Comparaison de séquences d'éléments transposables et de gènes d'hôte chez cinq espèces : A. thaliana, C. elegans, D. melanogaster, H. sapiens et S. cerevisiae

Emmanuelle Lerat

N° d'ordre 182-2001 Année 2001

THÈSE

présentée devant

l'UNIVERSITE LYON 1 - CLAUDE BERNARD

pour l'obtention

du DIPLÔME DE DOCTORAT

(arrêté du 30 mars 1992)

présentée et soutenue publiquement le 26 octobre 2001

par

Emmanuelle LERAT

Comparaison de séquences d'éléments transposables et de gènes d'hôte chez cinq espèces : *A. thaliana, C. elegans, D. melanogaster, H. sapiens* et *S. cerevisiae*

JURY : M. Christian Biémont

M. Pierre Capy

M. John Brookfield (rapporteur)

M. Serge Hazout (rapporteur)

Mme Dominique Mouchiroud

M. Christophe Terzian

Je tiens à remercier chaleureusement Pierre Capy et Christian Biémont pour m'avoir accueillie et encadrée dans leurs laboratoires respectifs.

Un grand merci aux étudiants présents lors de mon DEA et de la première partie de ma thèse au laboratoire Populations, Génétiques et Évolution : Coco, Emmanuelle B, Christine, Stéphane, Vincent et Olivier (même si il n'était plus étudiant !) pour toutes les pauses, les bouffes et les soirées partagées ! Un nouveau merci à Olivier Langella pour m'avoir incitée à me mettre au C++, pour toutes les bibliothèques de programmes qu'il a crées et qu'il m'a permis d'utiliser, et pour avoir toujours très gentiment répondu à toutes mes questions en bioinformatique et en informatique.

Ayant eu la possibilité de « vivre » dans deux laboratoires, je tiens à remercier toutes les personnes du laboratoire Biométrie et Biologie Évolutive qui m'ont accueillies et qui m'ont permis de me sentir à l'aise : Cristina, Carène, Christiane, Nathalie, David, Catherine, Gabriel, encore Vincent, et tous les autres étudiants du « débarras » !

Un grand merci à toutes les personnes des deux laboratoires qui m'ont aidée et conseillée dans mon travail.

Un nouveau merci à Vincent pour m'avoir supportée ne serait-ce que pendant la 2ème année (ah la crise de la 2ème année !) et pour toutes les autres bonnes choses partagées !

Un merci particulier à mes parents pour m'avoir permis d'atteindre ce but et qui vont enfin ne plus avoir d'enfant à charge !

Je tiens à dédier cette thèse à Zouzou et Ea ...

1ère PARTIE : INTRODUCTION

1ère Partie : Introduction

Pendant la première moitié du XXème siècle, le génome était considéré comme une entité statique et peu malléable. Bien que l'idée du changement brusque des gènes par mutations soit acceptée, on estimait notamment que la localisation des gènes sur les chromosomes était fixe sur de longues périodes d'évolution, tout en acceptant la possibilité de changement par la recombinaison. Cette vision changea avec la découverte de BARBARA MCCLINTOCK dans les années cinquante, concernant des mutations instables chez le maïs (MCCLINTOCK 1950 ; MCCLINTOCK 1951). Elle eut l'idée de les attribuer à des éléments génétiques se déplaçant dans le génome qu'elle nomma « éléments de contrôle » à cause de leur capacité à « réguler l'expression des gènes ». Ils sont appelés aujourd'hui éléments transposables (ETs). Cette découverte ne commença à être véritablement admise que lorsque de tels éléments furent trouvés tout d'abord chez les bactéries (SHAPIRO 1969) puis chez la drosophile (KIDWELL *et al.* 1977) et lorsque l'on s'est aperçu qu'une grande partie de la plupart des génomes était constituée de séquences répétées non codantes. Finalement, la découverte de B. MCCLINTOCK fut officiellement reconnue en 1983 quand elle reçut le Prix Nobel de médecine. Ainsi, la vision statique du génome fut abandonnée pour une vision dynamique recelant un grand nombre de fluctuations. Dans cette nouvelle optique, les éléments transposables ont tout d'abord été considérés comme de simples parasites (ORGEL et CRICK 1980). En effet, la transposition leur permet une auto-réplication en utilisant les ressources moléculaires de l'hôte. Ainsi, sous cette hypothèse, seul l'avantage réplicatif des ETs pourrait expliquer leur augmentation et leur maintenance dans les génomes. Néanmoins, lors de leur mobilisation, les éléments transposables peuvent avoir des effets négatifs sur l'hôte en abaissant sa valeur sélective (NIKITIN et WOODRUFF 1995 ; WOODRUFF et NIKITIN 1995). Puis des études ont suggéré qu'ils pouvaient jouer un rôle fondamental dans l'évolution des génomes et augmenter le pouvoir adaptatif de certains individus (MCDONALD 1995). Ils ne doivent donc plus être considérés comme de simples « séquences égoïstes » définies comme de l'ADN pouvant augmenter en fréquence sans avoir d'effet sur les individus (DAWKINS 1990). Actuellement, de nombreuses études ont pour but d'étudier les ETs d'un point de vue structural et fonctionnel, ainsi que leurs effets sur les génomes hôtes car ils apparaissent de plus en plus comme des composés fondamentaux des génomes.

Les questions principales abordées dans ce travail ont pour but d'étudier comment les éléments transposables et leurs génomes hôtes interagissent en essayant de déterminer si les ETs ont des caractéristiques propres indépendantes de leur hôte.

1.1. Les éléments transposables

1.1.1. Définition et caractéristiques générales

Les éléments transposables sont classiquement définis comme des séquences moyennement répétées qui ont la capacité de se déplacer d'une position à une autre dans les génomes hôtes. Les éléments autonomes possèdent des gènes codant pour des protéines nécessaires à leur déplacement. On trouve aussi un certain nombre d'éléments non autonomes qui dépendent des éléments autonomes pour leur transposition. Jusqu'à présent, les ETs ont été trouvés dans tous les organismes vivants, procaryotes et eucaryotes, dans lesquels on les a cherchés. De plus, au sein d'un même génome, on peut trouver des éléments de type différent. Le nombre de copies des ETs peut varier de moins d'une dizaine à plusieurs centaines de milliers selon le type d'éléments et l'espèce hôte. Ainsi, ils peuvent former de 3 à 80 % des génomes. Par exemple, chez le maïs, ils représentent plus de 50% du génome total (SANMIGUEL *et al.* 1996) alors que chez la levure ils ne forment que 3% du génome (KIM *et al.* 1998). Lors de leurs déplacements, ils peuvent s'insérer dans les gènes ou dans des régions régulatrices et être responsables de mutations pouvant être délétères ou non. Ils sont aussi responsables d'importants remaniements chromosomiques comme des délétions, des inversions et des translocations (BERG et HOWE 1989 ; KIM *et al.* 1998 ; EVGEN'EV *et al.* 2000). Par exemple chez la drosophile, on estime qu'ils sont responsables de 80 % des mutations (MCDONALD 1995). Chez l'homme, les *Alus* sont responsables de 0,1 à 0,3% des maladies (ROY *et al.* 1999). Ils sont donc une source de variations génétiques importantes. Ainsi, les ETs ont un impact important sur leur génome hôte.

Malgré la caractéristique commune qui est de dupliquer leur site cible lors de leur insertion, les ETs ont des structures pouvant être très différentes. On peut ainsi les caractériser selon leur structure et selon l'intermédiaire qu'ils utilisent pour se déplacer (FINNEGAN 1989 ; FINNEGAN 1992). La Figure 1 représente une version simplifiée de la classification proposée par CAPY *et al.* (1997a). Cette classification implique des relations entre les ETs, ainsi, l'évolution des ETs pose de nombreuses questions à l'heure actuelle. Beaucoup de relations au niveau des séquences protéiques peuvent être établies entre ETs de classes différentes. On retrouve des signatures protéiques communes entre éléments de classe I et éléments de classe II (CAPY *et al.* 1997b) ou bien entre rétrotransposons et rétrovirus (XIONG et EICKBUSH 1990 ; LERAT et CAPY 1999). Une hypothèse avancée concerne la possibilité d'acquisition modulaire de différents domaines par des éléments (LERAT *et al.* 1999). Ainsi, les ETs pourraient évoluer les uns à partir des autres sans qu'une orientation évolutive puisse être privilégiée.

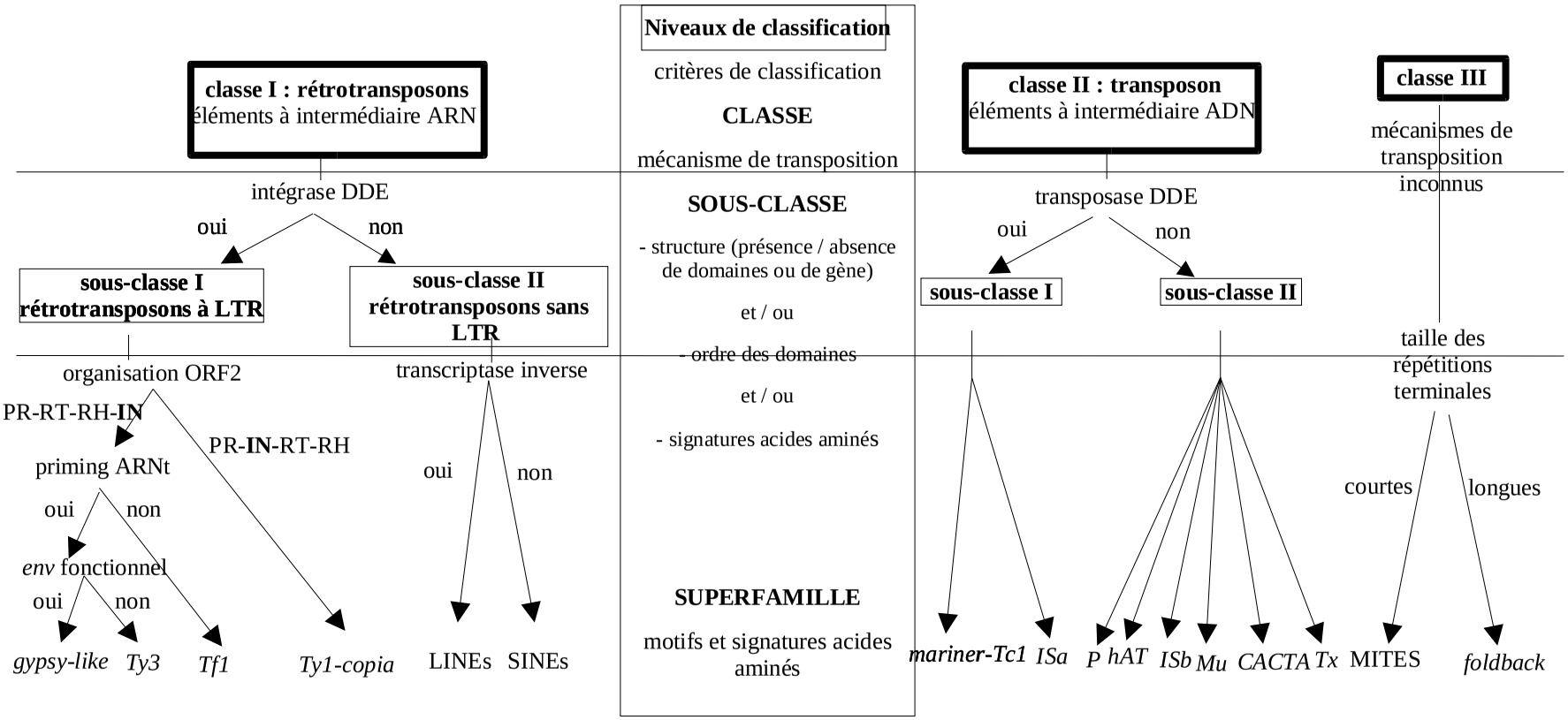


Figure 1 : classification simplifiée des ETs
LTR : Long Terminal Repeat ; PR : protéase ; RT : transcriptase inverse ; RH : Rnase H ; IN : intégrase ; DDE : signature protéique.

1.1.2. Les différentes classes d'ETs

1.1.2.1. Les éléments de classe I : les rétrotransposons (Figure 2)

Les rétrotransposons sont des éléments se déplaçant par un intermédiaire ARN. Ils sont subdivisés en deux sous-classes, suivant la présence ou l'absence de séquences « Long Terminal Repeat » (LTR). La sous-classe I comporte les rétrotransposons à LTR, qui sont des éléments de 5 à 9 kb. Les LTR sont des séquences en orientation directe de quelques centaines de bases présentes aux extrémités de l'élément. Ces séquences contiennent les régions promotrices et régulatrices des séquences codantes de l'élément et comportent trois domaines fonctionnels : U3, R et U5 (voir Figure 2). Lors de leur transposition, les rétrotransposons à LTR sont tout d'abord rétrotranscrits avant d'être insérés dans le génome. De par leur organisation, ces éléments sont proches des rétrovirus. Ainsi, on retrouve principalement deux ORFs (Open Reading Frame) présents chez les rétrovirus : les gènes *gag* et *pol*. Chez les rétrovirus, le gène *gag* code pour une polyprotéine qui donne trois protéines matures : la protéine de la matrice, la protéine de la capside et la protéine de la nucléocapside qui composent la capside virale (Varmus et Brown 1989). La nucléocapside est une protéine de liaison d'acide nucléique possédant des domaines en doigt à zinc, qui permet de lier l'ARN viral à la capside. Chez les rétrotransposons, l'utilité du gène *gag* n'est pas encore élucidée. Le gène *pol* code pour une polyprotéine qui, après protéolyse, forme les protéines requises pour la transposition, notamment une protéase, une transcriptase inverse nécessaire à la rétrotranscription, une Rnase H et une intégrase, qui permet l'insertion de l'élément dans le génome. Les rétrotransposons à LTR sont subdivisés en deux groupes principaux suivant l'ordre des domaines protéiques du gène *pol*. Chez les éléments de type *Ty1/copia*, l'intégrase se trouve du côté 5' de la transcriptase inverse alors que chez les éléments de type *gypsy/Ty3*, elle se place du côté 3' (voir Figure 2). Chez certains rétrotransposons à LTR du groupe *gypsy/Ty3*, on trouve aussi une troisième ORF, le gène *env*. Les rétrovirus possèdent ce gène qui code pour une polyprotéine formant après protéolyse deux protéines : une transmembranaire (TM) et une extracellulaire (SU). La protéine SU contient les sites de liaison à des récepteurs membranaires des cellules hôtes. Un complexe est formé entre ces deux protéines par des ponts disulfures. Ces protéines permettent l'adsorption et la pénétration dans la cellule cible du rétrovirus (Varmus et Brown 1989). Ce gène est parfois fonctionnel chez certains rétrotransposons comme par exemple les éléments *gypsy* (Kim *et al.* 1994 ; Teysset *et al.* 1998) et *ZAM* (Leblanc *et al.* 2000) chez *Drosophila melanogaster*. Des motifs protéiques communs sont retrouvés dans les intégrases et dans les transcriptases inverses entre

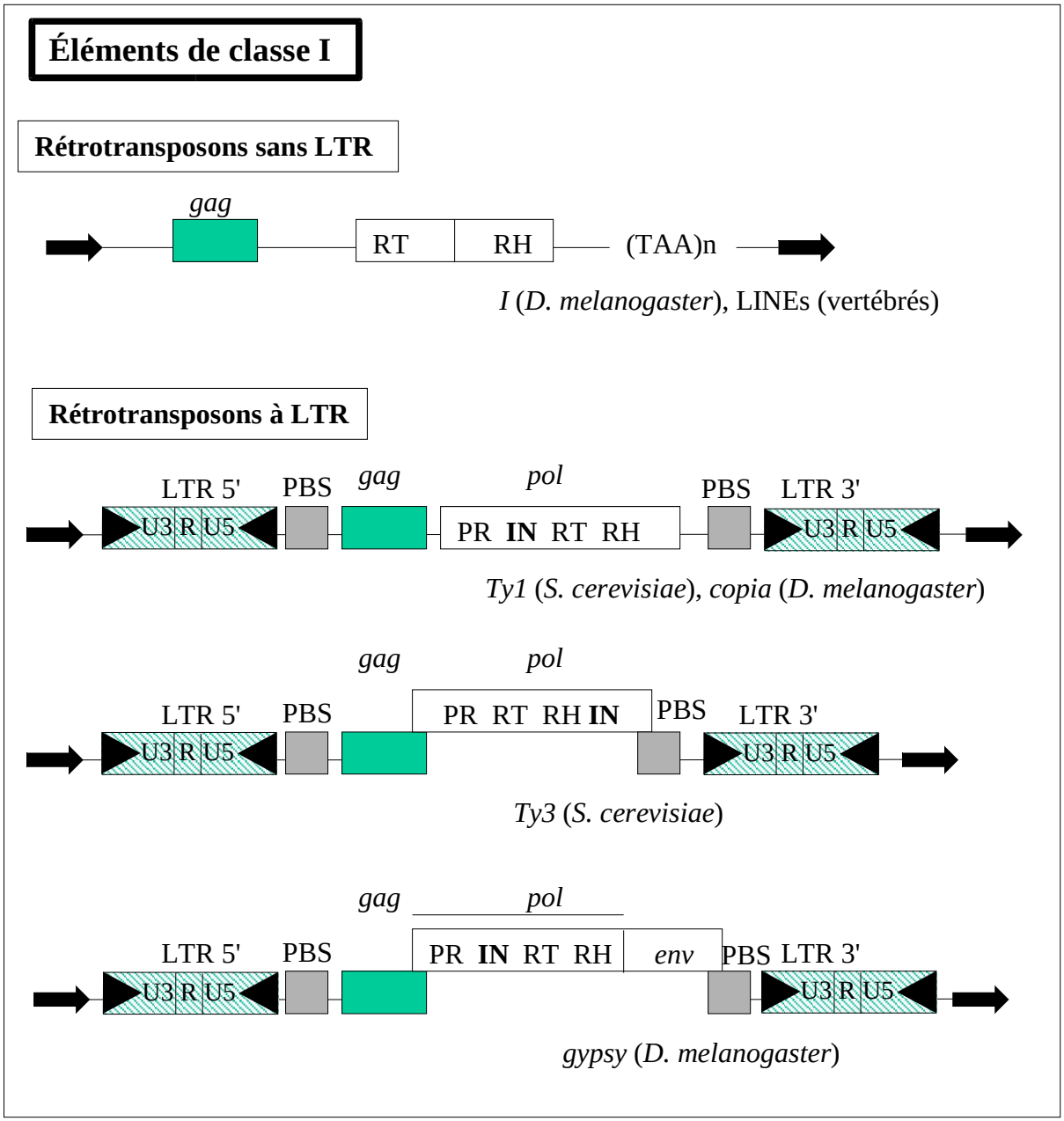


Figure 2 : structure de différents types d'éléments de classe I (voir figure 1 pour les abréviations)

rétrotransposons et rétrovirus (XIONG et EICKBUSH 1990 ; MCCLURE 1991 ; CAPY *et al.* 1996) ainsi qu'au niveau des séquences protéiques du gène *env* (LERAT et CAPY 1999). Ces différentes études suggèrent un lien étroit entre rétrovirus et rétrotransposons à LTR. L'hypothèse la plus couramment avancée est l'apparition de rétrovirus à partir de rétrotransposons à LTR ayant acquis un gène d'enveloppe (TEMIN 1980 ; FLAVELL 1981). Cependant, on ne peut exclure le phénomène inverse : un rétrovirus perdant son gène d'enveloppe devient un rétrotransposon. Actuellement, les rétrotransposons à LTR avec un gène *env* fonctionnel ou non sont rangés parmi une nouvelle catégorie : les errantivirus, qui font partie du groupe des *Metaviridae*, les éléments de type *Ty1/copia* faisant partie du groupe des *Pseudoviridae* (BOEKE *et al*. 1998).

La sous-classe II est formée par les rétrotransposons sans LTR (ou rétroposons) qui sont caractérisés par une queue poly A ou une région riche en A à leur extrémité 3', précédée par un signal de polyadénylation. Ils possèdent à chaque extrémité de courtes répétitions directes (voir Figure 2). La première sous-famille comporte des éléments possédant généralement deux ORF, correspondant aux gènes *gag* et *pol*. Il s'agit des éléments LINEs (Long Interspersed Elements). Une différence majeure avec les rétrotransposons à LTR est que le gène *pol* ne possède pas d'intégrase. Bien qu'ils soient rétrotranscrits, les éléments de la sous-classe II ont un système d'intégration complètement différent des rétrotransposons à LTR. En effet, la transcriptase inverse reconnaît le transcrit et initie la rétrotranscription, en même temps que l'intégration au niveau d'une coupure de l'ADN génomique effectuée par une endonucléase (LUAN *et al*. 1993). Les éléments LINE font généralement de 5 à 8 kb de long, bien que dans la plupart des espèces, un grand nombre de ces éléments soit tronqué (MARTIN 1991). Ils possèdent des promoteurs internes qui leur permettent d'être transcrits par une ARN polymérase II (MCLEAN *et al.* 1993 ; MINCHIOTTI et DINOCERA 1991).

Parmi les rétrotransposons sans LTR, on trouve des éléments ne possédant pas de cadre de lecture : il s'agit des éléments SINEs (Short Interspersed Elements). Ils font généralement environ 500 bp de longueur. Ils dérivent presque tous des ARN de transfert (OKADA 1991) excepté l'élément *Alu* des primates et la famille *B1* des rongeurs qui dérivent de l'ARN 7SL (ULLU et TSCHUDI 1984). Les SINEs dérivés d'ARNt ont une structure composite comportant une région homologue à un ARNt, une région « core » conservée de fonction inconnue, une région non homologue à un ARNt et une zone de taille variable AT riche en 3'. La zone homologue à un ARNt contient un promoteur interne d'ARN polymérase III qui est impliquée dans la transcription de l'élément. Ces éléments non autonomes utilisent vraisemblablement la transcriptase inverse des LINEs pour se déplacer (OKADA *et al.* 1997).

1.1.2.2. Les éléments de classe II : les transposons (Figure 3)

La classe II regroupe les éléments se déplaçant par un intermédiaire ADN. Ils sont directement excisés puis insérés à un autre endroit du génome grâce à une transposase. Ces éléments sont plus petits que les rétrotransposons. Ils possèdent à chacune de leurs extrémités de courtes séquences répétées inversées de quelques dizaines à quelques centaines de paires de base (ITR : Inverted Tandem Repeat). On trouve généralement dans leur séquence un seul cadre de lecture qui code la transposase. Celle-ci se fixe au niveau de séquences spécifiques près des ITR. Cette classe comporte un ensemble de superfamilles hétérogènes dont les éléments les plus caractéristiques sont l'élément *Ac* du maïs et les éléments *P* et *mariner* de la drosophile. On peut distinguer deux sous-classes suivant la présence ou l'absence dans la séquence protéique de la transposase de la signature $DD_{35}E$ (D pour aspartate et E pour glutamate). Cette signature est retrouvée dans les séquences protéiques des intégrases des rétrotransposons à LTR suggérant une origine commune, probablement procaryote, à ces deux gènes (Capy *et al.* 1996). On ne trouve que cette classe d'éléments chez les bactéries. Les bactéries peuvent comporter deux types de transposons : les *IS* (Insertion Sequence) (Mahillon et Chandler 1998) et les *Tn*, qui sont des éléments composites pouvant comporter un gène de résistance à un antibiotique ou à des métaux lourds.

1.1.2.3. Les éléments de classe III

Cette classe regroupe des éléments dont on connaît mal le mécanisme de transposition. C'est le cas par exemple pour les éléments *foldback* qui possèdent de grandes répétitions terminales inversées. Ils ont tout d'abord été détectés chez la drosophile (Truett *et al.* 1981) mais ils sont aussi présents chez d'autres organismes comme le nématode (élément *Tc4*) (Yuan *et al.* 1991) et l'oursin de mer (éléments *TU*) (Liebermann *et al.* 1983).

On place aussi dans cette classe la superfamille des MITEs (Miniature Inverted-repeat Transposable Elements) qui présente des éléments mal caractérisés et de très petite taille (100 à 500 bp), principalement trouvés chez les plantes comme l'arabette, le poivre vert ou le maïs et les champignons. Ils ont cependant été récemment détectés chez certains vertébrés (l'homme, le poisson zèbre et le xénope) et invertébrés (le nématode, le moustique *C. pipiens*, le moustique *A. aegypti* et la coccinelle *T. molitor*). Certaines caractéristiques comme des ITR et des duplications générées lors de l'insertion suggèrent qu'ils peuvent dériver d'éléments de classe II comme cela a été mis en évidence pour un élément d'*Arabidopsis*, il pourrait donc s'agir d'éléments de classe II dégénérés (Feschotte et Mouchès 2000).

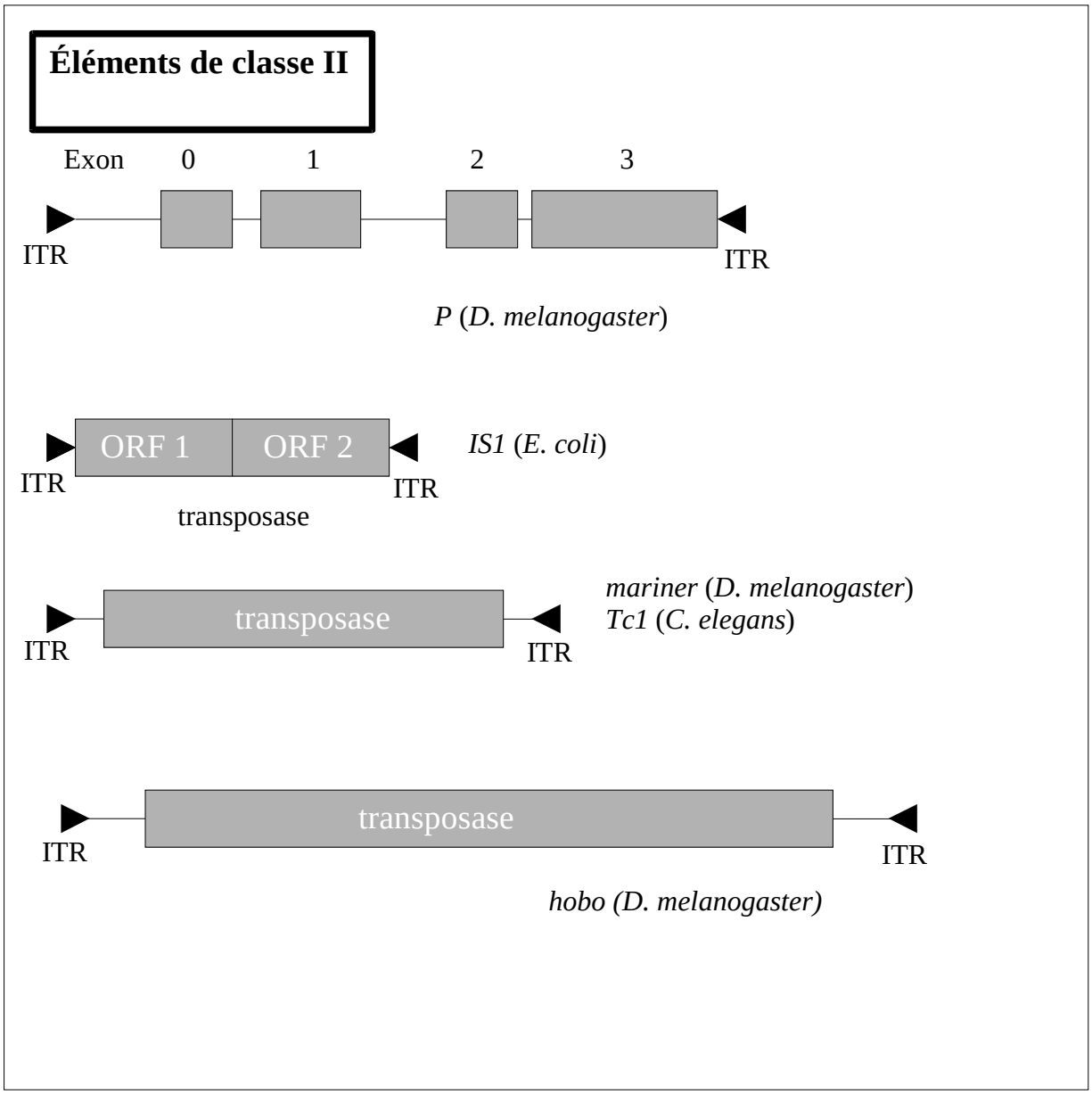


Figure 3 : structure de différents types d'éléments de classe II

1.1.3. Interactions avec le génome hôte

1.1.3.1. Influence des ETs sur le génome hôte

D'une manière générale, les ETs sont responsables d'un grand nombre de mutations et de réarrangements chromosomiques. Ces réarrangements sont probablement le produit de recombinaisons ectopiques* entre ETs ou de coupure près des sites d'insertion de deux éléments dans la même orientation. Par exemple, chez *Drosophila virilis*, l'élément *Pénélope*, un rétrotransposon sans LTR, et *Ulysses*, un rétrotransposon à LTR, sont responsables de multiples réarrangements chromosomiques pouvant avoir joué un rôle dans l'évolution de cette espèce (Evgen'ev *et al.* 2000). On trouve de nombreuses mutations associées à des *IS* chez *Escherichia coli* (Schneider *et al.* 2000). Dans des populations naturelles de *Drosophila melanogaster*, le transposon *hobo* semble être à l'origine d'inversions endémiques (Lyttle et Haymer 1992 ; Ladeveze *et al.* 2000). De même, l'étude du génome de la levure a montré que les rétrotransposons *Ty* sont trouvés près de séquences qui ont été dupliquées sur plusieurs chromosomes, indiquant qu'ils ont pu avoir un rôle dans les réarrangements du génome (Kim *et al.* 1998). Enfin, des séquences répétées sont impliquées dans des réarrangements au niveau de certains gènes de mammifères (Meuth 1989).

En plus des remaniements structuraux, les ETs peuvent provoquer des changements au niveau métabolique. En effet, possédant des séquences régulatrices, les ETs peuvent altérer l'expression des gènes proches de leurs sites d'insertion. Dans la plupart des cas, l'élément bloque l'expression mais parfois, l'expression peut être augmentée ou diminuée (Geyer *et al.* 1988). Certaines mutations d'insertions provoquées par les ETs semblent être favorisées par la sélection. Des études ont notamment montré que des différences évolutives au niveau de patron d'expression sont le résultat d'insertion d'ETs (McDonald 1995 ; Britten 1996). Il semble aussi que les rétrotransposons à LTR pourraient jouer un rôle dans l'évolution des enhancers* eucaryotes. En effet, les LTR de rétroéléments montrent des duplications régionales caractéristiques des enhancers* eucaryotes. Ainsi, lors de son insertion dans une région promotrice, un rétrotransposon à LTR peut provoquer une augmentation de la force de l'enhancer* existant, la sélection pouvant agir sur l'enhancer* résultant (McDonald *et al.* 1997). Il y a de nombreuses évidences qui montrent que certains introns proviennent de l'insertion d'ETs dans des séquences codantes (voir pour revue Purugganan 1993). Ainsi, des ETs peuvent altérer l'épissage des ARN pré-messagers de deux manières : (i) les ETs sont insérés dans des exons et sont capables de fonctionner comme de nouveaux introns, c'est par exemple le cas de certains éléments *P* ou d'éléments *412* chez *D.*

melanogaster (ii) les ETs sont insérés à la fois dans les exons et les introns, et peuvent modifier les patrons d'épissage constitutifs, c'est le cas pour des *Alus* chez l'homme. Ainsi, les ETs se comportant comme des introns peuvent contribuer à la diversification des gènes et des produits de gènes.

Dans la nature, les organismes peuvent subir des facteurs environnementaux auxquels ils doivent s'adapter. Ainsi, les stress augmentent la variabilité génétique des organismes (Imasheva *et al.* 1998). Dans certains cas, le stress induit des mécanismes de mutation qui vont permettre la mise en place de réponses à des changements environnementaux. Certains de ces mécanismes font appel à des éléments transposables (Capy *et al.* 2000). Ainsi, chez les bactéries, les ETs sont susceptibles de permettre des adaptations au milieu car certains éléments composites, les éléments *Tn*, possèdent des gènes de résistance à des antibiotiques (Berg et Howe 1989). De même chez le tabac, *Nicotiana tabacum*, il a été démontré que l'expression du rétrotransposon *Tnt1* est activée sous l'effet de certains stress qui déclenchent parallèlement le système de défense de la plante (Grandbastien *et al.* 1997). Enfin, des éléments de type rétrovirus semblent capables d'apporter une résistance à l'infection de virus (Gardner *et al.* 1991). Cela a été démontré chez la souris où le gène *Fv1* confère une résistance à l'infection de certaines souches du virus de la leucémie murine. Le produit de ce gène présente des homologies avec la protéine *gag*. Il s'est avéré que ce gène provient d'un rétrovirus endogène ou exogène (Benit *et al.* 1997). Les ETs pourraient donc être impliqués dans l'évolution des génomes aussi bien au niveau du fonctionnement que de la structure. Cette idée semble d'ailleurs de mieux en mieux acceptée et fait même son apparition dans le grand public, puisqu'elle est reprise par des auteurs de science-fiction (Bear 1999).

1.1.3.2. Influence du génome hôte sur les ETs

Nous venons de voir que les ETs ont un impact important sur l'organisation de leurs génomes hôtes. Afin de se protéger de ces effets, certains organismes ont développé des stratégies leur permettant de réguler leur activité. Ainsi, chez les plantes ou les mammifères, bien que les mécanismes soient différents suivant les espèces, la méthylation* des ETs servirait à prévenir la transposition (Yoder *et al.* 1997). Par exemple, des hybrides entre différentes espèces de wallaby présentent une hypométhylation dans l'ensemble de leurs génomes et on observe une amplification massive de rétroéléments (O'Neill *et al.* 1998). Cependant, le grand nombre d'ETs que l'on peut trouver dans les génomes des plantes tendrait à prouver que ce mécanisme est peu efficace. En fait, d'après Martienssen (1998), la méthylation* des ETs servirait à les « cacher » en favorisant des interactions avec des facteurs chromatiniens qui empêcheraient la transcription, plutôt que réguler la transposition ou la transcription. Ceci permettrait aux ETs de s'accumuler sans avoir d'effets

négatifs sur le génome. Un mécanisme faisant intervenir la méthylation* est aussi trouvé chez le champignon *Neurospora crassa* (RAND 1992). Ce phénomène, le RIP (Repeat-Induced Point mutation) est un processus de mutagenèse qui se produit pendant la phase sexuelle du champignon. Par ce mécanisme, les séquences répétées sont détectées et il y a mutation de G:C vers A:T, les résidus cytosines restantes étant toujours méthylées. Ainsi, ce système rend le génome de ce champignon particulièrement inhospitalier pour les ETs. Un phénomène analogue est rencontré chez le cricket, le RAP (Repeat-Associated Point Mutation), mais dans ce cas, il n'y a pas intervention de la méthylation* (RAND 1992). Chez *Drosophila melanogaster*, un certain nombre de gènes d'hôte paraisse réguler l'activité de certains ETs (LOZOVSKAYA *et al.* 1995). Par exemple, le rétrotransposon *gypsy* est mobilisé lors de la mutation d'un gène appelé *flamenco* (PRUD'HOMME *et al.* 1995).

Certains ETs sont domestiqués par leur génome hôte. Chez *Drosophila melanogaster* et chez *Bombyx mori*, deux éléments de type LINEs, *R1* et *R2*, sont insérés au niveau des gènes de l'ARN ribosomal 28S. Ces éléments servent à l'amplification de ces gènes (JACKUBCZAK *et al.* 1990 ; TAKAHASHI *et al.* 1997). Toujours chez la drosophile, des éléments LINE présents en multicopie au niveau des télomères des chromosomes ont été identifiés : les éléments *HeT-A* et *TART* (BIESSMANN *et al.* 1994 ; SHEEN et LEVIS 1994). Ils se transposent exclusivement au niveau des extrémités des chromosomes. La drosophile ne possède pas de système de réparation faisant appel à une télomérase, qui a pour fonction chez d'autres organismes, le maintien de l'intégrité des chromosomes. En effet, les télomères sont essentiels aux cellules car en protégeant les extrémités des chromosomes, ils empêchent leur raccourcissement et donc à terme l'apoptose* prématurée de la cellule. La télomérase est une ribonucléoprotéine qui utilise comme substrat son propre ARN et qui maintient les télomères en ajoutant des petites séquences répétées (GREIDER 1990 ; BLACKBURN 1991). Il semblerait que les éléments télomériques de la drosophile permettent en se transposant de régénérer les extrémités des chromosomes (BIESSMANN *et al.* 1992 ; PARDUE 1995). Les répétitions générées par ces éléments apparaissent comme plus complexes que celles générées par les télomérases (PARDUE *et al.* 1997). Chez les vertébrés, les gènes RAG1 et RAG2 sont essentiels pour le développement des lymphocytes en réalisant une réaction de recombinaison du système V(D)J* des immunoglobulines. Cette réaction est nécessaire à la production d'immunoglobulines possédant une grande variabilité. Or, il y a peu de temps, AGRAWAL, EASTMAN et SCHATZ (1998) ont montré que cette recombinaison est un mécanisme de transposition, similaire à celui provoqué par d'autres transposases *Tc1*-like. Ceci semble indiquer que ces deux gènes devaient à l'origine faire partie d'un ET.

Ces différents exemples montrent bien que les ETs peuvent jouer des rôles très importants dans l'évolution des organismes hôtes et promouvoir une diversité nécessaire à l'adaptation aux

changements de l'environnement.

1.2. Présentation des cinq espèces utilisées

Afin d'étudier les relations entre séquences d'éléments transposables et génomes hôtes, nous avons étudié cinq espèces eucaryotes différentes chez lesquelles un nombre important d'ETs sont caractérisés : une plante, *Arabidopsis thaliana*, un insecte, *Drosophila melanogaster*, un champignon, *Saccharomyces cerevisiae*, un nématode, *Caenorhabditis elegans* et un vertébré, *Homo sapiens*. La Figure 4 montre la position phylogénétique de ces différentes espèces les unes par rapport aux autres. Nous avons délibérément choisi de ne pas utiliser de procaryotes. En effet, une analyse effectuée en utilisant *E. coli* (résultats non présentés) a montré que cet organisme se différencie complètement des autres espèces ce qui entraîne une perte d'informations concernant les analyses multivariées. Dans les paragraphes suivants, je vais faire un rappel sur ce que l'on sait des génomes utilisés et des ETs qu'ils contiennent.

1.2.1. *Arabidopsis thaliana*

L'arabette est une plante à fleurs de la famille des brassicacées très largement distribuée en Europe, en Asie et en Amérique du Nord. Elle est étudiée comme organisme modèle depuis les années 80 tout d'abord en tant que modèle en génétique puis en physiologie, biochimie et développement. Le choix d'utiliser *A. thaliana* comme modèle vient notamment du fait qu'elle possède un très petit génome pour une plante (120 Mb reparties sur 5 chromosomes) avec peu de séquences répétées (10% du génome), un cycle de vie très court de 6 semaines et qu'elle est très prolifique (Meinke *et al.* 1998). C'est donc assez logiquement qu'*Arabidopsis* fait partie du programme de séquençage des génomes complets. Les premiers chromosomes à dévoiler leurs séquences furent les chromosomes 2 (Lin *et al.* 1999) et 4 (Mayer *et al.* 1999), puis un an après, les séquences des chromosomes 1 (Theologis *et al.* 2000), 3 (Salanoubat *et al.* 2000) et 5 (Tabata *et al.* 2000). Au cours de l'évolution, 70% du génome de l'arabette a été dupliqué. Ainsi, ses 26 000 gènes sont pour la plupart en double exemplaire (Walbot 2000). Globalement, *Arabidopsis* a un génome plutôt riche en bases A+T (environ 65% d'AT toutes régions confondues et 56% d'AT pour les parties codantes des gènes), (The Arabidopsis Genome Initiative 2000).

L'arabette possède un pourcentage d'ETs relativement peu élevé pour une plante (14%). Pour comparaison, le génome du maïs contient 50% d'ETs (SanMiguel *et al.* 1996). L'étude du génome complet a permis de montrer que les ETs sont localisés principalement dans les régions pauvres en

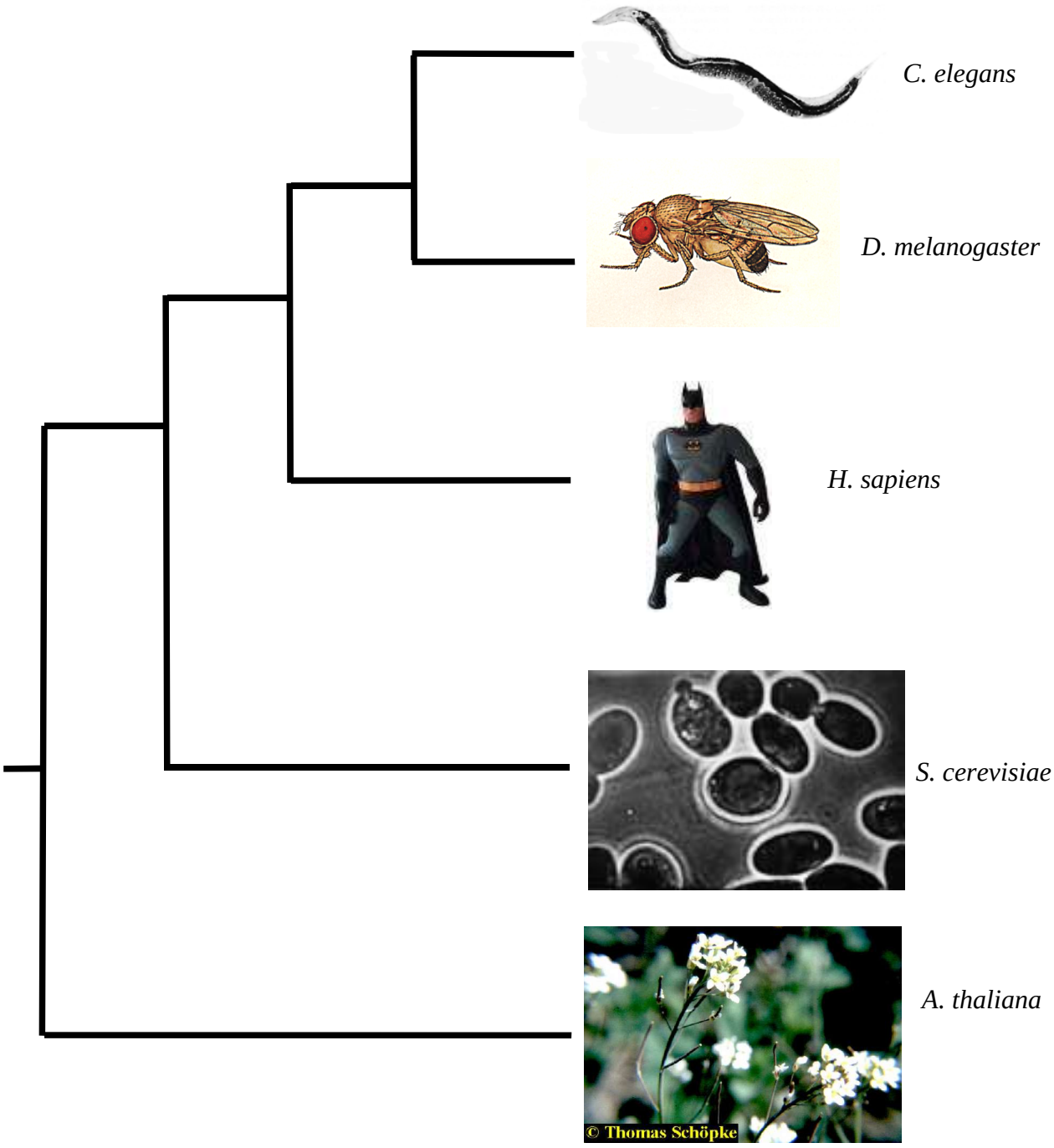


Figure 4 : position relative des cinq espèces (d'après Lecointre et Le Guyader 2001).

gènes, proches des centromères (The Arabidopsis Genome Initiative 2000). Par des études préliminaires, une large gamme d'ETs de familles différentes a été détectée : des rétrotransposons à LTR de type *copia* et *gypsy*, des rétrotransposons sans LTR (LINEs et SINEs), des transposons de la famille *hAT*, de la famille CACTA, des éléments MITE, des MULEs (*Mutator*-Like Éléments), ainsi que des éléments appartenant à une nouvelle famille de transposons : les éléments *Basho* (Le *et al.* 2000). D'après l'analyse du génome, les éléments de classe I sont moins abondants que chez les plantes de plus grands génomes (SanMiguel *et al.* 1996) et sont principalement localisés au niveau du centromère. Par contre, les éléments *Basho* et les transposons sont les plus représentés et sont localisés à la périphérie des domaines péricentromériques.

1.2.2. *Caenorhabditis elegans*

Ce nématode est le premier organisme multicellulaire à avoir été séquencé en 1998 (The *C. elegans* Sequencing Consortium 1998). Ce ver transparent de quelques millimètres a ceci de particulier qu'il est constitué d'un nombre connu de cellules, très exactement 959, dont on connaît parfaitement le devenir (Sulston *et al.* 1983). Il a donc été possible de déterminer la filiation exacte de chacune des cellules de cet organisme de la fécondation jusqu'au stade adulte. *C. elegans* est un organisme modèle pour le développement et pour le système nerveux depuis un grand nombre d'années (Brenner 1974). On connaît d'ailleurs exactement l'ensemble des jonctions synaptiques établies par chacun des 302 neurones, ce qui est loin d'être le cas pour tous les autres organismes. Son génome a une taille de 97 Mb reparties sur 6 chromosomes (5 autosomes et un chromosome sexuel) et comporte environ 19 000 gènes. Tout comme *Arabidopsis,* il s'agit d'un génome riche en AT (64% d'AT au total) qui comporte environ 17% de séquences répétées.

Chez le nématode, des rétrotransposons à LTR ont été récemment décrits grâce à l'analyse du génome complet (Bowen et McDonald 1999). Jusqu'à présent, on estimait que ce génome contenait principalement des transposons de type *Tc1-mariner,* des rétrotransposons sans LTR et des MITEs (Surzycki et Belknap 2000). L'élément *Tc1* est le mieux connu des ETs du nématode et a été le premier élément caractérisé dans cette espèce (Emmons *et al.* 1983). Cinq familles distinctes de transposons sont connus *Tc1, Tc2, Tc3, Tc5* et *Tc7*. On trouve aussi des familles de classe III représentées par les éléments *Tc4* et *Tc6*. Des éléments de classe I ont été identifiés plus récemment. Ainsi, en 1995, un élément de type *gypsy/Ty3* a été identifié dans le génome séquencé, l'élément *Cer1* (Britten 1995). Des rétrotransposons sans LTR ont aussi été identifiés soit *in vivo,* soit par homologie de séquence (Marin *et al.* 1998 ; Youngman *et al.* 1996). L'étude de leur répartition dans le génome complet montre que les transposons sont préférentiellement localisés dans les régions

possédant un fort taux de recombinaison alors que cette relation n'apparaît pas chez les rétrotransposons (Duret *et al.* 2000).

1.2.3. *Drosophila melanogaster*

La drosophile, plus connue sous le nom de mouche du vinaigre, est utilisée pour des études génétiques comme organisme modèle depuis le début du XX^{ème} siècle . Ces études ont permis la mise en place des premières bases pour la compréhension de la génétique des eucaryotes (Morgan *et al.* 1915). Notamment la théorie chromosomique de l'hérédité par T. H. Morgan, puis la découverte des effets mutagènes des rayons X sur les chromosomes par Muller (1927), la possibilité d'établir des localisations physiques de gènes grâce aux chromosomes polytènes et la découverte de très importantes voies de signalisation par l'étude du développement embryonnaire (voir pour revue Rubin et Lewis 2000). Finalement, en mars 2000, la première version du génome séquencé est publiée (Adams *et al.* 2000). Le génome de la drosophile comporte 180 Mb reparties en 5 chromosomes (3 autosomes et 2 chromosomes sexuels). La présence d'hétérochromatine* a empêché le séquençage de la totalité du génome. Ainsi, nous n'avons accès qu'aux 120 Mb que constitue l'euchromatine*. Celle-ci est principalement présente sur les deux grands chromosomes autosomaux 2 et 3, et sur le chromosome X, ainsi que sur une portion du petit chromosome 4, le chromosome Y étant presque entièrement constitué d'hétérochromatine*. La séquence génomique est formée d'un minimum de 14 332 gènes, ce qui est inférieur à ce que l'on trouve chez le nématode (19 000 gènes). Le génome de la drosophile possède une composition en bases globale de 57% de AT (Ashburner 1989), les régions codantes possédant un contenu en AT plus faible que les régions non codantes (Carulli *et al.* 1993).

La drosophile contient beaucoup d'ETs de types différents qui couvrent quasiment la gamme complète des différentes familles. De très nombreux éléments font l'objet d'études intensives. L'élément *I*, un rétrotransposon sans LTR et les transposons *hobo* et *P* sont responsables d'un phénomène connu sous le nom de dysgénésie* des hybrides. Lorsque l'on croise des mâles possédant l'un de ces éléments avec des femelles vides de l'élément, on peut observer dans la descendance femelle une très forte activité de l'élément (Sezutsu *et al.* 1995). Le fait que l'on ne trouve ces éléments que dans des souches récentes de laboratoire semble indiquer qu'il s'agit d'une acquisition récente. C'est donc un bon modèle pour étudier la dynamique de prolifération des ETs dans un génome. De plus, c'est chez cette espèce que l'on a pu caractériser un rétrotransposon à LTR possédant des propriétés infectieuses : l'élément *gypsy* (Kim *et al.* 1994).

1.2.4. *Homo sapiens*

Le génome de l'homme présente une caractéristique que l'on retrouve chez tous les vertébrés homéothermes (voir pour revue Bernardi 2000), ainsi que chez certains vertébrés poïkilothermes (Hugues *et al.* 1999) : il est constitué d'isochores. Il s'agit de longs segments d'ADN qui ont une composition en bases homogène mais différente entre régions. On en distingue différentes familles selon le pourcentage en base G+C (GC). Ainsi, le génome de l'homme est composé d'une mosaïque d'isochores. La densité en gènes est plus élevée principalement dans les isochores très riches en GC. Son génome est formé de 3 286 Mb répartie sur 24 chromosomes différents (22 autosomes et 2 chromosomes sexuels).

Le séquençage du génome humain s'est inséré dans le cadre de la recherche des maladies génétiques. Il est le premier génome vertébré à avoir été séquencé. L'achèvement de ce séquençage a récemment eu lieu à la fois pour le consortium public international (International Human Genome Sequencing Consortium 2001) et pour le groupe privé Celera (Venter *et al.* 2001). Un an auparavant, les séquences des deux chromosomes les plus petits, les chromosomes 21 et 22, avaient été publiés (Dunham *et al.* 2000 ; Hattori *et al.* 2000). Ces diverses études ont confirmés que le génome humain est très riche en séquences répétées (45% du génome) mais qu'il contient bien moins de gènes qu'on le prédisait (entre 30 000 et 40 000) au grand désarroi des adeptes de l'anthropocentrisme, ce qui n'accorde à l'homme, au minimum, que le double des gènes de la drosophile. De nouvelles hypothèses concernant la possibilité pour les gènes humains par épissage alternatif de produire un plus grand nombre de protéines sont avancées (International Human Genome Sequencing Consortium 2001).

Le génome humain contient une majorité de séquences répétées dérivées des ETs. On trouve représentés des éléments des classes I et II. La très grande majorité est composée d'éléments LINE et SINE (20,42% et 13,14% du génome séquencé, respectivement qui correspondent à un nombre de copies des SINEs d'environ 500 000 et pour les LINEs d'environ 10 000 copies (International Human Genome Sequencing Consortium 2001)). On trouve environ 8,3% de rétrotransposons à LTR et 2,8% de transposons (International Human Genome Sequencing Consortium 2001). Les éléments SINE sont principalement représentés par les éléments *Alus*. Ces éléments sont dans un grand nombre de cas impliqués dans des maladies, comme par exemple l'haemophilie B ou la dystrophie musculaire de Duchenne, lorsqu'ils sont insérés dans les parties codantes de gènes majeurs ou dans les régions régulatrices (voir pour revue Deragon et Capy 2000).

1.2.5. *Saccharomyces cerevisiae*

La levure est un champignon ascomycète communément utilisé comme ferment pour la production d'alcool, comme par exemple la bière, et par les boulangers pour faire lever le pain. Elle est étudiée comme organisme modèle depuis plusieurs décennies dans la recherche en génétique moléculaire car les mécanismes basiques de la réplication, de la recombinaison, de la division cellulaire et du métabolisme sont bien conservés entre la levure et les autres eucaryotes dont les mammifères. Le génome de la levure est le plus petit parmi ceux présentés dans cette étude puisqu'il représente 12 Mb réparties sur 16 chromosomes. Depuis avril 1996, la séquence complète de *Saccharomyces* est disponible au public (Goffeau *et al.* 1997). Le génome de *Saccharomyces* est globalement riche en AT avec en moyenne 62% d'AT, les parties codantes étant plus riches en GC.

La levure ne présente qu'une classe d'ETs : il s'agit de rétrotransposons à LTR. Les différentes séquences du génome complet ont été bien caractérisées : on connaît parfaitement leur degré d'homologie entre eux ainsi que leur localisation sur les chromosomes (Kim *et al.* 1998). Ces ETs se repartissent en cinq familles : *Ty1*, *Ty2*, *Ty4*, *Ty5* (qui font partie de la superfamille *Ty1-copia*) et *Ty3* (qui fait partie de la superfamille *Ty3-gypsy*). Ces éléments présentent des sites d'intégration très spécifiques. Par exemple, les éléments *Ty5* s'insèrent préférentiellement près des télomères car leur complexe d'intégration reconnaît un composant de la chromatine télomérique (Zou et Voytas 1997). Récemment, une étude des séquences du génome complet de la levure a mis en évidence l'existence d'éléments hybrides *Ty1-Ty2* résultats d'événements de recombinaison entre des éléments *Ty1* et *Ty2* (Jordan et McDonald 1999).

Tableau 1 : récapitulatif des données génomiques pour les 5 espèces.

espèces	taille totale du génome (Mb)	nombre de chromosomes par génome haploïde	nombre de gènes	%GC total	%ETs
A. thaliana	120	5	26 000	35	14
C. elegans	97	6	19 000	36	17
D. melanogaster	180	4	14 332	43	15
H. sapiens	3 286	23	30 000	42	45
S. cerevisiae	12	16	6 400	38	3

2ème PARTIE : L'USAGE DES CODONS DES ELEMENTS TRANSPOSABLES ET DES GENES D'HOTE

2ème partie : L'usage des codons des éléments transposables et des gènes d'hôte

Les ETs sont d'anciens composants des génomes puisqu'on les retrouve chez tous les organismes vivants. Ainsi, ils dépendent de la machinerie de leur hôte pour leur transcription et leur traduction. Donc, un moyen d'étudier si ces ETs subissent les mêmes contraintes que le génome hôte est d'analyser leur usage des codons et de le comparer à celui des gènes de l'hôte.

2.1. L'usage des codons

2.1.1. Définition

Les séquences codantes peuvent être considérées comme un enchaînement d'unités constituées de trois nucléotides adjacents : les codons, qui sont à mettre en relation avec les 20 acides aminés pouvant constituer les protéines. Comme il existe quatre bases différentes (A = l'Adénine, T = la Thymine, G = la Guanine et C= la Cytosine), il y a 4^4 triplets possibles, soit 64 codons différents. Chaque codon a une signification et ne peut désigner qu'un seul acide aminé, dans un organisme donné : le code génétique est donc non ambigu (Tableau 2). Cela implique aussi qu'il soit **dégénéré**, c'est-à-dire qu'un acide aminé peut être codé par plus d'un codon. Ainsi, les codons déterminant le même acide aminé sont appelés **codons synonymes**. Ces codons ne diffèrent généralement que par la base en 3ème position du codon. La dégénérescence du code génétique a des conséquences biologiques importantes : elle permet de minimiser l'effet de mutations et elle autorise une variation de la composition en bases sans changer la nature de la séquence protéique. Ainsi, si on prend l'exemple de la valine qui est codé par 4 codons (GTA, GTC, GTT, GTG), on peut constater qu'un changement au niveau de la 3ème position d'un des codons donnera toujours un codon déterminant la valine. Par contre, si on change la 1ère base G en A, on aura, suivant la 3ème base ATA, ATC, ATT (qui codent pour l'isoleucine) ou ATG (qui code pour la méthionine). De même un changement en 2ème position du T en C par exemple donnera, quelle que soit la 3ème base, des codons déterminant l'alanine. Ainsi, les positions 1 et 2 lorsqu'elles sont mutées vont entraîner dans la majorité des cas un changement de l'acide aminé codé.

Parmi les 64 codons possibles, trois ne codent pas pour des acides aminés. Ces codons, TGA, TAA et TAG, sont appelés codons stop et signifient la fin de la séquence codante chez la plupart des organismes. La dégénérescence du code a deux exceptions : la méthionine et le tryptophane qui ne sont codés chacun que par un seul codon. La particularité du codon spécifiant la

méthionine est qu'il est aussi le codon d'initiation de la traduction chez la plupart des organismes.

Tableau 2 : le code génétique universel

codon	***acide aminé***	***codon***	***acide aminé***	***codon***	***acide aminé***	***codon***	***acide aminé***
AAA	Lys	ATA	Ile	ACA	Thr	AGA	Arg
AAC	Lys	ATC	Ile	ACC	Thr	AGC	Ser
AAT	Asn	ATT	Ile	ACT	Thr	AGT	Ser
AAG	Asn	**ATG**	**Met**	ACG	Thr	AGG	Arg
TAA	***stop***	TTA	Leu	TCA	Ser	**TGA**	***stop***
TAC	Tyr	TTC	Phe	TCC	Ser	TGC	Cys
TAT	Tyr	TTT	Phe	TCT	Ser	TGT	Cys
TAG	***stop***	TTG	Leu	TCG	Ser	TGG	Trp
CAA	Gln	CTA	Leu	CCA	Pro	CGA	Arg
CAC	His	CTC	Leu	CCC	Pro	CGC	Arg
CAT	His	CTT	Leu	CCT	Pro	CGT	Arg
CAG	Gln	CTG	Leu	CCG	Pro	CGG	Arg
GAA	Glu	GTA	Val	GCA	Ala	GGA	Gly
GAC	Asp	GTC	Val	GCC	Ala	GGC	Gly
GAT	Asp	GTT	Val	GCT	Ala	GGT	Gly
GAG	Glu	GTG	Val	GCG	Ala	GGG	Gly

2.1.2. Utilisation et interprétation

La dégénérescence du code génétique implique donc l'existence de codons synonymes déterminant un même acide aminé. Ainsi, des mutations synonymes se produisant sur ces codons ne vont pas changer la séquence protéique. La sélection naturelle étant supposée agir au niveau protéique, les mutations synonymes sont des candidates parfaites pour être des mutations sélectivement neutres (KIMURA 1968). Cependant, sous cette hypothèse, les différents codons synonymes doivent être utilisés de manière équiprobable. Or, ce n'est pas le cas. Dans les années 80, GRANTHAM *et al.* (1980) ont montré que l'usage des codons synonymes n'est pas aléatoire à la fois chez les procaryotes et chez les eucaryotes. En effet, en effectuant une Analyse Factorielle des Correspondances (voir paragraphe 2.1.4.) sur la fréquence des 61 codons possibles de 90 séquences provenant d'organismes différents comme des virus, des phages, des bactéries et des mammifères, ils ont observé que les différents gènes se regroupaient suivant leur génome d'origine. Ainsi les

gènes d'un même organisme montrent un même « pattern » de choix des codons synonymes : c'est « **l'hypothèse du génome »** (GRANTHAM *et al.* 1980). Sous cette hypothèse, les gènes d'une espèce donnée utilisent la même stratégie codante c'est-à-dire que le biais d'usage des codons est spécifique d'une espèce. De plus, des espèces proches utilisent des stratégies codantes proches.

Cette hypothèse est généralement vraie pour les génomes de très petite taille comme les phages et les virus, mais il existe quand même une certaine hétérogénéité d'usage des codons entre gènes d'un même organisme. Ainsi, POST *et al.* (1979) ont montré que les gènes codant pour les protéines ribosomales d'*Escherichia coli* utilisent préférentiellement des codons correspondants aux ARN de transfert (ARNt) les plus abondants dans la cellule. Les différences de quantité en ARNt sont corrélées avec le nombre de gènes codant pour les ARNt (PERCUDANI *et al*. 1997 ; DURET 2000). La préférence d'utilisation des ARNt les plus abondants s'explique par la sélection naturelle de ces codons qui permettent une meilleure efficacité de la traduction en terme de vitesse, étant donné que les ARNt correspondants sont rapidement disponibles. IKEMURA (1981 ; 1982) a confirmé ce phénomène chez *E. coli* et chez *S. cerevisiae* en montrant que chez ces deux espèces, il existe une corrélation positive entre la fréquence relative des codons synonymes dans un gène et l'abondance relative en ARNt. Cette corrélation est particulièrement forte pour les gènes très exprimés c'est-à-dire les gènes donnant un grand nombre d'ARN messagers. De plus, il a été observé que le biais d'usage des codons est plus fort dans les gènes hautement exprimés que dans les gènes faiblement exprimés chez ces deux espèces. Cette différence s'explique par le fait que dans les gènes fortement exprimés, la sélection pour une bonne efficacité et une bonne fidélité de la traduction est plus importante. Ainsi, chez les gènes très exprimés de ces organismes, l'explication du biais d'utilisation des codons vient principalement d'une pression de sélection en faveur de l'optimisation de la vitesse de la traduction (SHARP et MATASSI 1994). Ce type de sélection est supposé faible et n'agissant réellement que chez les organismes possédant une grande taille efficace* (BULMER 1987 ; LI 1987). Ainsi, chez les mammifères, dont la taille efficace* est réduite, une telle sélection est négligeable devant l'effet de la dérive. De plus, le contenu en bases d'un mammifère varie beaucoup le long de son génome et l'usage des codons semble corrélé avec le contenu local en GC (BERNARDI *et al.* 1985). SUEOKA (1962) a proposé la notion de « pression mutationnelle dirigée » pour expliquer ces variations en G+C. Cette théorie fixe le taux de conversion des bases G et C vers A et T (u) et le taux de conversion des bases A et T vers G et C (v), et définit le taux de GC à l'équilibre de l'expression $v/(u+v)$. Les déviations par rapport à l'équilibre impliqueront une pression de sélection vers AT ou vers GC. Ainsi, l'usage des codons dans ce cas semble être le reflet de «patterns » de mutations qui peuvent varier le long du génome. Dans certains cas, c'est le biais mutationnel qui semble le principal responsable du biais de l'usage des codons.

Chez *Drosophila melanogaster*, qui possède une taille efficace* relativement grande, SHIELDS *et al.* (1988) ont montré que la sélection pour certains codons synonymes se retrouve, et que le biais est lié au niveau d'expression des gènes. D'après AKASHI (1994) cette sélection agit sur l'usage des codons pour permettre une augmentation de la fidélité de la traduction. En effet, l'auteur montre qu'il y a une corrélation entre le biais d'usage des codons et les contraintes fonctionnelles sur les protéines. Ainsi, au niveau de domaines très conservés, fonctionnellement importants, on trouve d'avantage de codons préférés*.

Le nématode montre également un biais d'usage des codons plus fort dans les gènes hautement exprimés que dans les gènes faiblement exprimés, qui semblent plutôt sujets à des biais mutationnels (STENICO *et al.* 1994). Ainsi, chez certains organismes, le biais d'usage des codons résulte d'un équilibre entre pression de sélection pour l'utilisation de codons optimaux* pour la traduction et biais mutationnel qui permet la persistante des codons non-optimaux. Cette théorie est appelée « selection-mutation-drift » (BULMER 1991).

Il existe aussi un lien entre usage des codons et taux de recombinaison. Ainsi, chez la drosophile, les gènes se trouvant dans des régions possédant un faible taux de recombinaison ont un biais d'usage des codons faible (KLIMAN et HEY 1993). Ce phénomène est expliqué par un modèle de génétique des populations qui prédit que la sélection naturelle ne peut agir dans les régions où la recombinaison est faible : il s'agit de l'effet Hill-Robertson (HILL et ROBERTSON 1966). Une relation est aussi trouvée entre le biais d'usage du code et la longueur des gènes. Ainsi, il y a une relation négative entre la longueur de la partie codante d'un gène et son biais d'usage des codons (COMERON *et al.* 1999 ; DURET et MOUCHIROUD 1999 ; MORIYAMA et POWELL 1998). Cependant, on observe la corrélation inverse chez *E. coli* (MORIYAMA et POWELL 1998). Il semble donc que les mécanismes agissant sur l'usage des codons chez des procaryotes ne soient pas les mêmes que ceux agissant chez les eucaryotes.

L'usage des codons est un reflet de l'ensemble des contraintes que subissent les gènes dans un organisme. L'étude du choix des codons des ETs peut donc nous aider à comprendre les contraintes qu'ils subissent. En particulier on doit pouvoir déterminer si les ETs se caractérisent par des contraintes spécifiques ou s'ils adoptent l'usage des codons du génome hôte ou de la région dans laquelle ils sont insérés.

2.1.3. Les indices de biais

Dans cette partie, les différents indices existants pour mesurer le biais de l'usage des codons seront présentés. Il en existe deux types : ceux qui prennent comme hypothèse H_0 l'utilisation

aléatoire des différents codons synonymes (CBI, c^2 normalisé, N_c) et ceux qui mesurent le biais en comparant la fréquence observée de différents codons synonymes avec la fréquence de codons optimaux (F_{op}, CAI).

2.1.3.1. Le « Codon Bias Index » (CBI)

Cet indice (Benetzen et Hall 1982 ; Morton 1993 ; Morton 1994) permet de mesurer la déviation par rapport à une utilisation uniforme de codons synonymes. Les valeurs peuvent aller de 0 (pas de biais) à 1 (biais extrême). L'expression de l'indice est la suivante :

$$CBI = \sum_{i=1}^{18} \left(\frac{n_i}{n_{tot}} \right) \sum_{j=1}^{s_i} \left(\frac{\left(1 - R_{ij}\right)^2}{(s_i - 1)} \right)$$

où $R_{ij} = \frac{n_{ij}}{n_{i\,max}}$ est l'utilisation relative des codons synonymes

Avec s_i = le nombre de codons synonymes pour le $i^{ème}$ acide aminé (dégénérescence du $i^{ème}$ acide aminé).

n_i = le nombre d'occurrences du $i^{ème}$ acide aminé.

n_{tot} = le nombre total d'acides aminés, exceptés la méthionine et le tryptophane qui ne sont codés chacun que par un seul codon.

n_{ij} = le nombre d'occurrences du $j^{ème}$ codon synonyme de l'acide aminé i.

n_{imax} = le nombre d'occurrences du codon synonyme le plus utilisé de l'acide aminé i.

2.1.3.2. Le χ^2 normalisé

Il s'agit d'un c^2 calculé à partir de la déviation par rapport à un usage des codons équiprobable à l'intérieur des groupes synonymes, puis divisé par le nombre total de codons dans le gène en excluant le tryptophane et la méthionine (Shields *et al.* 1988). Ainsi, cet indice est normalisé par la longueur du gène. C'est une mesure simple et indépendante de la longueur du gène pour des gènes de plus de 100 codons.

2.1.3.3. Le nombre efficace de codons (N_c)

Cet indice proposé par Wright (1990) permet de quantifier le nombre de codons également fréquents. Il a été établi en faisant un parallèle avec les fréquences alléliques observées à un locus

donné. Chaque acide aminé est considéré comme un locus présentant différents allèles et on calcule une homozygotie pour chaque acide aminé. On subdivise la table du code génétique selon le nombre de codons synonymes codant pour chaque acide aminé. On a ainsi cinq types : 2 acides aminés codés par 1 codon, 9 acides aminés codés par 2 codons, 1 acide aminé codé par 3 codons, 5 acides aminés codés par 4 codons et 3 acides aminés codés par 6 codons.

On mesure la contribution d'un acide aminé au biais global d'un gène :

$$\bar{F}_k = \frac{\left(n \sum_{i=1}^{k} p_i^2 - 1 \right)}{(n-1)}$$

avec n = le nombre de codons dans la séquence.

p_i = la fréquence relative du i^{ème} codon synonyme.

k = le nombre de codons synonymes pour l'acide aminé considéré.

Le N_c est obtenu en sommant les contributions de chacun des cinq types d'acide aminés :

$$N_c = 2 + \frac{9}{\overline{F}_2} + \frac{1}{\overline{F}_3} + \frac{5}{\overline{F}_4} + \frac{3}{\overline{F}_6}$$

avec `F_i = la moyenne de la contribution des acides aminés du type i.

On peut faire des ajustements pour les acides aminés rares ou absents. Cet indice permet de quantifier le biais d'un gène à partir d'une évaluation de la quantité de codons préférés* par rapport à un usage uniforme des codons synonymes. Il prend des valeurs comprises entre 20 (biais extrême) et 61 (pas de biais).

2.1.3.4. La fréquence des codons optimaux (F_{op})

Il s'agit d'une mesure espèce-spécifique du biais vers les codons optimaux pour la traduction (IKEMURA 1981). Cet indice prend des valeurs de 0 (pas de biais) à 1 (biais extrême). Il est proche de l'indice défini par BENNETZEN et HALL (1982).

On le calcule de la manière suivante : $F_{op} = \frac{x_i}{x_{tot}}$

avec x_i = le nombre d'occurrences des codons optimaux.

x_{tot} = le nombre total d'occurrences des 18 acides aminés possédant des codons synonymes.

2.1.3.5. Le « Codon Adaptation Index » (CAI)

Cet indice, proposé par SHARP et LI (1987), estime le degrés d'adaptation des codons synonymes d'un gène par rapport à l'usage optimal. Il teste l'hypothèse que l'usage des codons d'un gène est fortement biaisé. La première étape est de calculer l'usage relatif des codons synonymes des gènes fortement exprimés ($RSCU_{max}$). On obtient ainsi une table de référence. Le RSCU pour un codon est calculé comme suit :

$$RSCU_{ij} = \frac{n_{ij}}{\left(\frac{1}{n_i} \sum_{j=1}^{n_i} n_{ij} \right)}$$

avec n_{ij} = le nombre d'occurrences du j^{ème} codon pour le i^{ème} acide aminé.

n_i = le nombre de codons possibles pour le i^{ème} acide aminé (1, 2, 3, 4 ou 6).

Une valeur de RSCU proche de 1 indique un manque de biais pour le codon considéré.

On peut calculer l'adaptation relative d'un codon :

$$w_{ij} = \frac{RSCU_{ij}}{RSCU_{i\,max}} = \frac{n_{ij}}{n_{i\,max}}$$

où $RSCU_{i\,max}$ et $n_{i\,max}$ sont les valeurs de RSCU et de n pour le codon le plus utilisé pour le i^{ème} acide aminé.

Le CAI consiste en la moyenne géométrique des valeurs de RSCU correspondant à chaque codon utilisé dans un gène, divisée par le CAI maximal pour un gène de même composition en acides aminés.

$$CAI = \frac{CAI_{obs}}{CAI_{max}}$$

avec $$CAI_{obs} = \left(\prod_{k=1}^{L} RSCU_k \right)^{\frac{1}{L}}$$ et $$CAI_{max} = \left(\prod_{k=1}^{L} RSCU_{k\,max} \right)^{\frac{1}{L}}$$

où *L* est le nombre de codons dans le gène, $RSCU_k$ la valeur de RSCU pour le $k^{ème}$ codon dans le gène, et $RSCU_{k\ max}$ la valeur maximale de RSCU pour l'acide aminé codé par le codon *k* dans le gène.

Les valeurs de CAI sont comprises entre 0 et 1 suivant la similitude d'usage des codons du gène considéré avec l'ensemble de référence. L'ensemble de référence étant constitué de gènes hautement exprimés, cet indice permet une mesure indirecte du niveau d'expression du gène étudié.

2.1.3.6. Comparaison des différents indices

Nous avons vu que les différents indices de biais peuvent être repartis en deux classes : ceux qui comparent les données à une hypothèse de départ fondée sur l'utilisation aléatoire des codons, ceux qui quantifient le biais en comparant les fréquences observées aux fréquences des codons préférés.

COMERON et AGUADÉ (1998) ont effectué une comparaison d'un certain nombre de ces indices par simulation afin d'étudier l'influence de la longueur des gènes suivant un biais particulier sur les valeurs des indices. Ainsi, il apparaît que le CBI est moins sensible aux fluctuations que le χ^2 normalisé, à toutes les conditions de longueurs de gènes et de biais. Le N_c et le CAI ne semblent pas affectés par la longueur des gènes, quel que soit le biais. Ainsi, ces deux indices semblent particulièrement adaptés pour la comparaison de séquences de différentes longueurs. Dans le cas de séquences de même longueur, le N_c et le CBI donnent les meilleurs résultats. Le CAI est une mesure performante d'un biais par rapport à un ensemble de référence. Cependant, cet indice ne peut être réellement valable que dans le cas où il existe seulement deux catégories de gènes dans un organisme : les gènes fortement exprimés et les gènes faiblement exprimés. Ce n'est pas forcément le cas pour l'ensemble des organismes ou même au sein d'un organisme pluricellulaire dans lequel il peut y avoir des variations selon les tissus.

En ce qui concerne le F_{op}, il s'agit d'une mesure assez fine du biais qui présente cependant des inconvénients. Principalement, la définition des codons optimaux est délicate et n'est pas forcement généralisable chez toutes les espèces, ce qui peut poser problème si on veut comparer des espèces différentes. De plus, le calcul de l'indice est fortement dépendant de la définition de ces codons optimaux.

Ainsi, aucun des indices existants pour exprimer les biais n'est idéal. Ils sont souvent complémentaires et certains indices sont plus adaptés que d'autres selon la question que l'on se pose.

2.1.4. L'Analyse Factorielle des Correspondances comme outil

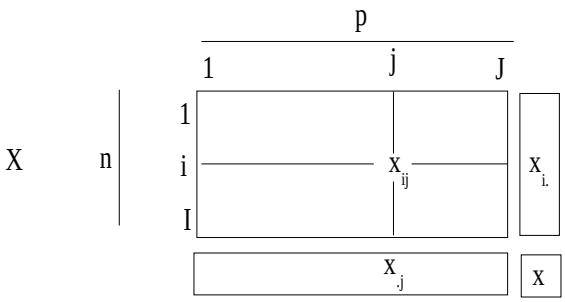
Les techniques d'analyses de données permettent la représentation graphique d'un grand nombre d'observations contenues dans des tableaux multidimensionnels dans un espace de dimensions réduites, tout en évitant au maximum la perte d'informations. Il s'agit de statistiques descriptives qui n'imposent pas de conditions sur les distributions des données de départ et ne modifient pas les tendances présentes dans les données de départ.

L'Analyse Factorielle des Correspondances (AFC) a été développée dans les années 60 par J-P Benzécri (1973). Elle a depuis été souvent utilisée pour l'étude de l'usage des codons. Bien qu'adaptée aux tableaux de contingence, elle peut aussi être appliquée à des tableaux où toutes les données sont positives et de même nature. Il s'agit donc d'une représentation graphique qui calcule les axes de projection les plus aptes à représenter le tableau de données, en minimisant la perte d'informations. L'AFC a ceci de particulier que les lignes et les colonnes du tableau jouent le même rôle. Elle permet la représentation simultanée des lignes et des colonnes sur le même graphique car il existe des relations simples entre les coordonnées factorielles des deux nuages.

Le but de cette partie est de décrire brièvement les principales étapes de calculs de l'AFC (pour plus de détails voir Lefebre 1983 et Lebart *et al*. 1997) et d'exposer son utilisation pour l'étude de l'usage des codons. La Figure 5 représente la synthèse des différents calculs.

2.1.4.1. Les différentes étapes de l'analyse

Soit un tableau de données X contenant *p* colonnes et *n* lignes



$$x_{i.} = \sum_{j}^{p} x_{ij}$$

$$x_{j.} = \sum_{i}^{n} x_{ij}$$

$$x = \sum_{i,j} x_{ij}$$

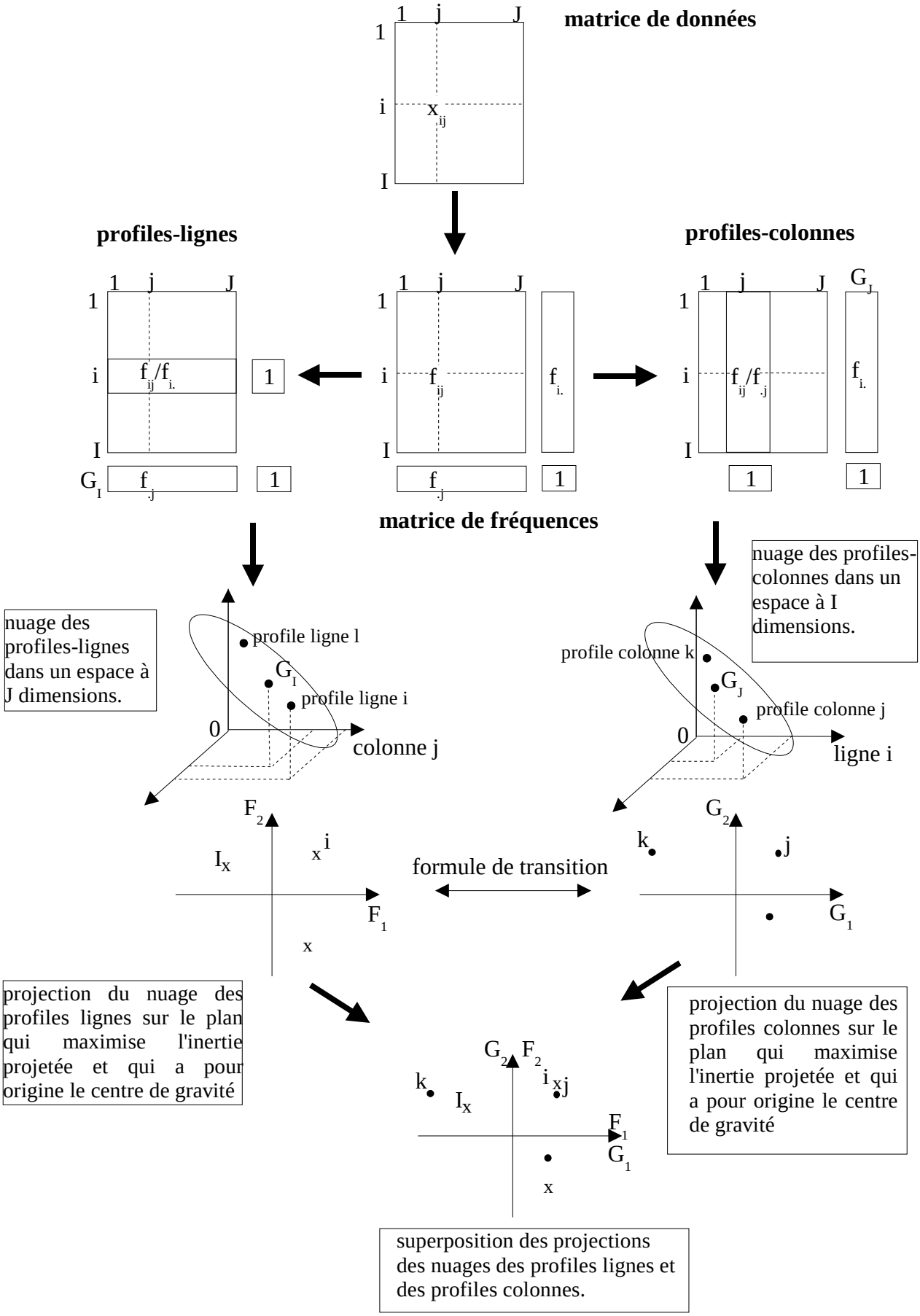


Figure 5 : schéma récapitulatif des différentes étapes de l'AFC

Calcul du tableau de fréquences

A partir du tableau X et des relations précédentes, on peut calculer les fréquences :

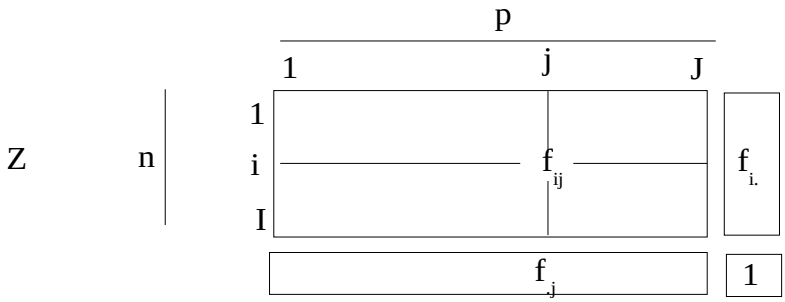
$$f_{ij} = \frac{x_{ij}}{x}$$

$f_{i.} = \sum_{j}^{p} f_{ij}$ la fréquence marginale de la ligne *i*

$f_{.j} = \sum_{i}^{n} f_{ij}$ la fréquence marginale de la colonne *j*

$$\sum_{i,j} f_{ij} = 1$$

on obtient donc le tableau Z suivant :



Deux variables sont indépendantes si pour chaque *i* et chaque *j* : $f_{ij} = f_{i.}\, f_{j}$

De ce tableau, on peut obtenir deux matrices diagonales : la matrice marges-lignes D_n (d'ordre (n,n)) et la matrice marges-colonnes D_p (d'ordre (p,p)). Ces matrices possèdent comme éléments diagonaux les marges en lignes $f_{i.}$ pour D_n et les marges en colonnes $f_{.j}$ pour D_p.

Construction des nuages

La matrice Z peut être analysée symétriquement selon les colonnes d'une part et selon les lignes d'autre part. On obtient donc deux nuages de points :

- le nuage des *n* lignes : l'ensemble des profils-lignes forme un nuage de *n* points dans l'espace des *p* colonnes. Chaque point possède une masse $f_{i.}$ et des coordonnées $f_{ij} / f_{i.}$
- le nuage des *p* colonnes : l'ensemble des profils-colonnes constitue de la même manière un nuage de *p* points dans l'espace des *n* lignes. Chaque point possède une masse $f_{.j}$ et des coordonnées $f_{ij} / f_{.j}$

On va chercher à représenter géométriquement les similitudes entre deux points-lignes ou entre

deux points-colonnes. Ainsi, on va calculer une distance entre deux points-lignes ou deux points-colonnes. Cette distance n'est pas une distance euclidienne classique mais une distance euclidienne pondérée par l'inverse de la masse de la colonne ou de la ligne que l'on appelle **distance du χ^2**.

Pour les profils lignes $$d^2(i,i') = \sum_{j=1}^{p} \left(\frac{1}{f_{.j}}\right)\left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}}\right)^2$$

Pour les profils colonnes $$d^2(j,j') = \sum_{i=1}^{n} \left(\frac{1}{f_{i.}}\right)\left(\frac{f_{i.}}{f_{.j}} - \frac{f_{ij'}}{f_{.j'}}\right)^2$$

Une propriété de cette distance est **l'équivalence distributionnelle**. La moyenne pondérée (c'est-à-dire le centre de gravité du nuage) des profiles lignes et, symétriquement des profils colonnes, est exactement la fréquence marginale $f_{.j}$ et $f_{i.}$, respectivement.

$$\sum_i f_{i.} \frac{f_{ij}}{f_{i.}} = f_{.j}$$
$$\sum_j f_{.j} \frac{f_{ij}}{f_{.j}} = f_{i.}$$

La dispersion du nuage des profils lignes autour de son centre de gravité G peut être mesurée par son inertie. Ainsi, l'inertie du point *i* est le produit de sa masse $f_{i.}$ par la distance carré $d^2(i,G)$. L'inertie du nuage des lignes est la somme des inerties de tous les points lignes.

$$Inertie(I) = \sum_i f_{i.} d^2(i,G) = \sum_i f_{i.} \sum_j \left(\frac{1}{f_{.j}}\right)\left(\frac{f_{ij}}{f_{i.}} - f_{.j}\right)^2 = \sum_{ij} \left(\frac{\left(f_{ij} - f_{i.} f_{.j}\right)^2}{(f_{i.} f_{.j})}\right)$$

Il y a une symétrie complète entre les indices *i* et *j*, donc l'inertie du nuage des colonnes est égale à la même expression : In(I) = In(J). La valeur de l'inertie est un indicateur de la dispersion du nuage et mesure la liaison entre les deux variables.

Calcul de la matrice d'inertie

Pour chaque nuage, l'origine est transférée au niveau du centre de gravité. Ainsi, on calcule la matrice d'inertie R entre colonnes comme suit :

$$R = Z' D_n^{-1} Z D_p^{-1}$$

soit de terme générale : $$R_{jj'} = \sum_{i=1}^{n} \left(\frac{(f_{ij} f_{ij'})}{(f_{i.} f_{.j'})}\right)$$

A partir de cette matrice vont être calculés les valeurs propres et les vecteurs propres. Le calcul de la matrice d'inertie entre lignes est aussi possible. En pratique, la solution choisie est celle qui est la plus simple : celle qui donne la matrice d'inertie de plus petite dimension.

Valeurs propres et vecteurs propres

La diagonalisation de la matrice d'inertie R permet le calcul de valeurs propres (inerties de chaque axe). A chaque valeur propre *l* est associée un vecteur propre *m* tel que *Rm = lm*. les vecteurs propres permettent de définir les axes de projection des nuages de lignes et de colonnes. Les valeurs propres permettent de quantifier la part d'information expliquée par chaque axe.

Calcul des coordonnées des points sur les axes factoriels

Les coordonnées des colonnes et des lignes sont calculées à partir des valeurs propres et des composantes des vecteurs propres.

Ainsi, on calcule la coordonnée d'un point colonne *j* suivant l'axe *k* :

$$x_j^k = m y_j^k \sqrt{\frac{\lambda_k}{f_j}}$$

avec μ_j^k = la $j^{ème}$ composante du $k^{ème}$ vecteur propre

λ_k = la $k^{ème}$ valeur propre

f_j = le total de la colonne *j* dans la matrice Z

On calcule la coordonnée du point ligne *i* suivant l'axe *k* à partir des coordonnées des points colonnes :

$$x_i^k = \frac{1}{\sqrt{\lambda_k}} \sum_{j=1}^{p} \left(\frac{f_{ij}}{f_i} \right) x_j^k$$

avec λ_k = la $k^{ème}$ valeur propre

p = le nombre total de colonnes

f_{ij}/f_i = la fréquence relative normalisée de la variable *j* chez l'individu *i*

x_j^k = la coordonnée de la variable *j* suivant l'axe *k*

La coordonnée d'un point-colonne est aussi calculable à partir des coordonnées des points lignes. Cette faculté permet de passer facilement des coordonnées des points lignes à celles des points colonnes grâce à une formule de transition. On peut donc superposer les projections des deux nuages de points.

Aides à l'interprétation

- Les **contributions absolues** : elles expriment la part prise par une variable dans la variance expliquée par un facteur. Ainsi, on cherche à connaître les éléments responsables de la construction de l'axe. Le calcul de la contribution de chaque point prend en compte les coordonnées de chacun des points ainsi que le poids de chaque point qui explique la variabilité observée sur cet axe. La contribution de la variable *j* à la variance expliquée par l'axe *k* est calculée comme suit :

$$c_j^k = (my_j^k)^2 \times 100 = \frac{(f_j (x_j^k)^2)}{\lambda_k} \times 100$$

avec μ_j^k = la j^{ème} composante du k^{ème} vecteur propre

f_j = le total de la colonne *j* dans le tableau Z

x_j^k = la coordonnée de la variable *j* suivant l'axe *k*

λ^k = la *k*^{ème} valeur propre

On peut calculer de même la contribution de l'individu i à la variance expliquée de l'axe k :

$$c_i^k = (my_i^k)^2 \times 100 = \frac{(f_i (x_i^k)^2)}{\lambda_k} \times 100$$

avec μ_i^k = la *i*^{ème} composante du *k*^{ème} vecteur propre

f_i = le total de la colonne *i* dans le tableau Z

x_i^k = la coordonnée de la variable *i* suivant l'axe *k*

- Les **cosinus carrés** : ils expriment la part prise par un facteur dans la dispersion d'une variable. On cherche à apprécier si un point est bien représenté sur un sous-espace. Le cosinus carré d'une variable *j* avec l'axe *k* est calculé comme suit :

$$\cos^2(j,k) = \frac{(x_j^k)^2}{\sum_{i=1}^{p} (x_j^i)^2}$$ avec x_j^k la coordonnée de la variable *j* suivant l'axe *k*.

Plus la valeur du cosinus carré est proche de 1, plus la position du point observé en projection est proche de la position réelle du point dans l'espace. On apprécie la qualité de la représentation d'un point dans un plan en faisant la somme des cosinus carrés sur les axes étudiés.

2.1.4.2. Application à l'usage des codons

Un certain nombre d'études ont utilisé l'AFC dans l'analyse du biais des codons. Originellement, c'est GRANTHAM *et al.* (1980) qui ont utilisé cette méthode pour l'étude de l'usage des codons de différents organismes. Elle est depuis couramment employée sur différents types de données représentant des indices de biais différents. Par exemple, en 1989, SHIELDS et SHARP ont effectué une étude sur l'usage des codons de gènes de *Drosophila melanogaster* et d'un petit nombre de séquences d'ETs où l'AFC a été réalisée sur les valeurs de RSCU des séquences. L'utilisation de cette méthode qui montre la répartition en trois classes des gènes d'*E. coli*, a été effectuée sur les fréquences relatives des codons (MÉDIGUE *et al.* 1991). Il en est de même lors d'une étude effectuée sur *A. thaliana* qui a permis de montrer une relation entre l'usage des codon et la fonction des gènes (CHIAPELLO *et al.* 1998).

Dans le cas de l'étude de l'usage des codons telle que nous allons le voir, la matrice de données de départ correspond à une matrice composée de *n* lignes correspondant aux *n* gènes analysés, et de 59 colonnes correspondant aux fréquences relatives des 59 codons synonymes. Le croisement entre une ligne et une colonne contient donc la fréquence relative d'un codon particulier dans un gène particulier. Le choix de la fréquence relative a été fait car elle permet de donner un même poids à chacune des séquences. On élimine ainsi tout biais relatif à la composition en acides aminés. Les analyses ont été effectuées à partir des fréquences absolues des codons c'est à dire les effectifs des codons. Les résultats obtenus sont identiques. Ainsi, la fréquence relative d'un codon est le rapport entre le nombre d'occurrences de ce codon et le nombre d'occurrences de l'acide aminé qu'il code. D'après le paragraphe précédent, dans le nuage des points lignes, chaque point représente un gène. Ainsi, les gènes qui auront un usage des codons semblable seront proches dans l'espace. De même, le déroulement de l'AFC nous permet la comparaison entre les codons suivant leur utilisation dans l'ensemble des gènes. En pratique, une fois l'AFC réalisée, les résultats sont interprétés en fonction de différents paramètres. Tout d'abord, il convient de choisir les axes qui seront examinés : nous choisirons ceux qui déterminent le plus grand pourcentage de variabilité. Ainsi, généralement, on n'examine pas plus de 3 ou 4 axes. Une fois la projection réalisée des gènes et des codons, l'analyse des contributions aide à déterminer la participation de chaque point à la formation des axes. Les calculs des fréquences des codons et de la composition en bases ont été obtenus à l'aide d'un programme en C++ et toutes les AFC ont été effectuées à l'aide du logiciel ADE-4 (THIOULOUSE *et al.* 1997). Les groupes observés avec l'AFC peuvent être analysés statistiquement afin de déterminer s'ils sont bien distincts les uns des autres. Pour cela, nous avons utilisé l'analyse de la variance multivariée (MANOVA) sur les coordonnées de chaque point sur les différents axes

considérés. Cette analyse effectue essentiellement des tests d'hypothèse sur les moyennes des différents groupes.

2.2. Les analyses

Dans ce paragraphe, je vais exposer des analyses effectuées sur l'usage des codons des ETs et des gènes de différents hôtes. En annexe sont reportés les articles correspondants aux différentes études.

2.2.1. Origine de l'élément *P*

Sous « l'hypothèse du génome », le biais de l'usage des codons est souvent utilisé pour caractériser des gènes acquis par transfert horizontal. Ainsi, les éléments transposables étant parfois capables d'être transmis horizontalement, un critère de preuve souvent avancé est que l'usage des codons de l'élément en question est soit proche de celui de l'espèce dont il provient (Powell et Gleason 1996) soit possède un usage proche de gènes dits « aliens » supposés provenir de transferts horizontaux (Médigue *et al.* 1991).

L'élément *P* est un transposon de 2,9 kb initialement découvert chez *Drosophila melanogaster*. Il code pour une protéine unique : la transposase, d'une taille de 2262 pb (764 acides aminés). Cet élément est impliqué dans un système de dysgénésie* des hybrides que l'on observe dans la descendance de croisement entre des mâles possédant cet élément et des femelles n'en possédant pas (Hiraizumi 1971). Des analyses ont montré que cet élément n'était pas présent dans les vieilles souches de laboratoire collectées avant 1950. Il en a été déduit que cet élément provenait d'une introduction récente suivie d'une invasion rapide. Brookield *et al.* (1984) ont montré que cet élément n'est présent, dans le sous-groupe *melanogaster*, que chez l'espèce *D. melanogaster*. Il a aussi été détecté dans les espèces des groupes *obscura*, *willistoni*, *saltans* et chez *Scaptomyza pallida*. La Figure 6 présente une phylogénie simplifiée des espèces de drosophiles utilisées dans cette étude. L'analyse de la séquence de l'élément *P* de *D. melanogaster* a montré que celui-ci ne diffère de l'élément de *D. willistoni* que par un nucléotide à la position 32, ce qui fait que les parties codantes sont rigoureusement identiques, alors que ces deux espèces ont divergé il y a environ 50 millions d'années (Daniels *et al.* 1990). Ceci est interprété par la transmission horizontale de l'élément *P* de *D. willistoni* vers *D. melanogaster*. Powell et Gleason (1996) ont analysé l'usage du code de l'élément *P* et l'ont comparé avec celui de gènes de *D. willistoni*. Ils ont conclu que l'usage

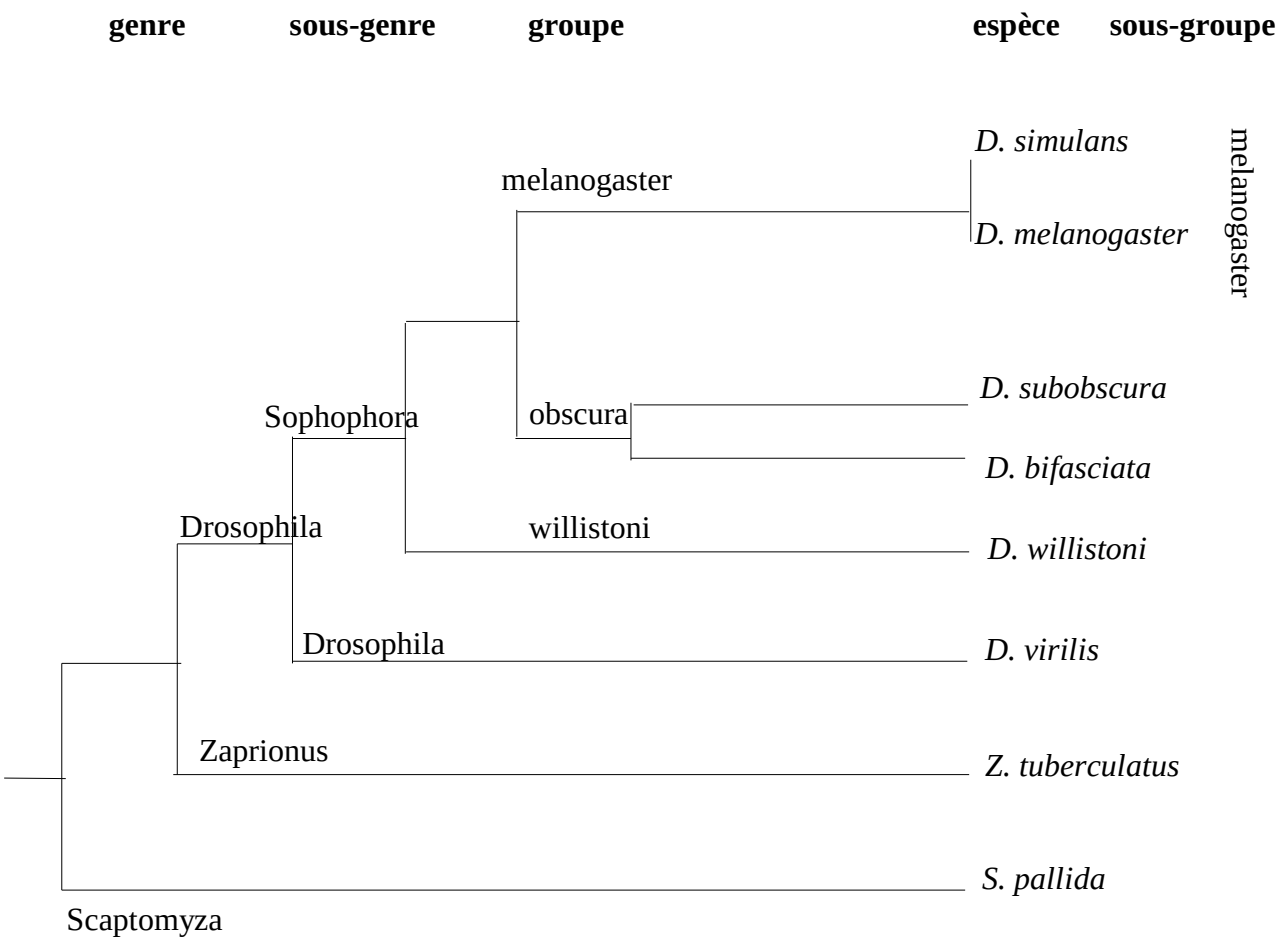


Figure 6 : phylogénie simplifiée des drosophiles pour l'étude de l'élément *P*

du code de l'élément *P* correspondait à celui des gènes de *D. willistoni*. Nous avons testé cette hypothèse en analysant l'usage du code de l'élément *P* par une AFC afin de déterminer si l'usage des codons est un bon indicateur pour la détection de transferts horizontaux d'éléments transposables. Les résultats de cette partie font l'objet de l'article 3 en annexe «*Codon usage and the origin of P elements* » paru dans Mol. Biol Evol.

2.2.1.1. Les données

D. willistoni est une espèce assez peu étudiée d'un point de vue moléculaire. Il n'existe dans les banques de données que six gènes complets ou partiels. Nous avons utilisé ces six gènes ainsi que leurs homologues chez *D. melanogaster, D. virilis, D. simulans* et *D. subobscura* (voir tableau 3 et l'article 3 en annexe). Les séquences de certains gènes étant partielles chez *D. willistoni, D. simulans et D. subobscura,* nous n'avons considéré que les régions homologues chez les espèces dans lesquelles ces gènes sont complets.

Tableau 3 : gènes et numéros d'accession dans Genbank

	D. melanogaster	***D. willistoni***	***D. virilis***	***D. simulans***	***D. subobscura***
amyrel	U69607	AF039560	/	U96159	U79724
Xdh	Y00307	AF058985*	/	/	Y08237
Adh	M11290	L08648	U26846	M36581	M55545
SOD	Y00367	L13281	X13831	X15685	U38233*
Gpdh	X80204	L37038*	D10697	AF085163*	U47877*
per	M30114	U51055*	X13877	L07829*	/

amyrel : *Amylase*-related gene, *Xdh* : Xanthine déshydrogénase, *Adh* : alcool déshydrogénase, *SOD* : Superoxide dismutase, *Gpdh* : Glycérol 3 phosphate deshydrogénase, *per* : period.

* indique les séquences partielles.

Nous avons aussi ajouté les séquences d'éléments *P* trouvées dans des espèces proches comme *D. bifasciata* (numéro d'accession : X60990), *D. subobscura* (numéro d'accession : S74793) et *S. pallida* (numéro d'accession : M63341 et M63342), ainsi que tous les éléments de classe I et de classe II décrits chez *D. melanogaster, D. virilis, D. simulans* et *D. subobscura* (voir Tableau 3' en annexe pour les noms et les numéros d'accession).

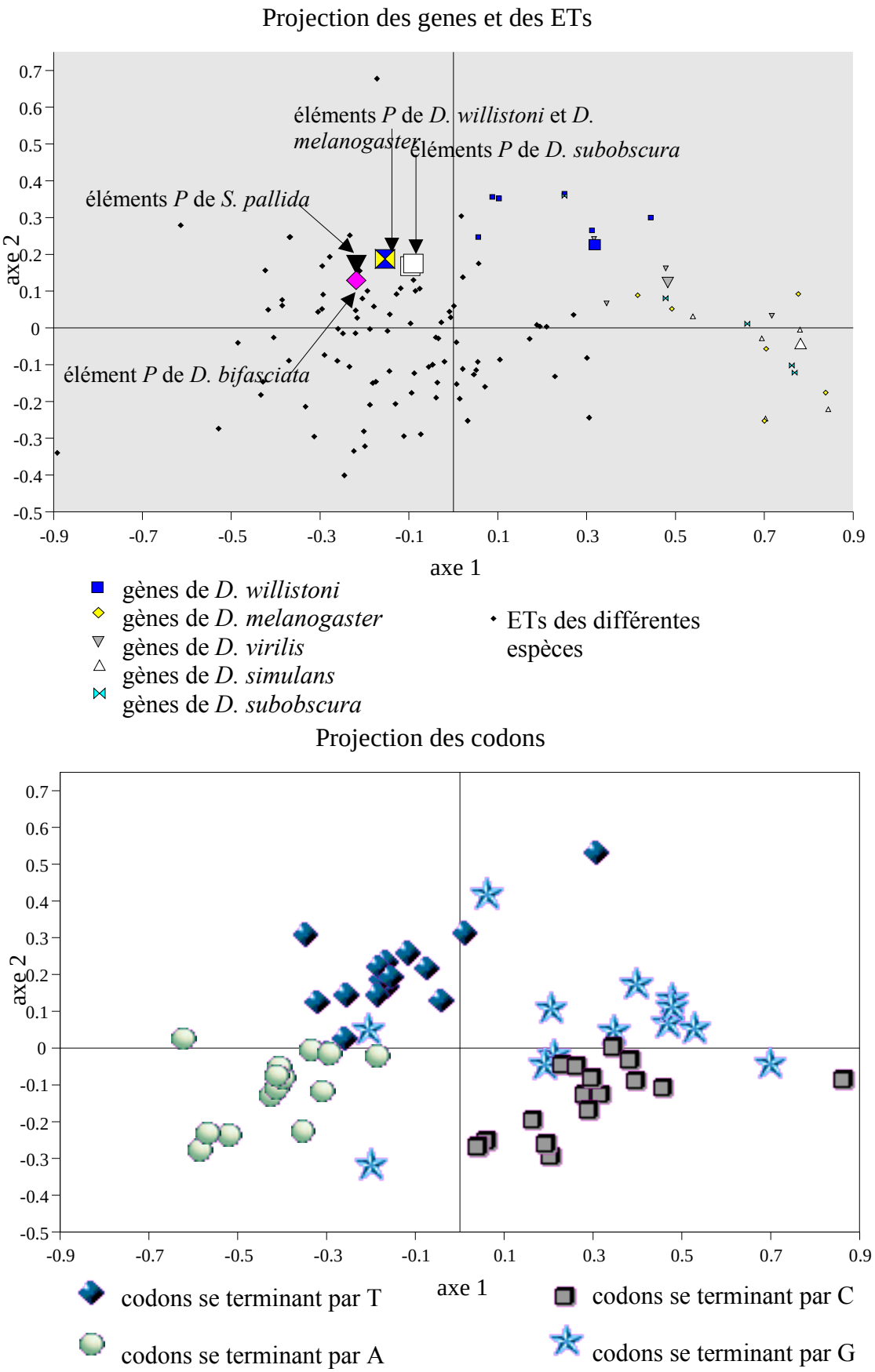


Figure 7 : Projection des gènes d'hôte, des ETs, des éléments P et des codons sur les deux premiers axes.

2.2.1.2. Les résultats

La Figure 7 présente la projection des gènes et des codons sur les deux 1ers axes de l'AFC qui représentent 44% de la variance totale (34% pour le 1er axe et 10% pour le 2ème). On observe un regroupement de tous les éléments transposables qui se séparent des différents gènes de drosophile. Les codons, projetés sur les mêmes axes, se regroupent suivant leur dernière base. Ainsi, les ETs utilisent préférentiellement les codons se terminant par A et par T. On peut constater que les gènes de *D. willistoni* sont clairement séparés des gènes de *D. melanogaster* : ces gènes utilisent préférentiellement des codons se terminant par T alors que ceux de *D. melanogaster* utilisent principalement des codons se terminant par G et par C. Concernant la position des différents éléments *P*, il est clair sur le graphe qu'ils se regroupent ensemble parmi le groupe d'ETs, et sont bien séparés des différents gènes de drosophile et en particulier de ceux de *D. willistoni*.

Le calcul du pourcentage en base au niveau de la 3ème position des codons (voir tableau 4) montre qu'en moyenne, les gènes de *D. willistoni* présentent un fort pourcentage en T_3 mais aussi un fort pourcentage en C_3, à la différence de tous les éléments *P* qui montrent un pourcentage élevé en T_3 mais aussi en A_3. Les autres espèces de drosophile montrent un fort pourcentage en C_3 et dans une moindre mesure en G_3. *D. virilis* s'en distingue en montrant de plus un pourcentage assez fort en T_3.

Tableau 4 : Fréquence en base de la 3ème position des codons (en pourcentage) des gènes et des éléments *P* pour les cinq espèces de drosophiles.

D. willistoni	16,02	32,80	29,33	21,85	48.81
D. virilis	13,73	27,06	29,80	29,41	40.78
D. simulans	4,28	13,62	52,14	29,96	17.90
D. *subobscura*	7,99	20,32	42,70	28,99	28.31
P de *D. melanogaster*	28,51	31,43	19,10	20,95	59.95
P de *D. willistoni*	28,51	31,43	19,10	20,95	59.95
P de *S. pallida*	30,61	30,91	18,25	20,23	61.52
P de *D. bifasciata*	31,96	29,05	17,77	21,22	61.01
P de *D. subobscura*	28,47	28,39	20,57	22,57	56.86

les cases grisées correspondent aux pourcentages supérieurs à ce que l'on attend à l'équiprobabilité.

2.2.1.3. Conclusion

Cette analyse montre que l'usage des codons de l'élément *P* de *D. melanogaste*r ne correspond pas à celui des gènes de *D. willistoni*. En fait, il semble que tous les ETs, quelle que soit leur espèce hôte, ont des caractéristiques communes, indépendantes de leur génome hôte, qui les regroupent ensemble dans l'AFC. On peut remarquer que les gènes de *D. willistoni* utilisent des codons se terminant par T et par C. Les éléments *P* utilisent plutôt des codons se terminant par A et par T. Ainsi, l'étude de POWELL et GLEASON (1996) n'a en fait détecté que la ressemblance d'utilisation des codons se terminant par T entre les gènes de *D. willistoni* et l'élément *P*.

Cette étude suggère que l'usage des codons n'est pas un bon indicateur pour démontrer des transferts horizontaux entre espèces de drosophiles. Une étude récente effectuée sur des bactéries a confirmé qu'un usage du code ou une composition en bases atypiques ne sont pas des indicateurs fiables de transferts horizontaux (KOSKI *et al.* 2001). Il semble que les ETs et les gènes d'hôte ne subissent pas les mêmes types de contraintes.

La suite de ce travail va consister à analyser l'usage des codons d'ETs et de gènes d'hôte chez différentes espèces afin de déterminer si l'observation selon laquelle les ETs semblent avoir des caractéristiques communes est généralisable à d'autres espèces que les drosophiles.

2.2.2. Analyse de l'usage des codons des ETs et des gènes des cinq espèces *A. thaliana*, *C. elegans*, *D. melanogaster*, *H. sapiens* et *S. cerevisiae*

Les résultats suivant, exceptés les analyses effectuées espèce par espèce, font l'objet de l'article 4 en annexe « *Codon usage by transposable elements and their host genes in five species* » accepté dans J. Mol. Evol. Les données des ETs ont été légèrement modifiées : un certain nombre d'ETs ont été rajoutés mais ceci ne modifie en rien les conclusions de l'article.

2.2.2.1. Les données

Nous avons extrait à partir de la base de données GenBank les parties codantes d'éléments transposables des cinq espèces présentées au paragraphe 1.2.. Pour les rétrotransposons, les différents ORFs *gag*, *pol* et *env* sont considérés séparément. Les séquences utilisées sont, dans la mesure du possible, complètes. Dans certains cas, les éléments sont caractérisés comme des éléments actifs. Certains éléments ont été découverts par similarité. Ils peuvent être considérés comme potentiellement actifs si leur structure correspond à celle d'éléments actifs bien que la

preuve de leur activité ne soit pas faite. Enfin, d'autres sont reconnus comme étant inactifs. Le jeu de données est donc constitué de 50 séquences d'ETs pour *Arabidopsis thaliana*, 25 pour *Caenorhabditis elegans*, 32 pour *Saccharomyces cerevisiae*, 58 pour *Drosophila melanogaster* et 18 pour *Homo sapiens*, soit un total de 183 séquences d'ETs toutes espèces confondues (voir Tableau 4' en annexe pour les noms et les numéros d'accession).

Les parties codantes des gènes d'hôte des différentes espèces ont été extraites à partir de GenBank. Il s'agit de séquences obtenues dans la majorité des cas lors d'études de biologie moléculaire et qui ont des fonctions bien définies. L'intérêt d'utiliser des gènes parfaitement caractérisés est de s'affranchir de biais artificiels qui pourraient apparaître avec des gènes mal prédits pour lesquels il y a des erreurs dans la prédiction des parties codantes. En effet, les différentes techniques existantes pour la prédiction de gènes n'ont pas des résultats fiables à 100%. C'est pour cette raison que nous n'avons pas utilisé les gènes prédits dans le cadre des génomes complets. Au total, nous avons donc sélectionné 1560 gènes pour *A. thaliana*, 592 pour *C. elegans*, 300 pour *D. melanogaster*, 300 pour *H. sapiens* et 1569 pour *S. cerevisiae*, pour lesquels la fonction est parfaitement connue.

2.2.2.2. Les résultats

2.2.2.2.1. Espèce par espèce

Nous avons effectué des analyses factorielles sur les différentes espèces séparément. Nous avons utilisé 300 gènes d'hôte tirés au sort dans les pools en contenant plus de 300, et les différentes séquences d'ETs précédemment décrites pour chacune des cinq espèces. Le tableau suivant résume pour chaque analyse la valeur de l'inertie totale ainsi que les pourcentages de variance expliquée des quatre premiers axes.

Tableau 5 : inertie totale et valeurs propres pour les cinq AFC

espèces	**inertie totale**	**pourcentage d'inertie**			
		axe 1	**axe 2**	**axe 3**	**axe 4**
A. thaliana	24.22%	13.83	7.10	5.20	5.00
C. elegans	26.74%	26.45	9.43	5.36	4.20
D. melanogaster	27.63%	40.76	7.16	4.88	3.68
H. sapiens	29.79%	36.76	4.65	4.37	3.68
S. cerevisiae	27.61%	22.75	9.06	6.13	5.18

On observe que le 1er axe pour chaque espèce possède un pourcentage de variance très élevé par rapport au 2ème. Ceci indique l'axe 1 est le facteur regroupant le maximum d'informations sur le nuage de points. Nous avons donc représenté les analyses sur les deux premiers axes. La Figure 8 montre les projections pour les gènes et les ETs, et les projections des codons sur les mêmes axes. On observe dans tous les cas que les ETs se regroupent. C'est particulièrement net chez la drosophile, l'homme et la levure.

La projection des codons montre dans le cas d'*Arabidopsis*, qu'il y a un groupe formé par les codons se terminant par A nettement séparé d'un groupe formé par les codons se terminant par C, les ETs d'*Arabidopsis* semblant « éviter » ces derniers. Chez *H. sapiens*, *C. elegans* et *D. melanogaster*, la projection des codons montre un groupement des codons se terminant par A et par T nettement séparés des codons se terminant par C et par G. On retrouve cette tendance chez la levure avec, de plus, une séparation entre codons se terminant par C et par G. Dans le cas de *A. thaliana*, les codons se terminant par C se séparent de l'ensemble des autres codons.

Dans tous les cas, les ETs des différentes espèces utilisent peu les codons se terminant par C et par G. Le regroupement des ETs n'est pas dû à des similarités au niveau des séquences. En effet, on peut observer que les différentes ORFs d'un même élément se regroupent entre elles. La Figure 9 montre la projection des ETs de *D. melanogaster* sur les mêmes axes que la Figure 8 pour cette espèce. Nous observons que d'une manière générale, les ORFs d'un même ET sont regroupés. Cependant, on peut noter dans le cas des éléments *17.6* et *nomad* que les gènes d'enveloppe de ces éléments ne se regroupent pas avec les deux autres gènes *gag* et *pol*. Les gènes *env* de ces deux éléments étant potentiellement actifs puisqu'ils présentent toutes les structures observées chez les éléments *gypsy* et *ZAM* (Terzian *et al*. 2001), il pourrait s'agir d'un indice permettant de montrer que ces éléments ont acquis un gène *env* récemment. On remarque aussi que le gène *gag* de l'élément *297* n'est pas regroupé avec les gènes *pol* et *env*. Les gènes *pol* et *env* de *297* sont proches du gène *env* de *17.6* alors que le gène *gag* de *297* est proche des gènes *gag* et *pol* de *17.6*. Peut être avons nous affaire à des éléments composites. Des études phylogénétiques ont montré que ces éléments sont très proches (Xiong et Eickbush 1990 ; Eickbush 1994 ; Lerat et Capy 1999). D'une manière générale, nous n'observons pas de regroupement des ORFs codant pour des fonctions similaires appartenant à des éléments différents. Ainsi, ce que nous observons provient de la caractéristique des éléments.

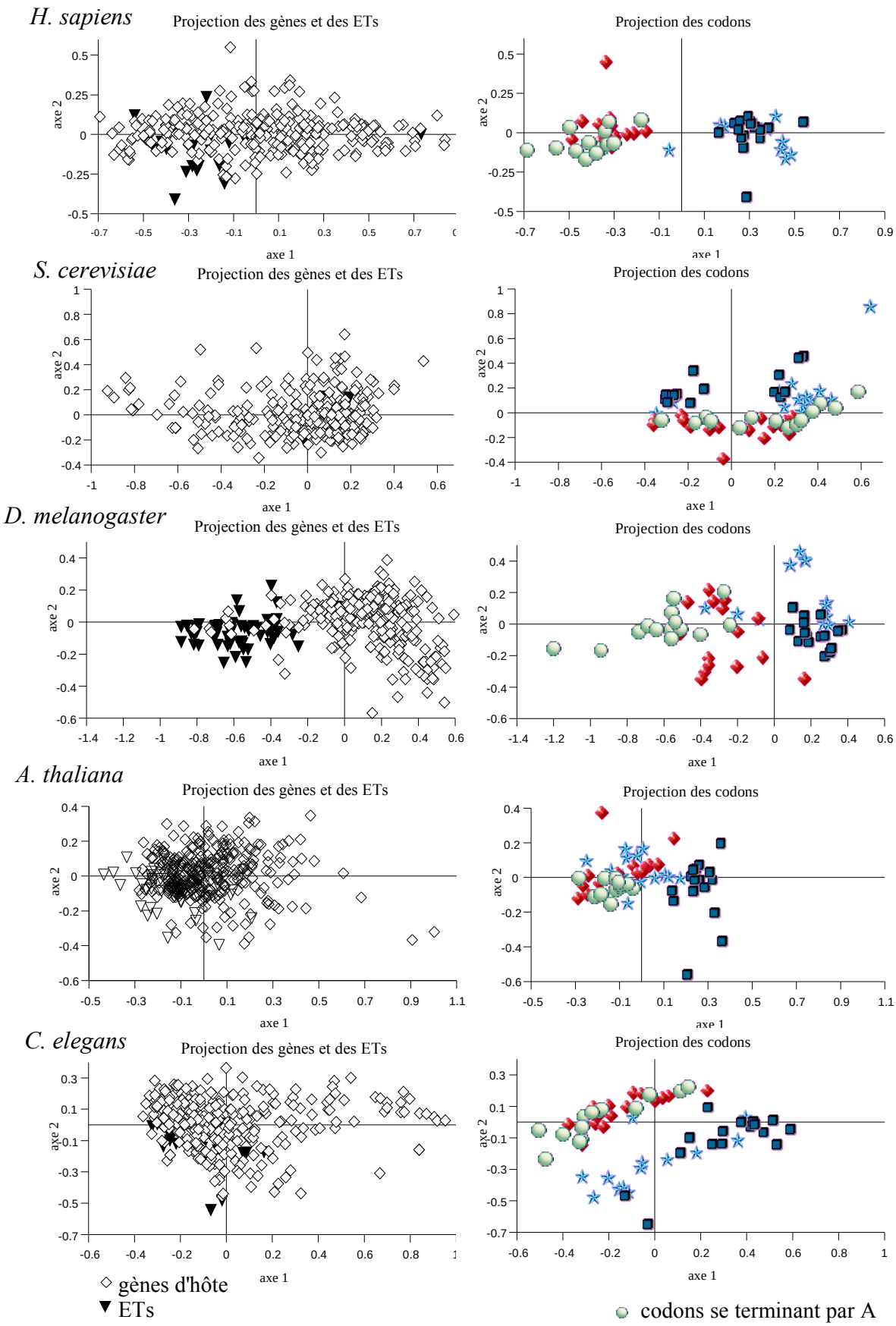


Figure 8 : Projections des AFC effectuées sur les fréquences des codons des gènes et des ETs des cinq espèces.

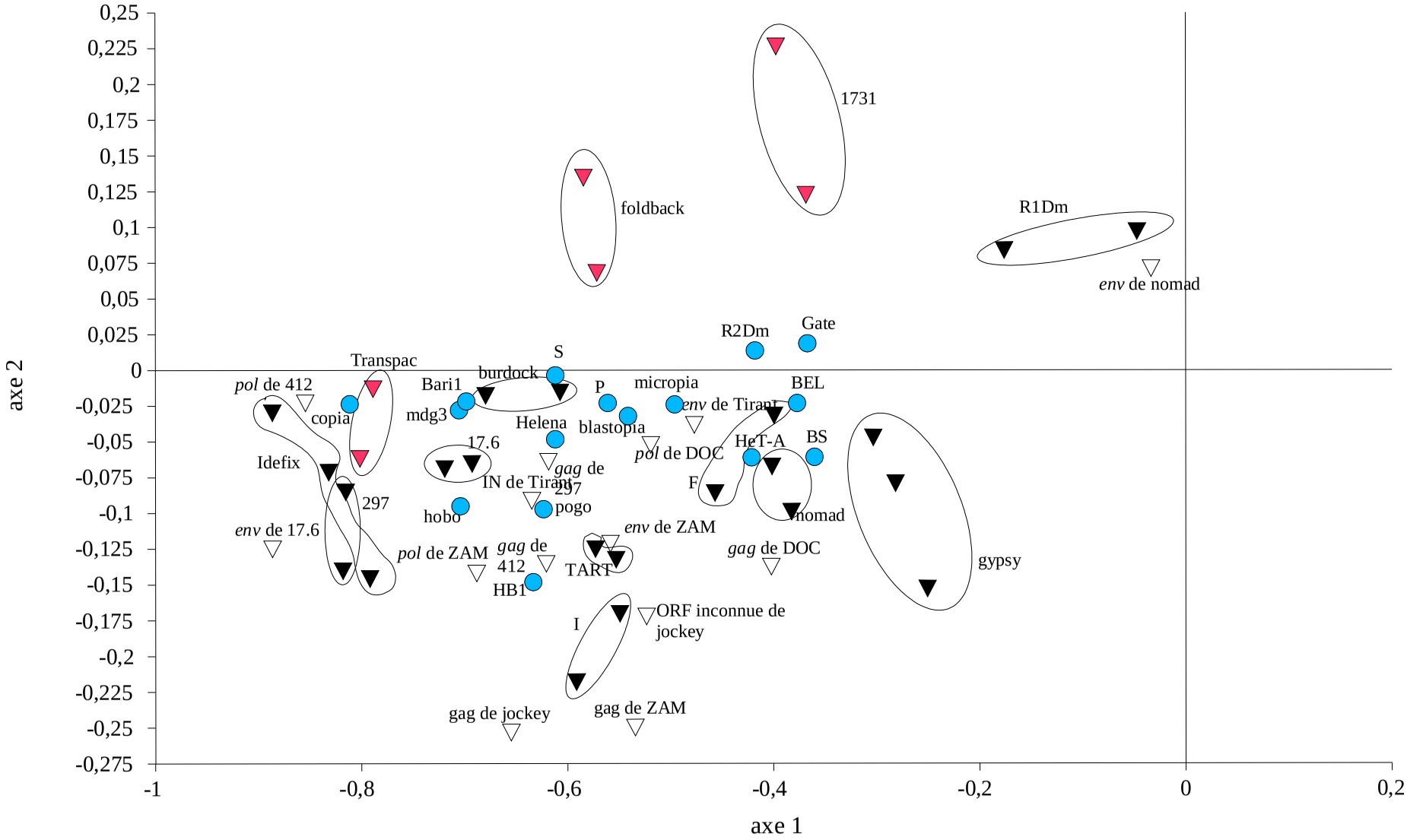


Figure 9 : Projection des ORFs des ETs seuls de *D. melanogaster* de la Figure 8

2.2.2.2.2. Les cinq espèces

L'AFC sur les ETs et les gènes d'hôte des cinq espèces va nous permettre d'observer à la fois le comportement des ETs par rapport aux gènes de tous les hôtes mais aussi d'étudier le comportement des ETs entre eux suivant leur espèce hôte. En effet, l'étude sur l'élément *P* a montré que les ETs de différentes espèces de drosophiles présentent des caractéristiques communes. Ceci peut être dû au fait que les espèces étudiées sont assez proches phylogénétiquement. C'est pour cette raison que nous avons choisi des espèces assez éloignées pour déterminer si on retrouve des caractéristiques communes entre ETs de ces espèces. Cette partie fait l'objet de l'article 4 «*Codon usage by transposable elements and host genes in five species*» présenté en annexe et accepté dans J Mol Evol. Nous allons en rappeler les principaux résultats.

2.2.2.2.2.1. L'AFC

Afin d'éviter des biais au niveau de l'AFC, nous avons constitué un échantillon de gènes d'hôte de taille comparable à celle des ETs pour chaque espèce. Ainsi, parmi tous les gènes rapatriés, nous avons tiré au sort 43 gènes pour chaque espèce (voir article 4 en annexe). Ceci nous donne au total 43 * 5 = 215 gènes d'hôte pour l'ensemble des cinq espèces. Nous avons donc constitué une matrice de données de 215 gènes d'hôte + 183 séquences d'ETs pour les lignes, croisées avec 59 colonnes correspondant aux 59 codons synonymes. Le tirage au sort des 43 gènes d'hôte a été répété 10 fois. Les AFC que l'on obtient sont similaires, c'est pourquoi nous ne parlerons ici que de la première analyse effectuée.

La tableau suivant donne le pourcentage de variance expliquée pour les quatre premiers axes obtenus et l'inertie totale. On peut observer que le 1er axe est particulièrement important par rapport aux trois autres.

Tableau 6 : pourcentage de variance expliquée pour quatre axes et pourcentage d'inertie totale de l'AFC effectuée sur la fréquence des 59 codons synonymes de 215 gènes d'hôte et 183 séquences d'ETs

% inertie totale	***% de variance expliquée***			
	axe 1	***axe 2***	***axe 3***	***axe 4***
33.07	30.34	7.00	6.14	5.21

Les Figures 10A et 10B montrent les projections sur les deux premiers axes des séquences et des codons. On observe que les ETs se regroupent, quelle que soit leur espèce hôte. La projection des codons montre qu'ils se regroupent suivant la dernière base du codon. Ainsi, les codons se terminant par A et par T se regroupent et se distinguent des codons se terminant par C et par G. Par conséquent, les ETs utilisent préférentiellement les codons se terminant par A et T. ll y a cependant quelques exceptions qui ressortent sur le graphe : les deux ORFs de l'élément *R1* et le gène *env* de *nomad,* de *D. melanogaster,* et l'élément *Tramp* d'*H. sapiens,* sont très éloignés des autres ETs et montrent une préférence pour les codons se terminant par G et par C (voir Figure 10A). Nous avons regardé la contribution absolue de chaque séquence sur le 1er axe. Les valeurs sont données dans le Tableau 7. Il est intéressant de noter que les séquences qui contribuent le plus à la formation de l'axe 1 sont des gènes d'*Homo sapiens* et de *Drosophila melanogaster*. Ces gènes sont particulièrement riches en GC. Ainsi, les différents regroupements que l'on observe sont dus à des différences de composition globale en base. L'axe 1 représente donc les différences de richesse en AT et en GC des différentes séquences analysées.

Afin de s'affranchir du biais de composition en bases, nous avons représenté les projections sur les axes 2 et 3. Les Figures 10C et 10D montrent les graphiques correspondant aux projections des séquences et des codons. Il y a, à nouveau, un regroupement des ETs. La direction du nuage de points des ETs est intéressante. En effet, elle est perpendiculaire à celle du nuage des gènes d'hôtes. La projection des codons montrent à nouveau un regroupement selon la 3ème base des codons. Ainsi, les codons se terminant par A et par C se regroupent et se séparent des codons se terminant par T et par G. Ainsi, le nuage de points des ETs montrent deux régions : l'une regroupant des ETs utilisant des codons se terminant par A et par C, l'autre utilisant principalement des codons se terminant par T. Pour clarifier la représentation, nous avons représenté sur la Figure 11 cette projection en remplaçant les différents symboles par des ellipses. Chacune des ellipses regroupent 90% des points de chaque catégorie et sont significativement différentes les unes des autres par MANOVA. On voit nettement que les directions des ellipses des ETs sont parallèles entre elles et perpendiculaires à celles des gènes d'hôtes. On peut cependant constater que les ETs ont une tendance à se regrouper suivant leur espèce hôte. Ainsi, les gènes d'*Arabidopsis* utilisent préférentiellement les codons se terminant par T et G alors que les autres espèces montrent en plus une utilisation des codons se terminant par C. Cette tendance se retrouve chez les ETs d'*Arabidopsis* mais pas chez les autres ETs.

2.2.2.2.2.2. Composition en bases et niveau d'expression

Le calcul de la composition en bases des gènes d'hôte et des ETs suivant les différentes

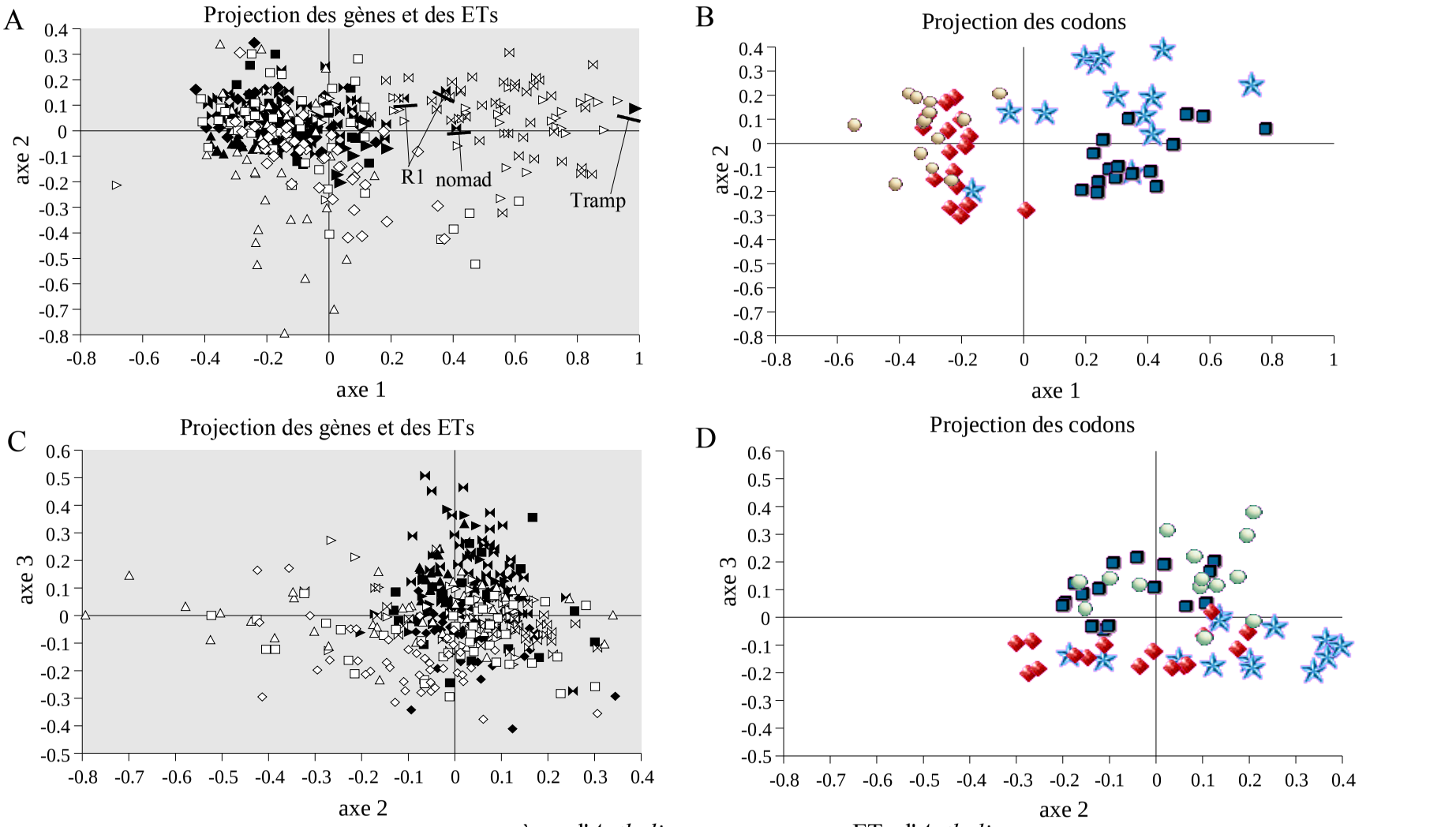


Figure 10 : Projection des gènes d'hôte, des ETs (A et C) et des codons (B et D) sur les trois premiers axes de l'AFC.

noms	axe1	noms	axe1	noms	axe1	noms	axe1	noms	axe1	noms	axe1	noms	axe1	noms	axe1	noms	axe1	noms	axe1
At	8	Ce	1	Sc	12	Dm	4	Hs	0	AtRE1	8	romani-5	7	Ty1-YD8142B-TyB	19	BEL	0	MLE-2178	0
At	9	Ce	30	Sc	39	**Dm**	**81**	Hs	14	AtRE1-63	2	romani-1	2	Ty1-YD8142B-TyA	11	S	6	MLE-2179	8
At	0	Ce	14	Sc	0	**Dm**	**110**	**Hs**	**77**	AtRE2-64	4	RT1	0	Ty2-ch3-TyA	23	Helena	6	MLE-2217	0
At	7	Ce	28	Sc	9	Dm	30	Hs	0	AtRE1-65	2	RT9b	4	Ty2-ch3-TyB	29	burdock.gag	16	humar1pcr	12
At	2	Ce	1	Sc	5	**Dm**	**180**	**Hs**	**141**	AtRE2-1-66	2	RT2	0	Ty2-ch2-TyA	33	burdock.pol	5	humar1g1	9
At	5	Ce	0	Sc	13	**Dm**	**167**	Hs	13	AtRE2-2-67	2	RT3	0	Ty2-ch2-TyB	22	DOC.gag	0	Tigger1.transposase	5
At	4	Ce	21	Sc	13	Dm	39	Hs	0	del-At	10	RT4	0	Ty2-ch15-TyA	24	DOC.pol	0	Tigger1.inc	4
At	11	Ce	31	Sc	5	Dm	39	Hs	10	Tag1	31	Tc1-transposase	1	Ty2-ch15-TyB	32	ZAM.gag	0	mariner2	23
At	4	Ce	5	Sc	0	**Dm**	**93**	**Hs**	**135**	Tag2	39	Tc1-represseur	4	Ty3-1A-gag	9	ZAM.pol	0	mariner1	1
At	1	Ce	2	Sc	4	**Dm**	**94**	Hs	7	Tat1-1	5	Tc1-Bristol	16	Ty3-1A-pol	5	ZAM.env	7	Hsmar1	0
At	0	Ce	0	Sc	6	**Dm**	**129**	Hs	48	Tat1-2	2	Tc4v	0	Ty3-2A-TyA	8	F.gag	1	**Tramp**	**245**
At	7	Ce	16	Sc	22	Dm	16	**Hs**	**77**	Tat1-3	2	T1-2	30	Ty3-2A-TyB	2	F.pol	0		
At	1	Ce	0	Sc	12	Dm	45	**Hs**	**165**	Athila ORF1	0	Tc2-RF4	3	Ty3-2-ORF3	23	gypsy.gag	4		
At	0	Ce	2	Sc	29	**Dm**	**65**	Hs	0	Athila ORF2	2	Tc2-RF1	23	Ty3-2-9910-TyA	10	gypsy.pol	8		
At	3	Ce	41	Sc	14	**Dm**	**79**	**Hs**	**189**	Ta11-1.F1	7	Tc2-RF5	1	Ty3-2-9910-TyB	2	gypsy.pol	8		
At	7	Ce	7	Sc	22	**Dm**	**80**	Hs	46	Ta11-1.F2	11	Tc2-RF2	3	Ty4-TyA	36	I.gag	1		
At	0	Ce	15	Sc	7	**Dm**	**102**	Hs	0	Ta12	14	Tc2-RF6	7	Ty4-TyB	30	I.pol	0		
At	4	Ce	8	Sc	0	Dm	8	Hs	84	Ta13	10	Tc2-RF3	6	Ty4-chX-TyA	36	jockey.inc	2		
At	0	Ce	51	Sc	13	**Dm**	**61**	Hs	0	Ta14	16	Tc5	2	Ty4-chX-TyB	32	jockey.gag	6		
At	1	Ce	3	Sc	10	Dm	12	Hs	82	Ta15	4	Cer1	22	Ty5-ch3-ORF1	10	Tirant-IN	7		
At	20	Ce	1	Sc	0	Dm	38	**Hs**	**129**	Ta18	20	Rte-1	1	Ty5-ch3-ORF2	15	Tirant-env	0		
At	1	Ce	7	Sc	14	**Dm**	**90**	Hs	4	Ta19	3	sam1	0	BARI1	17	297.gag	28		
At	0	Ce	2	Sc	15	Dm	45	Hs	0	Ta20	14	sam2	0	blastopia	2	297.pol	24		
At	1	Ce	0	Sc	29	Dm	35	**Hs**	**135**	Ta21	8	sam3	1	copia	32	297.env	1		
At	1	Ce	3	Sc	1	**Dm**	**115**	**Hs**	**197**	Ta22	6	sam4	2	micropia	1	BS	0		
At	19	Ce	8	Sc	13	Dm	27	**Hs**	**177**	Ta23	10	sam5	9	foldback.F1	4	nomad.gag	0		
At	34	Ce	38	Sc	9	Dm	85	**Hs**	**119**	Ta24	14	sam6	11	foldback.F2	5	nomad.pol	42		
At	0	Ce	2	Sc	1	**Dm**	**172**	**Hs**	**76**	Ta25	14	sam7	0	HeT-A	1	nomad.env	5		
At	5	Ce	0	Sc	24	**Dm**	**172**	**Hs**	**65**	Ta26	14	sam8	1	hobo	16	P	8		
At	3	Ce	35	Sc	0	**Dm**	**53**	Hs	0	Ta27	46	sam9	8	17.6.gag	14	pogo	5		
At	14	Ce	6	Sc	26	Dm	49	Hs	40	Ta28	14	frodo1	16	17.6.pol	11	Idefix.gag	25		
At	29	Ce	89	Sc	9	**Dm**	**106**	Hs	10	Ta1-3	10	frodo2	14	17.6.env	36	Idefix.pol	24		
At	1	Ce	3	Sc	3	**Dm**	**72**	**Hs**	**102**	Ta1-1	7	Ty1-H3-TyA	13	mdg3	14	Idefix.env	38		
At	0	Ce	33	Sc	17	**Dm**	**84**	Hs	0	Ta1-2-K-ORF1	9	Ty1-H3-TyB	22	R1-gag	35	Transpac.gag	26		
At	10	Ce	4	Sc	14	**Dm**	**109**	Hs	2	Ta1-2-K-ORF2	19	Ty1-9532-TyB	21	R1-pol	14	Transpac.pol	26		
At	20	Ce	25	Sc	3	**Dm**	**166**	Hs	10	Ta1-2-K-ORF3	7	Ty1-9532-TyA	25	R2	0	Gate.pol	1		
At	0	Ce	1	Sc	4	**Dm**	**59**	**Hs**	**74**	Ta1-2-L-ORF1	8	Ty1-9827-TyA	12	412.F3	7	L1.gp40	0		
At	4	Ce	56	Sc	0	Dm	44	Hs	4	Ta1-2-L-ORF2	13	Ty1-9827-TyB	20	412.F4	35	L1.gp150	1		
At	3	Ce	42	Sc	0	**Dm**	**132**	**Hs**	**141**	ATLN1	26	Ty1-inactif	4	HB1	12	LINE1.gp40	0		
At	2	Ce	14	Sc	11	Dm	10	**Hs**	**206**	ATLN3	2	Ty1-YD8142A-TyA1	32	1731.gag	0	LINE1.gp150	1		
At	11	Ce	18	Sc	17	**Dm**	**134**	Hs	0	ATLN4	21	Ty1-YD8142A-TyB1	21	1731.pol	0	L1-RT	1		
At	22	Ce	15	Sc	13	**Dm**	**112**	Hs	48	romani-2	6	Ty1-YD8142A-TyB2	19	TART.gag	1	MLE-2173	3		
At	9	Ce	0	Sc	24	**Dm**	**90**	**Hs**	**146**	romani-3	0	Ty1-YD8142A-TyA2	11	TART.pol	1	MLE-2177	0		

Tableau 7 : Contributions des lignes sur le 1er axe.
En gras sont représentées les contributions supérieures à 50.
At : gènes d'hôte d'Arabidopsis, Ce : gènes d'hôte de Caenorhabditis, Sc : gènes d'hôte de Saccharomyces, Dm : gènes d'hôte de Drosophila, Hs : gènes d'hôte de l'homme.

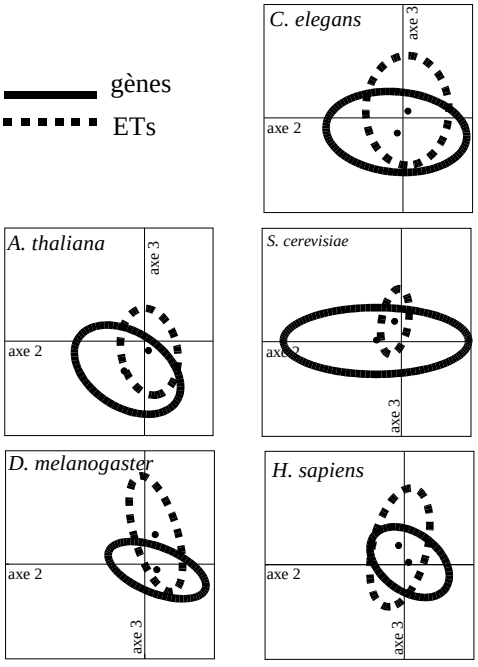


Figure 11 : Projection des éléments transposables et des gènes d'hôte selon l'espèce hôte sur les mêmes axes que la Figure 10C.

positions dans les codons nous amène des informations supplémentaires sur les contraintes des ETs. Le Tableau 8 (correspondant au Tableau 1 de l'article 4 en annexe) montre la composition en AT aux trois positions dans le codons et globalement suivant les classes d'ETs et pour les gènes d'hôtes. Nous avons indiqué les valeurs pour les rétrovirus de *H. sapiens* que l'on peut considérer comme des éléments de classe I. Le pourcentage en AT total montre bien que les ETs, quelle que soit leur espèce hôte, sont riches en AT. On constate aussi que les gènes de la drosophile et de l'homme sont en moyenne moins riches en AT que les gènes des trois autres espèces.

Au niveau des ETs, on remarque que le pourcentage en AT au niveau de la 3ème position des codons est plus fort que celui au niveau de la 1ère position. Cette tendance est aussi observée dans les gènes d'*Arabidopsis*, de *Caenorhabditis* et de *Saccharomyces* mais pas chez les deux autres espèces qui sont principalement GC riches à cette position. Cela confirme que la richesse en AT à cette position est une caractéristique des ETs, indépendamment du génome hôte.

Chez *Arabidopsis*, *Caenorhabditis*, *Saccharomyces* et *Drosophila*, il a été démontré que le biais d'usage des codons est lié au taux d'expression des gènes (Ikemura 1982 ; Shields *et al.* 1988 ; Stenico *et al.* 1994 ; Duret et Mouchiroud 1999). Afin de tester s'il peut y avoir un lien entre le niveau d'expression des ETs et le biais d'usage des codons, nous avons caractérisé les gènes d'hôte de l'étude suivant leur niveau d'expression (voir matériel et méthode de l'article 4 en annexe). Nous avons ensuite déterminé les fréquences d'utilisation des codons synonymes pour chaque acide aminé pour l'ensemble des ETs, pour des gènes fortement exprimés et pour des gènes faiblement exprimés d'une espèce. Le Tableau 9 résume les résultats obtenus (présentés dans le Tableau 2 de l'article 4 en annexe).

Tableau 8 : composition en A+T aux trois positions des codons pour les ETs et les gènes d'hôte et des zones non contraintes

	nombre de sequences	***1ère position***	***2ème position***	***3ème position***	***total***
A. thaliana					
Rétrotransposons à LTR	41	52.00	58.50	57.60	**56.03**
Rétrotransposons sans LTR	22	55.41	61.13	60.45	**59.00**
Transposons	2	53.67	63.46	68.88	**62.00**
gènes nucléaires	13859	49.72	59.76	57.33	**55.60**
zones non codantes	/	/	/	0.00	**68,00[$]**
C. elegans					
Rétrotransposons à LTR	6	51.07	63.58	57.01	**56.83**
Rétrotransposons sans LTR	13	58.31	61.64	58.87	**59.60**
Transposons	10	52.54	61.32	58.32	**57.38**
gènes nucléaires	14425	50.90	61.41	59.74	**57.35**
zones non codantes	/	/	/	/	**70,00[#]**
S. cerevisiae					
Rétrotransposons à LTR	32	56.77	63.44	66.23	**62.16**
gènes nucléaires	6301	55.39	63.00	60.03	**59.67**
zones non codantes	/	/	/	/	**65,00[$]**
D. melanogaster					
Rétrotransposons à LTR	31	53.99	65.35	60.01	**59.73**
Rétrotransposons sans LTR	16	51.31	57.41	53.66	**54.13**
Transposons	5	56.97	65.94	62.58	**61.83**
gènes nucléaires	14332	44.10	58.49	34.35	**45.65**
zones non codantes	/	/	/	/	**65,00[*]**
H. sapiens					
Rétrovirus	74	47.85	56.38	57.52	**53.90**
Rétrotransposons sans LTR	5	59.06	65.54	56.90	**60.50**
Transposons	12	52.03	60.96	57.30	**53.41**
gènes nucléaires	12227	44.01	57.46	42.89	**48.12**

Des tests de χ^2 ont été effectués afin de déterminer si les différences observées entre le %AT global et le %AT à chaque position sont significatives entre gènes d'hôte et ETs.

*Shields *et al.* 1988, Kliman et Hey 1994

[$] Lin *et al.* 1999

[#] Duret, communication personnelle

Tableau 9 : préférence des codons selon le taux d'expression des gènes d'hôte et selon les ETs

	A. thaliana	***C. elegans***	***D. melanogaster***	***S. cerevisiae***
gènes fortement exprimés	**codons G/C**	**codons G/C**	**codons G/C**	**codons G/C**
gènes faiblement exprimés	codons A/T	codons A/T	**codons G/C**	codons A/T
ETs	codons A/T	codons A/T	codons A/T	codons A/T

Chez l'arabette, le nématode et la levure, un test de signes permet de montrer que les ETs utilisent les mêmes codons que les gènes faiblement exprimés, c'est-à-dire les codons se terminant par A et T. Cependant on ne retrouve pas cette tendance chez la drosophile où les gènes faiblement exprimés utilisent des codons se terminant par G ou C, tout comme les gènes fortement exprimés.

2.2.2.3. Conclusion

Notre étude montre que les ETs ont globalement une caractéristique générale, indépendante de leur espèce hôte : leur richesse en AT. Celle-ci se répercute sur l'usage des codons. Ainsi, les ETs utilisent préférentiellement des codons se terminant par A et par T. Cette tendance est retrouvée chez les gènes faiblement exprimés de *A. thaliana, S. cerevisiae* et *C. elegans* mais pas chez *D. melanogaster*. Il semble donc que l'utilisation des codons se terminant par A et T chez les ETs soit liée à patron d'expression spécifique.

L'observation plus précise de la composition en bases suivant les positions dans le codon montre que les ETs sont plus riches en AT au niveau de la 3ème base qu'au niveau de la 1ère base. Cette tendance se retrouve, quelle que soit l'espèce hôte. Chez *D. melanogaster* et *H. sapiens*, les gènes nucléaires montrent la tendance inverse : la 3ème position des codons est moins riches en A+T que la 1ère. Ceci confirme que la tendance que l'on trouve chez les ETs est bien une caractéristique à part, indépendante de l'hôte.

Chez *C. elegans, S. cerevisiae, D. melanogaster* et *A. thaliana*, le pourcentage en AT des

régions non codantes a été déterminé dans la littérature. Celui-ci est donné dans le tableau 8. La composition des régions non codantes est le reflet des biais mutationnels qui tendent à enrichir les séquences en AT, ces régions ne subissant pas de pressions de sélection. On peut comparer la 3ème base des codons des gènes avec le pourcentage observé des zones non codantes puisque cette position dans les codons n'est pas contrainte. Ainsi, les ETs de *Caenorhabditis*, *Drosophila* et *Arabidopsis* montrent un pourcentage en AT au niveau de la 3ème base significativement inférieur à celui que l'on rencontre dans les zones non codantes, à l'exception des transposons d'*Arabidopsis* et de *Drosophila*, ainsi que les rétrotransposons à LTR de *Caenorhabditis*, probablement dû au faible effectif des échantillons pour ces classes. Ceci indique que les ETs ne subissent pas uniquement des « patterns » de mutations comme le supposaient SHIELDS et SHARP (1989) mais sont aussi soumis à des pressions de sélection. Chez la levure, le pourcentage de la 3ème base des codons des ETs est statistiquement plus élevé que celui les zones non codantes.

On peut se demander quels sont les facteurs qui permettent aux ETs d'avoir ces caractéristiques. Il peut s'agir des résultats d'interactions avec les génomes hôtes, ou avec des facteurs environnementaux au niveau de leur site d'insertion, ou encore de leur capacité intrinsèque à bouger, ou du mode de transposition, ou bien encore de leur origine.

2.2.3. Dégénérescence des ETs

Notre étude montre que les ETs semblent être sujets à des pressions de sélection et à des biais mutationnels qui leur confèrent une richesse en AT. Cette caractéristique est généralisable à l'ensemble des ETs quelle que soit leur espèce hôte. Le mécanisme impliqué peut donc être relatif à leur activité. En analysant des ETs d'une même famille mais plus ou moins actifs, on doit pouvoir déterminer si l'activité ou la perte de l'activité entraîne une modification dans l'usage des codons et dans la richesse en AT. L'organisme idéal pour cette analyse serait *Drosophila melanogaster* qui possède un grand choix d'ETs de familles différentes. Malheureusement, le génome complet publié de la drosophile ne comporte pour l'instant que des séquences consensus et mal définies des ETs. Il est donc impossible de faire des analyses concernant cette espèce actuellement. Par contre, les ETs du génome de *Saccharomyces* qui ont été recensés et annotés peuvent faire l'objet d'un telle analyse (KIM *et al.* 1998).

2.2.3.1. Les données

Les ETs de *Saccharomyces* ont été récupérés à partir du génome complet en utilisant la base

de données « Transposable elements ressource » qui donne les positions des ETs sur les 16 chromosomes de la levure (www.public.iastate.edu/~voytas/resources/resources.html). La levure possède uniquement des rétrotransposons à LTR subdivisés en cinq familles. Les familles *Ty1* et *Ty2* sont les plus représentées. Nous nous sommes donc intéressés aux ETs de ces familles. Ces éléments possèdent deux ORFs, TYA et TYB correspondant respectivement aux gènes *gag* et *pol*. Le génome de la levure possède 32 éléments complets *Ty1* et 13 éléments complets *Ty2*. Un élément *Ty1* a été identifié comme fonctionnel (Boeke *et al.* 1988). Cet élément, *Ty1H3,* peut donc être utilisé comme référence par rapport aux autres éléments étudiés. Parmi les éléments complets du génome, sept ont été identifiés comme étant inactifs soit par des frameshifts, soit par des délétions, soit par des duplications (Kim *et al.* 1998). Le Tableau 10 donne les noms de ces éléments. De plus, trois éléments de type *Ty1* sont définis comme les représentants d'une nouvelle sous-famille *Ty1*', il s'agit des éléments YNLCTy1-1, YBLWTy1-1 et YMLWTy1-3.

Tableau 10 : noms et mutations des éléments *Ty* connus comme inactifs du génome de la levure

noms	***ORF affectée***	***type de mutation***
YARCTY1-1	TYA	frameshift
YLRWTy1-4	TYB	frameshift
YDRCTy1-3	TYB	délétion
YHLCTy1-1	TYB	duplication
YGRCTy2-1	TYB	frameshift
YNLCTy2-1	TYA	frameshift
YLRCTy2-2	TYB	délétion

2.2.3.2. Les résultats

Une AFC a été effectuée sur les fréquences relatives des deux ORFs de chaque séquence d'ETs. Le tableau suivant donne le pourcentage de variance expliquée par chaque axe ainsi que le pourcentage d'inertie totale.

Tableau 11 : pourcentage de variance expliquée pour trois axes et pourcentage d'inertie totale de l'AFC effectuée sur la fréquence des 59 codons synonymes de 95 ORFs de *Ty1* et *Ty2*.

% inertie totale	***% variance expliquée***		
	axe 1	axe 2	axe 3
5.89	47.64	28.2	6.95

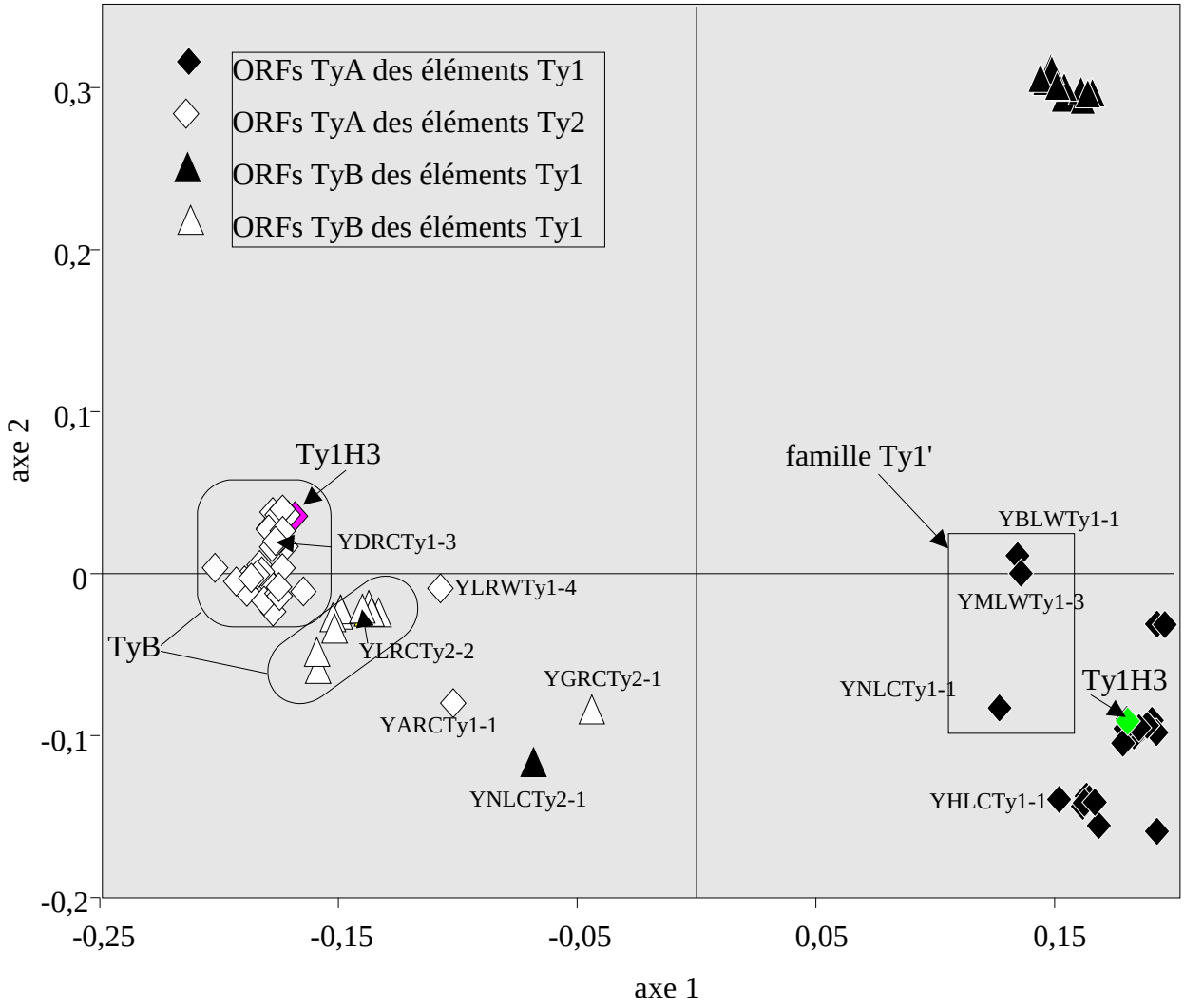


Figure 12 : Projection des ORFs des éléments *Ty1* et *Ty2* du génome de la levure.

Les deux premiers axes de l'AFC expliquent 76% de la variance totale. La projection sur ces deux axes est représentée sur la Figure 12. Quatre groupes principaux ressortent sur le graphique. Ils correspondent aux diverses ORFs des deux principales familles : les gènes TYA (losanges) se regroupent suivant la famille *Ty1* ou *Ty2*, de même pour les gènes TYB (triangles). Les gènes TYA sont plus divergents entre les deux familles que les gènes TYB.

La position des gènes de l'élément actif *Ty1H3* est indiquée. On observe que la majorité des gènes des autres éléments *Ty1* se positionnent à proximité des gènes de *Ty1H3*. Il y a plusieurs exceptions : ainsi, parmi les éléments décrits dans le Tableau 10 certains des gènes mutés ont des positions particulières. C'est le cas pour le gène TYA de YHLCTy1-1 et des gènes TYB des éléments YLRWTy1-4 et YARCTy1-1. Ces positions particulières sont principalement dues aux changements de cadre de lecture provoqués par les mutations. Pour YLRWTy1-4 et YARCTy1-1, si on corrige le frameshift, les gènes TYB de ces éléments se positionnent dans le groupe des TYB. En ce qui concerne le pourcentage en AT des divers éléments utilisés dans l'AFC, il est très semblable d'un élément à l'autre.

2.2.3.3. Conclusion

On constate une légère différence dans l'usage des codons entre les différents éléments d'une même famille *Ty1* et *Ty2*. Cependant, les divers éléments utilisés sont particulièrement proches du point de vue séquence. Ainsi, les identités au niveau des acides nucléiques au sein d'une même famille, pour un gène donné, ne sont pas inférieurs à 90%. On a donc des séquences très proches dont certaines sont cependant inactives. Dans le génome de la levure, en dehors des éléments complets utilisés dans cette analyse, on ne trouve que des LTR-solo ou des fragments de LTR. Il n'y a donc pas d'éléments incomplets qui pourraient être plus divergents par rapport à l'élément actif. Il est possible qu'il y ait un « turn over » rapide des éléments *Ty* empêchant la persistance d'éléments très dégénérés. Ceci nous empêche de déterminer l'évolution des pressions de sélection qui semblent agir sur les ETs de la levure. Il faut attendre les éléments du génome de la drosophile pour pouvoir déterminer si la disparition rapide des éléments trop divergents observée chez la levure, se retrouve chez d'autres espèces.

2.3. Conclusion

Nos analyses montrent que les ETs ont un comportement différent des gènes de leur hôte en ce qui concerne l'usage des codons. Les ETs montrent des caractéristiques communes,

indépendantes de leur génome hôte, à savoir la richesse en AT et l'utilisation préférentielle des codons se terminant par A et par T. Ce comportement est particulièrement marqué dans les espèces comportant des gènes riches en GC. Cependant, il est possible d'observer une influence du génome hôte qui se caractérise par le regroupement préférentiel des ETs d'un même génome dans une AFC (voir Figure 11).

Certaines séquences d'ETs de *D. melanogaster* et *H. sapiens* présentent un fort taux en GC. il s'agit des deux ORFs de *R1*, du gène *env* de *nomad* chez *D. melanogaster* et de l'élément *Tramp* chez *H. sapiens*. Les cas de *R1* et *Tramp* sont très intéressants car ces éléments semblent être le produit d'une domestication par leur génome hôte. Comme exposé dans l'introduction, *R1* est un rétrotransposon sans LTR inséré dans les gènes codant pour l'ARN ribosomique 28S et qui, lorsqu'il se transpose, permet l'amplification de ces gènes (JACKUBCZAK *et al.* 1990). Ainsi, cet élément qui a été recruté par le génome hôte, remplit une fonction précise. Ceci pourrait expliquer pourquoi ses ORFs présentent un usage des codons communs aux gènes d'hôte : elles subissent les mêmes types de contraintes que les gènes. Le cas de *Tramp* est similaire. Il s'agit un transposon inséré dans la région pseudoautosomal PAR située aux extrémités du bras p des chromosomes X et Y (ESPOSITO *et al.* 1999). Il ne peut pas se déplacer mais pourrait être impliqué dans des fonctions biologiques car sa protéine présente une homologie avec un facteur d'activation de promoteur de la drosophile. La position du gène *env* de *nomad* est assez surprenante, alors que ses ORFs *gag* et *pol* sont regroupées et se placent avec les autres ETs. Il s'agit d'un rétrotransposon à LTR de type *gypsy* qui semble être proche de l'élément *yoyo* de *Ceratitis capitata* et de l'élément *gypsy* de *Drosophila melanogaster* (WHALEN et GRIGLIATTI 1998). Le gène d'*env* de *nomad* est particulièrement riche en GC, ce qui explique sa position parmi les gènes d'hôte de la drosophile. Cependant, des signatures spécifiques aux protéines d'enveloppe ont été détectées indiquant qu'il s'agit bien d'un gène d'enveloppe mais produisant une protéine plus courte que celle des autres éléments de type *gypsy* (TERZIAN *et al.* 2001). Il ne semble pour l'instant n'y avoir aucune explication sur la richesse en GC de ce gène.

Plusieurs hypothèses peuvent être évoquées pour expliquer la richesse en AT des ETs. On sait que chez les rétrovirus, la transcriptase inverse fait des erreurs lors de son action sur l'ARN viral, ce qui provoque l'incorporation préférentielle de bases A et T. Ceci a pour conséquence un enrichissement des rétrovirus en AT (ZSÌROS *et al.* 1999). Un tel phénomène pourrait se retrouver chez les rétrotransposons qui, lors de leur cycle, utilisent une transcriptase inverse. Cependant, plusieurs contre-exemples peuvent être avancés. Les éléments SINEs, qui ne possèdent pas de partie codante, sont totalement dépendants des éléments LINEs pour se déplacer (WICHMAN *et al.* 1992 ; JURKA 1997). Alors que les éléments LINEs sont riches en AT, les SINEs sont riches en GC (SCHMID 1998). Ainsi, il ne semble pas que les erreurs de la transcriptase inverse soient une condition

suffisante pour expliquer la richesse en AT des rétrotransposons. Cela n'empêche pas la possibilité que la richesse en GC des SINEs puissent être une autre opportunité pour leur multiplication étant donné que ces éléments sont non autonomes. Cependant, l'hypothèse des erreurs de la transcriptase inverse n'explique pas non plus la richesse en GC du gène *env* de l'élément *nomad*. De plus, on observe cette même richesse en AT chez les transposons qui ont un mode de transposition complètement différent. Tout ceci suggère plutôt un mécanisme commun à l'ensemble des ETs indépendamment de leur classe.

Une hypothèse alternative serait que les ETs subissent l'influence du contexte génomique de la région dans laquelle ils sont insérés. En effet, chez *Homo sapiens* et chez *Drosophila melanogaster*, une corrélation est observée entre la composition en base des gènes, l'usage des codons et la position dans le génome, qui est due aux « patterns » de biais mutationnels (Kliman et Hey 1994 ; Sharp et Matassi 1994 ; Jabbari et Bernardi 2000). D'une manière générale, il semble que les ETs ont des sites d'insertion préférentiellement composés de A et de T. Par exemple, les rétrotransposons à LTR *297* et *17.6* ont un site d'insertion de type TATAT (Ikenaga et Saigo 1982 ; Inouye *et al.* 1984). De même, le transposon *S* ainsi que tous les éléments de type *Tc1-mariner* s'insèrent au niveau de dinucléotides TA (Merriman *et al.* 1995). Cependant, cela ne semble pas être une situation généralisable à tous les ETs. En effet, élément *P* de *Drosophila melanogaster* a récemment été montré comme s'insérant dans des régions GC riches (Liao *et al.* 2000) alors qu'il s'agit d'un élément riche en AT. De même, les rétrotransposons *Tirant* (Viggiano *et al.* 1997) et *ZAM* (Leblanc *et al.* 1999) s'insèrent aussi au niveau de sites cibles riches en GC. Enfin, la richesse en GC des éléments SINE ne semble pas s'expliquer par une insertion préférentielle dans des régions riches en GC puisque les *Alus* ne semblent pas cibler particulièrement des régions riches en GC mais s'y accumulent d'avantage au cours du temps (International Human Genome Sequencing Consortium 2001).

Si l'on compare la 3ème base des codons des ETs, qui est la position non contrainte, avec le pourcentage en base des zones non codantes, on peut savoir si les ETs subissent un biais mutationnel. Le Tableau 8 montre que, excepté chez certains ETs, la 3ème position des codons des ETs est significativement moins riche en AT que les zones non codantes, ce qui indique que les ETs ne subissent pas uniquement un biais mutationnel mais qu'ils subissent aussi des pressions de sélection. Kliman et Hey (1993) ont montré qu'il existe une relation entre l'usage des codons et le taux de recombinaison chez *D. melanogaster* avec un faible biais de codon associé à un faible taux de recombinaison (voir paragraphe 2.1.2.). Marais, Mouchiroud et Duret (2001) ont montré que ceci est lié à un biais mutationnel des régions hautement recombinantes. Ainsi, le faible biais d'usage des codons observé chez les ETs pourrait être dû au fait qu'ils s'insèrent préférentiellement

dans des régions à faible taux de recombinaison. Cependant, des études effectuées chez *D. melanogaster* et *C. elegans* n'ont pas permis de mettre en évidence une corrélation claire entre la densité en ETs et le taux de recombinaison le long des chromosomes (Biémont *et al.* 1997 ; Duret *et al.* 2000). Il est possible que l'insertion des ETs se fasse principalement sur des critères de conformation de la chromatine comme chez les rétrovirus (Gama Sosa *et al.* 1989), conformations qui pourraient aussi déterminer l'usage des codons.

Les ETs peuvent avoir des effets négatifs sur leur génome hôte. Cependant, nous avons noté dans la 1ère partie d'introduction que certains génomes développent des systèmes de défense. Ainsi chez le champignon *Neurospora crassa*, le mécanisme de RIP qui permet l'inactivation des éléments de cet organisme a pour conséquence de les enrichir en AT. Chez certains organismes comme *Arabidopsis* et l'homme, la méthylation* des CpG est un moyen de réguler l'activité des gènes. Les ETs pourraient donc échapper à ce type d'inactivation en devenant riches en AT. Cependant, la méthylation* n'est pas observée dans tous les organismes. On a détecté une faible activité de méthylation* très récemment chez la drosophile (Lyko *et al.* 2000) mais elle n'existe pas chez le nématode, ni chez la levure. De plus, le grand nombre d'ETs présents dans les plantes tendrait à montrer qu'il s'agit d'un moyen peu efficace pour empêcher l'amplification des ETs.

Chez *Arabidopsis*, *Saccharomyces*, *Caenorhabditis* et *Drosophila*, le biais d'usage des codons est corrélé positivement avec le taux d'expression des gènes. Nous avons vu que les ETs d'*Arabidopsis*, *Saccharomyces* et *Caenorhabditis* semblent utiliser les mêmes codons que les gènes d'hôte peu exprimés. Cependant, cette tendance n'apparaît pas chez la drosophile. Il se peut donc que ce que l'on observe chez *Arabidopsis*, *Saccharomyces* et *Caenorhabditis* soit dû au fait que les gènes d'hôte de ces espèces sont plutôt riches en AT et ont un usage des codons qui ressemble d'avantage à celui des ETs. Ainsi, il n'y aurait pas de relation entre le taux d'expression fort ou faible des ETs avec le biais d'usage des codons. Les ETs pourraient refléter au travers de leur biais un patron d'expression qui leur est propre. La richesse en AT pourrait alors être une propriété intrinsèque des ETs liée à leur capacité à transposer ou à leur mécanisme de transposition.

3ème PARTIE : L'ABONDANCE RELATIVE EN DI- ET EN TRINUCLEOTIDES

3ème partie : L'abondance relative en di- et en trinucléotides

Les ETs montrent au niveau de l'usage des codons un comportement particulier par rapport aux gènes de l'hôte. Cependant, nous avons pu constater qu'il existe quand même une légère influence du génome hôte (voir Figure 11). Pour aller plus loin dans cette analyse, nous avons décidé d'étudier l'abondance relative en di- et en trinucléotides. Elle détermine la fréquence de chaque di- et trinucléotide en la normalisant par le contenu en GC. Cette abondance relative détermine une signature spécifique de chaque espèce. Nous avons donc voulu savoir si les ETs possèdent cette signature génomique spécifique à l'hôte ou bien s'ils possèdent une signature propre.

3.1. Définition

En 1992, BURGE *et al.* ont défini des indices permettant de caractériser des séquences génomiques selon l'abondance relative en di- ou en trinucléotides. Le calcul de ces indices permet de s'affranchir de la composition en bases des génomes et peut prendre en compte à la fois les parties codantes et les parties non codantes. Ces indices permettent de déterminer pour chaque génome une signature particulière appelée « signature génomique », qui rend compte des « patterns » spécifiques de la sur- et de la sous-représentation de chaque di- ou trinucléotide pour un génome donné.

Le calcul des indices pour les dinucléotides se fait de la manière suivante. Soit XpY un dinucléotide, on calcule l'indice ρ_{XY} tel que $$\rho_{XY} = \frac{f_{XY}}{(f_X f_Y)}$$

avec f_{XY} = la fréquence du dinucléotide XpY.

f_X = la fréquence de la base X.

f_Y = la fréquence de la base Y.

Cet indice est valable pour les séquences simple brin comme par exemple les séquences codantes. Afin de tenir compte de la nature double brin de l'ADN et de pouvoir comparer des espèces différentes ou des séquences dont l'orientation sur l'ADN génomique n'est pas connue, la formule est modifiée pour prendre en compte la structure antiparallèle de l'ADN double brin. Ainsi, pour la base A et sa base associée T, on peut calculer les fréquences des bases A et T dans la séquence double brin : $$f_A^* = f_T^* = \frac{1}{2}(f_A + f_T)$$

avec f_A et f_T les fréquences des bases A et T dans la séquence simple brin.

De même pour les bases associées G et C, on peut calculer leurs fréquences dans une séquence double brin :

$$f^*_G = f^*_C = \frac{1}{2}(f_G + f_C)$$

avec f_G et f_C les fréquences des bases G et C dans la séquence simple brin.

On peut donc calculer de manière similaire la fréquence du dinucléotide GpA dans une séquence double brin comme :

$$f^*_{GA} = \frac{1}{2} f_{GA} + \frac{1}{2} f_{TC}$$

Ainsi, l'indice d'abondance relative en dinucléotide pour une séquence double brin est calculé comme :

$$\rho^*_{XY} = \frac{f^*_{XY}}{(f^*_X f^*_Y)}$$

Les déviations des indices d'abondance relative en dinucléotides par rapport à 1 rendent compte de contrastes entre la fréquence observée des dinucléotides et celle que l'on attend dans le cas d'associations au hasard des bases pour former les dinucléotides. KARLIN et BURGE (1995) ont estimé que $\rho^*_{XY} \leq 0{,}78$ et $\rho^*_{XY} \geq 1{,}23$ représentent significativement la sous- et la sur-représentation d'un dinucléotide XpY, respectivement, dans des séquences générées aléatoirement avec une probabilité de 0,001. Les valeurs prises par ρ^*_{XY} peuvent aller de 0,20 à 1,50 (KARLIN et MRÁZEK 1997).

Dans le cas de la mesure de l'abondance en trinucléotides, on calcule de la même manière pour un trinucléotide XYZ un indice γ^*_{XYZ} tel que :

$$\gamma^*_{XYZ} = \frac{(f^*_{XYZ} f^*_X f^*_Y f^*_Z)}{(f^*_{XY} f^*_{YZ} f^*_{XNZ})}$$

avec f^*_{XYZ} = la fréquence du trinucléotide XYZ dans la séquence double brin

f^*_X = la fréquence de la base X dans la séquence double brin

f^*_Y = la fréquence de la base Y dans la séquence double brin

f^*_Z = la fréquence de la base Z dans la séquence double brin

f^*_{XY} = la fréquence du dinucléotide XpY dans la séquence double brin

f^*_{YZ} = la fréquence du dinucléotide YpZ dans la séquence double brin

f^*_{XNZ} = la fréquence du trinucléotide XNZ dans la séquence double brin avec N n'importe laquelle des 4 bases.

Ainsi, on détermine la sur- ou la sous-représentation selon la déviation de γ^* par rapport à 1 : si $\gamma^*_{XYZ} > 1$ le trinucléotide XYZ est sur-représenté et si $\gamma^*_{XYZ} < 1$, le trinucléotide XYZ est sous-représenté.

On peut calculer une distance relative entre deux séquences f et g comme la somme de la valeur absolue des différences des indices ρ^*_{ij} pour chaque dinucléotide ij entre les deux séquences (KARLIN et LADUNGA 1994 ; KARLIN et MRÁZEK 1997) :

$$\delta^*(f,g) = \frac{1}{16}\sum_{ij}\left|\rho^*_{ij}(f) - \rho^*_{ij}(g)\right|$$

Une valeur proche de 0 de $\delta^*(f,g)$ indique que les séquences f et g possèdent une signature en dinucléotides proche ; plus la valeur de $\delta^*(f,g)$ est grande et plus les signatures en dinucléotides des séquences f et g sont différentes. KARLIN et LADUNGA (1994) ont défini des bornes pour distinguer des niveaux de distance : « aléatoire » (0,00 - 0,015), « très proche » (0,015 - 0,030), « proche » (0,030 - 0,045), « modérément proche » (0,046 - 0,065), « faiblement proche (0,065 - 0,095), « modérément distant » (0,095 - 0,140), « distant » (0,140 - 0,180) et « très distant » (≥ 0,180).

3.2. Utilisation

L'abondance relative en di- et en trinucléotides est une caractéristique d'un génome donné. De plus, elle est relativement constante le long des chromosomes (KARLIN 1998 ; GENTLES et KARLIN 2001). Des organismes proches d'un point de vue phylogénétique montrent des « patterns » plus similaires que des organismes éloignés. Ainsi, cette mesure indique que certains facteurs physiques typiques de chaque génome vont imposer des « patterns » de structures et de composition aux séquences (KARLIN et MRÁZEK 1997). L'abondance relative en dinucléotides semble rendre compte de facteurs influençant la structure de l'ADN comme les systèmes de réparation et de réplication, et les tendances d'empilement des bases (KARLIN et LADUNGA 1994 ; KARLIN *et al.* 1994). Un certain nombre de caractéristiques générales sont observées comme par exemple une sous-représentation du dinucléotide TpA chez les eucaryotes que l'on n'observe pas dans les génomes mitochondriaux et chloroplastiques. On peut aussi observer une sous-représentation du dinucléotide CpG particulièrement chez les vertébrés mais également dans les mitochondries animales, le génome des plantes dicotylédones et quelques protistes (KARLIN et MRÁZEK 1997). Les hypothèses avancées pour expliquer ces sous-représentations sont principalement d'ordre structural. Ainsi, l'évitement de TpA pourrait être lié au fait qu'il s'agit du dinucléotide possédant la plus faible énergie d'empilement.

Cela pourrait conférer une trop grande flexibilité à la double hélice. Il a aussi été démontré que UpA est une cible privilégiée pour les RNAses : en évitant ce dinucléotide, on a un bon moyen d'augmenter la stabilité des ARN (Beutler *et al.* 1989). Aussi, on trouve ce dinucléotide dans beaucoup de signaux de régulation comme par exemple les boites TATA, les terminateurs de transcription, de traduction ou les signaux de polyadénylation. Ainsi, il s'agirait d'éviter l'apparition inopportune de tels signaux n'importe où dans les séquences. Enfin, une étude récente montre que l'évitement des TpA serait directement lié à l'évitement des CpG et dépendrait du contenu en GC des séquences (Duret et Galtier 2000).

Dans le cas de la sous-représentation en CpG chez les vertébrés, il pourrait s'agir d'une conséquence relative à son implication dans le processus de méthylation* - désamination* - mutation qui entraîne le changement de CpG en TpG/CpA. Cette hypothèse n'est cependant pas valable pour les mitochondries qui ne possèdent pas de système de méthylation*. Une autre possibilité est que le dinucléotide CpG est celui qui possède la plus forte énergie d'empilement*, ainsi, une diminution en CpG pourrait faciliter la réplication et la transcription (Karlin et Burge 1995).

Karlin et Ladunga (1994) et Karlin et Mrázek (1997) ont utilisé cette méthode afin d'observer les relations phylogénétiques entre différents organismes et retrouvent globalement les mêmes relations que celles établies notamment pour l'ARN 16S. Une comparaison entre des génomes mitochondriaux et leurs génomes hôtes montre une grande différence de signature génomique. Cependant, il existe une relation parallèle entre les différences observées entre génomes nucléaires d'une part, et entre les génomes mitochondriaux correspondants d'autre part (Karlin et Mrázek 1997 ; Campbell *et al.* 1999). Par la même méthode, il est possible d'établir une signature de codons. Il s'agit de calculer les valeurs en dinucléotide à chacune des positions dans le codon. L'indice utilisé dans ce cas est celui correspondant à de l'ADN simple brin. Tout comme pour la signature génomique, on retrouve une grande stabilité dans la signature des codons pour un génome donné. De plus, signatures génomiques et signatures des codons sont très fortement corrélées (Karlin et Burge 1995).

Une étude effectuée sur des séquences de rétrovirus de mammifères afin de déterminer des signatures oligonucléotidiques particulières a montré qu'il existe deux types de motifs chevauchants dans ces séquences. Le premier constitue une caractéristique des rétrovirus et doit être le résultat d'événements évolutifs ancestraux. Il est composé de deux séquences consensus : CCTGG et CAGR (R = purine) Le deuxième est corrélé avec le biais de composition en bases des séquences rétrovirales et rendrait compte d'événements de duplication récents (Laprevotte *et al.* 1997). Une analyse sur des rétrotransposons a été effectuée par Terzian *et al.* (1997) en utilisant cette même

méthode, afin d'étudier la co-adaptation avec les génomes hôtes. Elle montre que les rétrotransposons ne possèdent pas la signature rétrovirale trouvée par Laprevotte *et al.* (1997), ce qui indiquerait que les rétrovirus de vertébrés forment un groupe distinct homogène. Cependant, les rétrotransposons semblent posséder une signature provenant de l'influence de leur hôte. Certains éléments provenant de la même espèce hôte ont en commun un certain nombre d'oligomères. Cette signature de l'hôte est principalement basée sur la sous-représentation du dinucléotide CpG qui apparaît chez les éléments de plantes et des mammifères mais pas chez ceux d'insectes et de levure.

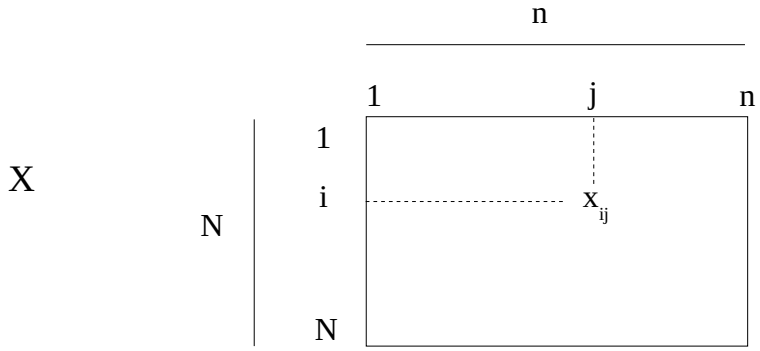
3.3. L'Analyse en Coordonnées Principales (ACO)

L'Analyse en Coordonnées Principales (ACO) permet de faire des représentations graphiques à partir d'une matrice de distance. Cette méthode a été développée par Gower en 1966. Elle est souvent utilisée afin de comparer des grands jeux de séquences d'ADN alignées. Globalement, elle permet de rendre une matrice de distance euclidienne, avant d'extraire les composantes principales pour faire une représentation graphique.

Dans les paragraphes suivants, je vais exposer les différentes étapes du déroulement de l'ACO ainsi que l'utilisation que l'on peut en faire pour l'étude de l'abondance relative en di- et en trinucléotides des ETs et des génomes hôtes.

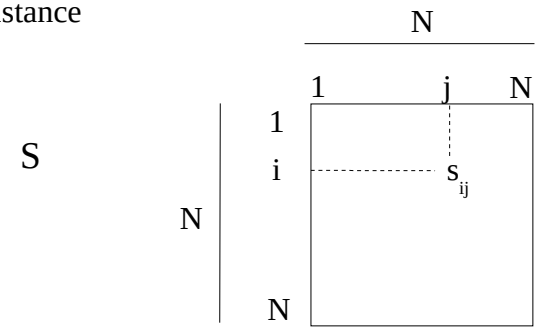
3.3.1. Les étapes principales

Soit une matrice X correspondant à *N* observations et *n* variables



Afin de situer les observations les unes par rapport aux autres, nous pouvons utiliser une matrice de dispersion ou de ressemblance des observations (matrice de distance).

Soit S la matrice de distance



Il s'agit d'une matrice carrée et symétrique, donc $s_{ij} = s_{ji}$. Dans le cas d'une matrice de dispersion, la diagonale est constituée de 0 et dans le cas d'une matrice de ressemblance, la diagonale est constituée de 1.

Soit $\lambda_1, \lambda_2, ..., \lambda_N$ les valeurs propres de la matrice S et $u_1, u_2, ..., u_N$ les vecteurs propres correspondants qui forment la matrice U suivante :

$$U = \begin{pmatrix} u_{11} & u_{21} & ..... & u_{N1} \\ .... & .... & .... & .... \\ .... & .... & .... & ... \\ u_{1N} & u_{2N} & .... & u_{NN} \end{pmatrix}$$

avec $u'_i \, u_i = \lambda_i$

Le carré de la distance entre deux points i et j est donné par la relation suivante :

$$\begin{aligned} d_{ij}^2 &= \sum_1^N (u_{ik} - u_{jk})^2 \\ &= \sum u_{ik}^2 + \sum u_{jk}^2 - 2 \sum u_{ik} u_{jk} \\ &= s_{ii} + s_{jj} - 2 s_{ij} \end{aligned}$$

Dans le cas d'une matrice de dispersion, $s_{ii} = s_{jj} = 0$, donc $d_{ij}^2 = -2 s_{ij}$

Après changement de coordonnées, nous avons : $t_{ik} = s_{ik} - \overline{s_i} - \overline{s_k} + \overline{s}$

avec $\overline{s_i}$ = la moyenne des éléments de la ligne i

$\overline{s_k}$ = la moyenne des éléments de la ligne k

$\overline{s}$ = la moyenne de tous les éléments de la matrice S

La nouvelle matrice T formée par les valeurs t_{ik} est symétrique, donc ses vecteurs propres

sont orthogonaux. Après diagonalisation de cette matrice, on peut calculer les valeurs propres ainsi que les vecteurs propres correspondants v_1, v_2, ..., v_N.

Les vecteurs propres normés de telle sorte que $v'_i \, v_i = \lambda_i$ forment la matrice V qui contient les coordonnées sur chaque axe des différentes observations.

3.3.2. Utilisation dans l'étude de l'abondance relative des di- et des trinucléotides

Une distance entre séquences peut être calculée par rapport à l'abondance relative en di- et en trinucléotides (voir paragraphe 3.1.1.). KARLIN et MRÁZEK (1997) ont utilisé ces distances calculées entre les abondances relatives en dinucléotides de séquences génomiques de différentes espèces afin de construire un arbre par la méthode UPGMA. Ceci permet une représentation graphique des relations de proximité en terme d'abondance relative en dinucléotides entre les différentes séquences analysées. La construction d'un tel arbre permet d'observer les relations entre les différentes séquences de manière plus simple qu'un tableau de données et peut éventuellement refléter des relations évolutives. Cependant, cette représentation peut poser des problèmes. Ainsi, il ne faut pas perdre de vue qu'il ne s'agit pas d'un arbre phylogénétique. De plus, ce type de représentation ne peut admettre qu'un nombre limité de données pour des raisons de clarté de figures.

Ainsi, pour pallier à ces inconvénients, nous avons décidé de traiter les matrices de distance par l'intermédiaire de l'ACO. Celle-ci va nous permettre de comparer de nombreuses séquences.

3.4. Les analyses

Dans cette partie, je vais exposer les différentes analyses effectuées sur l'abondance relative en di- et trinucléotides entre génomes hôte et ETs. Une partie de ces résultats fait l'objet de l'article 5 présenté en annexe « *Potentially infectious retroelements come out from relative abundance of dinculeotides* » actuellement en cours de soumission.

3.4.1. Chez les cinq génomes : *A. thaliana, C. elegans, D. melanogaster, H. sapiens* et *S. cerevisiae*

3.4.1.1. Les données

Les génomes complets de *Saccharomyces cerevisiae*, *Caenorhabditis elegans* et *Drosophila melanogaster*, ainsi que les chromosomes 2 et 4 d'*Arabidopsis thaliana* et les chromosomes 21 et 22 d'*Homo sapiens*, ont été extraits sur le site de la « Genome On Line Database » (wit.integratedgenomics.com/GOLD/, Kyrpides 1999). De plus, les séquences entières d'ETs (ETs complets) présentes dans la base de données Genbank pour les éléments de *C. elegans*, *D. melanogaster*, *H. sapiens* et *A. thaliana* ont été collectées. D'autres ETs d'*Arabidopsis* ont été rajoutés à partir de la base de données « The Arabidopsis Transposable Element Database » (soave.biol.mcgill.ca/clonebase/main.html). Les positions des ETs sur chacun des chromosomes séquencés de *Saccharomyces* ont été extraites sur le site « Transposable Element ressources » (www.public.iastate.edu/~voytas/resources/resources. html). Au total, notre jeu de données est composé de 40 ETs de drosophile, 50 de levure, 19 de nématode, 25 d'homme et 31 d'arabette (voir tableau 5' en annexe pour les noms et les numéros d'accession).

3.4.1.2. L'abondance relative en dinucléotides

3.4.1.2.1. Utilisation des génomes entiers et des ETs concaténés

A partir de chaque génome et des ETs concaténés pour chaque espèce, nous avons calculé les indices d'abondance relative en dinucléotides ρ^*. Les valeurs pour chaque dinucléotide et couple de dinucléotides sont données dans le Tableau 12.

Tableau 12 : Abondance relative en dinucléotides pour les génomes hôtes et les ETs

	CpG	GpC	TpA	ApT	CpC/GpG	ApA/TpT	TpG/CpA	ApG/CpT	ApC/GpT	TpC/GpA
Arabidopsis thaliana										
ETs	*0.712*	0.924	*0.682*	0.904	0.998	1.120	1.120	1.068	0.895	1.167
génome	*0.724*	0.923	*0.751*	0.905	1.044	1.130	1.101	1.030	0.909	1.110
Caenorhabditis elegans										
ETs	0.888	0.941	*0.639*	0.902	1.049	1.204	1.078	0.968	0.850	1.150
génome	0.975	1.051	*0.615*	0.853	1.055	**1.289**	1.089	0.893	0.853	1.087
Drosophila melanogaster										
ETs	0.884	1.157	*0.772*	0.909	1.036	1.171	1.105	0.957	0.916	0.967
génome	0.928	**1.278**	*0.752*	0.974	1.047	1.218	1.133	0.887	0.854	0.906
Homo sapiens										
ETs	*0.393*	1.035	*0.680*	0.862	1.172	1.170	1.199	1.117	0.837	1.010
génome	*0.263*	1.001	*0.708*	0.865	1.228	1.119	1.227	0.982	1.179	0.835
Saccharomyces cerevisiae										
ETs	0.781	*0.772*	0.862	1.066	1.021	0.967	1.169	0.946	0.997	1.124
génome	0.801	1.021	*0.770*	0.939	1.061	1.135	1.099	0.987	0.895	1.053

En italique sur fond gris et en gras sont représentés les indices déterminant un dinucléotide significativement sous- (≤ 0,78) et sur-représenté (≥ 1,23), respectivement.

D'après le Tableau 12, pour une espèce donnée, les ETs et les génomes montrent le même pattern en abondance relative des dinucléotides. Ceci est confirmé par le calcul de coefficients de corrélation entre ETs et séquences d'hôte. Ainsi, on obtient des coefficients r = 0,98 ($p<0,05$) pour *Arabidopsis*, r = 0,93 ($p<0,05$) pour *Caenorhabditis*, r = 0,94 ($p<0,05$) pour *Drosophila*, r = 0,87 ($p<0,05$) pour l'homme. En ce qui concerne la levure, le coefficient n'est pas significativement différent de zéro (r = 0,54, p = 0,40).

On remarque que, quelle que soit l'espèce, le dinucléotide TA est sous-représenté à la fois dans les génomes et dans les ETs, excepté chez les rétrotransposons à LTR de la levure. Le dinucléotide CG est sous-représenté chez *Arabidopsis thaliana* et *Homo sapiens*, à la fois dans les séquences génomiques et dans les ETs. Dans le génome de *Caenorhabditis*, on observe une sur-représentation du couple de dinucléotides AA/TT.

A partir des données du Tableau 12, nous pouvons calculer entre chaque séquence une

distance en abondance relative. La Tableau 13 montre le résultat de ce calcul.

Tableau 13 : Distance relative δ^* entre génomes et ETs (multipliée par 1000)

		A. thaliana		***C. elegans***		***D. melanogaster***		***H. sapiens***		***S. cerevisiae***	
		génome	ETs	génome	ETs	génome	ETs	génome	ETs	génome	ETs
A. thaliana	génome	0	**28.21**	85.11	49.44	104.68	60.93	153.05	100.50	**31.23**	85.00
	ETs		0	101.03	55.42	124.75	85.80	167.04	100.60	56.07	96.93
C. elegans	génome			0	**48.26**	72.14	75.75	197.89	124.84	70.42	142.07
	ETs				0	85.81	62.30	181.57	116.23	52.00	108.28
D. melanogaster	génome					0	**50.13**	176.77	134.10	79.83	145.03
	ETs						0	153.68	111.34	41.31	112.47
H. sapiens	génome							0	**110.54**	145.43	184.43
	ETs								0	96.59	168.44
S. cerevisiae	génome									0	91.88
	ETs										0

En gras sont représentées les distances les plus faibles pour chaque génome.

On constate d'après ce tableau que les distances les plus faibles observées pour chaque génome correspondent aux distances entre un génome et ses ETs. Il y a une exception : le génome de la levure montre une plus faible distance avec le génome d'*Arabidopsis*.

A partir des données de ce tableau, nous avons effectué une ACO. Le résultat est présenté sur la Figure 13.

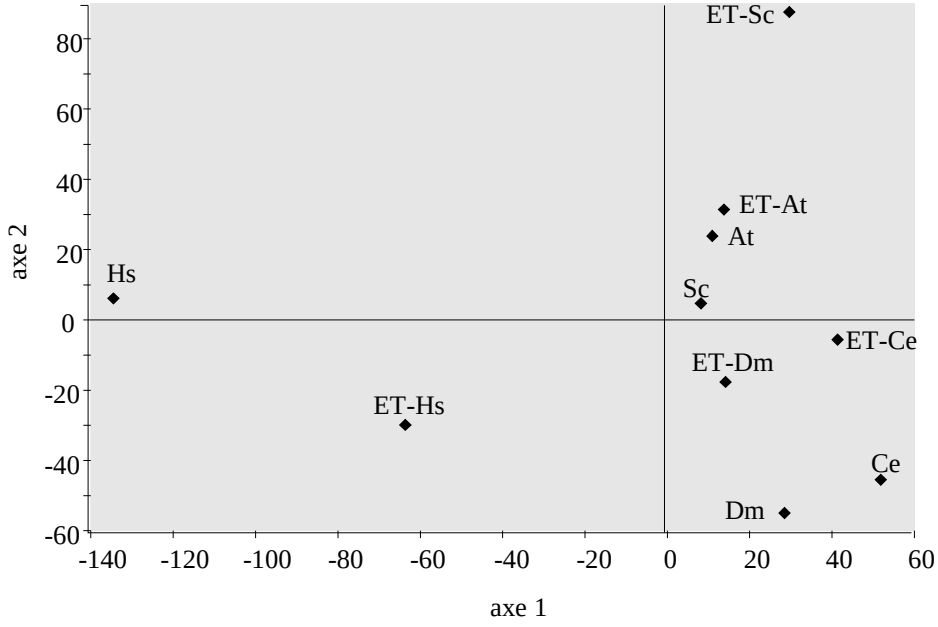


Figure 13 : Projection sur les deux premiers axes d'une ACO de la distance en dinucléotides des génomes et des ETs. Les pourcentage de variance des axes 1 et 2 sont 51% et 27% respectivement. ET-At = ETs d'*A. thaliana*, ET-Ce = ETs de *C. elegans*, ET-Dm = ETs de *D. melanogaster*, ET-Hs = ETs d'*H. sapiens*, ET-Sc = ETs de *S. cerevisiae*, At = génome d'*A. thaliana*, Ce = génome de *C. elegans*, Dm = génome de *D. melanogaster*, Hs = génome d'*H. sapiens* et Sc

= génome de *S. cerevisiae.*

La Figure 13 confirme bien les tendances observées dans le Tableau 13 : génomes et ETs se regroupent suivant l'espèce à laquelle ils appartiennent, excepté la levure qui montre une grande distance entre ETs et génome. Les ETs d'*H. sapiens* montrent une grande distance avec le génome hôte, cependant, elle reste inférieure à celle observée avec le reste des points de la figure.

3.4.1.2.2. Utilisation de fragments génomiques et des ETs complets

L'analyse précédente donne un aperçu global concernant l'ensemble des ETs et le génome total de chaque espèce. Afin de déterminer la variabilité de ce que l'on observe et de s'affranchir de biais pouvant résulter de la présence d'ETs dans les séquences génomiques, nous avons fractionné chaque génome en fragments de 9 000 bp. Cette taille est relativement équivalente à celle des ETs complets (parties codantes et non codantes). Un tirage au sort a permis de prélever 100 fragments pour chaque espèce. Afin d'éliminer les fragments possédant des ETs, nous avons réalisé un BlastN pour chaque espèce (ALTSCHUL *et al.* 1997) des 100 fragments avec les séquences d'ETs complets. Ceci nous a permis de garder au total 459 fragments génomiques pour les cinq espèces qui ne contiennent pas d'ETs. Le calcul des distances en dinucléotides nous a permis de réaliser une ACO sur les cinq espèces. Le résultat est présenté sur la Figure 14.

Les Figures 14A à 14E sont superposables. Pour plus de clarté, nous avons représenté la projection de l'ACO sur cinq graphes différents correspondant aux cinq espèces. On constate d'après ces figures que pour chaque espèce, excepté *S. cerevisiae*, ETs complets et fragments génomiques sont superposés. Des MANOVA effectuées sur les coordonnées des fragments génomiques et des ETs pour chaque espèce montrent cependant que ETs et fragments génomiques forment des groupes distincts, excepté chez *A. thaliana*. Une MANOVA sur toutes les espèces, en considérant ETs et fragments génomiques comme un seul groupe montre que tous les « groupes espèces » sont différents. Ainsi, les ETs et les fragments génomiques forment un groupe caractéristique de l'espèce, bien que ces deux sous-groupes soient distincts à l'intérieur d'une espèce.

Afin de comparer l'ensemble des Figure 14A à 14E avec la Figure 13, nous avons calculé la moyenne des coordonnées des fragments, d'une part, et des ETs complets, d'autre part, pour chacune des cinq espèces. La Figure 14F montre le même type de répartition que la Figure 13. Ainsi, ETs et fragments génomiques d'une espèce sont regroupés, excepté pour la levure. On retrouve la proximité entre le génome de la levure et celui de l'arabette, ainsi que la forte proximité entre ETs et génome de cette dernière. On observe aussi une forte proximité entre ETs de *Caenorhabditis elegans* et de *Drosophila melanogaster* d'une part, et entre fragments génomiques de ces deux

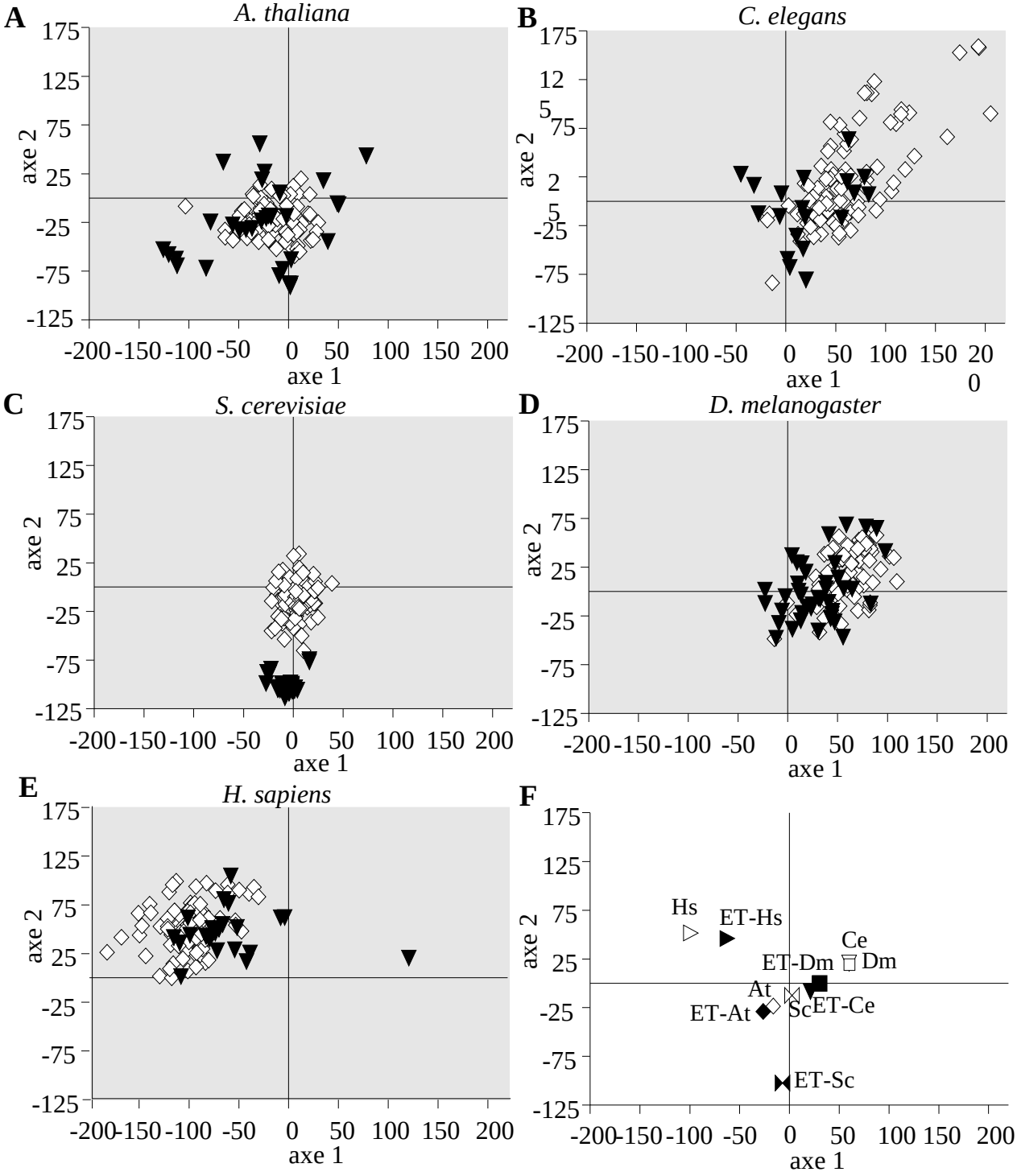


Figure 14 A à E : Projection d'une ACO de la distance relative en dinucléotides de fragments génomiques (losanges blancs) et de séquences d'ETs complets (triangles noirs). Chacune des cinq figures sont superposables. Le pourcentage de variabilité des axes 1 et 2 sont 29% et 19% respectivement. La **Figure 14 F** représente la position de la moyenne des coordonnées des fragments génomiques (symboles blancs) et des ETs (symboles noirs) pour chaque espèce.

	CG	GC	TA	AT	CC/GG	AA/TT	TG/CA	AG/CT	AC/GT	TC/GA
A. thaliana										
AtRE1	0.790	0.884	*0.655*	1.052	0.992	1.039	**1.236**	0.937	0.947	1.152
AtRE1-2	0.786	0.876	*0.664*	1.055	0.981	1.024	**1.231**	0.952	0.946	1.167
AtRE2	0.838	0.901	*0.717*	1.099	0.961	0.986	**1.235**	0.924	0.968	1.141
AtRE2-1	0.838	0.910	*0.719*	1.091	0.960	0.992	**1.242**	0.918	0.977	1.126
AtRE2-2	0.832	0.909	*0.718*	1.084	0.958	0.994	**1.239**	0.925	0.974	1.130
Tat1-1	1.170	0.930	*0.624*	0.822	0.930	**1.260**	0.837	1.074	0.833	**1.293**
Tat1-2	1.160	0.933	*0.601*	0.787	0.912	**1.270**	0.850	1.085	0.850	**1.292**
Tat1-3	1.069	0.943	*0.592*	0.783	0.962	**1.282**	0.812	1.157	*0.778*	1.314
Athila	*0.631*	0.927	*0.600*	0.925	0.970	1.095	1.188	1.113	0.860	1.217
Ta1-2	*0.387*	0.882	*0.624*	0.844	0.887	1.033	**1.255**	**1.251**	0.925	**1.237**
Ta1-2-2	*0.440*	0.852	*0.620*	0.828	0.875	1.047	1.227	**1.247**	0.933	**1.255**
Ta1-3	*0.458*	0.833	*0.626*	0.805	0.869	1.050	1.214	**1.252**	0.954	**1.252**
gypsy-likeI	*0.643*	1.040	*0.795*	0.923	1.040	1.112	1.140	1.016	0.913	1.048
gypsy-likeIII	*0.650*	0.883	*0.651*	0.874	1.147	1.201	1.046	1.085	0.798	1.183
gypsy-likeVIII	*0.676*	1.017	*0.751*	0.900	1.127	1.167	1.152	0.951	0.921	1.005
gypsy-likeXV	0.931	*0.738*	0.797	0.993	1.130	1.072	0.924	1.018	0.916	1.212
Ta11-1	*0.462*	0.848	*0.594*	0.781	1.142	1.206	1.075	1.210	0.807	1.201
TSCL	0.926	0.821	*0.565*	*0.732*	0.993	**1.251**	1.056	1.010	1.010	1.155
LINE-likeV	*0.726*	0.898	*0.598*	0.860	1.012	1.170	1.113	1.085	0.874	1.194
LINE-likeXXIII	*0.470*	0.894	0.792	0.984	0.972	0.986	1.218	1.120	0.930	1.151
LINE-likeXVIII	*0.558*	1.103	*0.777*	0.919	**1.272**	1.179	1.027	1.061	*0.750*	1.057
LINE-likeXI	*0.610*	1.115	*0.759*	0.915	1.111	1.154	1.059	1.105	*0.778*	1.089
Limpet	*0.665*	*0.549*	*0.712*	0.937	1.129	1.120	1.043	1.048	0.832	**1.319**
Tag1	*0.738*	1.029	*0.744*	0.974	1.052	1.121	1.179	0.940	0.893	1.058
Tag2	*0.522*	0.999	*0.714*	0.917	0.976	1.085	1.203	1.104	0.890	1.126
Ac-likeI	*0.581*	*0.772*	0.795	0.917	1.155	1.101	1.140	0.994	0.970	1.067
MITE-XI	*0.422*	0.921	*0.624*	0.873	*0.719*	1.055	**1.241**	**1.245**	0.883	**1.323**
MITE-V	1.169	1.126	*0.746*	0.939	1.126	**1.246**	1.027	0.802	0.871	0.988
CACTA-like1	*0.733*	1.066	*0.636*	0.788	0.801	1.162	1.142	1.162	0.916	1.170
CACTA-like2	*0.447*	0.991	*0.684*	0.828	0.948	1.098	**1.233**	1.166	0.943	1.098
CACTA-like3	*0.750*	1.053	*0.642*	0.794	0.828	1.165	1.122	1.157	0.906	1.172
C. elegans										
Tc1	1.055	1.063	*0.502*	0.865	1.032	**1.360**	1.177	*0.764*	0.904	1.028
Tc2	1.030	1.202	*0.542*	0.781	0.800	**1.302**	1.094	1.022	0.858	1.138
Tc5	1.013	1.009	*0.637*	0.829	0.917	**1.239**	1.171	0.874	1.020	1.025
TR-5	0.882	*0.740*	*0.655*	0.946	1.067	1.138	1.063	0.973	0.908	1.228
Tc6.1	1.099	0.972	*0.573*	*0.757*	1.031	**1.363**	1.071	0.849	0.960	1.034
Tc4	1.023	0.943	*0.489*	0.784	1.216	**1.412**	1.124	*0.759*	0.843	1.079
Tc4v	1.031	1.035	*0.590*	0.993	1.092	**1.254**	1.179	*0.748*	0.841	1.081
sam1	0.886	0.855	*0.723*	0.966	1.164	1.177	0.997	0.966	0.845	1.138
sam2	0.928	0.849	*0.592*	1.042	0.916	1.078	1.176	0.938	0.895	1.277
sam3	0.872	0.943	*0.640*	1.056	0.935	1.058	1.136	1.013	0.841	1.254
sam4	*0.769*	0.920	*0.691*	0.888	1.118	1.197	1.014	1.056	0.808	1.167
sam5.1	0.796	1.018	*0.677*	0.951	1.190	1.223	1.040	0.968	0.746	1.126
sam7.1	*0.765*	0.789	*0.681*	0.908	1.098	1.149	0.957	1.134	0.781	1.296
sam8	*0.629*	0.925	*0.743*	0.887	1.194	1.172	0.977	1.136	0.771	1.153
sam9	*0.721*	0.837	*0.710*	0.882	1.186	1.188	0.948	1.115	0.778	1.207

	CG	GC	TA	AT	CC/GG	AA/TT	TG/CA	AG/CT	AC/GT	TC/GA
Ce suite										
frodo1	*0.714*	1.092	0.780	0.896	1.021	1.136	1.157	0.996	0.949	0.987
frodo2.2	0.905	0.924	0.851	0.906	1.061	1.113	1.031	0.989	0.978	1.031
RTE1	0.943	0.797	0.521	0.882	0.914	1.118	1.055	1.077	0.923	1.340
Cer1	0.817	0.959	0.549	0.921	0.975	1.210	1.124	1.007	0.786	1.256
D. melanogaster										
nomad	1.032	1.063	0.821	0.881	1.075	**1.234**	0.946	0.965	0.898	0.988
Transpac	0.973	1.057	0.813	0.887	1.053	1.166	1.009	0.978	0.914	1.031
blastopia	0.896	1.161	*0.678*	0.868	0.920	1.201	1.164	0.966	0.937	1.007
copia	*0.664*	1.207	0.803	0.922	1.046	1.132	1.171	0.972	0.916	0.960
micropia	1.049	1.138	*0.758*	0.960	0.900	1.137	1.100	0.938	0.929	1.043
Gate	1.023	1.104	*0.686*	0.943	0.993	**1.233**	1.030	0.956	0.852	1.058
circe	1.004	1.056	*0.712*	0.976	0.974	1.157	1.122	0.893	0.909	1.072
17.6	*0.713*	1.076	*0.731*	0.904	1.097	1.180	**1.240**	0.860	0.980	0.929
297	0.831	0.966	*0.753*	0.881	1.206	1.208	1.098	0.883	0.937	0.977
MDG3	1.030	1.083	*0.769*	0.855	0.937	1.185	1.051	0.969	0.961	1.028
412	0.803	**1.234**	0.883	0.967	1.061	1.098	1.109	0.964	0.914	0.928
Tirant	1.084	0.877	0.904	0.964	1.079	1.091	1.091	0.803	1.113	0.916
1731	0.798	1.181	*0.705*	0.965	0.885	1.071	**1.250**	1.023	0.933	1.011
BEL	1.004	1.120	*0.660*	0.994	0.860	1.109	1.128	0.995	0.889	1.129
burdock	1.015	1.074	0.799	0.879	1.008	1.181	1.015	0.971	0.934	1.015
ZAM	0.880	0.837	0.872	0.871	1.177	1.142	1.089	0.877	1.103	0.889
gypsy	0.853	1.089	0.788	0.843	1.016	1.175	1.074	1.036	0.942	0.969
blood	0.945	**1.247**	0.925	0.983	1.013	1.083	1.064	0.961	0.925	0.923
HMS-beagle	0.957	1.041	*0.745*	0.926	1.086	1.203	1.042	0.929	0.880	1.035
roo	*0.725*	**1.230**	0.787	0.963	1.148	1.162	1.075	0.993	*0.775*	1.025
springer	0.949	1.094	*0.736*	0.946	1.026	1.199	1.058	0.965	0.878	1.011
Idefix	0.781	0.936	0.827	0.917	1.113	1.121	1.068	0.983	0.941	1.034
Het-A	0.958	**1.379**	*0.754*	0.891	1.115	**1.313**	1.073	0.877	0.827	0.831
R1Dm	0.889	1.119	*0.704*	1.001	0.895	0.991	1.177	1.093	0.913	1.068
R2Dm	0.914	1.002	0.829	1.001	1.049	1.090	1.040	0.994	0.905	1.049
TART-B1	*0.756*	1.221	*0.770*	0.882	1.065	1.209	1.107	1.036	0.851	0.919
DOC	0.857	1.179	*0.766*	0.840	0.944	1.177	1.146	1.001	0.978	0.931
I	0.784	0.917	0.836	0.883	1.223	1.171	1.111	0.884	1.032	0.879
jockey	0.789	1.217	0.908	0.980	0.985	1.049	1.144	1.019	0.939	0.917
F	0.827	1.165	*0.779*	0.868	0.955	1.145	1.184	0.989	0.993	0.913
G	*0.743*	1.035	*0.565*	0.782	0.994	1.174	1.174	1.126	0.913	1.054
BS	*0.684*	1.183	0.817	0.928	1.007	1.088	1.180	1.071	0.909	0.937
pogo	1.005	**1.733**	*0.666*	0.915	0.788	**1.259**	1.220	0.894	0.791	0.919
hobo	0.884	1.130	*0.719*	0.832	1.107	**1.267**	1.089	0.917	0.920	0.934
HB1	*0.665*	0.988	0.828	0.878	**1.387**	1.212	1.039	0.933	0.907	0.886
S	0.825	**1.337**	*0.719*	0.803	1.187	**1.311**	1.098	0.897	0.882	0.845
hoppel	1.126	**1.712**	0.890	0.989	0.811	1.122	1.066	0.967	0.816	0.914
hopper	0.883	**1.552**	*0.738*	0.845	1.123	**1.295**	1.112	0.886	0.836	0.823
P	1.071	1.177	*0.747*	0.940	0.995	1.198	1.070	0.892	0.872	1.025
Bari-1	*0.713*	**1.372**	*0.677*	0.815	1.122	**1.305**	1.190	0.904	0.884	0.839

	CG	GC	TA	AT	CC/GG	AA/TT	TG/CA	AG/CT	AC/GT	TC/GA
S. cerevisiae										
Ty1-chr1	*0.773*	*0.739*	0.863	1.077	1.036	0.956	1.168	0.946	0.998	1.136
Ty1-1-chr2	*0.757*	*0.766*	0.871	1.093	1.055	0.952	1.168	0.945	0.978	1.129
Ty1-2-chr2	0.810	*0.776*	0.881	1.110	1.044	0.945	1.150	0.938	0.968	1.140
Ty1-1-chr4	*0.764*	*0.774*	0.862	1.077	1.039	0.961	1.172	0.946	0.989	1.123
Ty1-2-chr4	*0.767*	*0.762*	0.865	1.074	1.029	0.957	1.170	0.951	0.996	1.129
Ty1-3-chr4	*0.773*	*0.747*	0.856	1.086	1.049	0.959	1.170	0.937	0.987	1.136
Ty1-4-chr4	*0.774*	*0.750*	0.874	1.101	1.060	0.949	1.166	0.933	0.981	1.132
Ty1-5-chr4	*0.745*	*0.773*	0.855	1.072	1.026	0.958	1.179	0.959	0.989	1.132
Ty1-1-chr5	*0.774*	*0.736*	0.866	1.080	1.036	0.954	1.165	0.950	0.997	1.140
Ty1-1-chr7	0.789	*0.774*	0.854	1.106	1.044	0.956	1.178	0.922	0.973	1.137
Ty1-2-chr7	*0.764*	*0.774*	0.865	1.081	1.031	0.955	1.171	0.952	0.988	1.129
Ty1-3-chr7	0.793	*0.769*	0.882	1.106	1.052	0.946	1.153	0.940	0.972	1.135
Ty1-1-chr8	*0.777*	*0.720*	0.876	1.090	1.057	0.947	1.159	0.941	0.998	1.136
Ty1-1-chr10	*0.771*	*0.738*	0.857	1.082	1.024	0.951	1.177	0.946	0.998	1.145
Ty1-2-chr10	*0.776*	*0.774*	0.856	1.095	1.039	0.957	1.178	0.933	0.978	1.135
Ty1-1-chr12	0.801	*0.763*	0.887	1.109	1.046	0.940	1.156	0.936	0.982	1.133
Ty1-2-chr12	0.781	*0.738*	0.866	1.079	1.037	0.954	1.162	0.947	0.996	1.139
Ty1-3-chr12	*0.744*	*0.773*	0.847	1.077	1.017	0.957	1.191	0.953	0.990	1.136
Ty1-4-chr12	*0.775*	*0.770*	0.857	1.059	1.028	0.970	1.167	0.952	1.000	1.122
Ty1-1-chr13	*0.763*	*0.744*	0.872	1.096	1.056	0.947	1.166	0.943	0.984	1.137
Ty1-2-chr13	*0.779*	*0.743*	0.867	1.080	1.026	0.950	1.166	0.952	0.998	1.141
Ty1-3-chr13	*0.750*	*0.769*	0.870	1.087	1.038	0.948	1.171	0.957	0.984	1.133
Ty1-4-chr13	0.796	*0.767*	0.877	1.107	1.054	0.949	1.166	0.924	0.982	1.126
Ty1-1-chr14	*0.765*	*0.765*	0.871	1.095	1.036	0.945	1.169	0.951	0.981	1.139
Ty1-2-chr14	0.783	*0.768*	0.846	1.072	0.990	0.958	1.178	0.958	0.993	1.152
Ty1-1-chr15	*0.778*	*0.735*	0.859	1.088	1.040	0.953	1.171	0.939	0.992	1.143
Ty1-2-chr15	0.797	*0.773*	0.874	1.108	1.056	0.952	1.158	0.930	0.969	1.134
Ty1-1-chr16	*0.777*	*0.739*	0.865	1.079	1.036	0.953	1.164	0.949	0.997	1.139
Ty1-2-chr16	*0.752*	*0.762*	0.848	1.072	1.020	0.960	1.186	0.951	0.996	1.136
Ty1-3-chr16	0.780	*0.746*	0.873	1.100	1.061	0.951	1.164	0.932	0.982	1.133
Ty1-4-chr16	*0.775*	*0.749*	0.854	1.098	1.040	0.962	1.182	0.933	0.986	1.144
Ty2-1-chr2	*0.778*	0.814	0.866	1.063	0.971	0.961	1.196	0.946	1.011	1.110
Ty2-1-chr3	0.814	0.788	0.861	1.058	0.990	0.973	1.182	0.930	1.015	1.111
Ty2-1-chr4	*0.778*	0.804	0.870	1.058	1.000	0.970	1.182	0.943	1.005	1.105
Ty2-2-chr4	0.799	0.814	0.870	1.054	0.990	0.973	1.183	0.937	1.016	1.096
Ty2-3-chr4	0.813	0.833	0.879	1.075	0.986	0.962	1.180	0.934	0.999	1.103
Ty2-1-chr6	0.793	0.788	0.866	1.053	0.981	0.967	1.187	0.940	1.023	1.107
Ty2-1-chr7	0.813	0.792	0.869	1.048	0.984	0.974	1.177	0.938	1.022	1.104
Ty2-2-chr7	0.796	0.791	0.868	1.052	0.982	0.968	1.184	0.942	1.022	1.106
Ty2-1-chr12	0.787	0.787	0.864	1.057	0.999	0.971	1.187	0.934	1.015	1.106
Ty2-2-chr12	0.824	0.788	0.865	1.064	0.996	0.972	1.172	0.930	1.006	1.116
Ty-2-1-chr14	0.819	0.798	0.868	1.048	0.990	0.978	1.173	0.935	1.018	1.101
Ty2-1-chr15	0.800	0.820	0.859	1.044	0.989	0.984	1.183	0.937	1.013	1.096
Ty2-2-chr15	0.796	0.811	0.865	1.057	0.984	0.970	1.191	0.935	1.016	1.101
Ty3-1-chr7	0.919	*0.768*	0.863	0.940	1.050	1.050	1.066	0.955	1.059	1.064
Ty3-1-chr9	0.920	*0.765*	0.851	0.928	1.049	1.059	1.066	0.956	1.063	1.066
Ty4-1-chr8	*0.678*	*0.743*	0.796	0.988	0.996	1.037	1.193	0.967	0.980	1.148
Ty4-1-chr10	*0.682*	*0.747*	0.794	0.982	1.005	1.044	1.188	0.966	0.980	1.142

	CG	GC	TA	AT	CC/GG	AA/TT	TG/CA	AG/CT	AC/GT	TC/GA
Sc suite										
Ty4-1-chr16	*0.667*	*0.744*	0.796	0.988	0.997	1.037	1.195	0.970	0.979	1.148
Ty5-1-chr3	*0.773*	0.911	0.906	0.998	0.936	0.953	1.137	1.054	1.020	1.081
H. sapiens										
HERVK-T47D	*0.255*	0.948	*0.719*	0.923	**1.279**	1.107	**1.253**	1.108	0.853	0.972
HERVK	*0.349*	0.942	*0.769*	0.903	**1.292**	1.131	1.157	1.103	0.849	0.981
HERV-K101	*0.346*	0.929	*0.770*	0.900	**1.313**	1.139	1.145	1.099	0.845	0.983
HERV-KC4	*0.286*	0.995	0.795	*0.774*	**1.406**	**1.232**	1.103	1.080	0.910	0.852
HIV1	*0.195*	1.048	0.862	0.898	**1.230**	1.041	1.172	**1.246**	0.836	0.962
HIV2	*0.284*	1.000	0.830	0.884	1.227	1.051	1.175	**1.234**	0.843	0.967
HTLV1	*0.549*	0.909	0.874	0.871	**1.294**	1.159	1.014	1.170	0.804	0.958
HTLV2	*0.500*	*0.759*	0.815	0.921	**1.380**	1.144	1.034	1.106	0.838	1.001
HSRV	*0.308*	0.906	0.904	0.950	**1.245**	1.017	1.106	1.169	0.885	1.023
v-onc	*0.281*	0.923	*0.774*	*0.765*	**1.295**	1.154	1.142	1.222	0.872	0.941
LINE-1	*0.318*	0.954	*0.624*	0.881	1.217	1.172	**1.230**	1.103	0.823	1.055
L1.5	*0.314*	0.977	*0.634*	0.890	1.178	1.143	**1.236**	1.137	0.819	1.066
L1.24	*0.311*	0.979	*0.630*	0.885	1.173	1.145	**1.246**	1.131	0.828	1.061
L1.33	*0.327*	0.967	*0.635*	0.889	1.186	1.146	**1.233**	1.127	0.825	1.062
L1.1	*0.359*	0.976	*0.625*	0.876	1.194	1.176	1.216	1.110	0.819	1.057
L1.2	*0.353*	0.966	*0.632*	0.879	1.194	1.165	1.218	1.115	0.826	1.057
Buster3	*0.418*	1.197	*0.703*	0.967	1.083	1.120	1.187	1.147	*0.729*	1.084
Buster1	*0.463*	1.217	*0.692*	0.902	0.974	1.137	**1.266**	1.074	0.863	1.019
Hsmar1	*0.373*	**1.408**	*0.528*	0.834	0.942	**1.249**	**1.469**	0.968	0.896	0.883
humar1pcr	*0.206*	1.065	*0.528*	*0.745*	1.099	**1.276**	**1.326**	1.080	0.886	1.013
humar1g1	*0.143*	1.181	*0.522*	0.816	1.110	**1.271**	**1.357**	1.090	0.801	1.026
mariner2	*0.554*	1.118	*0.472*	0.789	1.099	**1.343**	1.184	1.050	*0.755*	1.095
mariner1	1.223	1.207	*0.519*	0.808	0.856	**1.377**	1.112	0.828	0.929	1.024
Tigger1	*0.526*	**1.267**	*0.684*	0.903	0.975	1.143	1.178	1.178	*0.759*	1.066
Tramp	*0.581*	1.040	*0.439*	*0.702*	1.058	1.212	1.224	**1.240**	0.827	1.047

Tableau 14 : Abondance relative en dinucléotides pour les ETs complets des 5 espèces. Les traits de séparation correspondent aux classes ou familles principales des ETs pour une espèce

mêmes espèces d'autre part. Un test « Protected Least Significant Difference » (PLSD) de Fisher sur les deux axes montre en effet que les groupes ETs ne sont pas significativement différents entre ces deux espèces. Il en est de même pour les fragments génomiques. Enfin, les différents composants de l'homme sont clairement éloignés des autres espèces.

Nous avons détaillé les valeurs des indices de dinucléotides pour chaque ET complet (Tableau 14). Le dinucléotide CpG est fortement sous-représenté chez les éléments humains, chez la majorité des ETs d'*Arabidopsis* et dans les familles *Ty1*, *Ty4* et *Ty5* de la levure. Chez la drosophile et le nématode, cette sous-représentation n'existe que pour quelques éléments mais ne semble pas se généraliser. La sur-représentation du dinucléotide ApA/TpT observée dans le Tableau 12 dans le génome de *C. elegans*, est aussi présente dans les transposons et les éléments de classe III de cette espèce. Chez la levure et chez l'homme on observe que certaines tendances reflètent le type d'élément. Ainsi, chez la levure, on observe une sous-représentation de GpC dans les éléments des familles *Ty1*, *Ty3* et *Ty4*, qui n'existe pas dans les familles *Ty2* et *Ty5*. En revanche, les rétrovirus humains semblent caractérisés par une sur-représentation du couple CpC/GpG. Enfin, il y a une généralisation de la sous-représentation du dinucléotide TpA dans tous les ETs, à l'exception des éléments de la levure, des rétrovirus humains, chez la drosophile de nombreux rétrotransposons à LTR possédant un gène d'enveloppe. Il semble donc qu'il existe une caractéristique commune entre rétrovirus humains et rétrotransposons à LTR avec un gène *env* de la drosophile.

3.4.1.3. L'abondance relative en trinucléotides

3.4.1.3.1. Utilisation des génomes entiers et des ETs concaténés

A partir du jeu de données utilisé dans le paragraphe 3.4.1.2.1, nous avons calculé les indices d'abondance relative en trinucléotides g^*. Les valeurs pour chaque couple de trinucléotides sont données dans le Tableau 15. La sur- et la sous-représentation de chaque couple de trinucléotides est donnée par la variation de l'indice par rapport à 1. Ainsi, on peut définir les deux couples les plus sur-représentés et ceux les plus sous-représentés.

Tout comme nous l'avons observé dans le Tableau 12, les ETs et les génomes hôtes ont tendance à présenter le même type de « pattern ». Ainsi, chez *Arabidopsis*, ETs et génome présentent les mêmes couples les plus sur-représentés (CTC/GAG et CCG/CGG) et ont en commun le couple le plus sous-représenté (CCC/GGG). Nous observons le même type de comportement chez

Tableau 15 : Abondance relative en trinucléotides pour les génomes hôte et pour les ETs

	A. thaliana		*C. elegans*		*D. melanogaster*		*H. sapiens*		*S. cerevisiae*	
	ETs	génome	ETs	génome	ETs	génome	ETs	génome	ETs	génome
AAA/TTT	0.918	0.960	0.981	0.975	0.974	0.983	0.988	1.058	1.016	0.978
AAT/ATT	1.035	1.008	0.979	1.007	1.010	0.967	1.049	1.060	1.004	1.000
AAC/GTT	1.016	1.013	1.017	0.995	0.997	1.007	0.997	0.957	0.967	0.981
AAG/CTT	1.066	1.025	1.001	0.947	1.008	1.020	0.939	*0.862*	1.008	1.040
ATA/TAT	1.061	1.052	**1.114**	1.026	1.049	1.073	0.994	1.035	1.063	1.078
ATC/GAT	0.956	0.962	0.998	0.991	0.973	1.004	0.988	0.986	0.958	0.981
ATG/CAT	0.988	1.018	1.008	1.071	0.990	1.006	1.010	0.986	0.949	0.959
ACA/TGT	0.970	0.964	*0.890*	0.901	0.971	0.978	0.985	0.961	*0.913*	*0.908*
ACT/AGT	0.987	1.011	1.094	1.063	1.006	1.058	0.997	0.983	0.999	1.014
ACC/GGT	1.072	1.052	0.993	1.036	1.010	0.946	1.024	1.027	**1.122**	**1.087**
ACG/CGT	0.984	0.989	1.078	1.081	1.023	0.998	0.996	**1.104**	1.050	1.064
AGA/TCT	1.030	0.998	0.923	0.898	0.970	*0.899*	0.967	0.921	0.997	1.006
AGC/GCT	1.032	1.045	1.072	**1.119**	1.049	1.098	1.055	1.082	0.981	1.013
AGG/CCT	0.894	0.912	0.941	1.011	1.005	0.995	0.974	1.041	1.035	0.945
TAA/TTA	1.015	1.017	0.959	0.992	1.048	1.091	**1.081**	1.063	0.934	0.965
TAC/GTA	1.045	1.006	1.040	1.079	0.997	0.956	1.012	1.007	1.040	1.063
TAG/CTA	*0.882*	0.914	0.914	1.038	*0.904*	*0.814*	*0.931*	0.938	0.945	0.911
TTC/GGA	0.971	0.978	0.965	0.966	0.983	0.975	0.965	1.976	1.012	1.017
TTG/CAA	1.089	1.024	0.984	0.926	0.962	0.915	0.943	0.880	1.016	1.001
TCA/TGA	1.015	1.017	**1.122**	**1.176**	0.998	0.975	0.999	0.993	1.048	1.080
TCC/GGA	1.012	1.010	0.984	0.972	1.019	1.070	1.021	1.031	0.965	0.939
TCG/CGA	0.916	0.970	0.988	0.981	**1.054**	**1.141**	0.984	0.912	0.967	0.941
TGC/GCA	0.947	0.970	0.941	0.908	0.980	0.968	0.941	0.945	1.002	0.920
TGG/CCA	1.093	1.093	1.095	1.049	**1.073**	1.098	**1.105**	**1.165**	**1.079**	**1.115**
CAC/GTG	0.962	0.993	1.025	1.069	1.029	1.088	1.012	1.086	1.035	1.034
CAG/CTG	0.883	0.931	1.037	1.007	1.046	1.069	1.059	1.070	1.036	1.024
CTC/GAG	**1.123**	**1.122**	1.069	1.102	**1.054**	**1.103**	1.056	1.062	1.043	1.017
CCC/GGG	*0.826*	*0.839*	*0.896*	*0.876*	*0.926*	0.923	*0.933*	*0.870*	*0.809*	*0.910*
CCG/CGG	**1.264**	**1.194**	0.981	1.003	0.928	0.928	0.960	0.974	0.941	0.968
CGC/GCG	0.954	*0.888*	0.929	*0.888*	0.965	0.905	1.053	0.993	1.021	1.046
GAC/GTC	0.979	0.970	0.971	0.941	1.003	0.943	1.001	0.912	1.015	0.952
GCC/GGC	1.088	1.079	1.100	1.104	1.002	1.020	1.007	1.071	0.973	1.074
	r = 0,94		r = 0,80		r = 0,88		r = 0,76		r = 0,73	

En italique sur fond gris et en gras sont représentées les valeurs déterminant les quatre indices correspondant aux deux trinucléotides significativement sous- et sur-représentés, respectivement. La sous- et la sur-représentation sont déterminées en fonction de la déviation de l'indice par rapport à 1 (voir paragraphe 3.1.).

r indique le coefficient de corrélation entre les indices des ETs et des séquences génomiques pour une espèce.

les autres espèces : chez la levure, les deux couples les plus sur-représentés (ACC/GGT et TGG/CCA) et les deux couples les plus sous-représentés (CCC/GGG et ACA/TGT) sont retrouvés chez les ETs et le génome. il en est de même chez *C. elegans* et *H. sapiens,* le couple le plus sur-représenté (TCA/TGA pour le nématode et TGG/CCA pour l'homme) et celui le plus sous-représenté (TAG/CTA pour le nématode et CCC/GGG pour l'homme) sont retrouvés chez les ETs et les génomes. Chez la drosophile, les ETs et le génomes ont aussi en commun les couples les plus sur-représentés (TCG/CGA et CTC/GAG) et un des couples le plus sous-représenté (TAG/CTA).

D'une manière globale, le calcul des coefficients de corrélation entre ETs et génome montre une corrélation positive chez chaque espèce : r = 0,94 ($p<0,05$) pour *Arabidopsis*, r = 0,80 ($p<0,05$) pour *Caenorhabditis*, r = 0,88 ($p<0,05$) pour *Drosophila*, r = 0,76 ($p<0,05$) pour l'homme et r = 0,73 ($p<0,05$) pour *Saccharomyces*. Chez *A. thaliana, C. elegans, H. sapiens, S. cerevisiae* et les ETs de *D. melanogaster*, CCC/GGG fait partie des couples de trinucléotides le plus sous-représenté.

A partir des valeurs des indices d'abondance en trinucléotides, nous avons calculé des distances entre génomes et ETs pour chaque espèce. Elles sont données dans le Tableau 16.

Tableau 16 : Distance relative δ^* entre génomes et ETs (multipliée par 1000)

		A. thaliana		***C. elegans***		***D. melanogaster***		***H. sapiens***		***S. cerevisiae***	
		génome	ETs	génome	ETs	génome	ETs	génome	ETs	génome	ETs
A. thaliana	génome	0	**23.43**	56.70	46.05	58.69	41.20	61.97	47.24	48.22	53.13
	ETs		0	75.08	60.66	76.87	56.57	70.95	60.07	55.51	62.97
C. elegans	génome			0	**35.23**	59.08	54.91	53.24	54.91	51.26	61.53
	ETs				0	49.89	40.42	53.42	44.57	<u>34.08</u>	48.63
D. melanogaster	génome					0	**35.07**	54.55	44.78	66.36	70.87
	ETs						0	44.06	<u>23.14</u>	43.12	38.51
H. sapiens	génome							0	**37.24**	54.99	57.07
	ETs								0	45.16	47.82
S. cerevisiae	génome									0	**31.32**
	ETs										0

En gras sont représentées les distances les plus faibles pour chaque génome.

D'après ce tableau, le génome et les ETs d'une même espèce montrent la distance la plus faible par rapport aux génomes. On peut noter la faible distance entre les ETs de *Drosophila melanogaster* et les ETs d'*Homo sapiens,* ainsi qu'entre le génome de *Saccharomyces cerevisiae* et ceux de *Caenorhabditis elegans* (valeurs soulignées dans le Tableau 16).

A partir de ces données, nous avons réalisé une ACO. La Figure 15 montre la projection sur les deux premiers axes. Comme dans le cas de l'ACO sur les dinucléotides et comme le laissait supposer le Tableau 16, on observe une proximité des ETs et des génomes d'une même espèce. On observe que les ETs de *H. sapiens* et *D. melanogaster* sont très proches.

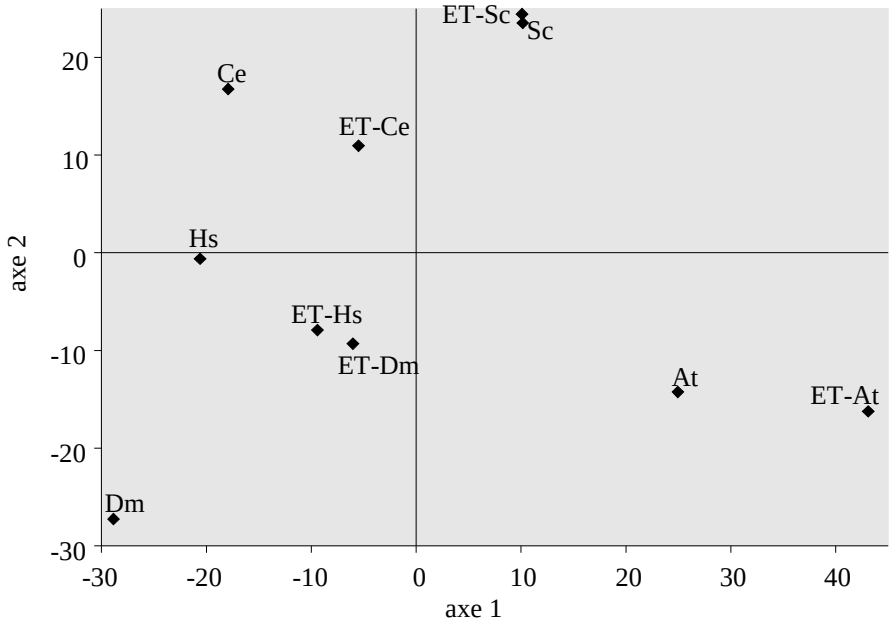


Figure 15 : Projection sur les deux premiers axes d'une ACO sur la distance en trinucléotides des génomes et des ETs. Les pourcentages des valeurs propres pour les axes 1 et 2 sont 36% et 23,5% respectivement. ET-At = ETs d'*A. thaliana*, ET-Ce = ETs de *C. elegans*, ET-Dm = ETs de *D. melanogaster*, ET-Hs = ETs d'*H. sapiens*, ET-Sc = ETs de *S. cerevisiae*, At = génome d'*A. thaliana*, Ce = génome de *C. elegans*, Dm = génome de *D. melanogaster*, Hs = génome d'*H. sapiens* et Sc = génome de *S. cerevisiae*.

3.4.1.2.3. Utilisation de fragments génomiques et des ETs complets

Afin d'avoir un aperçu de la variabilité de ce que l'on a observé au paragraphe précédent, nous avons à nouveau utilisé les mêmes fragments génomiques dépourvus d'ETs présentés au paragraphe 3.4.1.2.2.. Après calcul des indices d'abondance relative en trinucléotides des fragments génomiques et des ETs complets, nous avons calculé les distances δ^* entre chaque séquence, à partir desquelles nous avons effectué une ACO. Le résultat de cette analyse est présenté sur la Figure 16.

La projection de l'ACO est présentée sur cinq graphes différents (Figures 16 A à 16E) correspondant aux cinq espèces. Tous ces graphes sont superposables. On observe que les fragments génomiques et les ETs d'une espèce sont superposés. D'une manière générale, les points correspondant aux ETs sont répartis sur l'ensemble du nuage de points des fragments. On peut cependant noter que chez *D. melanogaster* et *H. sapiens*, les ETs se rassemblent et se superposent

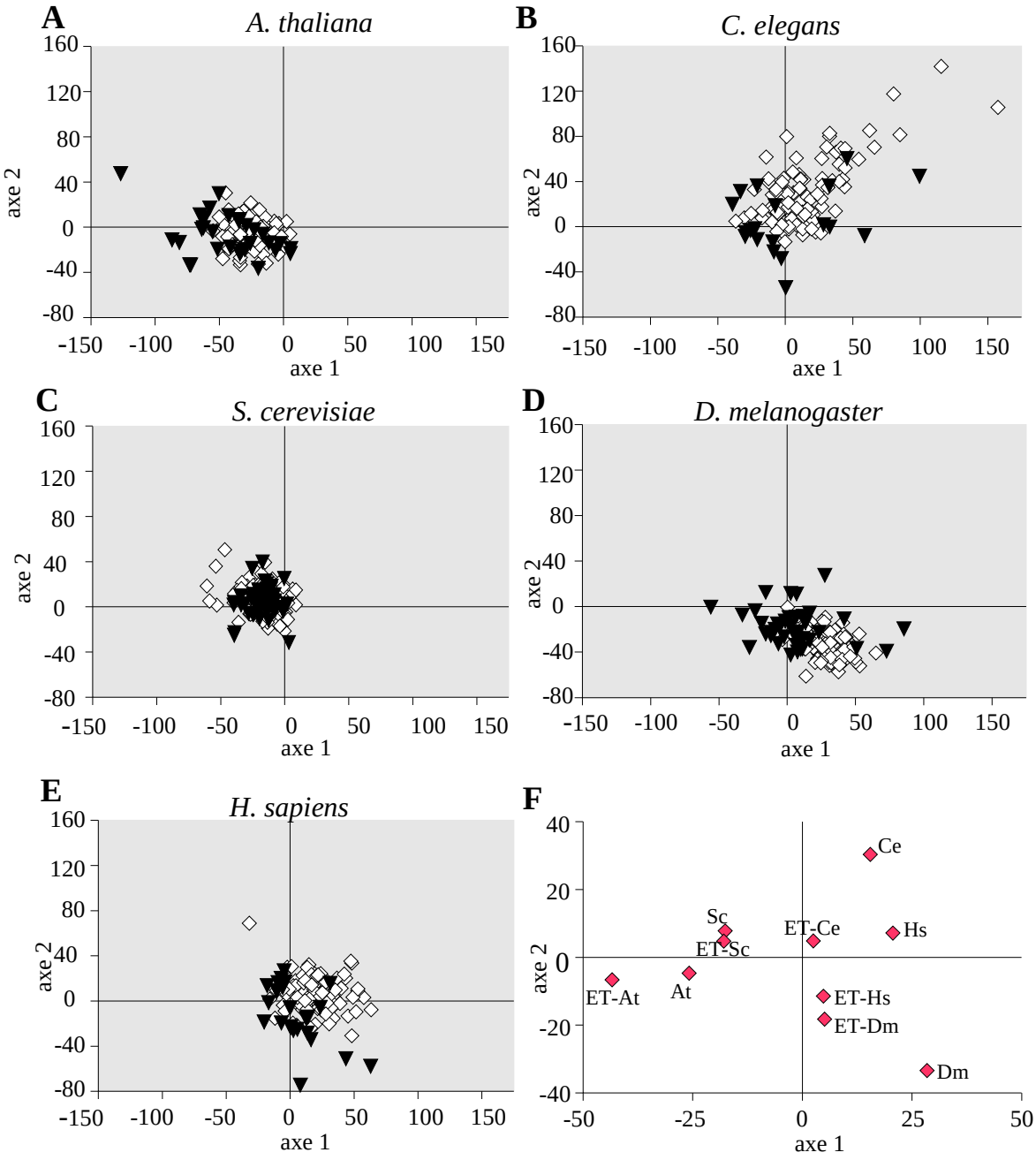


Figure 16 A à E : Projection d'une ACO sur la distance relative en trinucléotides de fragments génomiques (losanges blancs) et de séquences d'ETs complets (triangles noirs). Les cinq figures sont superposables. Le pourcentage de variabilité des axes 1 et 2 sont 13% et 10% respectivement. La **Figure 16 F** représente la position de la moyenne des coordonnées des fragments génomiques et des ETs pour chaque espèce.

au niveau d'une région sur le nuage de points des fragments génomiques. Ainsi des MANOVA effectuées sur les fragments génomiques et les ETs pour chaque espèce montrent que ETs et fragments forment des groupes distincts, excepté *S. cerevisiae*. Une MANOVA effectuée sur toutes les espèces en considérant comme au paragraphe 3.4.1.2.2. ETs et fragments génomiques d'une espèce donnée comme un seul groupe montre que tous les groupes espèces sont différents. Ainsi, ETs et fragments génomiques forment un groupe caractéristique de l'espèce tout en formant des sous-groupes distinct au sein de l'espèce.

La Figure 16F montre la projection de la moyenne des coordonnées pour les fragments génomiques et pour les ETs de chaque espèce. On retrouve la même disposition que dans la Figure 15 avec notamment une faible distance entre ETs de *D. melanogaster* et *H. sapiens*.

3.4.1.4. Conclusion

Ces études montrent que les ETs et les génomes d'hôte d'une espèce donnée présentent globalement le même « pattern » en dinucléotides et en trinucléotides. On observe une caractéristique partagée par l'ensemble des ETs et des génomes en ce qui concerne les dinucléotides, à savoir la sous-représentation du dinucléotide TpA. Cette sous-représentation est retrouvée pour un grand nombre d'organismes et pourrait provenir de contraintes structurales de l'ADN (Karlin et Mrázek 1997). Cependant, chez les rétrovirus humains, les ETs de la levure et certains rétrotransposons à LTR de la drosophile, une telle sous-représentation n'apparaît pas.

Chez *Arabidopsis* et l'homme, on peut remarquer une sous-représentation chez les ETs et dans le génome, du dinucléotide CpG. Ces espèces possèdent un système de méthylation* sur la Cytosine des doublets CpG qui a pour conséquence la mutation de la Cytosine en Thymine. On devrait donc s'attendre à une augmentation des couples TpG/CpA, ce que nous n'observons pas. On observe une sur-représentation de TpG/CpA chez certains ETs de l'arabette, ainsi que chez certains LINEs et certains transposons de l'homme (Tableau 14). Ceci pourrait indiquer que ces différents éléments subissent un contrôle par méthylation*. D'après Karlin et Burge (1995), qui retrouvent la sous-représentation en CpG dans le génome des mitochondries, cette réduction permettant de réduire les énergies d'empilement* de l'ADN pourrait faciliter la réplication et la transcription de l'ADN. Ce gain de rapidité pourrait être avantageux pour un génome de petite taille.

Pour les trinucléotides, on retrouve des caractéristiques généralisables à l'ensemble des organismes. Ainsi, la sous-représentation du couple CCC/GGG est présente dans l'ensemble des organismes. Cette sous-représentation est même une des deux plus importantes, excepté dans le génome de la drosophile. On retrouve aussi une tendance à la sur-représentation des couples

CTC/GAG et TGG/CCA chez les cinq organismes, dans les ETs et les génomes.

Ainsi, lorsqu'on considère la fréquence de dinucléotides ou de trinucléotides, indépendamment de la composition en bases, on observe que les ETs se comportent de la même manière que leur génome hôte. Ceci indique l'existence de contraintes structurales imposées par le génome hôte, indépendantes de celles qui permettent aux ETs d'avoir globalement la même composition en base.

3.4.2. Les rétrovirus et les éléments rétroviraux

Les études précédentes ont été réalisées en considérant d'une manière globale les parties codantes et non codantes des génomes et des ETs. Comme nous l'avons vu aux paragraphes 3.1. et 3.2., nous pouvons calculer un indice d'abondance relative en dinucléotides au niveau des parties codantes et en tenant compte de la position dans le codon. Ceci détermine une signature de codons.

3.4.2.1. Les données

Trois cents gènes pour chacune des cinq espèces ont été tirés au sort parmi ceux dont la description est donnée dans le paragraphe 2.2.2.1. (2ème partie). Il s'agit de gènes ayant une fonction bien définie et qui ne sont pas des gènes prédits.

Les parties codantes des ETs sont celles des éléments décrits au paragraphe 3.4.1.1., dont la liste est donnée en annexe dans le Tableau 5'. Les ETs ne possédant pas de parties codantes, comme par exemple les éléments CACTA ou SINE, ont été éliminés. Au total, nous obtenons 18 séquences d'ETs pour *C. elegans*, 59 pour *A. thaliana*, 58 pour *D. melanogaster*, 92 pour *S. cerevisiae* et 66 pour *H. sapiens* dont 46 sont des gènes de rétrovirus (*HIV1*, *HIV2*, *HTLV1*, *HTLV2*, *HSRV*, *HERV-K*, *HERV-KC4* et v-oncogène).

3.4.2.2. Signature des codons des gènes d'hôte et des ETs

Afin de déterminer une signature de codons pour les parties codantes des gènes d'hôte et des ETs, nous avons calculé l'abondance relative en dinucléotides au niveau des trois positions des codons : position 1-2 (1ère et 2ème bases du codon), position 2-3 (2ème et 3ème bases du codon) et position 3-1 (3ème base du codon et 1ère base du codon suivant). Les résultats sont présentés dans le Tableau 17.

Tableau 17 : Abondance relative en dinucléotides selon la position dans les codons pour les gènes d'ETs et les gènes des 5 espèces hôtes

	POSITION 1-2				
	At	Ce	Sc	Dm	Hs
AA	1.176	1.191	1.138	1.179	1.119
AT	0.985	1.113	0.937	1.093	0.904
AC	0.799	0.832	0.791	0.876	0.873
AG	1.017	*0.760*	1.126	*0.720*	1.098
TA	*0.524*	*0.533*	*0.462*	*0.600*	*0.579*
TT	**1.363**	**1.329**	**1.629**	**1.257**	**1.256**
TC	**1.317**	**1.313**	**1.276**	**1.335**	**1.241**
TG	0.898	0.959	*0.747*	0.962	1.034
CA	0.965	0.964	1.126	0.940	0.940
CT	1.141	0.991	0.864	1.103	**1.245**
CC	1.106	1.036	**1.274**	0.912	1.031
CG	*0.687*	1.050	*0.574*	1.094	*0.734*
GA	1.144	1.161	1.184	1.078	1.180
GT	*0.744*	*0.737*	*0.704*	*0.742*	*0.752*
GC	0.932	0.944	0.901	1.029	0.969
GG	1.221	1.165	**1.235**	1.176	1.088

	POSITION 2-3				
	At	Ce	Sc	Dm	Hs
AA	1.105	1.062	1.179	1.019	1.050
AT	0.799	0.937	0.861	1.089	0.913
AC	1.002	0.902	1.002	0.785	0.877
AG	1.175	1.108	0.989	1.195	1.136
TA	*0.573*	*0.342*	0.796	*0.615*	*0.561*
TT	1.027	1.153	0.967	0.940	0.966
TC	**1.256**	**1.362**	0.979	0.868	0.938
TG	1.159	**1.318**	**1.374**	**1.313**	**1.371**
CA	1.181	**1.287**	1.057	1.153	**1.364**
CT	**1.247**	0.970	1.119	0.794	**1.253**
CC	0.807	0.838	1.090	1.224	1.182
CG	*0.645*	0.792	*0.611*	*0.775*	*0.419*
GA	**1.253**	**1.506**	0.886	**1.303**	1.090
GT	0.993	0.928	1.195	1.222	0.900
GC	0.854	0.883	0.923	**1.320**	1.106
GG	0.927	*0.657*	0.954	0.482	0.923

	POSITION 3-1				
	At	Ce	Sc	Dm	Hs
AA	0.940	1.088	1.017	0.898	0.908
AT	0.925	1.083	1.013	1.208	0.789
AC	0.905	0.924	0.996	1.145	0.812
AG	1.122	0.893	0.947	0.827	**1.289**
TA	*0.656*	*0.592*	0.843	*0.609*	*0.586*
TT	1.018	0.958	1.071	0.904	0.975
TC	1.209	**1.294**	0.974	**1.241**	0.959
TG	1.177	1.209	1.144	1.210	**1.409**
CA	**1.256**	1.209	1.185	**1.320**	**1.348**
CT	1.169	1.026	0.966	1.133	**1.350**
CC	0.897	*0.748*	0.974	*0.707*	1.143
CG	*0.749*	0.974	0.856	0.888	*0.424*
GA	**1.275**	1.210	1.006	0.858	0.976
GT	0.933	0.933	0.921	0.837	0.810
GC	0.910	0.967	1.082	1.164	1.022
GG	0.861	0.885	1.006	1.089	1.091

	ET-At	ET-Ce	ET-Sc	ET-Dm	ET-Hs
AA	1.140	**1.291**	1.077	1.116	1.053
AT	0.941	0.944	1.059	0.931	0.970
AC	0.800	0.783	0.830	0.859	0.874
AG	1.083	0.886	1.090	1.053	1.062
TA	*0.519*	*0.557*	*0.580*	*0.545*	*0.562*
TT	**1.265**	1.189	**1.307**	**1.482**	**1.359**
TC	**1.281**	**1.290**	**1.406**	1.131	0.982
TG	1.102	1.110	*0.751*	0.998	**1.321**
CA	1.084	0.944	1.071	0.998	1.119
CT	1.080	1.036	0.822	1.007	1.053
CC	1.101	1.022	1.138	1.100	**1.254**
CG	*0.636*	0.973	*0.721*	0.839	*0.432*
GA	1.172	1.122	1.190	1.178	1.166
GT	0.794	0.872	0.817	*0.720*	*0.737*
GC	0.940	0.953	*0.756*	1.018	0.929
GG	1.077	1.108	**1.369**	1.084	1.215
r	0.93	0.90	0.91	0.71	0.78

	ET-At	ET-Ce	ET-Sc	ET-Dm	ET-Hs
AA	0.997	1.202	0.993	1.050	1.057
AT	0.972	0.921	0.981	0.940	0.955
AC	0.983	0.952	1.129	1.031	0.903
AG	1.048	0.880	0.890	0.946	1.120
TA	*0.711*	*0.695*	0.849	0.880	*0.774*
TT	1.039	1.007	0.979	1.125	1.108
TC	1.201	1.158	1.025	0.891	0.922
TG	1.130	1.204	**1.359**	1.155	**1.241**
CA	**1.278**	1.120	1.154	1.115	1.173
CT	1.131	1.064	1.015	0.970	1.146
CC	0.853	0.893	0.826	1.011	1.200
CG	*0.614*	0.848	0.841	0.794	*0.335*
GA	1.170	1.008	0.876	0.942	0.977
GT	0.856	1.018	1.101	0.934	*0.772*
GC	0.890	0.969	0.963	1.112	1.004
GG	1.086	1.095	1.088	1.176	**1.377**
	0.88	0.65	0.69	0.15	0.79

	ET-At	ET-Ce	ET-Sc	ET-Dm	ET-Hs
AA	1.023	1.219	0.845	1.051	1.067
AT	0.865	0.928	1.217	0.948	0.845
AC	0.927	0.815	1.148	0.952	0.816
AG	1.113	0.973	0.894	0.982	1.184
TA	*0.692*	*0.659*	0.910	0.899	0.791
TT	1.011	1.070	0.825	1.074	1.099
TC	**1.290**	1.214	1.070	0.942	1.043
TG	1.138	1.145	**1.236**	1.111	1.174
CA	**1.332**	1.079	**1.266**	1.102	**1.244**
CT	**1.234**	1.004	0.955	1.003	**1.250**
CC	0.878	0.995	0.845	1.048	1.168
CG	*0.499*	0.878	0.780	0.847	*0.379*
GA	1.092	1.080	1.188	0.900	0.899
GT	0.981	1.002	0.977	0.969	0.891
GC	0.860	0.975	*0.696*	1.115	1.029
GG	1.039	0.962	1.019	1.097	1.216
	0.88	0.77	0.43	0.39	0.91

En italique sur fond gris et en gras sont représentées les valeurs déterminant les indices significativement sous- (≤ 0,78) et sur-représentés (≥ 1,23), respectivement.
r indique le coefficient de corrélation entre les indices des ETs et les gènes pour une espèce à chaque position dans le codon.
At = gènes d'*A. thaliana*, Ce = gènes de *C. elegans*, Sc = gènes de *S. cerevisiae*, Dm = gènes de *D. melanogaster*, Hs = gènes de *H. sapiens*.
ET-At = ETs d'*A. thaliana*, ET-Ce = ETs de *C. elegans*, ET-Sc = ETs de *S. cerevisiae*, ET-Dm = ETs de *D. melanogaster*, ET-Hs = ETs de *H. sapiens*

Le Tableau 17 montre que d'une manière générale, on retrouve la sous-représentation du dinucléotide TpA, que l'on observait au niveau des séquences entières, dans les gènes et les ETs de toutes les espèces, principalement au niveau de la position 1-2. Le dinucléotide GpT est aussi significativement sous-représenté à la position 1-2 principalement au niveau des gènes d'hôte de toutes les espèces mais aussi dans les ETs de *D. melanogaster* et *H. sapiens*. On a aussi une forte sous-représentation du dinucléotide CpG principalement au niveau des positions 1-2 et 2-3 dans les gènes de toutes les espèces, excepté *C. elegans*. Cette sous-représentation est retrouvée à ces positions dans les ETs d'*A. thaliana, S. cerevisiae* et *H. sapiens*. On trouve une sur-représentation générale des dinucléotides TpT et TpC au niveau de la position 1-2 chez toutes les espèces, ETs et gènes d'hôte confondus. Les dinucléotides TpG et CpA sont sur-représentés d'une manière générale principalement au niveau de la position 2-3 mais aussi au niveau de la position 3-1, à la fois chez les ETs et les gènes d'hôte. Cette sur-représentation, qui existe chez *H. sapiens* et dans une moindre mesure chez *A. thaliana* peut s'expliquer par « l'évitement » du dinucléotide CpG, cible du système de méthylation*.

On observe un « pattern » légèrement différent entre la position 1-2 et les positions 2-3 et 3-1 pouvant être dû aux contraintes sur les acides aminés. En effet, comme nous l'avons vu dans la 2ème partie, du paragraphe 2.1.1., les deux premières positions du codon déterminent généralement à elles seules l'acide aminé qui sera codé.

On trouve à nouveau une bonne concordance entre le «pattern » de la signature des codons des ETs et des gènes de leur hôte. C'est particulièrement net au niveau de la position 1-2 où les valeurs des indices pour les gènes et pour les ETs d'une espèce donnée sont corrélées positivement. Pour les positions 2-3 et 3-1, il n'y a plus de corrélation entre ETs et gènes d'hôte chez *D. melanogaster* et *S. cerevisiae*.

3.4.2.3. Abondance relative en dinucléotide des gènes d'hôte et des ETs

Nous avons calculé les indices d'abondance en dinucléotide r pour les gènes et les parties codantes des ETs sans tenir compte cette fois de la position dans le codon. Pour chaque espèce, nous avons calculé les distances de l'abondance relative en dinucléotides d entre gènes et ETs, avant d'effectuer une ACO. Le résultat de ces analyses multivariées est présenté sur la Figure 17. Chaque graphe correspond à la projection des gènes et des ETs sur les deux premiers axes de l'ACO.

Pour *A. thaliana* et *C. elegans*, les ETs et les gènes d'hôte sont superposés. Des MANOVA effectuées pour chacune de ces deux espèces montrent que pour le nématode, les groupes des ETs et des gènes d'hôte ne sont pas différents. Pour *A. thaliana*, les deux groupes sont significativement

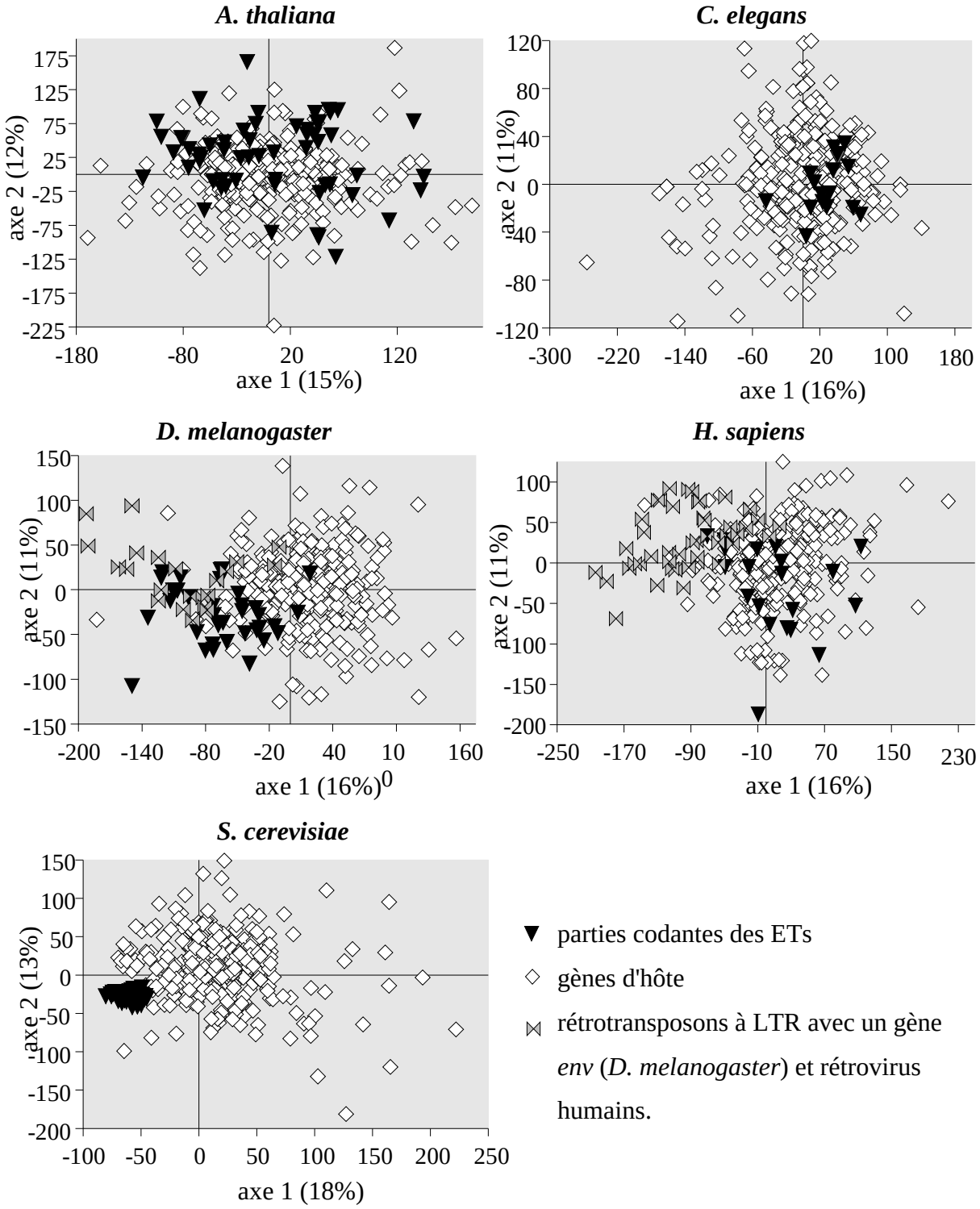


Figure 17 : Projection sur les deux premiers axes de cinq ACO différentes sur les distances relatives en dinucléotides pour les gènes d'hôte et pour les parties codantes des ETs. Les pourcentages de variabilité de chaque axe sont indiqués entre parenthèses

différents, cette différence étant due à une répartition différente de ces groupes sur l'axe 2. Ceci indique cependant un « pattern » proche. Dans le cas de *S. cerevisiae*, les ETs se regroupent et se séparent du nuage de points des gènes. Ce regroupement est principalement dû au fait que les séquences d'ETs utilisées présentent une forte homologie. Une MANOVA montre en effet que le groupe de ETs et celui des gènes d'hôte de la levure sont significativement différents. Dans le cas de *H. sapiens*, la plupart des ETs sont repartis dans le nuage de points des gènes, alors que les gènes de rétrovirus montrent une distance plus grande par rapport aux gènes d'hôte. Une MANOVA effectuée en considérant trois groupes : les rétrovirus, les ETs et les gènes d'hôte montrent que ces trois groupes sont significativement différents. Un test PLSD de Fisher sur l'axe 1 montre cependant que sur cet axe, ETs et gènes d'hôte ne sont pas significativement différents. Chez *D. melanogaster*, on retrouve un cas de figure similaire : certains ETs se regroupent au niveau du nuage de points des gènes, alors que les rétrotransposons à LTR avec un gène d'enveloppe montrent une distance plus importante. Une MANOVA a été effectuée en considérant trois groupes : les rétrotransposons à LTR avec un gène d'enveloppe, les autres ETs et les gènes d'hôte. Elle montre que ces trois groupes sont significativement différents. Ainsi, les rétroéléments de type infectieux se démarquent des autres ETs et des gènes d'hôte lors d'une ACO sur la distance relative en dinucléotides.

La Figure 18 montre le détail des ETs de *D. melanogaster* et *H. sapiens*. Ces deux graphes représentent la projection des ETs seuls d'après les graphes de la Figure 17. Dans le cas des éléments de la drosophile, les éléments les plus éloignés des gènes d'hôte sont les rétrotransposons *Tirant*, *ZAM*, *nomad* et *297*, et dans une moindre mesure, les éléments *gypsy* et *17.6*. Les éléments *Idefix*, *roo* et *BEL* sont les rétrotransposons à gènes d'enveloppe les plus près des gènes d'hôte. La projection des éléments d'*H. sapiens* montre que les rétrovirus les plus éloignés des gènes d'hôte sont *HTLV1* et *2*, *HERV* et certains gènes de *HIV1* et *2*. Les gènes du spuma rétrovirus *HSRV* montrent tous une distance faible avec les gènes d'hôte.

Afin d'établir la différence au niveau du « pattern » de dinucléotides entre rétroéléments de type infectieux, gènes d'hôte et ETs, nous avons indiqué dans le Tableau 18, la valeur des indices d'abondance relative en dinucléotides pour chaque classe ou famille d'ETs et pour les gènes d'hôte, chez les cinq espèces. On observe globalement une bonne adéquation entre les patterns des ETs et des gènes d'hôte, principalement chez *A. thaliana* et *C. elegans*. On retrouve la sous-représentation du dinucléotide CpG chez *Arabidopsis thaliana*, *Saccharomyces cerevisiae* et *Homo sapiens* pour les ETs et les gènes d'hôte. Une sous-représentation globale du dinucléotide TpA apparaît dans l'ensemble des ETs et des gènes d'hôte pour les cinq espèces, excepté les rétrovirus humains, les rétrotransposons à LTR avec un gène *env* et les rétrotransposons sans LTR de *D. melanogaster*, et les ETs de *S. cerevisiae*. L'absence de cette sous-représentation en TpA semble être la raison pour

Tableau 18 : Abondance relative en dinucléotides des gènes d'hôte et des parties codantes des ETs des cinq espèces suivant la classe et / ou la famille

	AA	AT	AC	AG	TA	TT	TC	TG	CA	CT	CC	CG	GA	GT	GC	GG
A. thaliana																
rétrotransposons LTR	1.045	0.909	0.948	1.090	*0.630*	1.064	**1.253**	1.190	**1.257**	1.184	0.846	*0.590*	1.130	0.911	0.934	0.986
rétrotransposons sans LTR	1.067	1.008	0.822	1.061	*0.675*	1.110	**1.256**	1.100	1.174	1.057	1.100	*0.568*	1.195	0.788	0.839	1.148
transposon	1.060	0.997	0.827	1.040	*0.684*	1.084	1.050	**1.303**	1.099	1.194	0.981	*0.601*	**1.265**	*0.760*	1.187	0.818
gènes d'hôte	1.068	0.910	0.900	1.092	*0.587*	1.075	**1.240**	1.189	1.127	1.200	0.933	*0.706*	**1.238**	0.863	0.923	0.952
C. elegans																
rétrotransposons LTR*	1.105	1.009	0.823	0.948	*0.599*	1.116	**1.256**	**1.293**	1.146	1.033	0.947	0.792	1.180	0.847	1.017	0.895
rétrotransposons sans LTR	**1.254**	0.939	0.819	0.933	*0.690*	1.044	1.202	1.172	0.984	1.074	1.010	0.877	1.129	0.913	0.961	1.004
transposon	1.224	0.912	0.943	0.855	*0.574*	1.114	**1.242**	1.189	1.190	0.934	0.869	0.958	1.008	1.033	0.934	1.028
gènes d'hôte	1.097	1.039	0.891	0.923	*0.488*	1.109	**1.318**	**1.242**	1.155	0.995	0.862	0.941	**1.273**	0.833	0.938	0.894
S. cerevisiae																
rétrotransposons LTR Ty1	0.976	1.087	1.001	0.913	0.798	0.957	1.165	**1.265**	1.146	0.994	0.950	*0.767*	1.151	0.896	0.806	1.118
rétrotransposons LTR Ty2	0.973	1.037	1.042	0.945	0.815	1.025	1.122	1.224	1.170	0.959	0.909	0.792	1.117	0.938	0.832	1.071
gènes d'hôte	1.112	0.930	0.933	0.983	*0.705*	1.162	1.045	1.200	1.107	1.024	1.086	*0.741*	1.105	0.874	0.985	1.012
D. melanogaster																
rétrotransposons LTR sans *env*	1.091	0.946	0.906	1.006	*0.736*	1.221	0.977	1.152	1.137	0.893	1.038	0.879	1.041	0.912	1.136	0.934
rétrotransposons LTR avec *env*	1.030	0.957	1.007	0.982	0.833	1.226	0.973	1.053	1.068	0.940	1.059	0.870	1.104	0.826	0.935	1.139
rétrotransposons sans LTR	1.052	0.955	0.948	1.009	0.808	1.069	1.004	1.216	1.065	1.140	0.968	0.792	1.005	0.832	1.114	1.065
transposon	1.214	0.912	0.833	0.899	*0.738*	**1.242**	0.928	1.200	1.029	0.987	1.161	0.819	0.946	0.846	1.229	1.092
gènes d'hôte	1.009	1.125	0.905	0.990	*0.568*	1.001	1.112	**1.269**	1.190	1.017	0.915	0.907	1.088	0.874	1.099	0.917
H. sapiens																
rétrovirus	1.006	0.938	0.869	1.194	0.828	1.154	0.957	1.134	1.152	1.145	**1.277**	*0.342*	1.022	0.782	0.945	**1.308**
rétrotransposons sans LTR	1.077	0.947	0.845	1.028	*0.586*	1.078	**1.254**	**1.474**	1.114	**1.434**	1.112	*0.264*	1.074	*0.602*	0.971	**1.321**
transposon	1.187	0.845	0.878	1.023	*0.472*	1.169	1.072	**1.462**	**1.241**	1.142	0.984	*0.503*	1.082	0.861	1.128	0.930
gènes d'hôte	1.019	0.904	0.875	1.156	*0.558*	1.016	1.044	**1.368**	1.196	**1.273**	1.103	*0.495*	1.134	0.812	1.003	1.018

* Seule une séquence de rétrotransposon à LTR chez *C. elegans* était disponible.

laquelle les rétroéléments potentiellement infectieux montrent une distance plus importante sur la Figure 18.

Afin de déterminer si on retrouve ce phénomène qui positionne les rétroéléments de type infectieux loin des gènes d'hôte, nous avons analysé les ETs et les gènes de *Drosophila virilis*. Les ETs sont présentés en annexe dans le Tableau 3'. Cette espèce de drosophile possède des rétroéléments de type *gypsy*. Après calcul de l'abondance relative en dinucléotides, nous avons calculé les distances entre gènes et ETs avant d'effectuer une ACO. Le résultat de la ACO est montré sur la Figure 19.

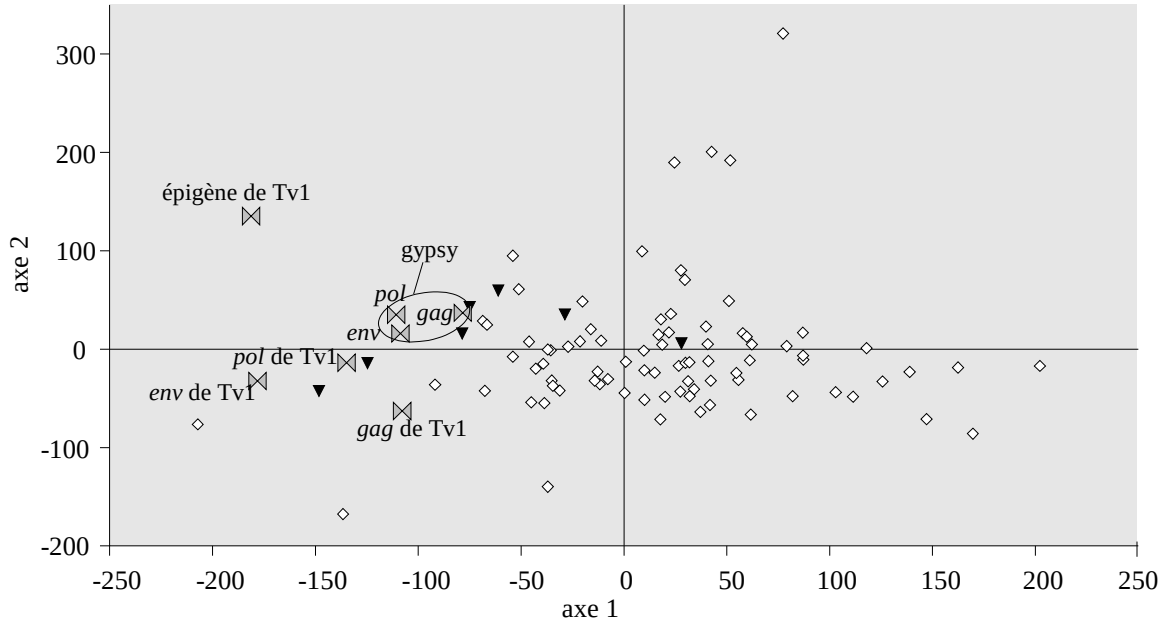


Figure 19 : Projection sur les deux premiers axes de l'ACO sur les distances relatives en dinucléotides pour les gènes d'hôte et pour les parties codantes des ETs de *Drosophila virilis*. Le pourcentage de variance pour les axes 1 et 2 sont 22% et 12% respectivement. Les triangles noirs correspondent aux ETs, les papillons gris aux ETs possédant un gène d'enveloppe et les losanges blancs aux gènes d'hôte.

On peut constater sur cette figure que les gènes des éléments *Tv1* et *gypsy*, deux rétrotransposons à LTR avec un gène *env*, montrent les distances les plus grandes par rapport aux gènes d'hôte. Ceci confirme la tendance des rétroéléments potentiellement infectieux à présenter une position particulière par rapport aux gènes d'hôte lors d'une ACO sur la distance en abondance relative en dinucléotides.

3.4.2.4. Conclusion

L'analyse de l'abondance relative en dinucléotides des gènes d'hôte et des parties codantes des ETs montre un « pattern » similaire au sein d'une espèce donnée. Cependant, chez l'homme et la drosophile, certaines séquences d'ETs présentent un « pattern » plus éloigné de celui des gènes d'hôte que les autres ETs. Il s'agit essentiellement des gènes des rétrovirus humains et des rétrotransposons à LTR avec un gène *env* de la drosophile. La position de ces rétroéléments semble due au fait que, contrairement aux séquences des gènes et des autres ETs, le dinucléotide TpA n'est pas particulièrement sous-représenté dans ces éléments. Ce regroupement particulier pourrait refléter la capacité infectieuse des rétrovirus. Ainsi, sous cette hypothèse, les rétrotransposons à LTR avec un gène *env* montrant un comportement similaire, à savoir une position éloignée des gènes d'hôte dans une ACO, pourraient se comporter comme des rétrovirus. Chez la drosophile, il a été démontré que l'élément *gypsy* a des propriétés infectieuses notamment par la présence d'un gène d'enveloppe actif (Kim *et al.* 1994). Cette position particulière des rétrovirus et des éléments de type rétroviraux n'est pas retrouvée quand on utilise les séquences entières, régions codantes et non codantes (voir paragraphe 3.4.1.2.). Ceci suggère des différences dans les mécanismes de transcription ou de traduction chez les rétrovirus par rapport aux gènes d'hôte.

On peut noter que les séquences de rétrovirus endogènes humains (*HERV*) sont aussi groupées avec les rétrovirus alors que ces éléments ne sont pas infectieux à cause de délétion ou de codons stop dans leurs parties codantes (Bock et Stoye 2000 ; Tristem 2000). Pourtant, un rétrovirus endogène du groupe *HERV-K* trouvé sur le chromosome 7 possède une seule mutation inactivant le gène *pol* alors que les gènes *gag* et *env* sont intacts (Mayer *et al.* 1999). La comparaison de ses LTR montre qu'elles diffèrent par 6 bp et que l'intégration dans le génome relativement récente, se serait produite il y a 1,2 millions d'années. Il pourrait ainsi se produire des *trans*-complémentations entre les éléments de ce groupe possédant seulement quelques mutations inactivantes, pour produire des éléments infectieux (Bock et Stoye 2000). Les éléments *17.6*, *Tirant*, *297*, *Idefix* et *nomad* montrent un « pattern » particulier qui les placent dans l'ACO loin des gènes d'hôte. De plus, ils ne présentent pas de sous-représentation du dinucléotide TpA. Ces éléments pourraient donc être infectieux ou avoir été infectieux il y a peu de temps. Par cette méthode de l'ACO sur l'abondance relative en dinucléotides, il semble donc possible de déterminer si certains rétroéléments peuvent ou non être infectieux. Il est possible que la position des *HERV* reflète une potentialité perdue récemment. Il peut donc en être de même pour les ETs de la drosophile : certains ETs sont peut être en train de devenir infectieux ou bien viennent de perdre cette capacité.

3.5. Conclusion

Les différentes études concernant l'abondance relative en di- et en trinucléotides montrent que lorsque l'on s'affranchit du biais de composition en bases, on observe un pattern similaire entre les ETs et le génome hôte d'une espèce donnée. Ce résultat est à mettre en relation avec le fait que les ETs se regroupent selon leur espèce hôte lorsque l'on observe leur usage des codons, indépendamment du biais de composition en base. La similitude de « pattern » entre ETs et génome hôte provient vraisemblablement de contraintes structurales imposées par le génome hôte. La structure de l'ADN est influencée par différents facteurs comme l'énergie d'empilement* des dinucléotides, la courbure, la superhélicité*, la méthylation* et les mécanismes de réparation (KARLIN et LADUNGA 1994). Ainsi, un génome donné va subir certains types de contraintes qui ne seront pas forcément les mêmes dans le génome d'une autre espèce. Un ET inséré dans un génome depuis longtemps subit aussi ce type de contraintes. Une manière de pouvoir envahir le génome hôte est de s'adapter aux contraintes et d'adopter un « pattern » similaire. Une étude a été effectuée sur la fréquence en oligomères 3-6 dans des virus humains et des bactériophages par rapport à des séquences humaines et bactériennes (BARRAI *et al.* 1990). Cette étude montre une corrélation positive au niveau de la fréquence en oligomères entre virus et hôte. Une hypothèse avancée pour expliquer cette corrélation est que les premiers gènes des virus sont transcrits par une enzyme de l'hôte. Ces gènes codent généralement pour une ADN polymérase spécifique du virus. Ainsi, pour pouvoir utiliser l'enzyme de l'hôte, le virus doit posséder des séquences signal reconnues par l'enzyme. Cependant, ce type de signaux est rare dans une séquence et n'apparaît généralement qu'une fois. Ce n'est donc pas suffisant pour expliquer la corrélation. Dans le cas des bactériophages, l'intérêt d'être constitués d'une séquence possédant les mêmes caractéristiques que les séquences d'hôte est un bon moyen d'échapper au système de défense constitué par les enzymes de restriction. Ainsi, posséder les mêmes caractéristiques que l'hôte peut être un avantage pour envahir le génome.

Une étude effectuée sur différents types de virus a montré qu'il existe une différence entre les virus possédant un petit génome et ceux possédant un grand génome en ce qui concerne la sous-représentation en CpG (KARLIN *et al.* 1994). Ainsi, on trouve une sous-représentation plus importante en CpG dans les virus de petits génomes. Cette suppression en CpG n'est pas retrouvée dans l'ensemble des ETs. On l'observe particulièrement dans les ETs d'*Arabidopsis*, les éléments *Ty1*, *Ty4* et *Ty5* de la levure et chez certains ETs du nématode. Chez la drosophile, très peu d'ETs présentent cette sous-représentation. Deux possibilités peuvent expliquer cette suppression : premièrement, il peut s'agir de la conséquence de la méthylation*. Cependant, la levure ne présente

pas de système de méthylation* et cela n'explique pas pourquoi il y a une relation entre taille des génomes viraux et suppression de CpG chez l'homme. Deuxièmement, l'évitement de ce dinucléotide peut être la conséquence de sa forte énergie d'empilement* qui en fait un dinucléotide peu malléable au niveau structural. Chez les virus de petite taille, l'évitement de CpG peut se révéler être un avantage afin d'augmenter la rapidité de leur réplication. Cependant, s'il y a un lien avec la taille, on devrait retrouver cette suppression chez les ETs.

L'analyse effectuée sur les parties codantes des gènes montrent qu'il y a d'avantage de différences entre ETs et gènes d'hôte que lorsque l'on considère les séquences entières. Ceci est particulièrement flagrant chez *D. melanogaster* et *H. sapiens*. Ainsi, les rétrovirus humains, infectieux et endogènes, et les rétrotransposons à LTR avec un gène d'enveloppe de la drosophile montrent une position particulière par rapport aux gènes d'hôte sur la projection d'une ACO. Cette position particulière semble être due au fait qu'il n'y a pas de suppression du dinucléotide TpA au niveau des gènes des rétrovirus et des rétrotransposons à LTR avec un gène d'enveloppe, alors que cette sous-représentation est généralisable à l'ensemble des organismes et des autres ETs, excepté les éléments de la levure. Ce dinucléotide TpA est le moins stable énergétiquement et il est associé à certaines distorsions de l'ADN. Son occurrence au niveau de sites de nucléation permet une ouverture plus facile de la double hélice. Les sites cibles de la plupart des ETs, notamment les éléments de classe II, contiennent souvent des successions de TpA. Il peut sans doute y avoir une relation entre le choix de ces sites et le fait que le dinucléotide TpA permet une ouverture plus facile de l'ADN. On trouve une suppression de TpA au niveau de la position 1-2 chez les rétrovirus humains et les rétrotransposons à LTR avec un gène *env* de la drosophile mais pas au niveau des deux autres positions (résultats non montrés). Une suppression à cette position peut être le résultat de l'évitement de l'occurrence inopportune de codons stop, TAA ou TAG, dans les parties codantes, cette contrainte ne se produisant pas au niveau des autres positions. Si cette caractéristique est due à l'infectiosité des rétrovirus, alors les rétroéléments de drosophile qui possèdent cette caractéristique sont susceptibles d'être infectieux. Par cette méthode, nous pouvons donc avoir un moyen de détecter des rétroéléments potentiellement infectieux. Des analyses expérimentales sur les autres rétroéléments possédant cette position particulière dans l'ACO permettraient de valider cette hypothèse.

4ème PARTIE : CONCLUSION GENERALE ET PERSPECTIVES

4ème partie : Conclusion générale et perspectives

Conclusion générale

Le but de ce travail de thèse a été d'analyser l'évolution moléculaire des ETs en relation avec d'éventuelles contraintes exercées par les génomes hôte. Ce travail s'est articulé en deux grandes parties : l'une relative à l'usage des codons des ETs en fonction de celui des gènes de l'hôte, l'autre traitant de la présence d'une signature génomique au sein des séquences d'ETs. Les résultats principaux de ces deux parties sont les suivants :

- l'analyse de l'usage des codons a permis de montrer que les ETs présentent une caractéristique commune. Quelle que soit l'espèce hôte, ils sont riches en AT. Il apparaît clairement que l'usage des codons ne peut pas être utilisé pour détecter d'éventuels transferts horizontaux entre espèces de drosophiles, notamment dans le cas de l'élément *P*.
- la richesse en AT est particulièrement marquée au niveau de la 3ème position des codons. Il s'agit de la position la moins contrainte. La comparaison des pourcentages en AT à cette position par rapport aux zones non codantes des génomes montre que les parties codantes des ETs ne subissent pas uniquement un biais mutationnel mais également des pressions de sélection. En effet, le pourcentage en AT à la 3ème position des codons de la plupart des ETs est inférieur à celui des zones non contraintes.
- il ne semble pas y avoir de lien entre la richesse en AT au niveau de la 3ème position des codons et le niveau d'expression des ETs, ce qui peut indiquer un patron d'expression particulier.
- si on élimine le biais de composition en base des ETs et des gènes d'hôte, on observe que les ETs présentent un usage des codons différents des gènes de l'hôte. Cependant, les ETs se regroupent suivant l'espèce hôte, ce qui indique l'existence de contraintes spécifiques à l'hôte.
- l'analyse de l'abondance relative en di- et trinucléotides révèlent des similitudes entre les ETs (parties codantes et non codantes) et le génome de leur hôte. En effet, il y a concordance entre le « pattern » de certains ETs et celui des gènes d'hôte. Cependant, on observe que les rétrovirus humains et les rétrotransposons à LTR de la drosophile contenant un gène d'enveloppe, se différencient des gènes d'hôte.

En 1989, Shields et Sharp ont analysé l'usage des codons de 22 ORFs de 11 ETs de *Drosophila melanogaster* (deux transposons, huit rétrotransposons et un élément de classe III). Ils conclurent que la richesse en AT qu'ils observaient chez les rétrotransposons provenait de « patterns » de mutation dus aux erreurs d'incorporation de la transcriptase inverse qui tend à

incorporer des A et des T. Cependant, ils n'expliquaient pas la richesse en AT des transposons en particulier à cause de la petite taille de l'échantillon. Notre étude montre que cette richesse en AT n'est pas uniquement le résultat d'un biais mutationnel mais que les ETs sont sujets à des contraintes sélectives. De plus, la généralisation de la richesse en AT aux ETs d'organismes différents semble indiquer qu'il s'agit d'un mécanisme propre aux ETs qui pourrait être en relation avec la capacité à transposer ou avec les mécanismes de transposition.

Une analyse récente de séquences de la famille de rétrovirus endogènes humains *K10* (*HERV-K10*) et de rétrovirus infectieux, a montré que ces éléments sont particulièrement riches en AT (Zsìros *et al.* 1999). De plus, alors que les *HERV* sont présents dans le génome des primates depuis environ 30 millions d'années, date de la divergence des singes du Nouveau Monde, ils possèdent un usage des codons différents de celui des gènes d'hôte et subissent une pression de sélection permettant la persistance de leurs ORFs. En effet, l'analyse des taux de mutations silencieuses et non silencieuses montre qu'il y a un biais pour les changements silencieux, c'est à dire les changements qui permettent de garder le même enchaînement d'acides aminés. Ainsi, il s'exerce ou s'est récemment exercée une pression de sélection sur les membres de la famille *HERV-K10*, qui est corrélée avec la persistance des ORFs de certains membres de cette famille. Très récemment, une séquence de rétrovirus endogène de la famille *HERV-W* a été trouvée dans le génome séquencé de l'homme. Ce rétrovirus a la particularité de posséder une ORF intacte codant pour une protéine d'enveloppe (Voisset *et al.* 2000). Ces résultats sont en accord avec nos observations et suggèrent l'existence d'une pression de sélection s'exerçant sur les ORFs des ETs.

Le fait que l'on n'observe pas de lien entre un taux d'expression fort ou faible et l'usage des codons des ETs suggère un patron d'expression spécifique aux ETs. Il y a sans doute un lien entre la persistance de la richesse en AT et leur taux d'expression. Malgré l'indépendance apparente des ETs concernant les compositions nucléotidiques et l'usage des codons, il existe néanmoins une influence de l'hôte lorsque l'on s'affranchit du biais de composition en bases. Cette influence que l'on observe au niveau des « patterns » d'abondance relative en di- et en trinucléotides provient vraisemblablement de contraintes structurales que subit le génome hôte. Ces contraintes peuvent provenir de mécanismes de réparation ou de réplication spécifiques des organismes. Ceci explique les différences de « pattern » que l'on observe entre espèces, ainsi que la spécificité d'un « pattern » donné pour une espèce. Ainsi, les ETs insérés dans le génome hôte vont subir les mêmes contraintes et adopter les mêmes types de « pattern ». Cette particularité pourrait éventuellement être utilisée pour la détection de transferts horizontaux entre espèces ayant des contraintes différentes. En effet, un élément introduit par transfert horizontal présenterait le même « pattern » que son hôte d'origine. Karlin *et al.* (1998) ont combiné signature génomique et biais de composition en bases afin de

détecter des portions de génomes chez des bactéries pouvant avoir été sujettes à des transferts horizontaux. Cependant, les espèces proches ont un « pattern » proche. Il est donc difficile de déterminer des différences significatives sauf entre espèces très éloignées. De plus, nous n'avons aucune indication sur le temps qu'il faut à un ET pour adopter le « pattern » de l'hôte.

Une différence de « pattern » a pu être observée en comparant les gènes d'ETs et les gènes d'hôte. Ainsi, les gènes des rétrovirus humains et des rétrotransposons à LTR de drosophile ayant un gène *env* semblent se comporter différemment des autres ETs et des gènes d'hôte. Cette différence de comportement pourrait s'expliquer dans le cas des rétrovirus par leur capacité infectieuse. L'infectiosité d'un virus est sa capacité à pouvoir pénétrer dans un hôte potentiel. Il est possible que des « patterns » en di- ou en trinucléotides particuliers reflètent des contraintes spécifiques que subit le virus pour permettre son infectiosité. Ainsi, les rétrotransposons à LTR de la drosophile présentant le même type de comportement pourraient être, ou avoir été infectieux il y a peu de temps.

Perspectives

Afin de valider ces résultats obtenus chez cinq espèces différentes, nous pouvons envisager d'étendre ce type d'analyses à d'autres organismes. Notamment, le séquençage du génome du moustique *Anopheles gambiae* va nous apporter des informations supplémentaires pour élargir notre étude. Une analyse effectuée sur 14 gènes et quatre rétrotransposons de l'anophèle a montré que gènes d'hôte et ETs n'ont pas le même type d'usage des codons (Besansky 1993). Il s'agit de l'étude d'un nombre très restreint de données qui mérite d'être approfondie. De plus, il serait très intéressant d'effectuer une comparaison entre insectes en comparant l'anophèle et la drosophile pour déterminer si on a les mêmes types de mécanismes dans des espèces proches. En effet, dans cette étude, nous nous sommes intéressé à des espèces éloignées qui ne présentent pas les mêmes types de contraintes.

Un autre point à approfondir serait de déterminer si la sélection agit bien sur les gènes des ETs. Pour cela, le calcul des taux de substitutions synonymes K_s et non synonymes K_a, peuvent être calculés entre des ETs d'une même famille, au sein d'une même espèce ou d'espèces différentes. Ceci implique d'avoir un nombre suffisant de séquences et que les ETs utilisés soient actifs ou aient été récemment actifs et qu'ils ne présentent pas une trop forte homologie de séquences, comme cela est le cas pour les éléments du génome séquencé de la levure. Ainsi, en observant le rapport K_a/K_s, nous pouvons être en mesure de déterminer si une pression de sélection s'exerce sur des ETs. Concrètement, lorsque K_a est inférieur à K_s, ceci indique qu'il y a d'avantage de substitutions silencieuses dans le gène, donc qu'une pression de sélection négative s'exerce sur le gène. Lorsque

les deux taux sont équivalents, ceci indique qu'il n'y a pas de sélection. Enfin, lorsque le K_a est supérieur au K_s, ceci indique que le gène subit une sélection positive qui conduit à une augmentation de la variabilité de la protéine codée, comme par exemple au niveau des gènes d'immunoglobulines. Cette étude n'a pu être effectuée en raison du manque de disponibilité de séquences, cependant la prochaine mise à jour du génome de la drosophile devrait nous permettre d'avoir accès aux séquences des ETs du génome. Il sera donc possible d'effectuer cette analyse sur ces séquences ainsi que d'observer comment se comportent les gènes des ETs suivant différents degrés de dégénérescence.

L'étude de la dégénérescence des ETs peut également nous apporter des éléments de réponse sur la question de la richesse en AT et notamment si cette composition particulière en bases a un lien avec l'activité de l'élément. Dans le cas où la richesse en AT est liée à la fonctionnalité ou à l'autonomie de l'élément, la question qui se pose est de savoir quel avantage peut conférer cette composition à un ET. Il pourrait s'agir d'avantages structuraux puisqu'une séquence ADN riche en AT est plus flexible qu'une séquence riche en GC (TRAVERS 1989). Cette propriété pourrait s'avérer intéressante pour une intégration plus facile des ETs dans le génome. Il est aussi possible qu'un ET adopte une structure tridimensionnelle particulière lors de son déplacement qui soit avantagée par une grande richesse en AT.

Les interactions des ETs et de leur génome hôte montrent qu'il s'agit de deux entités intimement liés. Notamment, les ETs peuvent dans certains cas jouer un rôle fondamental dans l'évolution du génome de l'hôte. Bien qu'il existe une influence certaine de l'hôte sur les ETs, les ETs présentent des caractéristiques indépendantes des génomes hôtes, comme la richesse en AT. Ceci montre que les ETs peuvent aussi être considérés comme un compartiment à part dans le génome.

REFERENCES BIBLIOGRAPHIQUES

Références bibliographiques

ADAMS MD, CELNIKER SE, HOLT RA, *et al.* (188 co-authors) (2000) The genome sequence of *Drosophila melanogaster*. Science 287:2185-2195

AGRAWAL A, EASTMAN QM, SCHATZ DG (1998) Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. Nature 394:744-751

AKASHI H (1994) Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. Genetics 136:927-935

ALTSCHUL SF, MADDEN TL, SCHAFFER AA, ZHANG JH, ZHANG Z, MILLER W, LIPMAN DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389-3402

ASHBURNER M (1989) Drosophila. A laboratory manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, N. Y.

BARRAI I, SCAPOLI C, BARALE R, VOLINIA S (1990) Oligonucleotide correlations between infector and host genomes hint at evolutionary relationships. Nucleic Acids Res 18:3021-3025

BEAR G (1999) Darwin's radio. Ballantine books, New York, USA.

BENNETZEN JL, HALL BD (1982) Codon selection in yeast. J Bio Chem 257:3026-3031

BENIT L, DE PARSEVAL N, CASELLA JF, CALLEBAUT I, CORDONNIER A, HEIDMANN T (1997) Cloning of a new murine endogenous retrovirus, MuERV-L, with strong similarity to the human HERV-L element and with a *gag* coding sequence closely related to the *Fv1* restriction gene. J Virol 71:5652-5657

BENZÉCRI J-P (1973) L'analyse de données. Dunod, Paris.

BERG D, HOWE MM (1989) Mobile DNA. American Society for microbiology, Washington DC, USA.

BERNARDI G (2000) Isochores and the evolutionary genomics of vertebrates. Gene 214:3-17

BERNARDI G, OLOFSSON B, FILIPSKI J, ZERIAL M, SALINAS J, CUNY G, MEUNIER-ROTIVAL M, RODIER F (1985) The mosaic genome of warm-blooded vertebrates. Science 228:953-958

BESANSKY NJ (1993) Codon usage patterns in chromosomal and retrotransposon genes of the mosquito *Anopheles gambiae.* Insect Mol Biol 1:171-178

BEUTLER E, GELBART T, HAN JH, KOZIOL JA, BEUTLER B (1989) Evolution of the genome and the genetic code: selection at the dinucleotide level by methylation and polyribonucleotide cleavage. Proc Natl Acad Sci USA 86:192-196

BIÉMONT C, TSITRONE A, VIEIRA C, HOOGLAND C (1997) Transposable elements distribution in Drosophila. Genetics 147:1997-1999

BIESSMANN H, KASRAVI B, BUI T, FUJIWARA G, CHAMPION LE, MASON JM (1994) Comparison of two active *HeT-A* retroposons of *Drosophila melanogaster*. Chromosoma 103:90-98

BIESSMANN H, VALGEIRSDOTTIR V, LOFSKY A, CHIN C, GINTHER B, LEVIS RW, PARDUE ML (1992) *HeT-A*, a transposable element specifically involved in healing broken chromosome ends in Drosophila. Mol Cell Biol 12:3910-3918

BLACKBURN EH (1991) Structure and function of telomere. Nature 350:569-573

BOEKE JD, EICHINGER D, CASTRILLON D, FINK GR (1988) The *Saccharomyces cerevisiae* genome contains functional and nonfunctional copies of transposon *Ty1*. Mol Cell Biol 8:1432-1442

BOEKE JD, EICKBUSH T, SANDMEYER SB, VOYTAS DF (1998) Metaviridae. In Murphy FA (ed) Virus taxonomy: ICTV VIIth report. Springer-Verlag, New York

BOCK M, STOYE JP (2000) Endogenous retroviruses and the human germline. Curr Opin Genet Dev 10:651-655

BOWEN NJ, MCDONALD (1999) Genomic analysis of *Caenorhabditis elegans* reveals ancient families of retroviral-like elements. Genome Res 9:924-935

BRENNER S (1974) The genetics of *Caenorhabditis elegans*. Genetics 77:71-94

BRITTEN RJ (1995) Active *gypsy/Ty3* retrotransposons or retroviruses in *Caenorhabditis elegans*. Proc Natl Acad Sci USA 92:599-601

BRITTEN RJ (1996) DNA sequence insertion and evolutionary variation in gene regulation. Proc Natl Acad Sci USA 93:9374-9377

BROOKFIELD JFY, MONTGOMERY E, LANGLEY C (1984) Apparent absence of transposable elements related to *P* elements in *Drosophila melanogaster* and other species of Drosophila. Nature 310:331-332

BULMER M (1987) Coevolution of codon usage and tRNA abundance. Nature 325:728-730

BULMER M (1991) The selection-mutation-drift theory of synonymous codon usage. Genetics 129:897-207

BURGE C, CAMPBELL AM, KARLIN S (1992) Over- and under-representation of short oligonucleotides in DNA sequences. Proc Natl Acad Sci USA 89:1358-1362

CAMPBELL A, MRÁZEK J, KARLIN S (1999) Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. Proc Natl Acad Sci USA 96:9184-9189

CAPY P, BAZIN C, HIGUET D, LANGIN T (1997a) Dynamic and evolution of transposable elements. R.G. Landes company, Austin, Texas, USA.

CAPY P, GASPERI G, BIÉMONT C, BAZIN C (2000) Stress and transposable elements: co-evolution or useful parasites? Heredity 85:101-106

CAPY P, LANGIN T, HIGUET D, MAURER P, BAZIN C (1997b) Does the integrase of LTR-retrotransposons and most of the transposases of class II elements share a common ancestor? Genetica 100:63-72

CAPY P, VITALIS R, LANGIN T, HIGUET D, BAZIN C (1996) Relationships between transposable elements based upon the integrase-transposase domains: is there a common ancestor? J Mol Evol 42:359-369

CARULLI JP, KRANE DE, HARTL DL, OCHMAN H (1993) Compositional heterogeneity and patterns of molecular evolution in the *Drosophila* genome. Genetics 134:837-845

CHIAPELLO H, LISACEK F, CABOCHE M, HÉNAUT A (1998) Codon usage and gene function are related in sequences of *Arabidopsis thaliana*. Gene 209:GC1-GC38

COMERON JM, AGUADÉ M (1998) An evaluation of measures of synonymous codon usage bias. J Mol Evol 47:268-274

COMERON JM, KREITMAN M, AGUADÉ M (1999) Natural selection on synonymous sites is correlated with gene

length and recombination rate in Drosophila. Genetics 151:239-249.

DANIELS SB, PETERSON KR, STRAUSBAUGH LD, KIDWELL MG, CHOVNICK A (1990) Evidence for horizontal transmission of the *P* element between Drosophila species. Genetics 124:339-355

DAWKINS R (1990) Le gène égoïste. Edts Armand Collin

DERAGON J-M, CAPY P (2000) Impact of transposable elements on the human genome. Ann Med 32:264-273

DUNHAM I, SHIMIZU N, ROE BA, *et al.* (217 co-authors) (2000) The DNA sequence of human chromosome 22. Nature 402:489-495

DURET L (2000) tRNA gene number and codon usage in the *C. elegans* genome are co-adapted for optimal translation of highly expressed genes. Trends Genet 16:287-289.

DURET L, GALTIER N (2000) The covariation between TpA deficiency, CpG deficiency, and G+C content of human isochores is due to a mathematical artifact. Mol Biol Evol 17:1620-1625

DURET L, MARAIS G, BIÉMONT C (2000) Transposons but not retrotransposons are found preferentially in regions of high recombination rate in *Caenorhabditis elegans.* Genetics 156:1661-1669

DURET L, MOUCHIROUD D (1999) Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. Proc Natl Acad Sci USA 96:4482-4487.

EICKBUSH TH (1994) Origin and evolutionary relationships of retroelements. in MORSE SS (ed) the evolutionary biology of viruses, Raven Press, Ltd. New York, pp. 121-157

EMMONS SW, YESNER L, RUAN KS, KATZENBERG D (1983) Evidence for a transposon in *Caenorhabditis elegans*. Cell 32:55-65

ESPOSITO T, GIANFRANCESCO F, CICCODICOLA A, MONTANINI L, MUMM S, D'URSO M, FORABOSCO A (1999) A novel pseudoautosomal human gene encodes a putative protein similar to *Ac*-like transposases. Hum Mol Genet 8:61-67

EVGEN'EV MB, ZELENTSOVA H, POLUECTOVA H, LYOZIN GT, VELEIKODVORSKAJA V, PYATKOV KI, ZHIVOTOVSKY LA, KIDWELL MG (2000) Mobile elements and chromosomal evolution in the *virilis* group of Drosophila. Proc Natl Acad Sci USA 97:11337-11342

FESCHOTTE C, MOUCHÈS C (2000) Evidence that a family of Miniature Inverted-repeat Transposable Elements (MITEs) from the *Arabidopsis thaliana* genome has arisen from a *pogo*-like DNA transposon. Mol Biol Evol 17:730-737

FINNEGAN DJ (1989) Eucaryotic transposable elements and genome evolution. Trends Genet 5:103-107

FINNEGAN DJ (1992) Transposable elements. Curr Opin Genet Dev 2:861-867

FLAVELL AJ (1981) Did retroviruses evolve from transposable elements? Nature 289:10-11

GAMA SOSA MA, HALL JC, SCHNEIDER KE, LUKASZEWICZ GC, RUPRECHT RM (1989) Unusual DNA structures at the intregration site of an *HIV* provirus. Biochem Biophys Res Commun 161:134-142

GARDNER MB, KUZAK CC, O'BRIEN PN (1991) The lake Casitas wild mouse: evolving genetic resistance to retroviral disease. Trends Genet 7:22-27

GENTLES AJ, KARLIN S (2001) Genome-scale compositional comparisons in eukaryotes. Genome Res 11:540-546

GEYER PK, GREEN MM, CORCES VG (1988) Mutant gene phenotypes mediated by a *Drosophila melanogaster* retrotransposon require sequences homologous to mammalian enhancers. Proc Natl Acad Sci USA 85:8593-8597

GOFFEAU A, AERT R, AGOSTINI-CARBONE ML, *et al.* (633 co-authors) (1997) The yeast genome directory. Nature 387(suppl.):1-105

GOWER JC (1966) Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53:325-338

GRANDBASTIEN M-A, LUCAS H, MOREL J-B, MHIRI C, VERNHETTES S, CASACUBERTA JM (1997) The expression of the tobacco *Tnt1* retrotransposon is linked to plant defense responses. Genetica 100:241-252

GRANTHAM R, GAUTIER C, GOUY M, MERCIER R, PAVÉ A (1980) Codon catalog and the genome hypothesis. Nucleic Acids Res 8:r49-r62

GREIDER CW (1990) Telomeres, telomerase and senescence. Bioessays 12:363-369

HATTORI M, FUJIYAMA A, Taylor TD, et al. (63 co-authors) (2000) The DNA sequence of human chromosome 21. Nature 405:311-319

HILL WG, ROBERTSON A (1966) The effect of linkage on limits to artificial selection. Genet Res 8:269-294

HIRAIZUMI Y (1971) Spontaneous recombination in *Drosophila melanogaster* males. Proc Natl Acad Sci USA 68:268-270

HUGUES S, ZELUS D, MOUCHIROUD D (1999) Warm-blooded isochore structure in Nile crocodile and turtle. Mol Biol Evol 16:1521-1527

IKEMURA T (1981) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optinal for the *E. coli* translational system. J Mol Biol 151:389-409

IKEMURA T (1982) Correlation between the abundance of yeast transfer RNAs and the occurence of the respective codons in its protein genes: differences in synonymous codon choice patterns of yeast and *Escherichia coli* with reference to the abundance of isoaccepting transfer RNAs. JMol Biol 158:573-597

IKENAGA H, SAIGO K (1982) Insertion of a movable genetic element *297*, into the T-A-T-A box for the H3 histone gene in *Drosophila melanogaster*. Proc Natl Acad Sci USA 79:4143-4147

IMASHEVA AG, LOESCHCKE V, ZHIVOTOVSKY LA, LAZEBNY OE (1998) Stress temperatures and quantitative variation in *Drosophila melanogaster*. Heredity 81:246-253

INTERNATIONAL HUMAN GENOME SEQUENCING CONSORTIUM (2001). Initial sequencing and analysis of the human genome. Nature 409:860-921

INOUYE S, YUKI S, SAIGO K (1984) Sequence-specific insertion of the Drosophila transposable element *17.6.* Nature 310:332-333

JABBARI K, BERNARDI G (2000) The distribution of genes in the Drosophila genome. Gene 247:287-292

JAKUBCZAK JL, XIONG Y, EICKBUSH TH (1990) Type I (*R1*) and type II (*R2*) DNA insertions of *Drosophila melanogaster* are retrotransposable elements closely related to those of *Bombyx mori.* J Mol Biol

212:37-52

JORDAN IK, MCDONALD JF (1999) Comparative genomics and evolutionary dynamics of *Saccharomyces cerevisiae Ty* elements. Genetica 107:3-13

JURKA J (1997) Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. Proc Natl Acad Sci USA 94:1872-1877

KARLIN S, BURGE C (1995) Dinucleotide relative abundance extremes: a genomic signature. Trends Genet 11:283-290

KARLIN S, CAMPBELL AM, MRÁZEK J (1998) Comparative DNA analysis across diverse genomes. Annu Rev Genet 32:185-225

KARLIN S, DOERFLER W, CARDON LR (1994) Why is CpG suppressed in the genomes of virtually all small eukaryotic viruses but not in those of large eukaryotic viruses? J Virol 68:2889-2897

KARLIN S, LADUNGA I (1994) Comparisons of eukaryotic genomic sequences. Proc Natl Acad Sci USA 91:12832-12836

KARLIN S, LADUNGA I, BLAISDELL BE (1994) Heterogeneity o genomes: measures and values. Proc Natl Acad Sci USA 91:12837-12841

KARLIN S, MRÁZEK J (1997) Compositional differences within and between eukaryotic genomes. Proc Natl Acad Sci USA 94:10227-10232

KARLIN S (1998) Global dinucleotide signatures and analysis of genomic heterogeneity. Curr Op Micro 1:598-610

KIDWELL MG, KIDWELL JF, SVED JA (1977) Hybrid dysgenesis in *Drosophila melanogaster*: a syndrome of aberrant traits including mutation, sterility, and male recombination. Genetics 86:813-833

KIM A, TERZIAN C, SANTAMARIA P, PÉLISSON A, PRUD'HOMME N, BUCHETON A (1994) Retroviruses in invertebrates: the *gypsy* retrotransposon is apparently an infectious retrovirus of *Drosophila melanogaster*. Proc Natl Acad Sci USA 91:1285-1289

KIM JM, VANGURI S, BOEKE JD, GABRIEL A, VOYTAS DF (1998) Transposable elements and genome organization: a comprehensive survey of retrotransposon revealed by the complete *Saccharomyces cerevisiae* genome sequence. Genome Res 8:464-478

KIMURA M (1968) Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. Genet Res 11:247-269

KLIMAN RM, HEY J (1993) Reduced natural selection associated with low recombination in *Drosophila melanogaster*. Mol Biol Evol 10:1239-1258

KLIMAN RM, HEY J (1994) The effects of mutation and natural selection on codon bias in the genes of Drosophila. Genetics 137:1049-1056

KOSKI LB, Morton RA, GOLDING GB (2001) Codon bias and base composition are poor indicators of horizontally transferred genes. Mol Biol Evol 18:404-412

KYRPIDES N (1999) Genomes OnLine Database (GOLD): a monitor of complete and ongoing genome projects world wide. Bioinformatics 15:773-774

Ladeveze V, Aulard S, Chaminade N, Biémont C, Périquet G, Lemeunier F (2000) Dynamics of the *hobo* transposable element in transgenic lines of *Drosophila melanogaster*. Genet Res 77:135-142

Laprevotte I, Brouillet S, Terzian C, Hénaut A (1997) Retroviral oligonucleotide distributions correlate with biased nucleotide compositions of retrovirus sequences, suggesting a duplicative stepwise molecular evolution. J Mol Evol 44:214-225

Le QH, Wright S, Yu Z, Bureau T (2000) Transposon diversity in *Arabidopsis thaliana*. Proc Natl Acad Sci USA 97:7376-7381

Lebart L, Morineau A, Piron M (1997) Statistique exploratoire multidimensionnelle. Dunod, Paris.

Leblanc P, Dastugue B, Vaury C (1999) The integration machinery of *ZAM*, a retroelement from *Drosophila melanogaster*, acts as a sequence-specific endonuclease. J. Virol 73:7061-7064

Leblanc P, Desset S, Giorgi F, Taddei AR, Fausto AM, Mazzini M, Dastugue B, Vaury C (2000) Life cycle of an endogenous retrovirus, *ZAM*, in *Drosophila melanogaster*. J Virol 74:10658-10669

Lecointre G, Le Guyader H (2001) Classification phylogénétique du vivant. Ed. Belin.

Lefebvre J (1983) Introduction aux analyses statistiques multidimensionnelles. Masson, Paris.

Lerat E, Capy P (1999) Retrotransposons and retroviruses: analysis of the envelope gene. Mol Biol Evol 16:1198-1207

Lerat E, Brunet F, Bazin C, Capy P (1999) Is the evolution of transposable elements modular? Genetica 107:15-25

Li W-H (1987) Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. J Mol Evol 24:337-345

Liao G-C, Rehm EJ, Rubin GM (2000) Insertion site preferences of the *P* transposable element in *Drosophila melanogaster*. Proc Natl Acad Sci USA 97:3347-3351

Liebermann D, Hoffman-Liebermann B, Weinthal J, Childs G, Maxson R, Mauron A, Cohen SN, Kedes LH (1983) An unusual transposon with long terminal inverted repeats in the sea urchin *Strongylocentrotus purpuratus.* Nature 306:342-347

Lin X, Kaul S, Rounsley S, *et al.* (37 co-authors) (1999) Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. Nature 402:761-768

Lozovskaya ER, Hartl DL, Petrov DA (1995) Genomic regulation of transposable elements in Drosophila. Curr Opin Genet Dev 5:768-773

Luan DD, Korman MH, Jakubczak JL, Eickbush TH (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

Lyko F, Ramsahoye BH, Jaenisch R (2000) DNA methylation in *Drosophila melanogaster*. Nature 408:538-540

Lyttle TW, Haymer DS (1992) The role of the transposable element *hobo* in the origin of endemic inversions in wild populations of *Drosophila melanogaster*. Genetica 86:113-126

Mahillon J, Chandler M (1998) Insertion sequences. Microbiol Mol Biol Rev 62:725-774

Marais G, Mouchiroud D, Duret L (2001) Does recombination improve selection on codon usage? Lessons

from nematode and fly complete genomes. Proc Natl Acad Sci USA 98:5688-5692

MARIN I, PLATA-RENGIFO P, LABRADOR M, FONTDEVILLA A (1998) Evolutionary relationships among the members of an ancient class of non-LTR retrotransposon found in the nematode *Caenorhabditis elegans.* Mol Biol Evol 15:1390-1402

MARTIENSSEN R (1998) Transposons, DNA methylation and gene control. Trends Genet 14:263-264

MARTIN SL (1991) LINEs. Curr Opin Genet Dev 1:505-508

MAYER K, SCHÜLLER C, WAMBUTT R, *et al.* (230 co-authors) (1999) Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana.* Nature 402:769-777

MAYER J, SAUTER M, RACZ A, SCHERER D, MUELLER-LANTZSCH N, MEESE E (1999) An almost-intact endogenous retrovirus K on human chromosome 7. Nat Genet 21:257-258

MCCLINTOCK B (1950) The origin and behavior of mutable loci in maize. Proc Natl Acad Sci USA 36:344-355

MCCLINTOCK B (1951) Chromosomal organisation and genic expression. Cold Spring Harbor Symp Quant Biol 16:13-47

MCCLURE MA (1991) Evolution of retroposons by acquisition or deletion of retrovirus-like genes. Mol Biol Evol 8:835-856

MCDONALD JF (1995) Transposable elements: possible catalysts of organismic evolution. Trends Ecol Evol 10:123-126

MCDONALD JF, MATYUNINA LV, WILSON S, JORDAN IK, BOWEN NJ, MILLER WJ (1997) LTR retrotransposons and the evolution of eukaryotic enhancers. Genetica 100:3-13

MCLEAN C, BUCHETON A, FINNEGAN DJ (1993) The 5''untranslated region of the *I* factor, a long interspersed nuclear elements-like retrotransposon of *Drosophila melanogaster*, contains an internal promoter and sequences that regulate expression. Mol Cell Biol 13:1042-1050

MÉDIGUE C, ROUXEL T, VIGIER P, HÉNAUT A, DANCHIN A (1991) Evidence for horizontal gene transfer in *Escherichia coli* speciation. J Mol Biol 222:851-856

MEINKE DW, CHERRY JM, DEAN C, ROUNSLEY SD, KOORNNEEF M (1998) *Arabidopsis thaliana*: a model plant for genome analysis. Science 282:662-682

MERRIMAN PJ, GRIMES CD, AMBROZIAK J, HACKETT DA, SKINNER P, SIMMONS MJ (1995) *S* elements: a family of *Tc1*-like transposons in the genome of *Drosophila melanogaster*. Genetics 141:1425-1438

MEUTH M (1989) Illegitimate recombination in mammalian cells. In BERG DE, HOWE M., (eds) Mobile DNA. American Society for microbiology, Washington D. C., USA, pp. 833-860

MINCHIOTTI G, DINOCERA PP (1991) Convergent transcription initiates from oppositely orientated promoters within 5''end regions of *Drosophila melanogaster F* elements. Mol Cell Biol 11:5171-5180

MORIYAMA EN, POWELL JR (1998) Gene length and codon usage biais in *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Escherichia coli*. Nucleic Acids Res 26:3188-3193.

MORTON BR (1993) Chloroplast DNA codon use: evidence for selection at the *psb A* locus based on tRNA availability. J Mol Evol 37:273-280

MORTON BR (1994) Codon use and the rate of divergence of land plant chloroplast genes. Mol Biol Evol 11:231-238

MORGAN TH, STURTEVANT AH, MULLER HJ, BRIDGES C (1915) The mechanism of mendelian heredity. New-York.

MULLER HJ (1927) Artificial transmutation of the gene. Science 66:84-87

NIKITIN AG, WOODRUFF RC (1995) Somatic movement of the mariner transposable element and lifespan of Drosophila species. Mutat Res 338:43-49.

OKADA N (1991) SINEs. Curr Opin Genet Dev 1:498-504

OKADA N, HAMADA M, OGIWARA I, OHSHIMA K (1997) SINEs and LINEs share common 3' sequences: a review. Gene 205:229-243

O'NEILL RJ, O'NEILL MJ, GRAVES JA (1998) Undermethylation associated with retroelement activation and chromosome remodelling in an interspecific mammalian hybrid. Nature 393:68-72

ORGEL LE, CRICK FH (1980) Selfish DNA: the ultimate parasite. Nature 284:604-607

PARDUE ML (1995) Drosophila telomeres: another way to end it all. In GREIDER C, BLACKBURN EH (eds) Telomeres. Cold Spring Harbor Laboratory Press, pp. 339-370

PARDUE ML, DANILEVSKAYA ON, TRAVERSE KL, LOWENHAUPT K (1997) Evolutionary links between telomeres and transposable elements. Genetica 100:73-84

PERCUDANI R, PAVESI A, OTTONELLO S (1997) Transfer RNA gene redundancy and translational selection in *Saccharomyces cerevisiae.* J Mol Biol 268:322-330

POST LE, STRYCHARZ GD, NOMURA M, LEWIS H, DENNIS PP (1979) Nucleotide sequence of the ribosomal protein gene cluster adjacent to the gene for RNA polymerase subunit b in *Escherichia coli.* Proc Natl Acad Sci USA 76:1697-1701

POWELL JR, GLEASON JM (1996) Codon usage and the origin of *P* elements. Mol Biol Evol 13:278-279

PRUD'HOMME N, MASSON GM, TERZIAN C, BUCHETON A (1995) *Flamenco,* a gene controlling the *gypsy* retrovirus of *Drosophila melanogaster.* Genetics 139:697-711

PURUGGANAN MD (1993) Transposable elements as introns: evolutionary connections. Trends Evol Ecol 8:239-243

RAND DM (1992) RIPping and RAPping at Berkeley. Genetics 132:1223-1224

ROY AM, CARROLL ML, KASS DH, NGUYEN SV, SALEM A-H, BATZER MA, DEININGER PL (1999) Recently integrated human Alu repeats: finding needles in the haystack. Genetica 107:149-161

RUBIN GM, LEWIS EB (2000) A brief history of Drosophila's contributions to genome research. Science 287:2216-2220

SALANOUBAT M, LEMCKE K, RIEGER M, *et al.* (137 co-authors) (2000) Sequence and analysis of chromosome 3 of the plant *Arabidopsis thaliana.* Nature 408:820-822

SANMIGUEL P, TIKHONOV A, JIN YK, MOTCHOULSKAIA N, ZAKHAROV D, MELAKE-BERHAN A, SPRINGER PS, EDWARDS KJ, LEE M, AVRAMOVA Z, BENNETZEN JL (1996) Nested retrotransposons in the intergenic regions of the maize genome. Science 274:765-768

SCHMID CW (1998) Does SINEs evolution precule *Alus* function? Nucleic Acids Res 26:4541-4550

SCHNEIDER D, DUPERCHY E, COURSANGE E, LENSKI RE, BLOT M (2000) Long-Term Experimental evolution in *Escherichia coli.* IX. Characterization of insertion sequence-mediated mutations and rearrangements. Genetics 156:477-488

SEZUTSU H, NITASAKA E, YAMASAKI T (1995) Evolution of the LINE-like *I* element in the *Drosophila melanogaster* specices subgroup. Mol Gen Genet 249:168-178

SHAPIRO JA (1969) Mutations caused by the insertion of genetic material into the galactose operon of *Escherichia coli.* J Mol Biol 40:93-105

SHARP PM, LI W-H (1987) The codon adaptation index - a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15:1281-1295

SHARP PM, MATASSI G (1994) Codon usage and genome evolution. Curr Opin Genet Dev 4:851-860

SHEEN FM, LEVIS RW (1994) Transposition of the *LINE*-like retrotransposon *TART* to Drosophila chromosome termini. Proc Natl Acad Sci USA 89:7591-7595

SHIELDS DC, SHARP PM, HIGGINS DG, WRIGHT F (1988) "Silent" sites in Drosophila genes are not neutral: evidence of selection among synonymous codons. Mol Biol Evol 5:704-716

SHIELDS DC, SHARP PM (1989) Evidence that mutations patterns vary among Drosophila transposable elements. J Mol Biol 207:843-846

STENICO M, LLOYD AT, SHARP PM (1994) Codon usage in C*aenorhabditis elegans:* delineation of translational selection and mutational biases. Nucleic Acids Res 22:2437-2446

SUEOKA N (1962) On the genetic basis of variation and heterogeneity of DNA base composition. Genetics 48:582-592

SULSTON JE, SCHIERENBERG E, WHITE JG, THOMSON JN (1983) The embryonic cell lineage of the nematode *Caenorhabditis elegans*. Dev Biol 100:64-119

SURZYCKI SA, BELKNAP WR (2000) Repetitive-DNA elements are similarly distributed on *Caenorhabditis elegans* autosomes. Proc Natl Acad Sci USA 97:245-249

TABATA S, KANEKO T, NAKAMURA Y, *et al.* (129 co-authors) (2000) Sequence and analysis of chromosome 5 of the plant *Arabidopsis thaliana.* Nature 408:823-826

TAKAHASHI H, OKAZAKI S, FUJIWARA H (1997) A new family of site-specific retrotransposons *SART-1*, is inserted into telomeric repeats of the silkworm *Bombyx mori.* Nucleic Acids Res 25:1578-1584

TEMIN HM (1980) Origin of retroviruses from cellular moveable genetic elements. Cell 21:599-600

TERZIAN C, LAPREVOTTE I, BROUILLET S, HÉNAUT A (1997) Genomic signatures: tracing the origin of retroelements at the nucleotide level. Genetica 100:271-279

TERZIAN C, PÉLISSON A, BUCHETON A (2001) Evolution and phylogeny of insect endogenous retroviruses. BMC Evol Biol sous presse.

TEYSSET L, BURNS JC, SHIKE H, SULLIVAN BL, BUCHETON A, TERZIAN C (1998) A moloney murine leukemia virus-based retroviral vector pseudotyped by the insect retroviral *gypsy* envelope can infect Drosophila cells. J Virol 72:853-856

THE ARABIDOPSIS GENOME INITIATIVE (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796-815

THE *C. ELEGANS* SEQUENCING CONSORTIUM (1998) Genome sequence of the nematode *C. elegans*. A platform for investigating biology. Science 282:2012-2018

THEOLOGIS A, ECKER JR, PALM CJ, *et al.* (89 co-authors) (2000) Sequence analysis of chromosome 1 of the plant *Arabidopsis thaliana*. Nature 408:816-820

THIOULOUSE J, CHESSEL D, DOLÉDEC S, OLIVIER JM (1997) ADE-4: a multivariate analysis and graphical display software. Stat Comput 7:75-83

TRAVERS AA (1989) DNA conformation and protein binding. Annu Rev Biochem 58:427-452

TRISTEM M (2000) Identification and characterization of novel human endogenous retrovirus families by phylogenetic screening of the human genome mapping project database. J Virol 74:3715-3730

TRUETT MA, JONES RS, POTTER SS (1981) Unusual structure of the FB family of transposable elements in Drosophila. Cell 24:753-763

ULLU E, TSCHUDI C (1984) *Alu* sequences are processed 7SL RNA genes. Nature 312:171-172

VENTER JC, ADAMS MD, MEYERS EW, *et al*. (274 co-authors) (2001) The sequence of the human genome. Science 291:1304-1351

VIGGIANO L, CAGGESE C, BARSANTI P, CAIZZI R (1997) Cloning and characterization of a copy of *Tirant* transposable element in *Drosophila melanogaster*. Gene 197:29-35

VOISSET C, BOUTON O, BEDIN F, DURET L, MANDRAND B, MALLET F, PARANHOS-BACCALA G (2000) Chromosomal distribution and coding capacity of the human endogenous retrovirus HERV-W family. AIDS Res Hum Retroviruses 16:731-740

WALBOT V (2000) A green chapter in the book of life. Nature 408:794-795

WARMUS H, BROWN P (1989) Retroviruses. In BERG DE, HOWE M (eds) Mobile DNA. American Society for microbiology, Washington DC, USA, pp. 53-108

WHALEN JH, GRIGLIATTI TA (1998) Molecular characterization of a retrotransposon in *Drosophila melanogaster*, *nomad*, and its relationship to other retrovirus-like mobile elements. Mol Gen Genet 260:401-409

WICHMAN HA, VAN DEN BUSSCHE RA, HAMILTON MJ, BAKER RJ (1992) Transposable elements and the evolution of genome organization in mammals. Genetica 86:287-293

WOODRUFF RC, NIKITIN AG (1995) *P* DNA element movement in somatic cells reduces lifespan in *Drosophila melanogaster*: evidence in support of the somatic mutation theory of aging. Mutat Res 338:35-42

WRIGHT F (1990) The 'effective number of codons' used in a gene. Gene 87:23-29.

XIONG Y, EICKBUSH TH (1990) Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J 9:3353-3362

YODER JA, WALSH CP, BESTOR TH (1997) Cystosine methylation and the ecology of intragenomic parasites. Trends Genet 13:335-340

YOUNGMAN S, VAN LUENEN HGAM, PLASTERK RHA (1996) *RTe-1*, a retrotransposon-like element in

Caenorhabditis elegans. FEBS Letters 380:1-7

YUAN JY, FINNEY M, TSUNG N, HORVITZ HR (1991) *Tc4*, a *Caenorhabditis elegans* transposable element with an unusual fold-back structure. Proc Natl Acad Sci USA 88:3334-3338

ZOU S, VOYTAS DF (1997) Silent chromatin determines target preference of the retrotransposon *Ty*5. Proc Natl Acad Sci USA 94:7412-7416

ZSÌROS J, JEBBINK MF, LUKASHOV VV, VOÛTE PA, BERKHOUT B (1999) Biased nucleotide composition of the genome of *HERV-K* related endogenous retroviruses and its evolutionary implications. J Mol Evol 48:102-111

ANNEXES

Lexique

Apoptose : mécanisme prédéterminé constituant la mort cellulaire programmée.

Codons préférés : il s'agit des codons préférentiellement utilisés par un organisme donné et par un gène donné. Ce sont les codons mis en évidence par les AFC.

Codons optimaux : il s'agit des codons dont l'utilisation augmente avec l'expression des gènes.

Désamination : il s'agit de la perte du groupement amine d'un acide nucléique.

Dysgénésie : anomalie de développement d'un organe ou d'un tissu.

Energie d'empilement des bases : interactions de Van der Waals entre des paires de base empilée.

Enhancer : il s'agit d'un élément de contrôle ADN généralement placé en 5' du site de démarrage d'un gène qui, lorsque un facteur de transcription spécifique s'y fixe, augmente les niveaux d'expression du gène mais n'est pas suffisant seul pour permettre l'expression.

Hétérochromatine : il s'agit d'une région de la chromatine fortement compactée et qui présente des caractéristiques particulières de colorabilité en cytologie par rapport à l'**euchromatine**, région beaucoup moins compactée et qui contient les gènes.

Méthylation : il s'agit d'une processus qui permet l'addition de groupements méthyls sur certains nucléotides de l'ADN. Ceci affecte l'expression des gènes car l'ADN méthylé ne peut pas être facilement transcrits. Ce mécanisme a donc un rôle important dans le contrôle de l'expression des gènes.

Recombinaison ectopique : elle se produit entre deux régions différentes d'un génome dont les séquences sont semblables mais non homologues. Elle peut conduire à des anomalies chromosomiques.

Superhélicité de l'ADN : sur-enroulement de la double hélice de l'ADN qui correspond à des structures secondaires et tertiaires de la molécule.

Système V(D)J : l'expression des gènes codant pour les immunoglobulines et le récepteur des cellules T nécessite l'assemblage préalable de l'exon variable, à partir de segments géniques distincts, V, D et J. Ces événements de recombinaison site-spécifique sont régulés au cours du développement des cellules lymphocytaires.

Taille efficace : il s'agit du nombre d'individus participant au processus reproductif à un moment donné dans une population.

Tableau 3' : ETs et numéros d'accession dans Genbank

Numéro d'accession	Nom de l'élément	Classe
Drosophila virilis		
U26847	Helena	rétrotransposon sans LTR
U49102	Penelope	rétrotransposon sans LTR
AF056940	Tv1	rétrotransposon LTR
M38438	gypsy	rétrotransposon LTR
U71250	blastopia	rétrotransposon LTR
X56645	Ulysses	rétrotransposon LTR
AF009439	Telemac	rétrotransposon LTR
U26938	Paris	transposon
Drosophila simulans		
AF012033	Helena	rétrotransposon sans LTR
U13028	R1sim4	rétrotransposon sans LTR
U13033	R2sim1	rétrotransposon sans LTR
U63747	Het-A	rétrotransposon sans LTR
U64957	R2	rétrotransposon sans LTR
S80353	I	rétrotransposon sans LTR
D10880	copia	rétrotransposon LTR
D83207	ninja	rétrotransposon LTR
X04846	297C	rétrotransposon LTR
X04847	297D	rétrotransposon LTR
X04848	297E	rétrotransposon LTR
AF037052	mariner-Mos6a	transposon
AF037054	mariner-BordA	transposon
Drosophila subobscura		
X72390	gypsy	rétrotransposon LTR
U73800	bilbo	rétrotransposon LTR
Drosophila melanogaster		
X68130	Het-A	rétrotransposon sans LTR
X51968	R1	rétrotransposon sans LTR
X51967	R2	rétrotransposon sans LTR
U14101	TART	rétrotransposon sans LTR
AF01030	Helena	rétrotransposon sans LTR
X17551	Doc	rétrotransposon sans LTR
M17214	F	rétrotransposon sans LTR
M14954	I	rétrotransposon sans LTR
M38643	jockey	rétrotransposon sans LTR
X77571	BS	rétrotransposon sans LTR
Z27119	blastopia	rétrotransposon LTR

Numéro d'accession	Nom de l'élément	Classe
X04456	copia	rétrotransposon LTR
X14037	micropia	rétrotransposon LTR
X01472	17.6	rétrotransposon LTR
X95908	mdg3	rétrotransposon LTR
X04132	412	rétrotransposon LTR
X07656	1731	rétrotransposon LTR
U23420	BEL	rétrotransposon LTR
U89994	burdock	rétrotransposon LTR
AJ000387	ZAM	rétrotransposon LTR
M12927	gypsy	rétrotransposon LTR
X03431	297	rétrotransposon LTR
X93507	Tirant	rétrotransposon LTR
AF039416	nomad	rétrotransposon LTR
X04705	hobo	transposon
X67681	BARI-1	transposon
X01748	HB1	transposon
U33463	S	transposon
X59837	pogo	transposon
X15469	foldback	classe III

Tableau 4': noms et numéros d'accession dans GenBank des éléments transposables des 5 espèces.

Numéros d'accession	Noms	classe
A. thaliana		
X91210	RT9b	rétrotransposon LTR
X91204	RT4	rétrotransposon LTR
X91203	RT3	rétrotransposon LTR
X91202	RT2	rétrotransposon LTR
X91201	RT1	rétrotransposon LTR
AJ002625	romani-1	rétrotransposon LTR
AF073829	romani-5	rétrotransposon LTR
AF073828	romani-3	rétrotransposon LTR
AF073827	romani-2	rétrotransposon LTR
X53976	Ta1-2-Landsberg	rétrotransposon LTR
X53975	Ta1-2-Kashmir	rétrotransposon LTR
X53973	Ta1-1	rétrotransposon LTR
X13291	Ta1-3	rétrotransposon LTR
X81801	Athila	rétrotransposon LTR
AF056633	Tat1-3	rétrotransposon LTR
AF056632	Tat1-2	rétrotransposon LTR
AF056631	Tat1-1	rétrotransposon LTR
AF030632	del-At	rétrotransposon LTR
AB021267	AtRe2-1-67	rétrotransposon LTR
AB021266	AtRE2-1-66	rétrotransposon LTR
AB021265	AtRE1-65	rétrotransposon LTR
AB021264	AtRE2-64	rétrotransposon LTR
AB021263	AtRE1-63	rétrotransposon LTR
AB014748	AtRE1	rétrotransposon LTR
AB016131	ATLN4	rétrotransposon sans LTR
AB016130	ATLN3	rétrotransposon sans LTR
AB016128	ATLN1	rétrotransposon sans LTR
L47194	Ta28	rétrotransposon sans LTR
L47192	Ta27	rétrotransposon sans LTR
L47195	Ta26	rétrotransposon sans LTR
L47183	Ta25	rétrotransposon sans LTR
L47187	Ta24	rétrotransposon sans LTR
L47181	Ta23	rétrotransposon sans LTR
L47184	Ta22	rétrotransposon sans LTR
L47188	Ta21	rétrotransposon sans LTR
L47185	Ta20	rétrotransposon sans LTR
L47290	Ta19	rétrotransposon sans LTR
L47190	Ta18	rétrotransposon sans LTR
L47186	Ta15	rétrotransposon sans LTR
L47191	Ta14	rétrotransposon sans LTR
L47189	Ta13	rétrotransposon sans LTR
L47182	Ta12	rétrotransposon sans LTR
L47193	Ta11-1	rétrotransposon sans LTR

Numéros d'accession	Noms	classe
AF120335	Tag2	transposon
AF051562	Tag1	transposon
C. elegans		
L00665	Tc4v	classe III
U15406	Cer1	rétrotransposon LTR
Z48009	frodo2	rétrotransposon sans LTR
Z70755	frodo1	rétrotransposon sans LTR
Z81064	sam9	rétrotransposon sans LTR
AF016663	sam8	rétrotransposon sans LTR
Z82090	sam7	rétrotransposon sans LTR
Z82275	sam6	rétrotransposon sans LTR
Z81092	sam5	rétrotransposon sans LTR
Z92978	sam4	rétrotransposon sans LTR
U46668	sam3	rétrotransposon sans LTR
U57054	sam2	rétrotransposon sans LTR
U13643	sam1	rétrotransposon sans LTR
AF054983	RTE-1	rétrotransposon sans LTR
L13200	T1-2	transposon
Z35400	Tc5	transposon
M22301	Tc1-Bristol	transposon
X01005	Tc1	transposon
X59156	Tc2	transposon
D. melanogaster		
X15469	Foldback	classe III
Z27119	Blastopia	rétrotransposon LTR
X04456	Copia	rétrotransposon LTR
X14037	Micropia	rétrotransposon LTR
X01472	17.6	rétrotransposon LTR
X95908	Mdg3	rétrotransposon LTR
X04132	412	rétrotransposon LTR
X07656	1731	rétrotransposon LTR
U23420	BEL	rétrotransposon LTR
U89994	Burdock	rétrotransposon LTR
AJ000387	ZAM	rétrotransposon LTR
M12927	Gypsy	rétrotransposon LTR
X03431	297	rétrotransposon LTR
X93507	Tirant	rétrotransposon LTR
AF039416	Nomad	rétrotransposon LTR
X70361	Aurora	rétrotransposon LTR
AJ009736	Idefix	rétrotransposon LTR
AJ010298	gate	rétrotransposon LTR
AF22049	Transpac	rétrotransposon LTR
X68130	HeT-A	rétrotransposon sans LTR
X51968	R1	rétrotransposon sans LTR
X51967	R2	rétrotransposon sans LTR

Numéros d'accession	Noms	classe
X05643	TART	rétrotransposon sans LTR
AF012030	Helena	rétrotransposon sans LTR
X17551	Doc	rétrotransposon sans LTR
M17214	F	rétrotransposon sans LTR
M14954	I	rétrotransposon sans LTR
M38643	jockey	rétrotransposon sans LTR
X77571	BS	transposon
X67681	Bari-1	transposon
X04705	Hobo	transposon
X01748	HB1	transposon
U33463	S	transposon
X06779	P	transposon
X59837	pogo	transposon
H. sapiens		
AF148856	L1	rétrotransposon sans LTR
X52235	LINE-1	rétrotransposon sans LTR
X07857	L1-RT	rétrotransposon sans LTR
X84285	MLE-2173	transposon
X84286	MLE-2177	transposon
X84287	MLE-2178	transposon
X84288	MLE-2179	transposon
X84289	MLE-2217	transposon
U38613	Humar1pcr	transposon
U38615	Humar1g1	transposon
U49973	Tigger1	transposon
U49974	Mariner2	transposon
U52077	Mariner1	transposon
U80776	Hsmar1	transposon
Y17156	Tramp	transposon
S. cerevisiae		
M18706	Ty1-H3	rétrotransposon LTR
Z48502	Ty1-9532	rétrotransposon LTR
Z47816	Ty1-9827	rétrotransposon LTR
L22015	Ty1-inactif	rétrotransposon LTR
Z68194	Ty1-YD8142A	rétrotransposon LTR
Z68195	Ty1-YD8142B	rétrotransposon LTR
X59720	Ty2-chr3	rétrotransposon LTR
X79489	Ty2-chr2	rétrotransposon LTR
X95720	Ty2-chr15	rétrotransposon LTR
M34549	Ty3-1	rétrotransposon LTR
M23367	Ty3-2	rétrotransposon LTR
Z46728	Ty3-2-9910	rétrotransposon LTR
M94164	Ty4	rétrotransposon LTR
Z49389	Ty4-chr10	rétrotransposon LTR

Numéros d'accession	Noms	classe
X59720	Ty5	rétrotransposon LTR

Tableau 5' : numéros d'accession des ETs complets dans Genbank

Numéro d'accession	Nom de l'élément	Classe
Arabidopsis thaliana		
AF069298	CACTA-like 3	CACTA
AF058825	CACTA-like 2	CACTA
AC002341	CACTA-like 1	CACTA
AC004705	MITEV	MITE
AC003058	MITEXI	MITE
AF077408	gypsy-like XV	rétrotransposon à LTR
AC004482	gypsy-like VIII	rétrotransposon à LTR
AB005248	gypsy-like III	rétrotransposon à LTR
AB005247	gypsy-like I	rétrotransposon à LTR
X13291	Ta1-3	rétrotransposon à LTR
X53976	Ta1-2-2	rétrotransposon à LTR
X53975	Ta1-2	rétrotransposon à LTR
X81801	Athila	rétrotransposon à LTR
AF056633	Tat1-3	rétrotransposon à LTR
AF056632	Tat1-2	rétrotransposon à LTR
AF056631	Tat1-1	rétrotransposon à LTR
AB021267	AtRE2-2	rétrotransposon à LTR
AB021266	AtRE2-1	rétrotransposon à LTR
AB021264	AtRE2	rétrotransposon à LTR
AB021265	AtRE1-2	rétrotransposon à LTR
AB021263	AtRE1	rétrotransposon à LTR
L47193	Ta11-1	rétrotransposon sans LTR
AF077409	LINE-like XI	rétrotransposon sans LTR
AF069298	LINE-like XVIII	rétrotransposon sans LTR
Z97336	LINE-like XXIII	rétrotransposon sans LTR
AC005315	LINE-like V	rétrotransposon sans LTR
U65470	TSCL	rétrotransposon sans LTR
AF120335	Tag2	transposon
L12220	Tag1	transposon
U76697	Limpet1	transposon
AB013393	Ac-likeI	transposon
Caenorhabditis elegans		
X55356	Tc6.1	classe III
M60788	Tc4	classe III
L00665	Tc4v	classe III
U15406	Cer1	rétrotransposon à LTR
Z70755	frodo1	rétrotransposon sans LTR
Z68135	frodo2.2	rétrotransposon sans LTR
AF054983	RTE-1	rétrotransposon sans LTR
U13643	sam1	rétrotransposon sans LTR
U57054	sam2	rétrotransposon sans LTR
U46668	sam3	rétrotransposon sans LTR

Numéro d'accession	Nom de l'élément	Classe
Z92972	sam4	rétrotransposon sans LTR
Z81092	sam5.1	rétrotransposon sans LTR
Z82090	sam7.1	rétrotransposon sans LTR
AF016663	sam8	rétrotransposon sans LTR
Z81064	sam9	rétrotransposon sans LTR
U86951	TR-5	transposon
Z35400	Tc5	transposon
X59156	Tc2	transposon
X01005	Tc1	transposon
Drosophila melanogaster		
AJ009736	Idefix	rétrotransposon à LTR
D17529	springer	rétrotransposon à LTR
U11691	roo	rétrotransposon à LTR
J01078	HMS-beagle	rétrotransposon à LTR
X04671	blood	rétrotransposon à LTR
M12927	gypsy	rétrotransposon à LTR
AJ000387	ZAM	rétrotransposon à LTR
U89994	burdock	rétrotransposon à LTR
U23420	BEL	rétrotransposon à LTR
X07656	1731	rétrotransposon à LTR
X93507	Tirant	rétrotransposon à LTR
X04132	412	rétrotransposon à LTR
X95908	mdg3	rétrotransposon à LTR
X03431	297	rétrotransposon à LTR
X01472	17.6	rétrotransposon à LTR
AJ132547	circe	rétrotransposon à LTR
AJ010298	Gate	rétrotransposon à LTR
X14037	micropia	rétrotransposon à LTR
X04456	copia	rétrotransposon à LTR
Z27119	blastopia	rétrotransposon à LTR
AF222049	Transpac	rétrotransposon à LTR
AF039416	nomad	rétrotransposon à LTR
AL035311	roo	rétrotransposon à LTR
X77571	BS	rétrotransposon sans LTR
X06950	G	rétrotransposon sans LTR
M17214	F	rétrotransposon sans LTR
M38643	jockey	rétrotransposon sans LTR
M14954	I	rétrotransposon sans LTR
X17551	Doc	rétrotransposon sans LTR
U14101	TART-B1	rétrotransposon sans LTR
X51967	R2	rétrotransposon sans LTR
X51968	R1	rétrotransposon sans LTR
U06920	Het-A	rétrotransposon sans LTR
AF012030	Helena	rétrotransposon sans LTR
X06779	P	transposon

Numéro d'accession	Nom de l'élément	Classe
X80025	Hopper	transposon
M55078	Hoppel	transposon
X04705	hobo	transposon
U33463	S	transposon
X01748	HB1	transposon
X59837	pogo	transposon
X67681	Bari-1	transposon
Homo sapiens		
M80343	L1.2	rétrotransposon sans LTR
M80340	L1.1	rétrotransposon sans LTR
U93573	L1.33	rétrotransposon sans LTR
U93571	L1.24	rétrotransposon sans LTR
U93562	L1.5	rétrotransposon sans LTR
X52235	LINE-1	rétrotransposon sans LTR
U21247	HSRV	rétrovirus
AF074965	HL2V	rétrovirus
U19949	HTLV1	rétrovirus
AF082339	HIV2	rétrovirus
AF004394	HIV1	rétrovirus
M74509	v-onc	rétrovirus
X80240	HERV-KC4	rétrovirus endogène
AF164609	HERV-K101	rétrovirus endogène
AF074086	HERVK	rétrovirus endogène
AF020092	HERV-K-T47D	rétrovirus endogène
Y17156	Tramp	transposon
U49973	Tigger1	transposon
U52077	mariner1	transposon
U49974	mariner2	transposon
U38615	humar1g1	transposon
U38613	humar1pcr	transposon
AJ009227	Hsmar1	transposon
AF205600	Buster1	transposon
AF205598	Buster3	transposon
Saccharomyces cerevisiae		
chromosome I	Ty1	rétrotransposon à LTR
chromosome II	Ty2-1	rétrotransposon à LTR
	Ty1-1	rétrotransposon à LTR
	Ty1-2	rétrotransposon à LTR
chromosome III	Ty5-1	rétrotransposon à LTR
	Ty2-1	rétrotransposon à LTR
chromosome IV	Ty2-1	rétrotransposon à LTR
	Ty1-1	rétrotransposon à LTR
	Ty2-2	rétrotransposon à LTR
	Ty1-2	rétrotransposon à LTR

Numéro d'accession	Nom de l'élément	Classe
	Ty2-3	rétrotransposon à LTR
	Ty1-3	rétrotransposon à LTR
	Ty1-4	rétrotransposon à LTR
	Ty1-5	rétrotransposon à LTR
chromosome V	Ty1-1	rétrotransposon à LTR
chromosome VI	Ty2-1	rétrotransposon à LTR
chromosome VII	Ty1-1	rétrotransposon à LTR
	Ty1-2	rétrotransposon à LTR
	Ty2-1	rétrotransposon à LTR
	Ty3-1	rétrotransposon à LTR
	Ty2-2	rétrotransposon à LTR
	Ty1-3	rétrotransposon à LTR
chromosome VIII	Ty4-1	rétrotransposon à LTR
	Ty1-1	rétrotransposon à LTR
chromosome IX	Ty3-1	rétrotransposon à LTR
chromosome X	Ty4-1	rétrotransposon à LTR
	Ty1-1	rétrotransposon à LTR
	Ty1-2	rétrotransposon à LTR
chromosome XII	Ty1-1	rétrotransposon à LTR
	Ty1-2	rétrotransposon à LTR
	Ty1-3	rétrotransposon à LTR
	Ty1-4	rétrotransposon à LTR
	Ty2-1	rétrotransposon à LTR
	Ty2-2	rétrotransposon à LTR
chromosome XIII	Ty1-1	rétrotransposon à LTR
	Ty1-2	rétrotransposon à LTR
	Ty1-3	rétrotransposon à LTR
	Ty1-4	rétrotransposon à LTR
chromosome XIV	Ty1-1	rétrotransposon à LTR
	Ty1-2	rétrotransposon à LTR
	Ty2-1	rétrotransposon à LTR
chromosome XV	Ty1-1	rétrotransposon à LTR
	Ty1-2	rétrotransposon à LTR
	Ty2-1	rétrotransposon à LTR
	Ty2-2	rétrotransposon à LTR
chromosome XVI	Ty1-1	rétrotransposon à LTR
	Ty4-1	rétrotransposon à LTR
	Ty1-2	rétrotransposon à LTR
	Ty1-3	rétrotransposon à LTR
	Ty1-4	rétrotransposon à LTR

Article 1 : Retrotransposons and Retroviruses: analysis of the envelope gene

Mol. Biol. Evol. 16(9):1198-1207, 1999

Retrotransposons and Retroviruses: Analysis of the Envelope Gene

Emmanuelle Lerat and Pierre Capy

Laboratoire Populations, Génétique et Evolution, Centre National de la Recherche Scientifique, Gif-sur-Yvette, France

Retroviruses and long terminal repeat (LTR) retrotransposons share a common structural organization. The main difference between these retroelements is the presence of a functional envelope (*env*) gene in retroviruses, which is absent or nonfunctional in LTR retrotransposons. Several similarities between these two groups of retroelements have been detected for the reverse transcriptase, *gag,* and integrase domains. Assuming that each of these domains shares a common ancestral sequence, several hypotheses could account for the emergence of retroviruses from LTR retrotransposons. In this context, the positions of elements such as *gypsy* and the members of the *Ty3* subfamily are not clear, since they are classified as retroviruses but phylogenetically they are assigned to the LTR retrotransposon group. We compared the *env* gene products of these retroelements and identified two similar motifs in retroviruses and LTR retrotransposons. These two regions do not occur in the same order. If we assume that they are derived from the same ancestral sequence, this could result from independent acquisition of the various domains rather than the single acquisition of the whole *env* gene. However, we cannot exclude the possibility that the *env* gene was reorganized after being acquired. Trees based on these regions show that these two groups of elements are clearly distinguished. These trees are similar to those obtained from reverse transcriptase or integrase. In trees based on reverse transcriptase, the retroviruses with complete or partial *env* genes can be distinguished from the other LTR retrotransposons.

Introduction

Transposable elements (TEs) are divided into two main classes, depending on whether their transposition mechanisms use RNA or DNA intermediates (Finnegan 1989), and into several subclasses based on the presence/absence or order of domains (Capy et al. 1997*b*). In spite of clear structural differences, some similarities between retroviruses and retrotransposons can be detected at the protein level for domains in *gag* and *pol* genes like reverse transcriptase, Rnase H, protease, and integrase (see Capy et al. 1997*a* and references therein). If these similarities are not due to convergence, they must reflect common ancestral sequences. Assuming this, they can be used to infer the evolutionary history of different domains (Xiong and Eickbush 1990; McClure 1991; Capy et al. 1996, 1997*a*).

One of the main questions raised by the evolution of TEs is that of their relationships to retroviruses. The possibility of such a connection was initially discussed by Temin (1980) and Flavell (1981). Long terminal repeat (LTR) retrotransposons, and, more specifically, those of the *Ty3* superfamily, share some common structures with retroviruses, such as LTR, *gag,* and *pol* genes and, in some cases, incomplete and nonfunctional *env* genes. Several evolutionary scenarios have already been suggested on the basis of these similarities. For instance, Xiong and Eickbush (1990) and McClure (1991) have studied reverse transcriptase similarities in several retroelements, including retrotransposons and retroviruses. These authors suggest that retroviruses may evolve from LTR retrotransposons by acquiring a functional *env* gene. Capy et al. (1996) have also suggested this on the basis of comparison of DDE integrase/transposase.

Key words: transposable elements, retrotransposons, retroviruses, envelope gene, evolution.

Address for correspondence and reprints: Pierre Capy, Laboratoire Populations, Génétique et Evolution, Centre National de la Recherche Scientifique, 91198 Gif-sur-Yvette Cedex, France. E-mail: capy@pge.cnrs-gif.fr.

Mol. Biol. Evol. 16(9):1198–1207. 1999
 ISSN: 0737-4038

Although the transition from LTR retrotransposons to retroviruses seems to predominate, some oscillation between them cannot be excluded. A retrovirus, which loses the activity of its envelope gene (*env*), becomes an LTR retrotransposon. A clear definition of both is required before we can analyze conversions between LTR retrotransposons and retroviruses. In this paper, we will consider two distinct retroelements: the retroviruses and the errantiviruses. Following the nomenclature proposed by Boeke et al. (1998), the errantiviruses are LTR retrotransposons with *env*-like genes. In this context, the *gypsy* element, first reported as an LTR-retrotransposon (Bayev et al. 1984), was then considered an insect retrovirus (Kim et al. 1994; Song et al. 1994). Phylogenies based on reverse transcriptase or integrase clearly show that this element is always classified as an LTR retrotransposon (Xiong and Eickbush 1990; Capy et al. 1996, 1997*a*). Therefore, in the present work, this element will be considered as an errantivirus.

Several domains of these two entities have already been compared, including reverse transcriptase, integrase, Rnase H, and the *gag* gene, but no comparisons of their *env* genes had been done. This was probably because of the rapid evolution and the high variability of this gene. In this study, we attempted to compare the *env*-like gene of LTR retrotransposons (errantiviruses), with the *env* gene of retroviruses.

The retroviral *env* polypeptide is composed by subunit: a transmembrane protein (TM) and a larger peptide (SU) containing the receptor-binding function and antigenic sites able to elicit neutralizing antibodies in the infected host. In the human immunodeficiency virus (HIV), the polyprotein gp160 shares particular characteristic of *env* proteins: a signal peptide at the N-terminal end, a proteolytic cleavage site leading to glycoproteins gp41 (TM) and gp120 (SU), a transmembrane domain near the C-terminal end, a leucine zipper motif near the N- terminal end of gp41 corresponding to a fusion peptide involved in membrane fusion, and several glycosylated sites. The glycoprotein gp120 contains con-

Table 1
Accession Numbers of the Different Elements Used in the Comparison of the *env* Proteins

Element	Host	Type	Accession No.
tirant	*Drosophila melanogaster*	LTR retrotransposon	X93507[a]
tom	*Drosophila ananassae*	LTR retrotransposon	Z24451[a]
TED	*Trichoplusia ni*	LTR retrotransposon	C36329[b]
297	*D. melanogaster*	LTR retrotransposon	C24872[b]
17.6	*D. melanogaster*	LTR retrotransposon	P04283[c]
gypsy-sub	*Drosophila subobscura*	LTR retrotransposon	X72390[a]
gypsy-vir	*Drosophila virilis*	LTR retrotransposon	M38438[a]
gypsy-mel	*D. melanogaster*	LTR retrotransposon	S52567[b]
yoyo	*Ceratitis capitata*	LTR retrotransposon	U60529[a]
B104	*D. melanogaster*	LTR retrotransposon	Z48503[a]
ZAM	*D. melanogaster*	LTR retrotransposon	AJ000387[a]
FIV	Feline immunodeficiency virus	Retrovirus	U56928[a]
Visna	Ovine/caprine lentivirus	Retrovirus	M60609[a]
PLV-14	Puma lentivirus	Retrovirus	U03982[a]
HIV1	Human immunodeficiency virus type 1	Retrovirus	AF063223[a]
HIV2	Human immunodeficiency virus type 2	Retrovirus	M15390[a]
SIVcpz	Simian immunodeficiency virus (chimpanzee)	Retrovirus	U84097[a]
SIVmnd	Simian immunodeficiency virus (mandrill)	Retrovirus	P22380[c]
SIVsm	Simian immunodeficiency virus (mangabey)	Retrovirus	M31325[a]
AGMgri	African green monkey (SIV–mangabey)	Retrovirus	Q02837[b]
AGMsab	African green monkey (SIV–mangabey)	Retrovirus	S46352[b]
AGMtan	African green monkey (SIV–mangabey)	Retrovirus	U58991[a]

[a] EMBL/GenBank.
[b] NBRF.
[c] SwissProt.

served cysteine residues involved in gp41–gp120 interaction (Varmus and Brown 1989; Coffin 1990). The evolution rate of the *env* genes can be very high to escape neutralizing antibody recognition. However, the *env* genes can be represented as a succession of variable domains flanked by conserved regions. In silico hybridization with low stringency parameters was used to detect these conserved regions. Two motifs were found and used to infer trees.

Materials and Methods

All the sequences used in this work were extracted from EMBL, GenBank, NBRF, and SwissProt, with SRS 5.0 on the INFOBIOGEN site. Their names and accession numbers are given in tables 1– 3. Similar motifs (adjacent amino acids) or signatures (nonadjacent amino acids) were detected using BLASTP 2.0, which can cope with gaps (Altschul et al. 1997). The parameters were chosen to obtain as many sequences as possible, even with a low percentage of similarity. For instance, the following settings were used: EXPECT = 1,000 instead of 10, DESCRIPTION = 500 instead of 100, and ALIGNMENT = 500 instead of 100. The alignments between all the sequences were then performed manually using the SeqPup editor, version 0.8c (Gilbert 1998). Using the MEME program (Bailey and Elkan 1994) with all the sequences used in the alignments, the only motifs found between retroviruses and errantiviruses correspond to the two motifs first detected with BLASTP.

To estimate the validity of the alignment, a program in C (available upon request) was used. This program compares the distance distributions between the manually aligned sequence set and a randomly aligned sequence set. Distances were computed using the amino acid classification proposed by Hall (see the PHYLIP package, version 3.5c; Felsenstein 1993). Distributions were then compared using a Kolmogorov and Smirnov test (Sokal and Rolf 1995, p. 887).

In order to assess the amount of the phylogenetic signal in our data, a PTP test (Swofford et al. 1996) was performed. The tests were highly significant ($P < 0.001$ with 1,000 replicates) for all the data sets used. This includes the alignments based upon the *env,* integrase, and reverse transcriptase genes.

Trees were inferred using PAUP, version 3.1.1 (Swofford 1993), which uses the parsimony method; Puzzle, version 3.1 (Strimmer and von Haeseler 1996), which carries out an analysis using the maximum-likelihood method based on the "quartet puzzling" algorithm; and the neighbor-joining method in the PHYLIP package. Three models of amino acid substitution were used by PHYLIP: the Dayhoff, Schwartz, and Orcutt (1978) PAM matrix, based on the probability of the switch from one amino acid to another, and the George-Hunt-Barker matrix (George, Hunt, and Barker 1988) and the Hall matrix, which are both based on amino acid classification. Since all methods give similar topologies, only the results obtained with PAUP are reported. The heuristic search was used with stepwise addition (random addition and 10 replicates) and tree bisection-reconnection branch-swapping options.

Results

The Motifs

Two conserved motifs were determined among the LTR retrotransposons. The KRG motif was initially detected in *Tirant* (*Drosophila melanogaster*), *tom* (*Dro-*

Table 2
Accession Numbers of the Elements Used in the Comparison of Reverse Transcriptase Sequences

Element	Host	Type	Accession No.
Ty1	*Vicia melanops*	LTR retrotransposon	ID 176457[a]
1731	*Drosophila melanogaster*	LTR retrotransposon	X07656[b]
Tpm1	*Ptyas mucosus*	LTR retrotransposon	X74337[b]
Tch1	*Clupea harengus*	LTR retrotransposon	S22461[c]
Tnt1	*Nicotiana tabacum*	LTR retrotransposon	P10978[d]
SIRE-1	*Glycine max*	LTR retrotransposon	AF053008[b]
Tap1	*Pyxicephalus adspersus*	LTR retrotransposon	X74338[b]
Tco1	*Conolophus subcristatus*	LTR retrotransposon	X74336[b]
gypsy	*D. melanogaster*	LTR retrotransposon	X03734[b]
TED	*Trichoplusia ni*	LTR retrotransposon	B36329[c]
yoyo	*Ceratitis capitata*	LTR retrotransposon	U60529[b]
tom	*Drosophila ananassae*	LTR retrotransposon	S34639[c]
297	*D. melanogaster*	LTR retrotransposon	P20825[d]
ZAM	*D. melanogaster*	LTR retrotransposon	AJ000387[b]
17.6	*D. melanogaster*	LTR retrotransposon	P04323[d]
MoMuLV	AKV murine leukemia virus	Retrovirus	P03356[d]
HERV	Human endogenous virus C type	Retrovirus	M74509[b]
RSV	Rous sarcoma virus	Retrovirus	P03354[d]
HFV	Human foamy virus	Retrovirus	Y07725[b]
HIV1	Human immunodeficiency virus type 1	Retrovirus	D86069[b]
HIV2	Human immunodeficiency virus type 2	Retrovirus	M15390[b]
AGMgri	African green monkey (SIV–managabey)	Retrovirus	M66437[b]
SIVcpz	Simian immunodeficiency virus (chimpanzee)	Retrovirus	X52154[b]
PLV-14	Puma lentivirus	Retrovirus	U03982[b]
FIV	Feline immunodeficiency virus	Retrovirus	U56928[b]

[a] GenBank journal scan.
[b] EMBL/GenBank.
[c] NBRF.
[d] SwissProt.

sophila ananassae), and *TED* (*Trichoplusia ni*). Then, 25 amino acids were aligned around this motif and used in a BLASTP search. This allowed us to detect several errantiviruses and retroviruses, including *HIV* and *SIV* (fig. 1*A*).

Using the complete sequence of the *env* gene product of *17.6* (*D. melanogaster*) in a BLASTP search, a second motif, LTPL, was defined. This motif was found in both LTR retrotransposons and retroviruses. The alignment for 46 amino acids found for seven retrotransposons is shown figure 1*B*. Each sequence of this alignment was then used in a BLASTP search. Using *yoyo* (*Ceratitis capitata*), a puma lentivirus (*PLV-14*) *env* protein was detected. Four new sequences were obtained, including *FIV, SIV-cpz* (chimpanzee), and *HIV1,* using the complete sequence of this protein.

Positions of the Motifs

In this paper, the *env* gene of *HIV* is used as a reference for the positions of the various motifs. In *HIV,* this gene encodes for a polyprotein, which, after proteolysis, yields two proteins (fig. 2): gp120 (a surface glycoprotein) and gp41 (a transmembrane glycoprotein). The two proteins are involved in the adsorption and pen-

Table 3
Accession Numbers of the Elements Used in the Comparison of Integrase Sequences

Element	Host	Type	Accession No.
gypsy	*Drosophila melanogaster*	LTR retrotransposon	X03734[a]
TED	*Trichoplusia ni*	LTR retrotransposon	B36329[b]
yoyo	*Ceratitis capitata*	LTR retrotransposon	U60529[a]
tom	*Drosophila ananassae*	LTR retrotransposon	S34639[b]
297	*D. melanogaster*	LTR retrotransposon	P20825[c]
ZAM	*D. melanogaster*	LTR retrotransposon	AJ000387[a]
17.6	*D. melanogaster*	LTR retrotransposon	P04323[c]
HIV1	Human immunodeficiency virus type 1	Retrovirus	D86069[a]
HIV2	Human immunodeficiency virus type 2	Retrovirus	M15390[a]
AGMgri	African green monkey (SIV–mangabey)	Retrovirus	M66437[a]
SIVcpz	Simian immunodeficiency virus (chimpanzee)	Retrovirus	X52154[a]
PLV-14	Puma lentivirus	Retrovirus	U03982[a]
FIV	Feline immunodeficiency virus	Retrovirus	U56928[a]

[a] EMBL/GenBank.
[b] NBRF.
[c] SwissProt.

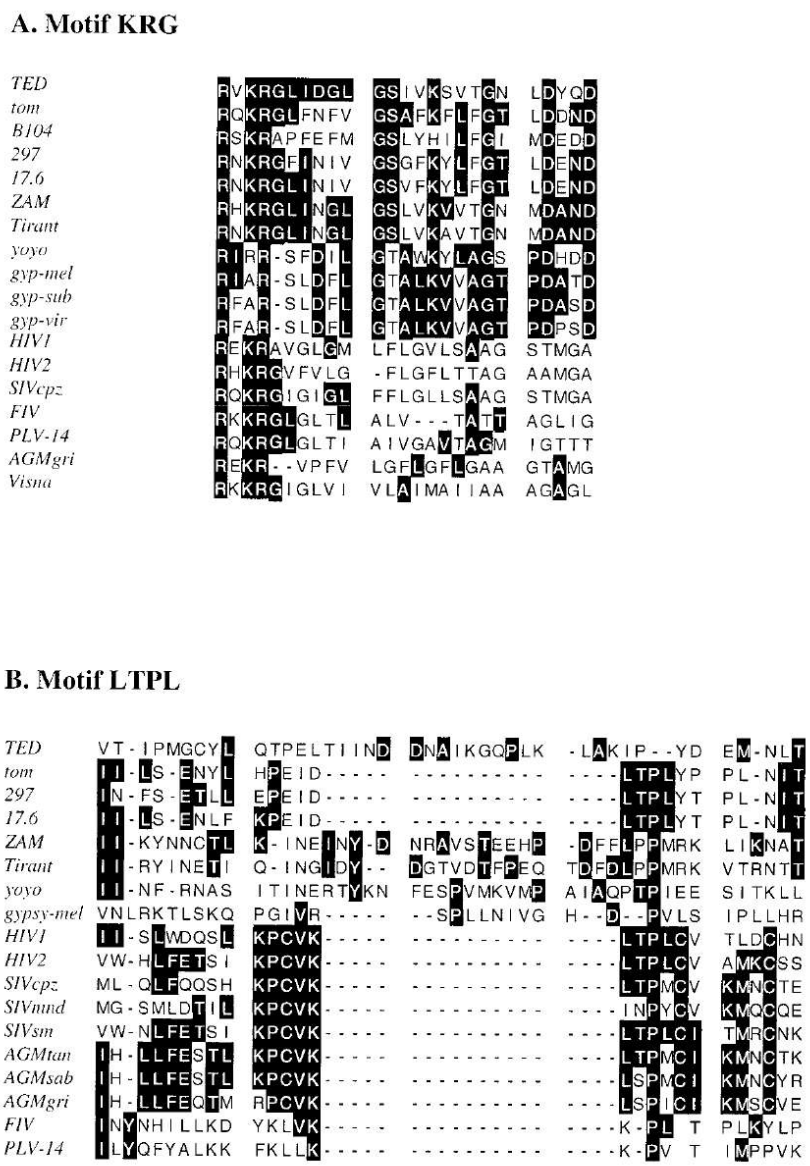


FIG. 1.—Alignments of LTR retrotransposons and retroviruses for the two motifs. In the trees of figure 3, the gaps were removed.

etration of the virus, i.e., its infectious potential. This gene is probably subjected to high selection pressures and so displays a high degree of variability (Hirsch and Curran 1990).

The two motifs were found in LTR retrotransposons, *gypsy*-like elements, and retroviruses, with the exception of the *visna* retroviruses (ovine/caprine lentivirus), which only contain the KRG motif. In *HIV* and *SIV,* the KRG motif is localized between the two glycoproteins at position 500 on the polyprotein (fig. 2). This corresponds to the cleavage site. In the LTR retrotransposons, this motif is found around position 100, in the N-terminal part of the protein. This does not correspond to the cleavage site of *gypsy*.

The position of the LTPL motif is more variable in all sequences. In *HIV* and *SIV,* this motif is found at the beginning of glycoprotein gp120, around position 100, in a region known as the "variable region," or, more precisely, in the V2 loop. In the puma retrovirus (*PLV-14*) and *FIV,* the LTPL and KRG regions are close together. In the retrotransposons and *gypsy*-like elements, the LTPL motif is localized in the C-terminal end of the protein, around position 400.

Other similarities were detected between the *env* gene of the sequences previously mentioned. These similarities were generally limited to a few retrotransposons, and most of them were localized in *HIV* gp120. The LTPL and KRG motifs were therefore used both separately and together to infer unrooted trees.

Trees Based on the Motifs

Using PAUP, four trees based on the KRG motif were obtained with a consistency index of 0.825. There is no topological difference between these four trees. Three distinct clusters were observed (fig. 3*A*): a cluster of retrotransposons including *ZAM, Tirant, TED, tom,*

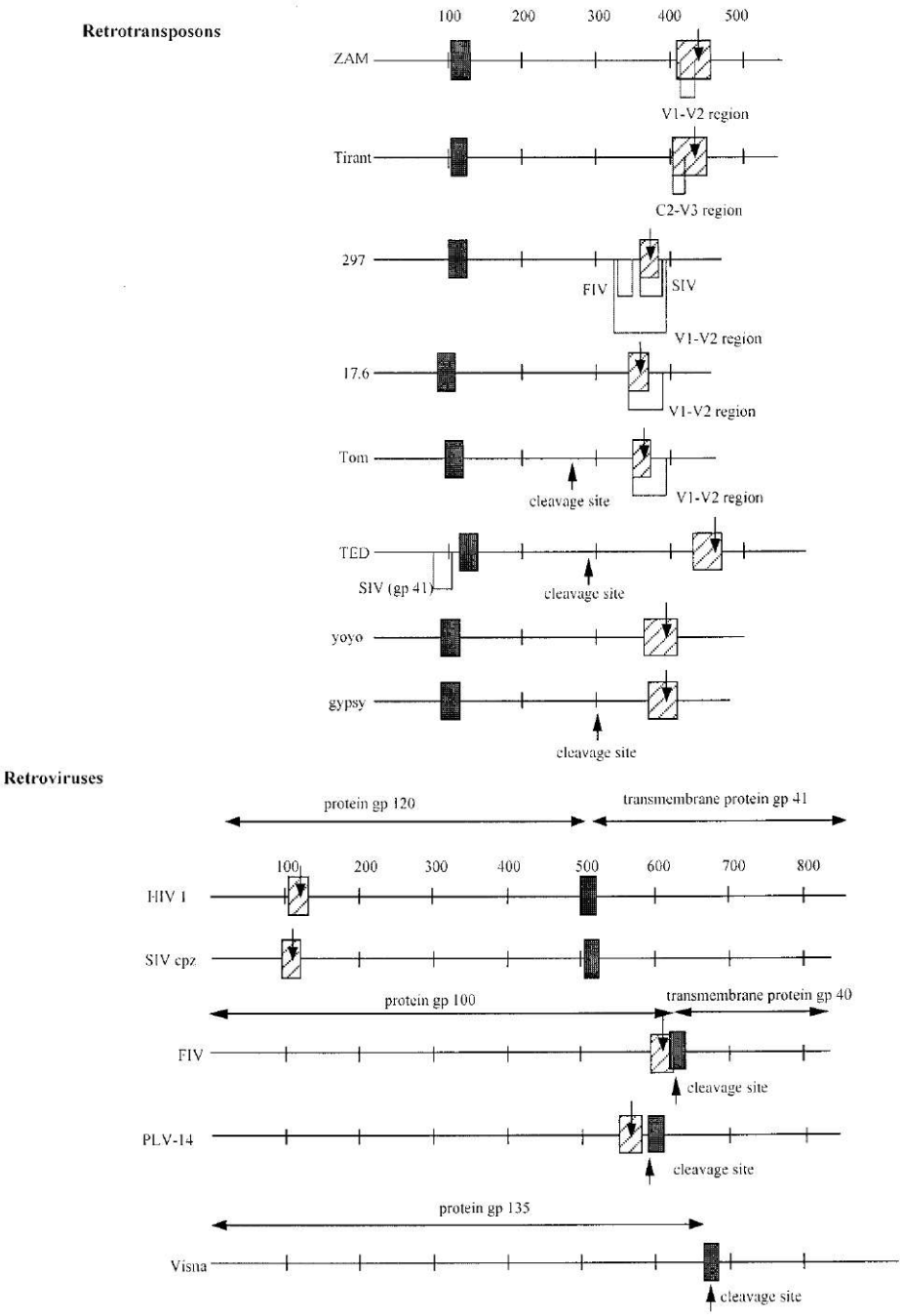


FIG. 2.—Positions of the two common motifs in retroviruses and LTR retrotransposons. The KRG motif is represented by a gray box, and the LTPL motif is represented by dashed box. Gp120 and gp41 are the glycoproteins of HIV1. V1, V2, and V3 are the variable loops of HIV1. C2 is the constant region between V2 and V3. ↓ = position of the first L of the LTPL motif. ↑ = position of the cleavage site.

B104, 297, 17.6, and *yoyo*; a cluster consisting of the *gypsy* elements of *Drosophila*; and a cluster of retroviruses. The trees obtained by the other methods are not fundamentally different, and the element classification remains the same. Any differences are due to local rearrangements of elements within each group, but in all cases the groups remain the same.

The topologies of the trees based on the LTPL motif are the same in the four trees obtained with PAUP. We found some changes inside the clusters previously defined. Two subgroups of retroelements can be distinguished, one consisting of *ZAM, Tirant,* and *TED,* and one consisting of *tom, 17.6* and *297* (fig. 3*B*). *Yoyo* occupies an intermediate position between these two groups. The last retroelement, *gypsy,* is associated with feline retroviruses (*FIV* and *PLV-14*). Similarly, some subgroups of retroviruses can be observed, such as *HIV2,* associated with *SIVsm,* and *HIV1,* associated with

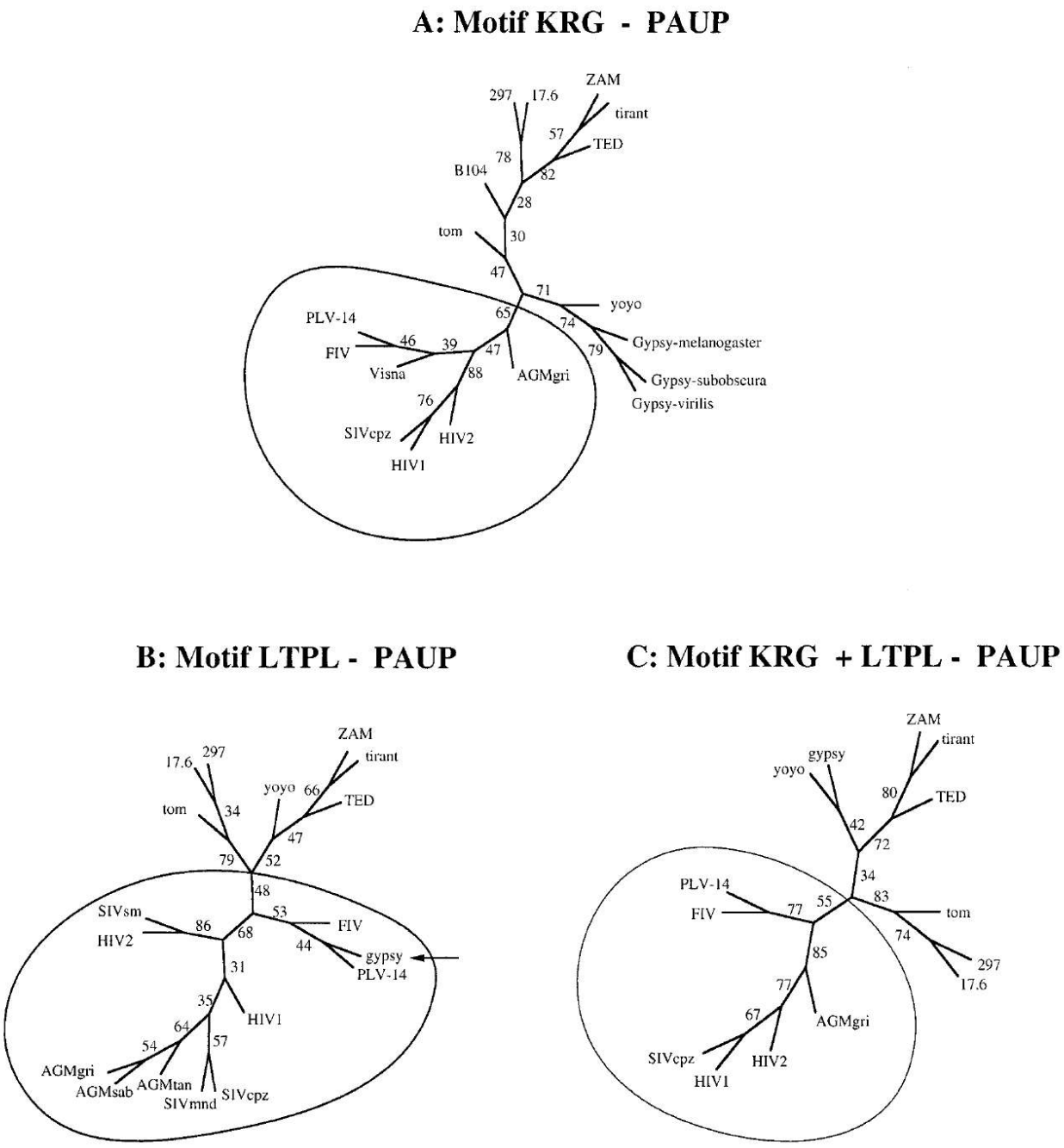


FIG. 3.—Trees obtained from the matrices given in figure 2 using the PAUP program. Trees obtained using the neighbor-joining method and those obtained using the maximum-likelihood method are similar. Numbers given along the branches are the bootstrap values after 100 repetitions. See table 1 for the accession numbers of the sequences. *Gypsy* is the *D. melanogaster* element. All the sequences used have the two motifs except the *B104* retrotransposon and the *visna* retrovirus.

green monkey retroviruses (*AGMgri, AGMsab, AGMtan*) and *SIV* (*SIVmnd* and *SIVcpz*).

Figure 3*C* shows the tree based on the two motifs. Clear segregation is again observed between the different types of elements. Among the retrotransposons, two groups can be identified: one consisting of *tom, 17.6,* and *297,* and one consisting of *ZAM, Tirant,* and *TED.* The *yoyo* and *gypsy* elements are grouped together. The feline retroviruses are closely related to the retrotransposons and to the *gypsy*-like elements, whereas human and monkey retroviruses are more distant.

Comparison with Other Trees

The tree based on the *env* gene and including both motifs was then compared with the trees obtained from reverse transcriptase and integrase (fig. 4). Whichever domain was used, the retroviruses were separated from the other elements. From the nonretroviruses, the *gypsy*/

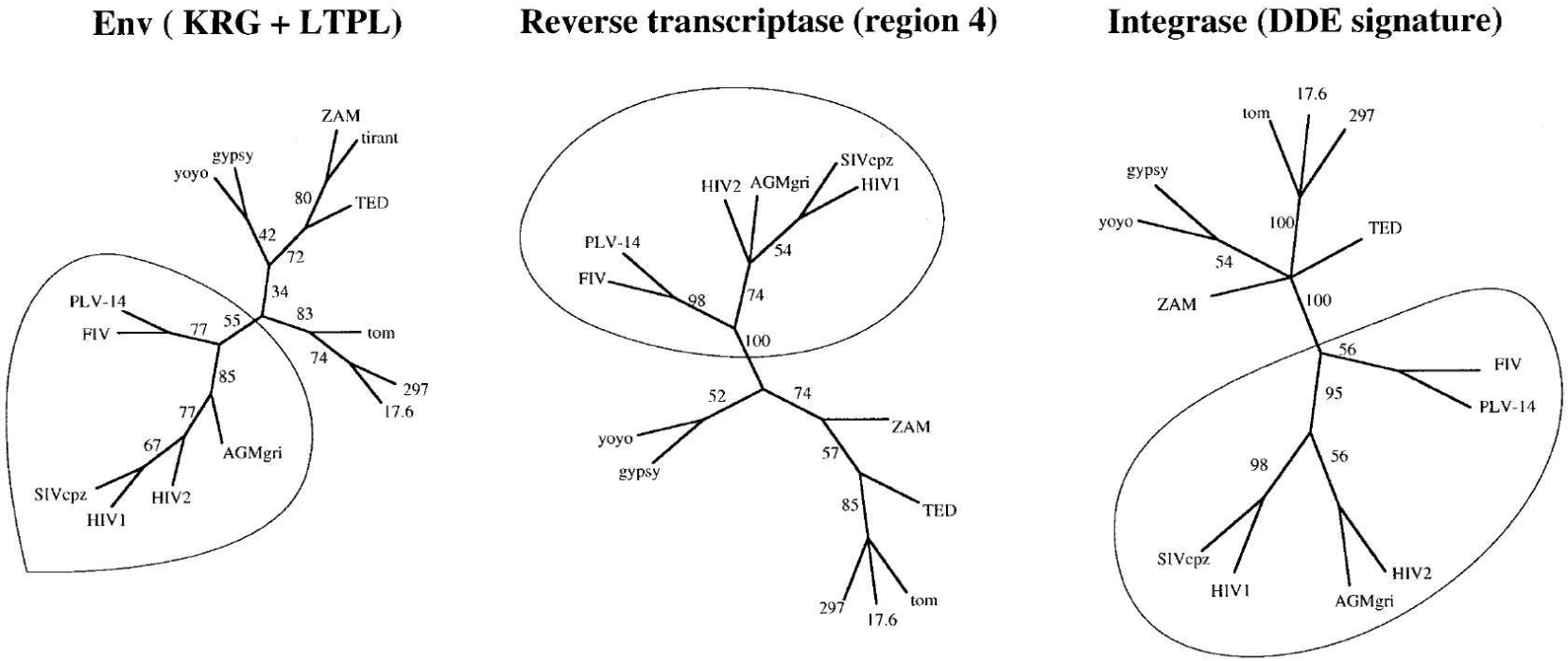


FIG. 4.—Trees obtained using the PAUP program. Trees obtained using the neighbor-noining method and those obtained using the the maximum-likelihood method are similar. Numbers given along the branches are the bootstrap values after 100 repetitions. See tables 1–3 for the accession numbers of the sequences. *Gypsy* is the *D. melanogaster* element.

yoyo and *tom/297/17.6* groups were clearly distinguished, whereas the positions of *ZAM* and *TED* were not stable. The bootstrap values were relatively weak, but the same classification of the elements was obtained regardless of the method and domain used.

Discussion

In this study, we investigated the relationships between nonretroviral transposable elements and retroviruses by comparing their *env* proteins. Two motifs were detected. Their locations within the *env* gene are clearly different. However, it must be stressed that the organization of the insect retroviral element *gypsy* is similar to that of the LTR retrotransposons. Whatever the methods used, trees obtained from an alignment of these motifs are similar to those deduced from other domains, such as reverse transcriptase and integrase.

It is puzzling that no endogenous retrovirus or avian retrovirus was detected. It is possible that the *env* gene of endogenous retroviruses is too degenerate to present detectable homology. This could be due to the high mutation rate and reorganization of the *env* gene. Furthermore, the two motifs were not detected in the envelope gene of the plant retrotransposon *SIRE-1* (Laten, Majumdar, and Gaucher 1998). This element, which has been detected in the soybean genome, is a member of the *copia/Ty1* superfamily.

The main homologies identified in retroviruses and nonretroviral elements have different protein locations in the *env* gene. They are frequently found in the variable zones of retroviruses, and especially in the V1-V2 loop. These zones are responsible for target cell recognition and retrovirus tropism (Rizzuto et al. 1998; Wyatt et al. 1998). If these motifs have functional roles, they are probably different in LTR retrotransposons and *gypsy*-like elements. On the one hand, the consensus KRG motif corresponding to a cleavage site in retroviruses (R/K-X-K/R-R, where X is any amino acid in *HIV1*; Bosch and Pawlita 1990) is not the cleavage site of *gypsy*. On the other hand, the LTPL motif is not well conserved in the *gypsy*-like elements (see, for instance, Alberola and de Frutos 1996), and there is no evidence that their functions are similar to those in retrovirus *env* genes.

Models of TE evolution suggest that LTR retrotransposons could have given rise to retroviruses (Temin 1980; Flavell 1981). More recently, this was also proposed on the basis of reverse transcriptase analysis carried out by Xiong and Eickbush (1990), by Flavell et al. (1995), and by McClure (1991). In this work, the trees obtained using the two motifs (KRG + LTPL) show that the two groups *gypsy/yoyo* and *tom/297/17.6* can be distinguished from retroviruses. Analysis of integrase and reverse transcriptase leads to a similar conclusion. However, in the case of the latter domain, the *gypsy/yoyo* group seems to be more closely related to the retroviruses (see fig. 4).

Can the *env* gene be acquired by retrotransposons and lost by retroviruses? In all trees, *gypsy* is always close to the LTR retrotransposons, and its status as an insect retrovirus is established (Kim et al. 1994; Song et al. 1994). A possible source of bias is that all of the retrotransposons were from arthropods and all of the retroviruses were from vertebrates. This was mainly due to the small number of known vertebrate LTR retrotransposons and the small number of retroviruses detected in invertebrates or plants. In vertebrates, LTR retrotransposons have been reported only from a small domain of the reverse transcriptase. These sequences and those of *gypsy,* retroviruses, and other LTR retrotransposons were used to infer a tree (fig. 5). This tree clearly shows that there are three groups of sequences: LTR retrotransposons without any ORF3 including those with an *env* gene without the two motifs detected here, like *SIRE-1,* those with a partial or complete *env* gene containing the two motifs, and the retroviruses. Moreover, the branch between retroviruses and the first group of LTR retrotransposons is supported by a bootstrap value of 100, whereas discrimination between retroviruses and the second group of LTR retrotransposons is less robust. This could be due to the evolutionary proximity of these two groups.

Another interesting feature is that *gypsy* and *yoyo* seem to be closely related to the spumavirus, *HFV.* This characteristic is also found when the *gag* region is analyzed (data not shown). For this gene, the major homology region (MHR) is present in all retroviruses and in some retrotransposons but not in the spumavirus or arthropod *gypsy*-like elements.

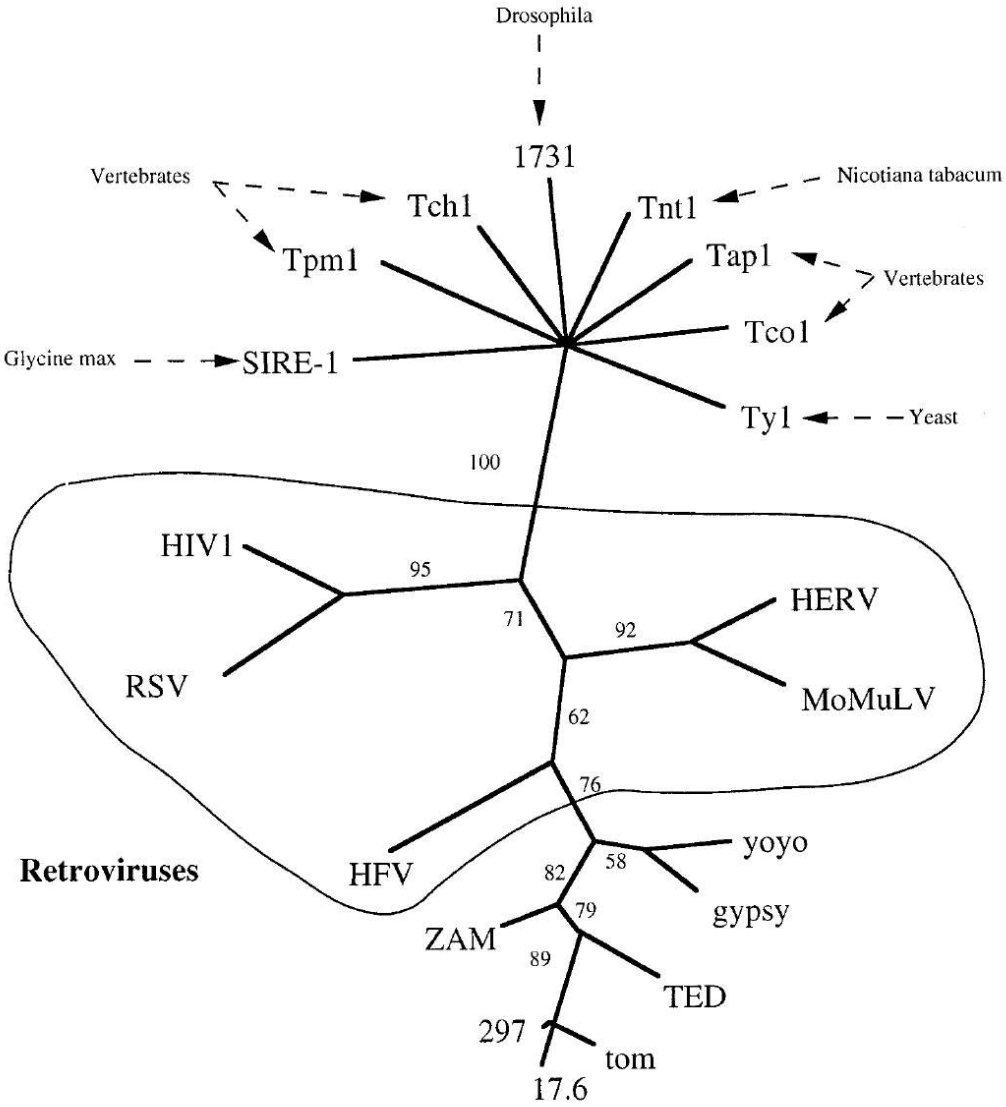
All of these findings suggest a direct relationship between nonretroviral elements and retroviruses. Evolution between these two entities probably occurs in both directions. On the one hand, inactivation of the *env* gene of retroviruses could have given rise to LTR retrotransposons. On the other hand, acquisition of a functional *env* gene by LTR retrotransposons could have given rise to retroviruses. This possibility was discussed by McClure (1991).

The variability of the motif positions in the *env* gene can be interpreted in at least two ways: it may result from rearrangements of different parts of the gene or from modular acquisition of the gene. The modification of the *gypsy* position in the LTPL tree, compared with the KRG tree (fig. 3*A* and *B*), could argue in favor of a modular acquisition of the two motifs. However, the alternative hypothesis is that the evolution rate of the LTPL motif in *gypsy* is different from that of the other errantiviruses.

For the moment, there is no decisive argument for or against rearrangements or modular acquisitions. However, there is a contrast between the apparent similarity of motif positions among the retrotransposons and *gypsy* elements and their variability in retroviruses. According to the modular-acquisition hypothesis, this could be because all of the retroelements with partial *env* gene and *gypsy*-like elements have a common ancestral retrovirus. This assumes that no rearrangement has occurred since the emergence of these elements.

We have not been able to answer the question raised by Flavell (1981) about the evolution of retroviruses from transposable elements. It is interesting to note

(*SIRE-1* element contain an *env* gene without the two motifs described in the other group of LTR-retrotransposons)



LTR-retrotransposons with complete or partial *env* gene

FIG. 5.—PAUP tree obtained from an alignment of region 4 of the reverse transcriptase defined by Xiong and Eickbush (1990) for LTR retrotransposons and retroviruses. See table 2 for the accession numbers of the sequences. *Gypsy* is the *D. melanogaster* element.

the intermediate position of retroviruses (fig. 5) between the LTR retrotransposons with no *env* genes and those with complete or partial genes. However, the question will probably be answered only when intermediate forms are identified.

Acknowledgments

We thank L. Bailey for the aligned sequences of the *env* genes of *HIV,* African green monkey retroviruses, and *SIV.* This work was supported by the Programme Génome no. 23.

LITERATURE CITED

ALBEROLA, T. M., and R. DE FRUTOS. 1996. Molecular structure of *gypsy* element of *Drosophila subobscura* (*gypsyDs*) constituting a degenerate form of insect retroviruses. Nucleic Acids Res. **24**:914–923.

ALTSCHUL, S. F., T. L. MADDEN, A. A. SCHÄFFER, J. ZHANG, Z. ZHANG, W. MILLER, and D. J. LIPMAN. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**:3389–3402.

BAILEY, T. L., and C. ELKAN. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Pp. 28–36 *in* R. ALTMAN, ed. Proceedings of the Sec-

ond International Conference on Intelligent Systems for Molecular Biology. AAAI Press, Menlo Park, Calif.

Bayev, A. A., N. V. Lyubomirskaya, E. B. Dzhmagaliev, E. V. Ananiev, I. G. Amiantova, and Y. V. Ilyin. 1984. Structural organization of transposable element *mdg4* from *Drosophila melanogaster* and a nucleotide sequence of its long terminal repeats. Nucleic Acids Res. **12**:3707–3723.

Boeke, J. D., T. Eickbush, S. B. Sandmeyer, and D. F. Voytas. 1998. Metaviridae. *In* F. A. Murphy, ed. Virus taxonomy: ICTV VIIth report. Springer-Verlag, New York.

Bosch, V., and M. Pawlita. 1990. Mutational analysis of the human immunodeficiency virus type 1 *env* gene product proteolytic cleavage site. J. Virol. **64**:2337–2344.

Capy, P., C. Bazin, D. Higuet, and T. Langin. 1997*a*. Dynamic and evolution of transposable elements. R. G. Landes Company, Austin, Tex.

Capy, P., T. Langin, D. Higuet, P. Maurer, and C. Bazin. 1997*b*. Does the integrase of LTR- retrotransposons and most of the transposases of class II elements share a common ancestor? Genetica **100**:63–72.

Capy, P., R. Vitalis, T. Langin, D. Higuet, and C. Bazin. 1996. Relationships between transposable elements based upon the integrase-transposase domains: is there a common ancestor? J. Mol. Evol. **42**:359–369.

Coffin, J. M. 1990. Retroviridae and their replication. Pp. 1437–1500 *in* B. N. Fields and D. M. Knipe, ed. Virology. Vol. 2. Raven Press, New York.

Dayhoff, M. O., R. M. Schwartz, and B. C. Orcutt. 1978. A model of evolutionary change in proteins. Pp. 345–352 *in* M. O. Dayhoff, ed. Atlas of protein sequence and structure. Vol. 5, suppl. 3. National Biomedical Research Foundation, Washington, D.C.

Felsenstein, J. 1993. PHYLIP (phylogeny inference package). Version 3.5.c. Distributed by the author, Department of Genetics, University of Washington, Seattle.

Finnegan, D. J. 1989. Eukaryotic transposable elements and genome evolution. Trends Genet. **5**:103–107.

Flavell, A. J. 1981. Did retroviruses evolve from transposable elements? Nature **289**:10–11.

Flavell, A. J., V. Jackson, M. P. Iqbal, I. Riach, and S. Waddell. 1995. *Ty1-copia* retrotransposon sequences in Amphibia and Reptilia. Mol. Gen. Genet. **246**:65–71.

George, D. G., L. T. Hunt, and W. C. Barker. 1988. Current methods in sequence comparison and analysis. Pp. 127–149 *in* D. H. Schlesinger, ed. Macromolecular sequencing and synthesis. A. R. Liss, New York.

Gilbert, D. G. 1998. SeqPup: a biosequence editor. Version 0.8c. Distributed by the author at seqpup@bio.indiana.edu.

Hirsch, M. S., and J. Curran. 1990. Human immunodeficiency viruses. Pp. 1545–1570 *in* B. N. Fields and D. M. Knipe, eds. Virology. Vol. 2. Raven Press, New York.

Kim, A., C. Terzian, P. Santamaria, A. Pélisson, N. Prud'homme, and A. Bucheton. 1994. Retroviruses in invertebrates: the *gypsy* retrotransposon is apparently an infectious retrovirus of *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA **91**:1285–1289.

Laten, H. M., A. Majumdar, and E. A. Gaucher. 1998. *SIRE-1,* a *copia/Ty1* retroelement from soybean, encodes a retroviral envelope-like protein. Proc. Natl. Acad. Sci. USA **95**:6897–6902.

McClure, M. A. 1991. Evolution of retroposons by acquisition or deletion of retrovirus-like genes. Mol. Biol. Evol. **8**: 835–856.

Rizzuto, C. D., R. Wyatt, N. Hernandez-Ramos, Y. Sun, P. D. Kwong, W. A. Hendrickson, and J. Sodroski. 1998. A conserved HIV gp 120 glycoprotein structure involved in chemokine receptor binding. Science **280**:1949–1953.

Sokal, R. R., and F. J. Rolf. 1995. Biometry. W. H. Freeman and Company, New York.

Song, S. U., T. Gerasimova, M. Kurkulos, J. D. Boeke, and V. G. Corces. 1994. An env-like protein encoded by a Drosophila retroelement: evidence that gypsy is an infectious retrovirus. Genes Dev. **8**:2046–2057.

Strimmer, K., and A. von Haeseler. 1996. Quartet puzzling: a quartet maximum likelihood method for reconstructing tree topologies. Mol. Biol. Evol. **13**:964–969.

Swofford, D. L. 1993. Phylogenetic analysis using parsimony. Version 3.1.1. Smithsonian Institution, Washington, D.C.

Swofford, D. L., G. J. Olsen, P. J. Waddell, and D. M. Hillis. 1996. Phylogenetic inference. Pp. 407–514 *in* D. M. Hillis, C. Moritz, and B. K. Mable, eds. Molecular systematics. 2nd edition. Sinauer, Sunderland, Mass.

Temin, H. M. 1980. Origin of retroviruses from cellular moveable genetic elements. Cell **21**:599–600.

Varmus, H. and P. Brown. 1989. Retroviruses. Pp. 53–108 *in* D. E. Berg and M. M. Howe, ed. Mobile DNA. American Society for Microbiology, Washington, D.C.

Wyatt, R., P. D. Lwong, E. Desjardin, R. W. Sweet, J. Robinson, W. A. Hendrickson, and J. G. Sodroski. 1998. The antigenic structure of the HIV gp 120 envelope glycoprotein. Nature **393**:705–711.

Xiong, Y., and T. H. Eickbush. 1990. Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J. **9**:3353–3362.

Manolo Gouy, reviewing editor

Accepted May 25, 1999

Article 2 : Is the evolution of transposable elements modular?

Genetica 107:15-25, 1999

Genetica **107:** 15–25, 1999.

Is the evolution of transposable elements modular?

Emmanuelle Lerat, Frédéric Brunet, Claude Bazin & Pierre Capy*
*Laboratoire Populations, Génétique et Evolution, Centre National de la Recherche Scientifique, 91198 Gif/Yvette Cedex France; *Author for correspondence (Phone: 33.1.69.82.37.09; Fax: 33.1.69.07.04.21; E-mail: capy@pge.cnrs-gif.fr)*

Accepted 11 February 2000

Key words: DDE signature, *env*, evolution, *gag*, transposable elements

Abstract

The evolution of transposable element structures can be analyzed in populations and species and by comparing the functional domains in the main classes of elements. We begin with a synthesis of what we know about the evolution of the *mariner* elements in the *Drosophilidae* family in terms of populations and species. We suggest that internal deletion does not occur at random, but appears to frequently occur between short internal repeats. We compared the functional domains of the DNA and/or amino acid sequences to detect similarities between the main classes of elements. This included the *gag*, reverse transcriptase, and *envelope* genes of retrotransposons and retroviruses, and the integrases of retrotransposons and retroviruses, and transposases of class II elements. We find that each domain can have its own evolutionary history. Thus, the evolution of transposable elements can be seen to be modular.

Introduction

Transposable elements (TEs) are present in all genomes and may account for a substantial part of them. They are divided into two main classes depending on their transposition mechanisms (Finnegan, 1989) – retrotransposons use an RNA intermediate, while transposons, *sensus stricto*, use a DNA intermediate. A detailed classification was recently proposed that uses the general structure of the elements, the existence of similar motifs (adjacent residues) and signatures (non-adjacent residues), and identities and similarities of DNA and amino acid sequences to define sub-classes, super-families, families and sub-families (Capy et al., 1997b).

The diversity of these elements, the distributions of the main families, or classes, and their genomic locations raised the question of their own evolution and their interactions with the host genomes. While they were classified as 'selfish DNA' or 'selfish gene' at the beginning of the 80's (Doolittle & Sapienza, 1980; Orgel & Crick, 1980), their interactions, their impact on genome function and structure, and on the adaptation and evolution of populations and species have changed our vision. Their mobility can certainly be deleterious for an individual, but they can be useful to populations and species. In other words, while they can be considered to be a genetic load, they can also be seen as useful parasites.

Their evolution can be investigated *via* their dynamics within and between species, their co-evolution with the host genomes, and their own structural evolution. Indeed, while TEs have an impact on the plasticity of the host genome, they are themselves very plastic. This paper is a synthesis of our results on the structural evolution of these elements.

The evolution of TE structure can be seen at several levels, from elements of the same family to a general comparison of all elements having homologous domains independently of their family or class. This evolution can be examined using the DNA and amino-acid sequences and by comparing them to secondary structures. This report focuses on the structural evolution of the *mariner* element in the *Drosophilidae*

family and then compares the homologous domains in at least two classes or super-families of elements, including the *gag*, integrase and *env* domains. The results obtained are compared to those of Xiong and Eickbush (1990) and of McClure (1991, 1993).

Materials and methods

Motif detection and phylogenetic analyses

All the sequences used in this work were extracted from the EMBL, GENBANK, NBRF and SWISS-PROT, with SRS 5.0 at the INFOBIOGEN site. Motifs (adjacent amino acids) and signatures (non-adjacent amino acids) were detected using BLASTP 2.0, which can cope with gaps (Altschul et al., 1997). The parameters were chosen to obtain even those sequences with a low score. The following settings were used: Expect = 1000 instead of 10, Description = 500 instead of 100 and Alignment = 500 instead of 100. No new motifs or signatures were detected using the MEME program (Bailey & Elkan, 1994) with all the sequences used in the alignments. The sequences were then aligned manually using the SeqPup editor (version 0.8c, Gilbert, 1998).

The phylogenetic signal in our data was assessed using a PTP test (Swofford et al., 1996). The tests were highly significant ($P < 0.001$ with 1000 replicates) in all cases.

Trees were inferred using PAUP (version 3.1.1, Swofford, 1993), based on the parsimony principle, Puzzle (version 3.1, Strimmer & von Haeseler, 1996), using the maximum likelihood method based on the 'quartet puzzling' algorithm, and by the Neighbor-Joining method in the PHYLIP package (Felsenstein, 1993). The models of amino acid substitution used were those proposed in PHYLIP: the Dayhoff PAM matrix (1978); the George–Hunt–Barker matrix (1988) and the Hall matrix. Since all the methods give similar topologies, only the results obtained with PAUP are reported. The heuristic search was used with stepwise addition (random addition and 10 replicates) and TBR branch swapping options.

Secondary structure analysis

The secondary structures of the protein domains were compared for few elements using the DRAWHCA program (Lemesle-Varloot et al., 1990). The HCA method (Hydophobic Cluster Analysis), developed by the group of J.P. Mornon, uses the hydrophobic cluster, which mainly corresponds to the internal part of the secondary structures. This method is particularly useful for proteins with little similarity (below 25–30%) when alignment becomes difficult by classical methods. The basic principal of this method is described in Gaboriaud et al. (1987).

Results

Evolution of Mariner *structure*

Mariner is a class II element encoding a 346 amino acid transposase. This element was firstly described in *D. mauritiana* (Jacobson, Medhora & Hartl, 1986) and has since been detected in many organisms like drosophilids (see Capy et al., 1994 for review), several mammals including humans (Auge-Gouillou et al., 1995; Morgan, 1995; Oosumi, Belknap & Garlick, 1995; Reiter et al., 1996; Robertson et al., 1996; Robertson & Martos, 1997; Robertson & Zumpano, 1997), a centipede (Robertson & MacLeod, 1993), nematodes (Sedensky et al., 1994; Wiley et al., 1997; Grenier et al., 1999), platyhelminths (Garcia-Fernàndez et al., 1993; Robertson, 1997), hydra (Robertson, 1997) and plants (Jarvik & Lark, 1998).

Mariner-like elements have mainly been detected by PCR amplification of a region lying between two conserved motifs used to design primers. Most of these elements are dead, since there is one or more stop codons and/or frameshifts in the translated sequences. Most of the elements amplified using primers from the inverted terminal repeats (ITRs), contain one or more internal deletions (Brunet et al., 1999), suggesting that they are inactive.

Translation of some of these deleted elements (*DTBZ1*, *DTBZ2*, *DTZW3*, *DTNB1*) in *D. teissieri*, shows that the only stop codon corresponds to the last three nucleotides of the element, that is, the last three nucleotides of the 3′ ITR. These elements have been found in geographically distant populations that are believed to have been isolated from each other for several thousands of years (Joly & Lachaise, personal communication). Thus, Brunet et al. (1996) suggested that such an element regulates the active elements in *D. teissieri* genome, like the putative active copy *teis32* element described by Maruyama and Hartl (1991).

Comparison of complete and deleted elements reveal that deletions frequently occur between short repeated motifs of a few base pairs. This is seen in

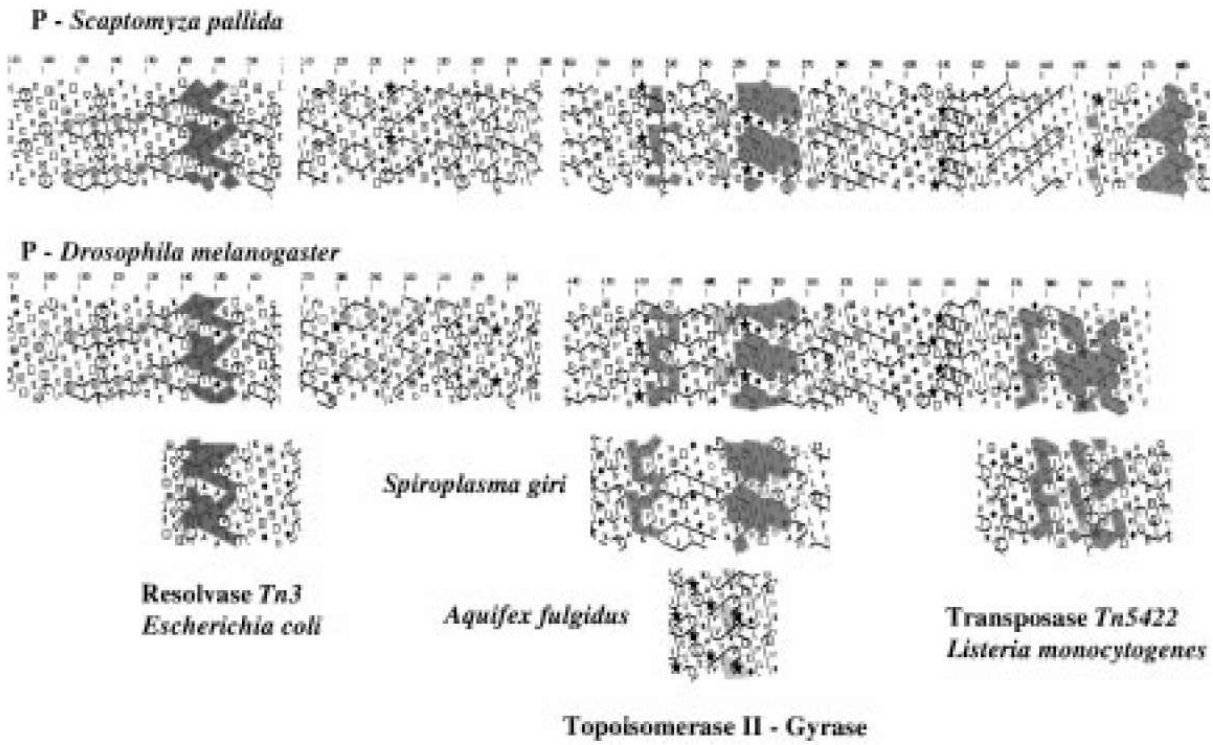


Figure 1. Alignment of secondary structures deduced from DrawHCA of the transposases of *P* elements and similarities for each exon.

the deleted elements of *D. teissieri* (Brunet et al., 1996), and in other *Drosophila* species (Brunet et al., in prep.). Two copies of a given motif are present in the complete elements, while a single copy is generally found in the deleted ones. This strongly suggests that internal recombination occurred. However, there may be a repetition of several bp at the breaking point of the deletion in some case, but no repeats in the complete element. This cannot be explained by internal recombination, but may occur via abortive gap-repair, as for the occurrence of *Ds* elements in *Zea mays* (Rubin & Levy, 1997).

Transposase–integrase comparison

The DDE elements

Most of the integrases of LTR retrotransposons and retroviruses, and the transposases of class II elements contain a DD(35)E or DD(34)D signature (Fayet et al., 1990; Doak et al., 1994; Capy et al., 1996; Capy et al., 1997c). Trees based on the alignment of this region show that the LTR-retrotransposons and retroviruses form a single group; the members of the *mariner-Tc1* super-family, including the elements of the *pogo-Fot1* family, form a monophyletic cluster; while the *IS* insertions of prokaryotes are spread throughout the tree. Such a distribution suggests that the class II prokaryotic elements could be the origin of the DDE transposases of eukaryotic class II elements, and of the integrases of LTR retrotransposons. The phylogenetic relatedness of *IS30* and retroelement integrase on one hand, and *IS630* and the transposases of the *mariner-Tc1* super-family on the other hand, are two arguments in favor of this. However, it is always possible that the DDE class II elements are derived from the LTR retrotransposon integrase.

The non-DDE elements

Members of the *P* and *hAT* super-families do not have the DD(35)E signature in their transposases. We looked for this signature using several techniques, including the BLASTP, MEME and HCA programs. BLASTP with low stringency parameters detected several sequences which do not correspond to TEs, but whose primary sequences show some similarities with them. These similarities were confirmed by analysis of the secondary structures using the DRAWHCA program.

Figure 1 summarizes the similarities between the last three exons of the *P* elements and several bacterial proteins. The exon 1 of *P* shows some similarit-

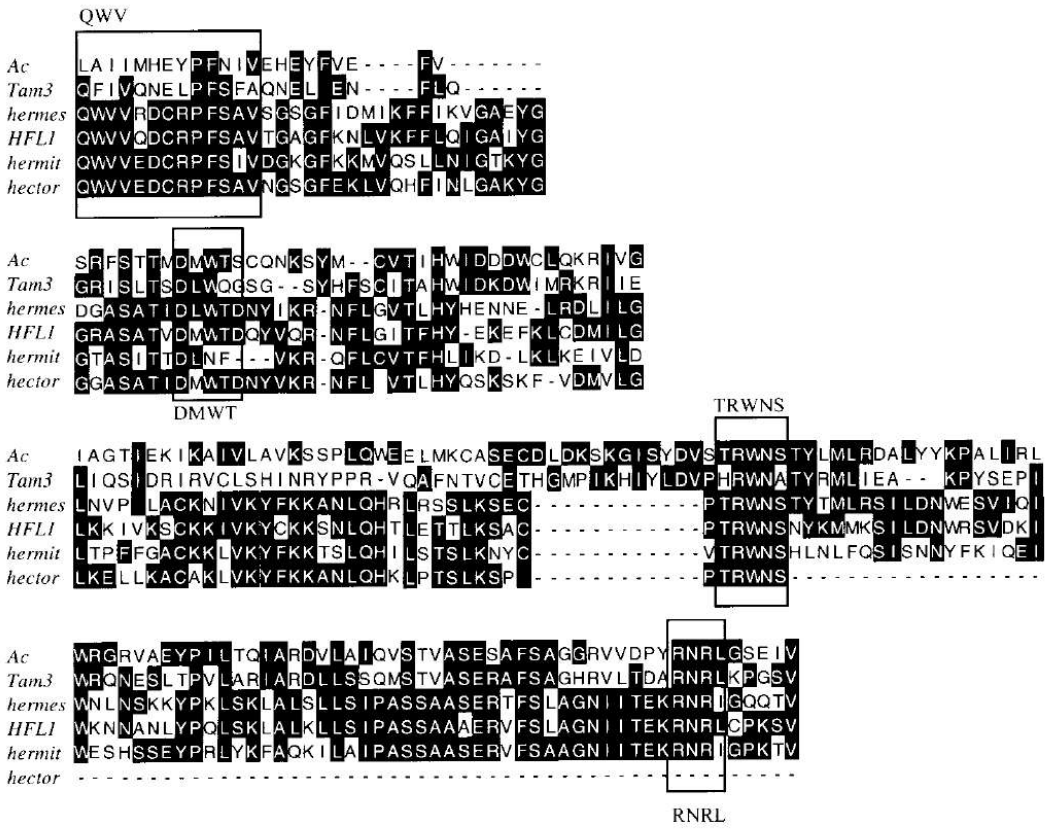


Figure 2. Alignments and conserved motifs in the transposases of the members of the *hAT* superfamily.

ies with the resolvase of the *Tn3* element of *E. coli* (Acc.: P03011), the exon 2 with Topoisomerase II of *Spiroplasma citri* (Acc: Z19108) and *Aquifex fulgidus* (Acc.: AE001072), and exon 3 with the transposase of the *Tn5422* element of *Listeria monocytogenes* (Acc.: L28104).

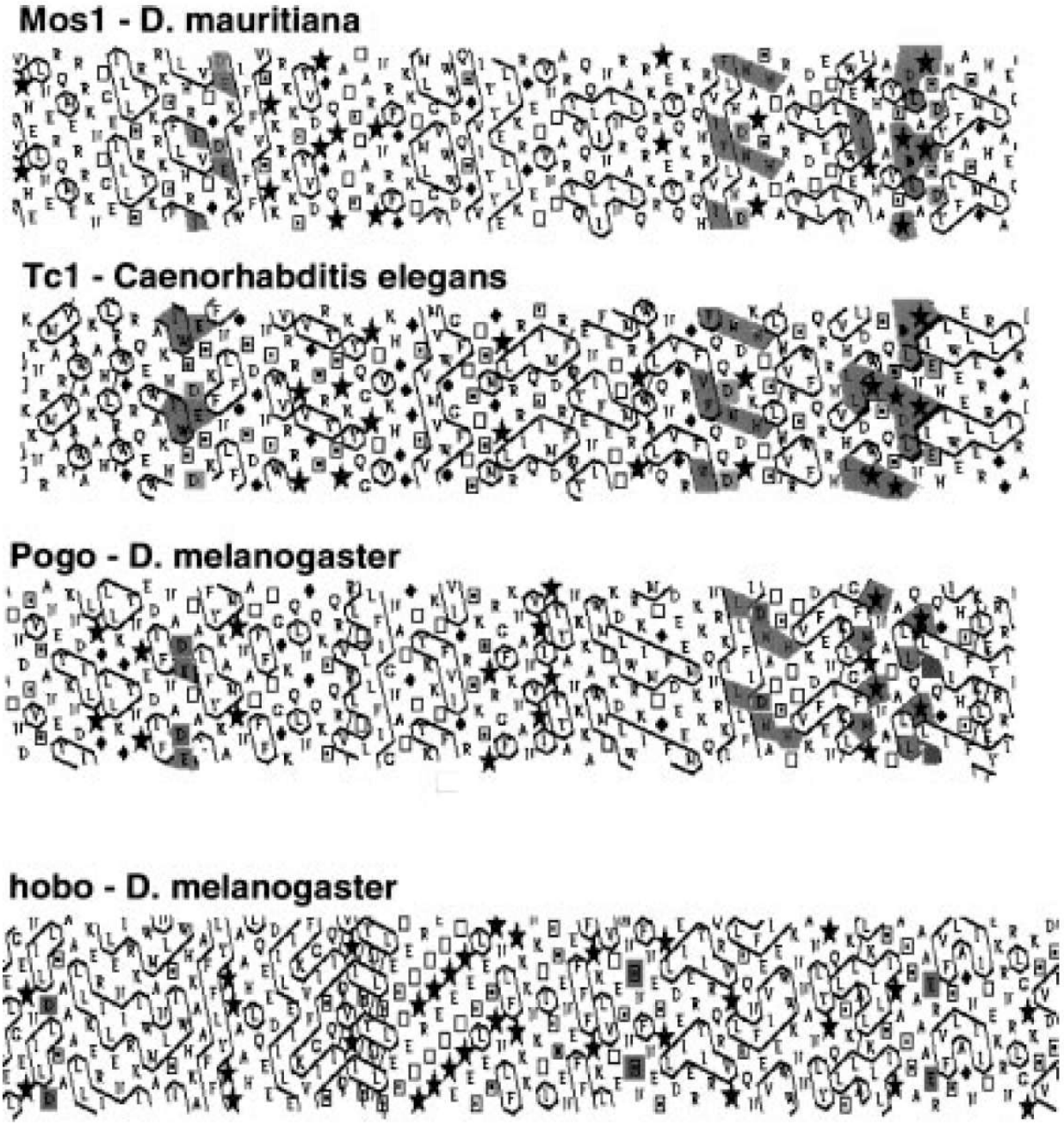
A new conserved region was identified in the members of the *hAT* super-family. This region, named QWV, includes about 32 amino acids (Figure 2) and is located upstream of the three conserved regions corresponding to the DMWT, TRWNS and RNRL motifs mentioned by Calvi et al. (1991). Analysis of the secondary structure indicated similarities with a gyrase of *Campylobacter fetus* (Acc: U25640). The region involved lies upstream of the DMWT motif. This corresponds to the DVDDLLP motif of the *hobo* element and DVEDLIP in the gyrase.

The secondary structure of the putative DDE signature suspected by Bigot, Augé-Gouillou & Periquet, (1996) in the *hAT* superfamily was anlysed. We did this by aligning members of the *mariner-Tc1* superfamily and integrase of *HIV*-1 from their secondary structure (not shown). We then tried to align the *hobo* secondary structure of the region mentioned by these authors with those of the *mariner-Tc1* superfamily (Figure 3), but no alignment was possible, suggesting that the DDE signature mentioned by Bigot, Augé-Gouillou & Periquet, (1996) does not have structure similar to that of the transposases of the *mariner-Tc1* superfamily and integrases of LTR retrotransposons and retroviruses.

gag *and* env *analyses*

The *gag* and *env* genes of retrotransposons and retoviruses were compared using the amino acid sequences. Two conserved regions of the *gag* gene were considered (Figure 4). Two adjacent regions, 3 to 9 amino acids apart, had the complete zinc knuckle motif ($CX_2CX_4HX_4C$). While the first motif, initially described by Covey (1986), was present in all elements, the second one is found in retroviruses but is absent from the *skippy* and *CfT-1* retrotransposons of *Fusarium oxysporum* and *Cladosporium fulvum*.

The second region lay upstream of the first. This is the MHR motif (major homology region) described



re 3. Alignment of secondary structures of the transposase of three members of the *mariner-Tc1* superfamily deduced from DrawHCA and DE signature suspected in the *hobo* element of *D. melanogaster*.

ıe telomeric elements by Pardue et al. (1997). This if is well conserved in retroviruses and is involved irus particle assembly (Mammano et al., 1994; ven et al., 1995). All those elements for which alignments were not convincing were eliminated ı the phylogenetic analysis. This included several ıents of the *hAT* super-family, the retrotransposon *aldo* of *Drosophila buzatii* (Labrador & Fontdev-1994) and the telomeric element *Tras* of *Bombyx i* (Okazaki, Ishikawa & Fujiwara, 1995).

The trees obtained from the sequences aligned are shown Figure 5. The position of the telomeric elements *TART* and *Het-A* of *Drosophila* changed according to the conserved regions used. The MHR signature (Figure 5) indicated that they are closely related to the retroviruses and the *CfT1* retrotransposons, while *SART1*, the telomeric element of *Bombyx mori* was within the cluster of the LINEs-like elements. The tree based on the zinc knuckle signature had a different topology. The telomeric elements of *Drosophila* and

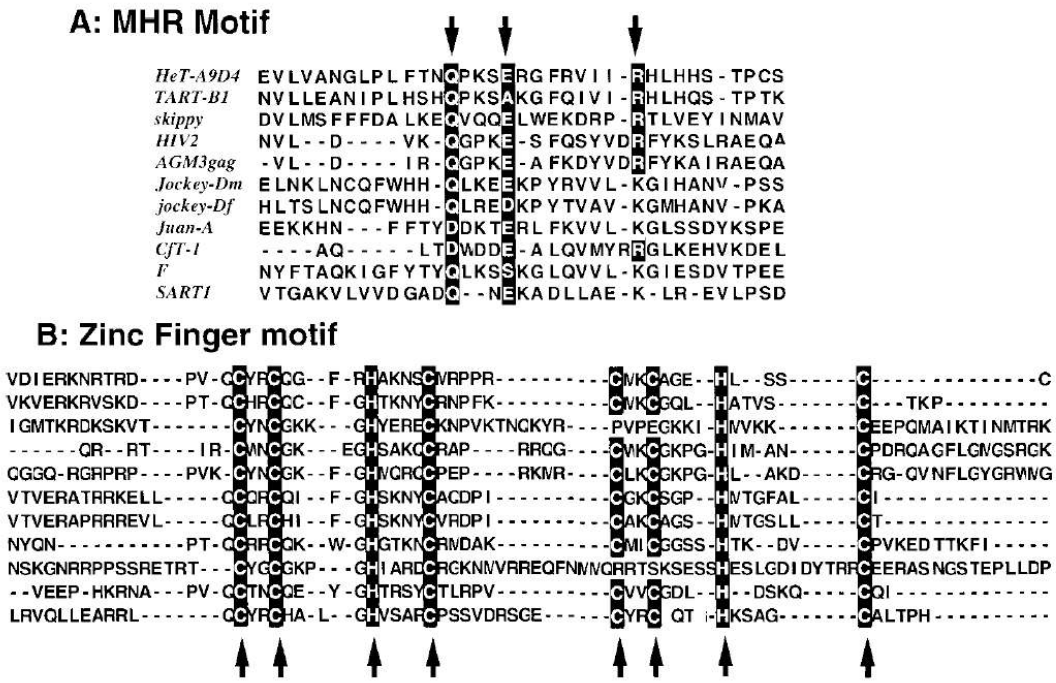


Figure 4. Alignment and conserved residues (arrows) in the *gag* protein of retroelements.

that of *B. mori* were grouped within the LINEs-like elements and, in particular, with the *jockey* elements of *D. funebris* and *D. melanogaster*. Using the two conserved regions together placed the telomeric elements of *Drosophila* in an intermediate position between the retroviruses and the LINEs-like elements, while the *SART1* element clearly belonged to the second group.

The *env* protein has recently been shown to have two regions conserved in retroviruses and some LTR retrotransposons (Lerat & Capy, 1999). The order of the two motifs are different in the two types of sequences: KRG and LTPL for the LTR-retrotransposons and the reverse order in the retroviruses. The trees obtained show no major modifications depending on the motifs used. The feline retroviruses (*FIV* and *PLV-14*) were always the closest related elements to the LTR retrotransposons. The *gypsy* element of *D. melanogaster* is included in the retrovirus group when the LTPL motif is used, while this element clearly belongs to the retrotransposons for the other motif. The tree based on the envelope motifs and those of region 4 of the reverse transcriptase (according to Xiong & Eickbush, 1990) and the DDE signature of the transposase or integrase of both classes of elements give similar classifications of the elements.

Finally, a tree based on the reverse transcriptase, including retroviruses and retrotransposons with or without the complete or partial *env* gene, shows that there is a clear discrimination between the retrotransposons. In this tree (Figure 6), the non-*env* retrotransposons are isolated from *env*-retrotransposons (the errantiviruses) by the retroviruses. The only exception is the *SIRE-1* elements of *Glycine max* (Laten, Majumdar & Gaucher, 1998). This element belongs to the *copia-Ty1* super-family, according to its *pol* gene structure (protease, integrase, reverse transcriptase, RnaseH). This element remains within this super-family in the reverse transcriptase tree, in spite of its *env* gene. However, none of the motifs detected in the *env* gene of the errantiviruses and retroviruses was found in this element.

Conclusions and discussion

The evolution of the structure of TEs within species suggests that deletion occurs at random. It seems that they usually arise between short internal repeats. This phenomenon has been mentioned for other elements, like *hobo* (Streck, MacGaffey & Beckendorf, 1986) and non-LTR retrotranposons (Petrov, Lozovskaya & Hartl, 1996). It can be very frequent in non-LTR retrotransposons. These elements 'create non-functional copies of themselves as a common by-product of their transpositional cycle' according to Petrov, Lo-

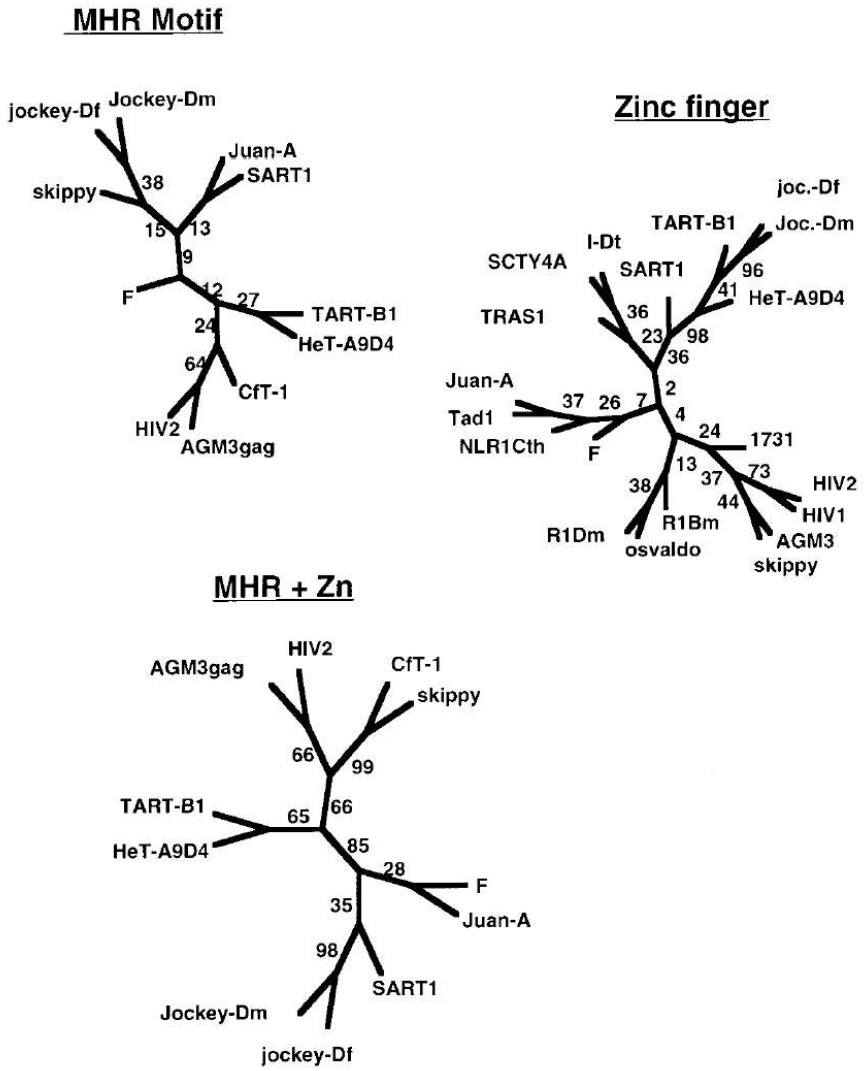


Figure 5. Parsimony trees of retroelements based upon the conserved regions of the *gag* protein (see Figure 4).

zovskaya & Hartl (1996), and they called them the Dead on Arrival elements (DOA). It is difficult to estimate the frequency of these deletions in class II elements. But such copies seem to be very frequent. These deleted copies may have an impact on the regulation of the activity of complete copies. Several reports have demonstrated or suggested such an impact for the *KP* element in the PM system (Black et al., 1987), some deleted elements of *D. teissieri* (Brunet et al., 1996) and for the non-LTR element *I* of *D. melanogaster* (Chaboissier, Bucheton & Finnegan, 1998; Jensen, Gassama & Heidmann, 1999).

The evolution of TE structure could also be due to recombination between elements or between elements and host genes, or again to their transposition mechanisms. No clear recombination has yet been demonstrated within species, although examples of hybrid elements between the *Ty1* and *Ty2* elements of yeast have been reported (Jordan & McDonald, 1999a, and Jordan & McDonald, 1999b).

Comparison of structures of the main classes of TEs shows that it is difficult to propose a general model for their evolution. Indeed, the different functional domains, and sometime the subdomains, may have different origins and histories.

The origins and histories of TE domains can be inferred from their presence/absence and from their physical order in the ORF considered. Figure 7 summarizes the conserved regions compared. The DDE signature in the integrase/transposase domain is shared by the transposases of the *mariner/Tc1* superfamily

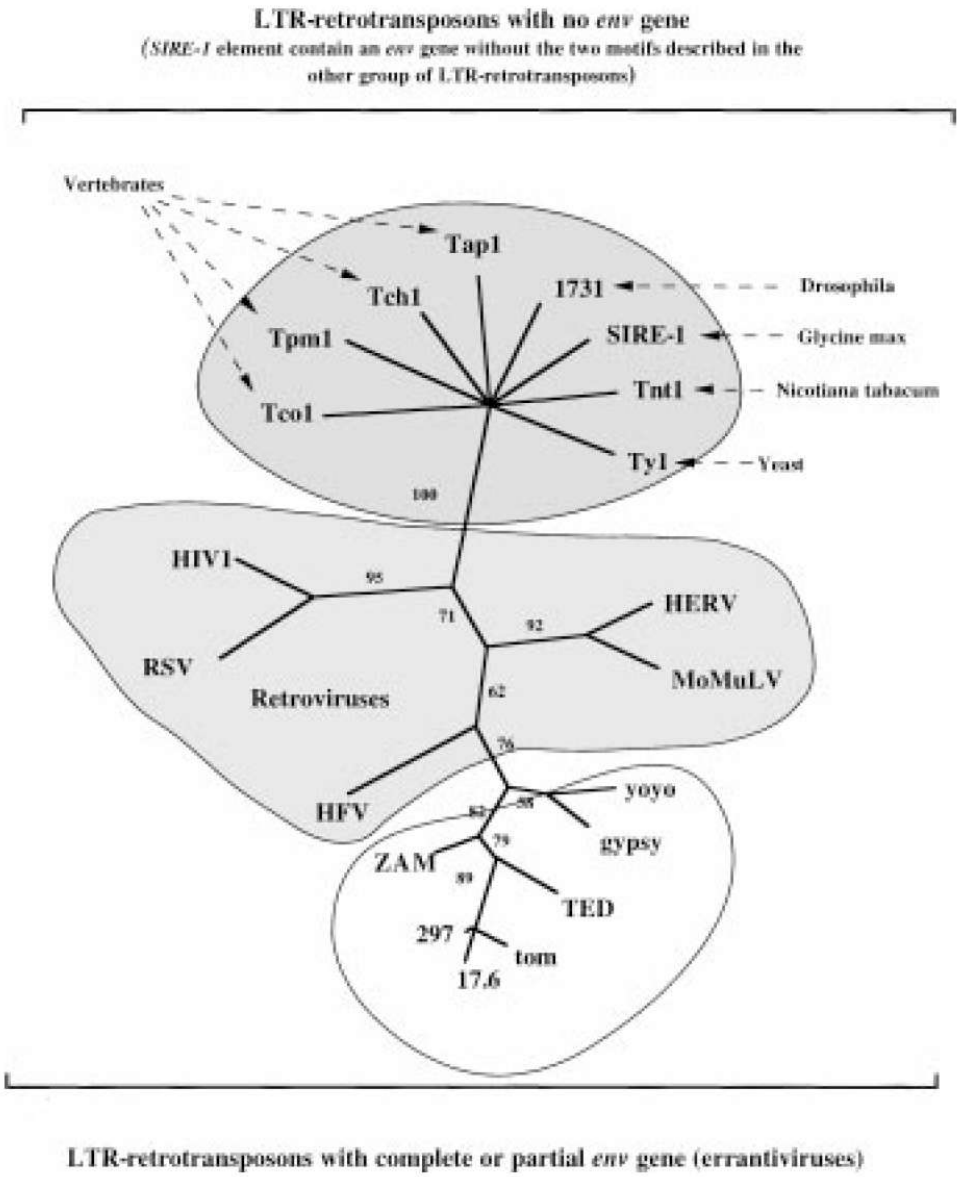


Figure 6. Parsimony tree of retroviruses and LTR retrotransposons with and without an *env* gene. This tree is based upon the conserved region 4 of the reverse transcriptase defined by Xiong and Eickbush (1990). Figure redrawn from Lerat and Capy (1999).

and the integrases of LTR retrotransposons and retroviruses, while the HHCC signature is present only in the retroelements. The non-DDE elements show some similarities between exons 1, 2 and 3, and bacterial transposons and genes, suggesting that these three exons may have different origins. Similarly, the QER signature (signature found in the MHR region) of the *gag* region is relatively well conserved in the retroviruses and *D. melanogaster* telomeric elements and less well conserved in LINEs-like elements.

The origins and histories of TE domains can also be determined by examining the changes of the topology of the trees according to a particular region. This is illustrated by the analysis of the *gag* protein. The trees obtained with the MHR and the zinc knuckle regions are clearly different. This mainly involves the telomeric elements of *D. melanogaster* which are more closely related to the LTR retrotransposons and retroviruses based on the MHR region, and more closely related to the LINEs-like elements according to the CCHC region.

The modular evolution of TEs is also the order of the 4 domains of the *pol* gene in the LTR retrotransposons. This order is different in the members

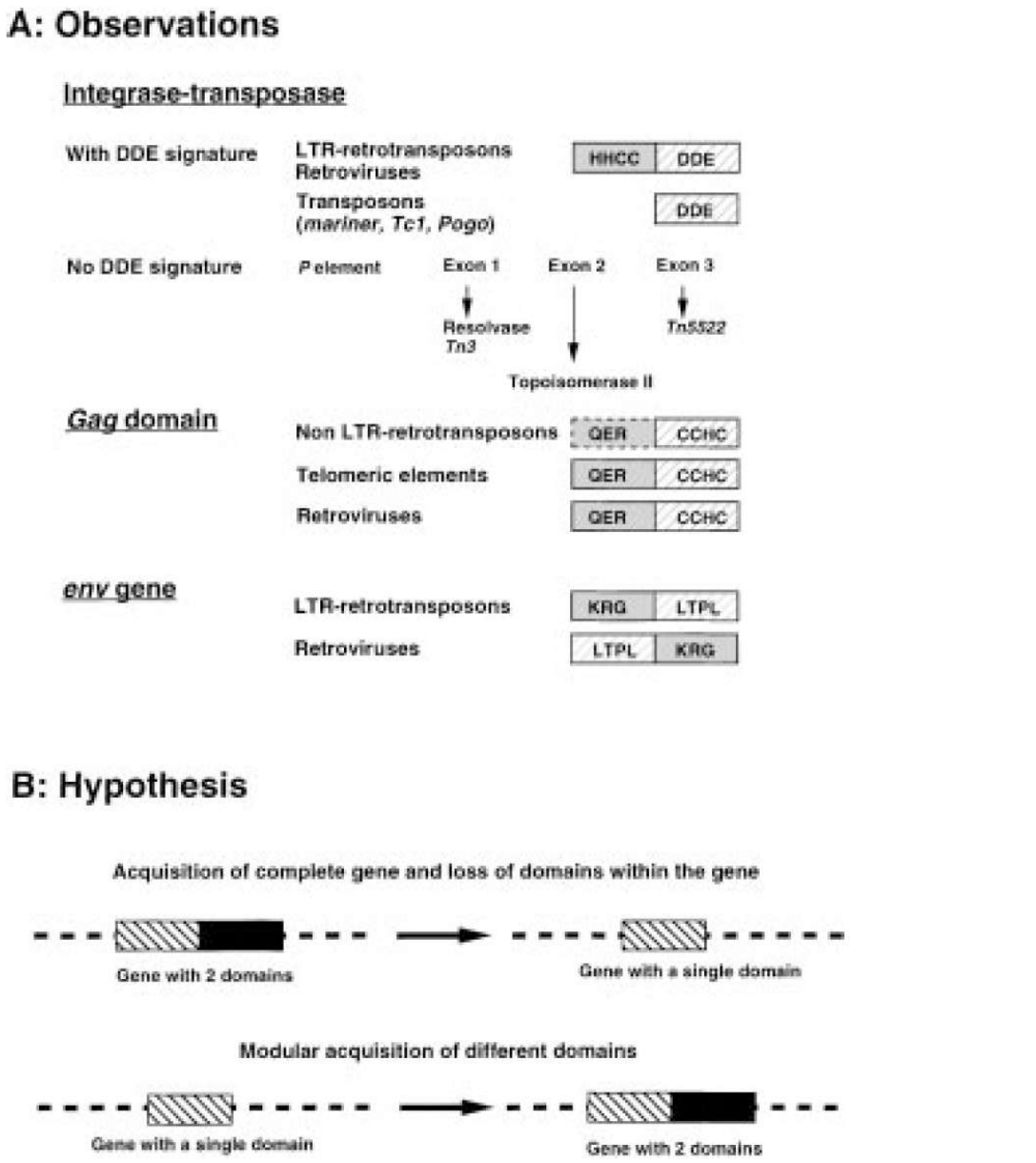


Figure 7. (A) Summary of the structural similarities and differences between the main classes of transposable elements, and (B) their possible origins.

of the *Ty1-copia* super-family (protease, integrase, reverse transcriptase, RNaseH) than it is in those of the *Ty3*, *Tf1* and *Gypsy* super-families (protease, reverse transcriptase, RNaseH, integrase) (see Capy et al., 1997 for the classification of the elements). The non-LTR retrotransposons have *gag* and reverse transcriptase domains, but they do not have the integrase domain of the LTR retrotransposons and retroviruses. The mechanism of transposition of these elements is clearly different from that of LTR retrotransposons and retroviruses (Luan et al., 1993), and an endonuclease domain was detected in non-LTR retrotransposons (Feng et al., 1996).

There are three possible explanations for the presence/absence of some domains. First is an initial acquisition of an ORF with several domains. This can be followed by the degeneration of some of them, due to the absence of selective pressures to maintain the initial function of the domain. Second is a modular acquisition of different domains. This could explain why the order of the domains in the *pol* genes are not all the same in the various super-families of LTR retro-

transposons. Third, the initial acquisition of an ORF with several domains could be followed by recombination, as suspected by McClure (1991). Of course, these three possibilities are not mutually exclusive.

The general models of TE evolution, as proposed by Xiong and Eickbush (1990), McClure (1991 and 1993) and Capy et al. (1996, 1997a and b), all suggest that there has been evolution from the most simple elements to the most complex ones. Several results, however, suggest that there may have been local 'toing and froing'. For instance, it was assumed that retroviruses emerged from LTR retrotransposons. However, a retrovirus which has lost its functional *env* gene can be viewed as an LTR retrotransposon, while a retrotransposon which has acquired a functional *env* gene and becomes infectious can be considered to be a retrovirus. It is also surprising that all class II elements with a DDE signature do not have any introns. This could be an argument in favor of their retrotransposon origin. Their transposases could be derived from retroelement integrases. The only class II elements with one or more introns are the members of the *P* and *hAT* super-families, which probably have different origins as suggested by the HCA analysis.

While these concluding remarks are highly speculative, their main objective is to stress that we must be prudent in our interpretation of the data on the relationships between homologous domains of the main classes of elements. All the domains used to build a TE can be found in prokaryotes. But this does not allow us to conclude that TEs were elaborated in prokaryotes. As long as the roots of the trees of each domain remain unknown it will be difficult to propose an evolutionary direction from our models.

Acknowledgements

This work was supported by the Ministère de la Recherche (ACC.SV3 - network #8) and a CNRS grant (Genome # 23). The English text was reviewed by Dr. Owen Parkes.

References

Altschul, S.F., T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller & D.J. Lipman, 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25: 3389–3402.

Auge-Gouillou, C., Y. Bigot, N. Pollet, M.H. Hamelin, M. Meunier-Rotival & G. Periquet, 1995. Human and other mammalian genomes contain transposons of the *mariner* family. FEBS Lett. 368: 541–546.

Bailey, T.L. & C. Elkan, 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers, pp. 28–36 in Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology. AAAI Press, Menlo Park, California.

Bigot, Y., C. Augé-Gouillou & G. Periquet, 1996. Computer analyses reveal a *hobo*-like element in the nematode *Caenorhabditis elegans*, which presents a conserved transposase domain common with the *Tc1-mariner* transposon family. Gene 174: 265–271.

Black, D.M., M.S. Jackson, M.G. Kidwell & G.A. Dover, 1987. *KP* elements repress hybrid dysgenesis in *Drosophila melanogaster*. EMBO J. 6: 4125.

Brunet, F., F. Godin, C. Bazin & P. Capy, 1999. Phylogenetic analysis of *Mos1*-like transposable elements in the *Drosophilidae*. J. Mol. Evol. 49: 760–768.

Brunet, F., F. Godin, C. Bazin, J.R. David & P. Capy, 1996. The *mariner* transposable element in natural populations of *Drosophila teissieri*. J. Mol. Evol. 42: 669–675.

Calvi, B.R., T.J. Hong, S.D. Findley & W.M. Gelbart, 1991. Evidence for a common evolutionary origin of inverted repeat transposons in *Drosophila* and plants: *hobo*, *Activator*, and *Tam3*. Cell 66: 465–471.

Capy, P., C. Bazin, D. Higuet & T. Langin, 1997a. Dynamic and Evolution of Transposable Elements. R.G. Landes Company, Austin, Texas, USA.

Capy, P., C. Bazin, D. Higuet & T. Langin, 1997b. Evolution and Impact of Transposable Elements. Kluwer Academic Publishers, Dordrecht.

Capy, P., T. Langin, Y. Bigot, F. Brunet, M.J. Daboussi, G. Periquet, J.R. David & D.L. Hartl, 1994. Horizontal transmission *versus* ancient origin: *mariner* in the witness box. Genetica 93: 161–170.

Capy, P., T. Langin, D. Higuet, P. Maurer & C. Bazin, 1997c. Does the integrase of LTR-retrotransposons and most of the transposases of class II elements share a common ancestor? Genetica 100: 63–72.

Capy, P., R. Vitalis, T. Langin, D. Higuet & C. Bazin, 1996. Relationships between transposable elements based upon the integrase-transposase domains: is there a common ancestor ? J. Mol. Evol. 42: 359–369.

Chaboissier, M.C., A. Bucheton & D.J. Finnegan, 1998. Copy number control of a transposable element, the I factor, a *LINE*-like element in *Drosophila*. Proc. Natl. Acad. Sci. USA 95: 11781–11785.

Covey, S.N., 1986. Amino acid sequence homology in *gag* region of reverse transcribing elements and the coat protein gene of cauliflower mosaic virus. Nucleic Acids Res. 14: 623–633.

Craven, R.C., A.E.L.-d. Pree, R.A.W. JR & J.W. Wills, 1995. Genetic analysis of the Major Homology Region for the Rous Sarcoma Virus *gag* protein. J. Virol. 69: 4213–4227.

Dayhoff, M.O., R.M. Schwartz & B.C. Orcutt, 1978. A model of evolutionary change in proteins, pp. 345–352 in Atlas of Protein Sequence and Structure, edited by M. O. Dayhoff. Natl. Biomed. Res. Found., Washington, DC.

Doak, T.G., F.P. Doerder, C.L. Jahn & G. Herrick, 1994. A proposed superfamily of transposase-related genes: new members in transposon-like elements of cilliated protozoa and a common 'D35E' motif. Proc. Natl. Acad. Sci. USA 91: 942–946.

Doolittle, W.F. & C. Sapienza, 1980. Selfish genes, the phenotype paradigm and genome evolution. Nature 284: 601–603.

Fayet, O., P. Ramond, P. Polard, M.F. Frère & M. Chandler, 1990. Functional similarities between retroviruses and the *IS3* family of bacterial insertion sequences ? Mol. Microbiol. 4: 1771–1777.

Felsenstein, J., 1993. PHYLIP (Phylogeny Inference Package). Version 3.5.c University of Washington, Seattle.

Feng, Q., J.V. Moran, H.J. Kazazian & J.D. Boeke, 1996. Human *L1* retrotransposon encodes a conserved endonuclease required for retrotransposition. Cell 87: 905–916.

Finnegan, D.J., 1989. Eukaryotic transposable elements and genome evolution. Trends Genet. 5: 103–107.

Gaboriaud, C., V. Bissery, T. Benchetrit & J.P. Mornon, 1987. Hydrophobic cluster analysis: an efficient new way to compare and analyse amino acid sequences. Febs Lett 224: 149–155.

Garcia-Fernàndez, J., G. Marfany, J. Bagunà & E. Salò, 1993. Infiltration of *mariner* elements. Nature 364: 109–110.

George, D.G., L.T. Hunt & W.C. Barker, 1988. Current methods in sequence comparison and analysis, pp. 127–149 in Macromolecular Sequencing and Synthesis, edited by D. H. Schlesinger. A.R. Liss, New York.

Gilbert, D.G., 1998. SeqPup: a biosequence editor. Version 0.8c. Distributed by the author at seqpup@bio.indiana.edu.

Grenier, E., M. Abadon, F. Brunet, P. Capy & P. Abad, 1999. A *mariner*-like transposable element in the entomopathogenic nematode *Heterorhabdis bacteriophora*, horizontal trasmission *versus* ancient origin. J. Mol. Evol. 48: 328–336.

Jacobson, J.W., M.M. Medhora & D.L. Hartl, 1986. Molecular structure of a somatically unstable element in *Drosophila*. Proc. Natl. Acad. Sci. USA 83: 8684–8688.

Jarvik, T. & K.G. Lark, 1998. Characterization of *Soymar1*, a *mariner* element in soybean. Genetics 149: 1569–1574.

Jensen, S., M.P. Gassama & T. Heidmann, 1999. Taming of transposable elements by homology-dependent gene silencing. Nat. Genet. 21: 209–212.

Jordan, I.K. & J.F. McDonald, 1999a. Phylogenetic perspective reveals abundant *Ty1/Ty2* hybrid elements in the *Saccharomyces cerevisiae* genome [letter]. Mol. Biol. Evol. 16: 419–422.

Jordan, I.K. & J.F. McDonald, 1999b. Comparative genomics and evolutionary dynamics of *Saccharomyces cerevisiae* Ty elements. Genetica 107: 3–13.

Labrador, M. & A. Fontdevila, 1994. High transposition rates of *Osvaldo*, a new *Drosophila buzzatii* retrotransposon. Mol. Gen. Genet. 245: 661–674.

Laten, H.M., A. Majumdar & E.A. Gaucher, 1998. *SIRE-1*, a *copia/Ty1* retroelement from soybean, encodes a retroviral envelope-like protein. Proc. Natl. Acad. Sci. USA 95: 6897–6902.

Lemesle-Varloot, L., B. Henrissat, C. Gaboriaud, V. Bissery, A. Morgat & J.P. Mornon, 1990. Hydrophobic cluster analysis: procedures to derive structural and functional information from 2-D-representation of protein sequences. Biochimie 72: 555–574.

Lerat, E. & P. Capy, 1999. Retrotransposons and retroviruses: analysis of the envelope gene. Mol. Biol. Evol. 16: 1198–1207.

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

Mammano, F., A. Öhagen, S. Höglund & H. Göttlinger, 1994. Role of the Major Homology Region of Human Immunodeficiency Virus type 1 in virion morphogenesis. J. Virol. 68: 4927–4936.

Maruyama, K. & D.L. Hartl, 1991. Evolution of the transposable element *mariner* in *Drosophila* species. Genetics 128: 319–329.

McClure, M., 1993. Evolutionary history of reverse transcriptase, pp. 425–444 in Reverse transcriptase, edited by M. Skalka, and S. P. Goff. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.

McClure, M.A., 1991. Evolution of retroposons by acquisition or deletion of retrovirus-like genes. Mol. Biol. Evol. 8: 835–856.

Morgan, G.T., 1995. Identification in the human genome of mobile elements spread by DNA-mediated transposition. J. Mol. Biol. 17: 1–5.

Okazaki, S., H. Ishikawa & H. Fujiwara, 1995. Structural analysis of TRAS1, a novel family of telomeric repeat- associated retrotransposons in the silkworm, Bombyx mori. Mol. Cell. Biol. 15: 4545–4552.

Oosumi, T., W.R. Belknap & B. Garlick, 1995. *Mariner* transposons in humans. Nature 378: 672–672.

Orgel, L.E. & F.H.C. Crick, 1980. Selfish DNA: the ultimate parasite. Nature 284: 604–607.

Pardue, M.-L., O.N. Danilevskaya, K.L. Traverse & K. Lowenhaupt, 1997. Evolutionary links between telomeres and transposable elements. Genetica 100: 73–84.

Petrov, D.A., E.R. Lozovskaya & D.L. Hartl, 1996. High intrinsic rate of DNA loss in *Drosophila*. Nature 384: 346–349.

Reiter, L.T., T. Murakami, T. Koeuth, L. Pentao, D.M. Muzny, R.A. Gibbs & J.R. Lupski, 1996. A recombination hotspot responsible for two inherited peripheral neuropathies is located near a *mariner* transposon-like element. Nature Genet. 12: 288–297.

Robertson, H.M., 1997. Multiple *mariner* transposons in flatworms and hydras are related to those of insects. J. Heredity 88: 195–201.

Robertson, H.M. & E.G. MacLeod, 1993. Five major subfamilies of *mariner* transposable elements in insects, including the Mediterranean fruit fly, and related arthropods. Insect Mol. Biol. 2: 125–139.

Robertson, H.M. & R. Martos, 1997. Molecular evolution of the second ancient human mariner transposon, *Hsmar2*, illustrates patterns of neutral evolution in the human genome lineage. Gene 205: 219–228.

Robertson, H.M. & K.L. Zumpano, 1997. Molecular evolution of an ancient *mariner* transposon, *Hsmar1*, in the human genome. Gene 205: 203–217.

Robertson, H.M., Z.L. Zumpano, A.R. Lohe & D.L. Hartl, 1996. Reconstruction of the ancient *mariners* of humans. Nature Genet. 12: 360–361.

Rubin, E. & A.A. Levy, 1997. Abortive gap repair: underlying mechanism for *Ds* element formation. Mol. Cell Biol. 17: 6294–6302.

Sedensky, M.M., S.J. Hudson, B. Everson & P.G. Morgan, 1994. Identification of a *mariner*-like repetitive sequence in *C. elegans*. Nucleic Acids Res. 22: 1719–1723.

Streck, R.D., J.E. MacGaffey & S.K. Beckendorf, 1986. The structure of *hobo* transposable elements and their insertion. EMBO J. 5: 3615–3623.

Strimmer, K. & A. vonHaeseler, 1996. Quartet puzzling: a quartet maximum likelihood method for reconstructing tree topologies. Mol. Biol. Evol. 13: 964–969.

Swofford, D.L., 1993. Phylogenetic analysis using parsimony. Version 3.1.1. Smithsonian Institution Washington DC.

Swofford, D.L., G.J. Olsen, P.J. Waddel & D.M. Hillis, 1996. Phylogenetic inference, pp. 407–514 in Molecular Systematics, edited by D. M. Hillis, Moritz and Mable. Sinauer.

Wiley, L.J., L.G. Riley, N.C. Sangster & A.S. Weiss, 1997. *mle-1*, a *mariner*-like transposable element in the nematode *Trichostrongylus colubriformis*. Gene 188: 235–237.

Xiong, Y. & T.H. Eickbush, 1990. Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J. 9: 3353–3362.

Article 3 : Codon usage and the origin of *P* element

Mol. Biol. Evol. 17(3):467-468, 2000

Letter to the Editor

Codon Usage and the Origin of *P* Elements

Emmanuelle Lerat,† Christian Biémont,† and Pierre Capy**

*Laboratoire Populations, Génétique et Evolution, Centre National de la Recherche Scientifique, Gif/Yvette, France; and †Laboratoire Biométrie et Biologie Évolutive, Centre National de la Recherche Scientifique, Université de Lyon I, Villeurbanne, France

The authors of a recent comparison of the *P* transposable element and three genes of *Drosophila melanogaster* and *Drosophila willistoni* suggested that the codon usage of the *D. melanogaster P* element is similar to that of *D. willistoni* genes (Powell and Gleason 1996). They concluded that this could be further evidence of the recent horizontal transfer of the *P* element from *D. willistoni* to *D. melanogaster* indicated by several previous findings (Clark and Kidwell 1997). More specifically, it was shown that *D. willistoni* genes tend to be T-ending-codon genes, whereas those of *D. melanogaster* tend to be C-ending-codon genes. The transposase genes of the *P* elements from both species was found to be AT-ending-codon genes, and one explanation for this may be that the *P* element of *D. melanogaster* originated from *D. willistoni.* This hypothesis assumes that the codon usage in transposable elements (TEs) and that in the host genome are similar. However, analysis of a large number of genes and TEs in *D. melanogaster* suggests that the T-ending-codon feature of the *P* element could be a general characteristic of all TEs in *Drosophila* species and independent of the host genome (Shields and Sharp 1989).

We extracted from the GenBank DNA sequence database the sequences of six genes common to *D. melanogaster* and *D. willistoni*: *Alcohol dehydrogenase* (*Adh,* M11290 and L08648), *Amylase*-related gene (*amyrel,* U69607 and AF039560), *superoxide dismutase* (*SOD,* Y00367 and L13281), *xanthine dehydrogenase* (*Xdh,* Y00307 and AF058985), *glycerol 3 phosphate dehydrogenase* (*Gpdh,* X80204 and L37038), and *period* (*per,* M30114 and U51055). The sequences of the last three genes have been only partially determined for *D. willistoni.* We therefore used only the homologous regions in both species to avoid bias due to differences in gene length (Moryiama and Powell 1998; Duret and Mouchiroud 1999). *P* elements, which have been reported in distantly related *Drosophila* species, were also added. These elements were described in *Drosophila bifasciata* (Hagemann, Miller, and Pinkser 1992), *Drosophila subobscura* (Paricio et al. 1991), and *Scaptomyza pallida* (Simonelig and Anxolabéhère 1991). All sequences of RNA (class I) and DNA (class II) elements described in the species previously mentioned were also used to compare *P* element features with those of other elements.

Key words: *P* element, transposable elements, *Drosophila,* codon usage, horizontal transfer.

Address for correspondence and reprints: Pierre Capy, Laboratoire Populations, Génétique et Evolution, Centre National de la Recherche Scientifique, UPR 9034, Bât. 13, 91198 Gif/Yvette Cedex, France. E-mail: capy@pge.cnrs-gif.fr.

Mol. Biol. Evol. 17(3):467–468. 2000

The relative codon frequencies were estimated for all sense codons (59 codons) for each gene and transposable element according to the formula

$$F_j = \frac{n_{ij}}{\sum_{j=1}^{s_i} n_{ij}},$$

where n_{ij} is the number of codon j observed for the amino acid i, and s_i is the number of synonymous codons for the the amino acid i. The 59 columns of the matrix are the variables of a factorial correspondence analysis (FCA). Because relative and absolute codon frequencies can be sensitive to several biases (Perrière and Thioulouze, personal communication), FCA was also performed using absolute codon frequencies and relative synonymous codon usage (RSCU), frequently used for such analyses (see, e.g., Shields and Sharp 1989). In all cases, the FCA gave similar topologies on the first two axes. Figure 1 shows a factor map crossing the first two axes when the relative frequencies given above are used.

The first two axes accounted for 44% of the total variance (34% on the first axis and 10% on the second). The percentage of the variance explained by the remaining axes was very low, close to 1% for the third and fourth axes, respectively, and $<1\%$ for the other ones. *Drosophila melanogaster* and *D. willistoni* genes are clearly separated. The projection of the codons shows that GC-ending codons (gray ellipses) are more frequent in the *D. melanogaster* genes than in those of *D. willistoni,* which display several T-ending codons. The TEs (black spots) of different *Drosophila* species clearly display a higher frequency of AT-ending codons. Codon usage patterns of *P* elements from different species are similar and clearly differ from those of the host genes. Moreover, the codon usage variability among genes from different species is lower than that between genes and TEs within the same species. A MANOVA (Statistica, version 3.0b, StatSoft) using the coordinates of each point on the first two axes as variables shows that the difference observed between genes of *D. melanogaster* and *D. willistoni* is significant, with a *P* value of 7.34×10^{-4}, while the difference between transposable elements and genes of *D. melanogaster* is significant, with a *P* value of $<10^{-7}$.

The characteristics of the *P* element in *D. willistoni* described by Powell and Gleason (1996) could be a general feature of TEs in *Drosophila,* and not attributable to the host. This was suggested by a previous analysis of *D. melanogaster* (Shields and Sharp 1989) and is confirmed by similarities in *P* elements from different hosts. Moreover, these elements and other DNA and RNA el-

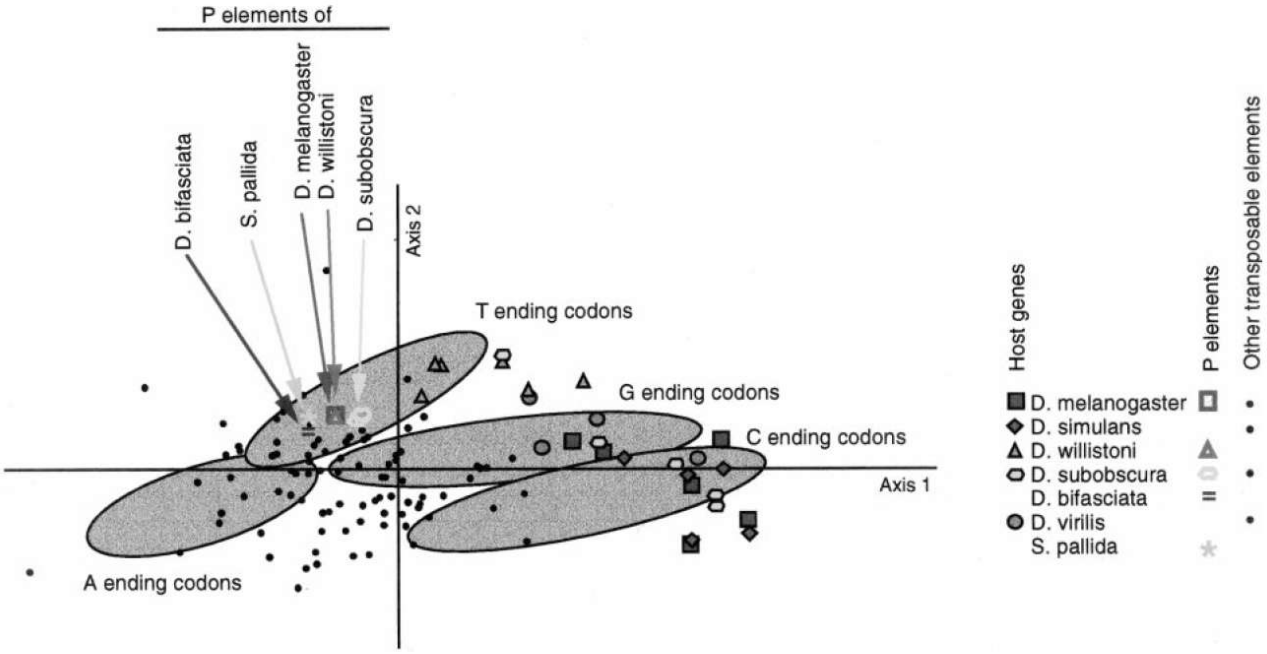


FIG. 1.—Projection on the first two axes on an FCA. The accession numbers of the *P* element sequences are X60990 for *jbifM3* of *Drosophila bifasciata* (Hagemann, Miller, and Pinkser 1992), S74793 for *A1* and *A2* of *Drosophila subobscura* (Paricio et al. 1991), and M63341 and M63342 for *PS2* and *PS18* of *Scaptomyza pallida* (Simonelig and Anxolabéhère 1991), respectively. Black spots represent the other DNA and RNA transposable elements described for different *Drosophila* species. Ellipses indicate the projections of the codons grouped according to their third bases.

ements described for several *Drosophila* species group together in our analysis (black spots in fig. 1).

These findings strongly suggest that codon usage cannot be employed to demonstrate horizontal transfers of TEs between *Drosophila* species. TEs and host genes may not be subject to the same constraints, but the possibility that the evolutions of these two entities are linked cannot be ruled out. For instance, the most frequent codon for the host gene could be the least frequent for TEs and vice versa. The only way to check this will be to analyze species using different codon usage strategies.

LITERATURE CITED

CLARK, J. B., and M. G. KIDWELL. 1997. A phylogenetic perspective on *P* transposable element evolution in *Drosophila*. Proc. Natl. Acad. Sci. USA **94**:11428–11433.

DURET, L., and D. MOUCHIROUD. 1999. Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis, Drosophila,* and *Arabidopsis.*. Proc. Natl. Acad. Sci. USA **96**:4482–4487.

HAGEMANN, S., W. J. MILLER, and W. PINSKER. 1992. Identification of a complete *P* element in the genome of *Drosophila bifasciata*. Nucleic Acids Res. **20**:409–413.

MORIYAMA, E. N., and J. R. POWELL. 1998. Gene length and codon usage bias in *Drosophila melanogaster, Saccharomyces cerevisiae* and *Escherichia coli*. Nucleic Acids Res. **26**:3188–3193.

PARICIO, N., A. M. PEREZ-ALONSO, M. J. MARTINEZ-SEBASTIAN, and R. FRUTOS. 1991. *P* sequences of *Drosophila subobscura* lack exon 3 and may encode a 66 kd repressor-like protein. Nucleic Acids Res. **19**:6713–6718.

POWELL, J. R., and J. M. GLEASON. 1996. Codon usage and the origin of *P* elements. Mol. Biol. Evol. **13**:278–279.

SHIELDS, D. C., and P. M. SHARP. 1989. Evidence that mutation patterns vary among *Drosophila* transposable elements. J. Mol. Biol. **207**:843–846.

SIMONELIG, M., and D. ANXOLABÉHÈRE. 1991. A *P* element of *Scaptomyza pallida* is active in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA **88**:6102–6106.

MANOLO GOUY, reviewing editor

Accepted November 18, 1999

Article 4 : Codon usage by transposable elements and their host genes in five species

J. Mol. Evol. sous presse

J Mol Evol (2002) 54:625–637
DOI: 10.1007/s00239-001-0059-0

JOURNAL OF MOLECULAR EVOLUTION

Codon Usage by Transposable Elements and Their Host Genes in Five Species

Emmanuelle Lerat,[1] Pierre Capy,[2] Christian Biémont[1]

[1] Laboratoire Biométrie et Biologie Évolutive, UMR CNRS 5558, Université Lyon 1, 69622 Villeurbanne Cedex, France
[2] Laboratoire Populations, Génétique et Évolution, UPR CNRS 9034, 91198 Gif/Yvette Cedex, France

Received: 2 May 2001 / Accepted: 29 October 2001

Abstract. We compared the codon usage of sequences of transposable elements (TEs) with that of host genes from the species *Drosophila melanogaster, Arabidopsis thaliana, Caenorhabditis elegans, Saccharomyces cerevisiae,* and *Homo sapiens.* Factorial correspondence analysis showed that, regardless of the base composition of the genome, the TEs differed from the genes of their host species by their AT-richness. In all species, the percentage of A + T on the third codon position of the TEs was higher than that on the first codon position and lower than that in the noncoding DNA of the genomes. This indicates that the codon choice is not simply the outcome of mutational bias but is also subject to selection constraints. A tendency toward higher A + T on the third position than on the first position was also found in the host genes of *A. thaliana, C. elegans,* and *S. cerevisiae* but not in those of *D. melanogaster* and *H. sapiens.* This strongly suggests that the AT choice is a host-independent characteristic common to all TEs. The codon usage of TEs generally appeared to be different from the mean of the host genes. In the AT-rich genomes of *Arabidopsis thaliana, Caenorhabditis elegans,* and *Saccharomyces cerevisiae,* the codon usage bias of TEs was similar to that of weakly expressed genes. In the GC-rich genome of *D. melanogaster,* however, the bias in codon usage of the TEs clearly differed from that of weakly expressed genes. These findings suggest that selection acts on TEs and that TEs may display specific behavior within the host genomes.

Key words: Codon usage — Transposable elements — Retrotransposons — Transposons

Correspondence to: C. Biémont; *email:* biemont@biomserv.univ-lyon1.fr

Introduction

Transposable elements (TEs) are repeated sequences found in all living organisms. They are inserted into the host genomes and can move from one position to another along the chromosomes. They are ancient components of the genome and were long considered to be just parasitic DNA but have now been shown to act as agents of genome restructuring and to have an impact on evolution. They have even led to the formation of novel host genes by molecular domestication and exon shuffling (Shapiro 1999; Tomilin 1999; Kidwell and Lisch 2001). TEs can be divided into two main classes on the basis of their structural organization and the intermediate component used for their transposition (Capy et al. 1997). Class I includes the retrotransposons, which move via an RNA intermediate and encode at least the reverse transcriptase necessary for their transposition. These retrotransposons are subdivided into two subclasses: the LTR retrotransposons, which possess long terminal repeats on their extremities, and the non-LTR retrotransposons (also called retroposons), which have a poly(A) tail. Class II includes the transposons, which code for a transposase and use a DNA intermediate to transpose. *Ac* in maize and *P* in *Drosophila* are typical examples of this class. MITEs

elements, which are abundant in some organisms and have no coding sequences (Wessler et al. 1995), seem to have originated from deleted class II elements (Feschotte and Mouchès 2000). TEs depend on the host machinery for their transcription and transposition, and so they must have been subjected to the constraints affecting the genome. Powell and Gleason (1996) therefore suggested that their codon usage is likely to be similar to that of the host genes.

The degeneracy of the genetic code means that most amino acids are encoded by more than one codon, and there is often a bias in the use of synonymous codons, with some codons being preferred (Grantham et al. 1980). This bias may result from the natural selection of codons on the basis of the availability of tRNA (the preferred codons correspond to the most abundant tRNA species) in a way that optimizes translation efficiency or from the nonrandom choice between codons ending in pyrimidines (Gouy and Gautier 1982). In *Drosophila melanogaster,* the bias of codon usage is correlated to tRNA availability and varies according to the function of the genes during development (Moriyama and Powell 1997). In this species, highly expressed genes display a stronger codon bias than less strongly expressed genes (Shields et al. 1988). In some species, codon usage bias is lower in genes that encode long proteins, suggesting that the process of differential selection acting on codon usage depends on the gene length (Duret and Mouchiroud 1999). An alternative but not necessarily incompatible explanation is that codon usage bias is attributable to mutational bias, and this would account for the extreme base composition of some bacterial genomes (Wright and Bibb 1992) and for the codon usage in less actively expressed genes (Sharp and Matassi 1994). Biased codon usage can therefore tell us about the interaction between the genome and its TEs. Any discrepancy in codon usage by the TEs and by the host genes could provide a way of detecting horizontal transfers between hosts with differing codon biases or could suggest that the TEs themselves have specific characteristics within the genome. However, the tendency of TEs to have more AT-ending codons than their host genes in *Drosophila* (Shields and Sharp 1989) shows that a difference in codon usage cannot be taken as conclusive evidence of horizontal transfers (Lerat et al. 1999). We have little information about TE codon usage in species other than *Drosophila,* except in human retroviruses (Kypr and Mrázek 1987). However, the molecular mechanism responsible for the AT-biased codon usage in human retroviruses has not yet been elucidated (Berkhout and Van Hemert 1994).

We compared the TE codon usage in *Arabidopsis thaliana, Caenorhabditis elegans, Saccharomyces cerevisiae, Drosophila melanogaster,* and *Homo sapiens* with that of host genes to obtain a wider view of the codon usage of TEs. Multivariate analysis of sequences of host genes and TEs is able to discriminate between these two kinds of DNA sequences with regard to codon usage, with a tendency for TEs to be AT-rich and to have a specific codon usage bias which differs from the average codon usage of the host genes, regardless of the base composition of the genomes. TEs therefore appear to display specific codon usage which reflects their specific behavior within the genome.

Materials and Methods

Data

TE sequences were obtained from the GenBank database using the ACNUC retrieval system (Gouy et al. 1985). The CDS of the *gag, pol,* and *env* genes of the retrotransposons were considered separately. The set of TE sequences (LTR retrotransposons, non-LTR retrotransposons, class II and class III elements) consisted of 65 sequences for *A. thaliana,* 33 for *C. elegans,* 32 for *S. cerevisiae,* 55 for *D. melanogaster,* and 18 for *H. sapiens,* for a total of 198 TE sequences.

Base composition was estimated for the TE sequences and for the genes identified in the complete genomes of *S. cerevisiae* (6301 genes), *C. elegans* (14,425 genes), and *D. melanogaster* (14,332 genes) and from chromosomes 1, 2, and 4 of *A. thaliana* (the coding sequences of chromosomes 3 and 5 not yet being available) (13,859 genes). The coding sequences of the complete genome of *H. sapiens* is not yet available, so we retrieved coding sequences from GenBank (12,227 genes). Any sequences with internal stop codons were eliminated. We also included retrovirus sequences from *H. sapiens,* which are class I elements and can be considered to be LTR retrotransposons. Thus, we obtained from GenBank 74 coding sequences from *HIV1, HIV2, HERV-K, HERV-H19, HERV3, HTLV, HSRV,* and one v-oncogene.

To calculate the average relative frequency of the codons, we retrieved well-defined host genes from GenBank, i.e., genes with a known function rather than being obtained by gene prediction. We selected 1560 genes for *A. thaliana,* 592 for *C. elegans,* 1569 for *S. cerevisiae,* 300 for *D. melanogaster,* and 300 for *H. sapiens.* The list of accession numbers of these genes is available upon request to the corresponding author.

Computation of Codon Frequencies

The three stop codons were not taken into account, because they appear only once in a coding sequence. Nor did we consider the ATG and TGG codons, which encode methionine and tryptophane and, so, are not degenerated. As a result, we used only 59 degenerated codons of the 64 existing. The relative frequency of each synonymous codon was computed as follows. If a given amino acid appears n times in a DNA sequence and is encoded by two codons, 1 and 2, appearing a and b times, respectively, then the relative frequency of the first codon is a/n and that of the second codon is b/n. This calculation avoids bias due to differences in the amino acid composition of the sequences (Chiapello et al. 1998).

Factorial Correspondence Analysis (FCA)

FCA is a multivariate analysis (Hirschfeld 1935) which is often used to analyze codon usage (Grantham et al. 1981; Shields and Sharp 1989). It gives a graphical representation of the best simultaneous representation of the two groups formed by the rows and columns of the data matrix. FCA calculates the position of the sequences in a multidimensional space according to codon usage. This method can detect differ-

Table 1. AT composition at each position of codons of transposable elements and host genes[a]

	Number of sequences	First position	Second position	Third position	Total
LTR retrotransposons					
A. thaliana	41	52.00	58.50	57.60	56.03
C. elegans[b]	6	51.07	63.58	57.01	56.83
S. cerevisiae	32	56.77	63.44	66.23	62.16
D. melanogaster	31	53.99	65.35	60.01	59.73
H. sapiens[c]	74	47.85	56.38	57.52	53.90
Non-LTR retrotransposons					
A. thaliana	22	55.41	61.13	60.45	59.00
C. elegans	13	58.31	61.64	58.87	59.60
S. cerevisiae[d]	/	/	/	/	/
D. melanogaster	16	51.31	57.41	53.66	54.13
H. sapiens	5	59.06	65.54	56.90	60.50
Transposons					
A. thaliana	2	53.67	63.46	68.88	62.00
C. elegans	10	52.51	61.32	58.32	57.38
S. cerevisiae[d]	/	/	/	/	/
D. melanogaster	5	56.97	65.94	62.58	61.83
H. sapiens[e]	12	52.03	60.96	57.30	53.41
Nuclear genes					
A. thaliana	13,859	49.72	59.76	57.33	55.60
C. elegans	14,425	50.90	61.41	59.74	57.35
S. cerevisiae	6,301	55.39	63.00	60.63	59.67
D. melanogaster	14,332	44.10	58.49	34.35	45.65
H. sapiens	12,227	44.01	57.46	42.89	48.12

[a] The values for class III elements from *D. melanogaster* (three sequences) and *C. elegans* (four sequences) are not represented.
[b] In addition to *Cerl* (AC: U15406), we included the elements *Cer4*, *Cer5*, *Cer5–1*, *Cer7*, and *Cer9* among the 15 sequences discovered by Bowen and McDonald (1999). Ten of the sequences were not used because of the strong degeneracy of their coding sequences.
[c] Proviruses sequences of *Homo sapiens: HIV1*, *HIV2*, *HERV-K*, *HERV-H19*, *HERV3*, *HTLV*, *HSRV*, and one v-oncogene.
[d] The genome of *S. cerevisiae* passes only LTR retrotransposons (Kim et al. 1998).
[e] Without *Tramp* element (see Discussion).

ences in codon usage between sequences and identify the codons involved. Sequences in which a given codon is used in a similar fashion lie close to each other on the graph. We used the ADE-4 software package (Thioulouse et al. 1997) to perform the FCA of our data. There were many more host genes than TEs, and so we randomly selected 43 genes per species, so as to avoid statistical bias in the multivariate analysis (see below). In this way, we obtained a total of 215 host genes (43 genes per species) and 203 TE sequences. The relative frequencies of codons were arranged in a matrix consisting of 418 rows (215 genes + 203 TE sequences) and 59 columns (59 synonymous codons). The random selection of genes was repeated 10 times. As the results obtained by the multivariate analysis were similar for all randomizations, we only used the data from the first randomization. To estimate the validity of the grouping of sequences obtained by FCA, we estimated the associated p values by multivariate analysis of variance done directly on the coordinates of the points on the FCA graph. This analysis was performed with Statview (version 5.0; SAS Institute Inc.).

Expression Data

Among the well-characterized host genes, we identified the weakly and highly expressed genes for each species, except for *H. sapiens,* in which there is no relationship between the level of gene expressivity and the codon usage bias (Shields et al. 1988). For *D. melanogaster, A. thaliana,* and *C. elegans,* gene expressivity was determined as described by Duret and Mouchiroud (1999). The gene sequences were compared to EST (expressed sequence tag) sequences with BLASTN (Altschul et al. 1997), and the level of expressivity of a given gene was estimated from the number of ESTs that matched the sequence of this gene. For *S. cerevisiae,* for which too few ESTs were available, we estimated the effective number of codons (N_c) (Wright 1990) for each gene sequence. The value found ranged from 20 (high codon bias) to 61 (no bias). According to Sharp and Cowe (1991), the sequences in which N_c was less than 30 and those in which N_c was greater than 55 were considered to correspond to highly and weakly expressed genes, respectively. This was justified because in *S. cerevisiae,* the codon bias is correlated with the gene expression level, with highly biased genes being highly expressed (Sharp et al. 1986).

Results

Base Composition of TEs and Host Genes

Table 1 shows the percentages of AT for the first, second, and third codon positions of coding sequences of TEs (*gag, pol,* and *env* for retrotransposons and transposase for class II elements), according to their family and host species, and of the host genes according to species. The values of the global percentage of AT for nuclear genes were in agreement with the values obtained from the sequenced genomes in *D. melanogaster* (Adams et al. 2000), *C. elegans* (*C. elegans* Sequencing Consortium 2000), and *S. cerevisiae* (Bowman et al. 1997; Bussey et al. 1997; Churcher et al. 1997; Dietrich et al. 1997; Dujon et al. 1997; Jacq et al. 1997; Johnston et al. 1997; Philippsen et al. 1997; Tettelin et al. 1997), from chromosomes 21 and 22 of *H. sapiens* (Dunham et al. 1999; Hattori et al. 2000), and chromosomes 2 and 4 of *A. thaliana* (Lin et al. 1999; Mayer et al. 1999). Table

Table 2. Average relative frequency of the 59 degenerated codons for highly and weakly expressed host genes and transposable elements (TEs), according to species[a]

		At			*Ce*			*Sc*			*Dm*		
		Genes			Genes			Genes			Genes		
		High	Weak	TEs	High	Weak	TEs	High	Weak	TEs	High	Weak	TEs
		92[b]	221[b]	65[b]	80[b]	112[b]	29[b]	66[b]	227[b]	32[b]	74[b]	65[b]	52[b]
Amino acids	Codons	76 (22–258)[c]	0[c]	/	86 (41–271)[c]	0[c]	/	/	/	/	72 (33–367)[c]	0[c]	/
K	AAA	0.364	0.490	**0.539**	0.324	**0.639**	**0.636**	0.147	**0.557**	**0.711**	0.231	0.282	**0.685**
K	AAG	**0.636***	**0.510**	0.461	**0.676***	0.361	0.364	**0.853***	0.443	0.289	**0.769***	**0.718**	0.315
N	AAT	0.323	**0.512**	**0.569**	0.498	**0.619**	**0.567**	0.085	**0.560**	**0.566**	0.339	0.476	**0.534**
N	AAC	**0.677***	0.488	0.431	**0.502***	0.381	0.433	**0.915***	0.440	0.434	**0.661***	**0.524**	0.466
I	ATA	0.087	0.251	0.239	0.055	0.167	0.244	0.005	0.326	0.349	0.110	0.208	0.370
I	ATT	0.392	**0.402**	**0.408**	0.431	**0.558**	**0.416**	0.473	**0.406**	**0.381**	0.362	0.331	**0.385**
I	ATC	**0.521***	0.347	0.353	**0.514***	0.275	0.340	**0.522***	0.268	0.270	**0.528***	**0.461**	0.245
T	ACA	0.209	0.305	**0.337**	0.237	**0.386**	**0.327**	0.016	**0.305**	**0.387**	0.145	0.191	**0.381**
T	ACT	**0.369**	**0.323**	0.313	0.307	0.296	0.298	**0.534***	0.286	0.282	0.192	0.145	0.257
T	ACC	0.331*	0.209	0.212	**0.326***	0.153	0.207	**0.449***	0.211	0.220	**0.463***	**0.384**	0.237
T	ACG	0.091	0.163	0.138	0.130	0.165	0.168	0.001	0.198	0.111	0.200	0.280	0.125
R	AGA	**0.315**	**0.360**	**0.346**	0.270	**0.281**	**0.314**	**0.854***	**0.335**	**0.476**	0.046	0.097	**0.334**
R	AGG	0.242*	0.203	0.182	0.031	0.086	0.112	0.009	0.235	0.104	0.080	0.121	0.174
R	CGA	0.062	0.115	0.131	0.151	0.251	0.172	0.002	0.117	0.145	0.115	0.163	0.174
R	CGT	0.267*	0.172	0.140	**0.338***	0.199	0.184	0.130	0.143	0.177	0.241*	0.150	0.113
R	CGC	0.071	0.061	0.102	0.161*	0.078	0.125	0.005	0.090	0.071	**0.417***	**0.312**	0.128
R	CGG	0.043	0.089	0.100	0.049	0.105	0.093	0.000	0.080	0.027	0.101	0.157	0.077
S	AGT	0.124	0.162	0.143	0.101	0.159	0.124	0.029	0.166	0.171	0.101	0.151	0.167
S	AGC	0.152*	0.134	0.109	0.097	0.096	0.106	0.023	0.138	0.071	0.192	0.244	0.189
S	TCA	0.174	0.206	0.240	0.211	**0.285**	0.229	0.033	0.203	**0.284**	0.083	0.087	**0.218**
S	TCT	**0.280**	**0.269**	**0.279**	**0.216***	0.192	**0.241**	**0.560***	**0.210**	0.234	0.119	0.065	0.164
S	TCC	0.176*	0.117	0.122	0.189*	0.120	0.179	0.355*	0.158	0.154	**0.292***	**0.238**	0.155
S	TCG	0.094	0.112	0.105	0.186*	0.148	0.121	0.000	0.125	0.086	0.213*	0.215	0.107
Y	TAT	0.280	**0.523**	**0.594**	0.381	**0.587**	**0.528**	0.079	**0.527**	**0.553**	0.296	0.383	0.483
Y	TAC	**0.720***	0.477	0.406	**0.619***	0.413	0.472	**0.921***	0.473	0.447	**0.704***	**0.617**	**0.517**
L	TTA	0.059	0.132	0.163	0.051	0.152	0.161	0.160	0.222	**0.322**	0.050	0.041	**0.214**
L	TTG	0.217	0.226	0.205	0.201	0.220	0.189	**0.762***	**0.253**	0.137	0.174	0.184	0.157
L	CTA	0.080	0.107	0.133	0.042	0.103	0.137	0.060	0.148	0.186	0.074	0.100	0.186
L	CTT	**0.287**	**0.256**	**0.209**	0.296*	**0.226**	**0.192**	0.017	0.133	0.174	0.100	0.085	0.178
L	CTC	**0.281***	0.178	0.163	**0.306***	0.161	0.178	0.000	0.083	0.093	0.148*	0.157	0.129
L	CTG	0.076	0.101	0.127	0.104	0.138	0.143	0.001	0.161	0.089	**0.454***	**0.433**	0.136
F	TTT	0.358	**0.527**	**0.554**	0.253	**0.522**	0.486	0.173	**0.559**	**0.582**	0.280	0.350	0.577
F	TTC	**0.642***	0.473	0.446	**0.747***	0.478	**0.514**	**0.827***	0.441	0.418	**0.720***	**0.650**	0.423
C	TGT	**0.556**	**0.597**	**0.575**	0.435	**0.568**	**0.536**	**0.944***	**0.557**	**0.521**	0.262	0.275	0.437
C	TGC	0.444*	0.403	0.425	**0.565***	0.432	0.464	0.056	0.443	0.479	**0.738***	**0.725**	**0.563**
Q	CAA	0.452	**0.577**	**0.603**	**0.607**	**0.655**	**0.689**	**0.984***	**0.597**	**0.738**	0.273	0.296	**0.640**
Q	CAG	**0.548***	0.423	0.397	0.393*	0.345	0.311	0.016	0.403	0.262	**0.727***	**0.704**	0.360
H	CAT	0.403	**0.632**	**0.655**	**0.525**	**0.637**	**0.574**	0.248	**0.600**	**0.593**	0.373	0.409	0.492
H	CAC	**0.597***	0.368	0.345	0.475*	0.363	0.426	**0.752***	0.400	0.407	**0.627***	**0.591**	**0.508**
P	CCA	**0.373**	0.335	**0.401**	**0.681***	**0.513**	**0.360**	**0.901***	**0.328**	**0.429**	0.227	0.234	**0.388**
P	CCT	0.345	**0.381**	0.265	0.103	0.192	0.263	0.094	0.290	0.301	0.126	0.112	0.222
P	CCC	0.131*	0.105	0.142	0.055	0.091	0.192	0.004	0.208	0.131	**0.391***	**0.315**	0.225
P	CCG	0.151	0.179	0.192	0.161	0.204	0.185	0.001	0.174	0.139	0.256	0.339	0.116
E	GAA	0.396	**0.514**	**0.604**	0.494	**0.661**	**0.680**	**0.977***	**0.618**	**0.783**	0.285	0.296	**0.656**
E	GAG	**0.604***	0.486	0.396	**0.506***	0.339	0.320	0.023	0.382	0.217	**0.715***	**0.704**	0.344

Table 2. Continued

		At			*Ce*			*Sc*			*Dm*		
		Genes			Genes			Genes			Genes		
		High	Weak	TEs	High	Weak	TEs	High	Weak	TEs	High	Weak	TEs
		92[b]	221[b]	65[b]	80[b]	112[b]	29[b]	66[b]	227[b]	32[b]	74[b]	65[b]	52[b]
Amino acids	Codons	76 (22–258)[c]	0[c]	/	86 (41–271)[c]	0[c]	/	/	/	/	72 (33–367)[c]	0[c]	/
D	GAT	**0.539**	**0.667**	**0.657**	**0.608**	**0.680**	**0.584**	0.408	**0.603**	**0.575**	0.491	**0.537**	0.464
D	GAC	0.461*	0.333	0.343	0.392*	0.320	0.416	**0.592***	0.397	0.425	**0.509***	0.463	**0.536**
V	GTA	0.081	0.150	0.176	0.900	0.188	0.237	0.003	0.236	0.305	0.089	0.098	0.272
V	GTT	**0.390**	**0.407**	**0.341**	0.369	**0.371**	**0.283**	0.572	**0.295**	**0.343**	0.186	0.098	**0.282**
V	GTC	0.271*	0.180	0.224	**0.374***	0.204	0.231	0.415*	0.217	0.228	0.262*	0.238	0.227
V	GTC	0.258	0.263	0.260	0.167	0.237	0.249	0.010	0.252	0.123	**0.463***	**0.486**	0.219
A	GCA	0.184	0.265	0.320	0.183	**0.367**	0.341	0.018	0.286	0.369	0.118	0.162	**0.337**
A	GCT	**0.483***	**0.422**	**0.387**	**0.398***	0.315	**0.335**	**0.738***	**0.302**	**0.381**	0.237	0.175	0.273
A	GCC	0.235*	0.176	0.186	0.341*	0.176	0.195	0.243	0.247	0.180	**0.520***	**0.470**	0.261
A	GCG	0.098	0.137	0.106	0.078	0.142	0.129	0.001	0.165	0.069	0.125	0.193	0.129
G	GGA	0.382	**0.363**	**0.364**	**0.733***	**0.555**	**0.429**	0.012	0.253	0.347	0.246	0.274	**0.365**
G	GGT	**0.416***	0.333	0.319	0.148	0.223	0.256	**0.950***	**0.316**	**0.440**	0.247	0.217	0.237
G	GGC	0.114*	0.142	0.166	0.081	0.133	0.169	0.036	0.262	0.161	**0.460***	**0.434**	0.257
G	GGG	0.088	0.162	0.151	0.038	0.089	0.146	0.002	0.169	0.052	0.047	0.075	0.141

[a]The highest frequency of codons for each amino acid is in boldface. *At, A. thaliana; Ce, C. elegans; Sc, S. cerevisiae; Dm, D. melanogaster.* Asterisks indicate optimal codons as found by Sharp and Crowe (1991), Sharp and Lloyd (1993), Chiapello et al. (1998), and Duret and Mouchiroud (1999).

[b]Number of sequences used for the analysis.

[c]mRNA abundance × 10^5, with the range in parentheses.

1 reveals a preference for AT in TE and retrovirus coding sequences, regardless of the AT-richness of the host species. χ^2 tests were done for each species to determine whether the observed differences between host genes and TEs in the global percentage of AT (%AT) and percentage of AT (%AT) in each codon position were statistically significant. Overall, the %AT in the TEs of the AT-rich genomes of *A. thaliana, S. cerevisiae,* and *C. elegans* was as high as in the TEs of the GC-rich genomes of *D. melanogaster* and *H. sapiens,* but the difference between the AT-richness of TEs and genes was greater in the last two species. Only the transposons in *C. elegans* had a global %AT that was not significantly different from that of the host genes. The %AT at the first position was lower than the %AT at the third position for both TEs and host genes from *A. thaliana, C. elegans,* and *S. cerevisiae.* The host genes of *H. sapiens* and *D. melanogaster* showed the opposite tendency, as did the non-LTR retrotransposons of *H. sapiens,* but the number of elements analyzed was small (five sequences).

Table 2 shows the average frequency of the 59 synonymous codons for the highly and weakly expressed genes and the TEs for each species. We have compared these data with the published preferred codons for the highly and weakly expressed genes in *S. cerevisiae* (Sharp and Cowe 1991), *C. elegans* (Stenico et al. 1994; Duret and Mouchiroud 1999), *D. melanogaster* (Sharp and Lloyd 1993; Duret and Mouchiroud 1999), and *A. thaliana* (Chiapello et al. 1998; Duret and Mouchiroud 1999). The codons preferentially used in the TEs also seem to be those preferentially used in the weakly expressed genes in *Arabidopsis, Saccharomyces,* and *Caenorhabditis.* However, TEs in *Drosophila* did not follow this pattern. In this species, as the data in Table 2 show, there was no difference between the highly and the weakly expressed genes with regard to the bias of codon usage, but these genes did differ from TEs with regard to their preferred codons. Note that we did not observe any bias between highly and weakly expressed genes in *D. melanogaster,* because only genes coding for short proteins have been shown to display such bias (Duret and Mouchiroud 1999). Since our sample consisted of a mixture of genes coding for proteins of different sizes, it was impossible to detect any bias between these two groups of genes. However, in the data of Duret and Mouchiroud (1999) the preferred codons in weakly expressed genes coding for short proteins were quite different from those used by the TEs in our study.

Codon Choice of TEs and Host Genes

Figure 1 shows the projection of TE and host gene sequences from all species on the plane determined by the

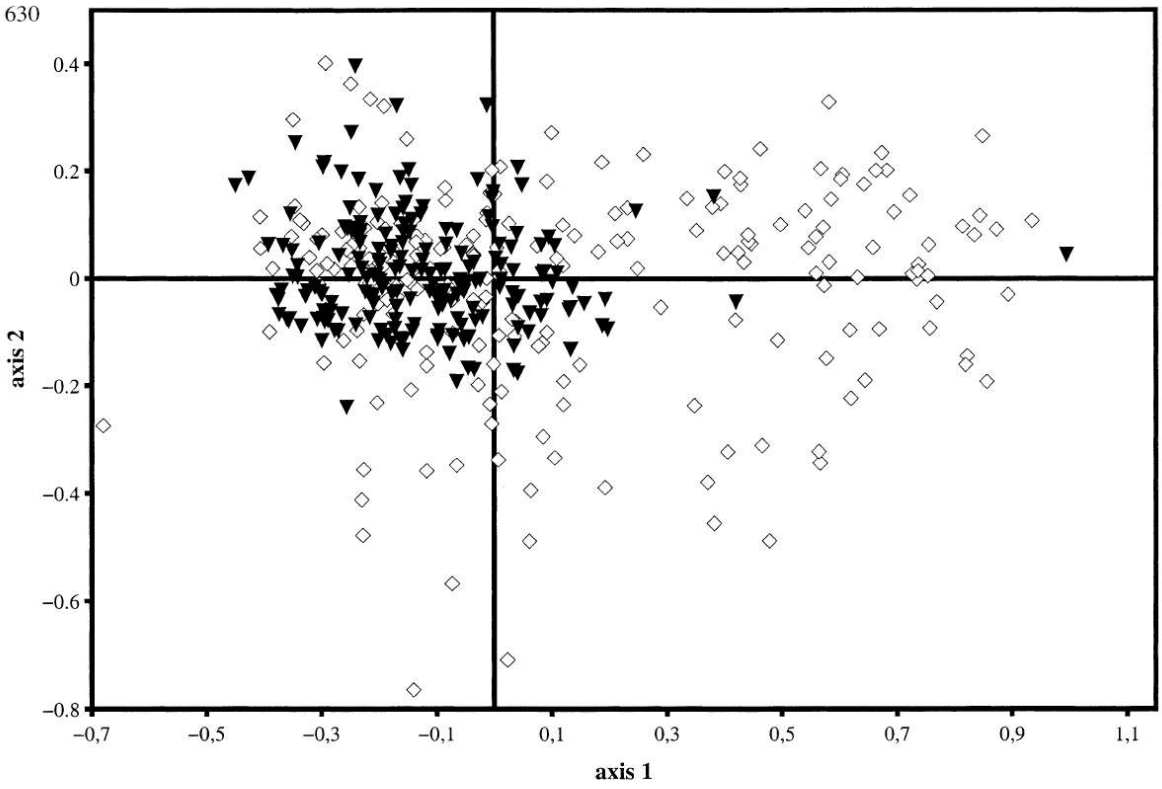


Fig. 1. Projection of the host genes (white diamonds) and transposable elements (black triangles) on the first two axes of FCA, which represent 27.7 and 6.4% of the total variance, respectively.

first two axes of the FCA done on the relative frequencies of the codons for one gene randomization run (see Materials and Methods). These axes represented 27.7 and 6.4% of the total variance of the analysis, respectively. As a group, TEs differed significantly from the group formed by the host genes ($p < 0.0001$ in multivariate analysis of variance). Four sequences seemed to be excluded from the TE group: *Tramp* (accession number Y17156) from *H. sapiens,* two ORFs corresponding to *gag* and reverse transcriptase genes of *R1* (accession number X51968), and the *env* gene of *nomad* (accession number AF039416) from *D. melanogaster.* Projecting the codons onto the same plane (Fig. 2) showed that the A-ending and T-ending codons were grouped together and could be distinguished from two other groups, the C-ending codons and the G-ending codons. The position on the plane of the A- and T-ending codons indicates that these codons were more frequent in TEs than in host genes for all species, with the exception of a few genes for each species. Analysis of the contributions of the sequences on the first axis showed that the genes of *D. melanogaster* and *H. sapiens* were the main contributors to the distribution of the sequences over the first axis. Since most of the CDS of these two genomes were GC-rich (Table 1), the discrimination on the first axis resulted mainly from differences in the overall GC composition of the genomes and the AT-richness of the TEs. This difference explains why many host gene sequences for *S. cerevisiae, A. thaliana,* and *C. elegans,* which are AT-rich, appeared to fall within the TE group.

To minimize the impact of the higher GC content of *H. sapiens* and *D. melanogaster* genes on the distribution of TEs and genes apparent in Fig. 1, in Fig. 3 we show the projection of the TE and gene sequences on axes 2 and 3 of the FCA (the third axis accounted for 6.0% of the total variance). Once again, the grouping of the TEs and host genes is significantly different ($p < 0.0001$ in multivariate analysis of variance), and this is independent of the high GC content of the host genes. Five TE sequences appeared to be slightly separated from the rest of the TE group, but this was attributable to the short length (200 to 500 bp) of these sequences. As shown in Fig. 4, the projection of the codons onto the plane shown in Fig. 3 suggests that the distribution of the genes over the second axis depended on their use of G-ending and C-ending codons and that the distribution of TEs over the third axis depended on their use of A-ending and T-ending codons. This analysis identified two groups of TEs, one rich in A-ending codons, and the other rich in T-ending codons, corresponding mainly to the *A. thaliana* elements.

Codon Choice in Each Species

To see whether the codons chosen were different in TEs and genes, we identified the TE and host gene sequences

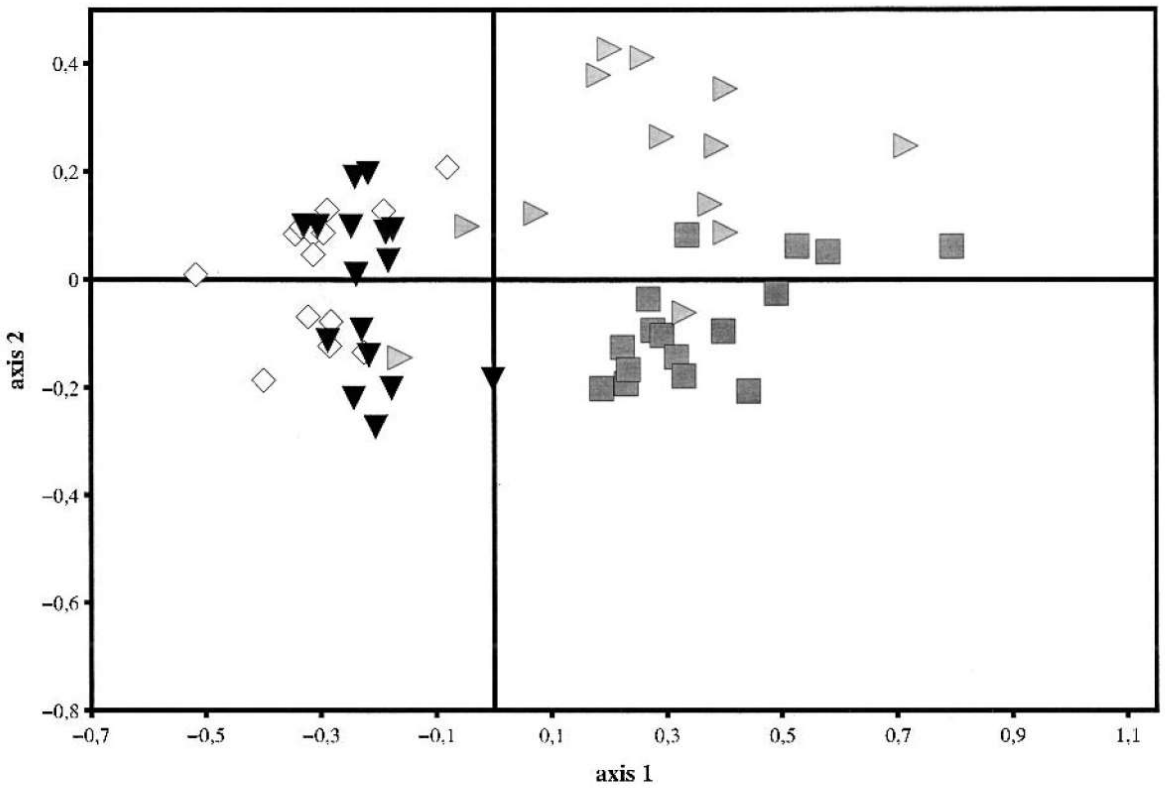


Fig. 2. Projection of the 59 degenerated codons on the first two axes of the FCA (see Fig. 1). White diamonds correspond to A-ending codons, black triangles to T-ending codons, gray squares to C-ending codons, and gray triangles to G-ending codons.

according to species in the cloud of points shown in Fig. 3. The species-specific groups were represented by ellipses that enclosed 90% of the points (Fig. 5). A multivariate analysis of variance showed that all groups were significantly different ($p < 0.0001$), indicating that codon usage in TEs and genes was distinct in each species. The centers of the ellipses of TEs for all species were on the same side of the plane and were displaced from the centers of the ellipses of the host genes. *D. melanogaster* showed the greatest separation between the ellipse centers of TEs and those of the host genes. Moreover, the TE ellipses were all oriented perpendicularly to the host gene ellipses, suggesting that the codons had a differential influence on the distribution of genes and TEs.

An FCA of the TE sequences confirmed the sequence grouping according to species for TEs but also revealed that the codon usage bias of TEs depended on the superfamily to which they belong (multivariate analysis of variance: $p < 0.0001$) (data not shown). Hence, the choice of codon differed in the superfamilies *gypsy/Ty3, Tyl/copia* (LTR retrotransposons), LINEs (non-LTR retrotransposons), *mariner-Tcl, hAT, P* (transposons), and *foldback* elements. In species such as *S. cerevisiae* and *H. sapiens,* most of the TEs belonged to just one or two superfamilies. For example, *S. cerevisiae* contained only TEs from the *Tyl/copia* superfamily (Kim et al. 1998). This means that the groups detected by the FCA could be attributable to the fact that most of the TEs found in one species were from a given superfamily.

Discussion

The transposable elements of the *Drosophila melanogaster, Arabidopsis thaliana, Caenorhabditis elegans, Saccharomyces cerevisiae,* and *Homo sapiens* genomes appear to be characterized by their global AT-richness, regardless of the base composition of their host genome. This confirms the AT-richness characteristic of retroviruses, especially the lentivirus group (Zsìros et al. 1999) and of the few TEs studied in *D. melanogaster,* sea urchin, and *Bombyx mori* (Shields and Sharp 1989; Springer et al. 1995) and also shows that it is a more general characteristic. The difference between TEs and host genes is especially marked in the *D. melanogaster* and *H. sapiens* genomes, which have predominantly GC-rich genes. Codon usage by the TEs in *A. thaliana, C. elegans,* and *S. cerevisiae* also differs from the average codon usage of the genes but is less biased. Moreover, the %AT on the first codon position is lower than that on the third codon position for all the TEs, which suggests that this is a specific characteristic of all TEs and independent of the host genes. TEs from *Arabidopsis* are distinguished by their preference for T-ending codons, suggesting that this is specific characteristic of this species.

The *R1* element and the *env* gene of the *nomad* retrotransposon of *D. melanogaster* and the *Tramp* element of *H. sapiens* were the only TE sequences we analyzed that did not fall into the major cluster of TEs shown in

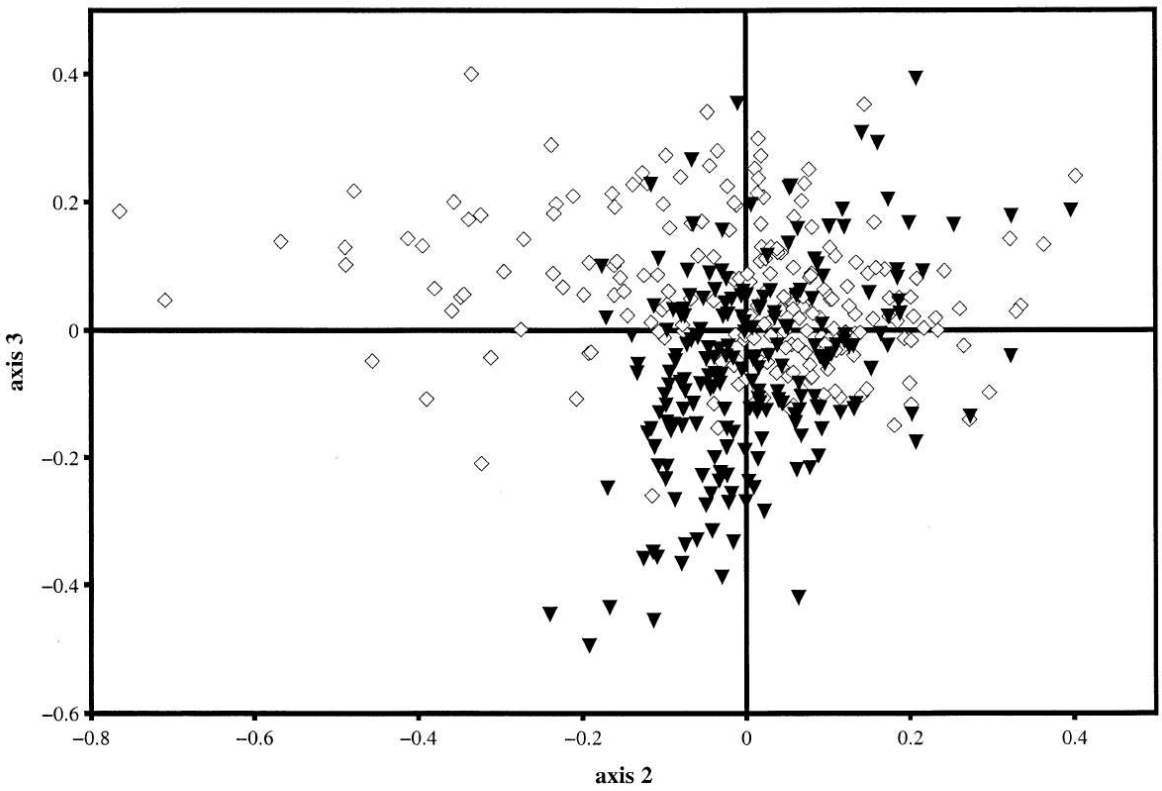


Fig. 3. Projection of the host genes (white squares) and transposable elements (black triangles) on axes 2 and 3 of FCA, which represent 6.4 and 6.0% of the total variance, respectively.

Fig. 1. The *R1* element is inserted in the 28S ribosomal RNA genes and is believed to have been stable in the *D. melanogaster* genome for several million years (Jakubczak et al. 1990). The nucleotide composition at the third codon position of the sequences of this element show excess GC, as in the host genes. According to Jakubczak et al. (1990), the presence of *R1* could be beneficial for rDNA amplification in the host. It is thus possible that the host genome views the *R1–28S* rDNA as a "host genomic entity." The *Tramp* element, which is inserted in the PAR region of the X and Y chromosomes (Esposito et al. 1999), is unable to move and could be involved in biological functions because its protein is homologous to DREF, a promoter-activating factor for *Drosophila* DNA replication-related genes. We detected no similarity between the sequence of the *env* gene of the *nomad* element and any host *Drosophila* genes. We can offer no simple explanation for why this TE sequence is GC-rich, despite the fact that its two other ORFs, *gag* and *pol,* are AT-rich and are located with the other TE sequences. We cannot rule out the possibility that the *env* gene of *nomad* has a specific function in *Drosophila,* perhaps like the envelope gene of the human endogenous retrovirus (*HERV-W*), which is expressed in placenta (Blond et al. 2000).

Codon Usage and Reverse Transcriptase Errors

Bias of codon usage is interpreted in terms of mutational bias or of natural selection acting on silent changes in DNA sequences (for a review see Sharp and Matassi 1994). Mutational bias is thus considered to have played a major role in shaping retroviral genomes and to be a driving force in their evolution (Bronson and Anderson 1994; Zsìros et al. 1999). The preference for G-to-A and C-to-T transitions in some lentiviruses (Zsìros et al. 1999) could reflect the action of the enzyme reverse transcriptase, which is involved in virus retrotranscription, and is known to be error-prone. Other molecular forces that it has been proposed could be responsible for the nucleotide bias in virus include imbalances in intracellular dCTP concentration (Vartanian et al. 1994) and aminoacyl-tRNA availability in host cells (Van Hemert and Berkhout 1995). The possibility of selection favoring particular features of the integrated provirus DNA genome or of the viral DNA itself (Zsìros et al. 1999) has also been proposed.

The mechanisms underlying the codon bias observed in viruses could account for the A-rich nucleotide bias of LTR and non-LTR retrotransposons, which, like retroviruses, move via an RNA intermediate. However, SINE transposable elements (short interspersed elements), which are non-LTR retrotransposons without ORFs, are GC-rich (Schmid 1998). To achieve their own transposition, these elements need reverse transcriptase, which is provided in *trans* by LINE elements (Wichman et al. 1992; Jurka 1997), so they should also reflect the errors made by the enzyme, leading to an overincorporation of A and T bases. The high GC content of the SINE ele-

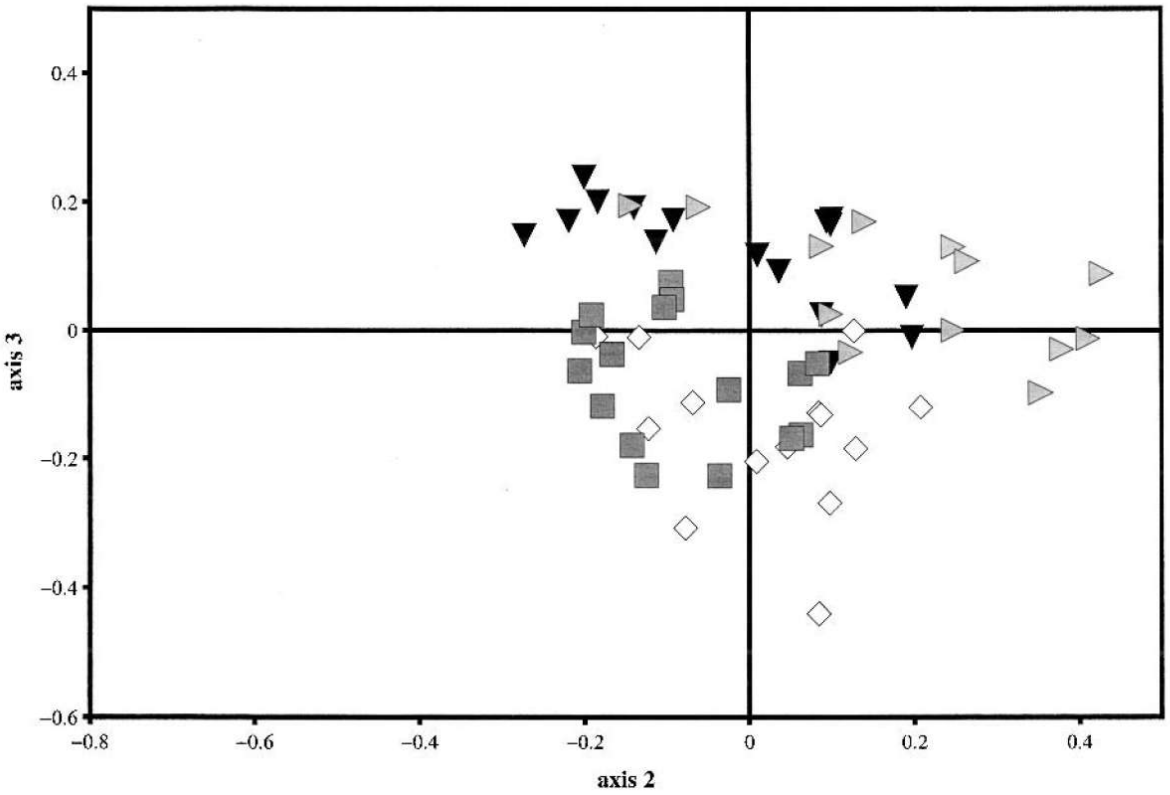


Fig. 4. Projection of the 59 degenerated codons on axes 2 and 3 of the FCA (see Fig. 3). White diamonds correspond to A-ending codons, black triangles to T-ending codons, gray squares to C-ending codons, and gray triangles to G-ending codons.

ments and the LINE-like *R1* element of *Drosophila* does not, therefore, suggest that reverse transcriptase plays any major, general role in generating the specific codon usage of LTR and non-LTR retrotransposons. In addition, the AT-richness of transposons, which are DNA elements involving a transposase for their transposition, means that the bias in codon usage of TEs could be due to a common mechanism or to different mechanisms specific to each TE class.

In the absence of selection, a mutational bias could be expected to have shifted the base composition of the TEs toward that of the noncoding DNA of the genome, i.e., 65% AT in *D. melanogaster* (Shields et al. 1988; Kliman and Hey 1994), 68% in *A. thaliana,* 65% in *S. cerevisiae* (Lin et al. 1999), and 70% in *C. elegans* (Duret, personal communication). The higher AT values at the third position of the codons in the TEs therefore point instead to selective constraints acting on this third codon base, as in the genes *env* and *pol* of *HIV-1* (Yan et al. 2000).

TE Codon Usage and Genomic Environment

As discussed by Sharp and Matassi in their review (1994), codon usage in the mammalian genome could reflect the physical location of the genes, which in turn may simply reflect differences in mutation patterns. Many TEs seem to be inserted into AT-rich, late-replicating DNA regions (Le et al. 2000), and their insertion sites are often successions of A and T bases (Merriman et al. 1995). The base composition of the TEs could therefore mimic that of the region in which they are inserted. This has been found to be the case for retrotransposons *1731* and *17.6* of *D. melanogaster,* TRIP of sea urchin, and *Mag* of *Bombyx mori* (Springer et al. 1995) and certain retrovirus sequences (Bernardi et al. 1985). This parallel is also illustrated by the GC-rich SINE elements, which are located in GC-rich regions (Korenberg and Rykowski 1988; Boyle et al. 1990; Jurka 1997). In addition, there is a link between codon usage bias and regional base composition in the *Drosophila* genome (Kliman and Hey 1994; Jabbari and Bernadi 2000). However, an analysis of the insertion site preferences of *P* elements in *Drosophila* has shown that 500-bp regions around the elements are GC-rich (Liao et al. 2000). Hence, the base composition of the region in which the TEs are inserted is not in itself sufficient to account for the specific codon usage of TEs. Codon usage may therefore be associated with the capacity of TEs to have moved and still to be moving around the genome. This could be tested by using the reverse transcriptase sequences of group II introns (data not available for a reliable statistical analysis), which can move into the genome using a mechanism similar to that used by LINE elements (Grivell 1996).

Kliman and Hey (1993) have shown that in *D. mela-*

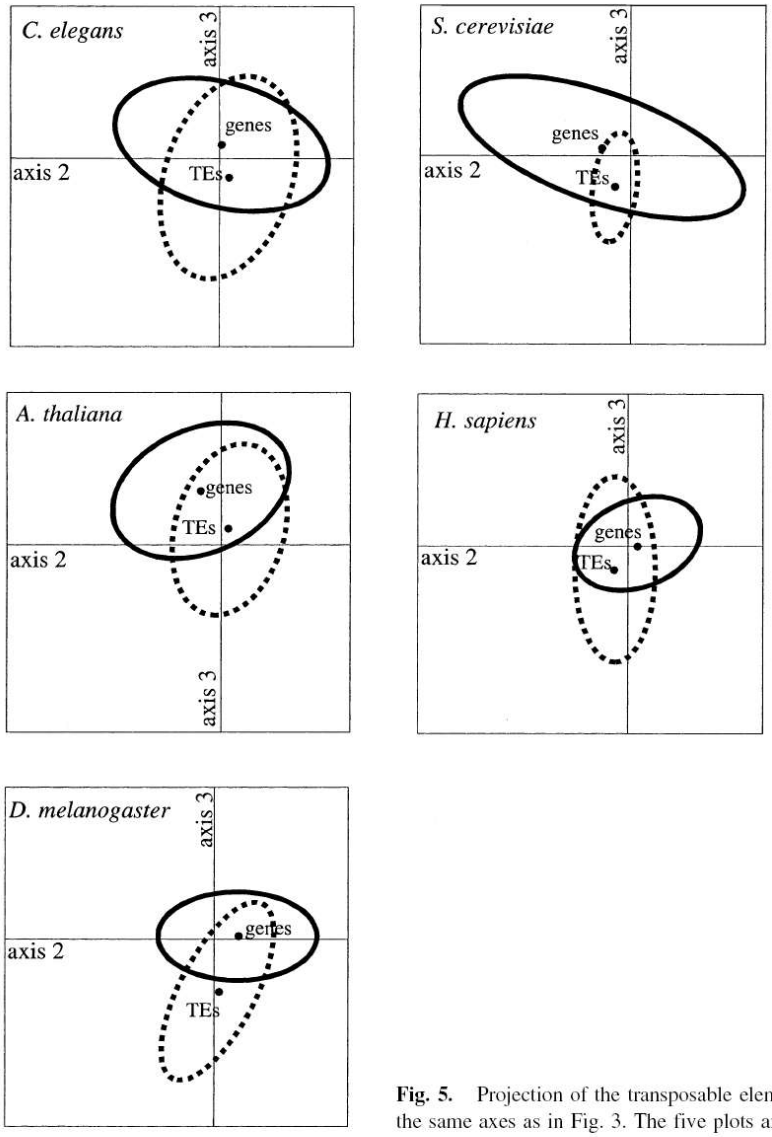


Fig. 5. Projection of the transposable elements and host genes according to their host species on the same axes as in Fig. 3. The five plots are superimposable. TEs, ---------; host genes, ——.

nogaster codon bias is significantly lower in regions where there is little or no recombination. This is because selection at any particular silent site can be expected to be less effective in regions of the genome with reduced recombination rates (The Hill-Robertson effect, Kliman and Hey 1993). There is no clear evidence, however, of a negative correlation between TE density and recombination rate along the chromosomes, either in *D. melanogaster* (Biémont et al. 1997) or in *C. elegans* (Duret et al. 2000). Thus it is proposed that TEs could be inserted preferentially within region of specific chromatin conformation, which determines codon usage, as in retroviruses (Gama Sosa et al. 1989).

Gene Silencing, mRNA Destruction, and Methylation

Methylation, chromatin-mediated silencing, homology-dependent gene silencing (cosuppression) by RNA interference are all responses of the host genomes to invasion by TEs (Yoder et al. 1997; McDonald 1998; Jensen et al. 1999; Ketting et al. 1999). Plant and mammal genomes regulate TEs by methylation of cytosine, hypomethylation being associated with active TEs (Bird 1997; Bender 1998; O'Neil et al. 1998). The AT-richness of the TEs could thus result from these inactivating processes or from a way of blocking these inactivation mechanisms so as to remain active. It has also been proposed that methylation is only a way of masking the effects of TEs and hiding them from the genome (Martienssen 1998). Cosuppression, especially by destruction of TE mRNAs, is therefore an attractive possibility as a process for limiting TE transposition in germlines. It remains to be determined whether the generation of dsRNA (double-stranded RNA) intervening to silence repetitive sequences by RNA interference is favored or disfavored by the AT-richness of these sequences.

Codon Usage, Gene Length, and Gene Expression

Codon bias is positively correlated with gene expression and negatively correlated with gene length in *Drosophila* and in the nematode *Caenorhabditis* (Moriyama and Powell 1998; Comeron et al. 1999; Duret and Mouchiroud 1999; Duret 2000). We have seen that TEs preferentially use the same preferred codons as weakly expressed genes in *Arabidopsis, Saccharomyces,* and *Caenorhabditis.* But this tendency does not appear in the *Drosophila* genome, suggesting that the codon usage bias of the TEs is not strongly associated with level of expressivity. The similar codon usage bias observed in TEs and in weakly expressed genes in *Arabidopsis, Saccharomyces,* and *Caenorhabditis* should therefore be due to the high A–T content of these genomes.

AT-biased genes have been reported to have distinct biological properties in *Arabidopsis* and in other plants species (Carels and Bernardi 2000). In *Arabidopsis* such genes, unlike the housekeeping genes, show patterns of expression that are tissue specific in many cases (Akashi 1997; Chiapello et al. 1998) and respond to various stressful conditions (dehydration, temperature, pathogens, heavy metal stress). Thus, the codon usage of the TEs may reflect their specific pattern of expression within the genome.

Acknowledgments, We would like to thank D. Chessel and L. Duret for their helpful comments, N.J. Bowen and J.F. McDonald for the gift of the *Cer* sequences, and M. Gosh for the English correction. This work was funded by the CNRS (Programme Génome, GDR 2157) and the Association pour la Recherche sur le Cancer (Contract 5428).

References

Adams MD, Celniker SE, Holt RA, et al. (2000) The genome sequence of *Drosophila melanogaster.* Science 287:2185–2195

Akashi H (1997) Codon bias evolution in Drosophila. Population genetics of mutation–selection drift. Gene 205:269–278

Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Bender J (1998) Cytosine methylation of repeated sequences in eukaryotes: The role of the DNA pairing. Trends Biochem Sci 23: 252–256

Berkhout B, Van Hemert FJ (1994) The unusual nucleotide content of the HIV RNA genome results in a biased amino acid composition of HIV proteins. Nucleic Acids Res 22:1705–1711

Bernardi G, Olofsson B, Filipski J, Zerial M, Salinas J, Cuny G, Meunier-Rotival M, Rodier F (1985) The mosaic genome of warm-blooded vertebrates. Science 228:953–958

Biémont C, Tsitrone A, Vieira C, Hoogland C (1997) Transposable element distribution in Drosophila. Genetics 147:1997–1999

Bird A (1997) Does DNA methylation control transposition of selfish elements in the germline? Trends Genet 13:469–470

Blond JL, Lavillette D, Cheynet V, Bouton O, Oriol G, Chapel-Fernandes S, Mandrand B, Mallet F, Cosset FL (2000) An envelope glycoprotein of the human endogenous retrovirus HERV-W is expressed in the human placenta and fuses cells expressing the type D mammalian retrovirus receptor. J Virol 74:3321–3329

Bowman S, Churcher C, Badcock K, et al. (22 co-authors) (1997) The nucleotide sequence of *Saccharomyces cerevisiae* chromosome XIII. Nature 387:90–93

Boyle AL, Ballard SG, Ward DC (1990) Differential distribution of long and short interspersed element sequences in the mouse genome—Chromosome karyotyping by fluorescence *in situ* hybridization. Proc Natl Acad Sci USA 87:7757–7761

Bowen NJ, MacDonald JF (1999) Genomic analysis of *Caenorhabditis elegans* reveals ancient families of retroviral-like elements. Genome Res 9:924–935

Bronson EC, Anderson JN (1994) Nucleotide composition as a driving force in the evolution of retroviruses. J Mol Evol 38:506–532

Bussey H, Storms RK, Ahmed A, et al. (1997) The nucleotide sequence of *Saccharomyces cerevisiae* chromosome XVI. Nature 387:103–105

Capy P, Bazin C, Higuet D, Langin T (1997) Dynamic and evolution of transposable elements. R.G. Landes, Austin, TX

Carels N, Bernardi G (2000) Two classes of genes in plants. Genetics 154:1819–1825

C. elegans Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans.* A platform for investigating biology. Science 282:2012–2018

Chiapello H, Lisacek F, Caboche M, Hénaut A (1998) Codon usage and gene function are related in sequences of *Arabidopsis thaliana.* Gene 209:GC1–GC38

Churcher C, Bowman S, Badcock K, et al. (1997) The nucleotide sequence of *Saccharomyces cerevisiae* chromosome IX. Nature 387:84–87

Comeron JM, Kreitman M, Aguadé M (1999) Natural selection on synonymous sites is correlated with gene length and recombination in Drosophila. Genetics 151:239–249

Dietrich FS, Mulligan J, Hennessy K, et al. (1997) The nucleotide sequence of *Saccharomyces cerevisiae* chromosome V. Nature 387: 78–81

Dong H, Nilsson L, Kurland CG (1996) Co-variation of tRNA abundance and codon usage in *Escherichia coli* at different growth rates. J Mol Biol 260:649–663

Dujon B, Albermann K, Aldea M, et al. (1997) The nucleotide sequence of *Saccharomyces cerevisiae* chromosome XV. Nature 387: 98–102

Dunham I, Shimizu N, Roe BA, et al. (1999) The DNA sequence of human chromosome 22. Nature 402:489–495

Duret L (2000) tRNA gene number and codon usage in the *C. elegans* genome are co-adapted for optimal translation of highly expressed genes. Trends Genet 16:287–289

Duret L, Mouchiroud D (1999) Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. Proc Natl Acad Sci USA 13:4482–4487

Duret L, Marais G, Biémont C (2000) Transposons but not retrotransposons are found preferentially in regions of high recombination rate in *Caenorhabditis elegans.* Genetics 156:1661–1669

Esposito T, Gianfrancesco F, Ciccodicola A, Montanini L, Mumm S, D'Urso M, Forabosco A (1999) A novel pseudoautosomal human gene encodes a putative protein similar to *Ac*-like transposases. Hum Mol Genet 8:61–67

Feschotte C, Mouchès C (2000) Evidence that a family of miniature inverted-repeat transposable elements (MITEs) from the *Arabidopsis thaliana* genome has arisen from a *pogo*-like DNA transposon. Mol Biol Evol 17:730–737

Gama Sosa MA, Hall JC, Schneider KE, Lukaszewicz GC, Ruprecht RM (1989) Unusual DNA structures at the integration site of an *HIV* provirus. Biochem Biophys Res Commun 161:134–142

Gouy M, Gautier C (1982) Codon usage in bacteria: Correlation with gene expressivity. Nucleic Acids Res 10:7055–7074

Gouy M, Gautier C, Attimonelli M, Lavane C, di Paola G (1985)

ACNUC—a portable retrieval system for nucleic acid sequence databases: Logical and physical designs and usage. Comp Appl Biosci 1:167–172

Grantham R, Gautier C, Gouy M, Mercier R, Pavé A (1980) Codon catalog usage and the genome hypothesis. Nucleic Acids Res 8: 49–62

Grantham R, Gautier C, Gouy M, Jacobzone M, Mercier R (1981) Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res 9:r43–r74

Grivell LA (1996) Transposition: Mobile introns get into line. Curr Biol 6:48–51

Hattori M, Fujiyama A, Taylor TD, et al. (2000) The DNA sequence of human chromosome 21. Nature 405:311–319

Hirschfeld HO (1935) A connection between correlation and contingency. Proc Cambr Philos Soc Math Phys Sci 31:520–524

Jabbari K, Bernardi G (2000) The distribution of genes in the Drosophila genome. Gene 247:287–292

Jacq C, Alt-Mörbe J, Andre B, et al. (1997) The nucleotide sequence of *Saccharomyces cerevisiae* chromosome IV. Nature 387:75–78

Jakubczak JL, Xiong Y, Eickbush TH (1990) Type I (R1) and type II (R2) Ribosomal DNA insertions of *Drosophila melanogaster* are retrotransposable elements closely related to those of *Bombyx mori.* J Mol Biol 212:37–52

Jensen S, Gassama MP, Heidmann T (1999) Cosuppression of *I* transposon activity in Drosophila by *I*-containing sense and antisense transgenes. Genetics 153:1767–1774

Johnston M, Hillier L, Riles L, et al. (1997) The nucleotide sequence of *Saccharomyces cerevisiae* chromosome XII. Nature 387:87–90

Jurka J (1997) Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. Proc Natl Acad Sci USA 94:1872–1877

Kanaya S, Yamada Y, Kudo Y, Ikemura T (1999) Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: Gene expression level and species-specific diversity of codon usage based on multivariate analysis. Gene 238:143–155

Ketting RF, Haverkamp THA, Van Luenen HGAM, Plasterk RHA (1999) mut-7 of *C. elegans,* required for transposon silencing and RNA interference, is a homolog of Werner syndrome helicase and RnaseD. Cell 99:133–141

Kidwell MG, Lisch DR (2001) Transposable elements, parasitic DNA and genome evolution. Evolution 55:1–24

Kim JM, Vanguri S, Boeke JD, Gabriel A, Voytas DF (1998) Transposable elements and genome organization: A comprehensive survey of retrotransposons revealed by the complete *Saccharomyces cerevisiae* genome sequence. Genome Res 8:464–478

Kliman RM, Hey J (1993) Reduced natural selection associated with low recombination in *Drosophila melanogaster.* Mol Biol Evol 10:1239–1258

Kliman RM, Hey J (1994) The effects of mutation and natural selection on codon bias in the genes of *Drosophila.* Genetics 137:1049–1056

Korenberg JR, Rykowski MC (1988) Human genome organization: *Alu,* LINEs, and the molecular structure of metaphase chromosome bands. Cell 53:391–400

Kypr J, Mrázek J (1987) Unusual codon usage of HIV. Nature 327:20

Le QH, Wright S, Yu Z, Bureau T (2000) Transposon diversity in *Arabidopsis thaliana.* Proc Natl Acad Sci USA 97:7376–7381

Lerat E, Biémont C, Capy P (2000) Codon usage and the origin of *P* elements. Mol Biol Evol 17:467–468

Liao G-C, Rehm EJ, Rubin GM (2000) Insertion site preferences of the *P* transposable element in *Drosophila melanogaster.* Proc Natl Acad Sci USA 97:3347–3351

Lin X, Kaul S, Rounsley S, et al. (1999) Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana.* Nature 402:761–768

Martienssen R (1998) Transposons, DNA methylation and gene control. Trends Genet 14:264–265

Mayer K, Schüller C, Wambutt R, et al. (1999) Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana.* Nature 402: 769–777

McDonald JF (1998) Transposable elements, gene silencing and macroevolution. Trends Ecol Evol 13:94–95

Merriman PJ, Grimes CD, Ambroziak J, Hackett DA, Skinner P, Simmons MJ (1995) *S* elements: A family of *Tc1*-like transposons in the genome of *Drosophila melanogaster.* Genetics 141:1425–1438

Moriyama EN, Powell JR (1997) Codon usage bias and tRNA abundance in Drosophila. J Mol Evol 45:514–523

Moriyama EN, Powell JR (1998) Gene length and codon usage bias in *Drosophila melanogaster, Saccharomyces cerevisiae* and *Escherichia coli.* Nucleic Acids Res 26:3188–3193

O'Neill RJ, O'Neill MJ, Graves JA (1998) Undermethylation associated with retroelement activation and chromosome remodelling in an interspecific mammalian hybrid. Nature 393:68–72

Philippsen P, Kleine K, Pöhlmann R, et al. (1997) The nucleotide sequence of *Saccharomyces cerevisiae* chromosome XIV. Nature 387:93–98

Powell JR, Gleason JM (1996) Codon usage and the origin of *P* elements. Mol Biol Evol 13:278–279

Schmid CW (1998) Does SINEs evolution preclude *Alus* function? Nucleic Acids Res 26:4541–4550

Shapiro JA (1999) Transposable elements as the key to a 21st century view of evolution. Genetica 107:171–179

Sharp PM, Cowe E (1991) Synonymous codon usage in *Saccharomyces cerevisiae.* Yeast 7:657–678

Sharp PM, Lloyd AT (1993) Codon usage. In Maroni G (ed) An Atlas of Drosophila genes: Sequences and molecular features. Oxford University Press, New York, pp 378–397

Sharp PM, Matassi G (1994) Codon usage and genome evolution. Curr Opin Genet Dev 4:851–860

Sharp PM, Tuohy TMF, Mosurski KR (1986) Codon usage in yeast: Cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res 14:5125–5143

Shields DC, Sharp PM (1989) Evidence that mutation patterns vary among Drosophila transposable elements. J Mol Biol 207:843–846

Shields DC, Sharp PM, Higgins DG, Wright F (1988) "Silent" sites in Drosophila genes are not neutral: evidence of selection among synonymous codons. Mol Biol Evol 5:704–716

Springer MS, Tusneem NA, Davidson EH, Britten RJ (1995) Phylogeny, rates of evolution, and patterns of codon usage among sea urchin retroviral-like elements, with implications for the recognition of horizontal transfer. Mol Biol Evol 12:219–230

Stenico M, Lloyd AT, Sharp PM (1994) Codon usage in *Caenorhabditis elegans:* Delineation of translational selection and mutational biases. Nucleic Acids Res 22:2437–2446

Tettelin H, Agostoni Carbone ML, Albermann K, et al. (1997) The nucleotide sequence of *Saccharomyces cerevisiae* chromosome VII. Nature 387:81–84

Thioulouse J, Chessel D, Dolédec S, Olivier JM (1997) ADE-4: A multivariate analysis and graphical display software. Stat Comput 7:75–83

Tomilin NV (1999) Control of genes by mammalian retroposons. Int Rev Cytol 186:1–48

Van Hemert FJ, Berkhout B (1995) The tendency of lentiviral open reading frames to become A-rich: Constraints imposed by viral genome organization and cellular tRNA availability. J Mol Evol 41:132–140

Vartanian JP, Meyerhans A, Sala M, Wain-Hobson S (1994) G–A hypermutation of the human immunodeficiency virus type 1 genome: evidence for dCTP pool imbalance during reverse transcription. Proc Natl Acad Sci USA 91:3092–3096

Wessler SR, Bureau TE, White SE (1995) LTR-retrotransposons and MITEs: Important players in the evolution of plant genomes. Curr Opin Genet Dev 5:814–821

Wichman HA, Van Den Bussche RA, Hamilton MJ, Baker RJ (1992) Transposable elements and the evolution of genome organization in mammals. Genetica 86:287–293

Wright F (1990) The 'effective number of codons' used in a gene. Gene 87:23–29

Wright F, Bibb MJ (1992) Codon usage in the G+C-rich Streptomyces genome. Gene 113:55–65

Yan Z, Nielsen R, Goldman N, Krabbe Pedersen AM (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155:431–449

Yoder JA, Walsh CP, Bestor TH (1997) Cytosine methylation and the ecology of intragenomic parasites. Trends Genet 13:335–340

Zsìros J, Jebbink MF, Lukashov VV, Voûte PA, Berkhout B (1999) Biased nucleotide composition of the genome of *HERV-K*-related endogenous retroviruses and its evolutionary implications. J Mol Evol 48:102–111

Article 5 : The relative abundance of dinucleotides in transposable elements in five species

Mol. Biol. Evol. sous presse

Letter to the Editor

The Relative Abundance of Dinucleotides in Transposable Elements in Five Species

Emmanuelle Lerat, Pierre Capy,† and Christian Biémont**

*Laboratoire Biométrie et Biologie Évolutive, UMR CNRS 5558, Université Lyon 1, 69622 Villeurbanne Cedex, France; and †Laboratoire Populations, Génétique et Évolution, UPR CNRS 9034, 91198 Gif/Yvette Cedex, France

Burge, Campbell, and Karlin (1992) observed that the relative frequencies of di- and trinucleotides characterize a genome, independent of its base composition and the coding and noncoding capacity of the regions analyzed. Species thus differ with regard to this genomic signature, which is constant in a given genome and shows similarities between related species (Gentles and Karlin 2001). The variation in the relative abundance of dinucleotides is interpreted as reflecting differences between species in the cellular machinery for replication and repair, which may select specific dinucleotides in the sequence (Campbell, Mrázek, and Karlin 1999). A tendency toward the suppression of CG is often observed and is interpreted as resulting from the action of methylation activities (Bird 1986). The dinucleotides pattern of the mitochondrial genome has also been shown to differ from that of the nuclear genome, and the explanation suggests that nuclear and mitochondrial genomes use independent DNA polymerase machinery and different methods of replication (Campbell, Mrázek, and Karlin 1999). We therefore wanted to find out whether transposable elements (TEs), which have been shown to have a greater AT content than their host genes in various species (Shields and Sharp 1989; Lerat, Capy, and Biémont 2002), have the same dinucleotides pattern as their host.

TEs are repeated sequences that are able to move from one position to another along chromosomes. They were first discovered in maize by Barbara McClintock (1984) in the 1950s and seem to exist in all living organisms. They are divided into two main classes, according to the transposition intermediate they use (Capy et al. 1997, pp. 1–197). Class I consists of retrotransposons that use an RNA intermediate and are subdivided into two subclasses according to whether they do or do not have long terminal repeats (LTRs) at their extremities, LTR retrotransposons and non-LTR retrotransposons, respectively. Class II consists of transposons that use a DNA intermediate for transposition and code for a transposase. There is a third class that consists of foldback elements and MITEs, the transposition mechanism of which has not yet been elucidated.

Key words: transposable elements, retrovirus, dinucleotide abundance.

Address for correspondence and reprints: Christian Biémont, Laboratoire de Biométrie et Biologie Évolutive, UMR CNRS 5558, Université Lyon 1, 69622 Villeurbanne Cedex, France. E-mail: biemont@biomserv.univ-lyon1.fr.

Mol. Biol. Evol. 19(6):964–967. 2002

The complete genomes of *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, and *Drosophila melanogaster*, chromosomes 2 and 4 of *Arabidopsis thaliana*, and chromosomes 21 and 22 of *Homo sapiens* were downloaded from the Genome On Line Database site (wit.integratedgenomics.com/GOLD/) (Kyrpides 1999). Entire sequences of transposons, LTR retrotransposons and non-LTR retrotransposons, and of class-III elements from *C. elegans*, *D. melanogaster*, *H. sapiens*, and *A. thaliana* were downloaded from GenBank. Other *Arabidopsis* TEs were obtained from the Arabidopsis transposable element database (soave.biol.mcgill.ca/clonebase/main.html). The positions of TEs in the sequenced genome of *Saccharomyces* were obtained from the site transposable element resources (www.public.iastate.edu/~voytas/resources/resources.html). The TE data set, thus available, consisted of 40 sequences from *D. melanogaster*, 50 from *S. cerevisiae*, 19 from *C. elegans*, 25 from *H. sapiens*, and 31 from *A. thaliana*. The TE sequences for each species were concatenated. Of the 25 TE sequences from *H. sapiens*, 10 were retroviruses (*HERV-K, HERV-K-T47D, HERV-K101, HERV-KC4, HIV1, HIV2, HTLV1, HTLV2, HSRV,* and v-oncogene), which are class-I elements and can be considered to belong to the LTR retrotransposon family.

We used the indices defined by Burge, Campbell, and Karlin (1992). For a dinucleotide XY, the indices $\rho_{XY} = f_{XY}/f_X f_Y$ were computed for each sequence, where f_X and f_Y are the frequencies of bases X and Y, respectively, and f_{XY} the frequency of the dinucleotide XY. When the coding sequences of TEs and genes were used, the indices were only calculated from single-stranded DNA. For complete sequences, we took into account the antiparallel and complementary structure of double-stranded DNA (Burge, Campbell, and Karlin 1992). We thus computed $f^*_A = f^*_T = 1/2(f_A + f_T)$ for base A and its associated T nucleotide in the double-stranded sequence and $f^*_G = f^*_C = 1/2(f_G + f_C)$ for base G and its associated C nucleotide. The frequency of the GT dinucleotide was computed as $f^*_{GT} = (1/2f_{GT} + 1/2f_{AC})$, and the indices $\rho^*_{XY} = f^*_{XY}/f^*_X f^*_Y$ were estimated. According to Karlin and Burge (1995), the XY dinucleotide was considered to be underrepresented if $\rho^*_{XY} \leq 0.78$ and overrepresented if $\rho^*_{XY} \geq 1.23$.

The relative distance between two sequences, f and g, was calculated as the sum of the differences between the ρ^*_{ij} indices for each ij dinucleotide between the two sequences: $\delta^*(f,g) = (1/16)\Sigma_{ij} |\rho^*_{ij}(f) - \rho^*_{ij}(g)|$ (Karlin and Ladunga 1994; Karlin and Mrázek 1997). Relative distances were computed for the genomic sequences and the concatenated TEs for all species, the fragments of genomic sequences and complete TEs for all species,

and the host genes and coding parts of TEs for each species separately. The distance matrix obtained was analyzed using a principal coordinates analysis, a specific multivariate analysis which transforms distance matrices into euclidean matrices before extracting the principal components (Gower 1966). This analysis makes it possible to visualize neighboring sequences in terms of their relative abundance of dinucleotides. These analyses were done using the ADE-4 package (Thioulouse et al. 1997).

The relative abundances of dinucleotides in TE and genomic sequences were calculated for the five species listed previously (detailed data available upon request). Whatever the species, the dinucleotide TA appeared to be underrepresented in both genomes and TEs, except in the yeast retrotransposons. The dinucleotide CG was underrepresented in both genomes and TEs in *A. thaliana* and *H. sapiens* and in the LTR retrotransposons Ty1, Ty4, and Ty5 in *Saccharomyces*. In the Caenorhabditis and Drosophila genomes, AA/TT was overrepresented. For a given species, the TE and genomic sequences displayed the same global pattern of relative dinucleotides abundance, as revealed by the positive correlation coefficients for the relative abundance of dinucleotides between TEs and host genomes ($r = 0.98$, $P < 0.05$ for Arabidopsis; $r = 0.93$, $P < 0.05$ for Caenorhabditis; $r = 0.94$, $P < 0.05$ for Drosophila; $r = 0.87$, $P < 0.05$ for *H. sapiens*). For Saccharomyces, the coefficient of correlation between the genome and TEs was not different from zero ($r = 0.54$, $P = 0.40$).

To check for a codon signature in coding regions, we calculated the relative abundance of dinucleotides according to their position in codons along the single-stranded DNA (data available upon request). The strong positive correlation detected at position 1–2 of codons between genes and TEs for each species ($r = 0.93$, $P < 0.05$ for Arabidopsis; $r = 0.90$, $P < 0.05$ for Caenorhabditis; $r = 0.70$, $P < 0.05$ for Drosophila; $r = 0.77$, $P < 0.05$ for human; $r = 0.91$, $P < 0.05$ for Saccharomyces) suggests that there were only a few differences between TE and gene sequences in the relative abundances patterns of dinucleotides. The correlation was also positive at position 2–3 for Arabidopsis ($r = 0.88$, $P < 0.05$), for Caenorhabditis ($r = 0.64$, $P < 0.05$), for human ($r = 0.80$, $P < 0.05$), and for Saccharomyces ($r = 0.64$, $P < 0.05$) but was not statistically different from zero in *D. melanogaster* ($r = 0.17$, $P = 0.40$). In *D. melanogaster* and *S. cerevisiae*, the relative abundance of dinucleotides at position 3–1 ($r = 0.40$, $P = 0.40$; $r = 0.50$, $P = 0.40$ for Drosophila and Saccharomyces, respectively) showed no correlation to that found in other species ($r = 0.87$, $P < 0.05$ for Arabidopsis; $r = 0.77$, $P < 0.05$ for Caenorhabditis; $r = 0.90$, $P < 0.05$ for human). The dinucleotide TA was strongly underrepresented at all positions in both genes and TEs in all the species, except Saccharomyces, where TA was underrepresented only at position 1–2 of the codons. TT and TC were strongly overrepresented, and CG and GT were underrepresented at position 1–2 in all the data sets. The TG and CA dinucleotides were well represented at position 2–3 and 3–1: ρ_{TG} and ρ_{CA} were often greater than 1 and sometimes reached values indicative of overrepresentation ($\rho > 1.23$).

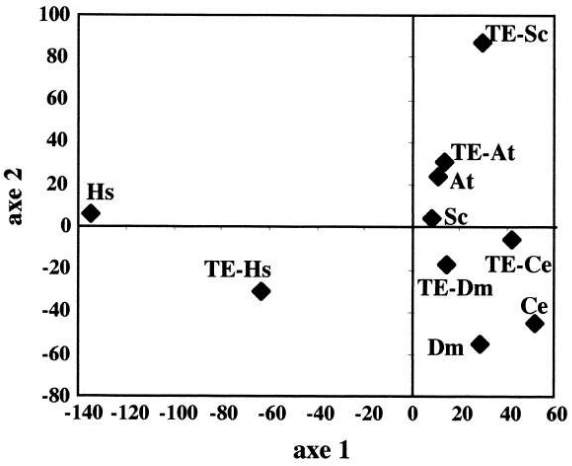


FIG. 1.—Plot of the two first axes of the principal coordinates analysis of dinucleotide distances of genomes and TEs. TE-At = TEs of *A. thaliana*, TE-Ce = TEs of *C. elegans*, TE-Sc = TEs of *S. cerevisiae*, TE-Dm = TEs of *D. melanogaster*, TE-Hs = TEs of *H. sapiens*, At = genome of *A. thaliana*, Ce = genome of *C. elegans*, Sc = genome of *S. cerevisiae*, Dm = genome of *D. melanogaster*, Hs = genome of *H. sapiens*.

Figure 1 shows the projection of TEs and genomes onto the plane defined by the two first axes of a principal coordinates analysis of the distance matrix between the dinucleotide relative abundance indices of genomic and TE sequences. TE and genomic sequences from one species were close, except for Saccharomyces, which presented no correlation between TE and genomic sequences for dinucleotide relative abundance. In this analysis, we compared TE sequences from genomic sequences likely to include TEs, and we therefore carried out a more detailed principal coordinates analysis on complete TE sequences and on TE-free genomic fragments. To do this, genomic sequences were broken down into genomic fragments of 9,000 bp size, which was roughly equivalent to the mean length of the complete TEs. For each species, 100 fragments were randomly selected and a BLASTN analysis (Altschul et al. 1997) was done to compare the genomic fragments and TE sequences and allow us to eliminate the genomic fragments including TEs. In this way, we obtained a total of 459 TE-free genomic fragments and 165 complete TE sequences for the five species. The distances between the indices of relative dinucleotides abundance were then computed. The relative abundances of dinucleotides in the genomic fragments were nearly the same as the values obtained for the overall genomic sequences. With the exception of Saccharomyces, TE sequences and genomic fragments from a given species were found to be clustered (figure available upon request).

Figure 2 shows the plot of the dinucleotide relative abundance distances between genes and coding parts of TEs for each species separately. Coding regions of the TEs and host genes appeared to be located together in Caenorhabditis and Arabidopsis. In *H. sapiens*, some of the TEs were located with the host genes, whereas the rest, corresponding to retrovirus sequences, formed a

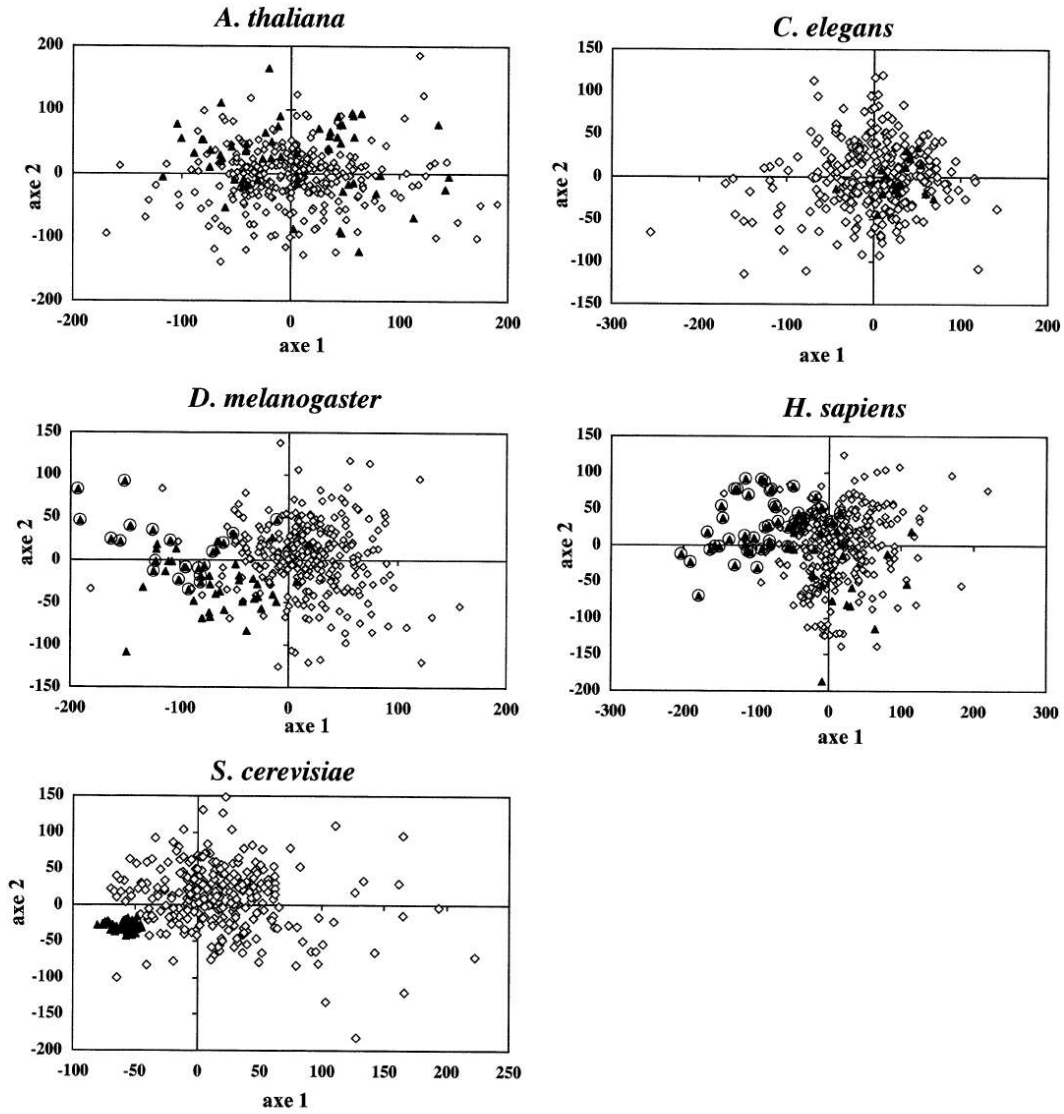


FIG. 2.—Plot of the two first axes of five different principal coordinates analysis of dinucleotide distances of host genes and coding parts of TEs, for each species analyzed individually. Black triangles correspond to coding parts of TEs and white squares to host genes. Circled black triangles correspond to the coding parts of LTR retrotransposons with an *env* gene in *D. melanogaster* and to the coding parts of human retroviruses.

distinct group. In Drosophila and Saccharomyces, the TEs were not located with host genes. In Drosophila, the TEs furthest from the host genes corresponded to LTR retrotransposons with an *env* gene, e.g., retrovirus-like elements (*Tirant, 297, ZAM* and in a lowest way *17.6, gypsy, idefix,* and *nomad*).

In the five species analyzed, *A. thaliana*, *C. elegans*, *S. cerevisiae*, *D. melanogaster*, and *H. sapiens*, TEs appear to display a similar pattern of the relative abundances of dinucleotides as their host genome. In all our analyses, we found that the TA dinucleotide was underrepresented in both genomes and TEs. Such underrepresentation of TA, which seems to be a general feature, is attributed to (1) the avoidance of the inappropriate terminate codons TAA or TAG in coding sequences, (2) the selection of mRNA stability by avoiding UpA, which is susceptible to RNAse activity (Beutler et al. 1989), or (3) the avoidance of having too many transcription signals (Burge, Campbell, and Karlin 1992). We also observed CG suppression in both genomes and TEs in Arabidopsis and human. Such global CG suppression is believed to reduce the stacking energies of DNA, thus facilitating replication and transcription (Karlin and Burge 1995). The fact that no CG suppression was observed in *C. elegans*, *S. cerevisiae*, and *D. melanogaster* suggests, however, that this explanation is far from universally applicable. We show here that CG suppression, which has been already reported in small eukaryotic viruses (Karlin, Doerfler, and Cardon 1994), also exists in the elements Ty1, Ty4, and Ty5 of Saccharomyces, in many LTR retrotransposons of Arabidopsis, and in all the LTR retrotransposons of *H. sapiens*. In Drosophila, however, LTR retrotransposons with an *env* gene do not exhibit this underrepre-

sentation of CG. The combination of these findings suggests that CG suppression does not affect all kinds of transposable elements and is not related to the size of the TE sequence.

Multivariate analysis showed that the retroviruses of *H. sapiens* and the LTR retrotransposons with *env* genes of Drosophila were very distant from their host genes. This specific grouping of the coding parts of retrovirus-like elements and of retroviruses relative to the host genes was not found when entire sequences were used, suggesting that there are differences in the transcription mechanisms for the coding parts of these elements. The coding parts of *HERV* (human endogenous retrovirus) were also located with the other retroviruses, although such endogenous retroviruses are not infectious because of deletions or the presence of stop codons in their coding parts (Bock and Stoye 2000; Tristen 2000). It has been shown, however, that the *HERV-K* element can theoretically be *trans*-complemented and then becomes infectious (Bock and Stoye 2000). If the large dinucleotide relative abundance distances observed between host genes and retroviruses and some LTR retrotransposon genes is an indication of their infectivity, then we can expect the Drosophila elements, *297, Tirant, 17.6,* and *idefix* to be infectious or to have been infectious in the recent past. Infectious capacity has been clearly demonstrated for *gypsy* (Kim et al. 1994), but the other five elements are only suspected of being retroviruses (Dessat et al. 1999; Canizares et al. 2000). Experimental evidences are therefore required to test the theoretical expectation of the present analysis.

LITERATURE CITED

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. H. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**:3389–3402.

Beutler, E., T. Gelbart, J. H. Han, J. A. Koziol, and B. Beulter. 1989. Evolution of the genome and the genetic code: selection at the dinucleotide level by methylation and polyribonucleotide cleavage. Proc. Natl. Acad. Sci. USA **86**:192–196.

Bird, A. P. 1986. CpG-rich islands and the function of DNA methylation. Nature **321**:209–213.

Bock, M., and J. P. Stoye. 2000. Endogenous retroviruses and the human germline. Curr. Opin. Genet. Dev. **10**:651–655.

Burge, C., A. M. Campbell, and S. Karlin. 1992. Over- and under-representation of short oligonucleotides in DNA sequences. Proc. Natl. Acad. Sci. USA **89**:1358–1362.

Campbell, A., J. Mrázek, and S. Karlin. 1999. Genome signature comparisons among prokaryote, plasmid, and mictochondrial DNA. Proc. Natl. Acad. Sci. USA **96**:9184–9189.

Canizares, J., M. Grau, N. Paricio, and M. D. Molto. 2000. Tirant is a new member of the gypsy family of retrotransposons in *Drosophila melanogaster*. Genome **43**:9–14.

Capy, P., C. Bazin, D. Higuet, and T. Langin. 1997. Dynamics and evolution of transposable elements. R. G. Landes Company, Austin, Tex.

Dessat, S., C. Conte, P. Dimitri, V. Calco, B. Dastugue, and C. Vaury. 1999. Mobilization of two retroelements ZAM and Idefix, in a novel instable line of *Drosophila melanogaster*. Mol. Biol. Evol. **16**:54–66.

Gentles, A. J., and S. Karlin. 2001. Genome-scale compositional comparisons in eukaryotes. Genome Res. **11**:540–546.

Gower, J. C. 1966. Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika **53**:325–338.

Karlin, S., and C. Burge. 1995. Dinucleotide relative abundance extremes: a genomic signature. Trends Genet. **11**: 283–290.

Karlin, S., W. Doerfler, and L. R. Cardon. 1994. Why is CpG suppressed in the genomes of virtually all small eukaryotic viruses but not in those of large eukaryotic viruses? J. Virol. **68**:2889–2897.

Karlin, S., and I. Ladunga. 1994. Comparisons of eukaryotic genomic sequences. Proc. Natl. Acad. Sci. USA **91**:12832–12386.

Karlin, S., and J. Mrázek. 1997. Compositional differences within and between eukaryotic genomes. Proc. Natl. Acad. Sci. USA **94**:10227–10232.

Kim, A., C. Terzian, P. Santamaria, A. Pélisson, N. Prud'homme, and A. Bucheton. 1994. Retroviruses in invertebrates: the gypsy retrotransposon is apparently an infectious retrovirus of *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA **91**:1285–1289.

Kyrpides, N. 1999. Genomes OnLine Database (GOLD): a monitor of complete and ongoing genome projects world wide. Bioinformatics **15**:773–774.

Lerat, E., P. Capy, and C. Biémont. 2002. Codon usage by transposable elements and host genes in five species. J. Mol. Evol. (in press).

McClintock, B. 1984. The significance of responses of the genome to challenge. Science **226**:792–801.

Shields, D. C., and P. M. Sharp. 1989. Evidence that mutation patterns vary among Drosophila transposable elements. J. Mol. Biol. **207**:843–846.

Thioulouse, J., D. Chessel, S. Dolédec, and J. M. Olivier. 1997. ADE-4: a multivariate analysis and graphical display software. Stat. Comput. **7**:75–83.

Tristen, M. 2000. Identification and characterization of novel human endogenous retrovirus families by phylogenetic screening of the human genome mapping project database. J. Virol. **74**:3715–3730.

Wolfgang Stephan, reviewing editor

Accepted November 21, 2000

Résumé

Les éléments transposables (ETs), qui sont présents chez tous les organismes vivants et sont impliqués dans un grand nombre de mutations et de réarrangements chromosomiques, apparaissent comme des composants incontournables des génomes. Ils doivent alors être soumis aux mêmes contraintes que les gènes d'hôte. Afin de tester cette hypothèse, nous avons dans un premier temps analysé l'usage des codons des gènes d'ETs et des gènes d'hôte chez cinq espèces : *A. thaliana, C. elegans, D. melanogaster, H. sapiens* et *S. cerevisiae*. Les résultats montrent que les ETs sont riches en AT quelle que soit l'espèce hôte : il s'agit donc d'une propriété intrinsèque aux ETs. L'analyse de la composition en bases aux différentes positions des codons montre que la richesse en AT à la 3ème position est inférieure aux valeurs des régions non contraintes des génomes. Ainsi, les ETs ne subissent pas uniquement des biais mutationnels mais sont aussi soumis à de la sélection. L'usage des codons des ETs n'est cependant pas lié à un taux d'expression fort ou faible, ce qui suggère un pattern d'expression particulier pour ces éléments.

Dans un deuxième temps, l'analyse de l'abondance relative en di- et en trinucléotides des ETs et des génomes hôtes montre que les ETs possèdent un pattern d'abondance similaire à celui de leur hôte, indépendamment du biais de composition en bases. Cette analyse montre cependant que les gènes de rétrovirus humains et de rétrotransposons à LTR avec un gène *env* de drosophile ont un pattern différent des gènes d'hôte. L'abondance en dinucléotides semble être un moyen de détecter les rétroéléments potentiellement infectieux.

Ce travail suggère un comportement spécifique des ETs qui semblent soumis à des contraintes de sélection particulières permettant le maintien de leur richesse en AT. Cependant, ils subissent aussi une empreinte du génome hôte, probablement de nature structurale.

Comparison of transposable elements sequences and host genes in five species: *A. thaliana*, *C. elegans*, *D. melanogaster*, *H. sapiens*, and *S. cerevisiae*.

Transposable elements (TEs), which are found in all living organisms and are implied in many mutations and chromosome rearrangements, appear to be major components of genomes. They should be submitted to the same constraints than the host genes. To test this hypothesis, we thus at first time have analyzed codon usage of TEs and host genes in five species: *A. thaliana*, *C. elegans*, *D. melanogaster*, *H. sapiens*, and *S. cerevisiae*. Results show that TEs are AT rich whatever the host species: this seems to be an intrinsic characteristic of the TEs. The analysis of the base composition at the different positions of the codons shows that the AT richness on the third position is lower than on the unconstrained regions of the genomes. TEs do not thus only undergo mutational bias but are submitted to selection. However, codon usage in TEs is not related to a low or high expression level, which suggests a particular expression pattern for these elements.

An analysis of relative abundance of di- and trinucleotides of TE sequences and host genomes shows that TEs have pattern of abundance similar to that of the host, independently to the base composition bias. This analysis shows, however, that human retrovirus genes and LTR retrotransposons with an *env* gene of Drosophila have a pattern of relative abundance different to those of the host genes. Dinucleotide abundance seems thus to be a good way to detect potentially infectious retroelements.

This work suggests a specific behavior of TEs, which seem submitted to particular selection constraints allowing the keeping of AT richness. However, they undergo host genome stamp, probably of structural type.

Mots clés
elements transposables, usage des codons, abondance relative en di- et en trinucléotides

Discipline
bioinformatique génomique

Laboratoire Biométrie et Biologie Evolutive, UMR CNRS 5558, Université Lyon I, Villeurbanne.