transporté vers le noyau par un mécanisme non connu. Le site d'intégration est généralement riche en adénines, suggérant qu'il fixe la terminaison 5' riche en thymines du brin complémentaire d'ADN. L'insertion amorce la synthèse du second brin. La seconde partie de ce mécanisme (étape nucléaire) permet aussi d'expliquer la présence des rétropseudogènes dans les génomes car les ARN messagers possèdent une terminaison poly-A (Moos et Gallwitz, 1983). Ce premier modèle imposant un passage par le cytoplasme permet aussi d'expliquer, pour les rétropseudogènes, l'élimination des introns.

Ce modèle est en partie soutenu par les expériences récentes qui ont permis de détecter la transcription inverse d'ARN synthétique ou messagers par la transcriptase inverse de L1 (Deragon *et al.*, 1990; Dhellin *et al.*, 1997). Les expériences déterminent que l'ORF2 est suffisant pour la synthèse d'ADNc. De plus un domaine minimum a été identifié et est contenu dans la partie centrale du gène qui comprend les 7 domaines

FIGURE E : Mécanisme de rétroposition tiré de l'étude des éléments SINE.

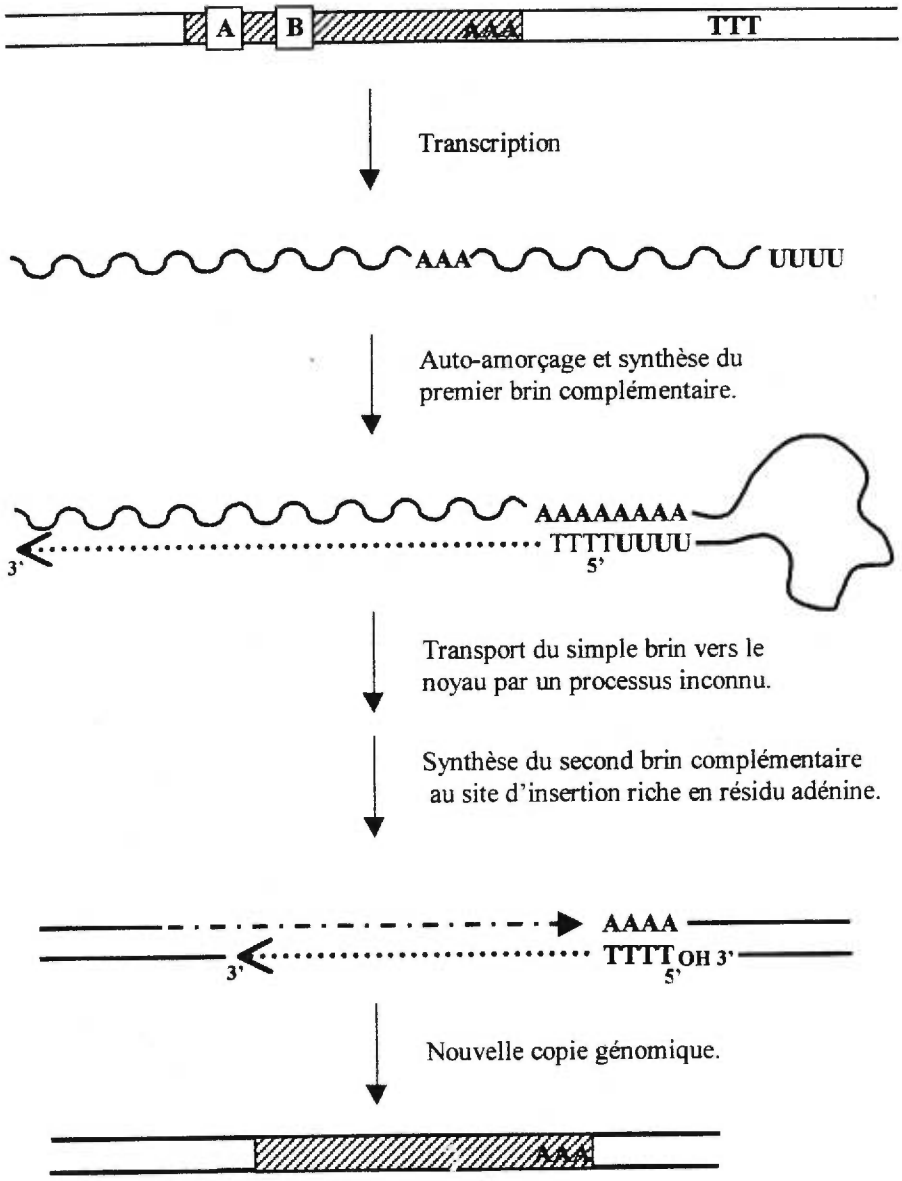


Figure établie à partir des références Jagadeeswaran *et al.* (1981) et Moos et Gallwitz (1983).

conservés chez toutes les transcriptases inverses. Dans ces expériences, la majorité des ADNc synthétisés est cytoplasmique et n'intègre pas le génome (Dhellin *et al.*, 1997). Ces expériences vont à l'encontre des conclusions du groupe de Kazazian (Moran *et al.*, 1996; Naas *et al.*, 1998) qui veulent que les protéines issues du transcrit des LINE agissent préférentiellement en *cis*. De la même façon, des expériences de complémentations menées chez la drosophile, avec l'élément I, proposent que la transriptase inverse peut être fournie en *trans* pour l'amplification du rétroposon (Busseau *et al.*, 1998). Ce modèle de *trans* activation des protéines issues du LINE L1 avait déjà été proposé pour expliquer l'amplification des rétroposons SINE *Alu* (Sinnett *et al.*, 1992).

Le second modèle propose que l'étape de transcription inverse s'effectue au site d'insertion du rétroposon. Ce modèle a été essentiellement construit à partir de l'étude des éléments LINE d'insectes, le facteur I de la drosophile (Bucheton, 1990) et le rétroposon R2 du ver à soie (Eickbush, 1992; Luan *et al.*, 1993). Les preuves directes de certaines étapes du mécanisme ont été établies à partir de l'étude de l'élément R2. L'unique ORF de cet élément code, entre autres, pour une endonucléase capable de cliver un site spécifique se trouvant dans le gène ribosomique 28S des insectes (Xiong et Eickbush, 1988a). Le transcrit de R2 est reconnu, à son extrémité 3', par le produit de son gène et possiblement par d'autres protéines pour former un complexe ribonucléoprotéique (Figure F) (Luan *et al.*, 1993). Ce complexe est ensuite dirigé vers le noyau par un mécanisme inconnu. Le site de clivage est préférentiellement reconnu par le domaine endonucléase codé par l'ORF et dans une première étape celle-ci coupe un seul brin de l'ADN cible. À partir de cette coupure va être initiée la synthèse du premier brin d'ADNc par la transcriptase inverse aussi codée par l'ORF. La synthèse complétée, le second brin du site est coupé et le second

FIGURE F : Mécanisme de rétroposition tiré de l'étude des éléments LINE.

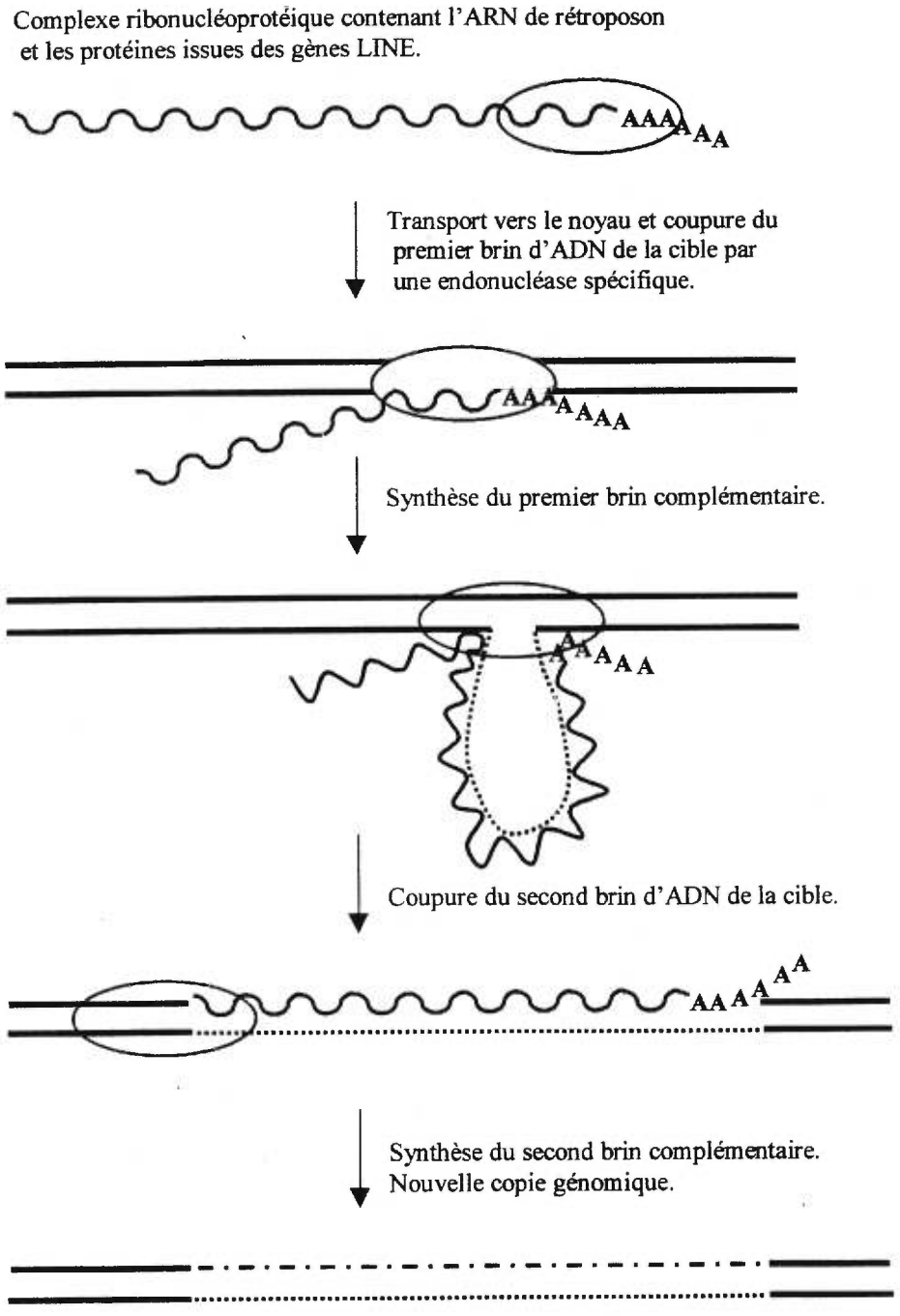


Figure établie à partir de la référence Luan *et al.* (1993).

brin de l'ADNc est synthétisé. Le temps qui sépare les deux coupures peut varier entre 30 et 120 minutes (Luan *et al.*, 1993). La distance qui sépare les deux coupures peut être variable, elle détermine la taille des répétitions directes flanquantes. Dans le cas des éléments rétroposons L1, *Alu*, B1 et ID, il a été démontré que la première coupure se fait préférentiellement après les deux thymines d'une séquence TTAAAA. Le second site de coupure est moins conservé mais il semble être préférentiellement situé juste après une séquence TYTN (Figure G)(Feng *et al.*, 1996; Jurka, 1997).

Dans les deux modèles présentés, la synthèse de l'ADNc est initiée à l'extrémité 3' de l'ARN du LINE ou du SINE. Cette observation permet d'expliquer le fait que la synthèse des ADNc des éléments LINE est rarement complète et qu'ils sont le plus souvent tronqués du coté 5'.

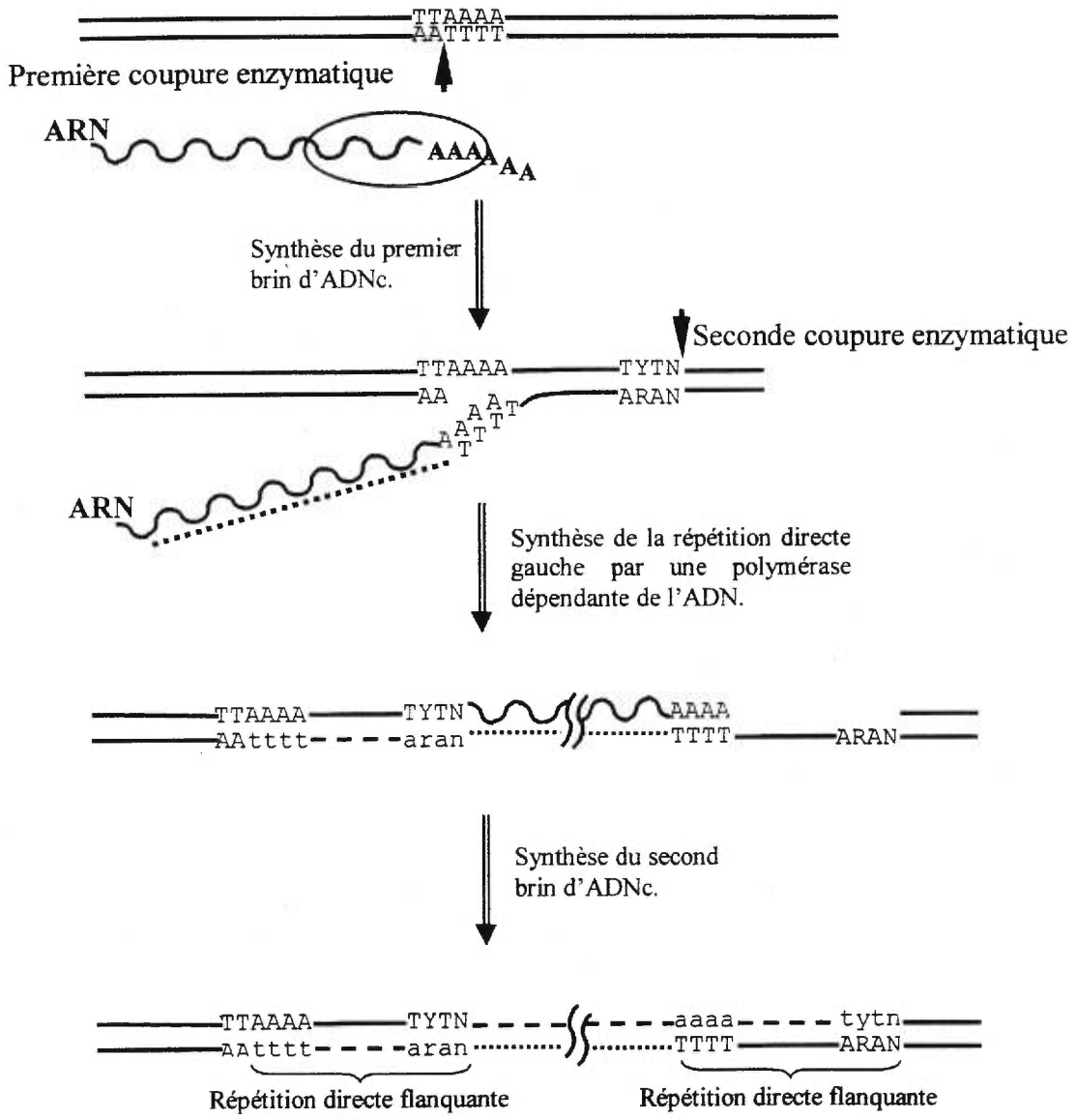
Cependant, les deux modèles proposés sont incomplets. L'étape de transfert vers le noyau est méconnue ainsi que les protéines qui participent à la rétroposition. Pour donner de nouveaux éléments de réponse, les études actuelles portent essentiellement sur les deux cadres de lecture des éléments LINE L1. Ces deux gènes possèdent de fortes similarités avec les deux premiers ORF des rétrovirus, *gag* et *pol*.

1-3.4- Les protéines impliquées dans la rétroposition.

La rétroposition est liée, chez les LINE, à l'activité enzymatique des protéines qui découlent des cadres de lecture. La présence d'éléments actifs "pleine longueur" dans le génome hôte est donc indispensable à l'amplification par rétroposition.

L'ORF1 correspond aux protéines gag des rétrovirus, mais aucune activité spécifique n'a encore été attribuée aux protéines qui en découlent car elles ne possèdent pas

FIGURE G : Modèle d'intégration des rétroposons L1, Alu, B1 et ID.



Complexe ribonucléoprotéique contenant l'ARN de rétroposon et les protéines issues des gènes LINE.

Figure établie à partir de la référence Jurka (1997).

d'homologie avec d'autres protéines connues (Holmes *et al.*, 1992). Cependant, par analogie avec les gènes *gag* des rétrovirus, il a été proposé que ORF1 code pour une protéine de liaison aux acides nucléiques qui jouerait un rôle dans la préparation d'une particule pseudo-virale pour la transcription inverse (Eickbush, 1992). Il est connu que l'intégrité du premier cadre de lecture est nécessaire à l'activité de rétroposition puisque des mutations dans sa séquence arrêtent la transposition (Moran *et al.*, 1996). De récents travaux ont démontré que le produit de l'ORF1 du LINE L1 est une protéine de 40 kDa, p40 (Leibold *et al.*, 1990; Bratthauer et Fanning, 1992). Elle forme par multimérisation un complexe ribonucléoprotéique cytoplasmique d'environ 200 kDa, contenant l'ARN du LINE L1 (Martin, 1991; Hohjoh et Singer, 1996; Hohjoh et Singer, 1997a). Il a été démontré que la protéine p40 se fixe de façon spécifique sur deux régions de l'ARN de l'élément LINE, situé dans le second cadre de lecture (Hohjoh et Singer, 1997b). Des résultats similaires ont été obtenus pour l'élément L1 homologue chez la souris (Kolosha et Martin, 1997). L'étude de l'ORF1 du rétroposon Tx1L, présent dans le génome du xénope, démontre aussi sa participation à la formation d'un complexe ribonucléoprotéique contenant le transcrit de l'élément LINE (Pont-Kingdon *et al.*, 1997). Cependant, le rôle du complexe contenant le produit de l'ORF1 dans le mécanisme de rétroposition n'a pas été identifié.

L'ORF2, analogue du cadre de lecture *pol* des rétrovirus, code pour les activités de transcription inverse et d'endonucléase. La première évidence directe d'une activité enzymatique portée par ORF2 a été fournie par l'étude de l'élément R2Bm du génome de *Bombix mori*. Cet élément possède un domaine endonucléase qui génère une coupure double brin sur un site unique d'insertion dans le gène de l'ARN ribosomique 28S (Xiong

et Eickbush, 1988a). D'autres éléments LINE semblent aussi posséder des sites d'intégration spécifiques : l'élément R1 s'intègre, comme R2, dans le gène ribosomique 28S (Jakubczak *et al.*, 1991), l'élément DRE, du génome de *Dictyostelium discoideum*, se retrouve en amont des gènes d'ARN de transfert (Marschalek *et al.*, 1992), le rétroposon L1Bm (*Bombyx mori*) a pour site préférentiel la portion 3' des gènes H2B (unité H2B des histones) (Ichimura *et al.*, 1997), les éléments HeT-A et TART (drosophile) ont pour site d'intégration les régions télomériques (Pardue *et al.*, 1996; Biessmann et Mason, 1997), et enfin les rétroposons Doc et F (drosophile) s'intègrent préférentiellement dans les régions de l'hétérochromatine (Pimpinelli *et al.*, 1995). Les deux LINE du xénope, Tx1L et Tx2L, semblent avoir pour site d'insertion préférentiel les transposons Tx1D et Tx2D, respectivement. Cependant, il n'a pas été décrit si l'amplification découle de plusieurs insertions des éléments LINE par rétroposition ou de la transposition des éléments Tx1D ou Tx2D avec un élément LINE intégré (Garrett *et al.*, 1989). En complément de ces observations, il a été récemment démontré que les éléments L1Hs possèdent un domaine endonucléase situé dans la région 5' de l'ORF2 (Feng *et al.*, 1996). Un domaine similaire a aussi été trouvé dans l'ORF1 de l'élément L1Tc (*Trypanosoma cruzi*) (Martin *et al.*, 1995). Ce domaine est similaire aux endonucléases apurinique/apyrimidique (AP endonucléase) (Martin *et al.*, 1995). Une revue des principaux LINE a permis d'identifier un domaine AP endonucléase dans le domaine 5' du deuxième cadre de lecture de chacun d'entre eux (Malik et Eickbush, 1998). L'analyse des sites d'insertions, par méthode directe en étudiant l'activité endonucléasique de l'ORF2, ou par méthode indirecte en recensant les intégrations de L1 dans les banques de données, révèlent un site préférentiel d'intégration des LINE L1 du génome humain, dont le consensus est le suivant; (Py)n|(Pu)n (le sigle |

représente le site de coupure) (Feng *et al.*, 1996). L'analyse des sites d'insertion des séquences *Alu* montrent aussi le même consensus TT|AAAA (Jurka et Klonowski, 1996; Jurka, 1997; Jurka *et al.*, 1998). Il est toutefois intéressant de noter que même si les deux rétroposons L1 et *Alu* possèdent le même site préférentiel d'insertion dans le génome, ils ne suivent pas forcément la même distribution sur les chromosomes. Les éléments *Alu* montrent une préférence pour les régions génomiques possédant des domaines de chromatine ouverts (Slagel *et al.*, 1987) alors que les LINE ont une préférence pour les domaines plus condensés (Holmquist et Caston, 1986; Korenberg et Rykowski, 1988). Cette différence peut être expliquée par une sélection négative des éléments LINE dans les régions riches en gènes (Deininger et Batzer, 1993).

L'activité de transcription inverse associée aux éléments L1 a été décrite pour la première fois dans des cellules humaines Ntera2D1 (Deragon *et al.*, 1990). Cette activité de transcription inverse, observée dans des expériences faites chez la levure par transfection d'éléments chimèriques Ty1/L1, pouvait provenir du deuxième cadre de lecture des éléments actifs LINE L1 (Mathias *et al.*, 1991). Ces deux dernières années, plusieurs travaux ont démontré, notamment par des expériences de transfections d'éléments LINE complets en aval d'un promoteur fort, que l'activité transcriptase inverse est issue de l'ORF2 de l'élément LINE (Moran *et al.*, 1996; Dhellin *et al.*, 1997; Busseau *et al.*, 1998; Naas *et al.*, 1998). Pour les éléments L1Hs, l'activité est plus particulièrement liée à la sous-famille L1Ta (Sassaman *et al.*, 1997).

1-4- L'origine des rétroposons chez les eucaryotes.

Les séquences répétées de type rétroposons ont été découvertes en premier lieu dans les génomes mammifères, et de ce fait ont été longtemps considérées spécifiques de ces génomes. Les 15 dernières années nous ont démontré que la rétroposition et les rétroposons sont distribués dans tous les groupes eucaryotes, allant des protozoaires jusqu'aux mammifères. Ces éléments rétroposons constituent donc des séquences dont les caractéristiques d'amplification ont été fortement conservées au cours de l'évolution.

1-4.1- Origine des éléments LINE.

Pour déterminer l'origine des rétroposons de type LINE, différents chercheurs ont utilisé la séquence protéique de la transcriptase inverse présente dans le deuxième cadre de lecture de ces éléments (Xiong et Eickbush, 1988b; Eickbush, 1997; Nakamura et Cech, 1998). Cette séquence présente l'avantage de posséder 7 domaines fortement conservés entre toutes les transcriptases inverses des éléments LINE, mais aussi entre celles d'autres éléments que sont les rétrovirus, les rétrotransposons, les télomérases, les introns du groupe II et les transcriptases inverses associées aux copies multiples d'ADN simple brin des bactéries (ms-DNA) (Xiong et Eickbush, 1988b; Xiong et Eickbush, 1990). De l'analyse phylogénétique des 7 domaines conservés découlent deux principales hypothèses sur l'origine des éléments possédant une transcriptase inverse. La première hypothèse, qui utilise comme racine de l'arbre phylogénétique les ARN polymérases dépendantes de l'ARN, place les rétrotransposons à l'origine de tous les rétrovirus, et la télomérase à l'origine des rétroposons, des intron du groupe II et des ms-DNA. La seconde hypothèse

utilise comme racine de l'arbre les éléments procaryotiques, et place à l'origine des éléments mobiles eucaryotes les rétroposons. À l'heure actuelle il est toujours impossible de déterminer l'origine exacte des rétroposons de type LINE. Cependant les analyses phylogénétiques et la présence de ces éléments dans l'ensemble des génomes eucaryotes favorise l'hypothèse selon laquelle les rétroposons prendraient leur origine chez les premiers eucaryotes (Eickbush, 1997; Nakamura et Cech, 1998).

1-4.2- Origine des éléments SINE.

Les éléments SINE ne possèdent aucune séquence codante. De ce fait, ils sont obligatoirement dépendants d'une source externe de transcriptase inverse. L'origine des éléments doit donc forcement être postérieure à un des éléments cités dans le paragraphe précédent, et notamment aux éléments LINE.

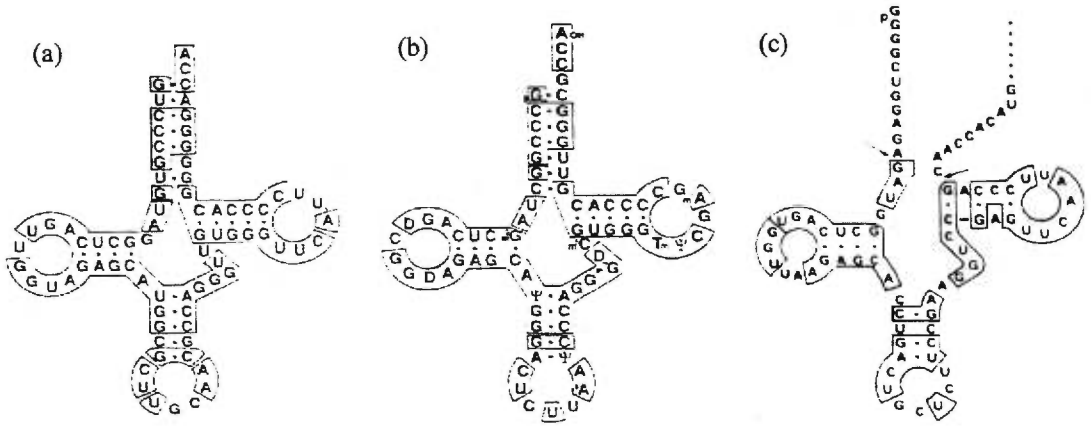
Tous les éléments SINE possèdent des structures communes, décrites dans le paragraphe 1-2.3-, qui ont permis leur amplification. Malgré cette conservation structurale, les SINE peuvent être classés en deux groupes. Le premier groupe comprend la majorité des éléments SINE, et est constitué des séquences dérivant des ARN de transfert ; le second groupe comprend les éléments dont l'origine est autre. À la fin de ce chapitre, un tableau regroupe les éléments SINE découverts à ce jour, et propose leurs origines (Tableau B, page 64).

1-4.2.1- Les SINE dérivés des ARNt.

Comme cela a été décrit précédemment, cette catégorie regroupe la grande majorité des éléments SINE. Dès 1985, plusieurs auteurs décrivent les premières familles

SINE dans les génomes mammifères. Pour une famille, provenant des génomes des ruminants (bovin, ovin, caprin), il a été possible de reconstruire, pour le segment 5' de la séquence, la structure secondaire en trèfle des ARNt (Rogers, 1985). De cette façon, un certain nombre de segments 5' de rétroposons SINE ont été associés à des ARNt spécifiques. Par exemple la famille SINE *Alu* type II du génome des galagos est associée à l'ARNt de la méthionine (Daniels et Deininger, 1985). À la même époque, Sakamoto et Okada présentent les identités de plusieurs SINE tels que B2, ID, C et Bov-tA, avec différents ARNt tels que l'ARNt lysine, l'ARNt phénylalanine et l'ARNt glycine, et peuvent rétablir dans une certaine mesure la structure secondaire des ARNt avec ces SINE (Sakamoto et Okada, 1985; Okada, 1991b). Deux exemples de reconstruction de la structure secondaire des promoteurs de SINE dérivés d'ARNt sont présentés sur la figure H. L'origine des rétroposons SINE dérivés d'ARNt n'est en fait pas toujours facile à démontrer. Lorsque la structure en trèfle est reconstruite, l'énergie d'appariement des bases pour constituer les structures est plus faible que pour les ARNt. Cette différence pourrait s'expliquer par le fait que les éléments SINE ne subissent pas de pression de sélection et sont sujets à l'évolution neutre, contrairement aux ARNt (Weiner *et al.*, 1986). En effet, pour certains éléments, la connexion avec les ARNt se limite essentiellement à la présence des boîtes A et B du promoteur interne de l'ARN polymérase III, dont la conservation est indispensable pour une transcription efficace. Par ailleurs, il a été démontré que malgré sa similarité avec les ARNt (Daniels et Deininger, 1985), l'élément ID présent dans le génome des rats a pour origine l'ARN BC1 (DeChiara et Brosius, 1987; Deininger et Batzer, 1995).

FIGURE H : Structure secondaire possible des promoteurs de SINE dérivés d'ARNt.



(a) et (c) : Représentation en feuille de trèfle des segments dérivés de l'ARNt lysine des rétroposons SINE *Fok1* (poissons) et B2 (rongeurs).

(b) : Structure secondaire de l'ARNt lysine.

Les nucléotides Ψ, m^2G, m^5G, m^7G, m^1A, t^6A et Tm sont des nucléotides modifiés des ARNt. Les nucléotides conservés sont encadrés.

Figure établie à partir des références Sakamoto et Okada (1985) et Okada (1991b).

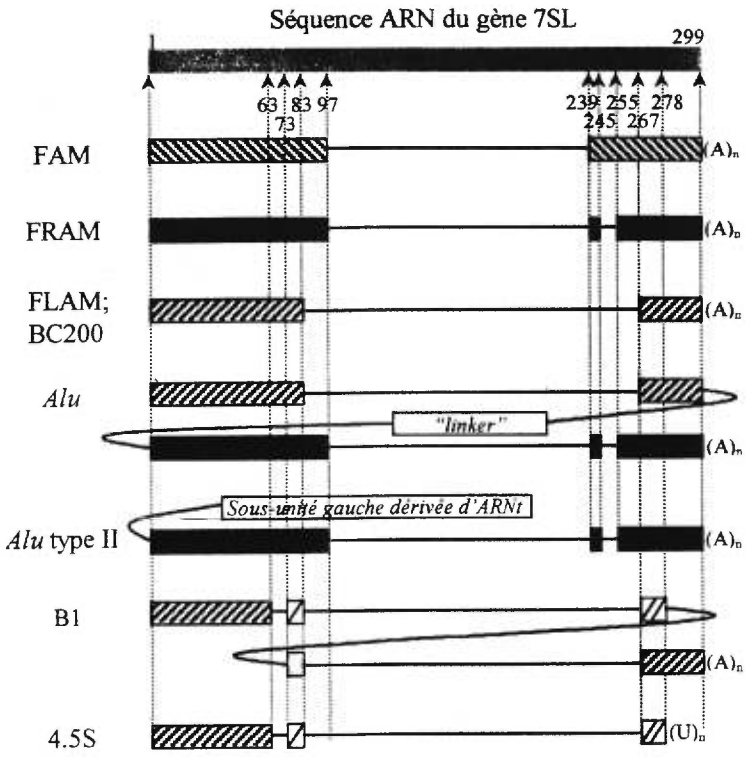
1-4.2.2- Les autres SINE.

Pour tous les éléments SINE ne dérivant pas d'ARNt, il a été possible de déterminer l'origine structurale exacte. C'est le cas des SINE *Alu*, B1, ID, 4,5SI et Bov-A2.

1-4.2.2.1- Les familles *Alu* des primates.

La famille *Alu* est à l'origine de la découverte et de la caractérisation des séquences répétées que l'on appelle aujourd'hui rétroposons. Les expériences de cinétique de renaturation de l'ADN humain suggéraient l'existence d'une répétition de 300 nucléotides représentant près de 6% du génome total (Houck *et al.*, 1979; Jelinek *et al.*, 1980) (Rinehart *et al.*, 1981). Tous les membres de cette répétition possédaient le site de restriction pour l'enzyme *AluI* (Houck *et al.*, 1979), ce qui a donné le nom à la séquence. Un certain nombre de séquences ont été obtenues et il a été possible de créer un premier consensus de la famille des éléments *Alu* du génome humain (Deininger *et al.*, 1981). Ce consensus a permis de déterminer que l'origine de la séquence provient du gène 7SL (Ullu *et al.*, 1982). Des études plus approfondies sur la structure des séquences *Alu* ont permis de déterminer que ces éléments ont une structure dimérique. Les deux sous-unités, gauche et droite, sont fortement homologues mais non identiques. Elles sont reliées entre elles par une séquence de liaison riche en résidus adénine, appelée "linker". Les deux sous-unités sont dérivées de l'ARN du gène 7SL ayant subit une délétion du segment interne, délétion différente pour chacune des deux sous-unités (Figure I) (Ullu et Tschudi, 1984). La forme dimérique est la plus présente dans les génomes de primate. Cependant, les deux sous-unités existent à l'état monomérique et sont appelées FLAM et FRAM (pour Free Left ou Right Alu Monomers). Ces deux sous-unités monomériques sont considérées comme étant

FIGURE I : Sequences SINE dérivées du gène 7SL.



Les blocs désignent les segments de séquence homologue à l'ARN 7SL. Les lignes qui relient les blocs indiquent les délétions et les dimérisations. Les lignes verticales ainsi que les chiffres correspondent aux différents points de cassure par rapport à l'ARN 7SL. Les noms des familles SINE sont indiqués sur la gauche.

Figure établie à partir de la référence Labuda et Zietkiewicz (1994).

les précurseurs de la famille *Alu* (Jurka et Zuckerkandl, 1991; Quentin, 1992a) et proviennent toutes deux d'un unique précurseur, le monomère FAM (Fossil Alu Monomer) qui dérive de l'ARN 7SL (Quentin, 1992b).

Dans le génome des galagos, une autre famille SINE dérivée des éléments *Alu*, nommée *Alu* type II, a été décrite. La séquence de ces éléments possède la sous-unité droite de l'élément *Alu*, soit l'équivalent de FRAM, précédée d'une unité pouvant dériver de l'ARNt méthionine (Figure I) (Daniels et Deininger, 1983; Daniels et Deininger, 1985; Rogers, 1985; Slagel et Deininger, 1989). L'identité entre la nouvelle unité gauche de la séquence *Alu* type II avec l'ARNt méthionine est proche de 68%. Cette unité gauche, encore appelée "monomère", a aussi été retrouvée seule dans le génome des prosimiens. Ceci suggère qu'elle a également été un rétroposon SINE indépendant (Daniels et Deininger, 1991). La famille *Alu* type II serait apparue après ses deux fondateurs, sans doute par l'insertion d'un "monomère" dans la séquence de liaison d'un élément *Alu* (Daniels et Deininger, 1983; Daniels et Deininger, 1991). Il est à noter que ces deux dernières familles décrites, *Alu* type II et "monomère", appartiennent au groupe des SINE dérivés d'ARNt (Tableau B, page 64).

1-4.2.2.2- La famille B1.

Il existe dans les génomes des rongeurs de nombreux rétroposons différents. Parmi ceux-ci le SINE B1 est présent chez tous les rongeurs. Cette famille est apparentée à la famille *Alu* des génomes des primates (Krayev *et al.*, 1980; Haynes et Jelinek, 1981). Cette relation est due à une origine commune aux deux familles, le gène 7SL. Le SINE B1 est lui aussi dérivé de l'ARN du gène 7SL ayant subit une délétion de

son segment central (Ullu et Tschudi, 1984). La différence avec les éléments *Alu* est que les séquences B1 sont des monomères dont la structure est proche du monomère FLAM. Une délétion de 10 nucléotides et une duplication de plus de 20 nucléotides différencient les éléments B1 de FLAM (Figure I) (Labuda *et al.*, 1991; Labuda et Zietkiewicz, 1994).

1-4.2.2.3- La famille 4,5SI.

Cette famille est spécifique au génome du rat et est fortement homologue au gène 4,5S, qui est un snRNA (Saba *et al.*, 1985; Takeuchi et Harada, 1986). Il faut noter aussi que le gène 4,5S dérive lui même de l'ARN du gène 7SL. La structure du gène est fortement homologue à la famille B1 (Figure I) (Labuda et Zietkiewicz, 1994).

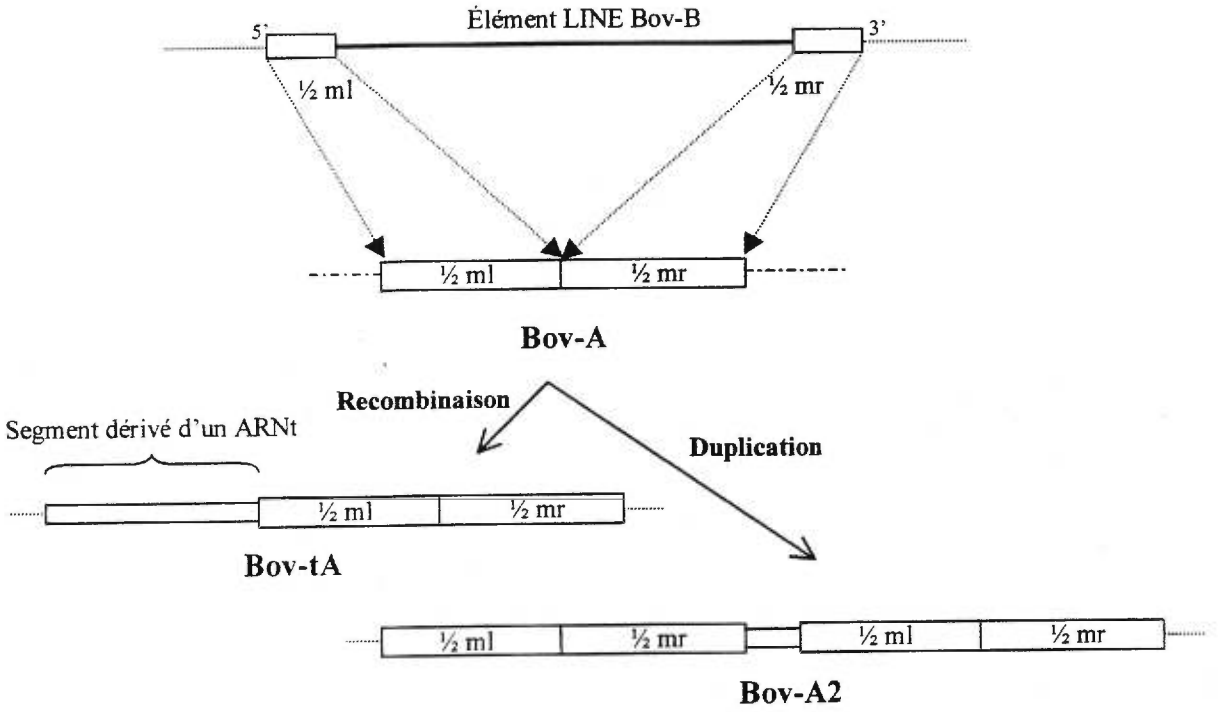
1-4.2.2.4- La famille ID.

Une autre famille SINE des rongeurs, ID, a été décrite initialement dans le génome des rats. Ces séquences ont été caractérisées à partir de régions non traduites d'ARN spécifiques du cerveau, d'où leur nom ID pour séquence identificatrice de la spécificité tissulaire (Sutcliffe *et al.*, 1982). Plusieurs ARNt, tels que l'ARNt alanine ou l'ARNt phenylalanine, ont été proposés comme étant à l'origine de la séquence ID (Sakamoto et Okada, 1985; Daniels et Deininger, 1985; Lawrence *et al.*, 1985). Il semble cependant que la véritable origine de la famille est l'ARN du gène BC1 (Matera *et al.*, 1990; Deininger et Batzer, 1995).

1-4.2.2.5- Les familles Bov-A.

Les premières études sur le génome des bovins laissaient supposer l'existence de plusieurs familles de courtes séquences répétées (Mayfield *et al.*, 1980). Il a été trouvé un segment de 115 nucléotides fortement répété chez les *Bovidae* (Watanabe *et al.*, 1982). Ce segment se retrouve soit à l'état dimérique, les deux unités étant reliées par une séquence riche en résidus thymine, soit précédé par une séquence dérivée d'ARNt glycine (Figure J) (Watanabe *et al.*, 1982; Sakamoto et Okada, 1985; Spence *et al.*, 1985; Zelnick *et al.*, 1987). Le monomère a été appelé BCS (Spence *et al.*, 1985), BMF (Zelnick *et al.*, 1987) ou encore Bov-A (Lenstra *et al.*, 1993), et c'est cette dernière nomenclature qui a été retenue. Les deux familles SINE sont donc appelées Bov-tA, pour ARNt suivi du monomère, et Bov-A2 pour le dimère (Lenstra *et al.*, 1993). Le segment Bov-A n'a pas été retrouvé à l'état unique dans les génomes (Lenstra *et al.*, 1993). Un autre élément rétroposon existe chez les artiodactyles, il est appelé Art-2 (Duncan, 1987), *Pst*I (Majewska *et al.*, 1988), ou encore Bov-B (Lenstra *et al.*, 1993). À l'origine, ce rétroposon était considéré comme appartenant aux SINE. Cependant, sa structure "pleine longueur" a été décrite récemment et il en a été conclu que les premières séquences analysées sont en fait des éléments tronqués d'une nouvelle famille de rétroposons LINE (Szemraj *et al.*, 1995). Cette famille, Bov-B (aussi appelé BDDF), serait à l'origine du monomère Bov-A. En effet ce dernier est issu de la délétion de toute la partie centrale de la séquence LINE Bov-B, ne laissant qu'une soixantaine de nucléotides provenant du coté 5' et autant de l'extrémité 3' (Figure J) (Szemraj *et al.*, 1995; Okada et Hamada, 1997a).

FIGURE J : Origine des éléments SINE bovins.



L'élément Bov-B a une taille d'environ 3,1 Kb. Il est constitué à ses extrémités par les segments ½ml et ½mr (½ monomer left et ½ monomer right). La délétion de la partie centrale de l'élément, schématisée par les flèches, a donné naissance au monomère Bov-A. Le monomère a créé le SINE Bov-tA, en s'associant avec une séquence dérivée d'un ARNt, et le SINE Bov-A2 par duplication. Le lien entre les deux monomères de Bov-A2 est une séquence riche en résidu T de 25 nucléotides qui pourrait dériver d'une répétition simple $(CACTTT)_4$.

Figure établie à partir des références Szemraj *et al.* (1995) et Okada et Hamada (1997a).

1-4.2.3- Le lien avec les LINE.

Pour la grande majorité des rétroposons SINE dérivés d'ARNt, l'origine n'est décrite que pour le segment 5', promoteur de la transcription par l'ARN polymérase III. Des travaux récents ont mis en évidence de fortes identités entre les SINE et les LINE. En effet, un certain nombre de segments 3' de séquences SINE possèdent de fortes identités avec l'extrémité 3' des séquences LINE (Vandergon et Reitman, 1994; Ohshima *et al.*, 1996; Okada et Hamada, 1997a; Okada *et al.*, 1997b). La taille de ces identités est de quelques dizaines de nucléotides (Tableau C, voir à la fin du chapitre page 66). Un rôle peut être attribué à cette identité entre les LINE et les SINE. D'après le modèle de rétroposition proposé à partir de l'étude de différents LINE, présenté dans le paragraphe 1-3.5-, où la transcription et l'intégration se fait sur le site d'insertion, le lien avec les LINE serait le moyen pour les SINE de fixer à l'extrémité 3' de son ARN le complexe pseudoviral et ainsi induire son amplification (Ohshima *et al.*, 1996) (Figure K). Pour les éléments *Alu* et B1, l'efficacité de rétroposition n'est pas associée à une identité de séquence avec les LINE, mais semble être liée à la conservation de la structure secondaire de l'ARN, très proche de celle des ARN du gène 7SL (Labuda *et al.*, 1991; Sinnett *et al.*, 1991; Labuda et Zietkiewicz, 1994). Les ARN ayant conservé leur structure secondaire participent au complexe SRP qui fixe les ribosomes. Les protéines des éléments LINE peuvent ainsi reconnaître l'extrémité poly-A des éléments *Alu* fixés sur le ribosome qui les traduit, c'est l'hypothèse de la "polyA connection" (Boeke, 1997).

L'origine de deux segments, 5' et 3', pour un certain nombre de rétroposons SINE dérivés des ARNt a été décrite dans les paragraphes précédents. Il reste cependant, pour un certain nombre d'éléments, un segment central dont l'origine est inconnue. Il a été

FIGURE K : Rétroposition des éléments SINE possédant une identité avec les LINE.

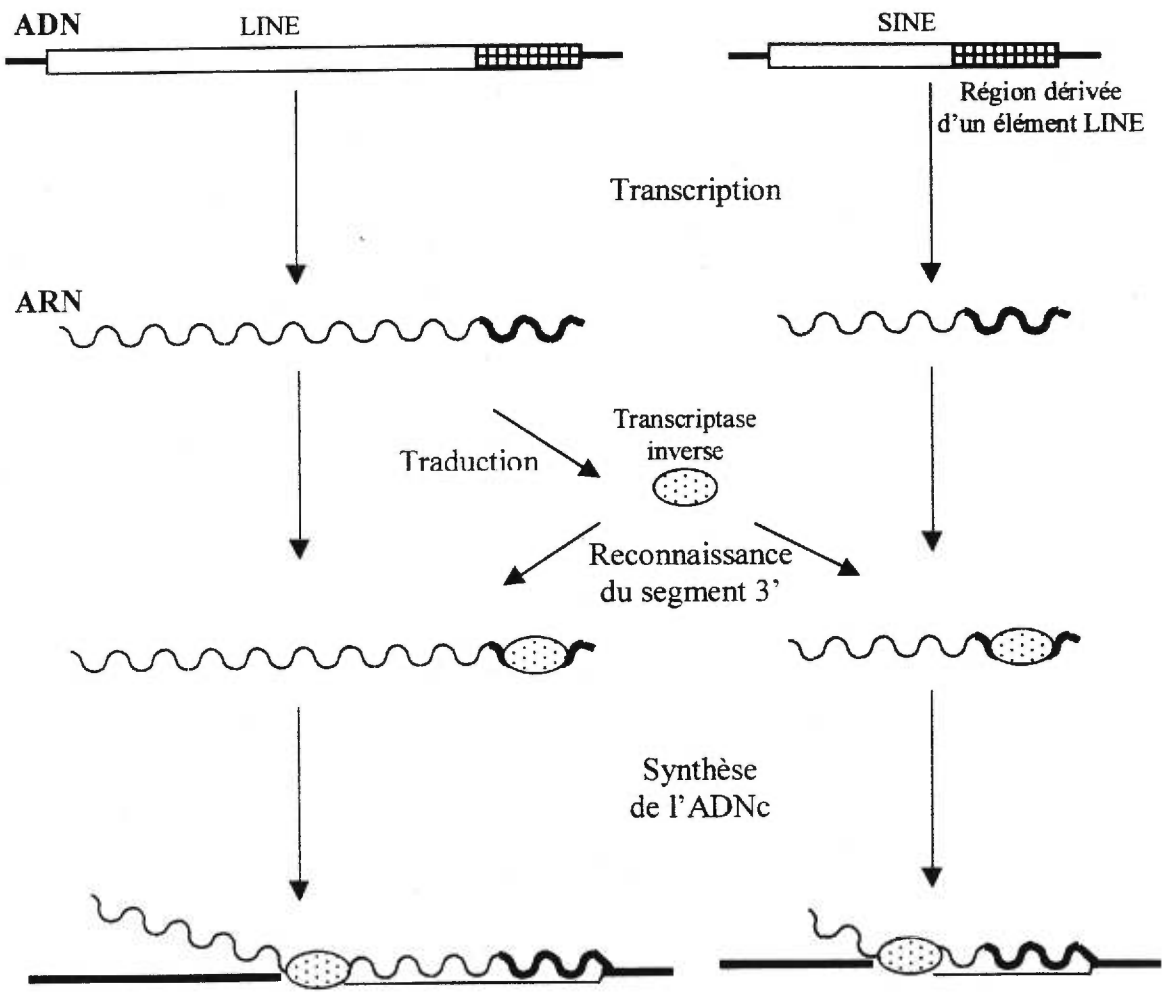


Figure établie à partir de la référence Ohshima *et al.* (1996).

proposé que ceux-ci proviennent du segment "strong stop" créé lors de l'initiation de la transcription inverse des éléments rétrotransposons (Figure L) (Ohshima *et al.*, 1993). L'analyse de cinq éléments SINE dérivés de l'ARNt lysine, FokI, SmaI, SK, B2 et TORT, décèle la présence de deux courts fragments conservés de 6 et 3 nucléotides dans le segment central. Ces fragments sont retrouvés peu conservés dans les régions U5 de LTR de différents rétrovirus, lesquels utilisent l'ARNt lysine pour amorcer la transcription inverse. Ce modèle reste cependant très fragile, car il n'est basé que sur l'identité d'à peine 10 nucléotides sur un segment qui peut comprendre plus de 80 nucléotides.

1-5- Le rétroposon MIR.

Le modèle que nous avons utilisé au laboratoire pour l'étude de la rétroposition chez les mammifères est l'élément MIR (Mammalian-wide Interspersed Repeat).

Peu de choses sont connues à l'heure actuelle sur le rétroposon MIR. En 1987, Degen et Davie découvrent une courte séquence répétitive d'environ 100 nucléotides dans le gène de la prothrombine chez l'humain (Degen et Davie, 1987). Un premier consensus de cette répétition, comprenant 70 nucléotides, est établi en 1989. Cependant, il ne peut être associé à aucun groupe de séquences répétées connues (Donehower *et al.*, 1989). En 1995, les expériences d'inter-MIR PCR (voir paragraphe 2-2.12-), faites dans notre laboratoire, démontrent que cette séquence répétée est présente dans tous les génomes mammifères (placentaires, marsupiaux et monotrèmes), d'où son nom MIR pour "Mammalian-wide Interspersed Repeat" (Jurka *et al.*, 1995). Au même moment Smit et Riggs (1995) complètent le premier consensus, à partir de chaque extrémité, pour identifier la structure

FIGURE L : Origine possible des segments centraux des SINE dérivés d'ARNt.

Initiation de la transcription inverse d'un rétroélément de type viral.

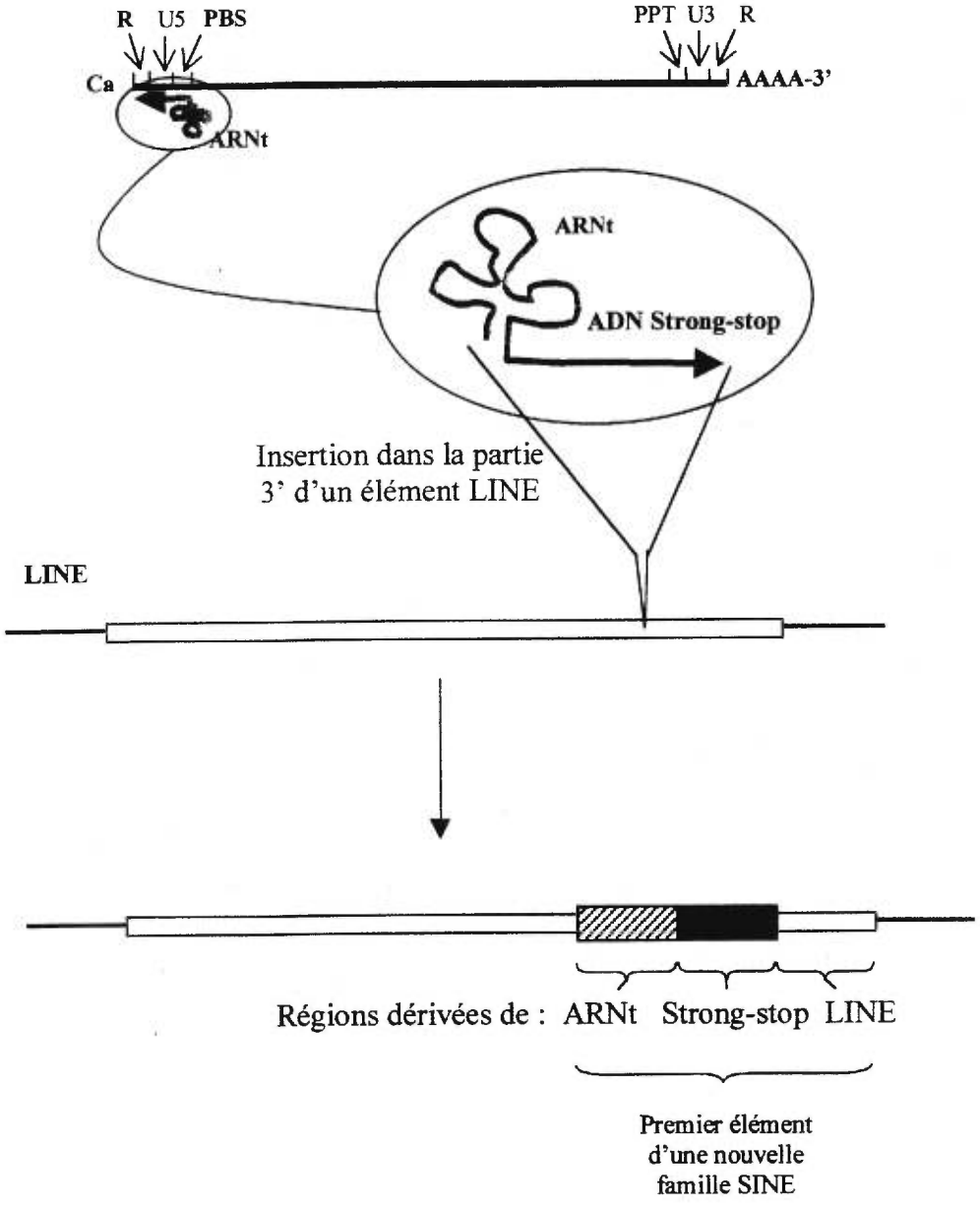
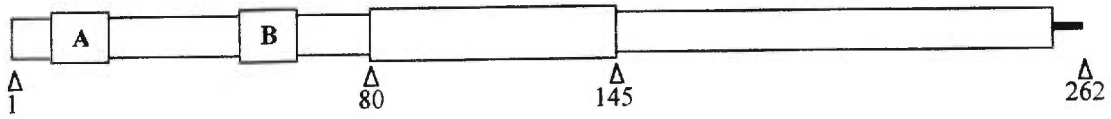


Figure établie à partir des références Ohshima *et al.* (1993) et Ohshima *et al.* (1996).

"pleine longueur" de la répétition. Ce second consensus, établi à partir de plus de 400 séquences provenant essentiellement du génome humain, décrit un élément de 262 nucléotides qui appartient à la famille des rétroposons de type SINE (Smit et Riggs, 1995). L'élément MIR possède un promoteur à ARN polymérase III identifié par les boîtes A et B (Figure D et M). Ce promoteur pourrait dériver d'un ARNt mais celui-ci ne peut être identifié avec exactitude. Un certain nombre d'éléments possèdent à leur extrémité 3' une répétition simple de type TTA. Le segment 3' des éléments possède une forte identité avec le segment 3' des éléments LINE L2 des génomes mammifères (Smit et Riggs, 1995; Smit, 1996). Le domaine central de MIR de 65 nucléotides correspond à la répétition identifiée par Donehower en 1989. C'est la région la mieux conservée entre les éléments et est appelée "core".

Toutes les séquences identifiées sont fortement divergentes par rapport à leur consensus (entre 25 et 35 %) et sont considérées, chez les mammifères placentaires, comme fossiles moléculaires, donc inactives (Jurka *et al.*, 1995). Du fait de leur ancienneté aucune séquence répétée directe n'a été identifiée pour déterminer une duplication du site d'insertion (Smit et Riggs, 1995). Le nombre total d'éléments MIR chez l'humain a été estimé, à partir des banques de données, entre 120 000 et 400 000 copies (Smit et Riggs, 1995; Smit, 1996). Leur distribution à travers le génome est aléatoire mais ne semble pas uniforme. Comme pour le retroposon *Alu*, les éléments MIR sont retrouvés à plus forte densité dans les régions riches en gènes (Matassi *et al.*, 1998). Quelques éléments ont été identifiés comme faisant partie de séquences codantes (Murnane et Morales, 1995).

FIGURE M : Structure schématisée du rétroposon MIR.



région promoteur dérivée d'un ARNt

domaine core 65 pb

segment dérivé du LINE L2

Répétition Simple (TTA)

1-6- Impact de la rétroposition sur la variabilité des génomes.

Les éléments mobiles des génomes ont été considérés, dès leur découverte dans les années 80, comme des séquences égoïstes, des parasites génomiques qui consomment de l'énergie pour leur amplification. Cependant, il est de plus en plus observé que les séquences mobiles hautement répétées peuvent avoir un rôle important dans la variabilité et la plasticité des génomes. Ce rôle est induit par deux mécanismes majeurs qui sont, l'insertion d'un élément par la rétroposition, et la recombinaison homologue, légitime ou non, induite par le grand nombre de copies fortement identiques.

Il a aussi été observé que certains rétroposons ont acquis des fonctions indispensables à la survie des cellules. En effet les rétroposons LINE HeT-A et TART des génomes des drosophiles ont remplacé la fonction de la télomérase qui n'est plus fonctionnelle (Pardue *et al.*, 1996; Biessmann et Mason, 1997). Ces deux rétroposons s'intègrent de façon spécifique dans les régions télomériques des chromosomes et permettent ainsi leurs extensions.

1-6.1- Effets directs des insertions de rétroposons dans les génomes.

La majorité des insertions des rétroposons se fait dans des régions non traduites telles que les introns ou l'hétérochromatine, elles n'ont donc pas en général d'effet dramatique sur la cellule hôte. Cependant, il arrive que l'intégration des éléments mobiles créent des perturbations génomiques. En effet, les éléments transposables peuvent s'insérer dans des séquences codantes et provoquer des mutations qui induisent des maladies génétiques (pour

les maladies induites par l'élément *Alu* voir article de revue Labuda *et al.*, 1995). Il existe, à l'heure actuelle, 19 cas recensés d'insertion d'élément *Alu* ou L1 dans le génome humain qui ont provoqué des maladies génétiques (Kazazian et Moran, 1998). Par exemple un cas d'hémophilie B induite par l'insertion d'une séquence *Alu* dans l'exon 5 du gène du facteur IX détruisant ainsi le cadre de lecture (Vidaud *et al.*, 1993). Ou encore l'insertion d'un élément L1 dans le gène du facteur VIII provoquant une hémophilie A (Kazazian *et al.*, 1988). Le rétroposon *Alu* peut aussi créer des dysfonctionnements de certains gènes lorsqu'il est intégré dans des introns, notamment en orientation inverse, en induisant un nouveau site d'épissage (Makalowski *et al.*, 1994).

Mais les insertions ne sont pas uniquement impliquées dans des modifications défavorables pour le génome hôte. Il a été démontré que la présence de rétroposons dans les parties codantes de certains gènes participent à l'évolution des protéines. En effet différents éléments LINE, comme L1, ou SINE, comme B2 (rongeur), C (lapin) et MIR (mammifère), ont généré des signaux de polyadénylation aux extrémité 3' de gènes (Rothkopf *et al.*, 1986; Harendza et Johnson, 1990; Krane et Hardison, 1990; Murnane et Morales, 1995). Il a aussi été observé le cas d'un gène humain, *RMSA-1*, possédant 2 éléments *Alu* "pleine longueur" représentant environ 40% du cDNA. Ce gène a un rôle dans le contrôle du cycle cellulaire. La queue poly-A du premier élément *Alu* a fourni un signal de localisation nucléaire indispensable à la fonction du gène (Margalit *et al.*, 1994).

1-6.2- Effet de la recombinaison induite par les rétroposons.

Chaque famille de rétroposons est, en général, représentée dans les génomes par 50000 à 400000 copies, exception faite pour l'élément *Alu* qui est présent à environ 1

million de copies. Ces répétitions, distribuées à travers les génomes, favorisent les recombinaisons homologues. Les recombinaisons illégitimes entre deux éléments qui ne sont pas situés sur la même position allélique peuvent induire des délétions et des duplications voir même des translocations lorsque la recombinaison a lieu entre deux chromosomes différents. Un certain nombre de ces événements a créé des maladies génétiques comme par exemple l'hyper-cholestérolémie familiale qui est induite par différentes recombinaisons entre les séquences *Alu* présentes dans différents introns du gène du récepteur des LDL (Low Density Lipoprotéine receptor). Les recombinaisons entre les séquences *Alu* sont favorisées par la présence d'une structure core de 26 nucléotides qui est considérée comme point chaud de recombinaison. Ce core possède une forte identité avec la structure *chi* des procaryotes qui stimule la recombinaison médiée par *rec*BC chez *E. coli* (Rudiger *et al.*, 1995). Les principales maladies induites par la présence des séquences *Alu* ont été regroupées dans un chapitre de livre écrit par Labuda *et al.* (1995).

Comme les insertions, les recombinaisons illégitimes n'ont pas forcément d'effets négatifs sur le génome. Certains événements sont neutres et participent à la réorganisation des génomes au cours de l'évolution. Par exemple, la recombinaison entre 2 éléments L1 chez l'Homme est responsable de l'inversion d'un segment du chromosome Y qui le différencie de celui des autres primates (Schwartz *et al.*, 1998).

1-7- Les Objectifs.

Les rétroposons sont les éléments mobiles les plus représentés dans le génome humain (près de 35% de la masse génomique). Cependant, il est difficile d'évaluer

l'importance de ces séquences dans l'organisation du génome et dans son évolution. Comme nous l'avons vu dans le paragraphe précédent, même si les éléments sont inactifs d'un point de vue rétropositionnel, ils peuvent intervenir par des mécanismes divers, notamment par recombinaison homologue, sur la structure du génome.

Malgrés de nombreuses informations accumulées au cours de ces dernières années, peu de choses sont encore connues sur l'origine des rétroposons de type SINE et sur leur évolution dans les génomes eucaryotes. Nous avons donc décidé d'utiliser comme modèle d'étude un ancient élément répété du génome humain, MIR, conservé à l'état de fossile moléculaire, pour apporter des compléments d'information.

Dans un premier temps nous allons utiliser des représentants des génomes mammifères placentaires, marsupiaux et monotrèmes pour vérifier si MIR possède effectivement toutes les caractéristiques d'un rétroposon de type SINE. Nous établirons d'ailleurs que ce dernier fait partie d'une "super-famille" de retroposons que nous appellerons CORE-SINE.

Une fois la structure établie, le deuxième objectif de notre travail sera de déterminer l'origine des éléments que nous aurons caractérisés. Nous établirons séparément l'origine des différents segments, 5', core et 3', qui composent les répétitions CORE-SINE. Nous nous interrogerons sur le rôle de chacun des segments dans le mécanisme de la retroposition.

Finalement, nous essayerons d'apporter des informations supplémentaires sur l'origine des rétroposons et sur le mécanisme général de rétroposition. Nous tenterons de confirmer l'importance de la rétroposition dans l'organisation et l'évolution des génomes eucaryotes.

TABLEAU A : Les éléments LINE.

Noms	espèces	groupe	références
L1	mammifères	mammifère	259
L2	mammifères		262
Bov-B	ruminants		274
Bov-B	reptiles		140
Bov-B	marsupiaux		(cette thèse)
CR1	oiseaux		252
CR1-like	amphibiens, poissons		34
CR1-like	reptiles		288
PsCR1	tortues		203
Tx1L	*Xenopus laevis*	amphibien	84
RSg1	*Salmonidae*	poisson	300
Eel LINE-like	*Eel*		203
ZEB. LINE	*Danio rerio*		209
Ci LINE 2	*Cichlidae*		282
SW1	teleost		68
BGR2	*Biomphalaria glabrata*	mollusque	136
R1	insectes	insecte	307
R2	insectes		35
Dong	*Bombyx mori*		309
TRAS1	*Bombyx mori*		210
L1Bm	*Bombyx mori*		106
?	*Bombyx mori*		209
I	*Drosophila melanogaster*		74
F	*Drosophila melanogaster*		62
Jockey	*Drosophila melanogaster*		220
G	*Drosophila melanogaster*		61
Doc	*Drosophila melanogaster*		202
Het-A	*Drosophila melanogaster*		47
TART	*Drosophila melanogaster*		249
LOA	*Drosophila silvestris*		75
bilbo	*Drosophila subobscura*		19
T1	*Anopheles gambiae*		14
Q	*Anopheles gambiae*		15

JuanA	*Aedes aegypti*	insecte	193
NLR1Cth	*Chironomus thummi*		20
Sam	*Caenorhabditis elegans*	nematode	172
Frodo	*Caenorhabditis elegans*		172
SR1	*Schistosoma mansoni*	ver plat	65
Cin4	*Zea mays*	plante	247
Ta11-1	*Arabidopsis Thaliana*		304
LINE potentiel	tomate et pomme de terre		203
del2	*Lilium speciosum*		158
BNR1	*Beta vulgaris*		244
LINE potentiel	mono- et dicotylédone		147
Zeep	*Chlorella vulgaris*	algue	95
Tad	*Neurospora crassa*	champignon	134
Ch	*Magnaporthe grisea*		122
CgT1	*Glomeralla cingulata*		pas d'article publié
DRE	*Dictyostelium discoideum*		173
Ingi	*Trypanosoma brucei*	protozoaire	133
SLACS	*Trypanosoma brucei*		4
CZAR	*Trypanosoma cruzi*		292
L1Tc	*Trypanosoma cruzi*		174
CRE1	*Crithidia fasciculata*		81

TABLEAU A (suite)

TABLEAU B : Les éléments SINE.

Noms	espèces	groupe	Origine possible du segment 5'	références
Ther-1 (MIR)	mammifères	mammifère	ARNt	264
Ther-2	thériens		ARNt	(cette thèse)
Mar-1	marsupiaux		ARNt	(cette thèse)
Opo-1	opossum		ARNt	(cette thèse)
Mon-1	monotrèmes		ARNt	(cette thèse)
Alu	primates		7SL	54
Alu type II	procimiens		ARNt (Lysine) et 7SL	44
"monomère"	procimiens		ARNt	228
B1	rongeurs		7SL	93
B2	rongeurs		ARNt (Lysine)	146
ID	rongeurs		BC1	271
MT	souris			11
4,5SI	rat		7SL	233
G-repeat	hamster		ARNt	184
Bov-tA	pecorans		ARNt et monomère A	160
Bov-A2	artiodactyles		monomère A	67 - 160
PRE-1	suidés		ARNt (Arginine)	253
PRE-2	suidés		ARNt	253
ARE-1	artiodactyles		ARNt	5
ARE-2	artiodactyles		ARNt	5
CHR-1	cétacés, ruminant, hippopotame		ARNt	251
CHR-2	cétacés, ruminant, hippopotame		ARNt	251
C	lapin		ARNt (Glycine)	40
Can	*canoidea*		ARNt (Lysine)	42
ERE-1	équins		ARNt	234
Pol III/SINE	tortue	reptile	ARNt (Lysine)	73
Pol III/TAN	salamandre	amphibien	ARNt (Glutamate)	197
OAX	xenope		ARNt	197
Hpa I	salmonidés	poisson	ARNt (Lysine)	129
Ava III	salmonidés		ARNt (Lysine)	130
Sma I	salmonidés		ARNt (Lysine)	67
Fok I	charr		ARNt (Lysine)	179
Ron-1	cichlidés		ARNt	31
AFC	cichlidés		ARNt	276

SK	Squid	mollusque marin	ARNt (Lysine)	204
OK	octopus		ARNt (Lysine)	205
OR1	octopus		ARNt (Arginine)	205
OR2	octopus		ARNt (Arginine)	205
SURF1-1	Sea urchin	échinoderme	ARNt	200
Bm-1	*Bombyx mori*	insecte	ARNt	1
Lm1	*Locusta migratoria*		ARNt	25
SM α family	*Schistosoma mansoni*	ver plat	ARNt (Arginine)	269
p-SINE1	riz	plante	ARNt (Glycine)	187
TS	tabac		ARNt (Lysine)	310
S1	brassicace		ARNt	58
SINE potentiel	Erysiphe graminis	champignon parasite	ARNt	225
Mg-SINE	*Magnaporthe grisea*		ARNt	122

TABLEAU B (suite)

TABLEAU C : Identité entre les segments 3' de LINE et de SINE.

Nom du SINE	Nom du LINE	Taille du segment partagé (pb)	Identité	Identité SINE/LINE (refs.)
Bov-tA et Bov-A2	Bov-B	111	75% - 89%	262,208
Pol III/SINE	CR1-like	100	65% - 70%	288, 203
HpaI	RSg-1	57	77%	203
TS	solanaceous LINE	105	77%	203
SmaI	eel LINE-like	40	94%	203
MIR	L2	52	90%	262
Bm-1	élément LINE potentiel	81	78%	203
Mg-SINE	Ch	204	83%	203
AFC	Ci LINE 2	60	85%	282

Tableau établi à partir des articles Okada *et al.* (1997) et (Gilbert et Labuda, 1999).

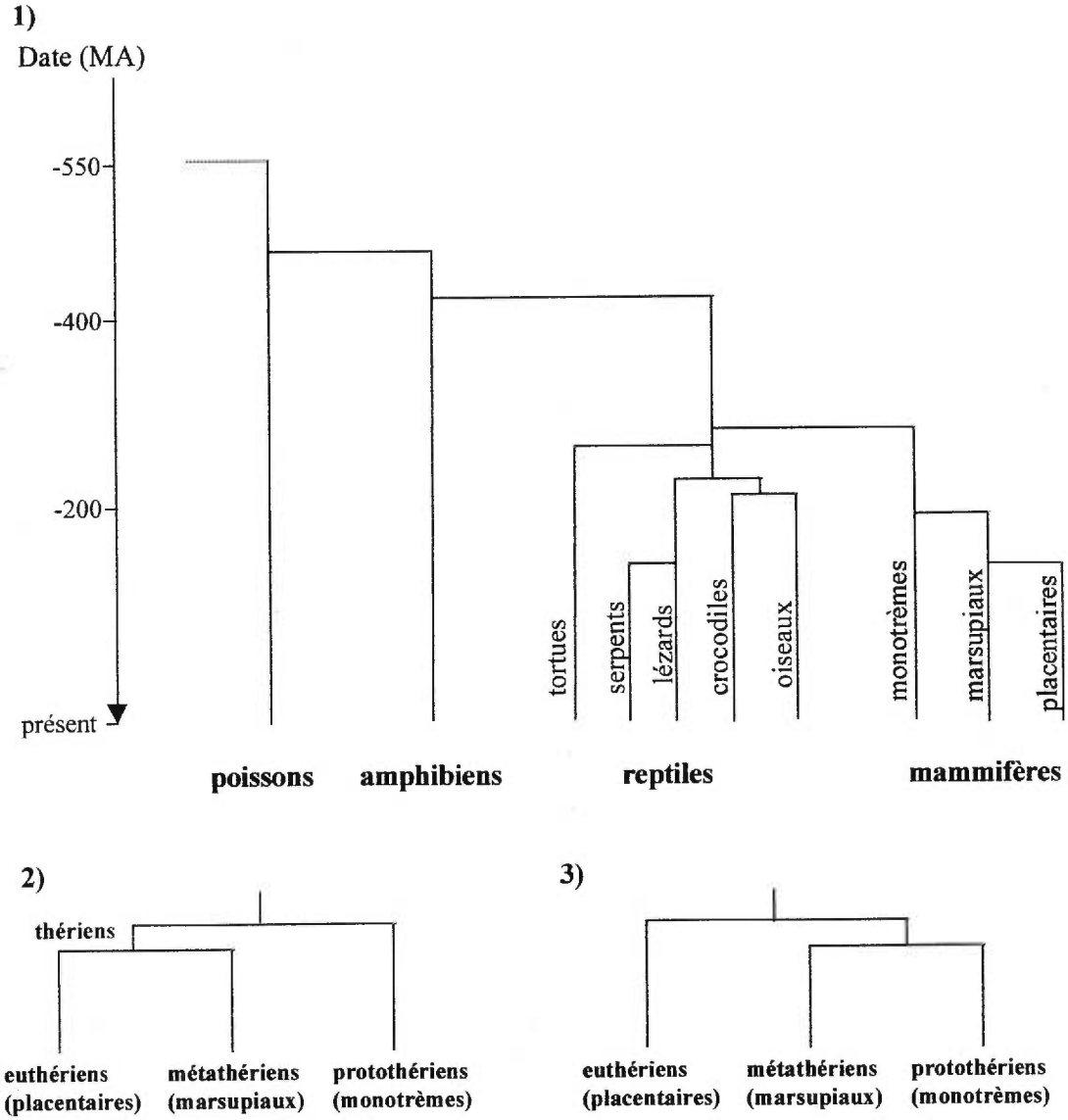
MATÉRIEL ET MÉTHODES.

2-1- MATÉRIEL.

2-1.1- Phylogénie des eucaryotes supérieurs.

Ce paragraphe est un survol rapide de la phylogénie des génomes eucaryotes. L'origine des eucaryotes remonte à environ 2000 millions d'années (MA). Vers 600 MA apparaissent les premiers invertébrés, suivis des premiers poissons (550 MA). Les premiers amphibiens et reptiles sont apparus entre 350 et 400 MA. Il existe plusieurs lignées de reptiles ; les tortues (300 MA), les crocodiles (250 MA) et les serpents et lézards (230 MA). De la lignée des dinosaures, proche parent des crocodiles, sont apparus les oiseaux (200 MA) (Gorr *et al.*, 1998). Descendants des reptiles, l'apparition des premiers mammifères date d'environs de 250 MA (Futuyma, 1998) (Figure N_1). De cette branche sont nés trois groupes. Deux hypothèses sont actuellement débattues : i) La plus ancienne (Figure N_2) place les protothériens (monotrèmes) comme les premiers mammifères d'où se sont séparés les thériens (150 à 200 MA). De l'ordre des thériens sont nés les sous-ordres des métathériens (marsupiaux) et des euthériens (placentaires) (110 à 170 MA). Cette hypothèse est fondée sur des données morphologiques de fossiles et sur des données moléculaires nucléaires (Kirsch *et al.*, 1997; Kumar et Hedges, 1998; Rougier *et al.*, 1998). ii) La deuxième hypothèse regroupe les protothériens et les métathériens dans une même branche séparée des mammifères placentaires (Figure N_3). Cette théorie, très controversée, s'appuie essentiellement sur des données moléculaires provenant de l'ADN mitochondrial (Janke *et al.*, 1996; Penny et Hasegawa, 1997).

FIGURE N : Arbres phylogénétiques des vertébrés et mammifères.



1) phylogénie des vertébrés, l'échelle de temps se trouve sur la gauche de l 'arbre. **2)** et **3)** phylogénie des mammifères suivant les deux théories qui s 'opposent.

Figure établie à partir des références Futuyma (1998), Gorr *et al.* (1998), Kumar et Hedges (1998) et Penny et Hasegawa (1997).

2-1.2- ADN et bactéries.

Les ADN génomiques utilisés représentent tous les groupes de vertébrés : les mammifères, les reptiles, les oiseaux, et les poissons. De l'ADN génomique de levure est utilisé comme contrôle négatif. La liste détaillée des espèces qui ont été analysées est décrite dans le Tableau D.

Certains échantillons d'ADN nous ont été généreusement donnés par les Docteurs Jennifer M. A. Graves (La Trobe University, Melbourne, Australia) et Chris Collet (Queensland University of Technology, Brisbane, Australia). Les autres ont été isolés au laboratoire à partir de tissus de foie ou de sang périphérique. L'acquisition d'échantillons de sang de certaines espèces a été possible grâce à la collaboration de la société zoologique de Granby par le Dr. Clément Lanthier, de l'écomuséum par le Dr. Bider et David Rodrigue, de la clinique vétérinaire Lachine par le Dr. Jean Gauvin et de l'école médecine vétérinaire de Sainte-Hyacinthe par la Dr. Pascale Benoîts (voir Tableau D).

La souche des bactéries *E. coli* DH5α à haute efficacité de transformation (Library Efficiency DH5α Competent Cells, Gibco BRL), > 10^8 transformants/μg, ont été utilisées pour la création des banques subgénomiques. Ces bactéries possèdent le gène pemettant l'α complémentation avec le gène LacZ du vecteur plasmidique qui reconstitue ainsi la βgalactosidase. L'activité de cette dernière permet la sélection des bactéries transformées.

TABLEAU D : Liste des espèces vertébrées utilisées.

Classes	*Sous-classes* (Infraclasse)	*Ordres* (genre)	*Espèces* (nom latin)	*Espèces* (nom vulgaire)	
Osteichthyes	*Actinopterygii*	*Salmoniforme*	*Salmo trutta*	truite	*a*
			Oncorhynchus keta	saumon	*c*
Amphibiens	*Lissamphibia*	*Caudata*	*Salamender sp.*	salamandre	*a*
		Anura	*Raua sp.*	grenouille	*a*
Reptiles	*Anapsida*	*Chelonia*	*Graptemys geographica*	tortue géographique	*a*
	Diapsida	*Squamata*	*Tiliqua scincoides*	lézard à langue bleue	*b*
			Pseudonaja textilis	serpent brun	*b*
Oiseaux	*Ornithurae*	*Gallinacé*	*Gallus gallus*	poule	*a*
		Anseriforme	*Alopochen aegyptiacus*	oie	*b*
		Columbiforme	*Columbia livia*	pigeon	*b*
		Passériforme	*Passer domesticus*	moineau	*b*
		Sturnidé	*Acridotherus tritis*	mainate indien	*b*
		Rallidé	*Porphyrio porphyrio*	« moorhen » (échassier)	*b*
Mammifères	*Protothérien*	*Monotreme*	*Ornithorhynchus anatinus*	Ornithorynque	*b*
	Thérien (Metathérien)	*Marsupial*	*Wallabia bicolor*	wallaby	*b*
			Sminthopsis macroura	souris marsupiale	*b*
			Isoodon macrorous	« bandicoot »	*b*
			Didelphis virginia	opossum de Virginie	*b*
	Thérien (Eutherien)	*Placentaire (Rongeur)*	*Mus musculus*	souris	*a*
		Placentaire (Artiodactyle)	*Bos taurus*	vache	*a*
		Placentaire (Chiroptère)	*Myotis sp.*	chauve-souris	*a*
		Placentaire (Primate)	*Homo sapiens*	humain	*a*

[a]: ADN isolés au laboratoire
[b]: ADN donnés par les Drs. Jennifer M. A. Graves et Chris Collet.
[c]: ADN acheté à la compagnie Gibco BRL

2-1.3- Plasmides et oligonucléotides.

Le vecteur utilisé pour les banques subgénomiques est le plasmide pBlueScript KS+ (pBS) (Pharmacia). Il possède un gène de résistance à l'ampicilline et une région de clonage avec plusieurs sites de restrictions uniques en amont du gène LacZ (ce gène représente les 146 premiers nucléotides du gène de la β–galactosidase). La région de clonage ne détruit pas le cadre de lecture du gène.

Deux oligonucléotides ont été utilisés comme sondes pour la sélection des éléments CORE-SINE à partir des banques subgénomiques; *Omir17* (17-mère) 5'-ACC TTG AGC AAG TCA CT-3' et *Omil17* (17-mère) 5'-GAT GAG GAA ACT GAG GC-3' (Jurka *et al.*, 1995).

Cinq oligonucléotides, de 40 bases chacun, ont été synthétisés pour l'hybridation par Southern-blot: T1 (5'-GAT AAT AAT AGC ACC TAC CTC CCA GGG TTG TTG TGA GGA T-3'), T2 (5'-TTG GAT TAG ATG GCC TCT AAG GTC CCT TTC AGT TCT AAA T-3'), M1 (5'-TGA GCT GGA GAA GGA AAT GGC AAA CCA CTC CAG TAT CTT T-3'), O1 (5'-TCC CAT TGC CTA GTC CTT NCC ACT TTT CTG CCT TGG AAC C-3') et Mo1 (5'-GAT TAA GAC TGT GAG CCC CAT GTG GGA CAG GGA CTG TGT C-3'). Chacun de ces oligonucléotides est spécifique pour une famille d'éléments CORE-SINE mammifère (respectivement Ther-1, Ther-2, Mar-1, Opo-1 et Mon-1) et provient de la région 3' terminale du consensus de la famille considérée.

Six oligonucléotides ont servi à la détection d'éléments LINE similaires à Bov-B ou CR1 chez les vertébrés. Quatre sont spécifiques de Bov-B: B5 (20-mère) 5'-AAG GCT ATG GTT TTT CCA GT-3', B3 (20-mère) 5'-CCA GCC ATC TCA TCC TCT GT-3', Br3 (17-mère) 5'-GTC GTG TCC GAC CCA TC-3' et Bm3 (17-mère) 5'-TCG TGT CCG ACT

CTT TG-3'. Parmi ces 4 oligonucléotides, les deux derniers proviennent de l'extrémité 3' des séquences LINE "Bov-B-like" des reptiles (Br3) ou du segment spécifique des éléments CORE-SINE de la famille Mar-1 des marsupiaux (Bm3). Les deux oligonucléotides utilisés pour la détection de CR1 sont CR1L5 (18-mère) 5'-CGG AGG GCA ACA AAA ATG-3' et CR1L3 (18-mère) 5'-TAG AGT TGG AAG GGA CCT-3'.

Deux oligonucléotides standards ont été utilisés pour le séquençage des éléments CORE-SINE; *P79* (17-mère) 5'- GGT GGC GGC CGC TCT AG -3' et *KS20m* (20-mère) 5'- CCT CGA GGT CGA CGG TAT CG -3'.

Deux autres oligonucléotides standards ont servi pour le séquençage des segments d'éléments LINE de type Bov-B; "sens" (Forward, 18-mère) 5'-GTT TTC CCA GTC ACG ACG-3' et "anti-sens" (Reverse, 18-mère) 5'-GAA TTG TGA GCG GAT AAC-3'. Le séquençage a été réalisé par la compagnie BioS&T à Montréal.

2-1.4- Les enzymes.

2-1.4.1- Les enzymes de restriction.

Les enzymes de restriction utilisés pour la digestion des plasmides sélectionnés dans les banques subgénomiques ou des ADN génomiques de différentes espèces vertébrées sont les suivants :

SmaI : site de restriction 5'-CCC|GGG-3' (bouts francs), utilisé pour créer le site de clonage des fragments d'ADN générés par nébulisation.

PvuII : site de restriction 5'-CAG|CTG-3' (bouts francs), utilisé pour la vérification de l'insertion de fragments d'ADN dans les vecteurs plasmidiques.

TaqI : site de restriction 5'-T|CGA-3' (extrémités 3' sortantes), utilisé pour la

digestion des ADN génomiques des Southern-blots.

Tous ces enzymes proviennent de la compagnie Gibco BRL.

2-1.4.2- Les enzymes de modification.

Les réactions de PCR (Réaction en Chaîne de Polymérisation - voir paragraphe 2.11) ont été réalisées avec l'ADN Polymérase *Taq* (Gibco BRL). La Polynucléotide Kinase T4 (Gibco BRL) a été utilisée pour la phosphorylation des extrémités des fragments d'ADN "nébulisés" et le marquage de sondes radioactives ; le fragment large "Klenow" de l'ADN Polymérase I (Gibco BRL) pour la réparation les extrémités des fragments issus de la nébulisation ; et enfin, l'ADN Ligase T4 (Gibco BRL) pour la ligature des fragments d'ADN "nébulisés" avec le plasmide pBS.

2-1.4.3- Les autres enzymes.

Une Phosphatase alcaline (Pharmacia) a été utilisée pour ôter les groupements phosphates aux extrémités 5' des plasmides pBS digérés par SmaI. La RNAse A et la protéinase K (Pharmacia) ont été utilisées lors des extractions d'ADN pour éliminer l'ARN et les protéines qui peuvent nuire aux expérimentations.

2-2- MÉTHODES.

2-2.1- Isolement d'ADN génomique de haut poids moléculaire.

L'ADN de haut poids moléculaire est isolé à partir de sang périphérique ou encore de tissus de foie. Les espèces dont l'ADN a été extrait au laboratoire sont énumérées dans le

Tableau D. Le sang périphérique est récolté dans des tubes contenant de l'EDTA (anticoagulant) et conservé à 4°C. Les tissus de foie, une fois extraits de l'animal, sont rapidement congelés à –80°C pour éviter la dégradation de l'ADN. Le tissu congelé est broyé dans un mortier, contenant de la glace sèche, à l'aide d'un pilon. Le broyât, maintenu à l'état congelé, prend l'aspect d'une poudre. Les étapes d'extraction qui vont suivre sont les mêmes pour le sang ou les tissus. La poudre ou le sang est alors ajouté à un tampon de lyse NE (EDTA 25 mM ; NaCl 75 mM; SDS 0,5%; pH 8,0) contenant de la protéinase K (0,1mg/ml), pour éliminer les protéines qui forment l'environnement de l'ADN, de la RNAse A (40µg/ml) pour éliminer la grande quantité d'ARN cellulaire. La quantité de tampon de lyse est de 1,2 ml pour 100 mg de tissus ou 100 µl de sang. La solution est incubée au moins 24 heures à 37°C avec agitation légère. Deux volumes d'eau et un volume de solution saturée en NaCl (6 M) sont ensuite ajoutés. Après agitation par inversion, la solution est centrifugée à 2400 g pendant 30 minutes pour éliminer les débris cellulaires. L'ADN contenu dans le surnageant est récupéré et précipité dans deux volumes d'éthanol à 95%. L'ADN est récupéré à l'aide d'une pipette pasteur dont l'extrémité à été recourbée, puis lavé à l'éthanol 70%. L'ADN récolté est resuspendu dans 250 µl d'une solution TE 1X (Tris-HCl 10 mM; EDTA 1 mM, pH 8,0). Pour que la resuspension soit totale, il est parfois nécessaire d'incuber la solution d'ADN à 37°C pendant quelques heures.

2-2.2- Isolement d'ADN de plasmide.

L'ADN plasmidique est extrait de mini-preparations de bactéries transformées (voir paragraphe 2-2.9.2-). Pour chacune des mini-préparations, une colonie de bactéries est isolée. À l'aide d'un cure-dent, un volume de 3 ml de milieu de culture TB 1X (Bacto-

tryptone 1,2% ; extrait de levure 2,4% ; glycérol 0,4% ; KH_2PO_4 17 mM ; K_2HPO_4 72 mM) contenant 100 µg/ml d'ampicilline est ensemencé à partir d'une colonie. On laisse croître les bactéries sous agitation (250 tours/min.) une nuit à 37 °C. Une centrifugation de 3 minutes à vitesse maximale du milieu de culture permet d'obtenir un culot de bactéries dans un tube Eppendorf de 1,5 ml. Le culot est resuspendu dans 200 µl de GTE (glucose 50 mM ; Tris-HCl 25 mM, pH 8,0 ; EDTA 10 mM). 400 µl de solution NS (NaOH 0,2 M ; SDS 1%) sont ajoutés et le mélange est agité délicatement par inversion du tube. Le SDS provoque la lyse des cellules et le NaOH dénature l'ADN. On ajoute enfin 200 µl de solution "hight salt" (Acétate de potassium 3 M ; Acide formique 1,8 M). Cette solution neutralise le milieu et permet la renaturation de l'ADN plasmidique, l'ADN génomique de la bactérie restant en complexe insoluble avec les débris cellulaires. Une courte centrifugation à vitesse maximale permet de recupérer l'ADN plasmidique dans le surnageant. L'ADN est alors précipité en présence de 0,6 volume d'isopropanol (480 µl). Après centrifugation pendant 10 minutes (vitesse maximale) le culot est resuspendu dans 100 µl de Tris-HCl 50 mM, pH 7,5. On y ajoute 50 µl d'acétate d'ammonium 7,5 M pour effectuer une seconde précipitation, mais cette fois avec de l'éthanol 95% (2 volumes). Après une centrifugation de 15 minutes à vitesse maximale le culot de plasmide est lavé dans de l'éthanol 70% puis séché. Les plasmides sont resuspendus dans 30 µl de solution TE 1X.

2-2.3- Précipitation des acides nucléiques.

Lorsque l'on veut précipiter des acides nucléiques pour effectuer une purification, on ajoute dans notre phase aqueuse contenant l'ADN des sels. La solution finale doit contenir 0,3 M d'acétate de sodium, pH 6,0, ou 2,5 M d'acétate d'ammonium, pH 7,5. On effectue

ensuite deux extractions à volume égal de phénol/C/AI (phénol : Chloroforme : Alcool Isoamylique, 50:48:2) et une troisième extraction à volume égal de C/AI (chloroforme : alcool isoamylique, 24:1). Pour chaque extraction la phase aqueuse contenant l'ADN est mélangée à la phase organique et récupérée après centrifugations de 30 minutes à 2400g. L'ADN est alors précipité avec deux volumes d'éthanol à 95%. Après une centrifugation de 30 minutes à 2400g, le culot est lavé une fois à l'éthanol 70% puis séché sous vide et enfin resuspendu dans un tampon TE 1X.

Pour augmenter la concentration d'une solution d'ADN on peut effectuer simplement les étapes de précipitation dans l'éthanol 95%. Si la concentration d'ADN de départ est très faible, on peut ajouter un vecteur de précipitation qui est le glycogène. Ce dernier est un polysaccharide qui permet une récupération plus complète de l'ADN sans altérer ses qualités, c'est-à-dire qu'il n'entrave pas aux digestions enzymatiques, ni ne perturbe la migration de l'ADN sur gel d'agarose.

2-2.4- Digestion de l'ADN et séparation sur gel d'électrophorèse.

L'ADN génomique ou plasmidique est digéré avec un excès d'enzyme de restriction suivant les conditions expérimentales recommandées par le manufacturier. Généralement, la digestion de l'ADN génomique est réalisée en 2 étapes, une première incubation de 12 à 18 heures en présence de l'enzyme de restriction (5 unités/µg), et une seconde incubation de 2 heures, après ajout d'enzyme, pour s'assurer d'une digestion complète. Lorsqu'il s'agit d'ADN plasmidique, l'incubation dure une heure et la quantité d'enzyme utilisée suit le ratio 1 unité pour 1 µg de plasmide.

La vérification de la digestion se fait par électrophorèse sur gel d'agarose 1% (p/v)

dans le tampon TBE 1X (Tris-borate 90 mM ; EDTA 2 mM, pH 8,3). Le temps et l'intensité de migration du gel peuvent varier, mais de façon générale le voltage est de 2 à 5 V/cm pendant 2 à 5 heures. La digestion génère, pour l'ADN génomique, une multitude de fragments qui ne peuvent se distinguer les uns des autres et qui donnent l'apparence d'une trainée ("smear"), visible à l'aide d'une coloration au bromure d'éthidium (0,5 µg/ml). Le bromure d'éthidium est un agent intercalant qui se fixe entre les bases de la chaîne d'ADN et qui émet une fluorescence lorsqu'il est excité par une lumière UV. Pour l'ADN de plasmide, la digestion génère des fragments de restriction spécifiques qui peuvent être très bien séparés les uns des autres sur le gel d'agarose. Les marqueurs de masse moléculaire permettent de déterminer la taille des fragments observés. Les marqueurs communément utilisés sont les marqueurs 1 Kb (GibcoBRL) et 100 pb (GibcoBRL).

2-2.5- Fractionnement d'ADN génomique par nébulisation.

La digestion de l'ADN est une technique de fractionnement non aléatoire, dépendante de la présence de sites de restriction. Au contraire le fractionnement mécanique permet un découpage aléatoire. Plusieurs techniques sont disponibles dont la nébulisation (Bodenteich et al 1994), qui produit des fragments de taille homogène. Il existe un instrument de laboratoire conçu pour la nébulisation. Cependant, les mêmes résultats peuvent être obtenus, à moindre frais (3$ ou 12FF), en utilisant un nébuliseur plastique prescrit aux asthmatiques. Trois paramètres sont importants pour sélectionner la taille des fragments : la viscosité du milieu, l'intensité et la durée de la pression. La viscosité restreint l'amplitude de la longueur des fragments autour d'une taille moyenne imposée par la pression. La durée affecte la taille moyenne et la dispertion des fragments. Un volume de 1 ml contenant 2 à 3 µg d'ADN

génomique est soumis aux conditions de nébulisation suivantes : viscosité, 25% glycérol; pression 2 kg/cm^2; durée 150 secondes. Ces conditions génèrent des fragments d'une taille moyenne de 300 nucléotides et variant entre 200 et 700 nucléotides. La solution issue de la nébulisation est récupérée et précipitée à l'acétate de potassium en présence de glycogène (voir paragraphe 2.3). Cette étape élimine le glycérol et augmente la concentration en ADN. Le culot d'ADN obtenu est resuspendu dans 9 µl de Tris-HCl 10 mM. Les fragments d'ADN sont ensuite phosphorylés pendant 30 minutes à 37°C avec 5 unités de kinase T4 (Gibco BRL) en présence du tampon "kinase forward" 5X fourni par le manufacturier (Tris-HCl 350 mM, pH 7,6 ; $MgCl_2$ 50 mM; KCl 500 mM ; 2-mercaptoethanol 5 mM) et 3 mM d'ATP dans un volume total de 12,5 µl. Les extrémités des fragments sont ensuite réparées dans le même tampon en rajoutant 2 unités du fragment large "Klenow" de l'ADN polymérase I (Gibco BRL) et les 4 dNTP (0,4 mM chaque). Après une 1 heure d'incubation à 37°C, la solution est ramenée à 100 µl pour purifier l'ADN par deux extractions au Phénol/C/AI (voir paragraphe 2-1). La précipitation qui suit est effectuée en présence d'acétate d'ammonium et de glycogène (voir paragraphe 2.3). L'ADN est resuspendu dans 20 µl de TE 1X.

L'inconvénient majeur de la technique de nébulisation est la perte d'environ 30% de l'ADN contenu au départ de l'expérience.

2-2.6- Transfert d'ADN sur support solide.

Les techniques de transfert sur support solide permettent la recherche d'identités entre l'ADN et des sondes nucléotidiques simple brin radiomarquées.

2-2.6.1- Southern-blot.

Après digestion de l'ADN, la technique de Southern-blot est utilisée pour transférer sur support solide les fragments séparés par électrophorèse sur gel d'agarose (Southern, 1975). Le support utilisé est une membrane de nylon chargée positivement, Hybon-N^+ (Amersham), permettant un transfert alcalin. Le gel d'agarose est immergé dans une solution d'HCl 0,25 M (dépurination) pendant 10 minutes, rincé à l'eau, et transféré par buvardage avec une solution alcaline (NaOH 0,4 M) (le montage standard est illustré dans les protocoles du livre "Molecular cloning, a laboratory manual", (Maniatis *et al.*, 1982)). Après le transfert, la membrane est rincée avec une solution de SSC 2X (NaCl 0,3 M ; citrate de sodium 30 mM, pH 7,0) puis l'ADN est fixé par exposition à des rayonnements UV (254 nm et à 120000 $\mu J/cm^2$ dans un transilluminateur Stratalinker-1800) (Stratagene).

2-2.6.2- Dot-blot.

La technique du dot-blot est utilisée pour fixer sur support solide de l'ADN qui n'est pas obligatoirement digéré. La fixation est effectuée avec l'appareil "Minifold®I dot-blot system" (Schleicher & Schuell).

L'ADN est dissout dans 100 µl d'une solution SSC 10X (NaCl 1,5 M; citrate de sodium 150 mM, pH 7,0) et est appliqué sur une membrane Hybon-N^+ (Amersham) préalablement humidifiée avec une solution SSC 10X. Après l'application, la membrane est rincée avec du SSC 10X, dénaturée dans une solution NaCl 1,5 M, NaOH 0,5 M pendant 10 minutes, puis laissée 15 minutes dans une solution de neutralisation (NaCl 1,5 M ; Tris-HCl 0,5 M, pH 7,2). L'ADN est fixé à la membrane par rayonnement UV de la même manière que pour le Southern-blot.

2-2.7- Hybridation moléculaire de l'ADN.

2-2.7.1- Hybridation avec une sonde oligonucléotidique.

La membrane de nylon sur laquelle est fixé l'ADN est préhybridée pendant une heure dans une solution de SSPE 1X (NaCl 150 mM ; NaH_2PO4 10 mM ; EDTA 1 mM, pH 7.4), NaCl 0,75 M, Tris-HCl 70 mM pH 7,4, SDS 1% et 200 ng/ml d'héparine. L'héparine bloque les sites non spécifiques. L'hybridation est faite en ajoutant directement 10 pmoles de la sonde oligonucléotide radiomarquée (paragraphe 2-2.13-) dans la solution de préhybridation (correspondant environ à 10 millions de cpm). La durée d'hybridation est de 2 à 4 heures et se fait à la même température que la préhybridation ; 37°C lorsque la sonde est de 17 nucléotides (*Omil*, *Omir*) et 50°C pour 40 nucléotides (T1, T2, M1, O1 et Mo1). Les hybridations sont effectuées à faibles stringence (plus de dix degrés en dessous du Tm) à cause des divergences importantes des séquences par rapport à leur consensus. Trois lavages suivent l'hybridation avec une solution contenant du SSPE 2X et du SDS 0,1%, un premier rapide à la température de la pièce et deux de 10 minutes à la température d'hybridation. Si le signal de radioactivité est très fort après le troisième lavage (vérifié au compteur Geiger), un lavage supplémentaire peut être effectué à la même température ou à une température supérieure de un degré. Toutes ces étapes sont effectuées dans un four à hybridation rotatif (Fisher). Les membranes lavées sont plastifiées et exposées sur un film sensible aux rayons X (Fuji HR-G ou Kodak BIOMAX) à -75°C, ou dans une cassette PhosphoImager à température de la pièce.

2-2.7.2- Hybridation avec une sonde PCR.

L'hybridation avec une sonde PCR de petite taille, c'est-à-dire inférieure à 60 nucléotides, ne diffère de l'hybridation avec une sonde oligonucléotide que par une étape de dénaturation. Cette étape consiste à faire chauffer la sonde pendant 3 minutes à 95°C juste avant de l'ajouter à la solution de préhybridation. Cela permet de séparer les deux brins du fragment de PCR et rend ainsi possible l'hybridation de la sonde avec l'ADN fixé à la membrane. La température d'hybridation varie entre 38 et 45°C.

L'hybridation avec une sonde PCR de plus grande taille nécessite aussi cette étape de dénaturation. La solution de préhybridation est différente des autres, elle contient du SSPE 6X, du Denhardt 1X (Ficoll 0,02%; polyvinyl pyrolidone 0,02%; albumine de sérum de boeuf 0,02%), du SDS 0,5% et de l'héparine (200 ng/ml). Les autres étapes sont les mêmes à l'exception de la température d'hybridation qui est de 65°C.

Le marquage des sondes est décrit dans le paragraphe 2.13

2-2.8- Quantification d'un signal radioactif par PhosphoImager.

Les cassettes PhosphoImager possèdent des écrans sensibles aux rayonnements X émis par les sondes radioactives utilisées pour les expériences d'hybridation. Ces écrans sont plus sensibles que les films d'autoradiographie. Le résultat de l'hybridation, le signal radioactif, est détecté par un scanner : PhosphoImager™ SI (Molecular Dynamics). Le second avantage de détection par PhosphoImager est que le signal peut être quantifié à l'aide du logiciel Image QuaNT™ (Version 4.0, Molecular Dynamics). Ceci permet d'effectuer des dosages semi-quantitatifs pour estimer le nombre de copies d'une séquence répétitive d'un génome donné.

Les séquences quantifiées étant très divergentes des sondes utilisées, le signal observé par hybridation ne représente pas la totalité des copies génomiques. Nous savons que nous détectons uniquement les séquences n'ayant pas plus de deux mésappariements avec les sondes. En considérant ces deux paramètres (divergence et mésappariement) nous avons pu établir une correction de l'estimation du nombre de copies dans les génomes. A l'aide d'une distribution binomiale, à un taux de mutations donné, nous pouvons savoir quel est le pourcentage du nombre d'éléments détectés par hybridation.

2-2.9- Création de banques subgénomiques.

Des banques subgénomiques de trois espèces mammifères (humain, kangourou et ornithorynque) ont été créées à partir d'ADN ayant subit un fractionnement mécanique par nébulisation (voir paragraphe 2-2.5-).

2-2.9.1- Ligature de fragments d'ADN à un vecteur plasmidique.

60 ng de fragments d'ADN sont liés à 100 ng de vecteur pBlueScript KS+ (Pharmacia) préalablement digérés par l'enzyme de restriction SmaI (Gibco BRL) et déphosphorylés aux extrémités 5'. La déphosphorylation s'effectue dans les conditions suivantes : 30 minutes à 37°C dans un tampon "one for all" (Tris-acétate 10 mM; Mg-acétate 10 mM; K-acétate 50 mM) en présence de 5 unités de phosphatase alcaline (Pharmacia). Cette réaction limite la fermeture du vecteur sur lui-même lors de l'étape de ligature. Le vecteur ne possède qu'un seul site de restriction SmaI (CCC|GGG) dans la région de clonage, en amont du segment lac Z, fraction du gène de la βgalactosidase. La réaction de ligature est effectuée à 15°C pendant une nuit dans 10 µl de tampon (Tris-HCl

50 mM, pH 7,6 ; $MgCl_2$ 10 mM ; ATP 1mM ; DTT 1mM ; polyéthylène glycol-8000 5% (masse/v)) en présence d'une unité de ligase T4 (Gibco BRL). Le produit obtenu est dilué 50 fois avant la transformation de bactéries *E. coli* compétentes "DH5α library efficiency" (Gibco BRL) et est conservé à -20°C.

Les produits de PCR peuvent aussi être intégrés dans le vecteur plasmidique. La quantité de fragment d'ADN mis en solution avec le plasmide digéré par l'enzyme SmaI dépend de sa taille et suit l'équation suivante: $\left(3xQx\left(\frac{l}{L}\right)\right)$. Le facteur 3 est utilisé pour favoriser l'insertion d'un fragment. Il augmente la probabilité de contact entre l'extrémité d'un fragment et celle d'un plasmide. Q représente la quantité de vecteur en solution (100 ng) , l la taille de l'insert (dans notre cas la moyenne est de 300 nucléotides) et L la taille du vecteur (2950 nucléotides).

2-2.9.2- Transformation de cellules bactériennes compétentes.

La transformation permet l'introduction d'ADN circulaire (plasmide) possédant un fragment d'ADN génomique étranger d'interêt dans des bactéries.

Les bactéries compétentes DH5α (voir paragraphe 1.1) sont décongelées lentement sur glace. Un volume de 1µl de la dilution du produit de ligature (voir paragraphe 2.9.1) est ajouté à 50 µl de bactéries. Après une incubation d'au moins 30 minutes à 4°C, on procède à un choc thermique de 45 secondes à 42°C pour faciliter la pénétration des plasmides dans les cellules. Les bactéries sont ensuite incubées à 37°C pendant 1 heure sous agitation (225 rpm) dans 500 µl de milieu SOB (bacto-tryptone 2 % ; extrait de levure 0,5% ; NaCl 10 mM ; $MgCl_2$ 10 mM). Les bactéries sont alors étalées sur des boîtes de pétri

contenant un milieu semi-solide LB-agar (bacto-tryptone 1%; extrait de levure 0,5%; NaCl 0,17 M, pH 7,4, agar 1.5%). Sur chaque boîte est étalé préalablement 2 mg d'ampicilline, 800 µg d'IPTG (Isopropylthio-β-D-galactoside) et 800 µg de X-Gal (5-bromo-4-chloro-3-indolyl-β-galactoside). Seules les bactéries qui ont intégré un plasmide codant pour la résistance à l'ampicilline peuvent pousser sur le milieu. L'IPTG est un inducteur du promoteur de la βgalactosidase présent en amont du site de clonage et du gène LacZ. X-Gal est le réactif clivé par la βgalactosidase produisant ainsi une coloration bleue des bactéries. Si le vecteur a intégré un insert d'ADN dans le site Sma I, il n'y a pas coloration puisque le gène de la βgalactosidase n'est pas reconstitué. Ces bactéries "blanches" sont donc sélectionnées et représentent la banque subgénomique.

2-2.10- Sélection des bactéries par hybridation moléculaire.

De la banque subgénomique, seules les colonies ayant un fragment d'ADN contenant la séquence d'intérêt doivent être sélectionnées. Pour cela, la totalité des colonies se trouvant sur les boîtes de pétri est transférée sur support solide par simple application d'une membrane Hybond-N+ sur le milieu semi-solide. Pour chaque boîte, il est fait un duplicata. Après le transfert, les colonies sont remises en incubation à 37°C pendant quelques heures. Les membranes sont disposées sur une solution de dénaturation (NaOH 0,5 M ; NaCl 1,5 M) pendant 10 minutes pour permettre l'extraction de l'ADN et sa fixation sur la membrane. Les membranes sont ensuite neutralisées (Tris-HCl 0,5 M, pH 8,0 ; NaCl 1,5 M ; EDTA 1 mM) pendant 15 minutes et exposées aux rayonnements UV pour fixer irréversiblement l'ADN. Les membranes ainsi créées sont hybridées avec des sondes oligonucléotidiques de la même façon que les Southern-blots ou les dot-blots. Les colonies montrant un signal

radioactif après l'hybridation sont directement sélectionnées sur la boîte de pétri qui est à l'origine de la membrane.

2-2.11- Réaction en chaîne de polymérisation (PCR).

La réaction en chaîne de polymérisation, ou PCR, est particulièrement utile pour détecter la présence d'une séquence donnée, dans un génome. Cet outil permet aussi d'obtenir en grande quantité un fragment précis d'ADN pouvant servir de sonde pour hybridation, ou pouvant être séquencé.

Les réactions de PCR sont effectuées dans un volume total de 20 µl. Ces 20 µl contiennent du tampon de PCR 1X (Tris-HCl 20 mM, pH 8,4; KCl 50 mM), du $MgCl_2$ 1,5 mM, les 4 dNTP (100 µM chaque), deux amorces (625 nM chacune) et une unité de polymérase *Taq* (GibcoBRL).

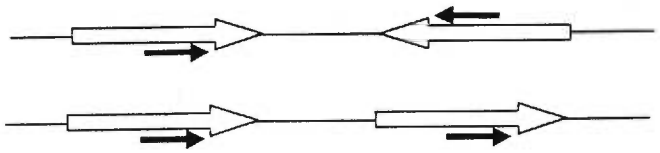
Les réactions s'effectuent en trois étapes. La première consiste en la dénaturation de l'ADN par la chaleur. La solution est exposée pendant 3 minutes à 94°C. La seconde étape est l'amplification cyclique du fragment. Le cycle suit les étapes suivantes : tout d'abord l'hybridation des amorces sur l'ADN, à une température (Tm) définie par la composition en nucléotides des amorces (composition (G+C) x 4°C + composition (A+T) x 2°C = Tm), pendant 30 secondes; ensuite l'élongation des amorces par l'enzyme à 72°C, pendant 1 minute; et enfin une dénaturation de 15 secondes à 94°C pour permettre un nouveau cycle. La dernière étape de la PCR sert à terminer toutes les élongations des cycles précédents: 30 secondes à la température Tm suivi de 5 minutes à 72°C. Les produits de PCR sont ensuite conservés à 4°C.

Toutes les PCR ont été initiées par "*hot-start*", c'est-à-dire que la polymérase est

ajoutée à la réaction pendant la dénaturation initiale à 94°C. Les machines à PCR utilisées sont, DNA Thermal Cycler de Perkin Elmer et HYBAID, Omnigene d'Interscience.

2-2.12- Réactions de PCR inter répétitions (*Inter-repeat PCR*).

Ces réactions ne diffèrent des PCR normales que par la présence d'une seule amorce d'amplification dans la solution au lieu de deux. L'amorce utilisée est spécifique d'un élément répété. Une amplification n'a lieue que si deux séquences répétées d'une même famille en orientation inverse sont proches (voir schéma).



2-2.13- Évaluation de la concentration des ADN.

Les concentrations d'ADN, aussi bien génomiques que plasmidiques, ont été évaluées par la technique de fluoromètrie, en utilisant le fluoromètre TKO 100 (Hoefer).

Une fraction de l'échantillon à doser est mélangée dans une solution de TNE 1X (NaCl 100 mM ; Tris HCl 10 mM, pH 8,0 ; EDTA 10 mM) et le fluorochrome Hoersht B3258 (100 ng). Le volume total de 50 µl est dosé et le résultat est comparé à une gamme étalon.

2-2.14- Marquage radioactif de sondes ADN.

Plusieurs sondes radioactives ont été utilisées dans les expériences d'hybridation. Toutes possèdent le ^{32}P comme isotope radioactif, qui provient du produit α-[^{32}P]-dCTP

(400 Ci/mmole, Amersham) ou γ-[^{32}P]-ATP (6000 Ci/mmole, Amersham). L'efficacité du marquage est vérifiée à l'aide d'un compteur à scintillation, et le résultat est donné en cpm (coups par minute).

2-2.14.1- Sondes oligonucléotidiques.

Le marquage des sondes oligonucléotidiques se fait à l'extrémité 5' déphosphorylée. Il s'agit donc de phosphoryler cette extrémité avec un groupement provenant de la molécule γ-[^{32}P]-ATP. Le marquage est réalisé dans les conditions suivantes : amorce 5 µM, tampon de kinase 1X (Tris-HCl 70 mM, pH 7,6 ; $MgCl_2$ 10 mM ; KCl 100 mM ; 2-mercaptoéthanol 1 mM), 10 unités de Polynucleotide Kinase T4 (GibcoBRL) et γ-[^{32}P]-ATP (33 µM). La réaction s'effectue à 37°C pendant 1 heure. La sonde est ensuite purifiée sur une colonne Sephadex-G25 (équilibrée avec une solution de Tris-HCl 10 mM, pH 7,4; EDTA 1 mM). Cette purification élimine tous les nucléotides marqués non incorporés.

Si on veut utiliser ces amorces marquées pour faire une PCR, la concentration finale des amorces après phosphorylation doit être 25 µM. Dans ces conditions de marquage, le γ-[^{32}P]-ATP est en quantité limitante (10 µM) pour éviter l'étape de purification.

2-2.14.2- Sondes PCR.

Il existe deux techniques pour radiomarquer des sondes PCR. La première est la technique de PCR (voir paragraphe 2.11) dite chaude. Elle nécessite la présence d'amorces marquées (voir paragraphe 2.13.1) lors de l'amplification. La seconde utilise l'α-[^{32}P]-dCTP

comme nucléotide marqué. Dans ce type de réaction un seul détail change par rapport à une PCR classique (voir paragraphe 2.11) ; de l'α-[^{32}P]-dCTP (330 nM) est ajouté pour remplacer partiellement le dCTP "froid" qui se trouve alors à une concentration de 50 µM.

Après les PCR, les produits sont purifiés sur colonne de Sephadex-G25 (paragraphe 2.13.1).

2-2.14.3- Marquage aléatoire.

La technique utilise un produit de PCR froid, dénaturé à 94°C, (environ 100 ng) avec lequel va être effectué un marquage aléatoire en utilisant le fragment large "Klenow" de l'ADN polymérase I (5 unités, Gibco BRL), des amorces hexanucléotides aléatoires (9 mmole, Amersham), trois dNTP (dATP, dGTPet dTTP, 150 µM chaque) et le nucléotide marqué α-[^{32}P]-dCTP (4µM). La réaction est faite dans 25 µl de tampon (Tris-HCl 50 mM, pH 8,0; $MgCl_2$ 10 mM; NaCl 50 mM) à 37°C pendant une heure. Encore une fois les produits de la réaction sont purifiés sur une colonne Sephadex-G25.

2-2.15- Séquençage automatique.

Le séquençage des clones des banques subgénomiques a été effectué à l'aide du séquenceur automatique ABI 373A (Perkin Elmer) et les logiciels "373A data collection" et "373A data analysis" fournis par Perkin Elmer. La réaction de séquençage des inserts a été réalisée à l'aide du kit "Taq DyeDeoxy Terminator Cycle Sequencing" (faisant référence à la technique de Sanger) et d'amorces standards du vecteur pBS, *P79* et *KS20m* (voir paragraphe 1.2). La réaction d'élongation est identique à celle d'une PCR. Cependant, une seule amorce est utilisée pour générer des fragments simples brin. Le volume réactionnel est

de 20 µl et contient la solution prémix (tampon, polymérase Taq, les 4 nucléotides didéoxy chacun couplé à un fluorochrome différent, Applied Biosystems, Inc), 1 µg d'ADN de plasmide (100 ng si on séquence un produit de PCR) et 5 pmoles d'une amorce. Cette solution subit 25 cycles de PCR dans les conditions suivantes : 30 secondes de dénaturation à 94°C, 15 secondes d'hybridation de l'amorce à 50°C et 4 minutes d'élongation à 60°C. Le produit d'élongation est ensuite conservé à 4°C.

La solution est purifiée sur une colonne Centri-Sep™ (Applied Biosystems, Inc) par centrifugation pour éliminer tous les nucléotides non incorporés. Le produit d'élongation est ensuite séché sous vide puis resuspendu dans 5 µl d'une solution formamide/EDTA 50 mM, pH 8,0 (v:v - 5:1).

Les réactions de séquençage sont ensuite déposées sur gel de polyacrylamide, dans le séquenceur ABI 373A, et les séquences sont collectées par laser couplé au logiciel "Data collection" pendant 10 heures. Par la suite le logiciel "Analisis" interprète les données et restitue les séquences.

Tous les clones ont été séquencés à partir de chacune des deux amorces *P79* et *KS20m*, permettant une double lecture et un contrôle des possibles erreurs de séquence. Les erreurs sont vérifiées en utilisant le logiciel Seqed (Seqed ABI Applied Biosystems version 1.0). Ce programme permet d'aligner plusieurs séquences et permet en même temps d'observer le chromatogramme de chacune des séquences comparées.

2-2.16- Comparaison et alignement de séquences par traitement informatique.

Pour identifier des régions d'identité entre deux séquences, j'ai utilisé le programme dot-matrix, du logiciel Mac DNAsis Pro V3.5 (Hitachi Software engineering CO. LTD.). Ce

programme détecte les identités suivant les paramètres suivant: nombre de nucléotides identiques dans une fenêtre de taille donnée. La condition la plus utilisée a été 9 nucléotides dans une fenêtre de 15.

Pour observer les régions d'identité, deux logiciels ont été utilisés, Mac DNAsis Pro V3.5 (Hitachi Software engineering CO. LTD.) et Multalin version 4.0 (Corpet, 1988). Ils permettent d'aligner un grand nombre de séquences et restituent un consensus pour les régions fortement identiques. Cependant, ils ne sont pas suffisamment performants pour donner le meilleur alignement. Ainsi pour chaque alignement effectué, un raffinement manuel à été nécessaire. Une séquence consensus a été créée pour chaque alignement en utilisant le nucléotide le plus fréquent pour chaque position.

2-2.17- Analyse phylogénétique de groupes de séquences.

L'analyse phylogénétique d'un groupe de séquences permet de déterminer les relations historiques qui peuvent exister entre ces séquences. Le programme de vraisemblance maximale (maximum likelihood) (DNAML) provenant du logiciel PHYLIP (Phylogeny Inference Package Version 3.5c) (Felsenstein, 1993) a servi à faire les analyses phylogénétiques, en utilisant les alignements de séquences établis par le programme Multalin et raffinés manuellement.

2-2.18- Recherche de séquences dans les banques de données.

L'Université de Montréal, via le serveur Cyclone, permet l'accès à l'environnement GCG sous UNIX (Wisconsin Package version 9.0, Genetics Computer Group, Madison, Wisc.) qui possède un grand nombre de programmes pour la recherche de séquences. Au

laboratoire, nous avons aussi accès, via internet, aux banques de données du site web de NCBI (National Center for Biotechnology Information) : http://www.ncbi.nlm.nih.gov.

2-2.18.1- Création de banques de données génomique.

Dans l'environnement GCG il est possible de créer ses propres banques de données de séquences à l'aide du programme "LookUp". Il nous permet de sélectionner les séquences provenant de toutes les banques de données possibles (EMBL, GenBank, etc.) suivant tous les critères possibles. Pour ma part, j'ai créé les banques en fonction des organismes d'intérêt. Une banque a été construite pour les séquences provenant des génomes monotrèmes en sélectionnant le mot *Monotremata* dans la section organisme. De la même façon, une banque a été construite pour les marsupiaux (*Metatheria*), les reptiles (*Squamata*), et les oiseaux (*Aves*) (Tableau E). Pour diminuer la taille de certaines banques, des sélections ont été faites afin d'éliminer les séquences mitochondriales. Il faut pour cela utiliser la section "définition" de la séquence et introduire le fait que l'on ne veut pas les séquences qui portent dans leur définition les mots mitochondrie ou mitochondrion.

La banque de donnée qui découle de cette manipulation est en fait une liste de noms (numéro d'accession avec la définition de la séquence) qui réfère à la banque de données originelle (GenBank ou EMBL). Ainsi lorsque nous effectuons une recherche sur nos banques, le programme sélectionne dans la banque originelle les séquences qui sont sur la liste. L'avantage de ce système est d'éviter l'accumulation d'un grand nombre de séquences sur notre compte et ainsi d'économiser beaucoup de mémoire, mais aussi de réduire le temps d'analyse.

TABLEAU E : Taille des banques génomiques créées dans GCG à partir du programme LookUp utilisant GenBank version 105.0.

Organisme	nb. de séquences [a]	Taille en nucléotides [b]	‰ de représentation du génome
monotrème	22 : 6 [c]	3 739	0,001
marsupial	524 : 355	231 136	0,07 [d]
reptile	1543 : 412	375 785	0,3 [d]
oiseau	6839 : 4741	5 668 148	4,7 [d]

[a] : Le premier chiffre correspond au nombre de séquences des organismes considérés dans GenBank version 105.0, le second chiffre donne le nombre de séquences nucléaires.

[b] : La taille est calculée en fonction des séquences nucléaires uniquement.

[c] : Dans la catégorie des monotrèmes, il a été possible d'éliminer les séquences redondantes (séquences homologues entre deux espèces), ce qui n'est pas le cas pour les autres données.

[d] : Les représentations des génomes sont des estimations (surévaluées) car il existe des redondances dans les banques.

2-2.18.2- Recherche de séquences dans GenBank et EMBL dans l'environnement GCG.

Dans l'environnement GCG, il est aussi possible de faire des recherches de séquences par similarité. Il existe deux programmes disponibles : FASTA, qui utilise la méthode Pearson et Lipman (Pearson et Lipman, 1988) et BLAST, ou Basic Local Alignment Search Tool, qui utilise la méthode de Altschul (Altschul *et al.*, 1990) pour chercher les similarités entre une séquence donnée et toutes les séquences d'une banque. Le programme FASTA est, dans certains cas, plus sensible que BLAST.

Avec le programme FASTA, la comparaison d'une séquence à la totalité des séquences de GenBank et de EMBL, nécéssite les termes par défauts : "GenEMBL:*". La comparaison d'une séquence à une sous-partie de GenBank est possible si la sous-partie est définie à l'origine par la banque (voir Tableau F). La comparaison d'une séquence à une banque de données créée par LookUp nécessite les termes suivants: "@nom de la banque".

Dans toutes les recherches par FASTA, les paramètres par défaut ont été utilisés. Le paramètre "taille du mot" (word size) est 6 (représentant la valeur maximale possible), et implique que l'identité minimale entre deux séquences est déterminée par au moins 6 nucléotides successifs. Le paramètre "E() score" est 2, valeur à partir de laquelle les identités entre deux séquences sont dues au hasard dans les conditions expérimentales utilisées (c'est à dire une séquence dont la taille varie entre 40 et 300 nucléotides).

TABLEAU F : Terme de recherche pour le programme FASTA dans l'environnement GCG (UNIX).

Division de la banque	Recherche dans GenBank et EMBL	Recherche dans GenBank	Recherche dans EMBL
La totalité	GenEMBL:* (GE:*)	GenBank:* (GB:*)	EMBL:* (EM:*)
Les invertebrés	Invertebrate:* (In:*)	GB_In:*	EM_In:*
Les vertébrés non-mammifères	Other_Vertebrate:* (Ov:*)	GB_Ov:*	EM_Ov:*
Les mammifères non-rongeurs et non-primates	Other_Mammalian:* (Om:*)	GB_Om*:	EM_Om:*
Les primates	Primate:* (Pr:*)	GB_Pr:*	EM_Pr:*
Les rongeurs	Rodent:* (Ro:*)	GB_Ro:*	EM_Ro:*
Banque de données créée à partir de LookUp	@nomdelabanque.list		

2-2.18.3- Recherche de séquences sur le site de NCBI.

Des recherches d'identité entre une séquence et la banque de données GenBank sont réalisables à partir du site web de NCBI (voir paragraphe 1.4). Le programme de recherche proposé est BLAST (voir paragraphe 2.17.2). Sa limite est l'impossibilité d'effectuer une recherche sur une fraction de la banque. Il n'a donc été utilisé que pour la recherche de séquences faiblement représentées dans les génomes.

Les séquences qui ont été sélectionnées à partir des recherches de similarité ont été extraites de la banque de donnée en utilisant le programme ENTREZ du site de NCBI. Il permet de sélectionner des séquences à partir d'un ou plusieurs termes. C'est-à-dire que l'on peut retrouver, par exemple, une séquence à partir de son numéro d'accession.

2-2.19- Création d'un programme d'analyse de séquences.

Pour calculer le pourcentage de divergence entre deux séquences, nous avons créé au laboratoire un programme à partir du logiciel Access de Microsoft (travail en collaboration avec D. Demers et J.F. Bideau, étudiants au baccalauréat du département informatique de l'Université de Montréal). Ce programme calcule la divergence d'une séquence par rapport à son consensus ainsi que la divergence moyenne d'un groupe de séquence par rapport à leur consensus. Il détermine le type de mutation d'une séquence par rapport à une autre (transition ou transvertion). Il donne également le nombre de mutations par position pour un alignement de séquences. A partir de ces résultats, il est possible de déterminer si la distribution des mutations d'un ensemble de séquences par rapport à leur consensus est aléatoire.

Le calcul d'identité d'une séquence par rapport à une autre suit l'équation suivante : $1-\left(\frac{S+G}{L}\right)$; où S dénote le nombre de substitutions de nucléotides ; G (gaps) le nombre d'insertions/délétions (compté comme 1 quelque soit la taille) et L la longueur de la séquence de référence (qui est la plupart du temps le consensus). Les "gaps" introduits dans la séquence de référence ne sont pas considérés dans le calcul de L.

RÉSULTATS

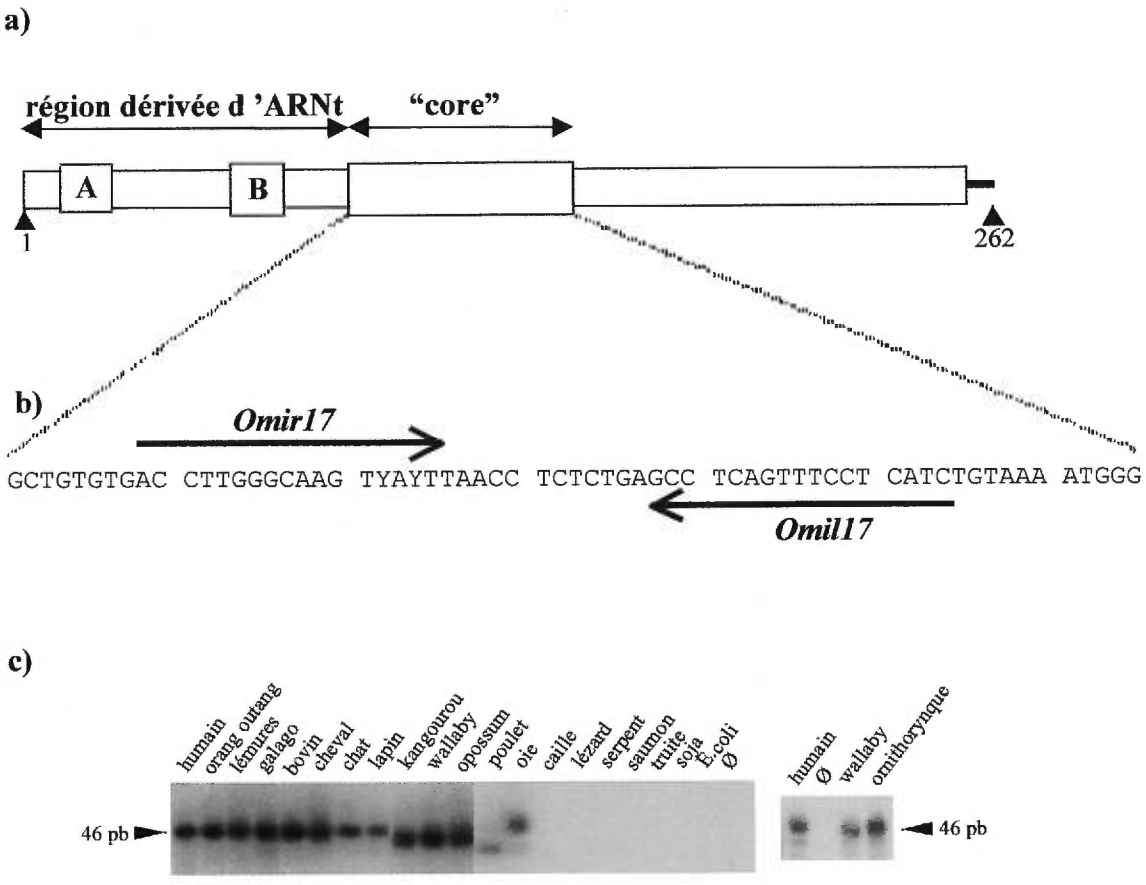
3-1- Caractérisation des éléments CORE-SINE chez les mammifères.

3-1.1- Identification et estimation semi-quantitative des segments "core" chez les mammifères.

Les expériences préalables PCR, réalisées au laboratoire, ont identifié la présence de séquences core dans tous les génomes mammifères : monotrèmes, marsupiaux et placentaires (Jurka *et al.*, 1995). Ces données ont été confirmées par des expériences "intra-MIR" PCR en utilisant les deux amorces *Omil17* et *Omir17*, qui avaient été utilisés pour les réactions "d'inter-MIR" PCR. Les résultats d'amplification révèlent un fragment de 46 nucléotides (Figure 1) correspondant à la taille attendue d'après le consensus MIR établi à partir des éléments placentaires (Donehower *et al.*, 1989; Jurka *et al.*, 1995; Smit et Riggs, 1995). Le fragment amplifié est plus court d'un ou deux nucléotides pour les génomes marsupiaux (Figure 1c). Nous observons aussi que le segment core semble être présent dans les génomes des oiseaux. Cependant, l'amplification n'est pas reproductible dans tous les essais. En effet, la figure 1 montre l'absence de séquence core dans le génome de la caille alors qu'une autre expérience révélait un fragment proche de 46 nucléotides. L'analyse des génomes d'oiseaux sera plus détaillée dans le paragraphe 3-2.1-.

Ayant ainsi identifié la présence d'un segment core chez les mammifères non-placentaires, nous avons fait une estimation du nombre de copies de ce segment chez tous les mammifères. Cette estimation a été obtenue par hybridation sur dot-blot en utilisant les oligonucléotides *Omil17* ou *Omir17* radiomarqués. Les expériences d'hybridation ont été répétées deux à quatre fois selon les génomes avec chacune des sondes, sur différentes

FIGURE 1 : Détection du domaine core par PCR.



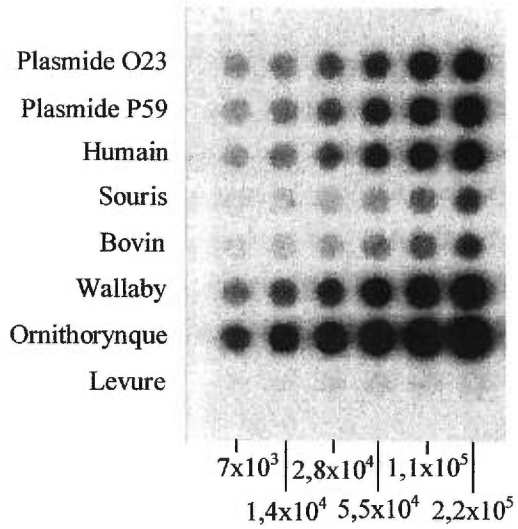
a) Représentation schématique du consensus de la famille MIR proposé par Smit et Riggs (1995). **b)** Détail de la séquence core avec identification des oligonucléotides utilisés pour les réactions de PCR. **c)** Résultat de PCR dite chaude (incorporation d 'α-[^{32}P]-dCTP) pour différents génomes mammifères, oiseaux, reptiles, poissons, plante et bactérie. Les PCR effectuées sans ADN sont notées Ø.

membranes. La figure 2 montre une membrane représentative de l'ensemble des hybridations réalisées. La quantification du signal effectuée à l'aide d'un PhosphoImager et du programme ImageQuant est résumée par la figure 3. Chez la souris et le bovin le signal d'hybridation, suivant les expériences, est entre 3 et 8 fois plus faible que chez l'humain. Ceci peut être expliqué par une horloge moléculaire plus rapide de ces génomes (Li et Tanimura, 1987). L'accumulation de mutations sur les séquences réduit l'affinité pour les sondes et ainsi le nombre de copies détectables diminue. Lors de l'étude préalable, il avait été observé que les éléments core avaient en moyenne 29% de divergence avec leur consensus chez l'Homme et près de 33% chez les bovins (Jurka *et al.*, 1995). En appliquant la correction des divergences aux estimations semi-quantitatives (voir Matériel et Méthodes paragraphe 2-2.8-) le nombre de copies pour les trois génomes placentaires devient similaire et les différences ne sont pas significatives. Ainsi les séquences core seraient au nombre de 300 000 chez l'humain (+/- 100 000) et 200 000 chez la souris ou le bovin (+/- 100000) (Figure 3).

À l'inverse de la souris et du bovin, le signal d'hybridation observé pour les génomes non-placentaires est de deux à quatre fois plus intense que celui observé pour le génome humain (Figure 2). Cela suggère que l'amplification a été plus efficace et (ou) plus récente, ou encore que les segments core ont été mieux conservés dans ces génomes. Cette dernière observation suggère aussi que les segments core de ces génomes puissent faire partie d'éléments mobiles encore actifs du point de vue de leur amplification. Nous avons donc décidé de cloner et d'analyser les segments core dans les génomes non placentaires pour déterminer s'ils font partie de la famille SINE MIR décrite par Smit et Riggs (1995).

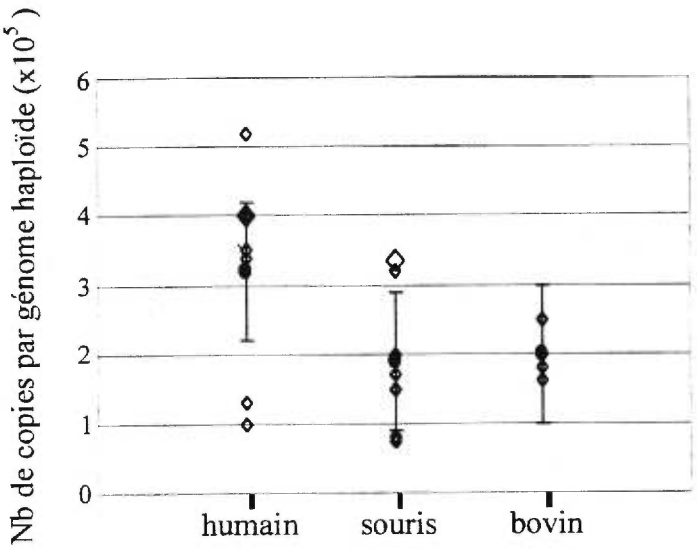
FIGURE 2 : Détection semi-quantitative du domaine core par hybridation.

(dot-blot hybridé avec la sonde *Omil17*)



Sur les deux premières lignes est déposée une quantité croissante de plasmides contenant une séquence core ($2{,}5x10^8$, $5x10^8$, 10^9, $2x10^9$, $4x10^9$, $8x10^9$ copies). Ensuite, a été déposé l'ADN de trois mammifères placentaires, d'un marsupial, d'un monotrème et d'une levure qui sert de témoin négatif à l'hybridation (Le nom des espèces est indiqué à gauche de chaque ligne). Le nombre de génomes haploïdes pour chaque dépôt est indiqué en dessous de chaque colonne.

FIGURE 3 : Estimation du nombre de copies de la séquence core.



Pour chaque expérience, un losange représente l'estimation du nombre de copies du domaine core dans le génome de l'espèce considérée. Les points noirs indiquent la moyenne calculée à partir des losanges et les écarts types sont représentés par les barres verticales. L'estimation du nombre total de copies par génome a été obtenue en considérant la divergence moyenne des éléments par rapport à leur consensus. Cette correction (voir Matériel et Méthodes, paragrahe 2-2.8-) est basée sur les résultats de Jurka *et al.* (1995) qui ont établi que les séquences core chez l'humain ont une divergence de 29% environ par rapport à leur consensus. Pour les séquences non humaines nous avons utilisé la valeur de 33%.

3-1.2- Création des banques subgénomiques et sélection des clones.

Pour déterminer les séquences des segments core identifiés par hybridation et PCR, nous avons décidé de créer des banques subgénomiques pour trois génomes de mammifères non placentaires. L'ornithorynque est le représentant des génomes des monotrèmes, l'opossum de Virginie des marsupiaux d'Amérique du Nord et un wallaby de ceux d'Australie. L'ADN nucléaire de chacune de ces trois espèces a été fractionné par nébulisation dans les conditions décrites au paragraphe 2-2.5-. Les fragments d'une taille moyenne de 300 nucléotides (variant de 200 à 800) ont été clonés dans le vecteur pBS KS+ pour donner les librairies subgénomiques. Les trois banques ont ensuite été criblées par hybridation en utilisant les sondes oligonucléotidiques radiomarquées, *Omil17* et *Omir17*, spécifiques du fragment core (Figure 1) (Jurka *et al.*, 1995). Nous avons sélectionné près de 100 clones pour les trois espèces et nous avons obtenu par séquençage automatique 36 éléments provenant de l'ornithorynque, 25 de l'opossum et 44 du wallaby. L'identification des segments core a été effectuée par comparaison dot-matrix avec le consensus MIR en utilisant le logiciel DNAsis Pro version 3.5.

3-1.3- Analyse des séquences core.

L'alignement des séquences pour chaque espèce a été initié par le segment core lui-même et a ensuite été prolongé vers les extrémités. Les alignements ont été construits grâce au programme Multalin (Corpet, 1988), mais ont été raffinés de façon manuelle pour optimiser les identités. La comparaison entre les éléments d'une même espèce ou entre ceux d'espèces différentes nous a permis d'établir l'existence de 5 familles de séquences distinctes. À partir de chaque famille nous avons pu créer un consensus (voir paragraphe 1-2.2-) qui représente la séquence ancestrale potentiellement active.

Les cinq familles ont été nommées en fonction de leur appartenance aux différents génomes étudiés, c'est-à-dire Mon-1, Opo-1, Mar-1, Ther-1 et Ther-2 pour monotrèmes, opossum, marsupiaux et thériens. La famille Mon-1, représentée par des éléments longs de 273 nucléotides, est spécifique des monotrèmes (Figure 4A). Opo-1, qui est plus court d'environ 90 nucléotides, possède un segment core tronqué du coté 3' et n'a été trouvé que dans le génome de l'opossum (Figure 4B). En revanche, les éléments Mar-1, dont le consensus fait 234 nucléotides de long, ont été identifiés aussi bien dans les génomes marsupiaux d'Amérique du Nord que d'Australie (Figure 4C). De la même façon, les éléments Ther-1 sont présents chez tous les marsupiaux (Figure 4D). Il faut noter que le consensus Ther-1 (264 nucléotides) est fortement identique au consensus de MIR, ces éléments sont donc très proches des séquences identifiées dans les génomes placentaires par Smit et Riggs (1995). Ther-2, quant à lui, a été retrouvé sous la forme de deux sous-familles très proches, une chez le wallaby, Wal Ther-2, et l'autre chez l'opossum, Opo Ther-2. Ces deux sous-familles ne diffèrent l'une de l'autre que par quelques positions diagnostiques dans le segment 5' (Figure 4E, Tableau 1).

Pour comparer les éléments homologues de Ther-1 des génomes placentaires avec ceux séquencés chez les non-placentaires, nous avons sélectionné les 30 éléments ayant une identité avec le domaine core à partir d'un segment d'ADN humain de près de 160 Kb provenant de 4 clones (U59962, U61375, U63313, U63312). Parmi ces 30 éléments, 28 appartiennent à la famille Ther-1 (Figure 4F), et leur consensus est presque identique à Mar Ther-1 et MIR (93% et 92% respectivement) (Figure 5). Les deux séquences restantes sont similaires à la famille Ther-2.

Tous les éléments analysés possèdent le segment core ainsi qu'un segment 5' dérivé d'ARNt décrit dans le consensus de MIR des génomes placentaires (appelée "région

FIGURE 4 : Alignements de séquences.

Les informations qui suivent sont valables pour tous les alignements. La première séquence, en général un consensus, sert de référence. Le nom des séquences est indiqué sur la gauche, et lorsque celles-ci proviennent de GenBank, c'est le numéro d'accession qui est indiqué. Sur l'alignement, les nucléotides identiques à la séquence de référence sont représentés par des points. Les trais d'unions représentent des délétions. Le caractère X est utilisé pour remplacer des insertions/délétions d'au moins 2 nucléotides. La nomenclature des nucléotides est indiquée dans la liste des abréviations.

Les informations pour la figure 4 uniquement sont les suivantes.
Les séquences de l'ornithorynque sont précédées de la lettre P, celles du wallaby de la lettre W et celles de l'opossum de la lettre O. Les séquences humaines provenant des clones U59962, U61313, U63313 et U63312 sont numérotées de 1 à 30. Les autres sont identifiées par leur numéro d'accession. Les nucléotides des consensus en caractère gras identifient les positions diagnostiques des sous-familles. Les segments core sont encadrés tandis que les régions flanquantes sont délimitées par des flèches. Les "blancs" aux extrémités des séquences sont dus à l'absence de donnée ou au fait que des segments ne possèdent aucune identité avec la séquence de référence. (Les séquences des trois espèces non placentaires ont été déposées dans GenBank et ont les numéros d'accessions suivants: AF055483-AF055511 pour l'ornithorynque, AF55512-AF55533 pour l'opossum et enfin AF055534-AF055577 pour le wallaby.)

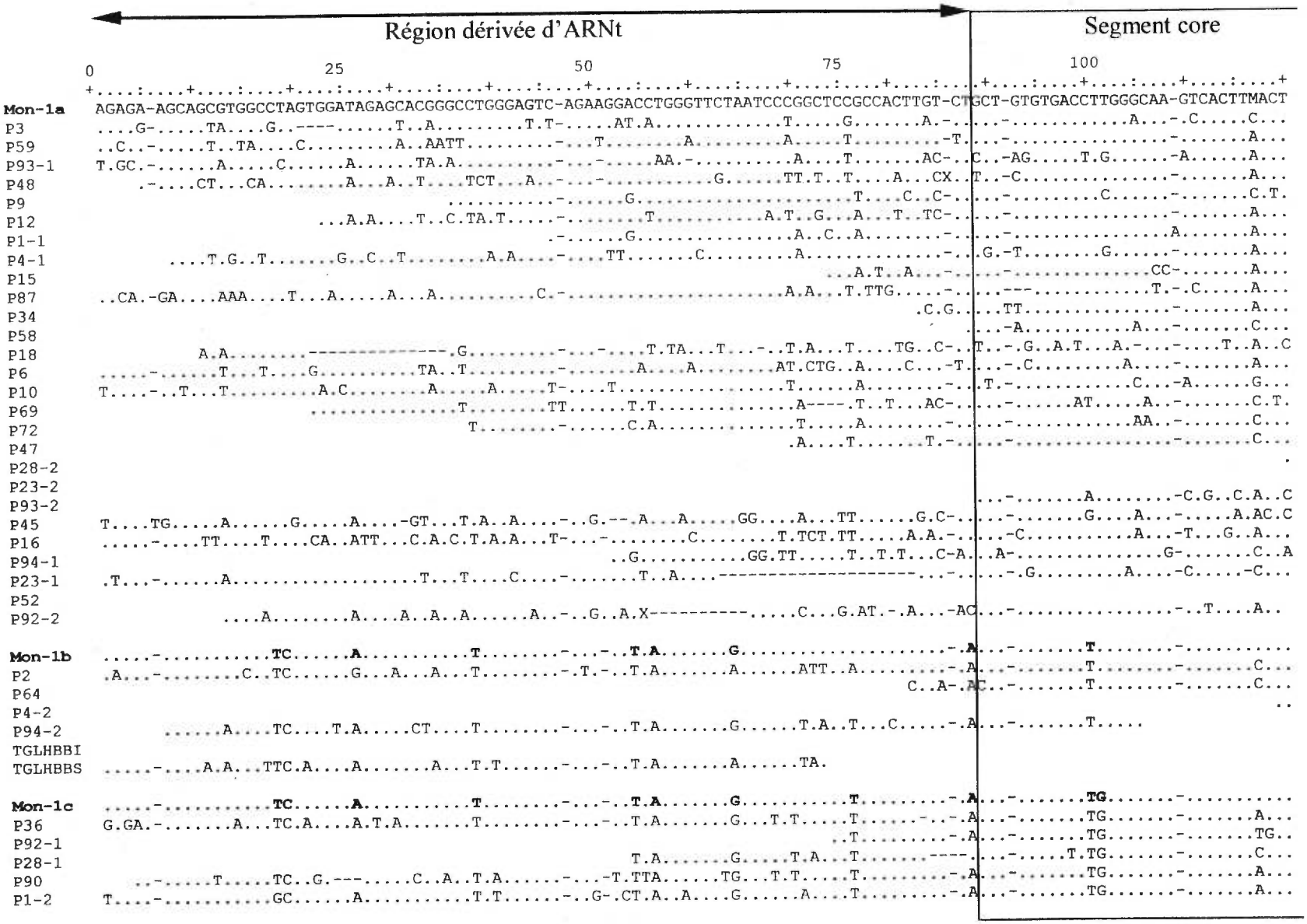


FIGURE 4A : Alignement des séquences de la famille Mon-1 (1/3).

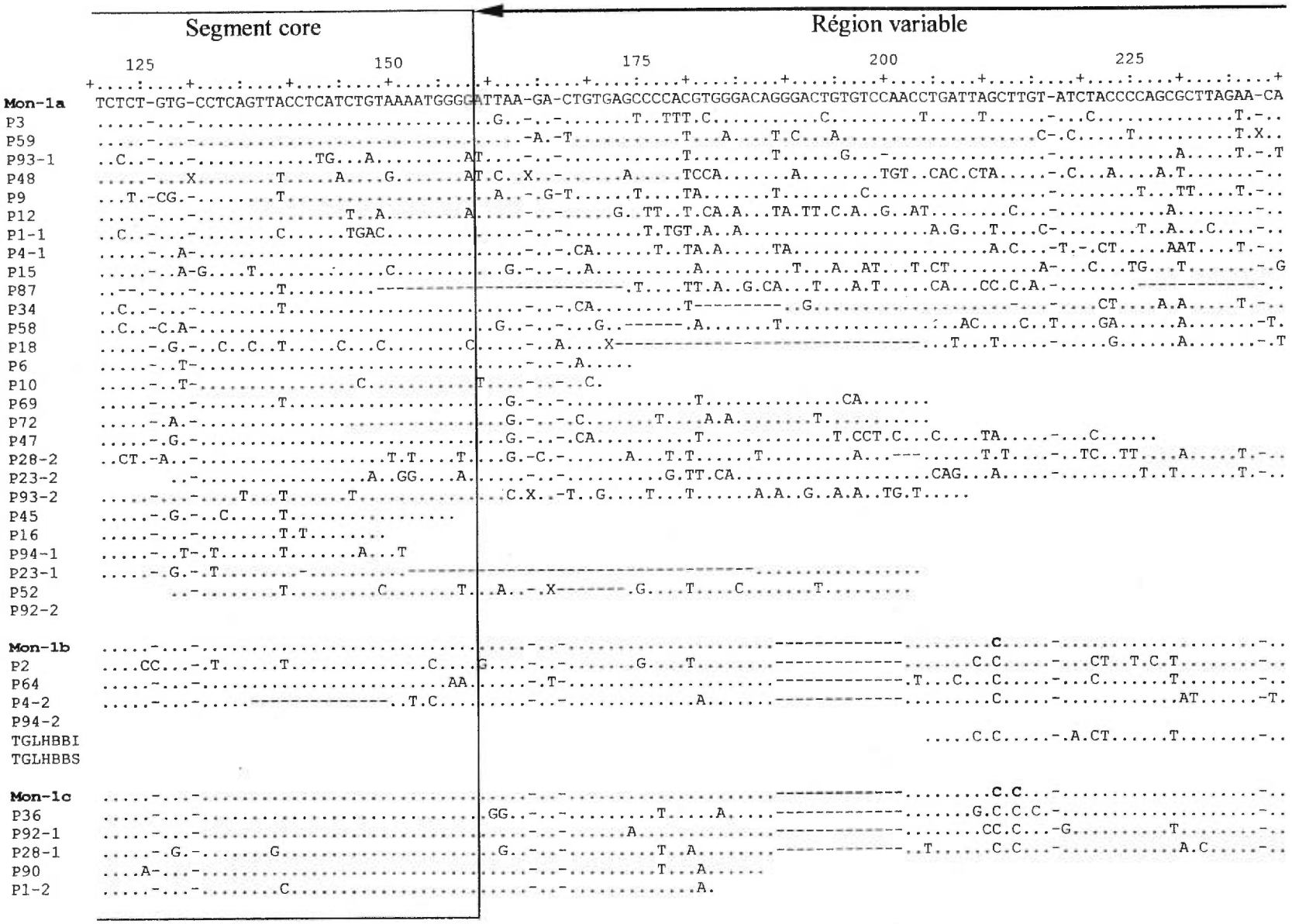


FIGURE 4A : Alignement des séquences de la famille Mon-1 (2/3).

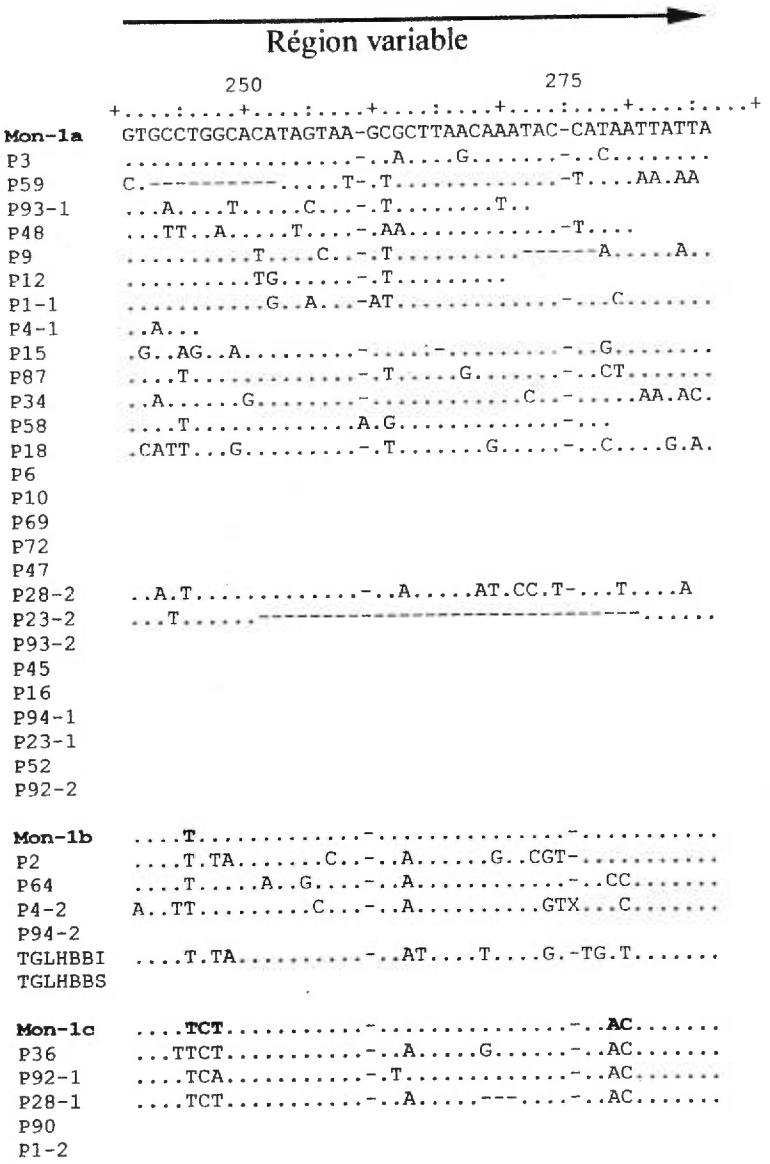


FIGURE 4A : Alignement des séquences de la famille Mon-1 (3/3).

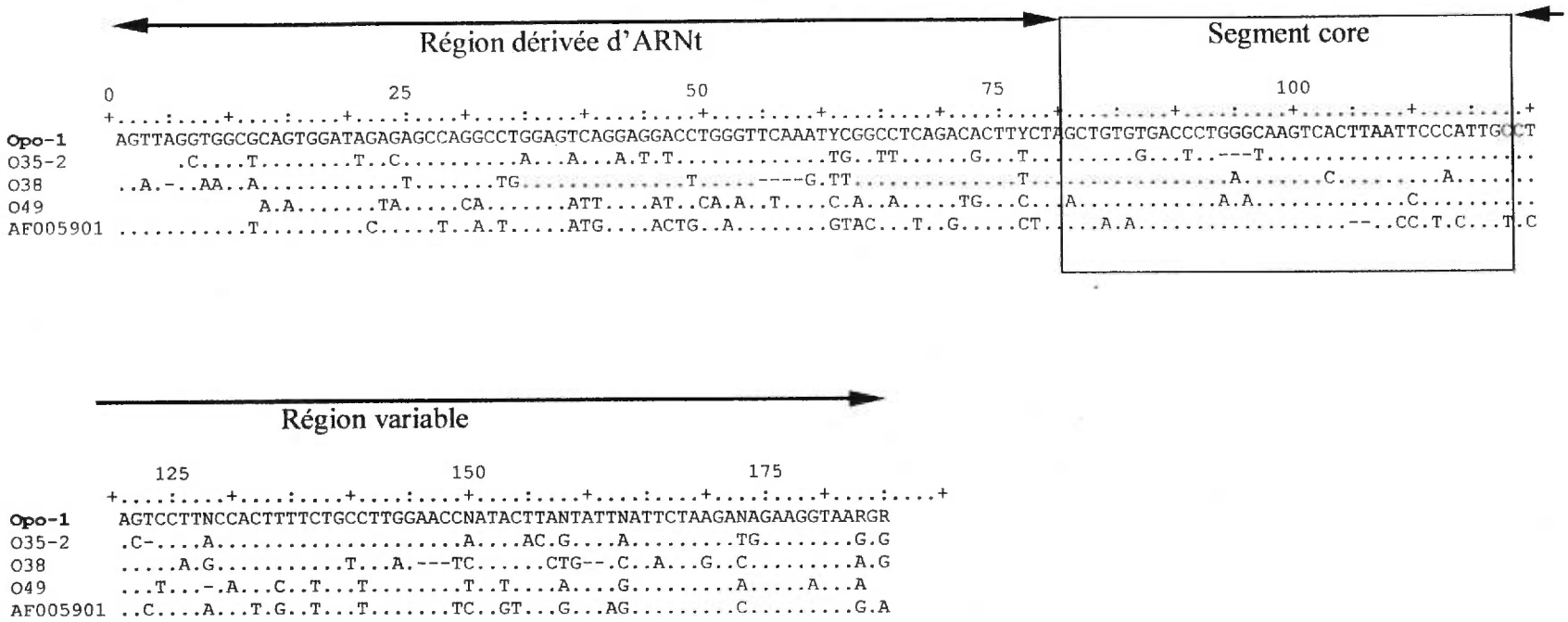


FIGURE 4B : Alignement des séquences de la famille Opo-1.

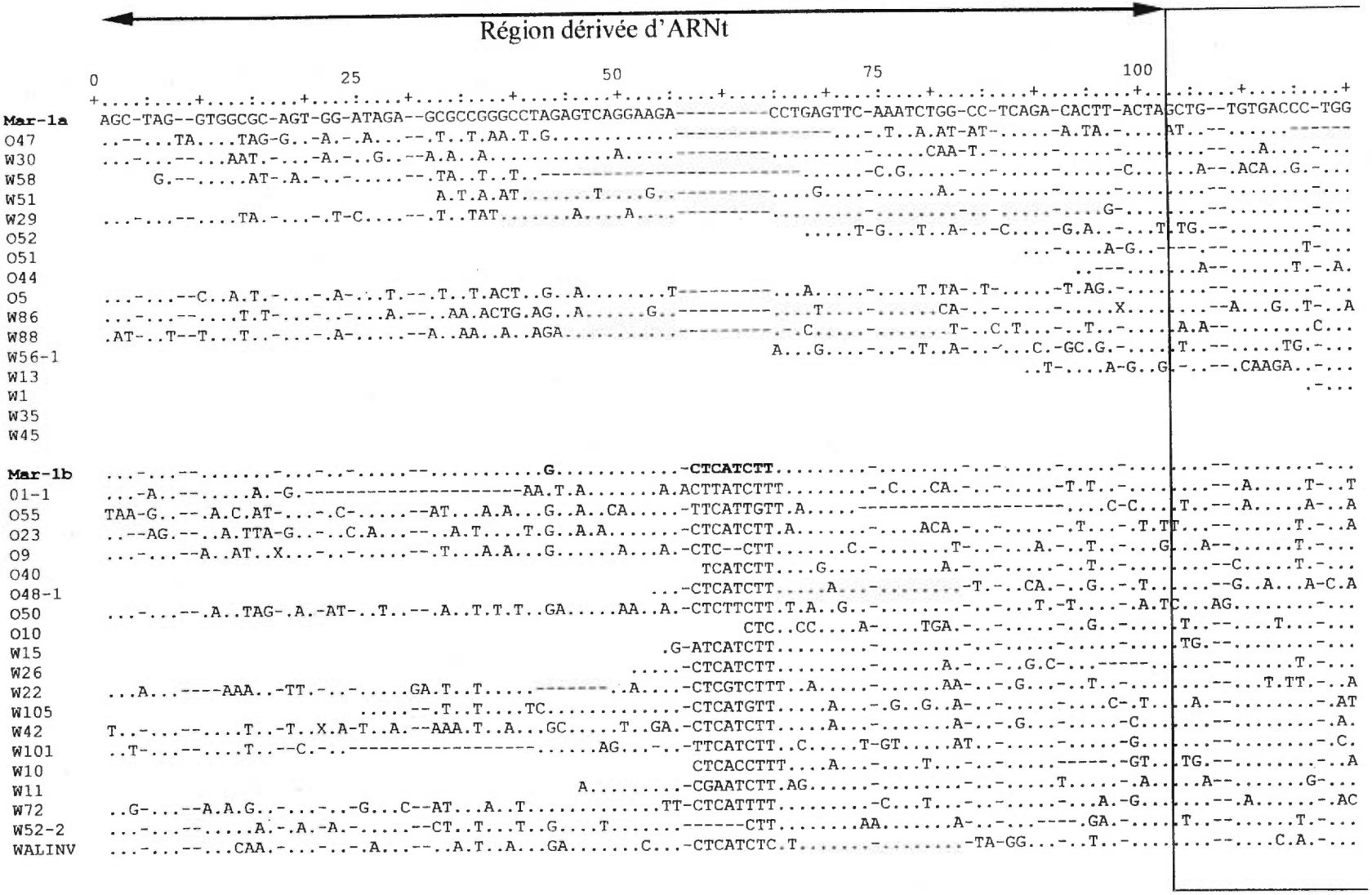


FIGURE 4C : Alignement des séquences de la famille Mar-1 (1/3).

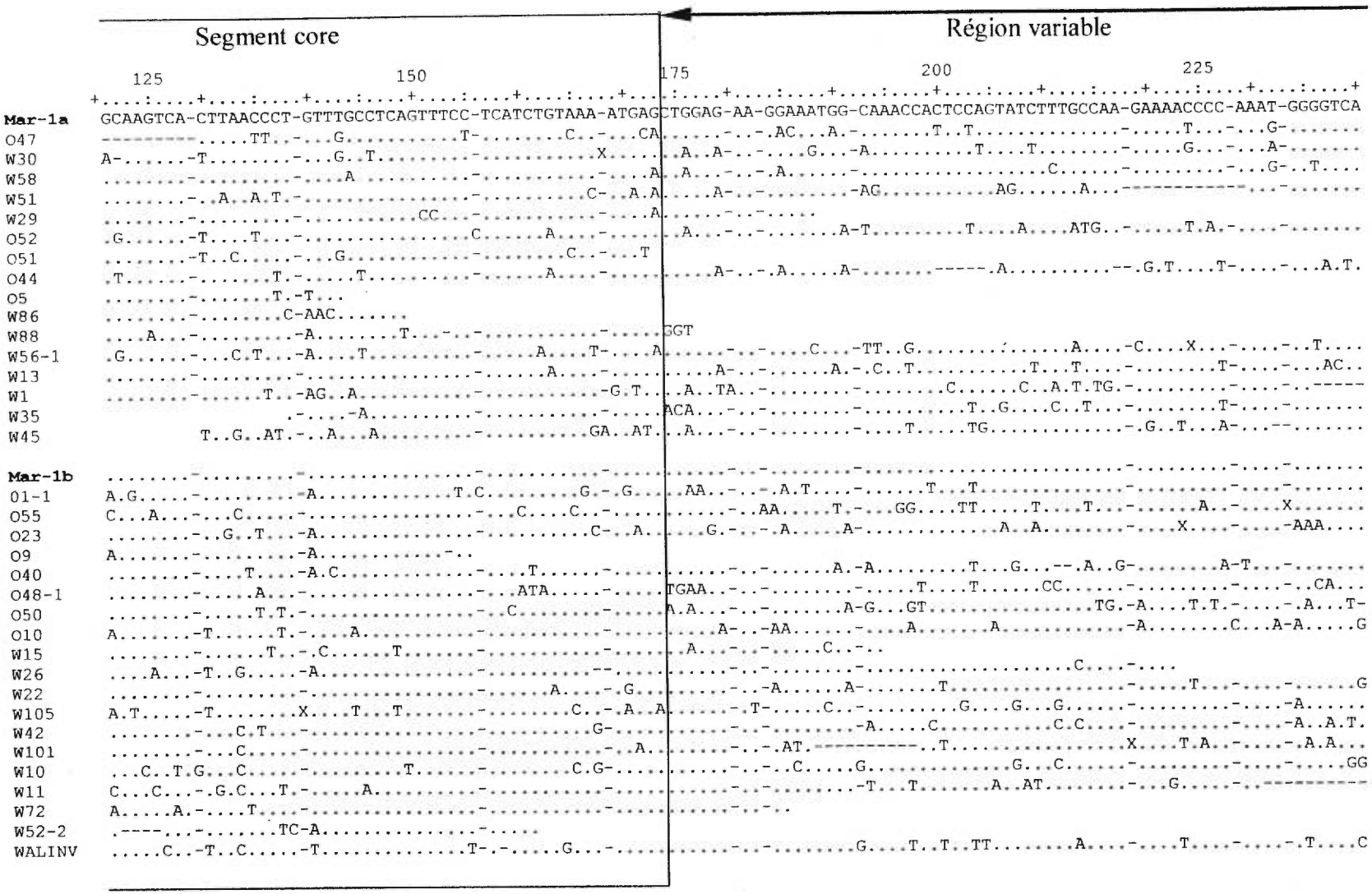


FIGURE 4C : Alignement des séquences de la famille Mar-1 (2/3).

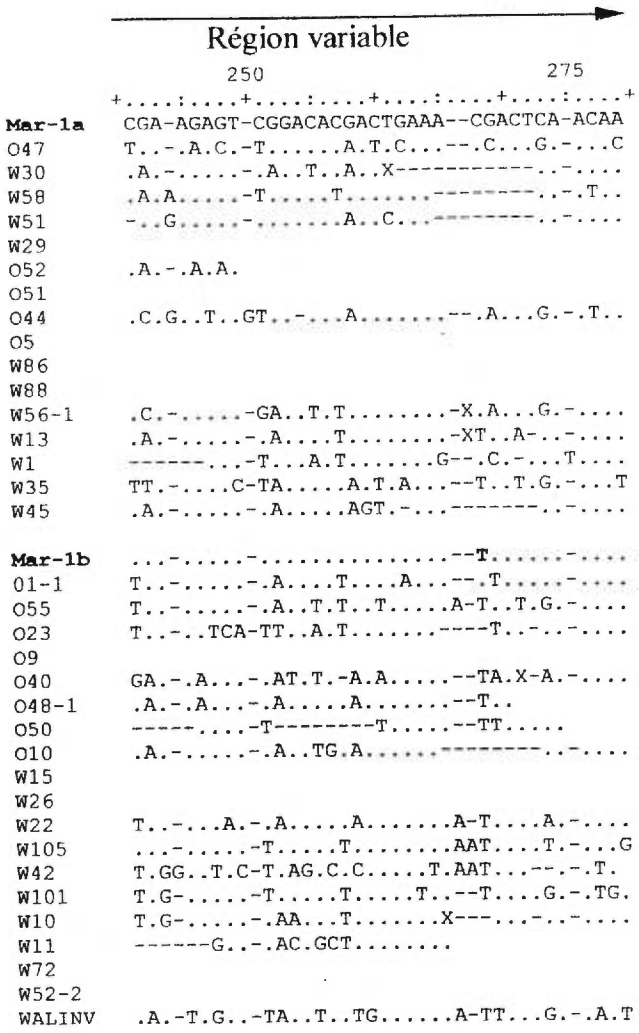


FIGURE 4C : Alignement des séquences de la famille Mar-1 (3/3).

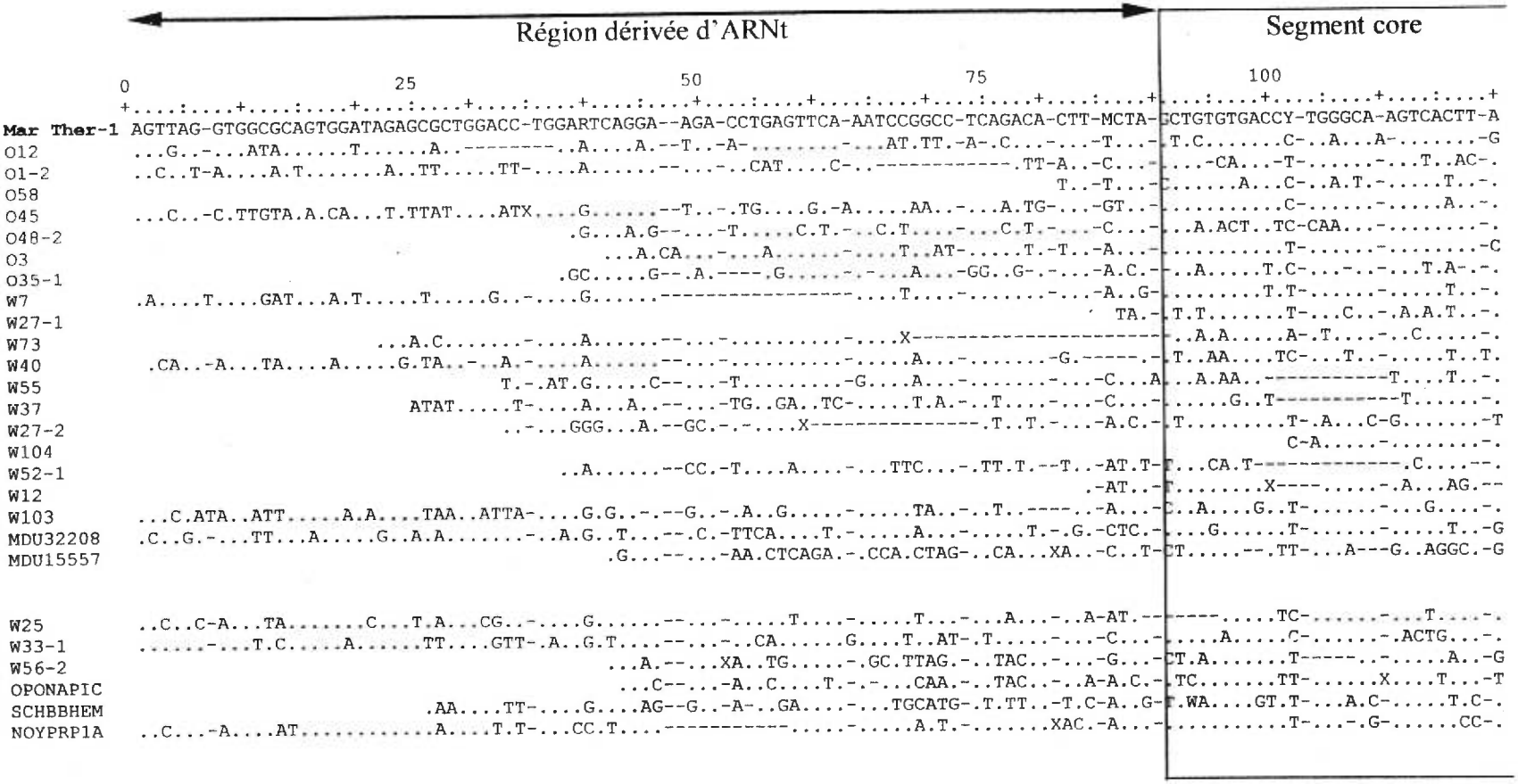


FIGURE 4D : Alignement des séquences de la famille Ther-1 (sous-famille Mar Ther-1) (1/3).

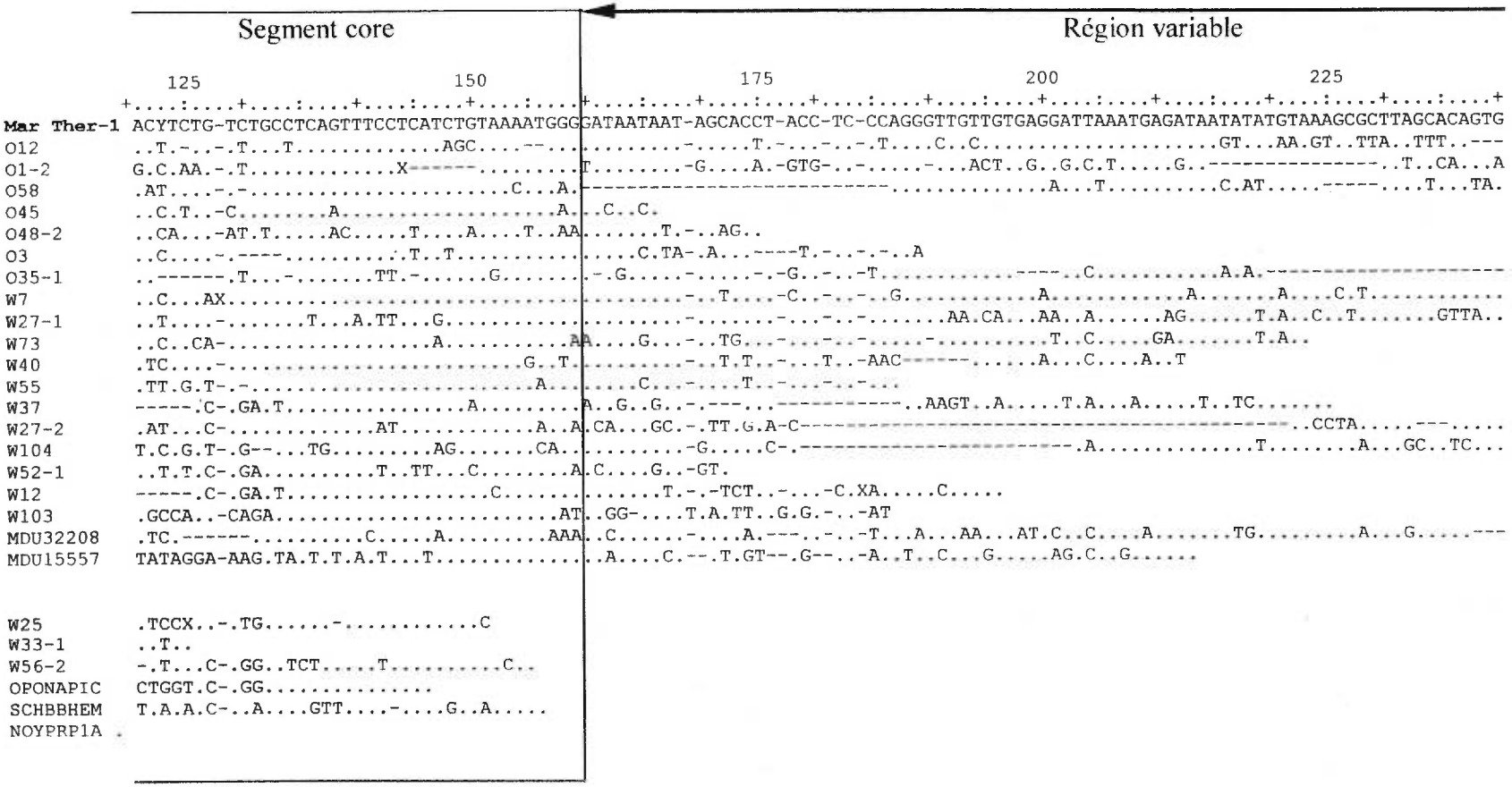


FIGURE 4D : Alignement des séquences de la famille Ther-1 (sous-famille Mar Ther-1) (2/3).

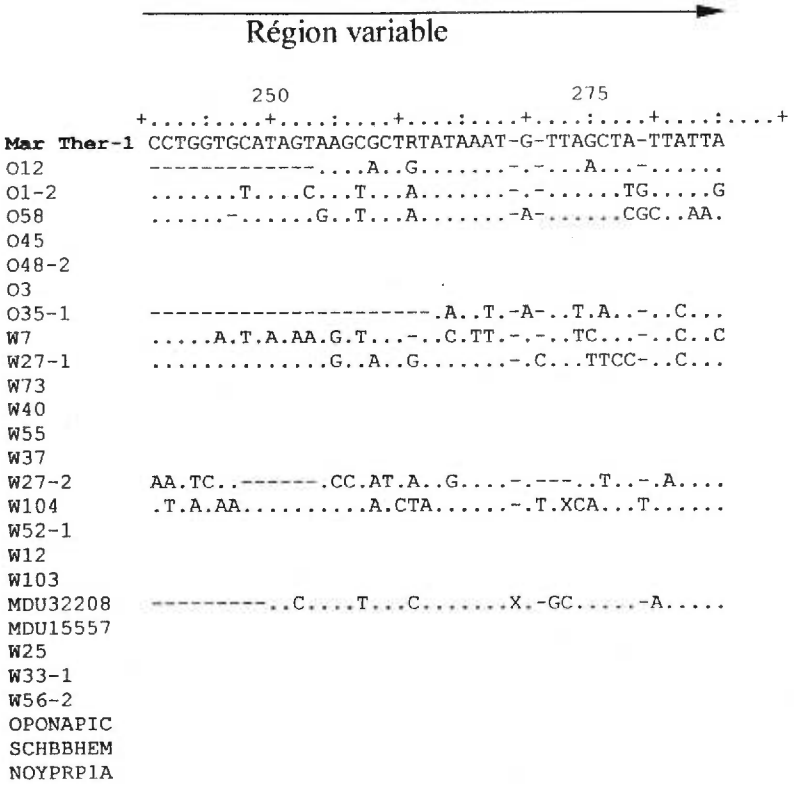


FIGURE 4D : Alignement des séquences de la famille Ther-1 (sous-famille Mar Ther-1) (3/3).

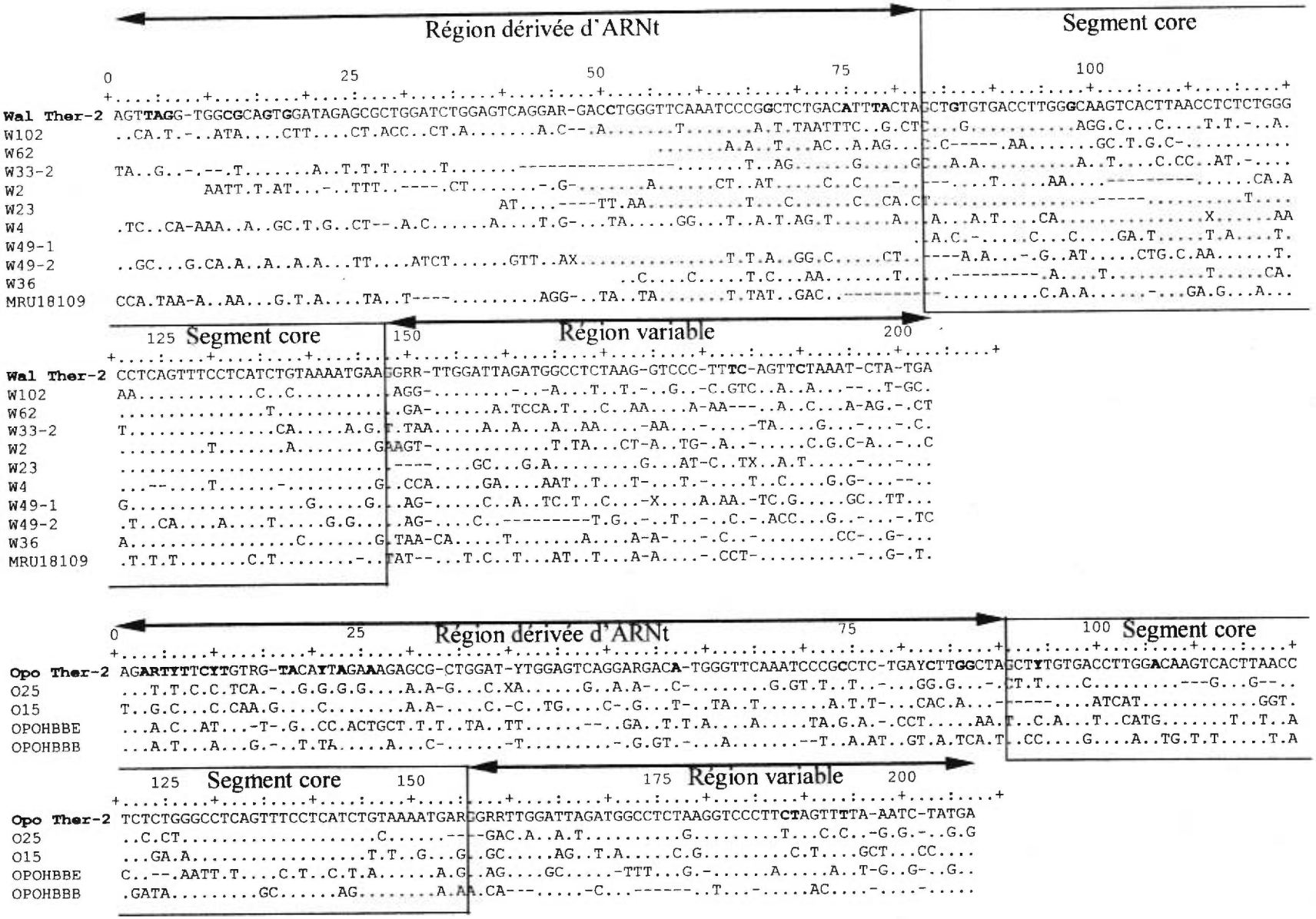


FIGURE 4E : Alignement des séquences de la famille Ther-2 (sous-familles Wal et Opo Ther-2).

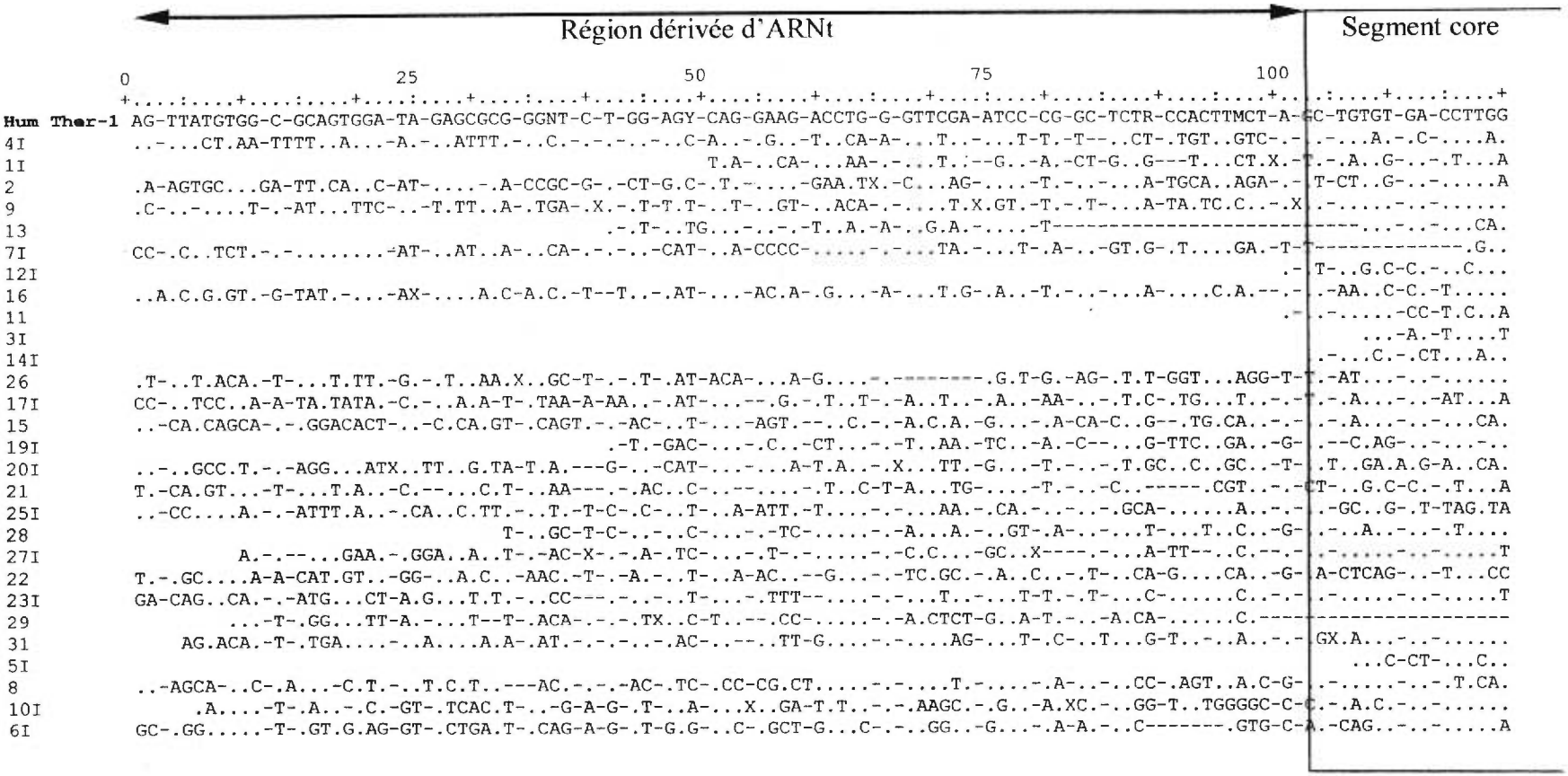


FIGURE 4F : Alignement des séquences de la famille Ther-1 (sous-famille Hum Ther-1) (1/3).

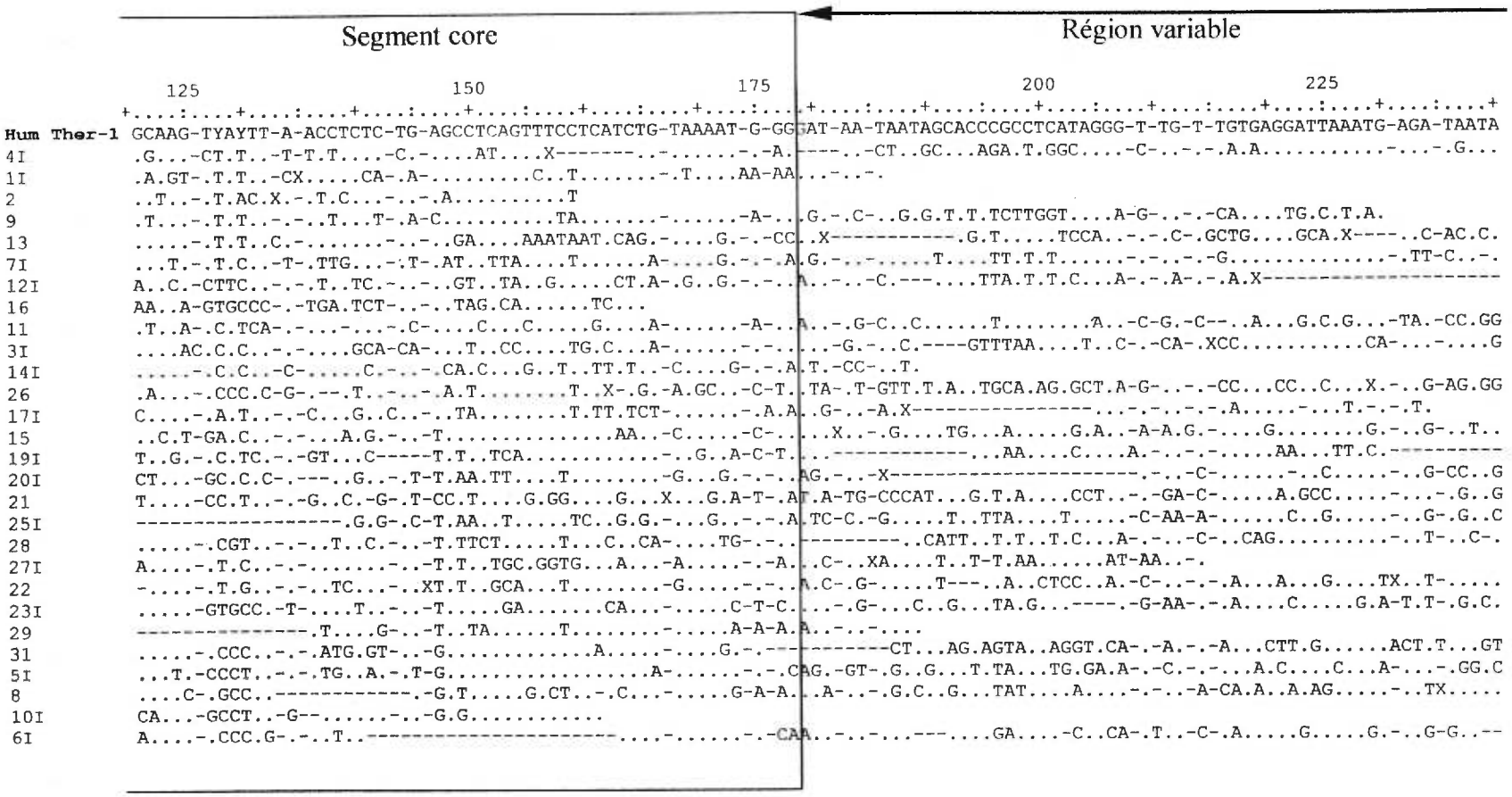


FIGURE 4F : Alignement des séquences de la famille Ther-1 (sous-famille Hum Ther-1) (2/3).

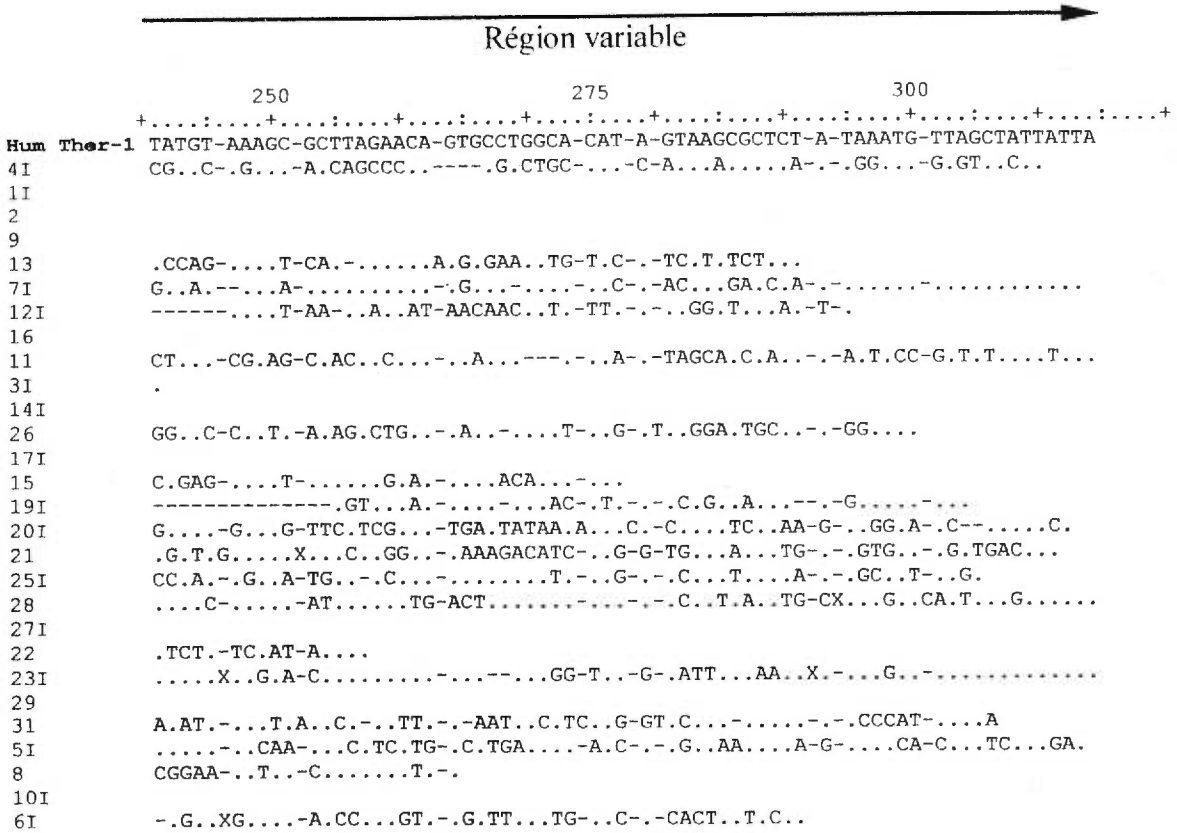


FIGURE 4F : Alignement des séquences de la famille Ther-1 (sous-famille Hum Ther-1) (3/3).

Ther-1	positions	6	29	33	53	58	64	70	72	115	117-118	168-169	217
Hum Ther-1		T	G	T	G	G	C	T	C	C	GA	AT	A
Mar Ther-1		G	T	C	A	A	G	A	A	G	CT	CC	C

Ther-2	positions [a]	3	4-6	15	17	20	50	66	74	77-78	85	97	189-190	195
Wal Ther-2		-	TAG	G	G	T	C	G	A	TA	G	G	TC	C
Opo Ther-2		ins. 5	CYT	Y	A	A	A	C	C	GG	Y	A	CT	T

Mar-1	positions	35	48
Mar-1a		A	-
Mar-1b		G	ins. 8

Mon-1	positions	17-18	25	37	49	52	54	62	73	85	96	97	180-192	202	204	239	240-241	265-266
Mon-1a		CT	T	C	A	A	C	T	C	T	C	T		G	T	C	TG	TA
Mon-1b		**TC**	**A**	**T**	-	**T**	**A**	**G**	C	**A**	**T**	T	**del.**	**C**	T	**T**	TG	TA
Mon-1c		**TC**	**A**	**T**	-	**T**	**A**	**G**	***T***	**A**	**T**	***G***	**del.**	**C**	***C***	**T**	***CT***	***AC***

Les positions données font référence au consensus de la famille considérée. Les lettres ins. et del. signifient respectivement insertion et délétion. Les chiffres après ins. correspondent aux nombres de nucléotides supplémentaires. Les positions diagnostiques communes à Mon-1b et Mon-1c sont notées en caractères gras, les positions diagnostiques spécifiques de Mon-1c en italique.

[a] : Pour la famille Ther-2 le consensus de référence est Wal Ther-2.

TABLEAU 1 : Positions diagnostiques des sous-familles.

dérivée d'ARNt") (Smit et Riggs, 1995). Seule la région 3' en aval du segment core diffère pour chacune des familles identifiées. Nous avons appelé ce segment la "région variable" (Figure 5). Sa taille varie entre 50 et 123 nucléotides et confère la spécificité à chacune des familles. Il existe le cas particulier des familles Mon-1 et Ther-1 qui partagent les 58 derniers nucléotides de leurs séquences (83% d'identité). Cependant, les 66 nucléotides restant, entre le core et ce segment de 58 pb, sont différents dans chacune des deux familles et confèrent ainsi la spécificité (Figure 5).

3-1.4- Caractéristiques propres aux rétroposons SINE.

Le promoteur à ARN Pol III est la première caractéristique des rétroposons de type SINE et est retrouvé dans le segment dérivé d'ARNt de toutes les familles identifiées ci-dessus (Figure 5).

La présence de répétitions simples en tandem, à l'extrémité 3' de la séquence, est la seconde caractéristique commune aux SINE. Les familles Ther-1 et Mon-1 partagent la même répétition TTA à leur extrémité et la famille Mar-1 présente la répétition AAC. Ces trinucléotides sont, en général, répétés deux à trois fois et exceptionnellement jusqu'à 10 fois. De façon relativement fréquente, nous avons observé des combinaisons de trinucléotides, tels que TTA et TTG pour les familles Ther-1 et Mon-1. Aucune répétition simple n'a été observée pour la famille Opo-1 et une seule combinaison de trinucléotides, ATC et TTC, a été trouvée pour un élément de la famille Ther-2.

Les répétitions terminales directes flanquantes (RTD), qui résultent de la duplication du site d'insertion, sont aussi une caractéristique des rétroposons, mais ne font pas partie intégrante de l'élément. Des RTD ont été reconnues pour un petit nombre d'éléments analysés. Seuls 4 éléments "pleine longueur", parmi les 7 provenant du génome de wallaby,

FIGURE 5 : Alignement des consensus de familles et sous-familles CORE-SINE mammifères.

Le nom de chaque consensus est indiqué sur la gauche. La séquence MIR est le consensus publié par Smit et Riggs (Smit et Riggs, 1995). Les boîtes A et B du promoteur bipartite de l'ARN Pol III sont encadrées. Le segment core est encadré en pointillé et les régions adjacentes sont délimitées par les flèches. Les oligonucléotides de 40 bases utilisés pour les Sourthern-blot sont soulignés : T1 sur la séquence Hum Ther-1, T2 sur Wal Ther-2, M1 sur Mar-1a, OP1 sur Opo-1 et MO1 sur Mon-1a.

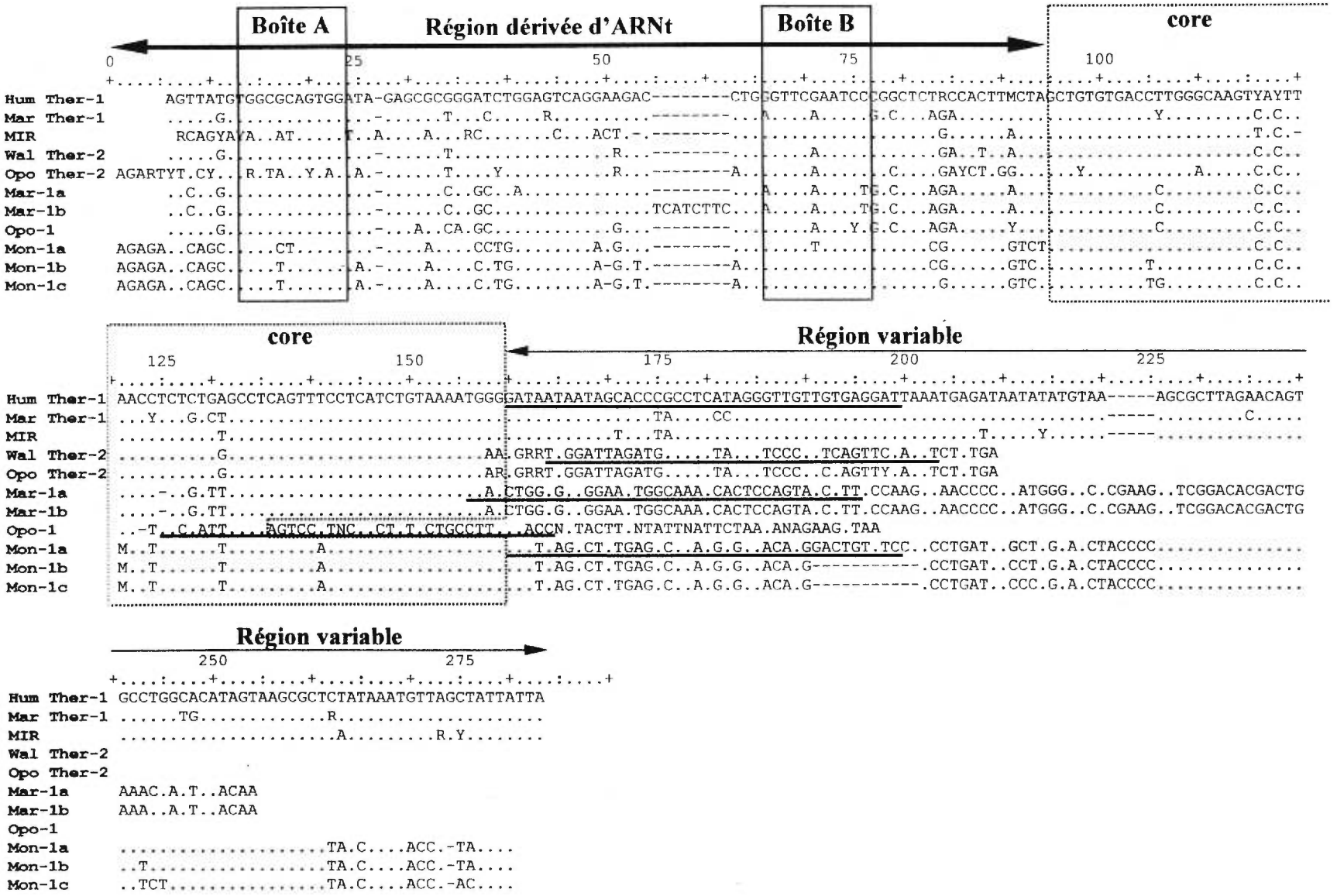

Boîte A
Région dérivée d'ARNt
Boîte B
core
Région variable
Hum Ther-1
Mar Ther-1
MIR
Wal Ther-2
Opo Ther-2
Mar-1a
Mar-1b
Opo-1
Mon-1a
Mon-1b
Mon-1c

FIGURE 5

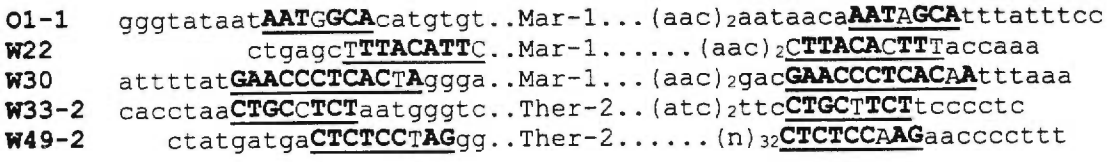
possèdent des RTD de 8 à 12 nucléotides de long avec des identités variant de 78 à 92% (Figure 6). Une autre RTD a été trouvée dans le cas du clone O1 (génome de l'opossum), où un élément de la famille Mar-1, O1-1, s'est intégré dans un élément de la famille Ther-1, O1-2. Le site d'insertion est reconnu par les RTD de 7 nucléotides qui possèdent une identité de 86% (Figure 6).

La dernière caractéristique des éléments rétroposables issus d'une longue période d'amplification, est leur organisation en sous-familles (Jurka et Smith, 1988; Quentin, 1988; Labuda et Striker, 1989). Les consensus des sous-familles sont des représentations approximatives des éléments actifs responsables de l'amplification et se distinguent les uns des autres par des positions diagnostiques. Cette subdivision a été possible pour toutes les familles (Figures 4 et 5, Tableau 1). Nous avons sélectionné deux types de sous-familles.

Les premières sont spécifiques du génome d'origine. Par exemple Ther-1 possède des sous-familles qui sont spécifiques soit des génomes mammifères placentaires soit des génomes marsupiaux. Elles sont distinguées dans la nomenclature en utilisant les trois premières lettres du génome d'origine : Hum Ther-1 ou Mar Ther-1 (pour humain ou marsupial) (Figure 4D et 4F, Tableau 1). La même distinction a été faite pour la famille Ther-2, où l'on sépare les éléments provenant du wallaby, Wal Ther-2, des éléments provenant de l'opossum, Opo Ther-2 (Figure 4E, Tableau 1). Pour la famille Ther-2, le nombre de séquences identifié étant faible, nous n'avons pas pu caractériser de sous-famille commune aux deux génomes.

Le second type de sous-familles est identifié à l'intérieur d'un même génome. Dans ces situations nous rajoutons pour la nomenclature des sous-familles une lettre après le nom. C'est le cas pour la famille Mar-1 qui peut être divisée en deux groupes, Mar-1a et Mar-1b (Figure 4C, Tableau 1). Il en est de même pour la famille Mon-1 où nous avons

FIGURE 6 : Identification de répétitions terminales directes (RTD).



Les noms des éléments sont indiqués sur la gauche. Le nom de la famille à laquelle appartient la séquence est indiqué au centre. Les caractères en majuscule et soulignés représentent les RTD. Lorsque les nucléotides ce celles-ci sont en caractères gras cela indique leur identité. Entre parenthèse sont indiqués les répétitions simples et le nombre de réitération.

distingué trois sous-familles (Mon-1a, Mon-1b et Mon-1c) (Figure 4A, Tableau 1). Nous avons constaté que ces dernières suivent un ordre séquentiel, c'est à dire que Mon-1c dérive de Mon-1b, qui elle-même aurait dérivé de Mon-1a. Cet ordre séquentiel est confirmé par l'analyse des divergences et la datation des séquences (paragraphe 3-1.7-).

Ayant trouvé 5 familles d'éléments dispersés possédant toutes les caractéristiques des SINE dérivés d'ARNt et toutes munies d'une structure centrale core conservée, nous avons décidé de définir une super-famille de rétroposons appelée CORE-SINE. Cette super-famille regroupe les éléments SINE possédant le domaine core de 65 nucléotides.

3-1.5- Distribution des familles CORE-SINE chez les mammifères.

Pour compléter l'identification des nouvelles familles CORE-SINE chez les mammifères nous avons effectué des recherches dans les banques de données accessibles par internet et nous avons vérifié leur distribution dans les différents génomes.

3-1.5.1- Recherche dans les banques de données.

Nous avons commencé par vérifier la présence des nouvelles familles core chez les mammifères placentaires, et de façon plus précise chez l'Homme. En utilisant le programme BLAST, chaque criblage de la base de donnée de GenBank avec un consensus "pleine longueur" d'une famille, a abouti sensiblement au même résultat dominé par la présence du domaine core. Nous avons ensuite utilisé pour la comparaison les extrémités 3' variables, spécifiques de chaque CORE-SINE. Aucun élément possédant une identité avec les familles Mon-1, Mar-1 et Opo-1 n'a été retrouvé chez les mammifères placentaires. A l'inverse, de nombreux éléments possédant une forte identité avec le fragment spécifique de

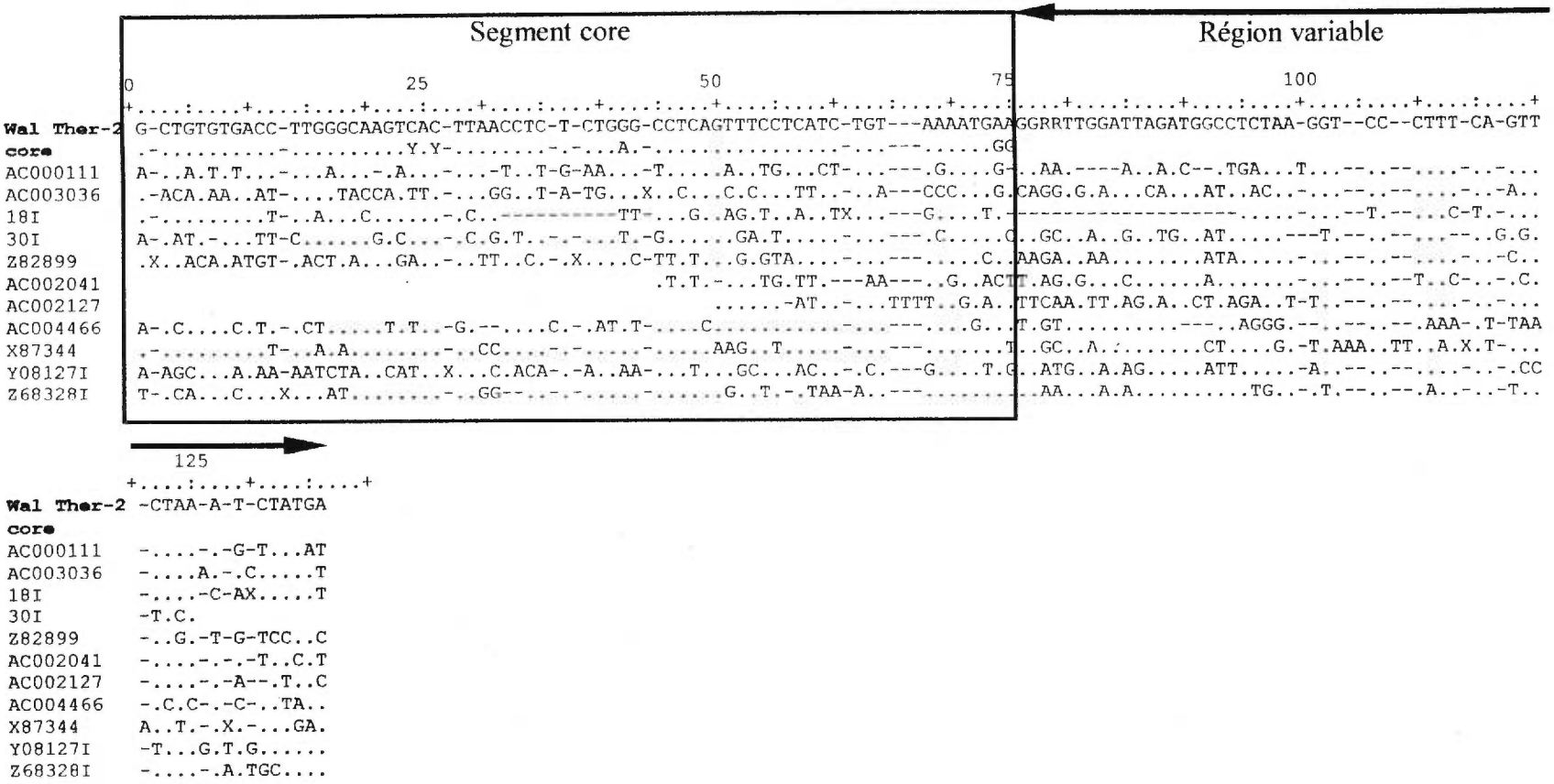
Ther-1 ont été détectés. Chez l'humain une dizaine de séquences a été trouvée présentant des similarités avec les 50 nucléotides du segment 3' de Ther-2.

Pour construire un alignement des éléments Ther-2 placentaires nous avons sélectionné 11 séquences humaines obtenues par la recherche avec le programme BLAST (Figure 7). Nous avons utilisé le consensus de la sous-famille Wal Ther-2 comme séquence de référence. L'alignement obtenu pour les éléments Ther-2 du génome humain ne permet pas de construire un consensus, essentiellement à cause du segment 5', en amont du core, qui est fort divergent. Cependant l'association du fragment spécifique avec le core est observée et démontre l'existence des éléments de la famille Ther-2 chez l'humain.

Les recherches de séquences dans les banques de données étant dominées par les séquences des génomes placentaires, nous avons créé nos propres banques de données spécifiques des génomes monotrèmes ou des génomes marsupiaux à l'aide du programme Lookup (voir paragraphe 2-2.18.1-). La recherche avec les séquences consensus entières ou avec les fragments 3' spécifiques des familles CORE-SINE a permis d'identifier douze nouveaux éléments : 2 séquences monotrèmes appartenant à la famille Mon-1 (Figure 4A), sept séquences marsupiales appartenant aux CORE-SINE Opo-1 (1), Mar-1 (1), Mar Ther-1 (2), Wal Ther-2 (1) et Opo Ther-2 (2) (Figures 4), et enfin trois autres séquences marsupiales qui n'ont pas pu être spécifiquement comparées à cause d'absence d'identité dans le segment 3' (Figure 4D).

3-1.5.2- Distribution des familles.

La représentation des génomes des mammifères non-placentaires est très faible dans les banques de données (voir Tableau E, paragraphe 2-2.18.1-). Nous avons donc utilisé des oligonucléotides de 40 bases (spécifique de chaque famille) (Figure 5) pour



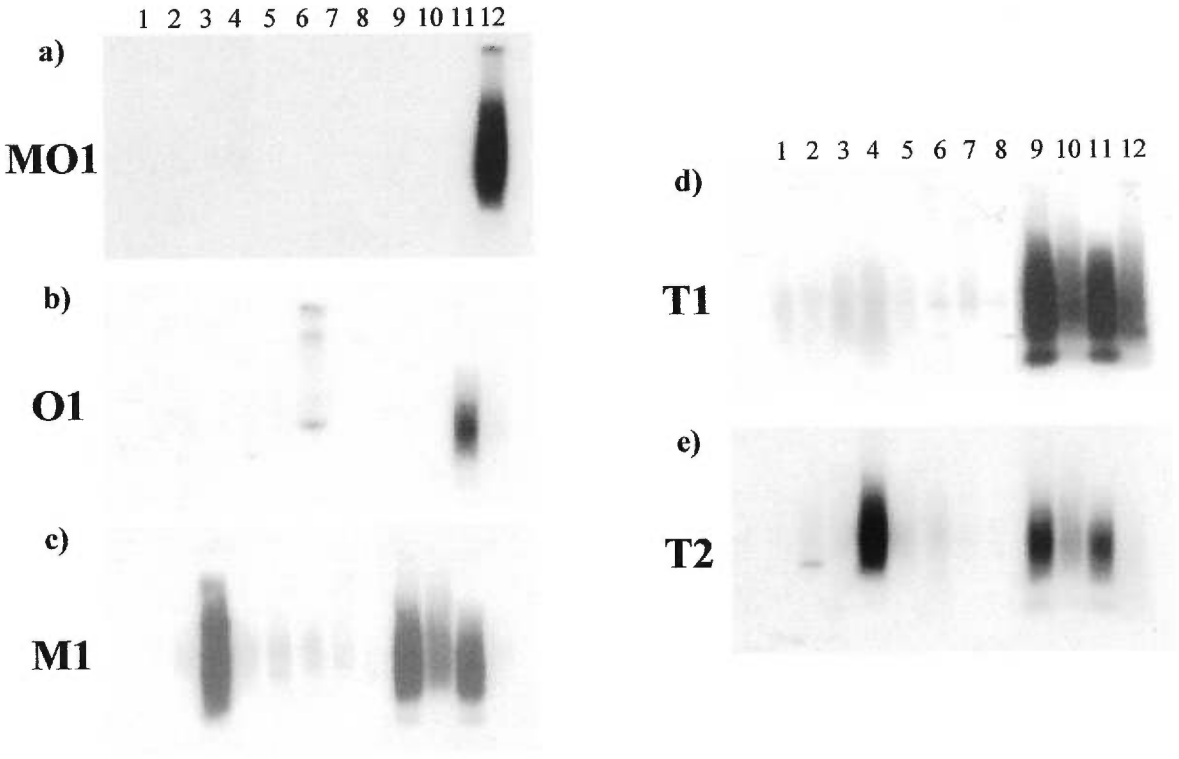
A cause de leur trop forte divergence, les segments 5' dérivé d'ARNt ont été exclus de l'alignement. Les séquences 18I et 30I proviennent des clones U59962, U61375, U63313 et U63312 ayant servi à établir l'alignement de la sous-famille Hum Ther-1.

FIGURE 7 : Alignement des séquences de la famille Ther-2 (sous-famille Hum Ther-2).

effectuer des Southern-blot et ainsi vérifier la distribution des familles CORE-SINE dans les génomes mammifères. Les expériences permettent de vérifier ainsi que la famille Mon-1 est spécifique des monotrèmes (Figure 8a) et Opo-1 spécifique du génome opossum (Figure 8b). Le signal observé avec la sonde OP1 sur l'ADN de pigeon (Figure 8b) pourrait être du à une identité entre la sonde et une séquence répétée en tandem, une exposition plus longue révèle un signal en échelle, spécifique de ce type de répétition. La sonde M1 s'hybride avec tous les génomes marsupiaux (Figure 8c). Avec les sondes T1 et T2, un signal est observé pour tous les génomes marsupiaux, mais aucun signal n'est détecté pour les génomes placentaires (Figure 8d, 8e), malgré leur présence démontrée par la recherche dans les banques de données. Les éléments CORE-SINE des génomes placentaires étant trop divergents par rapport à la sonde, la détection par hybridation manque de sensibilité. Une exposition plus longue révèle un faible signal pour tous les génomes utilisés dans l'expérience. La sonde T1 révèle la présence des membres de la famille Ther-1 dans les génomes monotrèmes (Figure 8d). La famille Ther-1 est donc présente dans tous les génomes mammifères.

Deux signaux d'hybridation très forts sont présents pour le génome du bovin avec la sonde M1 (Figure 8c) et pour le génome du serpent avec la sonde T2 (Figure 8e). Nous avons vérifié si ces signaux sont spécifiques des CORE-SINE correspondants dans les génomes ou s'ils proviennent d'une identité à d'autres séquences répétées. Les résultats, présentés dans le paragraphe 3-3.2-, nous indiquent que les signaux sont dus à une identité des sondes avec d'autres rétroposons.

FIGURE 8 : Southern-blot des segments spécifiques des familles CORE-SINE mammifères.



Les numéros de 1 à 12 correspondent aux espèces testées : 1 humain, 2 souris, 3 bovin, 4 serpent, 5 oie, 6 pigeon, 7 saumon, 8 levure, 9 wallaby, 10 souris marsupiale, 11 opossum, 12 ornithorynque. Le nom de la sonde utilisée pour chaque hybridation de la même membrane est indiqué à gauche : sonde spécifique de la famille **a)** Mon-1 (MO1), **b)** Opo-1 (OP1), **c)** Mar-1 (M1), **d)** Ther-1 (T1) et **e)** Ther-2 (T2).

3-1.6- Divergence des rétroposons CORE-SINE.

Nous avons observé que les éléments d'un même génome sont généralement plus similaires les uns par rapport aux autres qu'entre les différents génomes. De plus la divergence au sein de chaque famille est différente suivant le segment considéré. La partie centrale de la séquence, le core, est mieux conservée que chacun des segments flanquants. La divergence varie entre 2% et 27% pour le core tandis qu'elle est plus élevée pour les deux autres segments, variant de 7% à 40% (Tableau 2A). La présence de dinucléotides CpG dans les segments adjacents et leur absence dans le core explique en partie cet écart. En effet, il a été déterminé que ces dinucléotides ont un taux de mutation plus élevé que les autres nucléotides (Bird, 1980). Ceci est facilement observable sur les alignements de la figure 4. Cependant, l'élimination de ces positions dans le calcul de divergence ne permet pas de rétablir un équilibre entre les trois segments (Tableau 2B). Le contenu C+G n'intervient pas dans la différence observée puisque celui-ci est constant le long de la séquence et représente environ 50% des nucléotides. Le manque d'identification des positions diagnostiques est une deuxième raison pour laquelle la divergence avec le consensus est accrue. Dans les alignements présentés nous n'avons pu distinguer que deux ou trois sous-familles pour chaque famille CORE-SINE. Il est donc fort probable qu'il existe des positions diagnostiques non identifiées influençant la divergence. Elles doivent être prépondérantes dans les segments flanquants le core. Néanmoins, le core apparaît mieux conservé et la raison de cette pression de sélection reste encore inconnue.

Pour compléter l'étude de divergence des éléments de chaque famille CORE-SINE nous avons créé au laboratoire un logiciel d'analyse. Ce programme détermine la divergence de chaque élément par rapport à son consensus et permet d'observer si l'ensemble des mutations d'un groupe de séquences suit une distribution aléatoire. Le

TABLEAU 2 : Divergence des segments CORE-SINE par rapport à leur consensus.

A

Famille	Divergence		
	Région dérivée d'ARNt	"core"	Segment 3' spécifique
Hum Ther-1	40,4%	26,7%	38,1%
Mar Ther-1	26,2%	19,8%	23,8%
Wall Ther-2	34,5%	20,3%	31,1%
Opo Ther-2	28,6%	26,5%	25,4%
Mar-1a	21,8%	11,5%	17,6%
Mar-1b	20,0%	11,4%	17,3%
Opo-1	21,3%	12,7%	12,5%
Mon-1a	19,7%	8,8%	17,5%
Mon-1b	11,5%	6,4%	12,9%
Mon-1c	10,9%	2,2%	6,7%

B

Famille	Divergence sans les CpG		(nombre de CpG)
	Région dérivée d'ARNt	"core"	Segment 3' spécifique
Hum Ther-1	36,6% (5)	26,7%	36,0% (3)
Mar Ther-1	21,3% (3)	19,8%	22,9% (2)
Wal Ther-2	30,6% (3)	20,3%	31,1%
Opo Ther-2	26,8% (2)	26,5%	25,4%
Mar-1a	18,7% (3)	11,5%	14,6% (4)
Mar-1b	17,8% (3)	11,4%	14,6% (3)
Opo-1	19,7% (2)	12,7%	12,5%
Mon-1a	15,8% (4)	8,8%	15,7% (3)
Mon-1b	8.8% (5)	6,4%	10,5% (3)
Mon-1c	8,9% (4)	2,2%	5,0% (3)

A : La divergence est calculée suivant la formule suivante : $\left(\frac{S+G}{N}\right)$, où S représente les substitutions des nucléotides, G les insertions/délétions (compté comme un événement quelle que soit la taille) et N le nombre total de nucléotides dans l'alignement des séquences.

B : Dans ce tableau, les substitutions C vers T et G vers A aux positions CpG ne sont pas incluses dans le calcul.

résultat pour chaque alignement des sous-familles, en considérant la totalité de chaque séquence, permet de conclure à l'absence de distribution aléatoire des mutations. Nous pouvions nous attendre à ce résultat puisque la divergence des séquences n'est pas uniforme pour chacun des trois segments constitutifs des éléments et que la définition des sous-familles est incomplète.

Nous avons donc, dans un deuxième temps, analysé uniquement le segment central de la famille la mieux conservée, Mon-1, dans laquelle la divergence moyenne d'une séquence par rapport à son consensus est de 15%. De plus, nous avons considéré uniquement les éléments possédant le domaine core au complet pour éviter le biais créé par l'absence de données. Encore une fois nous avons observé que la distribution des substitutions des nucléotides n'est pas aléatoire lorsque toutes les sous-familles sont regroupées dans l'analyse (Figure 9a). Nous avons continué l'étude en considérant les sous-familles séparément. Les mutations de la sous-famille Mon-1c correspondent à plus de 85% à une distribution aléatoire suivant la loi de Poisson (Figure 9b). Par contre, la sous-famille Mon-1a (Figure 9c) suit une distribution similaire à celle de la famille au complet (Figure 9a). Nous n'avons pas intégré la sous-famille Mon-1b dans l'analyse car seulement deux séquences informatives sont disponibles. Pour Mon-1c, qui possède les éléments les moins divergents, la distribution des mutations est aléatoire.

Pour toutes les sous-familles des autres CORE-SINE, aucune distribution ne se rapproche des statistiques théoriques. De nombreuses raisons, tels que l'hétérogénéité des taux de mutations ou la conversion génique, peuvent expliquer ces résultats. Cependant, la principale semble être la caractérisation incomplète des sous-familles des CORE-SINE. Seule la sous-famille Mon-1c est bien identifiée. De plus, les éléments constitutifs des sous-

FIGURE 9 : Test de distribution aléatoire des mutations du segment core de la famille Mon-1.

a) Mon-1

mutations	Nb. de positions	Distribution binomial (7,5 %)
0	16	9
1	19	18
2	15	18
3	6	12
4	2	5
5	4	2
>6	3	1

Test Chi2	16,1
P	<0,5%

b) Mon-1c

mutations	Nb. de positions	Distribution de Poisson (2,15 %)
0	58	58,4
1	7	6,3
2	0	0,3

Test Chi2	0,025
P	>85%

c) Mon-1a

mutations	Nb. de positions	Distribution binomial (8,1 %)
0	24	13
1	15	22
2	12	17
3	7	9
>4	7	4

Test Chi2	15,7
P	<0,5%

La seconde colonne donne la somme des positions nucléotidiques qui ont un nombre donné de mutation dans l'alignement. La troisième donne le nombre correspondant attendu suivant une distribution aléatoire binomiale (pour la sous-famille Mon-1c nous avons utilisé la distribution de Poisson étant donné que la divergence est inférieure à 5%). Le test de Chi2 est effectué pour déterminer la ressemblance des positions observées par rapport à celles attendues. La valeur P indique le pourcentage de similitude entre les deux distributions. Analyse de ; **a)** 26 segments core pleine longueur de la famille Mon-1, **b)** 5 segments core de Mon-1c et **c)** 19 segments core de Mon-1a.

familles étant très anciens et très divergents de leur consensus, il sera difficile de les recréer.

3-1.7- Age des familles CORE-SINE.

Ayant déterminé la divergence moyenne des éléments CORE-SINE par rapport à leur consensus (Figure 10a), nous avons pu faire une estimation de l'âge de chaque famille. L'âge moyen des familles est difficile à estimer du fait du taux de mutation inconstant le long de la séquence des éléments. À partir des données moléculaires et morphologiques de fossiles, l'horloge moléculaire des mammifères est estimée autour de 0,4 mutations par site et par MA (Springer, 1995). Les analyses préalables faites au laboratoire sur le segment core uniquement donnait une horloge moléculaire plus lente, aux environs de 0,2 mutations par site et par MA (Jurka *et al.*, 1995). En utilisant les deux valeurs nous établissons une large fourchette de temps pour déterminer l'âge des familles en considérant la longueur totale des éléments. Les familles les plus vieilles sont Ther-1 et Ther-2 (entre 60 et 180 MA) et cela est en accord avec leur distribution dans les différents génomes mammifères (Figure 10b). Ensuite apparaissent les familles spécifiques aux marsupiaux (entre de 40 et 80 MA). La famille la plus jeune est Mon-1 des monotrèmes (entre 15 et 80 MA).

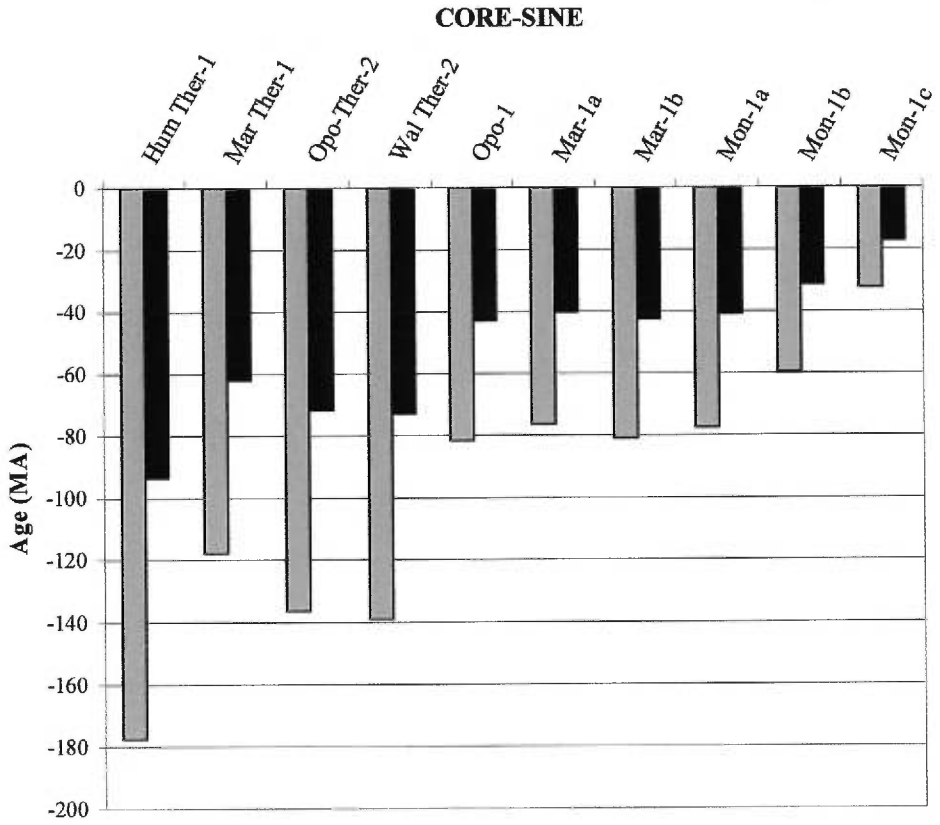
L'estimation de l'âge des sous-familles de Mon-1 nous permet de déterminer l'ordre séquentiel de leur apparition ; Mon-1c dérive de Mon-1b qui lui-même descend de Mon-1a. Cependant, il faut noter que Mon-1a ne constitue pas en temps que tel une sous-famille, mais regroupe les éléments dont les sous-familles ne sont pas encore bien identifiées.

FIGURE 10 : Divergence et âge moyen des sous-familles CORE-SINE mammifères.

a)

CORE-SINE	Hum Ther-1	Mar Ther-1	Opo Ther-2	Wal Ther-2	Opo-1	Mar-1a	Mar-1b	Mon-1a	Mon-1b	Mon-1c
Divergence (%)	35,5	23,6	27,3	27,8	16,3	15,3	16,2	15,5	11,9	6,5

b)



a) Divergence moyenne des éléments CORE-SINE par rapport à leur consensus.

b) Les estimations ont été obtenues en considérant pour chaque famille la totalité des séquences des éléments identifiés. Les colonnes en gris représentent l'âge calculé avec le taux de mutation de 0,2 par site et par MA (Jurka *et al.*, 1995), et les colonnes noires l'âge avec le taux de mutation de 0,4 par site et par MA (Springer, 1995).

3-2- Les Familles CORE-SINE chez les non-mammifères.

La présence d'éléments Ther-1 chez tous les mammifères et les réactions de PCR positives avec les amorces spécifiques du domaine core pour certains génomes d'oiseaux nous ont conduits à vérifier la présence d'éléments CORE-SINE chez les espèces non-mammifères.

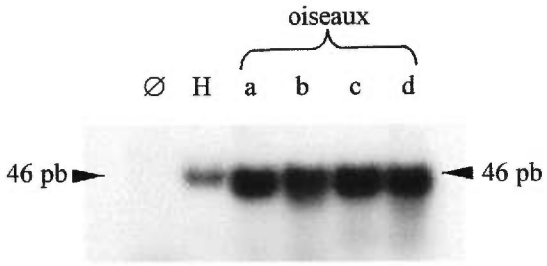
3-2.1- CORE-SINE chez les oiseaux et les reptiles.

Nous avons à nouveau effectué des PCR en utilisant quatre génomes d'oiseaux différents et l'amplification d'un fragment de la taille attendue a été obtenu (Figure 11). Les oiseaux possèdent dans leurs génomes des séquences avec le domaine core. Ces séquences ne sont pas détectables par hybridation. Nous avons en effet essayé de détecter la présence d'élément core par dot-blot, mais sans résultat.

Pour identifier des éléments à partir des données de GenBank, nous avons créé notre propre banque composée uniquement des séquences provenant de génomes d'oiseaux (voir Tableau E, paragraphe 2-2.18.1-). En utilisant le programme FASTA nous avons pu détecter la présence de 16 éléments Ther-1 tronqués (Figure 12) dont les identités varient entre 56 et 77 % avec le segment correspondant du consensus Hum Ther-1 (Tableau 3). La banque de séquences d'oiseaux que nous avons créée représentant au maximum 0.47 % du génome total, par extrapolation nous pouvons déterminer qu'il existe près de 3500 copies de la répétition Ther-1 détectables chez les oiseaux.

Le modèle le plus simple pour expliquer la présence d'un même rétroposon dans deux génomes distants est la transmission verticale, c'est à dire que le génome de l'ancêtre commun aux deux espèces doit contenir le rétroposon. Les reptiles, étant à l'origine des

FIGURE 11 : Amplification du segment core par PCR chez les oiseaux.



La taille du fragment est indiquée de chaque côté du gel. ∅ est le témoin de PCR sans ADN et H l'amplification avec l'ADN humain. Les génomes d'oiseaux utilisés sont les suivants : a- mainate indien, b- moineau, c- «moorhen» (échassier) et d- pigeon. Pour le puits H, 2 µl de la réaction de PCR (20 µl total) ont été déposés, pour les autres nous avons déposé 10 µl sur le gel.

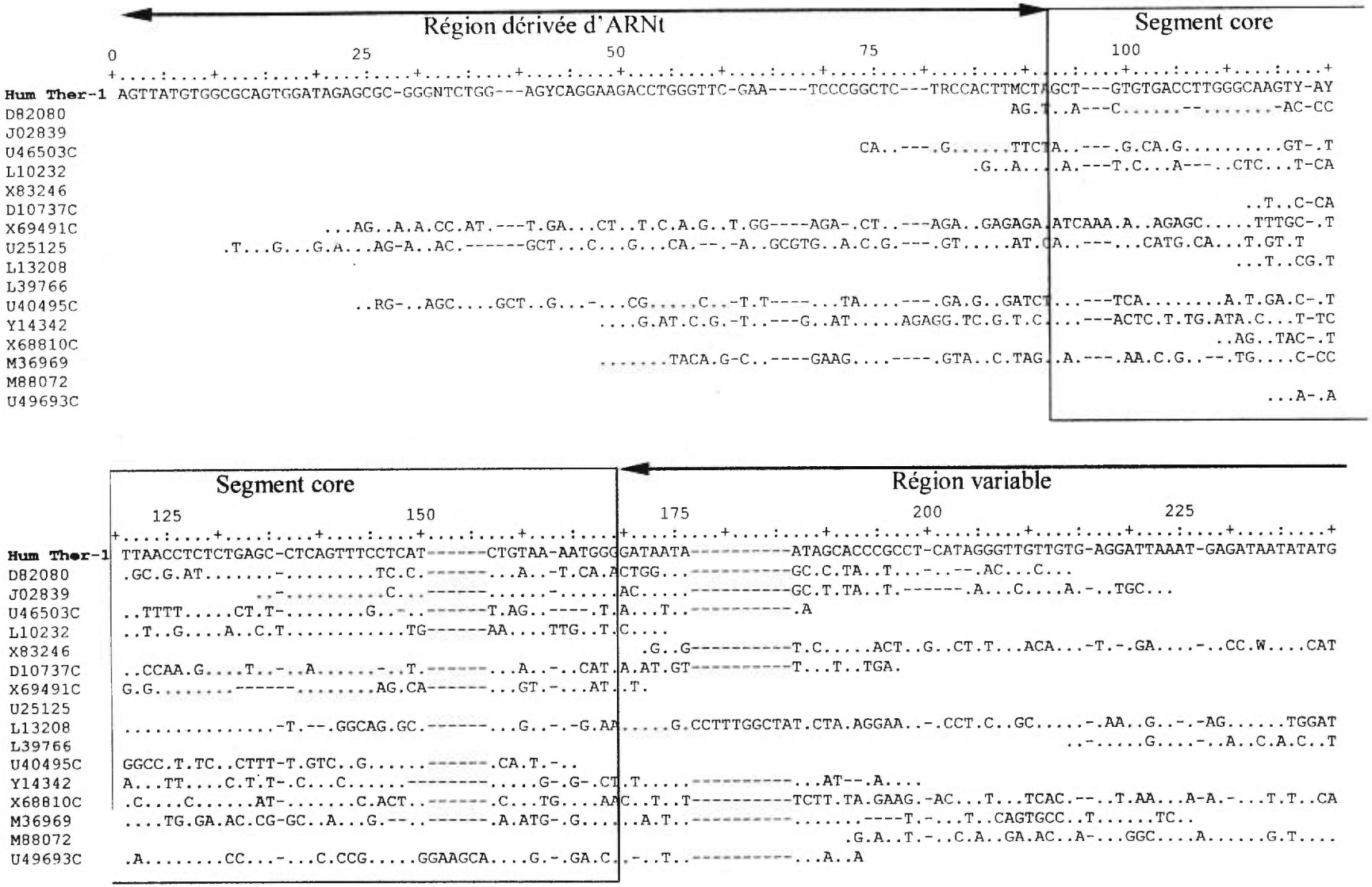
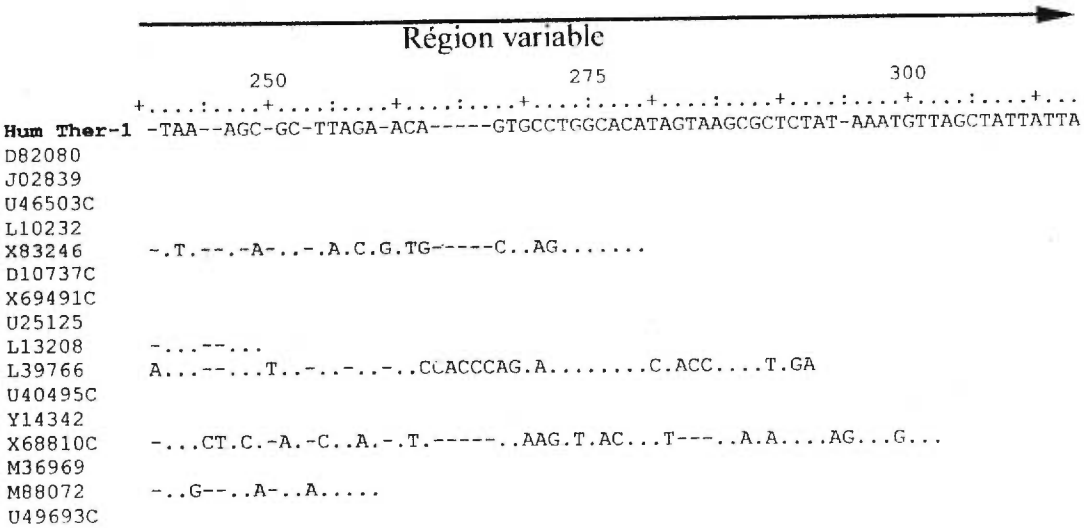


FIGURE 12 : Alignement des séquences CORE-SINE des génomes d'oiseaux (1/2).



La séquence de référence est le consensus de la sous-famille Hum Ther-1.

FIGURE 12 : Alignement des séquences CORE-SINE des génomes d'oiseaux (2/2).

TABLEAU 3 : CORE-SINE dans les génomes non-mammifères.

	Genomes	Sequences	taille	Similarité avec Ther-1 [a] (segments)	Identité [a]
Oiseaux	*Gallus gallus*	J02839	71	118 – 188 (core + région variable)	77.5%
		D82080	104	77 – 179 (core + région variable)	66.3%
		U46503	89	65 – 153 (région tRNA + core)	67.4%
		L10232	80	73-150 (région tRNA + core)	68.7%
		X83246	86	148-231 (région variable)	60.4%
		D10737	67	97-163 (core + région variable)	61.2%
		X69491	132	21-148 (région tRNA + core)	56.1%
		U25125	92	11-98 (région tRNA + core)	63.0%
		L13208	125	96-209 (core + région variable)	64.8%
		L39766	72	178-244 (région variable)	72.2%
		Y14342	126	44-165 (tout)	62.7%
		M88072	61	158-216 (région variable)	68.9%
		U49693	66	99-158 (core + région variable)	71.2%
	Eopsaltria australis	U40495	120	24-140 (région tRNA + core)	61.7%
	Anas platyrhynchos	X68810	164	94-253 (core + région variable)	59.8%
	Columba livia	M36969	149	44-190 (tout)	60.4%
Reptiles	*Anolis carolinensis*	L31503	115	8 – 124 (région tRNA + core)	72.3%
Poissons	*Salmonidés*	HpaI (segment central)	65	81 – 145 (core)	64.6%
		AvaIII (segment central)	44	81 – 124 (core)	68.2%
	Cichlidés	AFC (segment central)	42	81 – 122 (core)	73.8%
	Fundulus heteroclitus	U59855	200	8 – 207 (tout)	61.5%
Invertébrés	*Céphalopodes*	OR2 (segment central)	65	81 – 145 (core)	63.1%
		OR1 (segment central)	44	81 – 124 (core)	50.0%

[a] : Le consensus Ther-1 sert de référence pour la numérotation et le calcul d'identité.

oiseaux et des mammifères, devraient posséder cet élément dans leur génome. Nous avons donc recherché des séquences CORE-SINE par FASTA dans les banques de données. Nous n'avons trouvé qu'une seule séquence chez les reptiles présentant une identité de 72 % avec la région dérivée de l'ARNt et le domaine core (segment conservé des familles CORE-SINE) (Tableau 3). Etant donné la faible représentation du génome des reptiles dans GenBank (0,3‰), la séquence identifiée appartient probablement à une famille CORE-SINE. Ce postulat est soutenu par le fait que la proportion d'éléments détectés par rapport à la représentation du génome dans les banques est la même que pour les oiseaux.

3-2.2- CORE-SINE chez les poissons.

Nous avons continué nos recherches de séquences CORE-SINE chez tous les vertébrés en comparant, à l'aide du programme BLAST, le domaine conservé de la famille Ther-1, c'est-à-dire les nucléotides 1 à 145 du consensus (Figure 5). La section de GenBank comprenant les séquences d'organismes vertébrés non-mammifères a été criblée (voir Tableau F, paragraphe 2-2.18.2-). Une vingtaine de séquences reliées aux éléments CORE-SINE des mammifères a été identifiée. Toutes ces séquences ont la particularité d'appartenir à deux familles distinctes de rétroposons ; la famille *Hpa*I des salmonidés (Kido *et al.*, 1991) et la famille AFC des cichlidés (Takahashi *et al.*, 1998). Nous avons donc comparé les consensus de ces familles de SINE avec ceux des CORE-SINE mammifères (Figure 13a). Les domaines centraux des SINE *Hpa*I et AFC présentent respectivement 64,6 et 73,8 % d'identité avec le consensus du domaine core de Ther-1 (Tableau 3). Nous en avons déduit que ces rétroposons peuvent être associés à la super-famille des CORE-SINE. De plus ils possèdent en amont du domaine central un segment dérivé d'ARNt qui sert de promoteur.

FIGURE 13 : Identification des familles CORE-SINE non-mammifères.

a) Identité de séquence entre le segment core de la famille Ther-1 et les segments centraux des SINE *Hpa*I (Kido *et al.*, 1991) et *Ava*III (Kido et *al.*, 1994) des Salmonidés, AFC des Cichlidés (Takahashi *et al.*, 1998) et OR2 des Céphalopodes (Ohshima et Okada, 1994). La séquence du rétroposon OR1 des Céphalopodes (Ohshima et Okada, 1994) a été ajoutée à l'alignement en raison de son identité avec le SINE OR2. La numérotation correspond à celle du consensus de la famille Ther-1.

b) Comparaison des consensus deux à deux par dot-matrix. Le nom des SINE ou de la séquence U59855, représentant l'élément génomique du poisson *Fundulus heteroclitus*, est indiqué au-dessus ou sur la gauche des carrés. Les lignes qui traversent les carrés délimitent le segment core. Dans la comparaison des consensus OR1 : OR2, les 120 derniers nucléotides de OR1 ont été enlevés pour plus de clarté.

a)

```
          81          90          100         110         120         130         140  145
          |           |           |           |           |           |           |    |
core      GCTGTGT------GACC-TTGGGCAA-GTYAY-TTAA--CCTCTCTGAGCCTCAGTTTCCTCATCTG--TAAAATGGG
HpaI      AA.C...CGTTCT.C..-C..AA...-.GC.G-....--..-.A...TT...AG.--C.G....-..---.....AA.
AvaIII    AA.C...CGTTCT.C..-C..A....-.GC.G-....--..-.A...TT..C.G.
AFC       T.GT...------.T..-........-.AC.C-..C.--..-.GT..--...A
OR2       ..GT...------.TT.-...A....-AAC.C-..C.--TT..A.GTT..TC....CCA....G...GCA......A.
OR1       .-..A..------.TTTA...A..G.AAAC.CC....AG.TC.A.GAG..TC.G.
```

b)

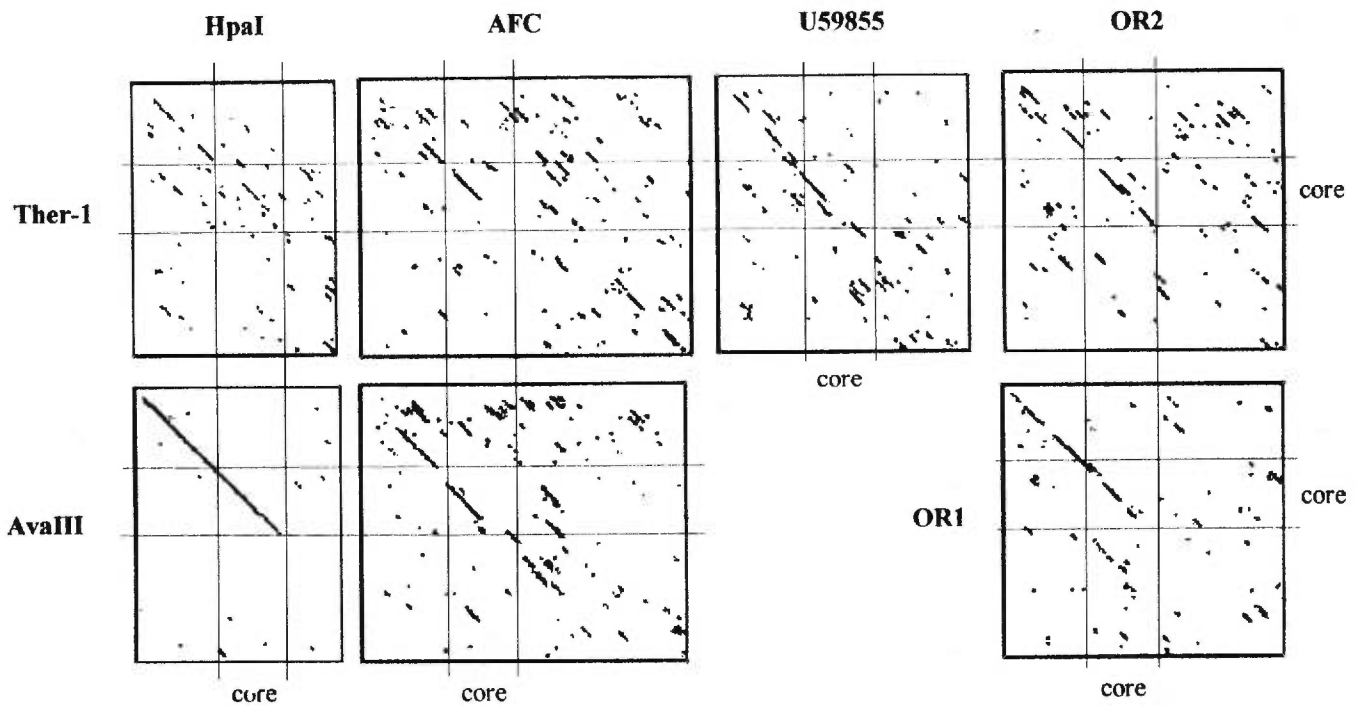


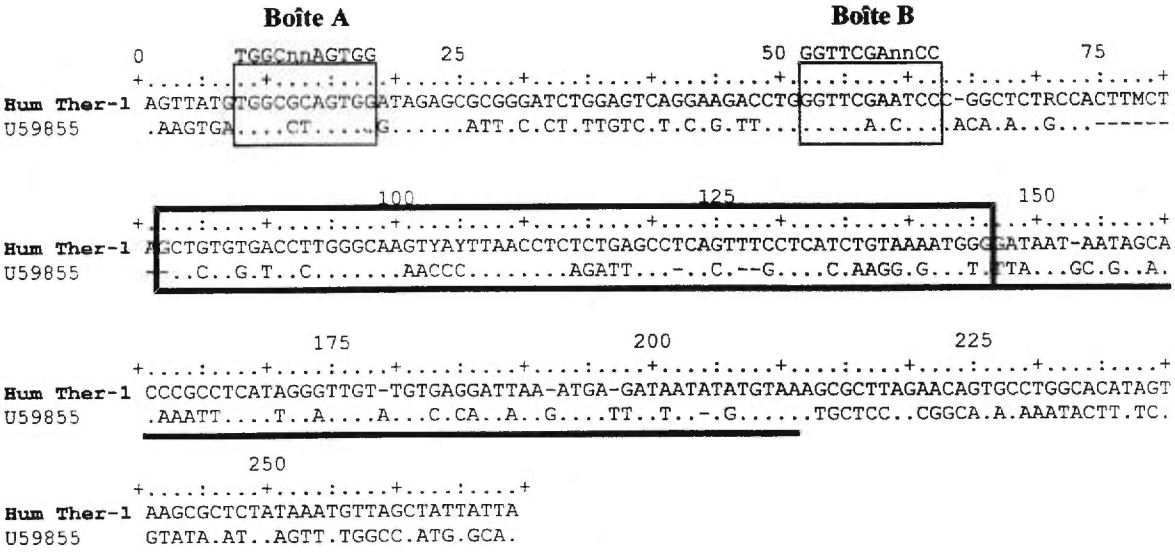
FIGURE 13.

La famille SINE *Ava*III des *Salmonidés* (Kido *et al.*, 1994) peut aussi être associée à la super-famille CORE-SINE du fait de son identité avec le SINE HpaI (Figure 13a). En effet, ces deux familles de rétroposons partagent une forte identité sur les 130 premiers nucléotides des séquences, comprenant le segment dérivé d'ARNt et 44 nucléotides du core (Kido *et al.*, 1994). Ces 44 nucléotides possèdent une identité de plus de 68 % avec les nucléotides correspondant de la famille Ther-1 (Tableau 3).

Par des recherches en utilisant le programme FASTA nous avons aussi trouvé un élément provenant du génome de *Fundulus heteroclitus* (poisson téléostéen) pouvant être associé aux CORE-SINE (Tableau 3). Cet élément, situé dans la région 5' flanquante du gène Ldh-B (Lactate dehydrogénase-B) (Schulte *et al.*, 1997), n'a pas encore été associé à une famille d'élément SINE. Cependant l'analyse de la séquence révèle une région 5' possédant les boîtes A et B du promoteur de l'ARN Pol III (pouvant dériver d'un ARNt) (Figure 14). Cette région est immédiatement suivie de la région d'identité avec le core. De plus la séquence présente une identité proche de 62 % avec le segment variable spécifique de la famille Ther-1, immédiatement après le domaine core (Figure 14).

En complément des alignements des consensus présentés sur la figure 13a, nous avons produit des comparaisons de ces mêmes séquences deux à deux par dot-matrix. Le résultat de cette approche indique que l'identité du domaine core entre les séquences n'est pas inférieure à l'identité observée dans les régions dérivées d'ARNt (Figure 13b). Cela suggère que la conservation de ce domaine central est similaire à celle du promoteur bipartite de l'ARN Pol III.

FIGURE 14 : Identité de la séquence du génome de *F. heteroclitus* avec le consensus Ther-1.



Les boîtes A et B du promoteur possible de la séquence de poisson sont encadrées et le consensus général de celles-ci est indiqué au dessus des cadres. Le cadre plus épais délimite le domaine core et la région spécifique de Ther-1 est soulignée.

3-2.3- CORE-SINE chez les Invertébrés.

Ayant identifié avec succès des éléments CORE-SINE dans chaque groupe de vertébrés (mammifères, oiseaux, reptiles et poissons), nous avons effectué la même recherche pour les génomes des espèces invertébrées. Aucune séquence n'a été obtenue par criblage des banques de données. Nous avons alors analysé les rétroposons des génomes invertébrés déjà publiés. La comparaison par alignement du rétroposon SINE OR2 dérivé d'ARNt (Ohshima et Okada, 1994), du génome du poulpe (céphalopode), avec le segment conservé de Ther-1 a révélé une identité de 63 % entre les segments centraux des deux consensus (Figure 13a et Tableau 3). Nous avons également inclus dans l'analyse le SINE OR1, présent aussi dans le génome du poulpe, du fait de sa forte identité avec OR2 dans les régions 5' et centrale du consensus (Ohshima et Okada, 1994) (Figure 13a et Tableau 3). De la même façon que pour les séquences provenant des génomes de poissons, nous avons illustré les identités par analyse dot-matrix (Figure 13b).

3-2.4- Relation phylogénétique des familles CORE-SINE.

Nous avons démontré la présence d'éléments rétroposons appartenant à la super-famille des CORE-SINE dans de nombreux génomes eucaryotes. Une analyse phylogénétique des domaines conservés, c'est-à-dire la région dérivée d'ARNt et le segment core (Figure 15), a été entreprise pour déterminer les relations possibles entre ces séquences. La figure 16 représente l'arbre phylogénétique, enraciné ou non, des séquences CORE-SINE. Nous observons que les trois groupes, céphalopodes, poissons et mammifères sont bien distincts et que leur origine commune est assez distante. Chez les mammifères, les séquences de chaque ordre ou sous-ordre sont aussi séparées, et le consensus Ther-1 semble à l'origine des familles CORE-SINE mammifères. Une seconde analyse phylogénétique a

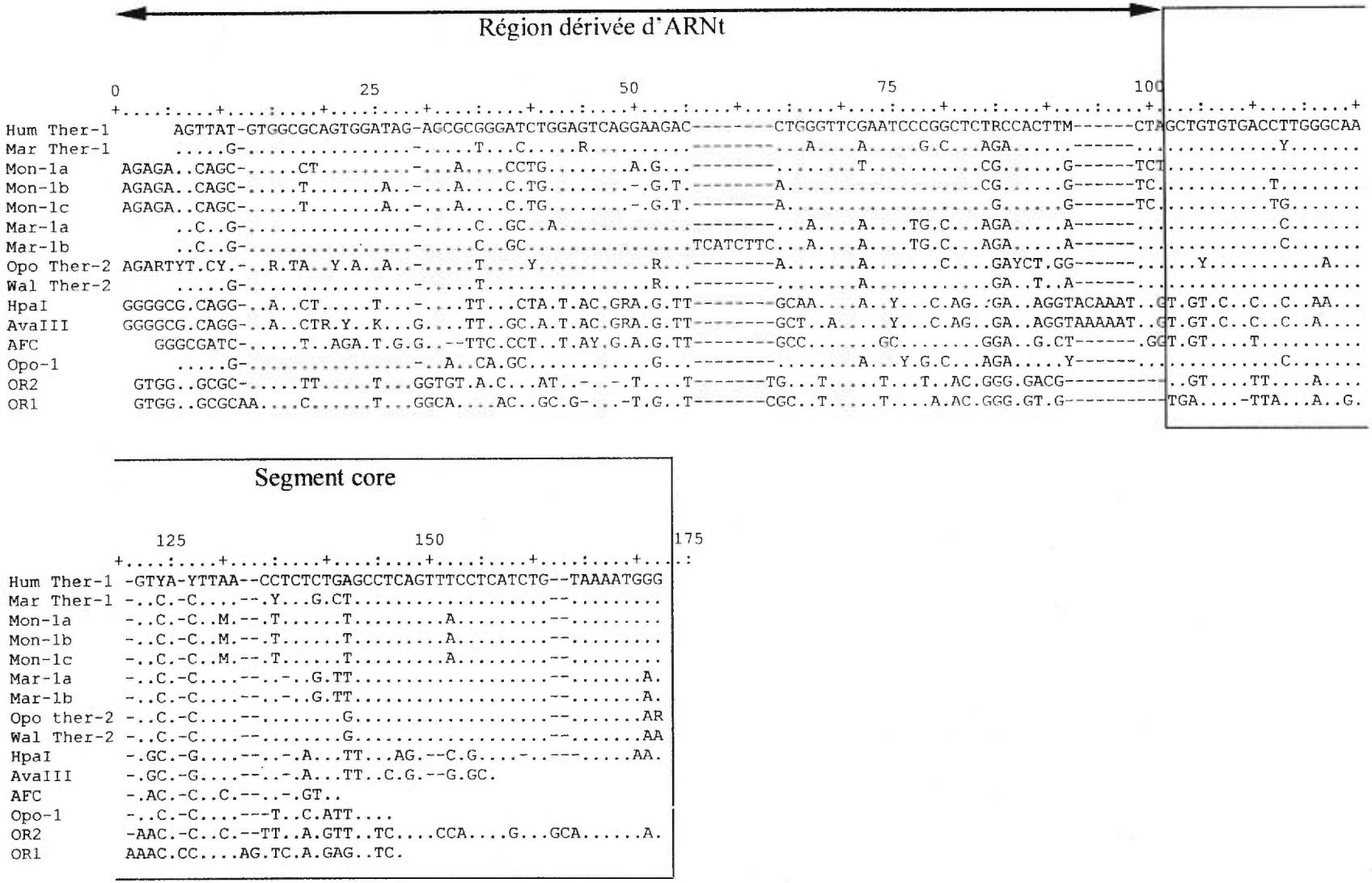
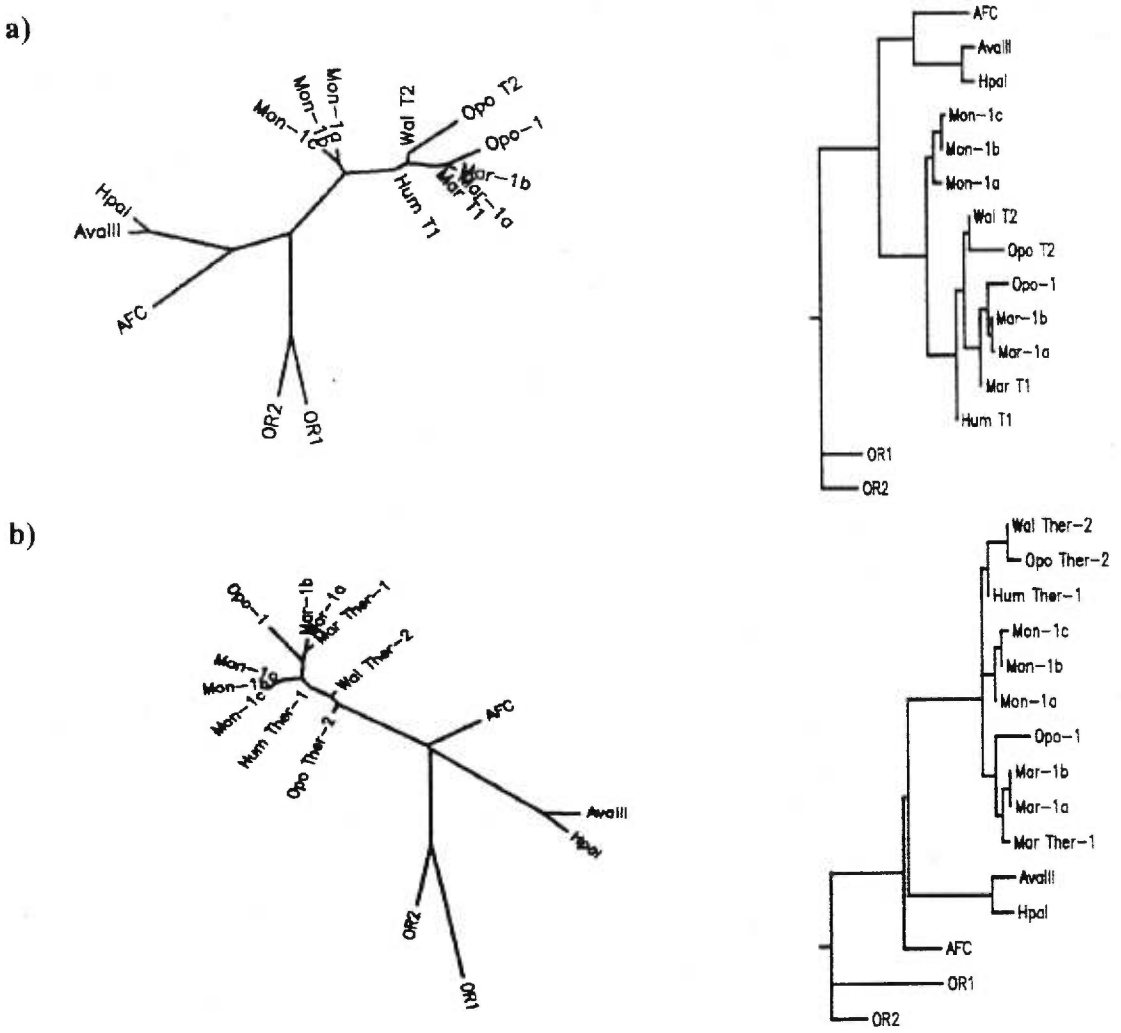


FIGURE 15 : Alignement de la région conservée des familles CORE-SINE.

FIGURE 16 : Arbres phylogénétiques des consensus CORE-SINE.



Arbres phylogénétiques établis par analyse de vraisemblance maximale.

a) Analyse effectuée en considérant la région dérivée d'ARNt et le segment core (Figure 15) ou **b)** avec le segment core uniquement.

La séquence qui sert de racine aux arbres situés à droite de la figure est OR2. La longueur des branches est proportionnelle à la divergence des séquences. Le nom de chaque famille CORE-SINE est inscrit aux extrémités des branches.

été effectuée, mais cette fois sans prendre en considération la région dérivée d'ARNt (Figure 16b). Nous obtenons sensiblement les mêmes résultats. Les trois groupes eucaryotes sont séparés et chez les mammifères les trois lignées le sont aussi. La seule différence est que la famille Ther-2 semble être à l'origine des CORE-SINE mammifères dans la seconde analyse.

3-3- Origine et rôle des segments des CORE-SINE.

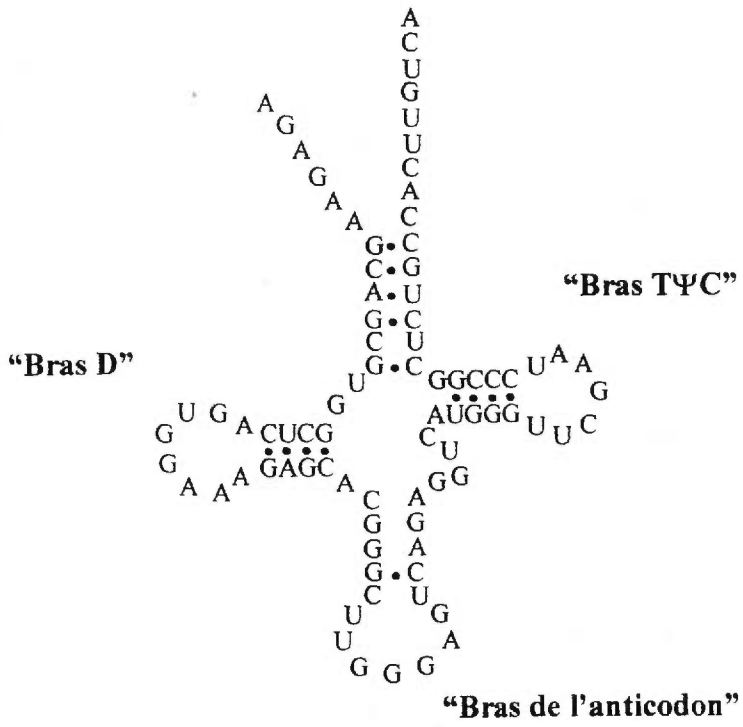
3-3.1- Le segment dérivé d'ARNt.

Pour un certain nombre de rétroposons SINE, il a été possible d'identifier l'origine ARNt du segment 5' (voir introduction). En revanche, les cinq familles CORE-SINE identifiées dans ces travaux n'ont pu être associées à un ARNt de façon spécifique. Comme l'illustre la figure 17, la structure secondaire en forme de trèfle, typique des ARNt, n'est pas reconstituée entièrement. Même si les bras D et TΨC de la structure sont facilement reconstruits, du fait de la conservation des boîtes A et B, le segment qui sépare les deux boîtes ne permet pas de former le bras de la boucle de l'anticodon à cause de l'absence de liaison entre les nucléotides. La structure secondaire du segment 5' des familles CORE-SINE ne semble donc pas conservée. Son rôle présumé serait uniquement d'assurer la transcription.

3-3.2- Le segment 3' spécifique.

Le segment 3' de la famille Ther-1 a été déterminé comme provenant de l'élément LINE des mammifères L2 (Smit et Riggs, 1995; Smit, 1996). La famille Mon-1 partageant

FIGURE 17 : Représentation en feuille de trèfle de la région dérivée d'ARNt de la famille Mon-1.



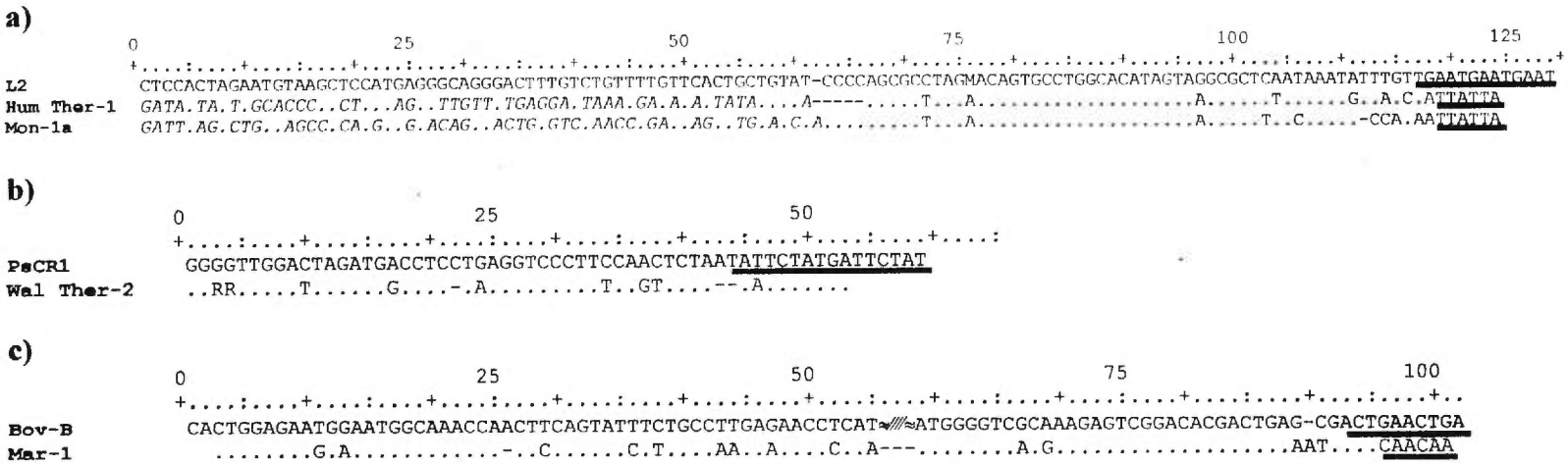
Les noms correspondants aux structures en trèfle des ARNt sont indiqués. Les points indiquent les liaisons possibles.

les mêmes 58 derniers nucléotides nous en avons conclu que ce segment dérive aussi du rétroposon L2 (Figure 18a, Tableau 4).

Par Southern-blot, avec les sondes T2 et M1 (Figure 8c et 8e) nous avons obtenu un signal d'hybridation avec les génomes de reptile et de bovin respectivement. Pour vérifier l'origine de ces signaux nous avons effectué une recherche par FASTA sur les banques d'ADN propre à chacun des génomes créées dans l'environnement GCG. Nous avons déterminé que l'hybridation n'est pas spécifique d'éléments CORE-SINE de la famille correspondant à la sonde, mais à l'identité des sondes avec des éléments LINE.

La sonde T2 présente une forte identité avec la séquence répétée LINE CR1. Ce rétroposon a été découvert à l'origine chez le poulet (Silva et Burch, 1989), mais est aussi présent chez les reptiles, les tortues et les amphibiens (Burch *et al.*, 1993; Vandergon et Reitman, 1994; Ohshima *et al.*, 1996). Le segment spécifique de la famille Ther-2 présente la plus forte identité (83 %) avec la séquence LINE provenant du génome de la tortue, PsCR1 (Ohshima *et al.*, 1996) (Figure 18b, Tableau 4).

La sonde M1 présente une forte identité avec 3 séquences répétées des génomes de bovins, dont deux SINE, Bov-tA et Bov-A2 (Lenstra *et al.*, 1993), et un LINE, Bov-B (Szemraj *et al.*, 1995). L'hybridation croisée est en fait générée par l'identité entre le segment spécifique de Mar-1 et le monomère A des bovins. Ce monomère prend son origine du LINE Bov-B et est le précurseur des deux SINE Bov-tA et Bov-A2 (voir introduction paragraphe 1-4.2.2.5-, Figure J). Le segment spécifique de Mar-1 est plus court de 28 nucléotides que le monomère Bov-A, dans la partie centrale (Figure 19a). L'apparition du segment pourrait donc provenir d'un autre événement de délétion du rétroposon Bov-B, plus que de la présence du monomère A chez les marsupiaux (Figure



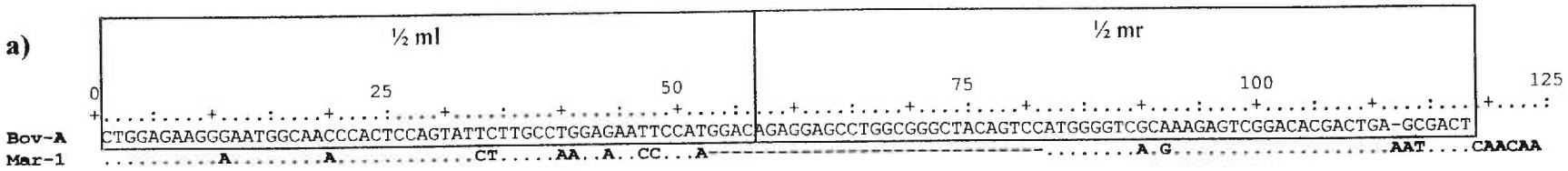
Séquences partagées entre : **a)** Ther-1, Mon-1 et le LINE L2, **b)** Ther-2 et le LINE PsCR1 et **c)** Mar-1 et le LINE Bov-B. Les extrémités 5' et 3' de Bov-B ont toutes deux contribuées à la création de la région variable de Mar-1. Le segment de séquence qui distingue Ther-1 et Mon-1 est en italique. Les répétitions simples des rétroposons sont soulignées.

FIGURE 18 : Comparaison des segments variables des familles CORE-SINE mammifères avec les extrémités 3' de LINE.

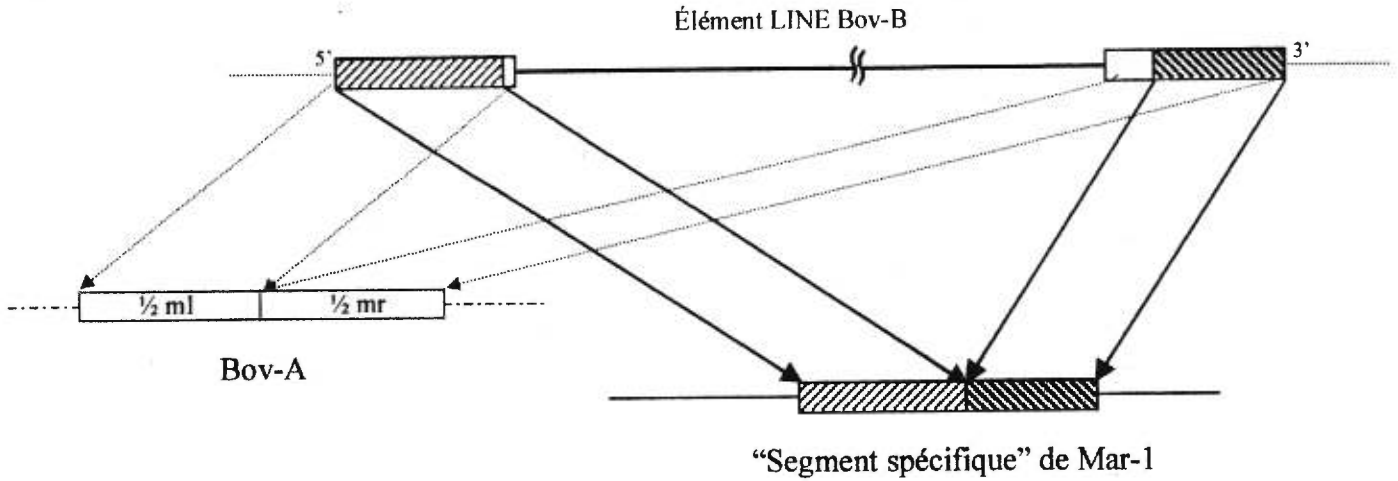
TABLEAU 4 : Identité entre les segments 3' de LINE et de CORE-SINE.

CORE-SINE	LINE	Taille du fragment partagé (pb)	Identité SINE/LINE
Ther-1	L2	52	86.5%
Mon-1	L2	56	82%
Ther-2	PsCR1	53	83%
Mar-1	Bov-B or BDDF	95	80%

a)



b)



a) Comparaison du fragment Bov-A avec le segment variable de Mar-1. Les deux segments du monomères sont encadrés. **b)** Origine possible du fragment variable du CORE-SINE Mar-1, issu d'une large délétion d'un élément Bov-B présent dans les génomes marsupiaux. Cette délétion est indépendante de celle qui a créé le monomère A des génomes bovins (Bov-A) (Voir figure J).

FIGURE 19 : Origine du fragment spécifique du CORE-SINE Mar-1.

19b). Le segment variable de Mar-1 possède 80 % d'identité avec les segments du LINE Bov-B (Figure 18c et Tableau 4).

Nous avons donc identifié pour 4 des 5 familles CORE-SINE mammifères un lien direct avec les rétroposons LINE par leurs extrémités 3'. Ce phénomène est aussi vérifié pour la majorité des familles CORE-SINE des génomes de poissons. En effet, des identités ont été démontrées entre les régions variables des SINE HpaI et AFC avec les extrémités 3' des LINE RSg-1 (Ohshima *et al.*, 1996) et Ci LINE 2 (Terai *et al.*, 1998) (voir introduction Tableau C).

L'efficacité de rétroposition des éléments CORE-SINE pourrait être expliquée par l'identité structurale des régions variables identiques aux éléments LINE actifs. En effet dans le modèle de rétroposition présenté par Luan *et al.* en 1993, la machinerie rétropositionelle vient fixer l'extrémité 3' de l'ARN de l'élément LINE pour initier la transcription inverse. Cette identité observée permettrait aux SINE d'être actifs grâce au complexe rétropositionel fourni en *trans* par les LINE (Ohshima *et al.*, 1996).

3-3.3- Le domaine core.

Le core, tout d'abord considéré comme spécifique des éléments MIR des génomes préplacentaires (Jurka *et al.*, 1995), s'avère être le constituant d'une vaste classe de rétroposons présents dans les génomes eucaryotes. Les recherches de similarité avec d'autres séquences ne nous ont pas permis de déterminer l'origine de ce domaine. Cependant, le fait de le retrouver dans de nombreux SINE et toujours en amont de segments dérivant de séquences LINE nous permet de postuler que le core servirait d'unité d'assemblage de segments déjà présents dans le génome.

3-4- Les LINE associés aux CORE-SINE.

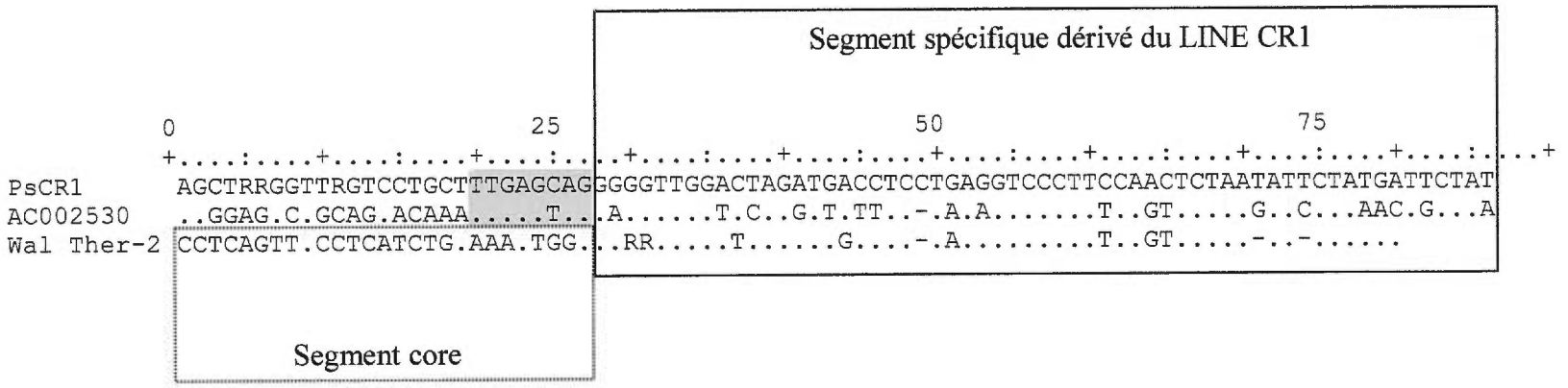
Pour expliquer la présence de segments identiques entre SINE et LINE, il faut que les deux éléments soient présents en même temps dans les mêmes génomes. Nous avons donc vérifié si cela était vrai pour chaque CORE-SINE mammifère.

3-4.1- Le LINE L2.

L'élément L2, nommé à l'origine MIR2 (Smit et Riggs, 1995), est retrouvé chez les mammifères mais aussi dans les génomes de reptiles et d'amphibiens (Smit, 1996). La faible représentation des génomes mammifères non-placentaire dans les banques de données n'a pas permis de vérifier la présence de séquences L2 dans ces espèces.

3-4.2- Le LINE CR1.

Le rétroposon LINE CR1, découvert à l'origine chez le poulet est présent dans tous les génomes d'oiseaux, de reptiles et d'amphibiens. Aucun élément n'a été identifié chez les mammifères à l'heure actuelle. Les éléments de la famille Ther-2, présents chez les mammifères placentaires et marsupiaux, possèdent une forte identité dans leur segment 3' (53 nucléotides) avec l'extrémité 3' du LINE PsCR1 (Figure 18b et Tableau 4). La recherche dans les banques de données de séquences humaines ne permet pas de caractériser de façon certaine un élément appartenant à la famille CR1. Cependant, une séquence du génome humain possède une identité avec le segment spécifique de la famille Ther-2 mais non avec le segment core. Cette séquence peut donc provenir d'un élément CR1. Son alignement avec le consensus de la famille PsCR1 montre une identité restreinte à 8 nucléotides en amont du segment commun aux deux rétroposons Ther-2 et CR1 (Figure



Les nucléotides de la séquence AC002530 strictements identiques au consensus LINE PsCR1 sont surlignés.

FIGURE 20 : Fragment de séquence humaine pouvant appartenir à la famille des éléments LINE CR1.

20). Cependant, la probabilité que cette identité soit aléatoire n'est pas négligeable. Nous avons essayé de détecter des éléments LINE CR1 chez les mammifères en effectuant des PCR avec des oligonucléotides spécifiques de la famille CR1. Aucun fragment d'amplification possédant la taille attendue n'a été obtenu.

3-4.3- Le LINE Bov-B.

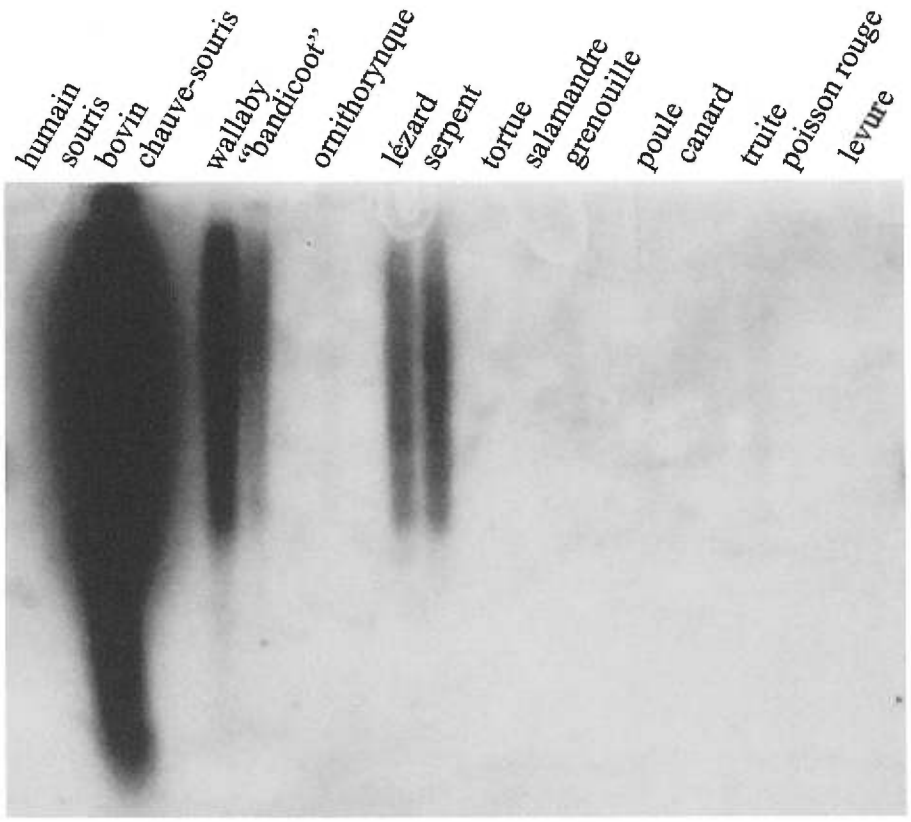
Le rétroposon Bov-B est présent dans les génomes des ruminants et ne se trouve dans aucun autre génome de mammifère placentaire. Des séquences possédant plus de 75 % d'identité avec ce LINE ont été identifiées dans les génomes des serpents et d'un certain nombre de lézards (Kordis et Gubensek, 1995; Kordis et Gubensek, 1997; Kordis et Gubensek, 1998).

La région variable de la famille Mar-1 présente une forte identité sur 95 nucléotides avec le LINE Bov-B. Par recherche dans les banques de données aucun élément n'a été trouvé chez les marsupiaux. En revanche, une expérience de Southern-blot avec différents génomes vertébrés utilisant une sonde spécifique de l'élément Bov-B (produit de PCR dite chaude avec les oligonucléotides B5 et B3) permet de détecter un signal d'hybridation pour les génomes marsupiaux (Figure 21). Le signal d'hybridation est entre 35 et 150 fois plus faible dans les génomes marsupiaux ou reptiliens que dans celui du bovin. Nous avons alors effectué des expériences de PCR en utilisant des amorces spécifiques des éléments Bov-B (Figure 22). L'amplification à partir des génomes représentant l'ensemble des lignées marsupiales produit un fragment de taille attendue. Le séquençage permet de conclure que des éléments homologues à Bov-B sont présents dans les génomes marsupiaux (Figure 23). La distribution de l'élément dans ces génomes montre que son

origine date d'environ 110 à 150 MA, avant l'expansion des espèces marsupiales (Kumar et Hedges, 1998; Rougier *et al.*, 1998).

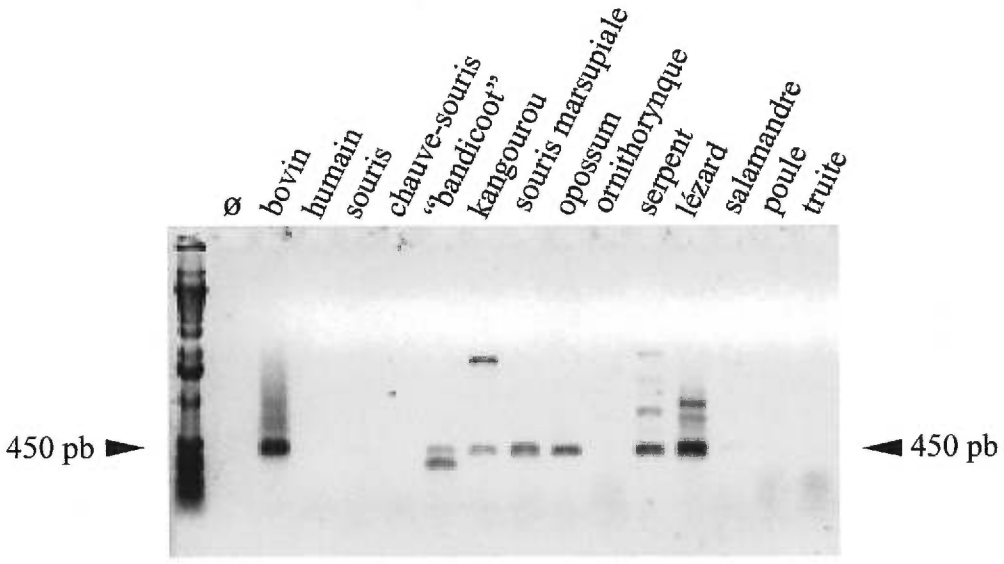
Les expériences de PCR (Figure 22) nous ont permis de détecter la présence d'éléments homologues à Bov-B chez la majorité des reptiles. Quelques produits de PCR chez le serpent et le bovin ont été séquencés (Figure 24). Afin de déterminer le lien entre les séquences Bov-B déjà publiées (bovins et reptiles) et les éléments que nous avons identifiés (marsupiaux, serpents et bovins) des études phylogénétiques ont été réalisées (Figure 25). La construction des arbres de vraisemblance maximale à partir des produits de PCR démontre la répartition des séquences en trois groupes distincts en fonction du génome d'origine (Figure 25a). Les éléments les plus divergents au sein d'un même groupe (ou d'un même génome) sont ceux des marsupiaux (wallaby); ce qui reflète leur origine plus ancienne. Les arbres phylogénétiques réalisés à partir des produits de PCR plus une sélection de séquences publiées (serpents et lézards) donne sensiblement les mêmes résultats (Figure 25b). Les trois groupes sont retrouvés et les nouvelles séquences des serpents n'augmentent pas la divergence au sein du groupe. Les séquences des lézards et des serpents anciens d'un point de vue évolutif (boa et python) nous apportent une information supplémentaire. Ces séquences s'ancrent dans le centre de l'arbre, indiquant leur lien plus proche avec la séquence originelle d'où proviennent les éléments Bov-B des trois groupes (Figure 25b). En regroupant toutes les séquences Bov-B des génomes des reptiles en un groupe phylogénétique, on constate que les éléments des génomes des bovins se trouvent intégrés à ce groupe (Figure 25b). La meilleure hypothèse pour expliquer cette situation est le transfert horizontal d'un élément Bov-B des reptiles vers l'ancêtre des génomes artiodactyles (Kordis et Gubensek, 1995; Kordis et Gubensek, 1998). En revanche, l'origine de l'élément chez les marsupiaux reste indéterminée.

FIGURE 21 : Détection des éléments Bov-B par Southern-Blot.

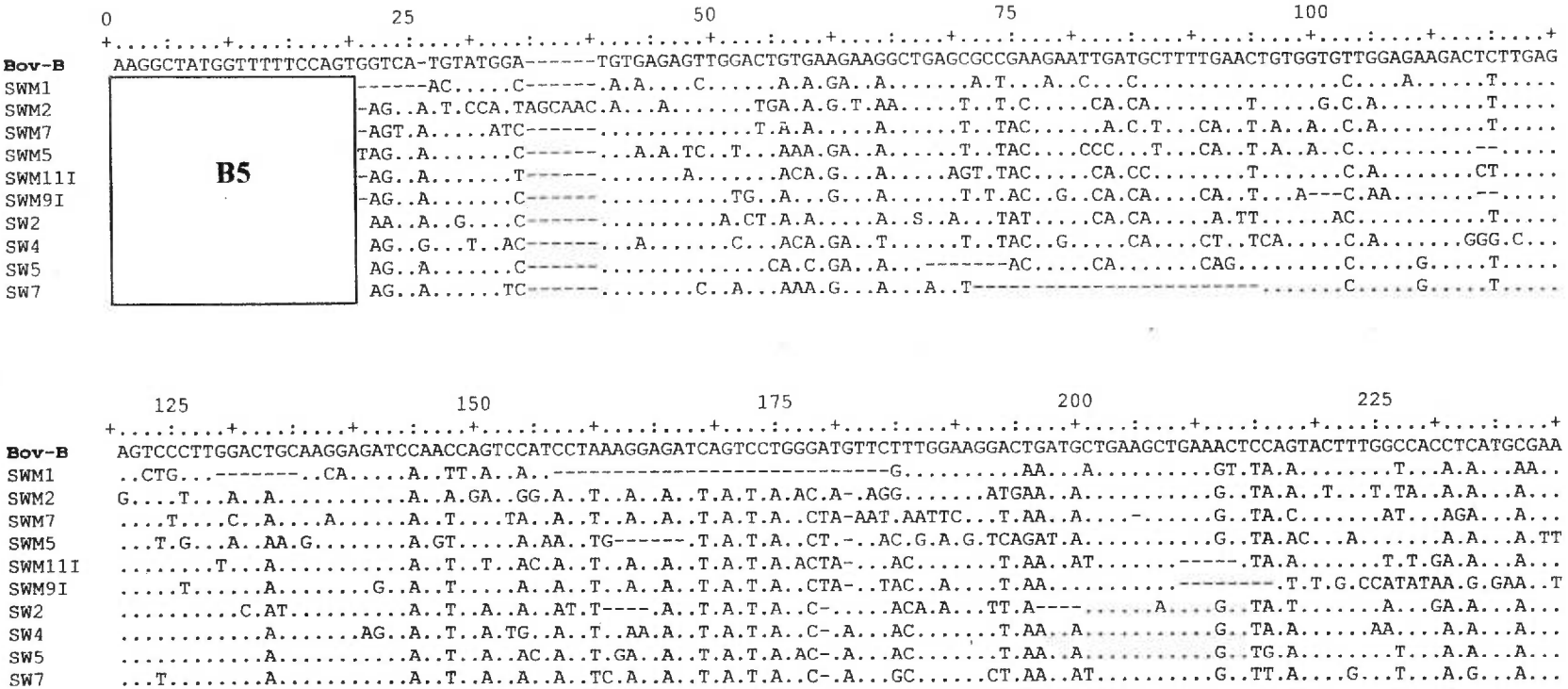


Les ADN génomiques ont subit une digestion complète par l'enzyme de restriction *Taq1*. La sonde est un produit de PCR radiomarqué spécifique des éléments Bov-B (les oligonucléotides utilisés sont B5 et B3).

FIGURE 22 : Détection des éléments Bov-B par PCR.



Ø désigne la réaction de PCR sans ADN, et la colonne de gauche représente l'échelle de poids moléculaire (1 Kb). La taille du fragment attendu est indiqué par les flèches. (La figure représente le négatif de la photo)



Les oligonucléotides utilisés pour les PCR sont indiqués aux extrémités des séquences. SW indique le génome utilisé (kangourou), M précise qu'il s'agit de la PCR avec l'oligonucléotide Bm3.

FIGURE 23 : Alignement des séquences marsupiales homologues à Bov-B (1/2).

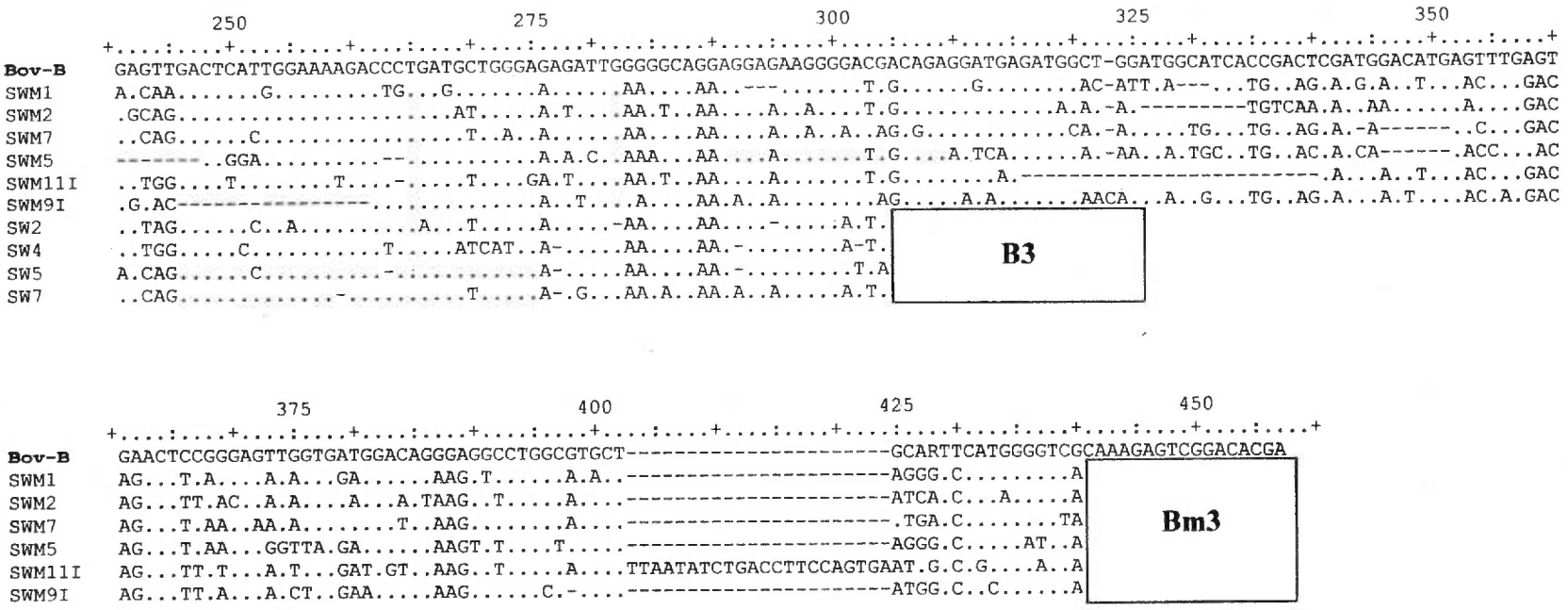
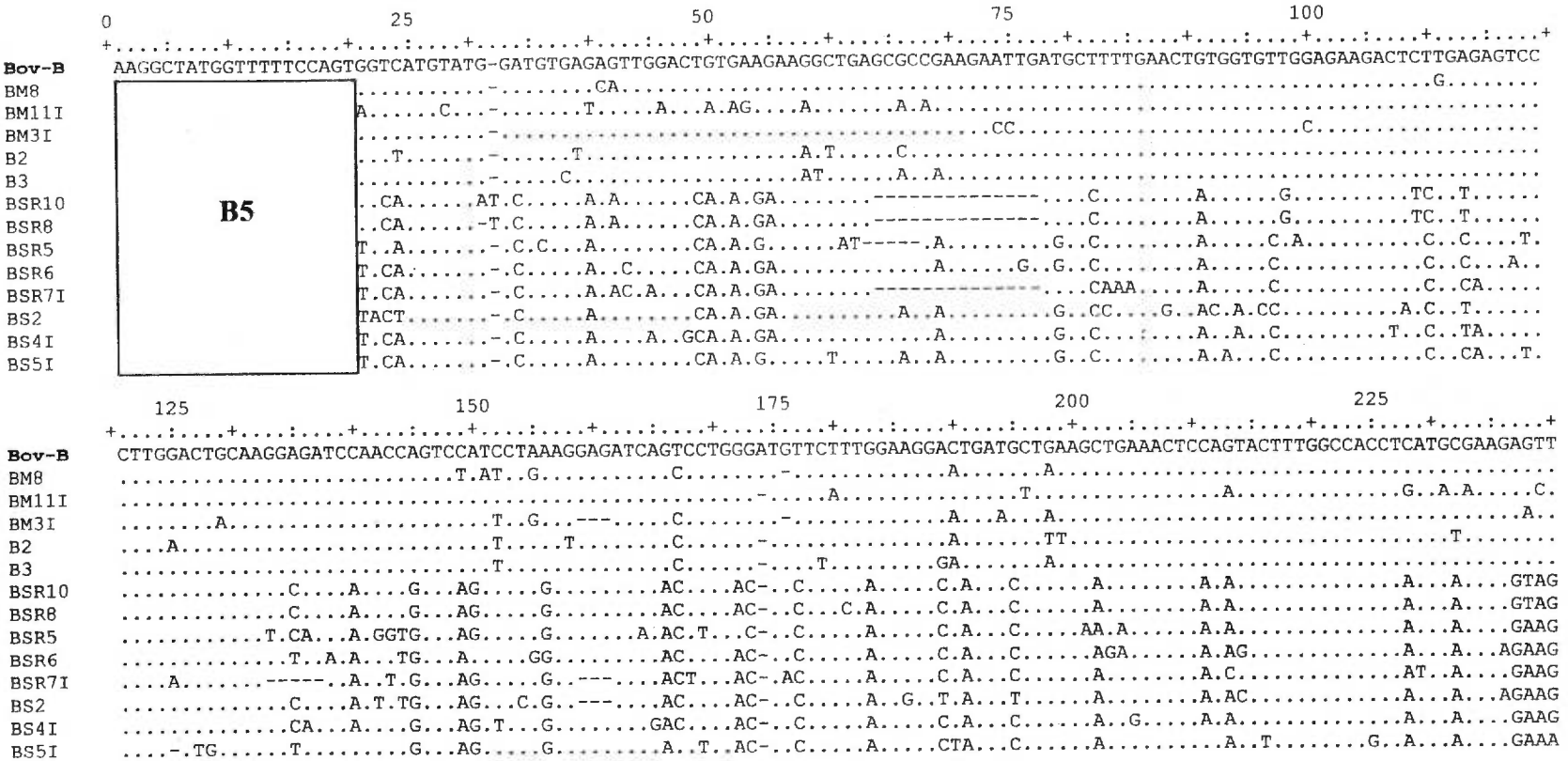


FIGURE 23 : Alignement des séquences marsupiales homologue à Bov-B (2/2).



Les oligonucléotides utilisés pour les PCR sont indiqués aux extrémités des séquences. B et BS indiquent les génomes utilisés (respectivement bovin et serpent), M et R précisent s'il s'agit des PCR avec l'oligonucléotide Bm3 ou Br3.

FIGURE 24 : Alignement des séquences de reptiles homologues à Bov-B avec des éléments Bov-B bovin (1/2).

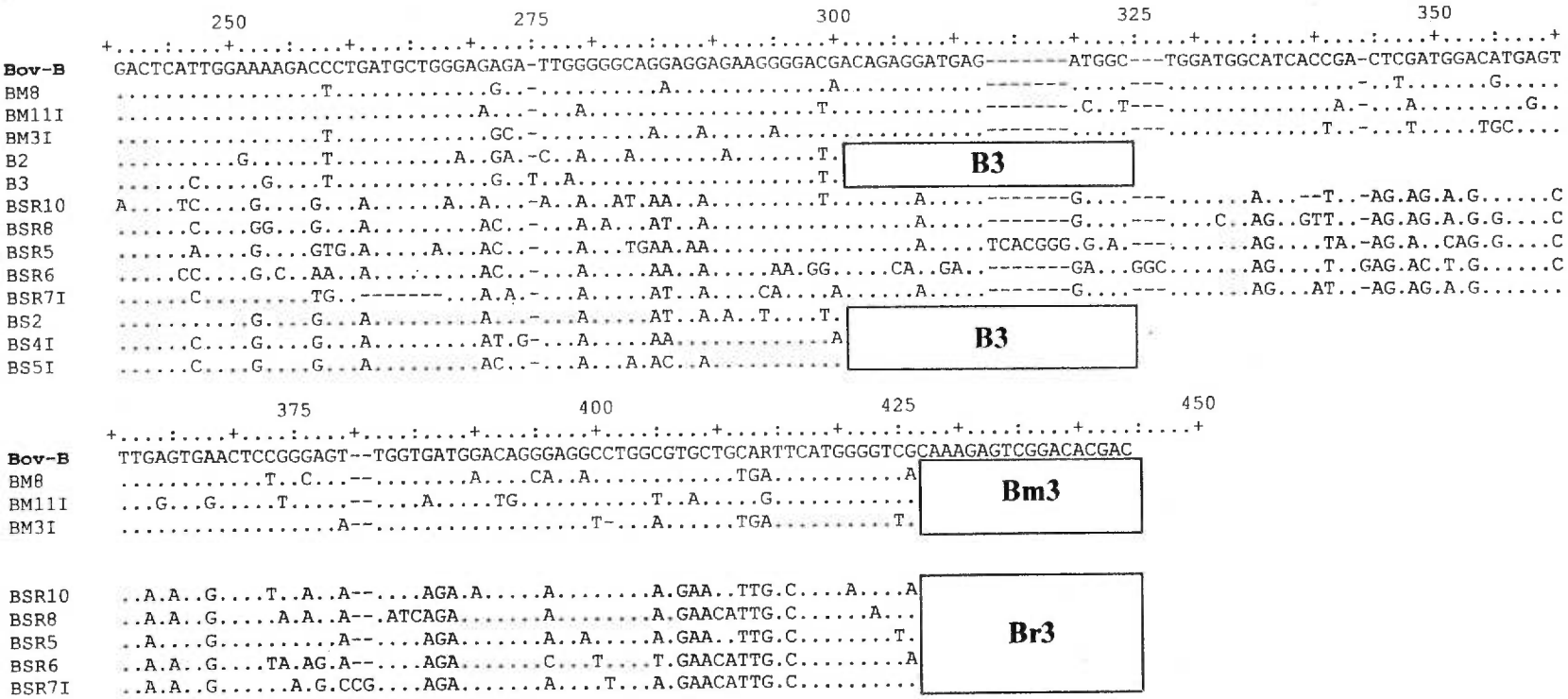


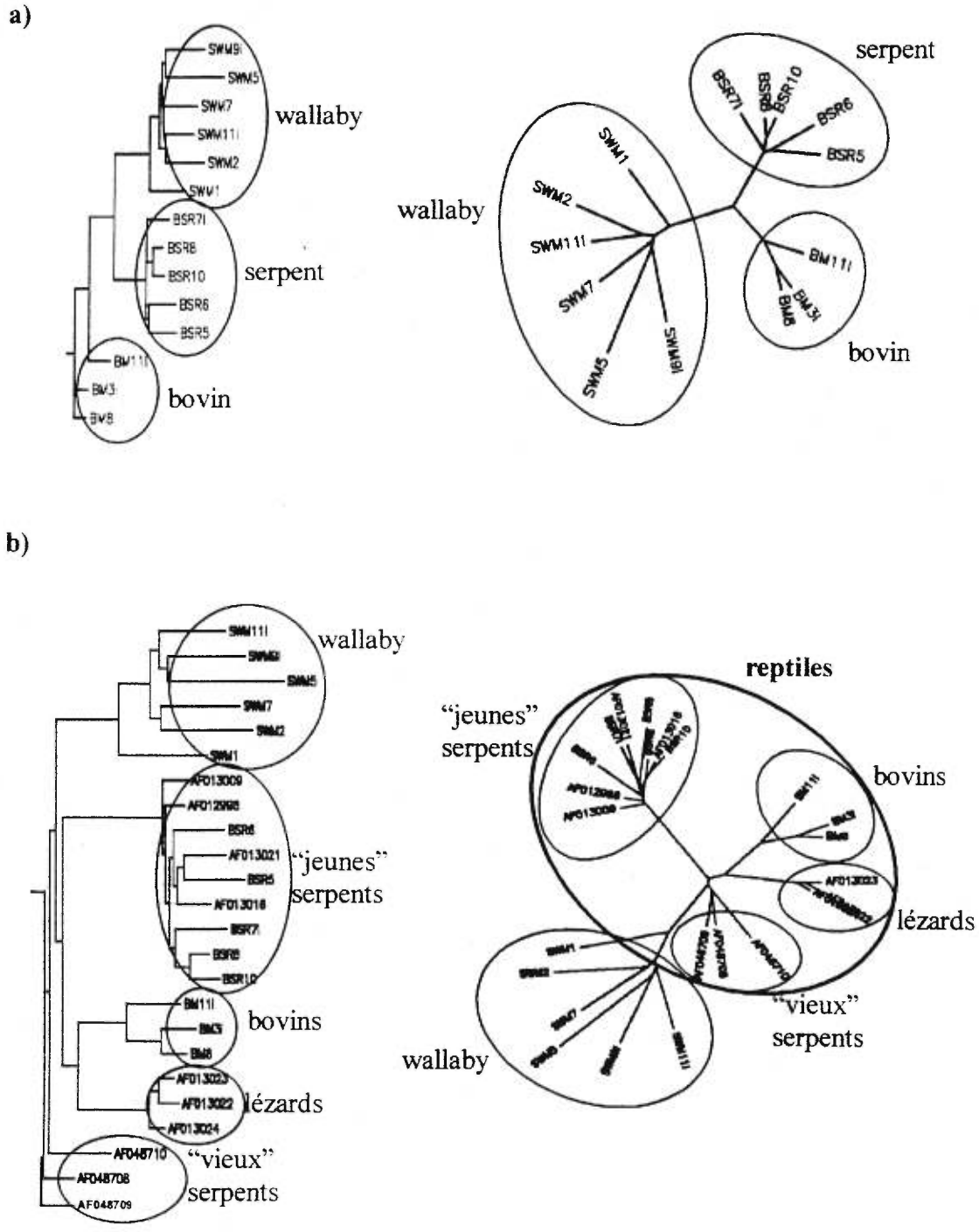
FIGURE 24 : Alignement des séquences de reptiles homologues à Bov-B avec des éléments Bov-B bovin (2/2).

FIGURE 25 : Analyse Phylogénétique des séquences Bov-B.

Arbres phylogénétiques établis par analyse de vraisemblance maximale : **a)** produit de PCR, **b)** produit de PCR plus séquences de reptiles publiées (Kordis et Gubensek, 1998). Les séquences qui servent de racine aux arbres situés à gauche de la figure sont BM8 (bovin) en **a)** et AF048708 (serpent) en **b).**

La longueur des branches est proportionnelle à la divergence des séquences. Les cercles regroupent les séquences d'un même génome et le nom des éléments est inscrit à chaque extrémité des branches. Les termes "vieux" et "jeunes" pour les séquences de serpents distinguent celles qui proviennent de génomes plus anciens d'un point de vue évolutif de celles des génomes plus récents.

FIGURE 25


a)
wallaby
serpent
bovin
b)
wallaby
"jeunes" serpents
bovins
lézards
"vieux" serpents
reptiles

DISCUSSION

4-1- Les familles CORE-SINE mammifères.

4-1.1- Les CORE-SINE sont des rétroposons dérivés d'ARNt.

Les travaux rapportés dans cette thèse ont permis d'identifier et de caractériser chez les mammifères 5 familles d'éléments SINE, très proches les unes des autres. Elles sont constituées d'une partie conservée, composée d'un promoteur de l'ARN Pol III pouvant dériver d'un ARNt suivi d'un domaine central appelé core, et d'une partie variable dérivant d'éléments LINE procurant à chaque famille sa spécificité.

Pour trois des familles (Mon-1, Mar-1 et Ther-1) des répétitions simples ont été trouvées aux extrémités 3'. Leur absence apparente dans les éléments de la famille Ther-2 s'explique par l'ancienneté des séquences. Pour la famille Opo-1 peu d'éléments ont été caractérisés et la séquence consensus décrite appartient probablement à un élément plus long du côté 3'.

Des répétitions terminales directes ont également été identifiées. Là encore, l'absence de RTD dans une majorité des éléments SINE (surtout placentaires) est liée à l'accumulation de mutations dans les séquences les plus anciennes empêchant leur détection.

Toutes ces caractéristiques indiquent l'appartenance de ces nouvelles familles à la classe des SINE dérivées d'ARNt. Nous avons appelé ce groupe de séquences les CORE-SINE parce qu'elles partagent une seconde caractéristique commune, le domaine core de 65 nucléotides. Le nombre de familles appartenant à ce groupe n'est peut être pas limité à 5 chez les mammifères. En effet, tous les génomes mammifères n'ayant pas été analysés

individuellement, et plus particulièrement ceux des marsupiaux, la liste des familles identifiées n'est pas exhaustive.

Pour chaque famille CORE-SINE, l'analyse de divergence des séquences n'a pas révélé de distribution aléatoire des mutations. Cette observation est en accord avec le modèle d'amplification des rétroposons selon lequel seule une petite portion des éléments d'une famille est active d'un point de vue rétropositionnel. Ce modèle devrait permettre de diviser chaque famille CORE-SINE en plusieurs sous-familles dans lesquelles la distribution des mutations deviendrait aléatoire. Nous avons ainsi identifié des sous-familles pour tous les CORE-SINE mammifères. Cependant, leur caractérisation reste incomplète. En effet, comme nous l'avons observé lors de l'analyse de divergence du segment core de la famille Mon-1 (Figure 9), les premières subdivisions ne sont pas suffisantes et il reste de nombreuses sous-familles non identifiées. Pour l'instant, le nombre d'éléments pleine longueur est trop faible pour définir d'autres positions diagnostiques dans les familles les plus récentes (familles non-placentaires). Pour les familles telles que Ther-1 ou Ther-2, l'ancienneté des séquences rend la distinction des positions diagnostiques presque impossible. Elles se confondent dans l'ensemble des mutations.

En général, les familles CORE-SINE mammifères sont anciennes et ne semblent plus actives. Une seule famille, spécifique du génome des monotrèmes, Mon-1, peut être encore active. En effet, l'analyse des sous-familles de Mon-1 et en particulier de Mon-1c, nous présente des éléments très peu divergeant de leur consensus. Une analyse plus détaillée de cette sous-famille, comprenant l'identification d'un plus grand nombre d'éléments et des expériences de RT-PCR, fournirait plus de renseignements à ce sujet.

4-1.2- Distribution des CORE-SINE chez les mammifères.

Étant donné la grande divergence des séquences par rapport à leur consensus, la détection des éléments dans les génomes ne représente qu'un échantillon. Nous avons néanmoins effectué une estimation expérimentale du nombre d'éléments CORE-SINE dans les génomes mammifères. Chez l'humain, nous avons dénombré près de 300 000 copies (Figure 3), évaluation proche de celle de Smit (400 000 copies) calculée à partir des séquences présentes dans les banques de données (Smit, 1996). Dans d'autres génomes mammifères placentaires (bovin et souris), le nombre d'éléments est similaire. Nous pouvons ainsi considérer que l'amplification des CORE-SINE détectables chez ces mammifères placentaires (familles Ther-1 et Ther-2) a précédé l'expansion des génomes. Chez les marsupiaux, les éléments des familles Ther-1 et Ther-2 identifiés sont moins divergents et laissent supposer que leur amplification s'est poursuivie dans ces génomes après la séparation des mammifères placentaires.

Dans les génomes mammifères non-placentaires les séquences CORE-SINE sont plus récentes et donc moins divergentes de leurs consensus. Cependant, ces génomes contenant au moins deux familles d'éléments CORE-SINE différentes (trois chez l'opossum), l'estimation du nombre de copies est difficile. En établissant une approximation de la divergence de l'ensemble des séquences (20 % chez les marsupiaux et 10 % chez les monotrèmes, voir les données pour chaque famille figure 10) et en quantifiant les signaux d'hybridation observés par dot-blot (tel que celui présenté sur la figure 2), nous avons évalué à 400 000 le nombre de copies d'éléments CORE-SINE dans ces génomes, ce qui représente environ 3 % de la masse génomique.

4-1.3- Modèle d'évolution des CORE-SINE mammifères.

La meilleure conservation du domaine core par rapport aux segments qui l'entourent biaise la datation des familles. Toutefois, si nous comparons les familles entre elles, ce biais est éliminé et nous pouvons alors établir la chronologie d'apparition des différentes séquences. L'âge estimé de la famille Ther-1, entre 60 et 180 MA, ainsi que la détection d'un signal d'hybridation avec une sonde spécifique de cette famille sur le génome de l'ornithorynque, nous permettent de conclure que l'origine des éléments précède la séparation des différentes lignées mammifères (datée au environ de 170 MA, (Kumar et Hedges, 1998)). Concernant Ther-2, l'âge calculé varie entre 70 et 140 MA, ce qui soutient le fait que les éléments de cette famille n'aient été observés que dans les génomes thériens. Cependant, pour cette famille nous avons que très peu de séquences et celles-ci ont été sélectionnées par hybridation. Nous avons donc a priori sélectionné les éléments les plus jeunes de cette famille. L'absence de signal par Southern-blot dans les génomes monotrèmes peut s'expliquer par une divergence trop forte des éléments avec la sonde utilisée, comme cela a été le cas pour les éléments de la famille Ther-1 qui n'ont pas été détectés par cette technique dans les génomes placentaires. De la même façon, la technique d'hybridation pour sélectionner les clones génomiques n'a peut-être pas été assez sensible pour détecter les éléments de la famille Ther-2 chez les monotrèmes. Il n'est donc pas impossible que les génomes de monotrème contiennent des éléments de la famille Ther-2.

Pour la famille Mar-1, l'âge moyen des séquences varie entre 40 et 80 MA. Ces éléments se trouvent dans tous les génomes marsupiaux, suggérant que leur apparition a précédé l'expansion des espèces de ce groupe de mammifères, c'est à dire entre 80 et 100 MA (Kirsch *et al.*, 1997).

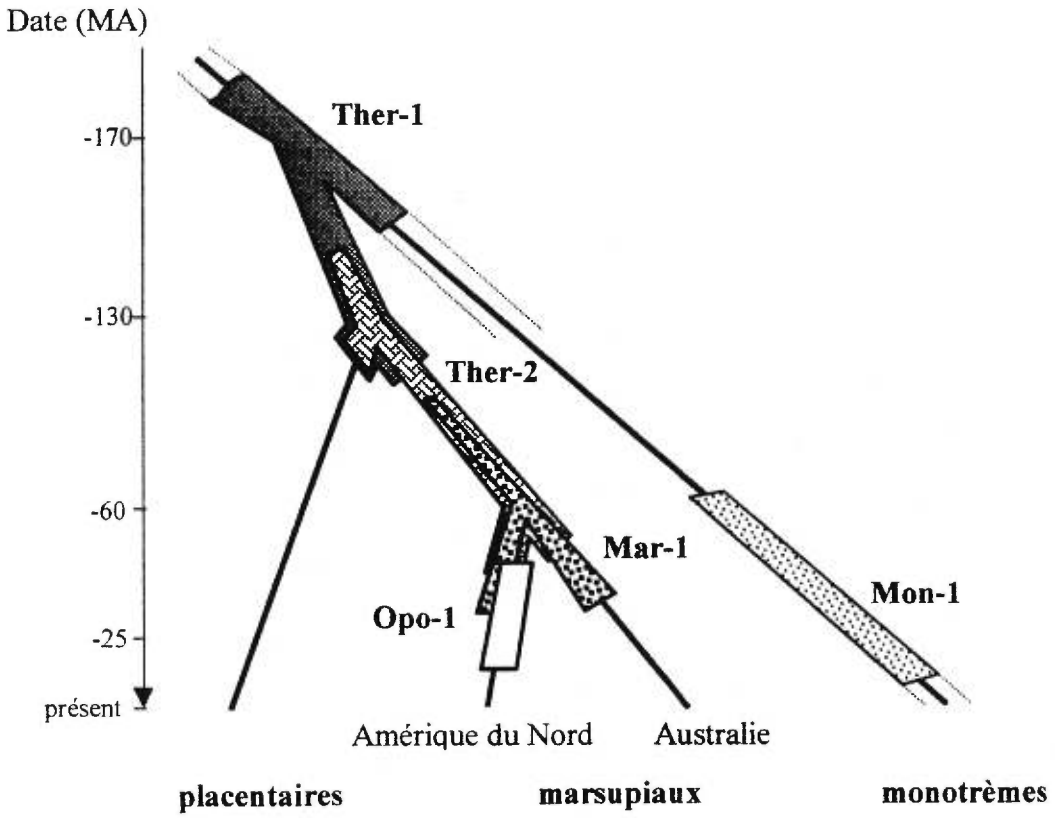
La famille Opo-1 possède le même âge moyen que Mar-1, mais doit être plus jeune puisqu'elle n'est pas retrouvée chez les marsupiaux d'Australie. Étant donné la divergence des éléments de cette famille par rapport à leur consensus (16,3%), nous pouvons supposer qu'ils ne sont pas spécifiques du génome de l'opossum de Virginie mais sont présents chez tous les marsupiaux d'Amérique du Nord. En effet, l'âge moyen estimé pour cette famille est de 40 à 80 MA et la séparation des marsupiaux d'Amérique du Nord de ceux d'Australie date d'environ 50 MA (Kirsch, 1997). La datation du début de l'expansion des espèces marsupiales d'Amérique du Nord à environ 40 MA renforce cette hypothèse.

Nous avons donc établi un modèle d'évolution de ces rétroposons chez les mammifères (Figure 26). La plus vielle famille identifiée, présente dans tous les génomes des mammifères, est Ther-1. Ensuite vient la famille Ther-2, détectée uniquement dans les génomes thériens. Après la séparation des marsupiaux et des mammifères placentaires, l'amplification des éléments CORE-SINE semble avoir cessé dans les gémones placentaires. En revanche, chez les marsupiaux, de nouvelles familles sont apparues. La première, Mar-1, est présente dans tous les génomes. La seconde, Opo-1, est spécifique des marsupiaux d'Amérique du Nord. Comme chez les marsupiaux, une nouvelle famille CORE-SINE est apparue dans les génomes monotrèmes (Mon-1).

4-2- Les familles CORE-SINE non-mammifères.

Dans ces travaux, nous avons aussi identifié des rétroposons appartenant à la super-famille des CORE-SINE chez de nombreux génomes non-mammifères. Il se peut également qu'il existe encore d'autres familles CORE-SINE non caractérisées à l'heure actuelle.

FIGURE 26 : Modèle d'évolution des familles CORE-SINE mammifères.



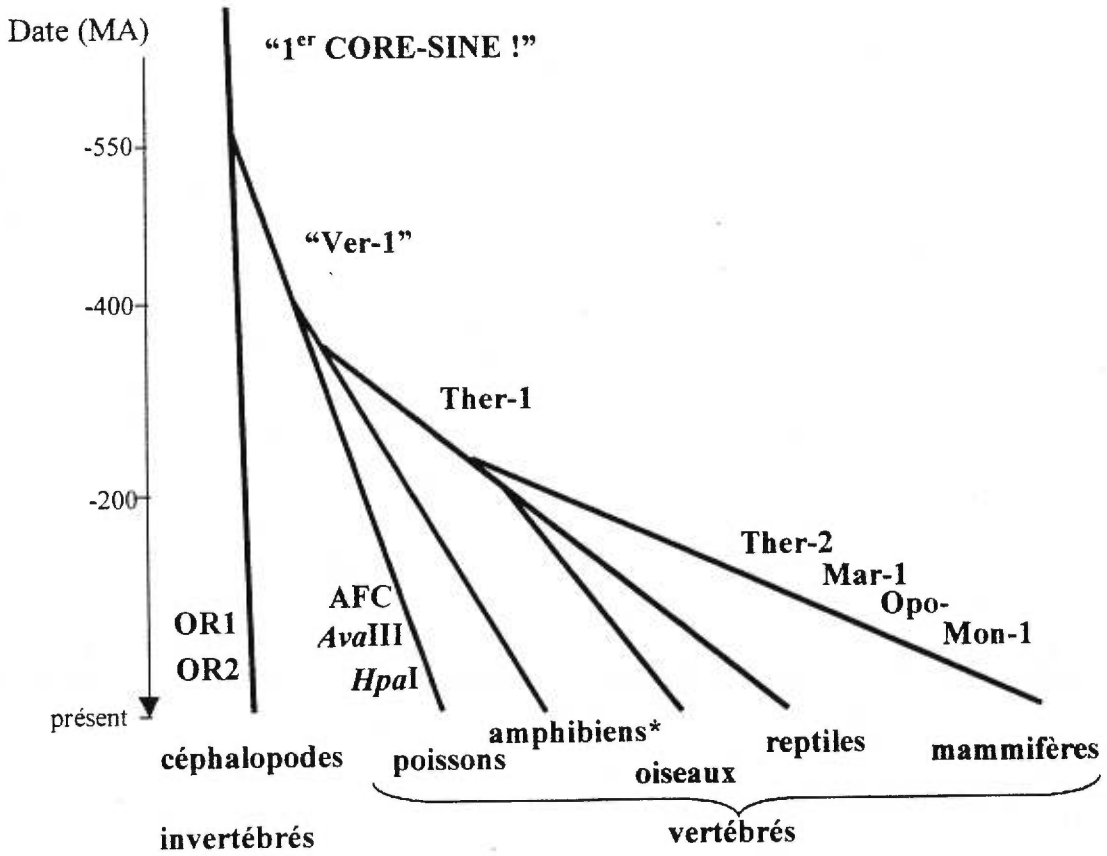
Chaque volume représente l'amplification d'une famille CORE-SINE. Les pointillés aux extrémités de Ther-1 indiquent la possibilité d'une amplification antérieure aux mammifères ou plus longue chez les monotrèmes. Les pointillés à l'extrémité de Mon-1 indiquent que cette famille est sans doute toujours active. L'échelle de temps se trouve sur la gauche.

Nous avons décrit des éléments de la famille Ther-1 dans les génomes d'oiseaux. Ce résultat suggère une amplification de ces mêmes rétroposons dans le génome des reptiles, ancêtre commun aux oiseaux et aux mammifères. Nous n'avons trouvé qu'un seul élément dans les banques de séquences de reptiles. Cependant, la représentation du génome des reptiles dans les banques de données est très faible (Tableau E, paragraphe 2-2.18.1). Par ailleurs, nous avons trouvé, en proportion, autant d'éléments chez les reptiles que chez les oiseaux, ce qui nous permet d'extrapoler le nombre d'éléments Ther-1 à quelques milliers. Ainsi, nous pouvons émettre l'hypothèse selon laquelle l'origine de la famille Ther-1 est antérieure aux mammifères et a pris son origine dans le génome des reptiles.

De façon surprenante nous avons aussi retrouvé le segment spécifique de la famille Ther-1 chez un poisson, *Fundulus heteroclitus* (Figure 14). Dans ce cas, la séquence n'a pas encore été caractérisée comme élément SINE, mais un certain nombre de structures spécifiques des rétroposons de petite taille sont identifiées (les boîtes A et B d'un promoteur à ARN Pol III). Il existe une différence entre cet élément et ceux de la famille Ther-1. En effet, les soixante derniers nucléotides, homologues à ceux du LINE L2, sont absents dans la séquence de poisson. Ainsi, l'apparition de la structure conservée des CORE-SINE suivie du "segment spécifique" de Ther-1 serait antérieure aux reptiles, et une extrémité 3' variable non caractérisée (dérivant probablement d'un élément LINE) s'y serait associée indépendamment. Cette découverte suggère l'existence d'une ancienne famille CORE-SINE commune aux vertébrés ("Ver-1", pour vertébré) (Figure 27). La démonstration de l'existence de la famille SINE dans le génome de *F. heteroclitus* serait une première étape pour donner des éléments de réponse à cette hypothèse.

Nous avons démontré, par comparaison de séquence du segment core, que les éléments CORE-SINE sont présents dans les génomes de poisson (*Hpa*I, *Ava*III et AFC).

FIGURE 27 : Modèle d'évolution des familles CORE-SINE eucaryotes.



Le nom de chaque famille CORE-SINE est associé à la branche à laquelle il appartient. Le "1er CORE-SINE" identifie un élément générique constitué d'une région conservée (ARNt plus core) associée à une région variable dérivée d'un élément LINE congénère. L'échelle de temps se trouve sur la gauche.

* Aucun élément CORE-SINE n'a été observé jusqu'à aujourd'hui dans les génomes amphibiens.

L'origine de cette super-famille de rétroposons serait même plus ancienne puisque le domaine core a aussi été retrouvé dans deux SINE du génome d'un céphalopode (invertébré). L'apparition des CORE-SINE remonte à au moins 550 MA, date estimée de la séparation des premiers vertébrés de la lignée des invertébrés (Futuyma, 1998). Un modèle d'évolution des familles CORE-SINE chez les eucaryotes est schématisé sur la figure 27.

4-3- Origine mosaïque des CORE-SINE.

4-3.1- Le segment dérivé d'ARNt, le promoteur.

Nous n'avons pas pu déterminer de façon certaine l'origine du segment 5' des familles CORE-SINE mammifères. Cependant, il semble que ce segment ait une origine unique pour toutes les familles. Dans aucun cas nous avons pu reconstruire la structure secondaire des ARNt. Nous n'avons pas, non plus, trouvé de conservation d'une structure secondaire quelconque entre les familles. A l'inverse, les éléments *Alu* et les autres SINE dérivés de l'ARN du gène 7SL ont subit des pressions de sélections pour conserver leur structure secondaire (Labuda et Striker, 1989; Labuda et Zietkiewicz, 1994; Zietkiewicz *et al.*, 1998). Celle-ci joue un rôle important pour la localisation des ARN au lieu de production des protéines intervenant dans la rétroposition. L'efficacité de rétroposition est ainsi étroitement liée à la structure secondaire de l'élément *Alu* (Sinnett *et al.*, 1991; Boeke, 1997). Dans le cas des séquences CORE-SINE, le modèle des éléments dérivés de 7SL ne peut pas être appliqué et le rôle du segment 5' serait ainsi limité à la transcription de l'élément.

4-3.2- Le domaine core, "l'échangeur".

Le domaine core est fortement conservé par rapport aux extrémités des séquences. Par alignement on observe qu'il l'est même entre les consensus des familles présentes dans des génomes très éloignés tels que ceux des poissons et des mammifères (Figure 13a). En complément, nous avons produit des comparaisons par dot-matrix (Figure 13b). Le résultat de cette approche indique que la conservation du domaine central est similaire à celle du promoteur. Ainsi, nous pouvons en conclure que le core est aussi important que le promoteur pour la survie et le maintien de l'activité rétropositionnelle. Il est préservé probablement parce qu'il augmente l'efficacité de rétroposition. Son rôle pourrait être directement lié à l'activité d'amplification en stimulant la transcription (enhancer) et/ou en stabilisant les transcrits. L'accumulation des transcrits favoriserait la compétition avec la machinerie rétropositionnelle endogène propre aux LINE. Le core pourrait aussi participer au routage de l'ARN vers les ribosomes, lieu de production des protéines de la machinerie rétropositionnelle. Il peut également jouer un rôle dans la promotion de la survie à long terme. Il servirait d'assembleur de segments potentiellement actifs pour la rétroposition et déjà présents dans le génome, c'est à dire des segments 5' promoteurs dérivés d'ARNt ou des segments 3' issus d'éléments LINE. L'acquisition de nouveaux domaines permet d'assurer la compétition avec des éléments LINE actifs. Le phénomène d'échange de segments 3' est observé dans différents génomes de mammifères, de poissons et de céphalopodes. Son mécanisme reste encore inconnu.

4-3.3- La région variable, le lien avec les rétroposons LINE.

Nous avons démontré que pour 4 des 5 familles CORE-SINE mammifères le segment variable dérive du fragment 3' d'éléments LINE. Cette caractéristique a été observée dans

plusieurs autres éléments dont les rétroposons CORE-SINE non-mammifères, tels que *Hpa*I et AFC (voir l'article de revue Okada *et al.* (1997) et Tableau C). L'identité entre les SINE et les LINE serait primordiale pour l'efficacité de la rétroposition. Elle permet la compétition avec les LINE pour la transcription inverse et l'intégration. Comme précédemment décrit dans l'introduction (paragraphes 1-3.3- et 1-4.2.3-, figures F et K), la machinerie rétropositionnelle reconnaît l'extrémité 3' des ARN pour initier la rétroposition et c'est cette portion que partagent les rétroposons SINE et LINE.

Comme pour certains CORE-SINE non-mammifères (*Ava*III, OR1 et OR2), le segment variable de la famille Opo-1 n'a apparemment pas d'identité avec un fragment 3' de LINE. Cependant, l'analyse d'identité a probablement été erronée par le peu de séquences répétées caractérisées chez les non-placentaires. Pour vérifier l'existence d'un rétroposon LINE possédant un segment 3' identique à Opo-1, un double criblage de la banque subgénomique de l'opossum pourrait être effectué. Le premier criblage, avec la sonde spécifique du segment variable (OP1) servirait à sélectionner les clones possédant le segment commun. Le second criblage, avec une sonde spécifique du segment core, permettrait d'éliminer les clones reconnus par les deux sondes. Par séquençage des clones uniquement positifs pour OP1 nous pourrons vérifier s'il existe une identité de séquence entre les clones en amont du segment variable. Si tel est le cas, il s'agira d'une nouvelle séquence répétée dispersée.

L'identification d'un segment partagé entre deux rétroposons implique que les deux soient présent dans le même génome. Pour les familles Ther-1 et Mon-1, la situation est vérifiée puisque le LINE L2, avec lequel elles partagent le segment 3', est présent chez tous les mammifères. Chez les marsupiaux, nous avons démontré l'existence d'un rétroposon LINE similaire à Bov-B, qui est à l'origine du segment variable de la famille Mar-1.

L'origine du segment variable de la famille Ther-2, fortement identique à l'extrémité des LINE CR1, reste énigmatique. Des expériences préliminaires de PCR n'ont pas révélé la présence de séquences identiques au LINE dans les génomes mammifères. De la même façon, nous n'avons pas identifié de séquences par criblage des banques de données des génomes placentaires. Pour démontrer la présence d'éléments identiques à CR1 chez les mammifères il faudrait approfondir les recherches dans le génome des marsupiaux, qui contiennent a priori, plus d'éléments Ther-2 mieux conservés. Nous pourrions identifier les éléments LINE à l'aide de la technique de double criblage décrite ci-dessus. La caractérisation d'un LINE identique à CR1 chez les marsupiaux permettrait d'extrapoler sa présence chez tous les mammifères puisque les reptiles, ancêtres communs aux deux groupes, contiennent l'élément. Ce serait le plus ancien rétroposon LINE des génomes vertébrés connu.

4-3.4- Origine du rétroposon CORE-SINE.

La question de l'origine du rétroposon CORE-SINE reste posée. Le promoteur dérive sans doute d'un ARNt et le segment 3' de différents LINE qui se sont succédés au cours de l'évolution. Cependant, nous ne connaissons pas l'origine du domaine central. Le core ne présente aucune identité particulière avec des séquences connues autres que celles des SINE. Une hypothèse serait que le domaine core ait à l'origine appartenu au segment 3' d'un élément LINE, quelques dizaines de nucléotides en amont de la répétition simple. Une séquence LINE tronquée juste en 5' du core, intégrée en aval d'un gène d'ARNt aurait pu générer le premier élément CORE-SINE. Un autre scénario serait qu'un rétrospeudogène d'ARNt se soit inséré en amont du core. Cette combinaison aurait augmenté l'efficacité de rétroposition et la capacité d'acquérir une extrémité 3' de LINE.

4-4- Le rétroposon LINE Bov-B.

Une étude phylogénétique a décrit qu'il y a environ 40 à 50 MA l'élément LINE Bov-B fut transmis de façon horizontale d'un génome de reptile vers un l'ancêtre des ruminants par un vecteur inconnu (Kordis et Gubensek, 1998). La présence du rétroposon Bov-B n'est pas observée chez tous les reptiles. D'après les résultats de PCR présenté par Kordis (Kordis, 1998) la distribution n'est pas continue d'un point de vue évolutif. Toutefois, les génomes possédant l'élément sont tous regroupés dans la lignée des squamés (serpents et lézards), ce qui permet de dater son origine entre 140 et 210 MA.

Nous avons démontré la présence dans tous les génomes marsupiaux d'un rétroposon LINE fortement identique à l'élément Bov-B. Nos résultats de PCR indiquent que l'origine de l'élément est antérieure à l'expansion de ces espèces, soit entre 110 et 170 MA (Kirsch, 1997; Kumar, 1998). Nous sommes donc dans une situation où trois groupes de génomes possèdent un rétroposon homologue, mais leur ancêtre commun semble en être dénue. Il reste ainsi à déterminer s'il s'agit aussi d'un transfert horizontal d'un élément des reptiles vers un ancêtre des marsupiaux (ou *vice versa*) ou si la discontinuité de la distribution est due à une "perte" de l'élément pour certains génomes. Le terme "perte" signifie que l'amplification de Bov-B n'a pas été efficace (très peu d'éléments) et s'est interrompue rapidement. De ce fait, les quelques séquences sont à l'état de fossiles et ne sont pas détectables par la technique de PCR utilisée. Des expériences sont en cours pour apporter des éléments de réponse à ces hypothèses.

4-5- Les SINE qui dérivent de Bov-B.

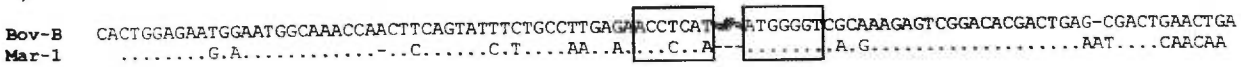
Il a été démontré que le rétroposon Bov-B est à l'origine de deux SINE dans le génome des bovins, Bov-A2 et Bov-tA. Ces deux SINE sont en fait issus d'un monomère A qui lui-même proviendrait de la délétion du fragment central du LINE (voir paragraphe 1-4.2.2.5-, Figure J) (Szemraj *et al.*, 1995).

Dans le génome des marsupiaux le même phénomène génétique s'est produit, à la différence près que la délétion n'est pas identique (Figure 19). Pour savoir si un monomère issu de cette délétion existe à l'état indépendant (non associé au core) dans le génome des marsupiaux, nous avons cherché dans les banques de données. Aucun monomère n'a été observé.

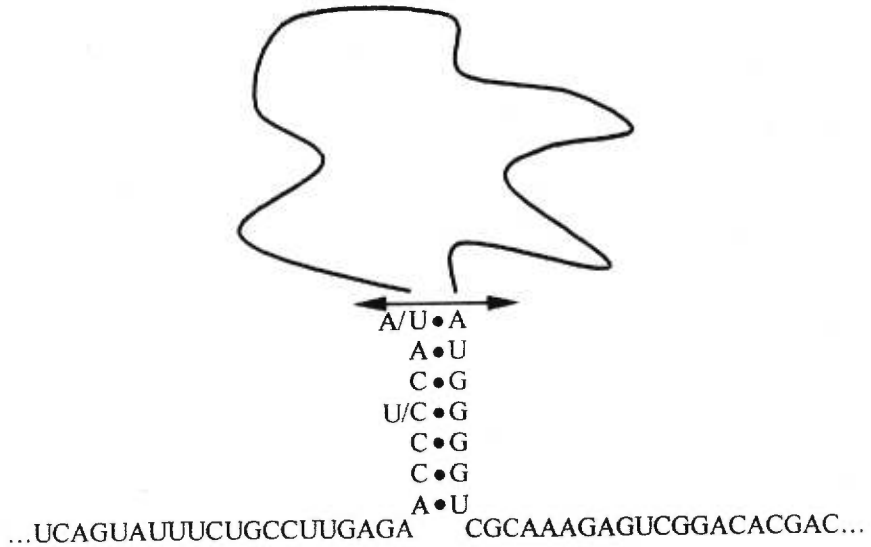
Pour expliquer ce mécanisme de délétion, la portion de séquence ARN du consensus Bov-B à l'origine du fragment variable du CORE-SINE Mar-1 a été analysée. Un appariement strict de 7 nucléotides (en direction sens) est possible juste au site de coupure (Figure 28). Cette structure secondaire permet de rapprocher les deux extrémités et une coupure de la "boucle" provoque la création du monomère. Le complexe rétropositionnel, reconnaissant les derniers nucléotides de la séquence, peut effectuer la transcription inverse du monomère et son intégration dans le génome. A priori, ce monomère n'a aucun pouvoir rétropositionnel car nous n'observons pas de promoteur bipartite de l'ARN Pol III. Cependant les 50 premiers nucléotides empruntés à la région 5' du rétroposon LINE Bov-B peut contenir un promoteur. Cette hypothèse est non négligeable étant donné que chez le bovin le monomère A est capable d'être amplifié par rétroposition, donc d'être transcrit.

FIGURE 28 : Origine du fragment variable du CORE-SINE Mar-1.

a)



b)



a) Alignement de la séquence Bov-B avec le segment variable de Mar-1.Les nucléotides participant à l'appariement sont encadrés. **b)** Schéma d'une structure secondaire possible au site de coupure (indiqué par la flèche). Les appariements de nucléotides sont indiqués par des points.

4-6- Évolution des rétroposons dans les génomes eucaryotes.

L'ensemble de nos résultats nous a amenés à la conclusion que les rétroposons détectables des génomes eucaryotes sont issus d'une longue évolution. Nous avons constaté que dans ces génomes, des rétroposons, aussi bien LINE que SINE, se succèdent au cours du temps. Dans cette évolution, ils conservent un certain nombre de fragments cruciaux. Dans le cas des LINE, il s'agit surtout des 7 domaines de la transcriptase inverse codée par l'ORF2, indispensables à la rétroposition (Xiong et Eickbush, 1990; Malik et Eickbush, 1998). Pour les CORE-SINE, nous avons démontré que c'est le domaine core qui est le mieux conservé. L'explication de ces conservations est sans doute liée à une sélection positive des cellules sur ces segments. Suite aux nombreux travaux réalisés sur les rétroposons et la rétroposition, il est reconnu que les LINE et les SINE sont des séquences "opportunistes" des génomes eucaryotes, mais aussi qu'elles jouent un rôle positif important dans l'évolution des génomes. L'observation de la grande variabilité créée par les rétroposons au sein d'une population démontre leur rôle dans le modelage et l'évolution des génomes (Sinnett, 1991). De plus, nous savons par l'identification de nombreuses maladies génétiques (Labuda *et al.*, 1995; Kazazian et Moran, 1998) que les rétroposons sont impliqués dans des mécanismes de recombinaisons qui induisent des insertions, des délétions, et des translocations. Il a aussi été démontré que la rétroposition pouvait être impliquée dans le mécanisme de réparation de l'ADN. Plusieurs travaux, chez la levure, indiquent que la transcriptase inverse est impliquée dans la réparation de coupure double brin d'ADN par insertion d'éléments mobiles tels que le rétrotransposon Ty1 ou l'élément chimère Ty1-L1 (ORF2 du rétroposon L1 dans une structure de Ty1 sous un promoteur de *GAL1*) (Boeke, 1996; Moore et Haber, 1996; Teng *et al.*, 1996). Enfin, un autre rôle de la

rétroposition dans l'évolution est que la transcriptase inverse intervient dans le retour de l'information dans le génome. Des travaux récents démontrent que la rétroposition d'un élément LINE peut induire la transcription inverse de séquences se trouvant en aval de l'élément (et dont la taille peut être de plusieurs Kb). Ce phénomène a pu être impliqué dans la dispersion de séquences telles que des promoteurs, des activateurs de transcription ou des exons (Moran *et al.*, 1999).

De la même façon qu'il existe une sélection des segments cruciaux de la rétroposition, il doit exister un contrôle de l'envahissement des génomes par les séquences mobiles. Chaque génome doit avoir un seuil de tolérance pour les séquences répétées d'une même famille. Cette tolérance peut être très variable suivant les espèces considérées (ou suivant la famille de rétroposon). En effet, le nombre de copies de l'élément Alu accepté par le génome humain est considérable (près de un million), alors que dans d'autres génomes le nombre de copies d'autres SINE se limite à quelques milliers. Cette tolérance est sans doute associée à la capacité de la cellule à contrôler les événements de recombinaisons entre séquences fortement identiques situées dans différentes localisations chromosomiques (recombinaison illégitime). Ce contrôle pourrait être en relation avec le mécanisme de réparation des mésappariements de l'ADN. Des expériences, chez la bactérie, indiquent que la recombinaison entre des séquences nucléotidiques variant de plus de 3% est fortement réduite si les protéines impliquées dans le mécanisme de réparation, MutS et MutL, sont intactes (Worth *et al.*, 1994). Un génome moins efficace pour la réparation des mésappariements subirait alors plus d'événements de recombinaisons susceptibles de le détruire si le nombre de copies est élevé.

A l'issue de nos travaux nous remarquons que l'apparition d'une nouvelle famille CORE-SINE est souvent associée à la période d'expansion des espèces qui contiennent le

nouvel élément. Par exemple : Ther-1 a accompagné le développement des génomes mammifères, ensuite Ther-2 celui des thériens (placentaire et marsupiaux), par la suite Mar-1 celui des marsupiaux et enfin Opo-1 celui des marsupiaux d'Amérique du Nord. Bien entendu nous ne pouvons pas définir si l'invasion génomique d'une nouvelle famille rétroposon est la cause ou l'effet de la divergence des espèces.

BIBLIOGRAPHIE

1. **Adams, D. S., Eickbush, T. H., Herrera, R. J., et Lizardi, P. M.** (1986). A highly reiterated family of transcribed oligo(A)-terminated, interspersed DNA elements in the genome of Bombyx mori. Journal of Molecular Biology **187:** 465-478.

2. **Adams, J. W., Kaufman, R. E., Kretschmer, P. J., Harrison, M., et Nienhuis, A. W.** (1980). A family of long reiterated DNA sequences, one copy of which is next to the human beta globin gene. Nucleic Acids Research **8:** 6113-6128.

3. **Adey, N. B., Schichman, S. A., Graham, D. K., Peterson, S. N., Edgell, M. H., et Hutchison, C. A.** (1994). Rodent L1 evolution has been driven by a single dominant lineage that has repeatedly acquired new transcriptional regulatory sequences. Molecular Biology & Evolution **11:** 778-789.

4. **Aksoy, S., Williams, S., Chang, S., et Richards, F. F.** (1990). SLACS retrotransposon from Trypanosoma brucei gambiense is similar to mammalian LINEs. Nucleic Acids Research **18:** 785-792.

5. **Alexander, L. J., Rohrer, G. A., Stone, R. T., et Beattie, C. W.** (1995). Porcine SINE-associated microsatellite markers: evidence for new artiodactyl SINEs. Mammalian Genome **6:** 464-468.

6. **Altschul, S. F., Gish, W., Miller, W., Myers, E. W., et Lipman, D. J.** (1990). Basic local alignment search tool. Journal of Molecular Biology **215:** 403-410.

7. **Anzai, K., Kobayashi, S., Suehiro, Y., et Goto, S.** (1987). Conservation of the ID sequence and its expression as small RNA in rodent brains: analysis with cDNA for mouse brain-specific small RNA. Brain Research **388:** 43-49.

8. **Arcot, S. S., Wang, Z., Weber, J. L., Deininger, P. L., et Batzer, M. A.** (1995). Alu repeats: a source for the genesis of primate microsatellites. Genomics **29:** 136-144.

9. **Baltimore, D.** (1970). Rna-dependent dna polymerase in virions of rna tumour viruses. Nature **226:** 1209-11.

10. **Barnett, T. R., Drake, L., et Pickle, W.** (1993). Human biliary glycoprotein gene: characterization of a family of novel alternatively spliced RNAs and their expressed proteins. Molecular & Cellular Biology **13:** 1273-1282.

11. **Bastien, L et Bourgaux, P.** (1987). The MT family os mouse DNA is made of short interspersed repeated elements. Gene **57:** 81-88.

12. **Batzer, M. A., Deininger, P. L., Hellmann-Blumberg, U., Jurka, J., Labuda, D., Rubin, C. M., Schmid, C. W., Zietkiewicz, E., et Zuckerkandl, E.** (1996). Standardized nomenclature for Alu repeats. Journal of Molecular Evolution **42:** 3-6.

13. **Batzer, M. A., Kilroy, G. E., Richard, P. E., Shaikh, T. H., Desselle, T. D., Hoppens, C. L., et Deininger, P. L.** (1990). Structure and variability of recently inserted Alu family members [published erratum appears in Nucleic Acids Res 1991 Feb 11;19(3):698-9]. Nucleic Acids Research **18:** 6793-6798.

14. **Besansky, N. J.** (1990). A retrotransposable element from the mosquito Anopheles gambiae [published erratum appears in Mol Cell Biol 1990 May;10(5):2442]. Molecular & Cellular Biology **10:** 863-871.

15. **Besansky, N. J., Bedell, J. A., et Mukabayire, O.** (1994). Q: a new retrotransposon from the mosquito Anopheles gambiae. Insect Molecular Biology **3:** 49-56.

16. **Biessmann, H. et Mason, J. M.** (1997). Telomere maintenance without telomerase. [Review] [85 refs]. Chromosoma **106:** 63-69.

17. **Bird, A. P.** (1980). DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Research **8:** 1499-1504.

18. **Blackburn, E. H.** (1991). Structure and function of telomeres. [Review] [74 refs]. Nature **350:** 569-573.

19. **Blesa, D. et Martinez-Sebastian, M. J.** (1997). bilbo, a non-LTR retrotransposon of Drosophila subobscura: a clue to the evolution of LINE-like elements in Drosophila. Molecular Biology & Evolution **14:** 1145-1153.

20. **Blinov, A. G., Sobanov, Y. V., Bogachev, S. S., Donchenko, A. P., et Filippova, M. A.** (1993). The Chironomus thummi genome contains a non-LTR retrotransposon. Molecular & General Genetics **237:** 412-420.

21. **Boeke, J. D.** (1996). DNA repair. A little help for my ends [news]. Nature **383:** 579.

22. **Boeke, J. D.** (1997). Lines and alus--the polya connection [news; comment]. Nature Genetics **16:** 6-7: 9.

23. **Boeke, J. D., Garfinkel, D. J., Styles, C. A., et Fink, G. R.** (1985). Ty elements transpose through an RNA intermediate. Cell **40:** 491-500.

24. **Boyes, J. et Bird, A.** (1991). DNA methylation inhibits transcription indirectly via a methyl-CpG binding protein. Cell **64:** 1123-1134.

25. **Bradfield, J. Y., Locke, J., et Wyatt, G. R.** (1985). An ubiquitous interspersed DNA sequence family in an insect. DNA **4:** 357-363.

26. **Bratthauer, G. L. et Fanning, T. G.** (1992). Active LINE-1 retrotransposons in human testicular cancer. Oncogene **7:** 507-510.

27. **Britten, R. J.** (1994). Evidence that most human Alu sequences were inserted in a process that ceased about 30 million years ago. Proceedings of the National Academy of Sciences of the United States of America **91:** 6148-6150.

28. **Britten, R. J., Baron, W. F., Stout, D. B., et Davidson, E. H.** (1988). Sources and evolution of human Alu repeated sequences. Proceedings of the National Academy of Sciences of the United States of America **85:** 4770-4774.

29. **Britten, R. J. et Davidson, E. H.** (1971). Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. Quarterly Review of Biology **46:** 111-38: 2.

30. **Britten, R. J. et Kohne, D. E.** (1968). Repeated sequences in DNA. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms. Science **161:** 529-40.

31. **Bryden, L. J., Denovan-Wright, E. M., et Wright, J. M.** (1999). ROn-1 SINEs: a tRNA-derived, short interspersed repetitive DNA element from *Oreochromis niloticus* and its species-specific distribution in Old World cichlid fishes. Molecular Marine Biology and Biotechnology **7:** 48-54.

32. **Bucheton, A.** (1990). I transposable elements and I-R hybrid dysgenesis in Drosophila. [Review] [36 refs]. Trends in Genetics **6:** 16-21.

33. **Buntjer, J. B., Hoff, I. A., et Lenstra, J. A.** (1997). Artiodactyl interspersed DNA repeats in cetacean genomes. Journal of Molecular Evolution **45:** 66-69.

34. **Burch, J. B., Davis, D. L., et Haas, N. B.** (1993). Chicken repeat 1 elements contain a pol-like open reading frame and belong to the non-long terminal repeat class of retrotransposons. Proceedings of the National Academy of Sciences of the United States of America **90:** 8199-8203.

35. **Burke, W. D., Calalang, C. C., et Eickbush, T. H.** (1987). The site-specific ribosomal insertion element type II of Bombyx mori (R2Bm) contains the coding sequence for a reverse transcriptase-like enzyme. Molecular & Cellular Biology **7:** 2221-2230.

36. **Busseau, I., Malinsky, S., Balakireva, M., Chaboissier, M. C., Teninges, D., et Bucheton, A.** (1998). A genetically marked I element in Drosophila melanogaster can be mobilized when ORF2 is provided in trans. Genetics **148:** 267-275.

37. **Campuzano, V., Montermini, L., Molto, M. D., Pianese, L., Cossee, M., Cavalcanti, F., Monros, E., Rodius, F., Duclos, F., et Monticelli, A.** (1996).

Friedreich's ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion [see comments]. Science **271:** 1423-1427.

38. **Carey, M. F. et Singh, K.** (1988). Enhanced B2 transcription in simian virus 40-transformed cells is mediated through the formation of RNA polymerase III transcription complexes on previously inactive genes. Proceedings of the National Academy of Sciences of the United States of America **85:** 7059-7063.

39. **Carey, M. F., Singh, K., Botchan, M., et Cozzarelli, N. R.** (1986). Induction of specific transcription by RNA polymerase III in transformed cells. Molecular & Cellular Biology **6:** 3068-3076.

40. **Cheng, J. F., Printz, R., Callaghan, T., Shuey, D., et Hardison, R. C.** (1984). The rabbit C family of short, interspersed repeats. Nucleotide sequence determination and transcriptional analysis. Journal of Molecular Biology **176:** 1-20.

41. **Chesnokov, I. et Schmid, C. W.** (1996). Flanking sequences of an Alu source stimulate transcription in vitro by interacting with sequence-specific transcription factors. Journal of Molecular Evolution **42:** 30-36.

42. **Coltman, D. W. et Wright, J. M.** (1994). Can SINEs: a family of tRNA-derived retroposons specific to the superfamily Canoidea. Nucleic Acids Research **22:** 2726-2730.

43. **Corpet, F.** (1988). Multiple sequence alignment with hierarchical clustering. Nucleic Acids Research **16:** 10881-10890.

44. **Daniels, G. R. et Deininger, P. L.** (1983). A second major class of Alu family repeated DNA sequences in a primate genome. Nucleic Acids Research **11:** 7595-7610.

45. **Daniels, G. R. et Deininger, P. L.** (1985). Repeat sequence families derived from mammalian tRNA genes. Nature **317:** 819-822.

46. **Daniels, G. R. et Deininger, P. L.** (1991). Characterization of a third major SINE family of repetitive sequences in the galago genome. Nucleic Acids Research **19:** 1649-1656.

47. **Danilevskaya, O., Slot, F., Pavlova, M., et Pardue, M. L.** (1994). Structure of the Drosophila HeT-A transposon: a retrotransposon-like element forming telomeres. Chromosoma **103:** 215-224.

48. **Das, G., Henning, D., Wright, D., et Reddy, R.** (1988). Upstream regulatory elements are necessary and sufficient for transcription of a U6 RNA gene by RNA polymerase III. EMBO Journal **7:** 503-512.

49. **DeChiara, T. M. et Brosius, J.** (1987). Neural BC1 RNA: cDNA clones reveal nonrepetitive sequence content [published erratum appears in Proc Natl Acad Sci U S A 1987 Jul;84(14):4895]. Proceedings of the National Academy of Sciences of the United States of America **84:** 2624-2628.

50. **Degen, S. J. et Davie, E. W.** (1987). Nucleotide sequence of the gene for human prothrombin. Biochemistry **26:** 6165-6177.

51. **Deininger, P. L. et Batzer, M. A.** (1993). Evolution of retroposons, dans Evolutionary Biology Hecht, MK and et al. New York. pp 157-196.

52. **Deininger, P. L. et Batzer, M. A.** (1995). SINE master genes and population biology, dans The impact of short interspersed elements (SINEs) on the host genome. Maraia, R. J. pp 43-60.

53. **Deininger, P. L., Batzer, M. A., Hutchison, C. A., et Edgell, M. H.** (1992). Master genes in mammalian repetitive DNA amplification. Trends in Genetics **8:** 307-311.

54. **Deininger, P. L., Jolly, D. J., Rubin, C. M., Friedmann, T., et Schmid, C. W.** (1981). Base sequence studies of 300 nucleotide renatured repeated human DNA clones. Journal of Molecular Biology **151:** 17-33: 12.

55. **Deininger, P. L. et Slagel, V. K.** (1988). Recently amplified Alu family members share a common parental Alu sequence. Molecular & Cellular Biology **8:** 4566-4569.

56. **Denison, R. A. et Weiner, A. M.** (1982). Human U1 RNA pseudogenes may be generated by both DNA- and RNA- mediated mechanisms. Molecular & Cellular Biology **2:** 815-828.

57. **Deragon, J. M., Gilbert, N., Rouquet, L., Lenoir, A., Arnaud, P., et Picard, G.** (1996). A transcriptional analysis of the S1Bn (Brassica napus) family of SINE retroposons. Plant Molecular Biology **32:** 869-878.

58. **Deragon, J. M., Landry, B. S., Pelissier, T., Tutois, S., Tourmente, S., et Picard, G.** (1994). An analysis of retroposition in plants based on a family of SINEs from Brassica napus. Journal of Molecular Evolution **39:** 378-386.

59. **Deragon, J. M., Sinnett, D., et Labuda, D.** (1990). Reverse transcriptase activity from human embryonal carcinoma cells NTera2D1. EMBO Journal **9:** 3363-3368.

60. **Dhellin, O., Maestre, J., et Heidmann, T.** (1997). Functional differences between the human LINE retrotransposon and retroviral reverse transcriptases for in vivo mRNA reverse transcription. EMBO Journal **16:** 6590-6602.

61. **Di Nocera, P. P.** (1988). Close relationship between non-viral retroposons in Drosophila melanogaster. Nucleic Acids Research **16:** 4041-4052.

62. **Di Nocera, P. P. et Casari, G.** (1987). Related polypeptides are encoded by Drosophila F elements, I factors, and mammalian L1 sequences. Proceedings of the National Academy of Sciences of the United States of America **84:** 5843-5847.

63. **Di Nocera, P. P. et Sakaki, Y.** (1990). LINEs: a superfamily of retrotransposable ubiquitous DNA elements. Trends in Genetics **6:** 29-30.

64. **Donehower, L. A., Slagle, B. L., Wilde, M., Darlington, G., et Butel, J. S.** (1989). Identification of a conserved sequence in the non-coding regions of many human genes. Nucleic Acids Research **17:** 699-710.

65. **Drew, A. C. et Brindley, P. J.** (1997). A retrotransposon of the non-long terminal repeat class from the human blood fluke Schistosoma mansoni. Similarities to the chicken-repeat-1-like elements of vertebrates. Molecular Biology & Evolution **14:** 602-610.

66. **Duncan, C., Biro, P. A., Choudary, P. V., Elder, J. T., Wang, R. R., Forget, B. G., de Riel, J. K., et Weissman, S. M.** (1979). RNA polymerase III transcriptional units are interspersed among human non-alpha-globin genes. Proceedings of the National Academy of Sciences of the United States of America **76:** 5095-5099.

67. **Duncan, C. H.** (1987). Novel Alu-type repeat in artiodactyls. Nucleic Acids Research **15:** 1340: 6.

68. **Duvernell, D. D. et Turner, B. J.** (1998). *Swimmer 1*,a New Low-Copy-Number Family in Teleost Genomes with Sequence Similarity to Mammalian L1. Molecular Biology & Evolution **15:** 1791-1793.

69. **Economou, E. P., Bergen, A. W., Warren, A. C., et Antonarakis, S. E.** (1990). The polydeoxyadenylate tract of Alu repetitive elements is polymorphic in the human genome. Proceedings of the National Academy of Sciences of the United States of America **87:** 2951-2954.

70. **Eickbush, T. H.** (1992). Transposing without ends: the non-LTR retrotransposable elements. [Review] [75 refs]. New Biologist **4:** 430-440.

71. **Eickbush, T. H.** (1997). Telomerase and retrotransposons: which came first? [comment]. Science **277:** 911-912.

72. **Elder, J. T., Pan, J., Duncan, C. H., et Weissman, S. M.** (1981). Transcriptional analysis of interspersed repetitive polymerase III transcription units in human DNA. Nucleic Acids Research **9:** 1171-1189.

73. **Endoh, H., Nagahashi, S., et Okada, N.** (1990). A highly repetitive and transcribable sequence in the tortoise genome is probably a retroposon. European Journal of Biochemistry **189:** 25-31.

74. **Fawcett, D. H., Lister, C. K., Kellett, E., et Finnegan, D. J.** (1986). Transposable elements controlling I-R hybrid dysgenesis in D. melanogaster are similar to mammalian LINEs. Cell **47:** 1007-1015.

75. **Felger, I. et Hunt, J. A.** (1992). A non-LTR retrotransposon from the Hawaiian Drosophila: the LOA element. Genetica **85:** 119-130.

76. **Felsenstein, J.** (1993). PHYLIP (Phylogeny Inference Package). version 3.5p. Distributed by the author. Department of genetics, University of Washington, Seattle.

77. **Feng, Q., Moran, J. V., Kazazian, H. H. Jr, et Boeke, J. D.** (1996). Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. Cell **87:** 905-916.

78. **Finnegan, D. J.** (1989). Eukaryotic transposable elements and genome evolution. [Review] [25 refs]. Trends in Genetics **5:** 103-107.

79. **Fuhrman, S. A., Deininger, P. L., LaPorte, P., Friedmann, T., et Geiduschek, E. P.** (1981). Analysis of transcription of the human Alu family ubiquitous repeating element by eukaryotic RNA polymerase III. Nucleic Acids Research **9:** 6439-6456.

80. **Futuyma, D. J.** (1998). Evolutionary Biology Sinauer Associates, Inc. Sunderland, Massachusetts

81. **Gabriel, A., Yen, T. J., Schwartz, D. C., Smith, C. L., Boeke, J. D., Sollner-Webb, B., et Cleveland, D. W.** (1990). A rapidly rearranging retrotransposon within the miniexon gene locus of Crithidia fasciculata. Molecular & Cellular Biology **10:** 615-624.

82. **Gabrielsen, O. S. et Sentenac, A.** (1991). RNA polymerase III (C) and its transcription factors. [Review] [40 refs]. Trends in Biochemical Sciences **16:** 412-416.

83. **Galli, G., Hofstetter, H., et Birnstiel, M. L.** (1981). Two conserved sequence blocks within eukaryotic tRNA genes are major promoter elements. Nature **294:** 626-631.

84. **Garrett, J. E., Knutzon, D. S., et Carroll, D.** (1989). Composite transposable elements in the Xenopus laevis genome. Molecular & Cellular Biology **9:** 3018-3027.

85. **Gilbert, N. et Labuda, D.** (1999). CORE-SINEs: Eukaryotic Short Interspersed Retroposing Elements with Common Sequence Motifs. Proceedings of the National Academy of Sciences of the United States of America **in press:**

86. **Gorr, T. H., Mable, B. K., et Kleinschmidt, T.** (1998). Phylogenetic Analysis of Reptilian Hemoglobins: Trees, Rates, and Divergences. Journal of Molecular Evolution **47:** 471-485.

87. **Grandbastien, M. A.** (1992). Retroelements in higher plants. [review] [41 refs]. Trends in Genetics **8:** 103-108.

88. **Greider, C. W. et Blackburn, E. H.** (1987). The telomere terminal transferase of Tetrahymena is a ribonucleoprotein enzyme with two kinds of primer specificity. Cell **51:** 887-898.

89. **Gutierrez-Hartmann, A., Lieberburg, I., Gardner, D., Baxter, J. D., et Cathala, G. G.** (1984). Transcription of two classes of rat growth hormone gene-associated repetitive DNA: differences in activity and effects of tandem repeat structure. Nucleic Acids Research **12:** 7153-7173.

90. **Hardies, S. C., Martin, S. L., Voliva, C. F., Hutchison, C. A., et Edgell, M. H.** (1986). An analysis of replacement and synonymous changes in the rodent L1 repeat family. Molecular Biology & Evolution **3:** 109-125.

91. **Harendza, C. J. et Johnson, L. F.** (1990). Polyadenylylation signal of the mouse thymidylate synthase gene was created by insertion of an L1 repetitive element downstream of the open reading frame. Proceedings of the National Academy of Sciences of the United States of America **87:** 2531-2535.

92. **Hata, K. et Sakaki, Y.** (1997). Identification of critical CpG sites for repression of L1 transcription by DNA methylation. Gene **189:** 227-234.

93. **Haynes, S. R. et Jelinek, W. R.** (1981). Low molecular weight RNAs transcribed in vitro by RNA polymerase III from Alu-type dispersed repeats in Chinese hamster DNA are also found in vivo. Proceedings of the National Academy of Sciences of the United States of America **78:** 6130-6134.

94. **Hentschel, C. C. et Birnstiel, M. L.** (1981). The organization and expression of histone gene families. [Review] [147 refs]. Cell **25:** 301-313.

95. **Higashiyama, T., Noutoshi, Y., Fujie, M., et Yamada, T.** (1997). Zepp, a LINE-like retrotransposon accumulated in the Chlorella telomeric region. EMBO Journal **16:** 3715-3723.

96. **Hohjoh, H. et Singer, M. F.** (1996). Cytoplasmic ribonucleoprotein complexes containing human LINE-1 protein and RNA. EMBO Journal **15:** 630-639.

97. **Hohjoh, H. et Singer, M. F.** (1997a). Ribonuclease and high salt sensitivity of the ribonucleoprotein complex formed by the human LINE-1 retrotransposon. Journal of Molecular Biology **271:** 7-12.

98. **Hohjoh, H. et Singer, M. F.** (1997b). Sequence-specific single-strand RNA binding protein encoded by the human LINE-1 retrotransposon. EMBO Journal **16:** 6034-6043.

99. **Hollis, G. F., Hieter, P. A., McBride, O. W., Swan, D., et Leder, P.** (1982). Processed genes: a dispersed human immunoglobulin gene bearing evidence of RNA-type processing. Nature **296:** 321-5: 9.

100. **Holmes, S. E., Singer, M. F., et Swergold, G. D.** (1992). Studies on p40, the leucine zipper motif-containing protein encoded by the first open reading frame of an active human LINE-1 transposable element. Journal of Biological Chemistry **267:** 19765-19768.

101. **Holmquist, G. P. et Caston, L. A.** (1986). Replication time of interspersed repetitive DNA sequences in hamsters. Biochimica et Biophysica Acta **868:** 164-177.

102. **Houck, C. M., Rinehart, F. P., et Schmid, C. W.** (1979). A ubiquitous family of repeated DNA sequences in the human genome. Journal of Molecular Biology **132:** 289-306.

103. **Hunter, A., Tsilfidis, C., Mettler, G., Jacob, P., Mahadevan, M., Surh, L., et Korneluk, R.** (1992). The correlation of age of onset with CTG trinucleotide repeat amplification in myotonic dystrophy. Journal of Medical Genetics **29:** 774-779.

104. **Hutchinson, G. B., Andrew, S. E., McDonald, H., Goldberg, Y. P., Graham, R., Rommens, J. M., et Hayden, M. R.** (1993). An Alu element retroposition in two families with Huntington disease defines a new active Alu subfamily. Nucleic Acids Research **21:** 3379-3383.

105. **Hutchison III, C. A., Hardies, S. C., Loeb, D. D. Shehee W. R., et Edgell, M. H.** (1989). LINEs and related retroposons: long interspersed repeated sequences in the eucaryotic genome., dans Mobile DNA Berg, D. E. and Howe, M. M. Washington D.C. pp 593-617.

106. **Ichimura, S., Mita, K., et Sugaya, K.** (1997). A major non-LTR retrotransposon of Bombyx mori, L1Bm. Journal of Molecular Evolution **45:** 253-264.

107. **Jagadeeswaran, P., Forget, B. G., et Weissman, S. M.** (1981). Short interspersed repetitive DNA elements in eucaryotes: transposable DNA elements generated by reverse transcription of RNA Pol III transcripts?. [review] [0 refs]. Cell **26:** 141-142.

108. **Jakubczak, J. L., Burke, W. D., et Eickbush, T. H.** (1991). Retrotransposable elements R1 and R2 interrupt the rRNA genes of most insects. Proceedings of the National Academy of Sciences of the United States of America **88:** 3295-3299.

109. **Janke, A., Gemmell, N. J., Feldmaier-Fuchs, G., von Haeseler, A, et Paabo, S.** (1996). The Mitochondrial Genome of a Monotreme - The Platypus (*Ornithorhynchus anatinus*). Journal of Molecular Evolution **42:** 153-159.

110. **Jeffreys, A. J., Wilson, V., et Thein, S. L.** (1985). Hypervariable 'minisatellite' regions in human DNA. Nature **314:** 67-73.

111. **Jelinek, W. R., Toomey, T. P., Leinwand, L., Duncan, C. H., Biro, P. A., Choudary, P. V., Weissman, S. M., Rubin, C. M., Houck, C. M., Deininger, P. L., et Schmid, C. W.** (1980). Ubiquitous, interspersed repeated sequences in mammalian genomes. Proceedings of the National Academy of Sciences of the United States of America **77:** 1398-1402.

112. **John, B. et Miklos, G. L.** (1979). Functional aspects of satellite dna and heterochromatin. [review] [259 refs]. International Review of Cytology **58:** 1-114: 16.

113. **Jurka, J.** (1993). A new subfamily of recently retroposed human Alu repeats. Nucleic Acids Research **21:** 2252.

114. **Jurka, J.** (1997). Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. Proceedings of the National Academy of Sciences of the United States of America **94:** 1872-1877.

115. **Jurka, J. et Klonowski, P.** (1996). Integration of retroposable elements in mammals: selection of target sites [letter]. Journal of Molecular Evolution **43:** 685-689.

116. **Jurka, J., Klonowski, P., et Trifonov, E. N.** (1998). Mammalian retroposons integrate at kinkable DNA sites. Journal of Biomolecular Structure & Dynamics **15:** 717-721.

117. **Jurka, J. et Milosavljevic, A.** (1991). Reconstruction and analysis of human Alu genes. Journal of Molecular Evolution **32:** 105-121.

118. **Jurka, J. et Smith, T.** (1988). A fundamental division in the Alu family of repeated sequences. Proceedings of the National Academy of Sciences of the United States of America **85:** 4775-4778.

119. **Jurka, J., Zietkiewicz, E., et Labuda, D.** (1995). Ubiquitous mammalian-wide interspersed repeats (MIRs) are molecular fossils from the mesozoic era. Nucleic Acids Research **23:** 170-175.

120. **Jurka, J. et Zuckerkandl, E.** (1991). Free left arms as precursor molecules in the evolution of Alu sequences. Journal of Molecular Evolution **33:** 49-56.

121. **Juttermann, R., Hosokawa, K., Kochanek, S., et Doerfler, W.** (1991). Adenovirus type 2 VAI RNA transcription by polymerase III is blocked by sequence-specific methylation. Journal of Virology **65:** 1735-1742.

122. **Kachroo, P., Leong, S. A., et Chattoo, B. B.** (1995). Mg-SINE: a short interspersed nuclear element from the rice blast fungus, Magnaporthe grisea. Proceedings of the National Academy of Sciences of the United States of America **92:** 11125-11129.

123. **Kajikawa, M., Ohshima, K., et Okada, N.** (1997). Determination of the entire sequence of turtle CR1: the first open reading frame of the turtle CR1 element encodes a protein with a novel zinc finger motif. Molecular Biology & Evolution **14:** 1206-1217.

124. **Kapitonov, V. V., Holmquist, G. P., et Jurka, J.** (1998). L1 repeat is a basic unit of heterochromatin satellites in cetaceans [letter]. Molecular Biology & Evolution **15:** 611-612.

125. **Kass, D. H., Kim, J., et Deininger, P. L.** (1996). Sporadic amplification of ID elements in rodents. Journal of Molecular Evolution **42:** 7-14.

126. **Kaufman, P. D. et Rio, D. C.** (1992). P element transposition in vitro proceeds by a cut-and-paste mechanism and uses GTP as a cofactor. Cell **69:** 27-39.

127. **Kazazian, H. H. Jr et Moran, J. V.** (1998). The impact of L1 retrotransposons on the human genome. [Review] [68 refs]. Nature Genetics **19:** 19-24.

128. **Kazazian, H. H. Jr, Wong, C., Youssoufian, H., Scott, A. F., Phillips, D. G., et Antonarakis, S. E.** (1988). Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. Nature **332:** 164-166.

129. **Kido, Y., Aono, M., Yamaki, T., Matsumoto, K., Murata, S., Saneyoshi, M., et Okada, N.** (1991). Shaping and reshaping of salmonid genomes by amplification of tRNA-derived retroposons during evolution. Proceedings of the National Academy of Sciences of the United States of America **88:** 2326-2330.

130. **Kido, Y., Himberg, M., Takasaki, N., et Okada, N.** (1994). Amplification of distinct subfamilies of short interspersed elements during evolution of the salmonidae. Journal of Molecular Biology **241:** 633-644.

131. **Kido, Y., Saitoh, M., Murata, S., et Okada, N.** (1995). Evolution of the active sequences of the HpaI short interspersed elements. Journal of Molecular Evolution **41:** 986-995.

132. **Kim, J., Kass, D. H., et Deininger, P. L.** (1995). Transcription and processing of the rodent ID repeat family in germline and somatic cells. Nucleic Acids Research **23:** 2245-2251.

133. **Kimmel, B. E., ole-MoiYoi, O. K., et Young, J. R.** (1987). Ingi, a 5.2-kb dispersed sequence element from Trypanosoma brucei that carries half of a smaller mobile element at either end and has homology with mammalian LINEs. Molecular & Cellular Biology **7:** 1465-1475.

134. **Kinsey, J. A.** (1990). Tad, a LINE-like transposable element of Neurospora, can transpose between nuclei in heterokaryons. Genetics **126:** 317-323.

135. **Kirsch, J. A. W., Lapointe, F.-J., et Springer, M. S.** (1997). DNA-hybridisation Studies of Marsupials and their Implications for Metatherian Classification. Australian Journal of Zoology **45:** 211-280.

136. **Knight, M., Miller, A., Raghavan, N., Richards, C., et Lewis, F.** (1992). Identification of a repetitive element in the snail Biomphalaria glabrata: relationship to the reverse transcriptase-encoding sequence in LINE-1 transposons. Gene **118:** 181-187.

137. **Kochanek, S., Renz, D., et Doerfler, W.** (1993). DNA methylation in the Alu sequences of diploid and haploid primary human cells. EMBO Journal **12:** 1141-1151.

138. **Kolosha, V. O. et Martin, S. L.** (1997). In vitro properties of the first ORF protein from mouse LINE-1 support its role in ribonucleoprotein particle formation during retrotransposition. Proceedings of the National Academy of Sciences of the United States of America **94:** 10155-10160.

139. **Kordis, D. et Gubensek, F.** (1995). Horizontal SINE transfer between vertebrate classes [letter]. Nature Genetics **10:** 131-132.

140. **Kordis, D. et Gubensek, F.** (1997). Bov-B long interspersed repeated DNA (LINE) sequences are present in Vipera ammodytes phospholipase A2 genes and in genomes of Viperidae snakes. European Journal of Biochemistry **246:** 772-779.

141. **Kordis, D. et Gubensek, F.** (1998). Unusual horizontal transfer of a long interspersed nuclear element between distant vertebrate classes. Proceedings of the National Academy of Sciences of the United States of America **95:** 10704-10709.

142. **Korenberg, J. R. et Rykowski, M. C.** (1988). Human genome organization: Alu, lines, and the molecular structure of metaphase chromosome bands. Cell **53:** 391-400.

143. **Kozak, M.** (1987). Effects of intercistronic length on the efficiency of reinitiation by eucaryotic ribosomes. Molecular & Cellular Biology **7:** 3438-3445.

144. **Krane, D. E. et Hardison, R. C.** (1990). Short interspersed repeats in rabbit DNA can provide functional polyadenylation signals. Molecular Biology & Evolution **7:** 1-8.

145. **Krayev, A. S., Kramerov, D. A., Skryabin, K. G., Ryskov, A. P., Bayev, A. A., et Georgiev, G. P.** (1980). The nucleotide sequence of the ubiquitous repetitive DNA sequence B1 complementary to the most abundant class of mouse fold-back RNA. Nucleic Acids Research **8:** 1201-1215.

146. **Krayev, A. S., Markusheva, T. V., Kramerov, D. A., Ryskov, A. P., Skryabin, K. G., Bayev, A. A., et Georgiev, G. P.** (1982). Ubiquitous transposon-like repeats B1 and B2 of the mouse genome: B2 sequencing. Nucleic Acids Research **10:** 7461-7475.

147. **Kubis, S. E., Heslop-Harrison, J. S., Desel, C., et Schmidt, T.** (1998). The genomic organization of non-LTR retrotransposons (LINEs) from three Beta species and five other angiosperms. Plant Molecular Biology **36:** 821-831.

148. **Kumar, S. et Hedges, B.** (1998). A molecular timescale for vertebrate evolution. Nature **392:** 917-919.

149. **Kurnit, D. M. et Maio, J. J.** (1973). Subnuclear redistribution of DNA species in confluent and growing mammalian cells. Chromosoma **42:** 23-36: 4.

150. **Kurose, K., Hata, K., Hattori, M., et Sakaki, Y.** (1995). RNA polymerase III dependence of the human L1 promoter and possible participation of the RNA polymerase II factor YY1 in the RNA polymerase III transcription system. Nucleic Acids Research **23:** 3704-3709.

151. **Labrador, M. et Corces, V. G.** (1997). Transposable element-host interactions: regulation of insertion and excision. [Review] [116 refs]. Annual Review of Genetics **31:** 381-404.

152. **Labuda, D., Sinnett, D., Richer, C., Deragon, J. M., et Striker, G.** (1991). Evolution of mouse B1 repeats: 7SL RNA folding pattern conserved. Journal of Molecular Evolution **32:** 405-414.

153. **Labuda, D. et Striker, G.** (1989). Sequence conservation in Alu evolution. Nucleic Acids Research **17:** 2477-2491.

154. **Labuda, D. et Zietkiewicz, E.** (1994). Evolution of secondary structure in the family of 7SL-like RNAs. Journal of Molecular Evolution **39:** 506-518.

155. **Labuda, D., Zietkiewicz, E., et Mitchell, G. A.** (1995). Alu elements as a source of genomic variation: Deleterious effects and evolutionary novelties., dans The

impact of short interspersed elements (SINEs) on the host genome. Maraia, R. J. pp 1-24.

156. **Lawrence, C. B., McDonnell, D. P., et Ramsey, W. J.** (1985). Analysis of repetitive sequence elements containing tRNA-like sequences. Nucleic Acids Research **13:** 4239-4252.

157. **Leeflang, E. P., Liu, W. M., Hashimoto, C., Choudary, P. V., et Schmid, C. W.** (1992). Phylogenetic evidence for multiple Alu source genes. Journal of Molecular Evolution **35:** 7-16.

158. **Leeton, P. R. et Smyth, D. R.** (1993). An abundant LINE-like element amplified in the genome of Lilium speciosum. Molecular & General Genetics **237:** 97-104.

159. **Leibold, D. M., Swergold, G. D., Singer, M. F., Thayer, R. E., Dombroski, B. A., et Fanning, T. G.** (1990). Translation of LINE-1 DNA elements in vitro and in human cells. Proceedings of the National Academy of Sciences of the United States of America **87:** 6990-6994.

160. **Lenstra, J. A., van Boxtel, J. A. F., Zwaagstra, K. A, et Schwerin, M** (1993). Short interspersed nuclear element (SINE) sequences of the bovidae. Animal Genetics **24:** 33-39.

161. **Li, W. H. et Tanimura, M.** (1987). The molecular clock runs more slowly in man than in apes and monkeys. Nature **326:** 93-96.

162. **Liu, W. M., Chu, W. M., Choudary, P. V., et Schmid, C. W.** (1995). Cell stress and translational inhibitors transiently increase the abundance of mammalian SINE transcripts. Nucleic Acids Research **23:** 1758-1765.

163. **Liu, W. M., Maraia, R. J., Rubin, C. M., et Schmid, C. W.** (1994). Alu transcripts: cytoplasmic localisation and regulation by DNA methylation. Nucleic Acids Research **22:** 1087-1095.

164. **Liu, W. M. et Schmid, C. W.** (1993). Proposed roles for DNA methylation in Alu transcriptional repression and mutational inactivation. Nucleic Acids Research **21:** 1351-1359.

165. **Luan, D. D., Korman, M. H., Jakubczak, J. L., et Eickbush, T. H.** (1993). Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. Cell **72:** 595-605.

166. **Majewska, K., Szemraj, J., Pluciennicak, G., Jaworski, J., et Pluciennicak, A.** (1988). A new family of dispersed, highly repetitive sequences in bovine genome. Biochimica et Biophysica Acta **949:** 119-124.

167. **Makalowski, W., Mitchell, G. A., et Labuda, D.** (1994). Alu sequences in the coding regions of mRNA: a source of protein variability. Trends in Genetics **10:** 188-193.

168. **Malik, H. S. et Eickbush, T. H.** (1998). The RTE class of non-LTR retrotransposons is widely distributed in animals and is the origin of many SINEs. Molecular Biology & Evolution **15:** 1123-1134.

169. **Maniatis, T., Fritsch, E. F., et Sambrook, J.** (1982). Molecular cloning, a laboratory manual Cold Spring Harbor laboratory. Cold Spring Harbor N.Y.

170. **Manuelidis, L.** (1976). Repeating restriction fragments of human DNA. Nucleic Acids Research **3:** 3063-3076.

171. **Margalit, H., Nadir, E., et Ben-Sasson, S. A.** (1994). A complete Alu element within the coding sequence of a central gene [letter]. Cell **78:** 173-174.

172. **Marin, I., Plata-Rengifo, P., Labrador, M., et Fontdevila, A.** (1998). Evolutionary Relationships Among the Members of an Ancient Class of Non-LTR Retrotransposons Found in the Nematode *Caenorhabditis elegans*. Molecular Biology & Evolution **15:** 1390-1402.

173. **Marschalek, R., Hofmann, J., Schumann, G., Gosseringer, R., et Dingermann, T.** (1992). Structure of DRE, a retrotransposable element which integrates with position specificity upstream of Dictyostelium discoideum tRNA genes. Molecular & Cellular Biology **12:** 229-239.

174. **Martin, F., Maranon, C., Olivares, M., Alonso, C., et Lopez, M. C.** (1995). Characterization of a non-long terminal repeat retrotransposon cDNA (L1Tc) from Trypanosoma cruzi: homology of the first ORF with the ape family of DNA repair enzymes. Journal of Molecular Biology **247:** 49-59.

175. **Martin, S. L.** (1991). Ribonucleoprotein particles with LINE-1 RNA in mouse embryonal carcinoma cells. Molecular & Cellular Biology **11:** 4804-4807.

176. **Matassi, G., Labuda, D., et Bernardi, G.** (1998). Distribution of the mammalian-wide interspersed repeats (MIRs) in the isochores of the human genome. FEBS **439:** 63-65.

177. **Matera, A. G., Hellmann, U., Hintz, M. F., et Schmid, C. W.** (1990). Recently transposed Alu repeats result from multiple source genes. Nucleic Acids Research **18:** 6019-6023.

178. **Mathias, S. L., Scott, A. F., Kazazian, H. H. Jr, Boeke, J. D., et Gabriel, A.** (1991). Reverse transcriptase encoded by a human transposable element. Science **254:** 1808-1810.

179. **Matsumoto, K., Murakami, K., et Okada, N.** (1986). Gene for lysine tRNA1 may be a progenitor of the highly repetitive and transcribable sequences present in the salmon genome. Proceedings of the National Academy of Sciences of the United States of America **83:** 3156-3160.

180. **Mayfield, J. E., McKenna, J. F., et Lessa, B. S.** (1980). The sequence organization of bovine DNA. Chromosoma **76:** 277-294.

181. **McMillan, J. P. et Singer, M. F.** (1993). Translation of the human LINE-1 element, L1Hs. Proceedings of the National Academy of Sciences of the United States of America **90:** 11533-11537.

182. **McNamara, P. T., Bolshoy, A., Trifonov, E. N., et Harrington, R. E.** (1990). Sequence-dependent kinks induced in curved DNA. Journal of Biomolecular Structure & Dynamics **8:** 529-538.

183. **Mellon, S. H., Baxter, J. D., et Gutierrez-Hartmann, A.** (1988). Cell-specific expression of transfected brain identifier repetitive DNAs. Nucleic Acids Research **16:** 3963-3976.

184. **Miles, C. et Meuth, M.** (1989). G-repeats: a novel hamster sine family. Nucleic Acids Research **17:** 7221-7228.

185. **Minchiotti, G. et Di Nocera, P. P.** (1991). Convergent transcription initiates from oppositely oriented promoters within the 5' end regions of Drosophila melanogaster F elements. Molecular & Cellular Biology **11:** 5171-5180.

186. **Mizrokhi, L. J., Georgieva, S. G., et Ilyin, Y. V.** (1988). jockey, a mobile Drosophila element similar to mammalian LINEs, is transcribed from the internal promoter by RNA polymerase II. Cell **54:** 685-691.

187. **Mochizuki, K., Umeda, M., Ohtsubo, H., et Ohtsubo, E.** (1992). Characterization of a plant SINE, p-SINE1, in rice genomes. Japanese Journal of Genetics **67:** 155-166.

188. **Moore, J. K. et Haber, J. E.** (1996). Capture of retrotransposon DNA at the sites of chromosomal double-strand breaks [see comments]. Nature **383:** 644-646.

189. **Moos, M. et Gallwitz, D.** (1983). Structure of two human beta-actin-related processed genes one of which is located next to a simple repetitive sequence. EMBO Journal **2:** 757-761.

190. **Moran, J. V., DeBerardinis, R. J., et Kazazian, H. H. Jr** (1999). L1 retrotransposition can shuffle exons. Science **283:** 1530-1534.

191. **Moran, J. V., Holmes, S. E., Naas, T. P., DeBerardinis, R. J., Boeke, J. D., et Kazazian, H. H. Jr** (1996). High frequency retrotransposition in cultured mammalian cells. Cell **87:** 917-927.

192. **Morin, G. B.** (1989). The human telomere terminal transferase enzyme is a ribonucleoprotein that synthesizes TTAGGG repeats. Cell **59:** 521-529.

193. **Mouches, C., Bensaadi, N., et Salvado, J. C.** (1992). Characterization of a LINE retroposon dispersed in the genome of three non-sibling Aedes mosquito species. Gene **120:** 183-190.

194. **Murnane, J. P. et Morales, J. F.** (1995). Use of a mammalian interspersed repetitive (MIR) element in the coding and processing sequences of mammalian genes. Nucleic Acids Research **23:** 2837-2839.

195. **Naas, T. P., DeBerardinis, R. J., Moran, J. V., Ostertag, E. M., Kingsmore, S. F., Seldin, M. F., Hayashizaki, Y., Martin, S. L., et Kazazian, H. H.** (1998). An actively retrotransposing, novel subfamily of mouse L1 elements. EMBO Journal **17:** 590-597.

196. **Nadir, E., Margalit, H., Gallily, T., et Ben-Sasson, S. A.** (1996). Microsatellite spreading in the human genome: evolutionary mechanisms and structural implications. Proceedings of the National Academy of Sciences of the United States of America **93:** 6470-6475.

197. **Nagahashi, S., Endoh, H., Suzuki, Y., et Okada, N.** (1991). Characterization of a tandemly repeated DNA sequence family originally derived by retroposition of tRNA(Glu) in the newt. Journal of Molecular Biology **222:** 391-404.

198. **Nakamura, T. M. et Cech, T. R.** (1998). Reversing time: origin of telomerase. [Review] [20 refs]. Cell **92:** 587-590.

199. **Nakamura, Y., Leppert, M., O'Connell, P., Wolff, R., Holm, T., Culver, M., Martin, C., Fujimoto, E., Hoff, M., Kumlin, E., et et al** (1987). Variable number of tandem repeat (VNTR) markers for human gene mapping. Science **235:** 1616-1622.

200. **Nisson, P. E., Hickey, R. J., Boshar, M. F., et Crain, W. R. Jr** (1988). Identification of a repeated sequence in the genome of the sea urchin which is transcribed by RNA polymerase III and contains the features of a retroposon. Nucleic Acids Research **16:** 1431-1452.

201. **Nur, I., Pascale, E., et Furano, A. V.** (1988). The left end of rat L1 (L1Rn, long interspersed repeated) DNA which is a CpG island can function as a promoter. Nucleic Acids Research **16:** 9233-9251.

202. **O'Hare, K., Alley, M. R., Cullingford, T. E., Driver, A., et Sanderson, M. J.** (1991). DNA sequence of the Doc retroposon in the white-one mutant of Drosophila melanogaster and of secondary insertions in the phenotypically altered derivatives white-honey and white-eosin. Molecular & General Genetics **225:** 17-24.

203. **Ohshima, K., Hamada, M., Terai, Y., et Okada, N.** (1996). The 3' ends of tRNA-derived short interspersed repetitive elements are derived from the 3' ends of long interspersed repetitive elements. Molecular & Cellular Biology **16:** 3756-3764.

204. **Ohshima, K., Koishi, R., Matsuo, M., et Okada, N.** (1993). Several short interspersed repetitive elements (SINEs) in distant species may have originated from a common ancestral retrovirus: characterization of a squid SINE and a possible mechanism for generation of tRNA-derived retroposons. Proceedings of the National Academy of Sciences of the United States of America **90:** 6260-6264.

205. **Ohshima, K. et Okada, N.** (1994). Generality of the tRNA origin of short interspersed repetitive elements (SINEs). Characterization of three different tRNA-derived retroposons in the octopus. Journal of Molecular Biology **243:** 25-37.

206. **Okada, N.** (1991a). SINEs. [Review] [33 refs]. Current Opinion in Genetics & Development **1:** 498-504.

207. **Okada, N.** (1991b). SINEs: Short Interspersed Repeated Elements of the Eukaryotic Genome. Trends in ecology and evolution **6:** 358-361.

208. **Okada, N. et Hamada, M.** (1997a). The 3' ends of tRNA-derived SINEs originated from the 3' ends of LINEs: a new example from the bovine genome. Journal of Molecular Evolution **44:** S52-S56.

209. **Okada, N., Hamada, M., Ogiwara, I., et Ohshima, K.** (1997b). SINEs and LINEs share common 3' sequences: a review. [review] [80 refs]. Gene **205:** 229-243.

210. **Okazaki, S., Ishikawa, H., et Fujiwara, H.** (1995). Structural analysis of TRAS1, a novel family of telomeric repeat-associated retrotransposons in the silkworm, Bombyx mori. Molecular & Cellular Biology **15:** 4545-4552.

211. **Oosumi, T., Belknap, W. R., et Garlick, B.** (1995). Mariner transposons in humans [letter]. Nature **378:** 672.

212. **Orr, H. T., Chung, M. Y., Banfi, S., Kwiatkowski, T. J. Jr, Servadio, A., Beaudet, A. L., McCall, A. E., Duvick, L. A., Ranum, L. P., et Zoghbi, H. Y.** (1993). Expansion of an unstable trinucleotide CAG repeat in spinocerebellar ataxia type 1. Nature Genetics **4:** 221-226.

213. **Pardue, M. L., Danilevskaya, O. N., Lowenhaupt, K., Slot, F., et Traverse, K. L.** (1996). Drosophila telomeres: new views on chromosome evolution. [Review] [19 refs]. Trends in Genetics **12:** 48-52.

214. **Paulson, K. E. et Schmid, C. W.** (1986). Transcriptional inactivity of Alu repeats in HeLa cells. Nucleic Acids Research **14:** 6145-6158.

215. **Peabody, D. S. et Berg, P.** (1986). Termination-reinitiation occurs in the translation of mammalian cell mRNAs. Molecular & Cellular Biology **6:** 2695-2703.

216. **Pearson, W. R. et Lipman, D. J.** (1988). Improved tools for biological sequence comparison. Proceedings of the National Academy of Sciences of the United States of America **85:** 2444-2448.

217. **Penny, D. et Hasegawa, M.** (1997). The platypus put in its place. Nature **387:** 549-550.

218. **Pimpinelli, S., Berloco, M., Fanti, L., Dimitri, P., Bonaccorsi, S., Marchetti, E., Caizzi, R., Caggese, C., et Gatti, M.** (1995). Transposable elements are stable structural components of Drosophila melanogaster heterochromatin. Proceedings of the National Academy of Sciences of the United States of America **92:** 3804-3808.

219. **Pont-Kingdon, G., Chi, E., Christensen, S., et Carroll, D.** (1997). Ribonucleoprotein formation by the ORF1 protein of the non-LTR retrotransposon Tx1L in Xenopus oocytes. Nucleic Acids Research **25:** 3088-3094.

220. **Priimagi, A. F., Mizrokhi, L. J., et Ilyin, Y. V.** (1988). The Drosophila mobile element jockey belongs to LINEs and contains coding sequences homologous to some retroviral proteins. Gene **70:** 253-262.

221. **Quentin, Y.** (1988). The Alu family developed through successive waves of fixation closely connected with primate lineage history. Journal of Molecular Evolution **27:** 194-202.

222. **Quentin, Y.** (1989). Successive waves of fixation of B1 variants in rodent lineage history. Journal of Molecular Evolution **28:** 299-305.

223. **Quentin, Y.** (1992a). Fusion of a free left Alu monomer and a free right Alu monomer at the origin of the Alu family in the primate genomes. Nucleic Acids Research **20:** 487-493.

224. **Quentin, Y.** (1992b). Origin of the Alu family: a family of Alu-like monomers gave birth to the left and the right arms of the Alu elements. Nucleic Acids Research **20:** 3397-3401.

225. **Rasmussen, M., Rossen, L., et Giese, H.** (1993). SINE-like properties of a highly repetitive element in the genome of the obligate parasitic fungus Erysiphe graminis f.sp. hordei. Molecular & General Genetics **239:** 298-303.

226. **Rinehart, F. P., Ritch, T. G., Deininger, P. L., et Schmid, C. W.** (1981). Renaturation rate studies of a single family of interspersed repeated sequences in human deoxyribonucleic acid. Biochemistry **20:** 3003-3010.

227. **Rogers, J.** (1983). Retroposons defined [letter]. Nature **301:** 460.

228. **Rogers, J. H.** (1985). The origin and evolution of retroposons. [Review] [631 refs]. International Review of Cytology **93:** 187-279.

229. **Rothkopf, G. S., Telakowski-Hopkins, C. A., Stotish, R. L., et Pickett, C. B.** (1986). Multiplicity of glutathione S-transferase genes in the rat and association with a type 2 Alu repetitive element. Biochemistry **25:** 993-1002.

230. **Rougier, G. W, Wible, J. R., et Novacek, M. J.** (1998). Implication of *Deltatheridium* specimens for early marsupial history. Nature **396:** 459-463.

231. **Rubin, C. M., VandeVoort, C. A., Teplitz, R. L., et Schmid, C. W.** (1994). Alu repeated DNAs are differentially methylated in primate germ cells. Nucleic Acids Research **22:** 5121-5127.

232. **Rudiger, N. S., Gregersen, N., et Kielland-Brandt, M. C.** (1995). One short well conserved region of Alu-sequences is involved in human gene rearrangements and has homology with prokaryotic chi. Nucleic Acids Research **23:** 256-260.

233. **Saba, J. A., Busch, H., et Reddy, R.** (1985). A new moderately repetitive rat DNA sequence detected by a cloned 4.5 SI DNA. Journal of Biological Chemistry **260:** 1354-1357.

234. **Sakagami, M., Ohshima, K., Mukoyama, H., Yasue, H., et Okada, N.** (1994). A novel tRNA species as an origin of short interspersed repetitive elements (SINEs). Equine SINEs may have originated from tRNA(Ser). Journal of Molecular Biology **239:** 731-735.

235. **Sakamoto, K. et Okada, N.** (1985). Rodent type 2 alu family, rat identifier sequence, rabbit c family, and bovine or goat 73-bp repeat may have evolved from tRNA genes. Journal of Molecular Evolution **22:** 134-140.

236. **SanMiguel, P., Tikhonov, A., Jin, Y. K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P. S., Edwards, K. J., Lee, M., Avramova, Z., et Bennetzen, J. L.** (1996). Nested retrotransposons in the intergenic regions of the maize genome [see comments]. Science **274:** 765-768.

237. **Sapienza, C. et St-Jacques, B.** (1986). 'Brain-specific' transcription and evolution of the identifier sequence. Nature **319:** 418-420.

238. **Sarrowa, J., Chang, D. Y., et Maraia, R. J.** (1997). The decline in human Alu retroposition was accompanied by an asymmetric decrease in SRP9/14 binding to dimeric Alu RNA and increased expression of small cytoplasmic Alu RNA. Molecular & Cellular Biology **17:** 1144-1151.

239. **Sassaman, D. M., Dombroski, B. A., Moran, J. V., Kimberland, M. L., Naas, T. P., DeBerardinis, R. J., Gabriel, A., Swergold, G. D., et Kazazian, H. H. Jr**

(1997). Many human L1 elements are capable of retrotransposition [see comments]. Nature Genetics **16:** 37-43.

240. **Schichman, S. A., Adey, N. B., Edgell, M. H., et Hutchison, C. A.** (1993). L1 A-monomer tandem arrays have expanded during the course of mouse L1 evolution. Molecular Biology & Evolution **10:** 552-570.

241. **Schmid, C. et Maraia, R.** (1992). Transcriptional regulation and transpositional selection of active SINE sequences. [Review] [61 refs]. Current Opinion in Genetics & Development **2:** 874-882.

242. **Schmid, C. W.** (1991). Human Alu subfamilies and their methylation revealed by blot hybridization. Nucleic Acids Research **19:** 5613-5617.

243. **Schmid, C. W. et Shen, C.-K. J** (1985). The evolution of interspersed repetitive DNA sequences in mammals and other vertebrates, dans Molecular evolutionary genetics MacIntyre, R. J. New York. pp 323-358.

244. **Schmidt, T., Kubis, S., et Heslop-Harrison, J. S.** (1995). Analysis and chromosomal localization of retrotransposons in sugar beet (Beta vulgaris L.): LINEs and Ty1-copia-like elements as major components of the genome. Chromosome Research **3:** 335-345.

245. **Schulte, P. M., Gomez-Chiarri, M., et Powers, D. A.** (1997). Structural and functional differences in the promoter and 5' flanking region of Ldh-B within and between populations of the teleost Fundulus heteroclitus. Genetics **145:** 759-769.

246. **Schwartz, A., Chan, D. C., Brown, L. G., Alagappan, R., Pettay, D., Disteche, C., McGillivray, B., de la Chapelle, A., et Page, D. C.** (1998). Reconstructing hominid Y evolution: X-homologous block, created by X-Y transposition, was disrupted by Yp inversion through LINE-LINE recombination. Human Molecular Genetics **7:** 1-11.

247. **Schwarz-Sommer, Z., Leclercq, L., Gobel, E., et Saedler, H** (1987). Cin4, an insert altering the structure of the *A1* gene in *Zea mays*, exhibits properties of nonviral retrotransposons. EMBO Journal **6:** 3873-3880.

248. **Sharp, P. A.** (1983). Conversion of RNA to DNA in mammals: alu-like elements and pseudogenes. Nature **301:** 471-472.

249. **Sheen, F. M. et Levis, R. W.** (1994). Transposition of the LINE-like retrotransposon TART to Drosophila chromosome termini. Proceedings of the National Academy of Sciences of the United States of America **91:** 12510-12514.

250. **Shen, M. R., Batzer, M. A., et Deininger, P. L.** (1991). Evolution of the master alu gene(s). Journal of Molecular Evolution **33:** 311-320.

251. **Shimamura, M., Yasue, H., Ohshima, K., Abe, H., Kato, H., Kishiro, T., Goto, M., Munechika, I., et Okada, N.** (1997). Molecular evidence from retroposons that whales form a clade within even-toed ungulates [see comments]. Nature **388:** 666-670.

252. **Silva, R. et Burch, J. B.** (1989). Evidence that chicken CR1 elements represent a novel family of retroposons. Molecular & Cellular Biology **9:** 3563-3566.

253. **Singer, D. S., Parent, L. J., et Ehrlich, R.** (1987). Identification and DNA sequence of an interspersed repetitive DNA element in the genome of the miniature swine. Nucleic Acids Research **15:** 2780.

254. **Singer, M. F.** (1982). SINEs and LINEs: highly repeated short and long interspersed sequences in mammalian genomes. Cell **28:** 433-4: 8.

255. **Sinnett, D.** (1991). Etude de la rétroposition des séquences Alu chez l'humain (Thèse de Doctorat). Département de Biochimie, Faculté de Médecine, Université de Montréal, Montréal.

256. **Sinnett, D., Richer, C., Deragon, J. M., et Labuda, D.** (1991). Alu RNA secondary structure consists of two independent 7 SL RNA-like folding units. Journal of Biological Chemistry **266:** 8675-8678.

257. **Sinnett, D., Richer, C., Deragon, J. M., et Labuda, D.** (1992). Alu RNA transcripts in human embryonal carcinoma cells. model of post-transcriptional selection of master sequences. Journal of Molecular Biology **226:** 689-706.

258. **Skowronski, J., Fanning, T. G., et Singer, M. F.** (1988). Unit-length line-1 transcripts in human teratocarcinoma cells. Molecular & Cellular Biology **8:** 1385-1397.

259. **Skowronski, J. et Singer, M. F.** (1985). Expression of a cytoplasmic LINE-1 transcript is regulated in a human teratocarcinoma cell line. Proceedings of the National Academy of Sciences of the United States of America **82:** 6050-6054.

260. **Slagel, V., Flemington, E., Traina-Dorge, V., Bradshaw, H., et Deininger, P.** (1987). Clustering and subfamily relationships of the Alu family in the human genome. Molecular Biology & Evolution **4:** 19-29.

261. **Slagel, V. K. et Deininger, P. L.** (1989). In vivo transcription of a cloned prosimian primate SINE sequence. Nucleic Acids Research **17:** 8669-8682.

262. **Smit, A. F.** (1996). The origin of interspersed repeats in the human genome. [review] [54 refs]. Current Opinion in Genetics & Development **6:** 743-748.

263. **Smit, A. F. et Riggs, A. D.** (1995). MIRs are classic, tRNA-derived SINEs that amplified before the mammalian radiation. Nucleic Acids Research **23:** 98-102.

264. **Smit, A. F. et Riggs, A. D.** (1996). Tiggers and DNA transposon fossils in the human genome. Proceedings of the National Academy of Sciences of the United States of America **93:** 1443-1448.

265. **Smit, A. F., Toth, G., Riggs, A. D., et Jurka, J.** (1995). Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. Journal of Molecular Biology **246:** 401-417.

266. **Smith, G. P.** (1976). Evolution of repeated DNA sequences by unequal crossover. Science **191:** 528-535.

267. **Southern, E. M.** (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis. Journal of Molecular Biology **98:** 503-517.

268. **Spence, S. E., Young, R. M., Garner, K. J., et Lingrel, J. B.** (1985). Localization and characterization of members of a family of repetitive sequences in the goat beta globin locus. Nucleic Acids Research **13:** 2171-2186.

269. **Spotila, L. D., Hirai, H., Rekosh, D. M., et Lo, Verde PT** (1989). A retroposon-like short repetitive DNA element in the genome of the human blood fluke, Schistosoma mansoni. Chromosoma **97:** 421-428.

270. **Springer, M. S.** (1995). Molecular Clock and the Incompleteness of the Fossil Records. Journal of Molecular Evolution **41:** 531-538.

271. **Sutcliffe, J. G., Milner, R. J., Bloom, F. E., et Lerner, R. A.** (1982). Common 82-nucleotide sequence unique to brain RNA. Proceedings of the National Academy of Sciences of the United States of America **79:** 4942-4946.

272. **Sutcliffe, J. G., Milner, R. J., Gottesfeld, J. M., et Lerner, R. A.** (1984). Identifier sequences are transcribed specifically in brain. Nature **308:** 237-241.

273. **Swergold, G. D.** (1990). Identification, characterization, and cell specificity of a human LINE-1 promoter. Molecular & Cellular Biology **10:** 6718-6729.

274. **Szemraj, J., Plucienniczak, G., Jaworski, J., et Plucienniczak, A.** (1995). Bovine alu-like sequences mediate transposition of a new site-specific retroelement. Gene **152:** 261-264.

275. **Tachida, H.** (1996). A population genetic study of the evolution of SINEs. II. Sequence evolution under the master copy model. Genetics **143:** 1033-1042.

276. **Takahashi, K., Terai, Y., Nishida, M., et Okada, N.** (1998). A novel family of short interspersed repetitive elements (SINEs) from Cichlids: The patterns of

insertion of SINEs at orthologous loci support the proposed monophyly of four major groups of Cichlids fishes in lake Tanganyika. Molecular Biology & Evolution **15:** 391-407.

277. **Takeuchi, Y. et Harada, F.** (1986). Cloning and characterization of rat 4.5S RNAI genes. Nucleic Acids Research **14:** 1643-1656.

278. **Tautz, D.** (1989). Hypervariability of simple sequences as a general source for polymorphic DNA markers. Nucleic Acids Research **17:** 6463-6471.

279. **Temin, H. M.** (1985). Reverse transcription in the eukaryotic genome: retroviruses, pararetroviruses, retrotransposons, and retrotranscripts. Molecular Biology & Evolution **2:** 455-468.

280. **Temin, H. M. et Mizutani, S.** (1970). RNA-dependent DNA polymerase in virions of rous sarcoma virus. Nature **226:** 1211-3: 2.

281. **Teng, S. C., Kim, B., et Gabriel, A.** (1996). Retrotransposon reverse-transcriptase-mediated repair of chromosomal breaks [see comments]. Nature **383:** 641-644.

282. **Terai, Y., Takahashi, K., et Okada, N.** (1998). SINE Cousins: The 3'-end tails of the two oldest and distantly related families of SINEs are descended from the 3' ends of LINEs with the same genealogical origin. Molecular Biology & Evolution **15(11):** 1471.

283. **Tourmente, S., Deragon, J. M., Lafleuriel, J., Tutois, S., Pelissier, T., Cuvillier, C., Espagnol, M. C., et Picard, G.** (1994). Characterization of minisatellites in Arabidopsis thaliana with sequence similarity to the human minisatellite core sequence. Nucleic Acids Research **22:** 3317-3321.

284. **Ullu, E., Murphy, S., et Melli, M.** (1982). Human 7SL RNA consists of a 140 nucleotide middle-repetitive sequence inserted in an alu sequence. Cell **29:** 195-202.

285. **Ullu, E. et Tschudi, C.** (1984). Alu sequences are processed 7SL RNA genes. Nature **312:** 171-172.

286. **Ullu, E. et Weiner, A. M.** (1985). Upstream sequences modulate the internal promoter of the human 7SL RNA gene. Nature **318:** 371-374.

287. **Van Arsdell, S. W., Denison, R. A., Bernstein, L. B., Weiner, A. M., Manser, T., et Gesteland, R. F.** (1981). Direct repeats flank three small nuclear RNA pseudogenes in the human genome. Cell **26:** Pt 1):11-7.

288. **Vandergon, T. L. et Reitman, M.** (1994). Evolution of chicken repeat 1 (CR1) elements: evidence for ancient subfamilies and multiple progenitors. Molecular Biology & Evolution **11:** 886-898.

289. **Varmus, H et Brown, P** (1989). Retroviruses, dans Mobile DNA Berg, D. E. and Howe, M. M. Washington D.C. pp 53-108.

290. **Varmus, H. E.** (1982). Form and function of retroviral proviruses. [Review] [85 refs]. Science **216:** 812-820.

291. **Vidaud, D., Vidaud, M., Bahnak, B. R., Siguret, V., Gispert, Sanchez S., Laurian, Y., Meyer, D., Goossens, M., et Lavergne, J. M.** (1993). Haemophilia B due to a de novo insertion of a human-specific Alu subfamily member within the coding region of the factor IX gene. European Journal of Human Genetics **1:** 30-36.

292. **Villanueva, M. S., Williams, S. P., Beard, C. B., Richards, F. F., et Aksoy, S.** (1991). A new member of a family of site-specific retrotransposons is present in the spliced leader RNA genes of Trypanosoma cruzi. Molecular & Cellular Biology **11:** 6139-6148.

293. **Watanabe, Y., Tsukada, T., Notake, M., Nakanishi, S., et Numa, S.** (1982). Structural analysis of repetitive DNA sequences in the bovine corticotropin-beta-lipotropin precursor gene region. Nucleic Acids Research **10:** 1459-1469.

294. **Weiner, A. M., Deininger, P. L., et Efstratiadis, A.** (1986). Nonviral retroposons: genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. [review] [179 refs]. Annual Review of Biochemistry **55:** 631-661.

295. **Weissenbach, J., Gyapay, G., Dib, C., Vignal, A., Morissette, J., Millasseau, P., Vaysseix, G., et Lathrop, M.** (1992). A second-generation linkage map of the human genome [see comments]. Nature **359:** 794-801.

296. **Wevrick, R., Earnshaw, W. C., Howard-Peebles, P. N., et Willard, H. F.** (1990). Partial deletion of alpha satellite DNA associated with reduced amounts of the centromere protein CENP-B in a mitotically stable human chromosome rearrangement. Molecular & Cellular Biology **10:** 6374-6380.

297. **Willard, C., Nguyen, H. T., et Schmid, C. W.** (1987). Existence of at least three distinct Alu subfamilies. Journal of Molecular Evolution **26:** 180-186.

298. **Willard, H. F.** (1991). Evolution of alpha satellite. [Review] [30 refs]. Current Opinion in Genetics & Development **1:** 509-514.

299. **Wilson, E. T., Condliffe, D. P., et Sprague, K. U.** (1988). Transcriptional properties of BmX, a moderately repetitive silkworm gene that is an RNA polymerase III template. Molecular & Cellular Biology **8:** 624-631.

300. **Winkfein, R. J., Moir, R. D., Krawetz, S. A., Blanco, J., States, J. C., et Dixon, G. H.** (1988). A new family of repetitive, retroposon-like sequences in the genome of the rainbow trout. European Journal of Biochemistry **176:** 255-264.

301. **Wong, Z., Wilson, V., Jeffreys, A. J., et Thein, S. L.** (1986). Cloning a selected fragment from a human DNA 'fingerprint': isolation of an extremely polymorphic minisatellite. Nucleic Acids Research **14:** 4605-4616.

302. **Woodcock, D. M., Lawler, C. B., Linsenmeyer, M. E., Doherty, J. P., et Warren, W. D.** (1997). Asymmetric methylation in the hypermethylated CpG promoter region of the human L1 retrotransposon. Journal of Biological Chemistry **272:** 7810-7816.

303. **Worth, L., Jr, Clark, S., Radman, M., Modrich, P.** (1994). Mismatch repair proteins MutS and MutL inhibit RecA-catalyzed strand transfer between diverged DNAs. Proceedings of the National Academy of Sciences of the United States of America **91:** 3238-3241.

304. **Wright, D. A., Ke, N., Smalle, J., Hauge, B. M., Goodman, H. M., et Voytas, D. F.** (1996). Multiple non-LTR retrotransposons in the genome of Arabidopsis thaliana. Genetics **142:** 569-578.

305. **Xiong, Y. et Eickbush, T. H.** (1988a). Functional expression of a sequence-specific endonuclease encoded by the retrotransposon R2Bm. Cell **55:** 235-246.

306. **Xiong, Y. et Eickbush, T. H.** (1988b). Similarity of reverse transcriptase-like sequences of viruses, transposable elements, and mitochondrial introns. Molecular Biology & Evolution **5:** 675-690.

307. **Xiong, Y. et Eickbush, T. H.** (1988c). The site-specific ribosomal DNA insertion element R1Bm belongs to a class of non-long-terminal-repeat retrotransposons. Molecular & Cellular Biology **8:** 114-123.

308. **Xiong, Y. et Eickbush, T. H.** (1990). Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO Journal **9:** 3353-3362.

309. **Xiong, Y. et Eickbush, T. H.** (1993). Dong, a non-long terminal repeat (non-LTR) retrotransposable element from Bombyx mori. Nucleic Acids Research **21:** 1318.

310. **Yoshioka, Y., Matsumoto, S., Kojima, S., Ohshima, K., Okada, N., et Machida, Y.** (1993). Molecular characterization of a short interspersed repetitive element from tobacco that exhibits sequence homology to specific tRNAs. Proceedings of the National Academy of Sciences of the United States of America **90:** 6562-6566.

311. **Zelnick, C. R., Burks, D. J., et Duncan, C. H.** (1987). A composite transposon 3' to the cow fetal globin gene binds a sequence specific factor. Nucleic Acids Research **15:** 10437-10453.

312. **Zietkiewicz, E. et Labuda, D.** (1996). Mosaic evolution of rodent B1 elements. Journal of Molecular Evolution **42:** 66-72.

313. **Zietkiewicz, E., Richer, C., Makalowski, W., Jurka, J., et Labuda, D.** (1994). A young Alu subfamily amplified independently in human and African great apes lineages. Nucleic Acids Research **22:** 5608-5612.

314. **Zietkiewicz, E., Richer, C., Sinnett, D., et Labuda, D.** (1998). Monophyletic origin of Alu elements in primates. Journal of Molecular Evolution **47:** 172-182.

315. **Zieve, G. W.** (1981). Two groups of small stable RNAs. [review] [18 refs]. Cell **25:** 296-297.