Étude de l'influence des éléments transposables sur la régulation des gènes chez les mammifères

Hussein Mortada

N° d'ordre **178-2011** Année 2011

THESE DE L'UNIVERSITE DE LYON

Délivrée par

L'UNIVERSITE CLAUDE BERNARD LYON 1

ECOLE DOCTORALE E2M2

DIPLOME DE DOCTORAT

(Arrêté du 7 août 2006)

Soutenue publiquement le 04 octobre 2011

par

Hussein MORTADA

TITRE :

Étude de l'influence des éléments transposables sur la régulation des gènes chez les mammifères.

Directeur de thèse : Emmanuelle LERAT

JURY :

Monsieur Christian GAUTIER	Président
Monsieur Richard CORDAUX	Rapporteur
Madame Aurélie HUA-VAN	Rapporteur
Madame Carène RIZZON	Examinateur
Madame Emmanuelle LERAT	Directeur
Madame Cristina VIEIRA	Co-directeur

Laboratoire de Biométrie et Biologie Evolutive
UMR CNRS 5558, Université Claude Bernard Lyon 1
43 bd du 11 novembre 1918 – 69622 Villeurbanne

UNIVERSITE CLAUDE BERNARD - LYON 1

Président de l'Université	**M. A. Bonmartin**
Vice-président du Conseil d'Administration	M. le Professeur G. Annat
Vice-président du Conseil des Etudes et de la Vie Universitaire	M. le Professeur D. Simon
Vice-président du Conseil Scientifique	M. le Professeur J-F. Mornex
Secrétaire Général	M. G. Gay

COMPOSANTES SANTE

Faculté de Médecine Lyon Est – Claude Bernard	Directeur : M. le Professeur J. Etienne
Faculté de Médecine et de Maïeutique Lyon Sud – Charles Mérieux	Directeur : M. le Professeur F-N. Gilly
UFR d'Odontologie	Directeur : M. le Professeur D. Bourgeois
Institut des Sciences Pharmaceutiques et Biologiques	Directeur : M. le Professeur F. Locher
Institut des Sciences et Techniques de la Réadaptation	Directeur : M. le Professeur Y. Matillon
Département de formation et Centre de Recherche en Biologie Humaine	Directeur : M. le Professeur P. Farge

COMPOSANTES ET DEPARTEMENTS DE SCIENCES ET TECHNOLOGIE

Faculté des Sciences et Technologies	Directeur : M. le Professeur F. Gieres
Département Biologie	Directeur : M. le Professeur F. Fleury
Département Chimie Biochimie	Directeur : Mme le Professeur H. Parrot
Département GEP	Directeur : M. N. Siauve
Département Informatique	Directeur : M. le Professeur S. Akkouche
Département Mathématiques	Directeur : M. le Professeur A. Goldman
Département Mécanique	Directeur : M. le Professeur H. Ben Hadid
Département Physique	Directeur : Mme S. Fleck
Département Sciences de la Terre	Directeur : Mme le Professeur I. Daniel
UFR Sciences et Techniques des Activités Physiques et Sportives	Directeur : M. C. Collignon
Observatoire de Lyon	Directeur : M. B. Guiderdoni
Ecole Polytechnique Universitaire de Lyon 1	Directeur : M. P. Fournier
Ecole Supérieure de Chimie Physique Electronique	Directeur : M. G. Pignault
Institut Universitaire de Technologie de Lyon 1	Directeur : M. le Professeur C. Coulet
Institut de Science Financière et d'Assurances	Directeur : M. le Professeur J-C. Augros
Institut Universitaire de Formation des Maîtres	Directeur : M. R. Bernard

À mes grands parents,

À mes parents,

À Mira,

Quand on commence la rédaction du manuscrit on ne pense qu'au moment où on arrivera aux remerciements mais une fois dedans on se rend compte qu'écrire ces remerciements est encore plus dur.

Mes premiers et énormes remerciements vont tout naturellement à Emmanuelle pour avoir cru en moi, pour m'avoir aidée pendant ces trois années de thèse dans l'apprentissage de la recherche et de la démarche scientifique, et pour m'avoir épaulée et guidée dans les moments difficiles,

Je remercie également Cristina Vieira et Christian Biémont pour leur patience, leurs conseils et tout le temps qu'ils ont passé à corriger mes manuscrits, articles, posters…sans oublier les présentations orales !!

Aurélie Hua-Van, Richard Cordaux, Carène Rizzon et Christian Gautier pour avoir accepté de faire partie de mon jury de thèse et pour leur intérêt pour ce travail,

Mamie, Hyoubert, Doudou, Marcounet ainsi que tous les anciens et actuels membres de l'équipe « TREEP » pour avoir rendu agréable le travail de tous les jours, pour avoir pris le temps d'écouter mes problèmes et soucis, pour avoir supporté mes blagues « à deux balles » et pour m'avoir laissé gagner dans tous les jeux (même si je sais que je reste le plus fort),

Tous les membres du LBBE pour l'ambiance géniale qui règne, pour le soutien qu'on peut trouver quel que soit le problème auquel on fait face… je pense surtout à Anne-Béatrice Dufour et Franck Picard pour tous les problèmes statistiques et à Simon Penel pour les problèmes de programmation,

Dominique Mouchiroud pour m'avoir guidée dans mon parcours universitaire, c'est grâce à vous que j'ai découvert la bioinformatique et que je m'y suis attaché (même s'il est vrai que j'étais nul en paillasse), et pour m'avoir encadrée et épaulée pendant mon stage de Master1,

Ali Machmouchi pour avoir endossé le rôle du frère toujours prêt à m'aider quand j'avais besoin de lui, pour m'avoir encouragé dans les moments difficiles, et pour m'avoir hébergé quand j'en avais besoin,

Mohammad, Ahmad (x2), Ali (x3), Hussein ainsi que tous les amis libanais de Lyon, pour le soutien que vous m'avez apporté, pour les barbec' improvisés mais réussis, pour les soirées PS3, et pour votre patience,

L'équipe de poussines au BCCL ainsi que les parents, pour avoir fait office de nouvelle famille, pour avoir supporté mes retards pendant trois ans, pour tous les pique-niques dans les petites villes paumées du Rhône, pour votre passion pour le basket, et pour toute la confiance que vous avez placé en moi,

Le président (Alain), Nico, toute l'équipe Sénior 4 du BCCL, ainsi que tous les membres du BCCL,

Rémi, Mathieu, Romain et Bastien pour toutes les séances de « jorky » sans oublier l'aventure du triathlon des 24 heures de l'INSA,

Mira pour avoir été mon ange gardien pendant ces deux dernières années, pour avoir pris le temps d'écouter mes problèmes (même si ta réponse était toujours « Hayni » ou « Bet Houn »), et pour tous les merveilleux moments que je vis grâce à ta présence,

Mes parents pour avoir cru en moi dès le départ, pour m'avoir fait confiance depuis tout petit, pour m'avoir guidé quand j'en avais besoin, pour avoir fait de moi ce que je suis aujourd'hui, pour m'avoir apporté le support financier nécessaire pendant six longues années à vos dépends… C'est grâce à vous que j'ai pu surmonter tous les problèmes auxquels j'ai pu faire face… pour tout ce que vous avez fait pour moi, je ne vous remercierai jamais assez !

On dit qu'on garde toujours le meilleur pour la fin … c'est pour cela que je garde mes derniers remerciements à mes grands-parents que j'ai perdu au cours de ma thèse… J'aurais aimé fêter la fin de ma thèse avec vous deux mais le destin en a décidé autrement… Vous n'aurez certainement pas l'occasion de lire ces remerciements mais sachez qu'au fond de moi je ne cesse de me répéter que je suis l'homme le plus chanceux du monde pour avoir été votre petit-fils… C'est pour cela que je vous dédie ce manuscrit.

RESUME

Les éléments transposables sont des séquences génomiques capables de se répliquer et de se déplacer dans les génomes. Leur capacité à s'insérer près des gènes et à produire des réarrangements chromosomiques par recombinaison entre copies, font des éléments transposables des agents mutagènes. Les éléments transposables sont de plus capables de modifier l'expression des gènes voisins grâce aux régions promotrices qu'ils possèdent. Les éléments transposables ont été trouvés dans la plupart des génomes dans lesquels ils ont été recherchés. Ils forment ainsi 45 % du génome de l'homme et peuvent représenter jusqu'à 90 % du génome de certaines plantes. Dans la première partie de ma thèse, je me suis penché sur les facteurs qui déterminent la distribution de ces éléments. Je me suis intéressé à un facteur particulier, qui est la fonction des gènes dans le voisinage des insertions d'éléments transposables. Dans la deuxième partie, j'ai essayé de déterminer l'impact de l'altération des modifications épigénétiques (modifications d'histones plus précisément) associées aux différents composants géniques, dont les éléments transposables, sur la variation de l'expression des gènes en condition tumorale.

TITLE

Study of transposable element influence on gene regulation in mammals

ABSTRACT

Transposable elements are genomic sequences able to replicate themselves and to move within genomes. Their ability to integrate near genes and to produce chromosomal rearrangements by recombination between copies, make transposable elements mutagens. Moreover, transposable elements are able to alter the expression of neighboring genes through their promoter regions. Transposable elements form 45% of the human genome and may represent up to 90% of certain plant genomes. In the first part of my thesis, I examined the factors that determine the distribution of these elements. I have been interested in a particular factor, which is the function of the genes in the vicinity of transposable element insertions. In the second part, I determined the impact of epigenetic modifications alterations (histone modifications) in different gene components, including transposable elements, on the variation of gene expression in tumoral conditions.

DISCIPLINE

E2M2 : Évolution, Écosystèmes, Microbiologie, Modélisation

MOTS-CLES

Élément transposable, fonction génique, pression de sélection, génome humain, mammifères, expression des gènes, modifications post-traductionnelles des histones, cancer, altérations épigénétiques

INTITULE ET ADRESSE DE L'U.F.R. OU DU LABORATOIRE :

Laboratoire de Biométrie et Biologie Evolutive
UMR CNRS 5558, Université Claude Bernard Lyon 1
43 bd du 11 novembre 1918 – 69622 Villeurbanne

Table des matières

Table des figures

Liste des tableaux

Résumé

Les éléments transposables (ET) sont des séquences génomiques capables de se répliquer et de se déplacer dans les génomes. Les ET ont été trouvés dans la plupart des génomes chez lesquels ils ont été recherchés. Leurs capacités à s'insérer dans les gènes et à produire des réarrangements chromosomiques par recombinaison entre copies, font des ET des agents mutagènes. Malgré leurs effets mutagènes, parfois délétères, leur implication dans l'évolution des génomes n'est plus contestée.

Dans la première partie, je me suis penché sur les facteurs qui peuvent influencer la distribution des ET. Je me suis intéressé à un facteur particulier, qui est la fonction des gènes dans le voisinage des insertions d'ET. La comparaison des fonctions des gènes a montré que la densité en ET dans le voisinage d'un gène dépend de sa fonction dans le sens où les gènes impliqués dans des fonctions de régulation et de développement possèdent peu, voire aucune insertion d'ET, alors que les gènes impliqués dans les processus métaboliques et dans le transport possèdent une forte densité en ET. J'ai expliqué la variation de la densité en ET entre les gènes par une pression de sélection purificatrice plus forte qui s'exerce non seulement au niveau des régions codantes des gènes mais également au niveau de leurs régions flanquantes. Enfin, la comparaison du niveau d'expression des gènes a permis de voir que les gènes riches en ET possèdent un niveau d'expression plus élevé, particulièrement dans les tissus du système immunitaire et les tissus tumoraux, supposant ainsi un rôle important des ET dans la régulation des gènes.

Dans la deuxième partie, je me suis intéressé à la dérégulation épigénétique des ET comme un moyen d'influencer l'expression des gènes. Afin d'améliorer notre compréhension de cette dérégulation, j'ai étudié les modifications d'histones dans deux lignées cellulaires, une tumorale et une autre normale. Je me suis intéressé à huit modifications d'histones différentes. Après avoir divisé les gènes humains en quatre classes différentes selon leur divergence d'expression entre les deux états normal et tumoral, j'ai regardé les variations d'enrichissement pour chaque modification d'histone. Les résultats obtenus montrent que toutes les modifications étudiées subissent des variations d'enrichissement entre l'état normal et l'état tumoral. Cependant, ces variations d'enrichissement n'expliquent qu'un petit pourcentage ($\sim$0,21%) de la divergence d'expression des gènes.

Abstract

Transposable elements (TEs) are DNA sequences able to move and duplicate within genomes. They can be found in various proportions depending on the genomes analyzed, and the reason why TEs are able to invade some genomes more than others are still unknown. TEs can induce ectopic recombination, gene disruption or deregulation and chromosome breakage. Despite those mutagenic and sometimes deleterious aspects, their implication in genome evolution is no longer contested.

The first part of my PhD work was to identify factors that can influence the distribution of TEs in the human genome. I have focused on a specific factor, the function of genes in the vicinity of TE insertions. This analysis showed that the proportion of TE insertions depend on the function of genes in their vicinity, *i.e.* genes that have less TE insertions have different functions compared to genes enriched in TEs. This bias was shown to be due to a selective pressure acting not only on the coding regions of the gene but also on the non-coding regions upstream and downstream the gene. Moreover, we showed that genes rich in TEs have higher expression levels than TE-poor genes, especially in immune system and tumoral tissues, suggesting an important role of TEs in gene regulation.

In the second part of my work I focused on the epigenetic histone modifications as a clue to explain the TE effect on gene expression. Indeed, epigenetic histone modifications are thought to serve as a defense in mammalian genomes against deleterious effects associated with TE activity. Thus, one would expect that in tumoral conditions a deregulation of epigenetic control of TEs that will activate these elements allowing them to modify the expression of neighboring genes. In order to better understand this deregulation, I have studied the modifications for two cell lineages, one normal and the other tumoral (cancer). After having classified genes in four different classes depending on the variation of the expression level between normal and tumoral conditions, I have studied the enrichment for eight histone modification marks for genes in each class. The results we obtained showed that there is global change in all the histone modifications studied and a little percentage ($\sim$0,21%) of the variation of expression is explained by the variation of the enrichment of histone modifications.

Préambule

2001 fut une année clée dans le monde de la recherche scientifique en biologie. En effet, c'est l'année de la publication de la première version du génome humain, une version "brouillon" qui contenait "seulement" 90% du génome, la séquence complète ayant été annoncée comme finalisée le 14 avril 2003. Cette publication a été le fruit d'un travail long d'une quinzaine d'années puisque le projet du séquençage du génome humain a été discuté en 1985 et conçu en 1990 [Paslier et Bernot, 2001]. Une fois le génome humain séquencé, une multitude de mystères pouvaient enfin être abordés. Un de ces mystères dont le nom est "Éléments Transposables" fut, pendant ces vingt dernières années, l'attraction de beaucoup de chercheurs. Il fut également mon attraction pendant mes trois années de thèse, période au cours de laquelle j'ai essayé d'apporter ma contribution à la compréhension de l'influence de ces éléments sur l'expression des gènes.

Les éléments transposables restent aujourd'hui un sujet de débat, ceci malgré tous les progrès qui ont été faits quant à la compréhension de leurs mécanismes de réplication et de transmission et de leurs modes d'action. Ce débat porte essentiellement sur le rôle que joue ces éléments, ce qui divise la communauté des chercheurs sur les éléments transposables en deux groupes majeurs : ceux qui considèrent les éléments comme des parasites qui ne s'intéressent qu'à assurer leur survie même si celle-ci se fait aux dépens de l'hôte, et ceux qui considèrent que ces éléments constituent un réservoir génétique qui permet l'évolution des génomes. Les travaux de recherche que j'ai menés soutiennent en partie chacune de ces hypothèses. Je me suis focalisé sur les primates d'une façon globale et sur l'homme plus spécifiquement. La première partie de ma thèse étant axée sur les facteurs qui déterminent la distribution des éléments transposables, je me suis intéressé à la fonction des gènes situés dans le voisinage de ces éléments transposables. Dans la deuxième partie, Je me suis penché sur l'activité de ces éléments en condition tumorale afin de voir s'ils échappent à la répression de leur expression en condition normale.

Chapitre 1

Introduction

1.1 Définition et historique

Les éléments transposables (**ET**) ou "gènes sauteurs" sont par définition des séquences d'ADN capables de se répliquer et de se déplacer (autrement dit transposer) d'une position chromosomique à une autre dans le génome. Ces éléments ont, pour la première fois, été mis en évidence dans les années 50 par les travaux de Barbara McClintock, publiés le 8 avril 1953 dans un article intitulé "The origin and behavior of mutable loci in maize" [McClintock, 1950]. Dans cette étude, l'auteur met en évidence que les ET sont responsables de la différence de pigmentation que l'on peut observer sur certains grains de maïs. En effet, le gène responsable de la pigmentation des grains du maïs possède une insertion de l'élément *Ds* (Dissociateur) qui inhibe l'expression de ce gène. Quand un élément *Ac* (Activateur) induit l'excision de l'élément *Ds*, l'activité du gène est rétablie permettant ainsi la synthèse de pigments. Cette découverte a bousculé la connaissance des biologistes de l'époque puisqu'elle montre non seulement que l'ADN n'est pas une structure stable, mais également que les gènes peuvent être influencés par les ET, qui peuvent modifier leur expression et par conséquent le phénotype qui en découle [Böhne *et al.*, 2008] . Cette idée fut très contestée par la communauté scientifique jusqu'à la découverte de la dysgénésie des hybrides chez la drosophile [Picard *et al.*, 1978]. Ce phénomène se produit lors d'un croisement entre lignées spécifiques de *Drosophila melanogaster* et conduit à divers changements génétiques incluant une stérilité et des taux élevés de mutation et de recombinaison. Il a été montré plus tard que ces effets étaient dus à une mobilisation d'ET [Rubin *et al.*, 1982] [Bingham *et al.*, 1982].

Aujourd'hui les ET sont largement étudiés au vue de leur présence quasi ubiquitaire dans les génomes dans lesquels ils ont été recherchés, mais aussi leur implication importante dans l'évolution des génomes hôtes [Biémont, 2010a]. La proportion des ET dans les génomes est très variable. En effet, les ET peuvent constituer la majeure partie d'un génome donné, comme c'est le cas chez le maïs chez lequel ils représentent jusqu'à 80% du génome. À l'opposé les ET peuvent former seulement une infime partie du génome hôte comme chez l'abeille chez laquelle ils représentent 1% du génome séquencé. La **Figure 1** présente un exemple de variation du pourcentage d'ET dans les génomes de certains eucaryotes. On peut voir que la proportion des ET chez les primates (homme, chimpanzé et macaque) est quasiment

identique autour de 45%. Cette proportion peut cependant varier entre deux espèces très proches comme c'est le cas entre *Drosophila melanogaster*, qui a 15% d'ET, alors que son espèce soeur *D. simulans* n'en a que 5%. Néanmoins, le séquençage massif de génomes a permis d'identifier des organismes eucaryotes unicellulaires pour lesquels aucune présence d'ET n'a pu être détectée [Pritham, 2009]. Le nombre total des génomes dépourvus d'ET est aujourd'hui estimé à huit (l'algue rouge ou rodophyte *Cyanidioschyzon merolae* [Misumi *et al.*, 2005], les sporozoaires : *Babesia bovis* [Brayton *et al.*, 2007], *Cryptosporidium hominis* [Xu *et al.*, 2004], *C. parvum* [Abrahamsen *et al.*, 2004], *Plasmodium falciparum* [Gardner *et al.*, 2002], *P. yoelli yoelli* [Carlton *et al.*, 2002] et *Theileria parva* [Gardner *et al.*, 2005] et enfin la microsporidie *Encephalitozoon cuniculi* [Katinka *et al.*, 2001]) mais va très certainement augmenter dans le futur avec l'augmentation du nombre de génomes séquencés.

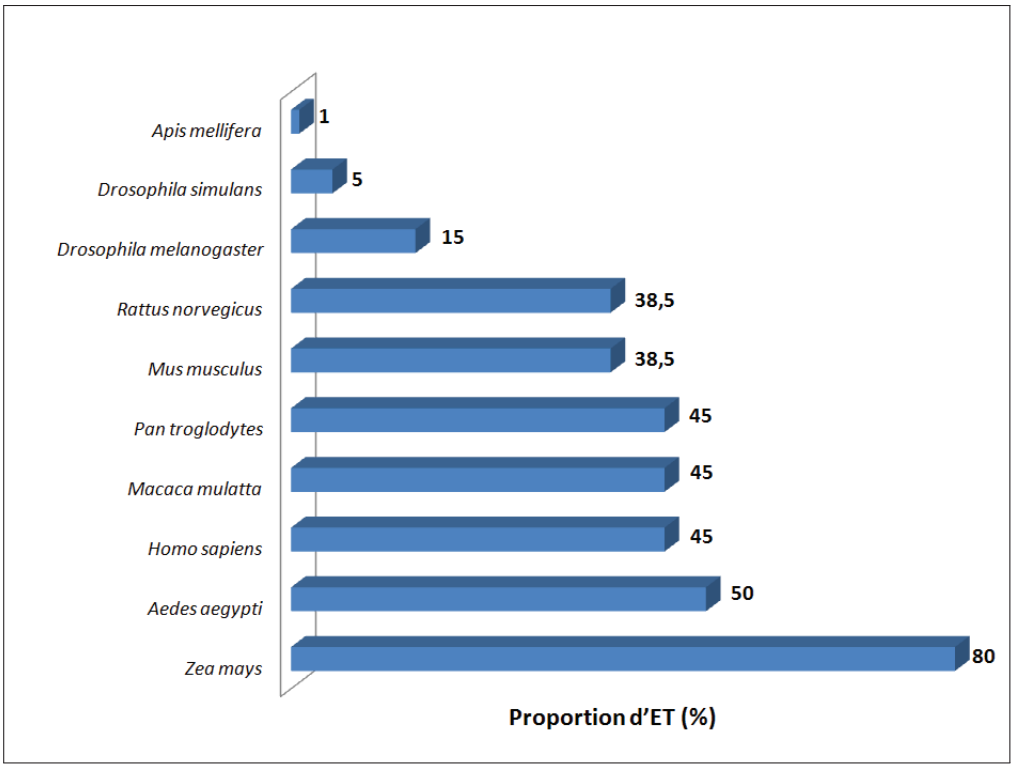

FIGURE 1: *Proportion d'éléments transposables dans dix génomes eucaryotes. Données obtenues à partir des données du séquençage de chacun de ces génomes (Aedes aegypti [Nene* et al.*, 2007], Apis mellifera [Honeybee Genome Sequencing Consortium, 2006], Drosophila simulans et D. melanogaster [Drosophila 12 Genomes Consortium , 2007], Homo sapiens [Lander* et al.*, 2001], Rattus norvegicus [Gibbs* et al.*, 2004], Zea mays [Schnable* et al.*, 2009], Mus musculus [Mouse Genome Sequencing Consortium, 2002], Pan troglodytes [Chimpanzee Sequencing and Analysis Consortium, 2005], Macaca mulatta [Rhesus Macaque Genome Sequencing and Analysis Consortium , 2007]).*

1.2 Classification des Éléments Transposables

Il existe diverses façons de classifier les ET selon les critères utilisés. Le premier système de classification des ET a été proposé par Finnegan en 1989 [Finnegan, 1989] (**Figure 2**) et utilise l'intermédiaire de transposition comme le critère principal qui permet de classifier les ET. Ainsi, les ET ont été divisés en deux classes principales :

- Les éléments de classe 1 ou rétrotransposons [Feschotte et Pritham, 2007] qui utilisent une molécule ARN comme intermédiaire de transposition. Dans ce cas, la séquence ADN de l'ET est d'abord transcrite en ARN, puis rétrotranscrite (*via* une transcriptase inverse ou RT, pour "Reverse Transcriptase") en une molécule d'ADN qui sera ensuite intégrée à une autre position du génome. On parle ainsi d'un mode de transposition en "copier-coller". Les rétrotransposons sont à leur tour divisés en deux sous-classes selon qu'ils possèdent ou non des longues répétitions terminales (LTR, pour "Long Terminal Repeat") à leurs extrémités. Les rétrotransposons à LTR ont une structure très proche de celle des rétrovirus et comportent les régions codantes nécessaires à leur transposition ainsi que des séquences régulatrices situées dans les LTR. Les rétrotransposons sans LTR sont divisés en deux groupes majeurs, les "Long Interspersed Nuclear Elements" (LINE) qui codent leur propre RT et sont de ce fait dits autonomes, et les "Short Interspersed Nuclear Elements" (SINE) qui sont non autonomes et utilisent la machinerie des LINE pour pouvoir transposer.

- Les éléments de classe 2 ou transposons à ADN [Feschotte et Pritham, 2007] utilisent, comme leur nom l'indique, une molécule d'ADN comme intermédiaire de transposition. Ils possèdent un cadre de lecture ouvert (ORF, pour "Open Reading Frame") qui code une transposase et qui est flanqué des deux côtés par des répétitions terminales inversées (TIR, pour "Terminal Inverted Repeats"). Cette transposase reconnaît la copie de l'ET au niveau de ses TIR, l'excise et l'insère ailleurs dans le génome. Ce mode de transposition en "couper-coller" est non réplicatif puisqu'il consiste en un déplacement de la même copie dans le génome.

Le développement des techniques de séquençage a permis d'accumuler une quantité importante de données génomiques dont beaucoup correspondent à des ET. Parmi ces

ET nouvellement identifiés, certains éléments ne peuvent pas être positionnés selon la classification de Finnegan. On peut citer pour exemple les *Helitrons* et les "Miniature Inverted Repeat Transposable Elements" (**MITE**) qui transposent par un mécanisme de "copier-coller" sans passer par un intermédiaire ARN. Ceci a poussé la communauté des chercheurs sur les ET à rediscuter cette classification, et un comité fut créé en avril 2006 dans ce but ("International Committee on the Classification of Transposable Elements"). Il faudra noter ici que les séquences consensus de la plupart des familles d'ET eucaryotes décrites (et les éléments répétés en général) sont regroupées dans une seule base, appellée **Repbase** [Jurka *et al.*, 2005] (`http://www.girinst.org/repbase/`) et créée en 1990. En 2007, Wicker et ses collaborateurs [Wicker *et al.*, 2007] ont proposé un nouveau système de classification des ET qui maintient les deux classes majeures de Finnegan en s'intéressant d'abord à l'intermédiaire de transposition, mais qui utilise comme critères supplémentaires les similarités de séquences et les relations de structure entre les ET. Il s'agit dans ce cas d'une classification hiérarchique qui commence par la classe de l'élément, suivie par sa sous-classe, son ordre, sa superfamille, sa famille et enfin sa sous-famille. En 2008, Kapitonov et Jurka [Kapitonov et Jurka, 2008] ont publié un article dans lequel ils affirment que la classification proposée par Wicker *et al.* est déjà celle qui est utilisée pour le regroupement des ET dans **Repbase**. En effet, les auteurs indiquent que la classification des ET utilisée dans **Repbase** prend en compte l'enzymologie et les similarités de structures et de séquences. De plus, la classification de Kapitonov et Jurka propose également une nouvelle nomenclature mentionnant le nom de la superfamille, un identifiant de structure et un identifiant d'espèce. À ce jour, la classification de Finnegan reste la plus utilisée car elle est simple bien qu'incomplète et c'est cette classification que j'utiliserai dans la suite de mon manuscrit.

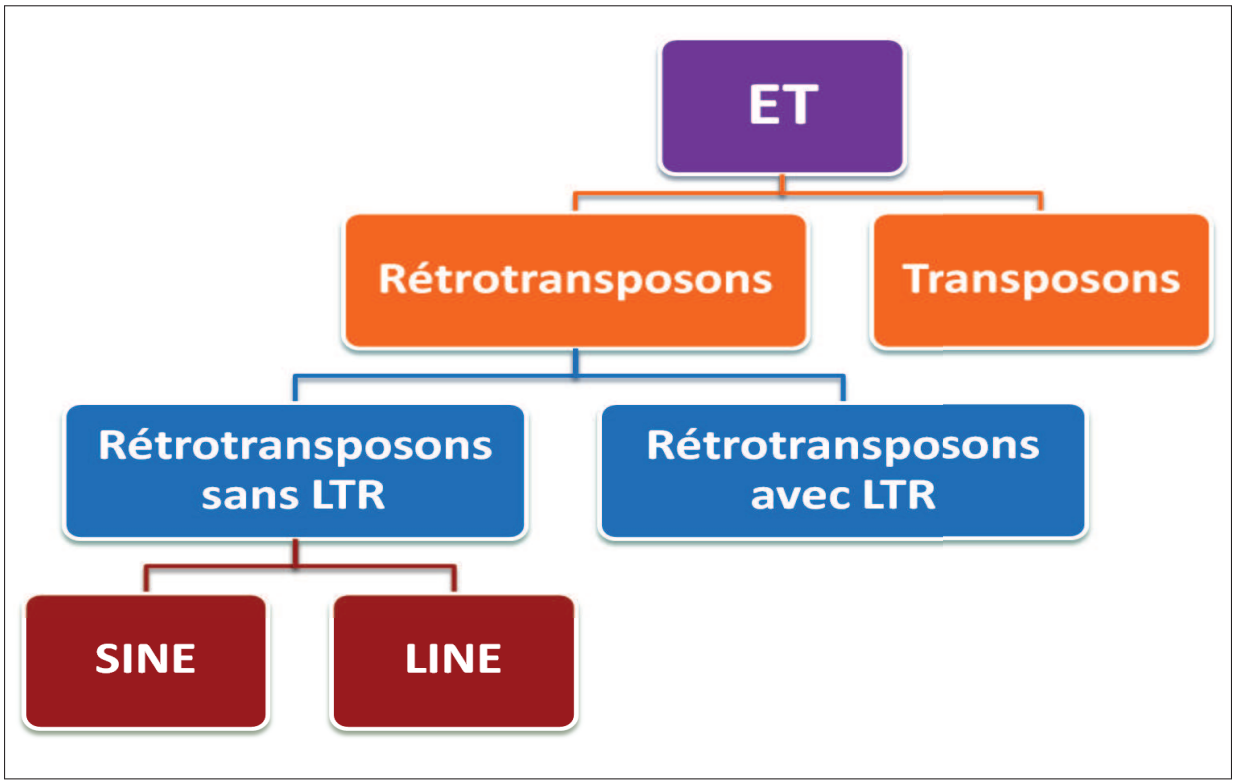

FIGURE 2: *La classification des ET selon Finnegan. Les ET sont divisés en deux classes majeures suivant l'intermédiaire de transposition. Les transposons (classe 2) utilisent un intermédiaire ADN et transposent via un mécanisme de "couper-coller" alors que les rétrotransposons (classe 1) utilisent un intermédiaire ARN via un mécanisme de "copier-coller".*

La contribution relative de chacune des classes d'ET au génome hôte est très variable. En effet, la proportion relative des transposons à ADN (égale au ratio du nombre total de transposons à ADN sur le nombre total d'ET) ainsi que celle des rétrotransposons sont variables suivant le génome considéré (**Figure 3**). Ceci est du à l'activité transpositionnelle et le mode de transposition variables entre les deux classes d'ET (réplicatif pour les rétrotransposons et conservatif pour les transposons à ADN). Ainsi chez certaines espèces comme la levure, les ET sont représentés uniquement par des rétrotransposons à LTR ce qui veut dire que ces derniers sont les seuls à avoir été actifs. La tendance inverse est observée chez le protozoaire flagellé du genre *Trichomonas* pour lequel la présence des ET se réduit à celles des transposons à ADN.

Chez l'homme, la proportion relative des rétrotransposons (42% de notre génome) est largement supérieure à celle des transposons à ADN (3% de notre génome)

[Lander *et al.*, 2001]. Ceci indique que l'activité des transposons à ADN a été (et reste aujourd'hui) très faible. Cette observation est validée par l'étude de Pace et Feschotte qui ont montré que le dernier transposon à ADN à avoir été actif chez les primates date d'au moins 40 millions d'années et que depuis aucune activité d'éléments de cette classe n'a pu être détectée [Pace et Feschotte, 2007]. Une autre étude visant à identifier les ET actifs dans le génome humain [Mills *et al.*, 2007] a montré que seuls certains rétrotransposons sont actifs. Cette étude a permis d'identifier 147 copies de *L1* (composant majeur des LINE) complètes donc potentiellement actives, plusieurs éléments *Alu* (composant majeur des SINE) et quelques éléments SVA (élément composite formé d'un SINE-R, une répétition en tandem (VNTR) et d'un élément *Alu* inséré en sens inverse). Tout ceci forme un ensemble de 40 sous-familles d'ET qui sont supposées être toujours actives dans le génome humain. Parmi ces sous-familles six appartiennent à la famille des LINE, six à la famille des SVA et 28 à la famille des *Alu*.

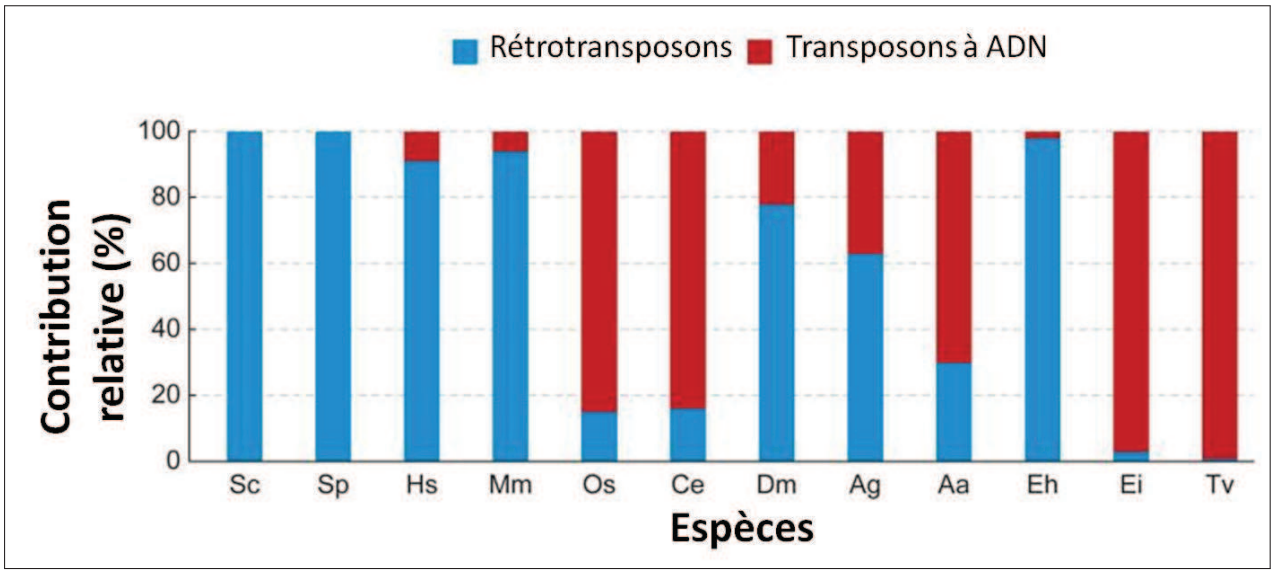

FIGURE 3: *Taux relatif des familles d'ET dans différents génomes eucaryotes. Ce graphe montre la contribution relative (en %) des transposons à ADN et des rétrotransposons au nombre total d'ET dans chacune des espèces. (Sc : Saccharomyces cerevisae ; Sp : Schizosaccharomyces pombe ; Hs : Homo sapiens ; Mm : Mus musculus ; Os : Oryza sativa ; Ce : Caenorhabditis elegans ; Dm : Drosophila melanogaster ; Ag : Anopheles gambiae ; Aa : Aedes aegypti ; Eh : Entamoeba histolytica ; Ei : Entamoeba invadens ; Tv : Trichomonas vaginalis). D'après [Feschotte et Pritham, 2007]*

1.3 Éléments Transposables et évolution des génomes

Après sa découverte des ET dans les années 50, B. Mc Clintock a émis l'hypothèse dans un article publié en 1956 [McClintock, 1956] que ces éléments avaient la capacité de re-sculpter le génome. Bien que cette idée fut très contestée, elle a accéléré les recherches sur les ET afin de mieux comprendre leurs structure, impact et fonction. Dans les années qui ont suivi, la discussion s'est surtout basée sur les rôles possibles que peuvent jouer les ET dans leurs génomes hôtes. En 1980, Doolittle et Sapienza [Doolittle et Sapienza, 1980] ont présenté l'idée du "paradigme phénotypique" de la théorie néo-Darwinienne, qui suppose que les gènes qui confèrent un avantage sélectif à leur organisme, vont assurer leur survie et leur présence dans les générations futures. À partir de là, on peut conclure que la présence des ET a pu être maintenue à travers l'évolution à cause de leur influence sur les génomes. Mais deux articles publiés par Orgel *et al.* dans la même année ([Orgel *et al.*, 1980], [Orgel et Crick, 1980]) supposent que l'émergence et la propagation des ET peuvent être expliquées uniquement par leur capacité à se répliquer dans le génome. Ainsi, selon cette hypothèse, la transmission des ET à travers les générations ne viendrait en aucun cas d'un avantage sélectif que conféraient ces éléments à leur hôte mais tout simplement de leur capacité de réplication autonome. Cette hypothèse fut validée par la démonstration théorique de la capacité des ET à se maintenir et se propager dans les populations naturelles malgré le désavantage sélectif qu'ils confèrent à ces populations [Hickey, 1982].

Les hypothèses d'Orgel et la découverte de Hickey [Hickey, 1982] ont été à l'origine de la théorie selon laquelle les ET seraient des éléments égoïstes qui se maintiennent dans les génomes malgré leurs effets négatifs, de la même manière que font les parasites. Cette théorie est connue sous le nom de "Selfish DNA" ou "Parasite DNA" [Dawkins, 2006]. Plus tard en 2002, Bowen et Jordan [Bowen et Jordan, 2002] ont argumenté que la cohérence et la logique sous-jacente de cette théorie étaient à tel point impérieuses qu'elle était incontestable à l'époque. Toutefois, l'acceptation de cette hypothèse est à l'origine de la stagnation de la recherche sur la significativité évolutive des ET.

François Jacob [Jacob, 1977] a comparé l'évolution à un bricoleur qui utilise tout le matériel disponible pour pouvoir créer de l'innovation génétique. Ainsi, les ET, abondants

et ubiquitaires, peuvent servir comme des blocs génétiques idéals que l'évolution pourrait utiliser pour "bricoler" et apporter des modifications génétiques. L'avancement des différentes méthodes de séquençage a permis de produire de plus en plus de données qui ont montré que la partie codante ne représentait qu'une partie minime du génome alors que la majeure partie était représentée par de l'ADN non-codant ou ADN poubelle ("Junk DNA"), dont les ET forment la majeure partie. Plusieurs forces, comme la dérive génétique, expliquent le maintien de tout cet ADN malgré le fait qu'il soit non-codant, mais certains ET sont également maintenus par la sélection. En effet, plusieurs exemples associent les ET à des rôles importants dans le génome (plusieurs de ces exemples seront montrès ultérieurement). De plus, l'hypothèse du maintien des ET par la sélection est vérifiée par la découverte de Cavalier-Smith et Beaton en 1999 [Cavalier-Smith et Beaton, 1999] qui ont mis en évidence l'absence complète d'ET dans le génome du nucléomorphe (petit noyau qui se trouve dans les chryptophytes) suite à la pression sélective empêchant l'insertion des ET dans ce génome, ce qui indique que les ET sont éliminés du génome hôte quand ils sont associés à des effets délétères. Ainsi, l'implication des ET dans l'évolution du génome est devenue de plus en plus indiscutable malgré leur nature parasitaire [Fedoroff, 1999].

Les ET sont actuellement reconnus comme des acteurs majeurs dans l'évolution de leurs génomes hôtes ([Oliver et Greene, 2009], [Biémont et Vieira, 2006]) et sont considérés comme responsables, en partie, de la grande diversité phénotypique qu'on observe, puisqu'ils peuvent causer des changements génétiques très variables [Böhne *et al.*, 2008]. Leur implication dans l'évolution peut avoir lieu de diverses façons ([Kazazian, 2004], [Deininger *et al.*, 2003]) qui seront détaillées plus loin, cependant il est important de s'intéresser auparavant aux caractéristiques qui font que les ET sont hautement propices à devenir des agents impliqués dans l'évolution, à la fois génotypique et phénotypique, de leur hôte.

1.3.1 Caractéristiques des Éléments Transposables

Les ET possèdent cinq traits importants qui leur confèrent une capacité à jouer un rôle important dans l'évolution de leur génome hôte :

1. **Mobilité** : La majorité des ET possèdent la capacité de se déplacer dans le génome.

Ce déplacement peut se faire *via* des protéines codées par l'élément lui-même, on parle dans ce cas d'éléments autonomes. D'autres éléments n'ont pas la machinerie nécessaire à leur déplacement, d'où le détournement de la machinerie des éléments autonomes pour promouvoir leur déplacement, comme par exemple les SINE.

2. **Universalité** : Les ET ont été trouvés dans la plupart des espèces, allant des bactéries jusqu'aux mammifères. Leur nature ubiquitaire vient de leur capacité à se disséminer dans un génome mais également à coloniser d'autres génomes par transferts horizontaux entre espèces : par exemple, le transposon à ADN, SPIN, a été transféré des tétrapodes à des mammifères [Pace *et al.*, 2008].
3. **Ancienneté** : L'origine des ET apparaît être très ancienne. En effet, on soupçonne aujourd'hui que certaines familles d'ET (transposons à ADN) étaient déjà présentes chez l'ancêtre commun des eucaryotes [Feschotte et Pritham, 2007]. De plus, la transcriptase inverse des introns de groupe II serait à l'origine de celle des rétotransposons à LTR et sans LTR [Capy *et al.*, 1997].
4. **Abondance** : Les ET peuvent représenter une part très importante des génomes eucaryotes. Les différents projets de séquençage ont confirmé cette tendance avec une proportion d'ET pouvant former la moitié du génome chez certains primates, voire plus de 80% du génome chez certaines plantes. Ainsi, les ET jouent un rôle important dans la taille du génome hôte [Kidwell, 2002]. Par contre, chez les bactéries, les nématodes et certains insectes les ET semblent constituer une plus faible proportion des génomes ($< 20\%$).
5. **Mutagènes** : Les ET sont des mutagènes puissants [Belancio *et al.*, 2008]. Les mutations provoquées par les ET peuvent être délétères, d'où leur élimination par la sélection naturelle, ou bien bénéfiques et donc conservées par la sélection. Plusieurs exemples de mutations bénéfiques apportées par les ET ([Bourque, 2009], [Muotri *et al.*, 2007]) seront présentés plus loin.

Si on se base sur le principe que l'évolution des génomes et des espèces implique trois mécanismes majeurs, que sont la duplication génique, l'évolution des séquences codantes et l'évolution des régions régulatrices [Carroll, 2005], on peut considérer que les ET peuvent être impliqués dans chacun de ces trois mécanismes jouant ainsi un rôle important dans l'évolution de leur génome hôte.

1.3.2 Implication des Éléments Transposables dans l'évolution des génomes

La participation des ET à l'évolution de leur génome hôte peut se faire de diverses façons. Tout d'abord les ET peuvent participer directement à la structure génomique (génique plus spécifiquement) de leur hôte. Ainsi la séquence d'un ET peut être "détournée" par le génome et assurer une fonction spécifique. Ce processus est appelé domestication [Gould et Vrba, 1982]. Plusieurs exemples de domestication d'ET sont connus aujourd'hui [Volff, 2006] parmi lesquels je ne citerai que deux exemples pour leur importance significative. Le premier exemple de domestication le plus connu chez l'homme, est celui du système de recombinaison V(D)J des immunoglobulines. Ce système est un mécanisme qui consiste en une recombinaison spécifique des sites V, D et J nécessaires pour l'assemblage de régions variables des récepteurs des cellules B (ou immunoglobulines) avec les récepteurs des cellules T provenant de plusieurs segments de gènes différents. La protéine RAG1 (RAG, pour "Recombinase Activating Genes") active cette recombinaison et le complexe protéique RAG1/RAG2 permet de l'initier. En effet, ce couple fonctionne comme une endonucléase introduisant des coupures doubles brin proches des signaux spécifiques de recombinaison. Ceci supporte l'hypothèse selon laquelle le complexe RAG1/RAG2 aurait comme origine une transposase, il y a 500 millions d'années [Sakano *et al.*, 1979]. De plus, la présence du motif DDE (présent au niveau de certaines transposases) au niveau de RAG1 supporte cette hypothèse. Enfin, une analyse récente du gène RAG1 a montré qu'il possède une forte similarité de séquence avec les éléments de type *transib*, présents chez les insectes et les nématodes, et dont le mode de transposition est de type "couper-coller" [Fugmann, 2010]. Le deuxième exemple de domestication concerne l'élongation des extrémités des chromosomes chez la drosophile. En effet, à la fin de chaque division mitotique, une télomèrase allonge les extrémités des chromosomes, un mécanisme qui existe chez plusieurs espèces. Cependant, chez *Drosophila melanogaster* cette élongation se fait par l'insertion de deux rétrotransposons sans LTR, *TART* et *Het-A*, qui par l'intermédiaire de leurs transpositions successives vont préserver l'intégrité des télomères [Pardue *et al.*, 2005].

Les ET sont également une source d'exons (complets ou partiels) de plusieurs gènes codant pour des protéines fonctionnelles [Britten, 2006]. Le mécanisme le plus

plausible pour expliquer cette "exonisation" des ET suppose que ces éléments vont tout d'abord s'insérer dans les introns des gènes (où la contrainte sélective est moins forte par rapport aux exons) et que cette insertion sera suivie par des mutations qui vont modifier l'épissage du gène incluant ainsi une partie (ou la totalité) de l'ET dans la protéine [Nekrutenko et Li, 2001]. Cette modification d'épissage est facilitée par la présence de plusieurs sites donneurs et accepteurs d'épissage au niveau de certaines familles d'ET, comme c'est le cas pour les *Alu* dont la participation aux exons se fait majoritairement, voire totalement, par épissages alternatifs [Sorek *et al.*, 2002]. Les rétrotransposons à LTR sont rarement soumis à cette "exonisation" dans des gènes humains [Piriyapongsa *et al.*, 2007b]. Enfin, les ET sont également à l'origine de plusieurs séquences extragéniques chez différentes espèces. Ainsi ils sont à l'origine des centromères chez certaines plantes, des télomères chez certains protozoaires et des régions S/MARs ("Scaffold/Matrix Associated Regions"), séquences régulatrices qui divisent le génome humain en régions distinctes ayant différentes conformations chromatiniennes [von Sternberg et Shapiro, 2005].

En plus de leur contribution structurale, les ET peuvent participer à la régulation des gènes qui sont dans leur voisinage. En effet, Jordan *et al.* (2003) ont montré que pas moins de 25% des régions promotrices contiennent des éléments qui dérivent des ET, et que les régions LCRs ("Locus Control Regions") et S/MARs, qui sont impliquées dans la régulation simultanée de plusieurs gènes, contiennent aussi plusieurs séquences qui dérivent des ET [Jordan *et al.*, 2003]. De plus, les ET possèdent dans leurs séquences des sites de fixation de divers facteurs de transcription [Bourque *et al.*, 2008] et peuvent être à l'origine de plusieurs sites hypersensibles à la DNaseI (régions régulatrices) [Mariño-Ramírez et Jordan, 2006]. La régulation ET-dépendante des gènes peut également se faire loin de ces derniers puisque les ET peuvent être à l'origine d'enhancer comme un élément SINE qui régule l'activité du gène *ISL1* depuis 410 millions d'années (avant la divergence des tétrapodes) [Bejerano *et al.*, 2006]. Enfin, les ET peuvent participer indirectement à la régulation des gènes *via* les micro-ARN (miRNAs). Les miRNAs sont des petits ARN (22 à 26 nucléotides) non codants qui régulent l'expression des gènes. Cette régulation se fait grâce à une complémentarité partielle entre le miRNA et la région 3' non traduite (UTR pour "UnTranslated Region")

de l'ARN messager (ARNm) du gène cible. Le complexe miRNA-RISC ("RNA-Induced Silencing Complex") fixe la molécule d'ARNm cible provoquant ainsi l'inhibition de la traduction de cet ARNm ou sa dégradation. Il a été montré que plusieurs miRNA dérivent des séquences d'ET [Piriyapongsa *et al.*, 2007a].

Les ET peuvent promouvoir le déplacement de certaines séquences géniques à un autre endroit du génome grâce à leur activité transpositionnelle. Ce mécanisme est connu sous le nom de transduction. Ce mécanisme a été mis en évidence tout d'abord sur des lignées cellulaires humaines par l'intermédiaire des éléments *L1* [Moran *et al.*, 1999]. La région transduite est située en 3' des *L1* et peut contenir des promoteurs, des enhancers ou des exons, ce qui permet de créer de nouveaux gènes ou de modifier la régulation de gènes préexistants. Plusieurs exemples existent comme par exemple la transduction de GLUD1 (gène de ménage exprimé dans plusieurs tissus) chez l'ancêtre commun des hominoïdes (23 millions d'années) qui est à l'origine du gène GLUD2 (expression spécifique dans les tissus nerveux et les testicules) [Burki et Kaessmann, 2004].

Enfin, le dernier mécanisme par lequel les ET peuvent influencer l'évolution de leur génome hôte est la recombinaison. En effet, du fait de leur grand nombre de copies, des recombinaisons entre régions homologues d'ET peuvent avoir lieu. Ces recombinaisons peuvent se faire entre des éléments situés sur un même chromosome, engendrant ainsi une duplication, une délétion ou des inversions des séquences situées entre les ET ; elles peuvent avoir lieu aussi entre des éléments situés sur des chromosomes différents ayant pour conséquence des translocations chromosomiques ou des réarrangements chromosomiques plus complexes.

Plusieurs études réalisées dans les génomes de l'homme et du chimpanzé ont montré des évènements de délétion et d'inversion suite à une recombinaison entre ET, dont la quasi totalité fait référence à une recombinaison entre les éléments *Alu*. Ce biais peut être expliqué tout d'abord par le grand nombre d'*Alu* présents dans le génome humain ($> 10\%$ du génome), mais également par la tendance de ces éléments à s'accumuler dans les régions riches en gènes [Jurka, 2004] et enfin par la présence de "points chauds" de recombinaison au niveau du monomère gauche des séquences *Alu* [Burwinkel et Kilimann, 1998].

Ainsi, les éléments *Alu* seraient à l'origine de duplications dans le génome humain [Bailey *et al.*, 2003], et seraient responsables de la translocation entre les chromosomes 11 et 22 de l'homme [Hill *et al.*, 2000]. Néanmoins, des recombinaisons entre les autres types d'ET ont été rapportées, comme par exemple l'inversion sur le chromosome Y de l'homme médiée par une recombinaison entre deux LINE [Schwartz *et al.*, 1998]. Chez *Drosophila virilis*, deux éléments de type rétrotransposons, *Penelope* (sans LTR) et *Ulysses* (avec LTR) montrent une insertion préférentielle au niveau de sites d'inversions, et sont ainsi à l'origine de divers réarrangements chromosomiques (inversions, translocations et délétions). On suppose que ces réarrangements induits par les ET jouent un rôle important dans l'évolution des espèces du groupe *virilis* [Evgen'ev *et al.*, 2000]. Enfin, chez les plantes, la recombinaison est majoritairement limitée aux gènes, ce qui peut expliquer le faible taux de recombinaison entre les ET [Bennetzen, 2000]. Ainsi, l'exemple le plus connu chez les plantes est celui de la recombinaison entre les LTR des rétrotransposons, qui est à l'origine de la suppression des séquences situées entre les deux LTR et la formation des solo-LTR [Tenaillon *et al.*, 2010]. Ce mécanisme n'est pas spécifique aux plantes mais concerne tous les rétrotransposons à LTR, quelle que soit l'espèce considérée.

1.4 Éléments Transposables et maladies humaines

L'activité des ET dans le génome humain peut être également à l'origine d'altérations génomiques. En effet, l'insertion et/ou la recombinaison des ET sont capables d'altérer l'activité des gènes voisins, cette altération aboutissant parfois à des maladies.

L'insertion des ET peut altérer la fonction des gènes de diverses façons. Cette insertion peut se faire soit au niveau de la région promotrice du gène (ou au niveau des régions régulatrices plus distantes telles que les "enhancers" et les "silencers") modifiant ainsi l'activité de ces régions régulatrices et par la suite l'activité des gènes. L'insertion des ET peut également avoir lieu au niveau des exons provoquant ainsi un décalage du cadre de lecture ("frameshift") ou une mutation non sens (codon stop). Elle peut enfin avoir lieu au niveau de la jonction correspondante à la fin de l'exon/début de l'intron, aboutissant ainsi à une modification du mécanisme d'excision/épissage lors de la maturation des

ARN messagers. Ces différentes insertions aboutissent à la formation d'une protéine tronquée (donc non fonctionnelle). Par exemple, l'insertion d'un élément *Alu* dans le gène BRCA2, suppresseur de tumeurs (27 exons ; 70 kb de séquence génomique), a provoqué une modification de l'excision/épissage, qui dans ce cas saute l'exon 22 et aboutit à un décalage du cadre de lecture. Ce décalage provoque une fin prématurée de la traduction de la séquence donnant ainsi une protéine tronquée plus courte que la protéine normale. Il semble que cette insertion soit l'une des causes de l'apparition d'un cancer du sein chez une patiente [Miki *et al.*, 1996]. De la même façon, l'insertion d'une partie d'un élément *L1* dans un exon au milieu du gène APC, considéré comme suppresseur de tumeurs associées aux développement des maladies colorectales, a ajouté une mutation non sens (codon stop) aboutissant ainsi à une protéine tronquée non fonctionnelle qui a favorisé l'apparition d'un cancer du colon chez un patient [Miki *et al.*, 1992]. Enfin, l'insertion de 632 pb d'un élément *SVA* au niveau de l'exon 5 du gène α-spectrin conduit à son épissage provoquant la formation d'une protéine tronquée et favorisant ainsi l'apparition de la maladie de l'elliptocytose héréditaire [Ostertag *et al.*, 2003].

La recombinaison entre les régions homologues des ET peut également altérer la structure de l'ADN génomique provoquant ainsi l'apparition de maladies. La plupart des maladies associées à des ET sont provoquées par des recombinaisons entre les éléments *Alu*, les recombinaisons entre ces derniers étant plus fréquentes que celles entre les *L1* (comme expliqué auparavant) et celles entre les autres classes d'ET dont le nombre de copies est largement inférieur. On estime que les recombinaisons *Alu/Alu* sont à l'origine de 33 mutations dans les cellules germinales et de 16 dans les cellules somatiques, ce qui représente environ 0,3% des maladies humaines [Deininger et Batzer, 1999]. D'un autre côté, on n'a pu identifier que deux maladies dues à la recombinaison entre deux éléments *L1*.

Il est évident qu'on ne peut pas savoir si l'activité transpositionnelle de l'ET est à l'origine ou la conséquence de l'apparition des maladies. Néanmoins il est clair que cette activité joue un rôle majeur dans ces maladies.

1.5 Distribution des Éléments Transposables dans les génomes

L'activité délétère des ET est à l'origine de leur contrôle génomique. Un des moyens simples permettant de réduire les effets des ET est d'organiser leur distribution. Ainsi, l'insertion des ET dans les régions pauvres en gènes du génome aura très probablement moins d'effets délétères par rapport à leur insertion dans les régions riches en gènes. Plusieurs études se sont intéressées à la distribution des ET dans les génomes. Ces études ont mis en évidence que l'abondance des ET varie localement et que la densité en ET varie entre les chromosomes, voire même entre les bras d'un seul chromosome. Ainsi, une étude faite chez l'homme sur les chromosomes 21 et 22 a montré que les éléments *Alu* représentent 40% de la taille totale de ces deux chromosomes et que le chromosome 22 a deux fois plus d'*Alu* que le chromosome 21 [Grover *et al.*, 2003]. De la même façon, différentes études menées sur plusieurs chromosomes du riz ont montré un pourcentage variable d'ET selon le chromosome considéré. En effet, les ET forment ∼18% du chromosome 4 [Feng *et al.*, 2002], ∼16% du chromosome 10 [Rice Chromosome 10 Sequencing Consortium, 2003] et ∼12% du chromosome 1 [Sasaki *et al.*, 2002]. De même, on observe chez la drosophile un large excès d'ET sur le chromosome 4 (∼7%) par rapport aux chromosomes 2R (∼2%) et X (∼1%) [Bartolomé *et al.*, 2002]. Cette abondance varie également suivant les différentes classes ou familles d'ET. On observe par exemple, un enrichissement deux fois plus grand des éléments *L1* sur le chromosome X humain par rapport aux autosomes [Bailey *et al.*, 2000]. Les auteurs expliquent cet enrichissement par une implication des éléments *L1* dans l'inactivation du chromosome X. De la même façon, l'étude de la distribution des différentes familles d'ET chez *C. elegans* a permis de mettre en évidence des différences d'abondance de diverses familles d'ET selon le chromosome considéré ([Rizzon *et al.*, 2003], [Surzycki et Belknap, 2000]) : certaines familles d'ET sont surreprésentées sur le chromosome X (par rapport aux autosomes) alors que d'autres présentent la tendance opposée [Duret *et al.*, 2000].

Toutes ces études montrent que les ET ne sont pas uniformément distribués dans un génome donné. Il existe donc une organisation importante de la distribution des ET dans

le génome. Cette organisation explique leurs insertions adjacentes (ou "clustering") mais également leurs insertions dans d'autres ET (ou "nesting"). Les facteurs qui influencent la distribution des ET sont souvent liés aux caractéristiques structurales de la région d'insertion. Les trois facteurs majeurs sont : (1) l'état de la chromatine, (2) la richesse en gènes et taux de GC des sites d'insertion (3) le taux de recombinaison.

État de la chromatine

La chromatine est par définition l'ensemble ADN-protéines. Cette chromatine existe à différents niveaux de compaction : l'euchromatine qui représente la forme non-condensée de la chromatine et qui est associée à une activité transcriptionnelle (c'est-à-dire que les gènes présents dans cette région sont actifs) et l'hétérochromatine qui représente la forme très condensée associée à une répression transcriptionnelle (gènes inactifs). Plusieurs études ont montré un enrichissement significatif des ET au niveau des régions hétérochromatiques des chromosomes par rapport aux régions euchromatiques. Ainsi, les ET représentent 60% de l'hétérochromatine chez *Anopheles gambiae* alors qu'ils ne forment que 16% de l'euchromatine [Holt *et al.*, 2002]. De la même façon, Rizzon *et al.* ont montré que les ET de *Drosophila melanogaster* avaient tendance à s'accumuler dans les régions péricentromériques et centromériques (toutes les deux de nature hétérochromatinienne) [Rizzon *et al.*, 2002]. Hoskins *et al.* ont montré que les ET représentent plus de 50% de l'hétérochromatine de *D. melanogaster* [Hoskins *et al.*, 2002]. Cette même tendance a été observée également chez les plantes chez qui une analyse faite chez *Arabidopsis thaliana* a permis de mettre en évidence des blocs hétérochromatiniens dans la région euchromatique. Ces blocs sont essentiellement composés d'ET, laissant les auteurs supposer que ce phénomène d'hétérochromatinisation est la conséquence de l'insertion des ET [Lippman *et al.*, 2004]. Enfin, même si ces études restent limitées chez les vertébrés, on a pu montrer que les ET chez *Tetraodon nigroviridis* sont groupés dans les parties hétérochromatiques des chromosomes [Dasilva *et al.*, 2002]. Ainsi, l'enrichissement significatif des ET dans l'hétérochromatine peut être expliqué par l'action de la sélection qui élimine les insertions d'ET dans les régions euchromatiniennes alors que les insertions neutres d'ET dans l'hétérochromatine sont maintenues. À ceci s'ajoute également le phénomène d'hétérochromatinisation induit par l'insertion d'une séquence d'ET dans l'euchromatine.

Richesse en gènes et taux de GC des sites d'insertion

La notion de richesse en gènes suit parfaitement l'état de la chromatine dans le sens où les régions hétérochromatiques sont globalement pauvres en gènes à l'opposé des régions euchromatiques qui en sont riches. Ainsi on pourrait supposer une insertion dirigée des ET dans les régions pauvres en gènes. Ceci est globalement vrai mais il existe quelques exceptions. Ainsi, on a montré chez l'homme que les *Alu* ont tendance à s'accumuler dans les régions riches en GC alors que les *L1* s'accumulent dans les régions riches en AT et que les rétrotransposons à LTR possèdent une tendance intermédiaire. De plus, les rétrotransposons à LTR et les *L1* sont exclus des régions géniques alors que les *Alu* sont surreprésentés dans le voisinage et à l'intérieur des gènes [Medstrand *et al.*, 2002]. Ces variations d'accumulation des familles d'ET entre les différentes régions peuvent être la conséquence de l'action de la sélection, qui par exemple élimine les insertions des éléments *L1* dans les régions riches en gènes alors que les insertions neutres de ces éléments dans les régions non codantes sont maintenues. Il est important de noter ici qu'on verra un peu plus loin que l'hypothèse qui associe la sélection à la variation d'insertion des éléments *Alu* et *L1* selon la richesse en GC n'est pas forcément vraie. Enfin, différentes études montrent une accumulation des ET dans les introns des gènes chez les animaux et dans les régions intergéniques des plantes ([SanMiguel *et al.*, 1996], [Wong *et al.*, 2000]). Même si ces résultats ne sont pas observés pour toutes les familles d'ET, au vue de ce qu'on connaît actuellement chez l'homme, il a été montré que les ET représentent plus de 50% des introns des gènes situés dans l'hétérochromatine de la drosophile alors que leur participation aux introns des gènes euchromatiques de cette espèce ne dépasse pas les 0,1% [Dimitri *et al.*, 2003].

Taux de recombinaison

L'accumulation des ET dans les régions à faible taux de recombinaison chez certaines espèces laissent supposer une localisation préférentielle des ET dans ces régions. Chez l'homme, l'étude de la distribution des paires d'éléments *Alu* en fonction de leur identité de séquences, orientation et distance par rapport aux gènes voisins, a montré que les paires d'*Alu* qui sont relativement identiques ont tendance à s'insérer dans la même orientation quand elles sont proches d'un gène [Stenger *et al.*, 2001]. Les auteurs expliquent ce

résultat comme le moyen d'induire plus de stabilité génique en minimisant la probabilité de recombinaison entre les deux *Alu*. Chez la drosophile, une seule étude a mis en évidence une forte association entre l'abondance des ET et le faible taux de recombinaison [Bartolomé *et al.*, 2002], résultats qui ont été restreints à la classe II des ET (transposons à ADN) [Rizzon *et al.*, 2002]. Tout ceci peut s'expliquer par une action de la sélection qui élimine les insertions d'ET dans les régions à fort taux de recombinaison à cause des effets délétères induits par la recombinaison entre les ET dans ces régions.

Les facteurs influençant la distribution des ET sont en réalité plus nombreux que les trois que je viens de citer. On peut par exemple rajouter l'âge de l'ET comme un facteur supplémentaire puisqu'on a trouvé que les "jeunes" éléments *Alu* montrent une localisation importante dans les régions riches en AT, alors que les vieux *Alu* sont plus fréquents dans les régions riches en GC [Medstrand *et al.*, 2002], ceci montre que les *Alu* ne ciblent pas les régions riches en GC comme cela a été proposé avant. De même, on sait que la fonction des gènes joue un rôle important dans l'accumulation ou non des ET dans le voisinage des gènes (chapitre 1 pour plus de détails) ainsi que leur niveau d'expression [Sironi *et al.*, 2006]. L'explication de la distribution des ET restera identique quel que soit le facteur impliqué. Si un ET (quelles que soient sa classe et sa famille) s'insère dans le voisinage d'un gène, il risque d'altérer la fonction du gène de diverses manières (altération de la régulation, épissage, niveau d'expression, etc.). La sélection tendra à éliminer cette insertion et interdire sa fixation. À l'opposé, si l'insertion d'un ET apporte quelque chose de bénéfique au génome par le biais des caractéristiques diverses que possèdent les ET (éléments de régulation, sites de fixation de facteurs de transcription, etc.), la sélection va favoriser la fixation de cette insertion même si elle se fait dans le voisinage des gènes.

La **Figure 4** indique les trois moyens par lesquels les ET peuvent induire des effets délétères :(i) la recombinaison ectopique, d'où leur présence dans les régions à faible taux de recombinaison ; (ii) l'interruption des gènes, d'où leur présence dans les régions les moins denses en gènes (iii) l'"intoxication" des gènes hôtes par les produits des ET, d'où leur présence dans les régions ayant un faible niveau d'expression. L'hétérochromatine possède les caractéristiques de ces trois régions, d'où l'accumulation des ET dans l'hétérochromatine. Il est important de noter que le cas opposé a été

également rapporté dans le sens où l'insertion des ET dans une région donnée est à l'origine d'une hétérochromatinisation locale au niveau de cette dernière.

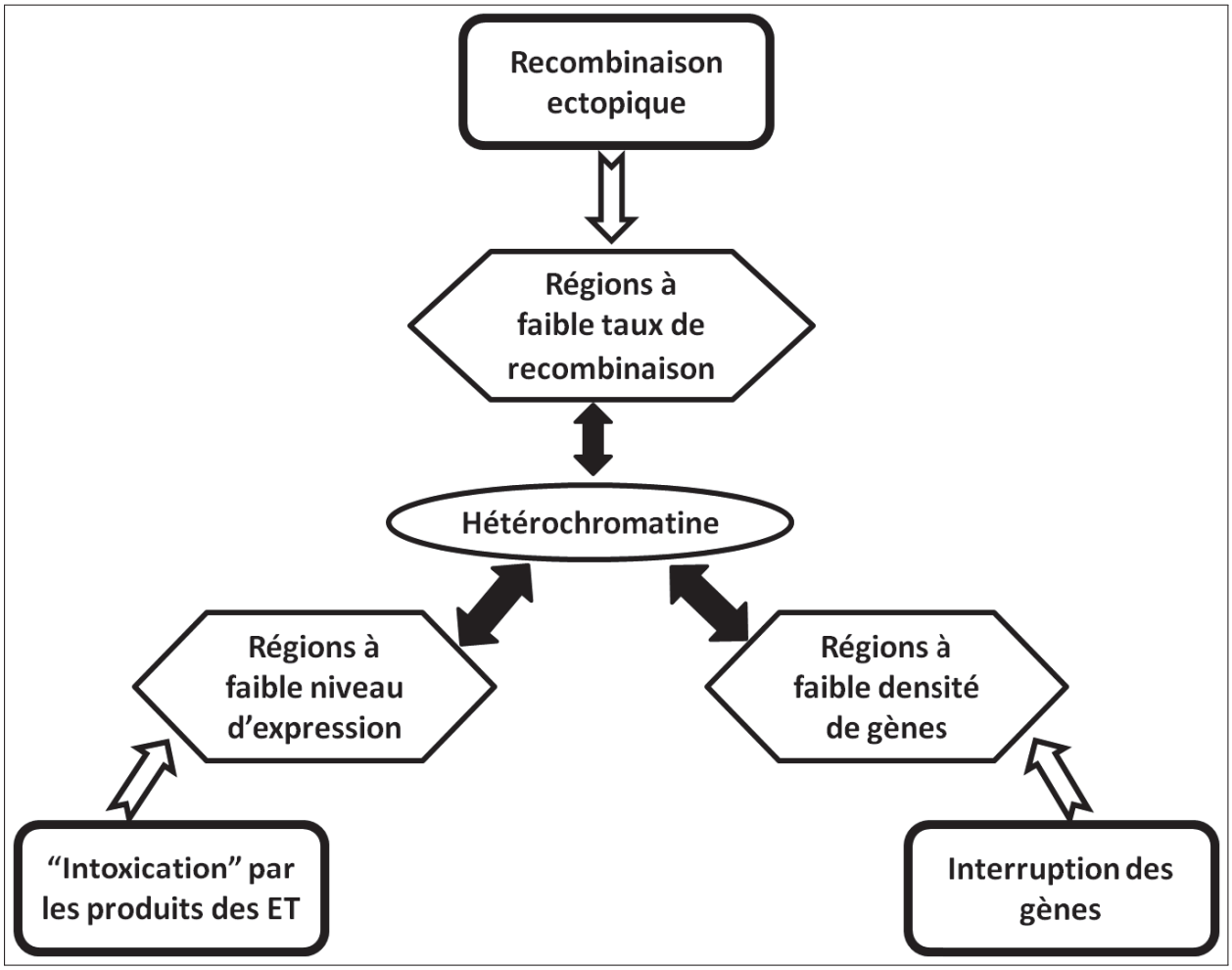

FIGURE 4: *Modèles d'accumulation des ET dans certaines régions chromosomiques. Cette figure montre les trois principaux modèles (rectangles) qui expliquent l'insertion des ET dans les différentes régions des chromosomes (octogones). D'après [Hua-Van* et al.*, 2005].*

1.6 "Silencing" des Éléments Transposables

Étant donné les divers moyens par lesquels les ET peuvent altérer l'activité du génome hôte, ils sont soumis à plusieurs mécanismes de régulation de leur activité, ce qui contre leur prolifération et réduit leur impact. Parmi ces différents mécanismes, je m'intéresserai particulièrement à la régulation épigénétique des ET qui apparaît comme le moyen le plus efficace pour réprimer leur activité.

1.6.1 Régulation épigénétique des Éléments Transposables

Définition de l'épigénétique

Le terme épigénétique a été introduit pour la première fois, par Conrad Waddington en 1942. Cet auteur a défini l'épigénétique comme étant "la branche de la biologie qui étudie les relations de cause à effet entre les gènes et leurs produits, faisant apparaître le phénotype" [Waddington, 1942], ce qui signifie que le terme épigénétique fait référence à toutes les voies moléculaires qui sont responsables de l'apparition d'un phénotype particulier à partir du génotype. Cette définition est restée inchangée pendant plusieurs décennies jusqu'en 1987, quand Robin Holliday a avancé que l'épigénétique serait plutôt associée aux changements de l'activité des gènes au cours du développement [Holliday, 1987]. Sept ans plus tard, Holliday apporte deux variations majeures à sa vision de l'épigénétique [Holliday, 1994]. Tout d'abord, il avance que les changements dans l'expression des gènes peuvent également se faire à l'âge adulte (et non seulement au cours du développement comme avancé auparavant) incluant ainsi la notion d'héritabilité, *via* la division mitotique, des modèles d'expression des gènes entre les cellules différenciées. Il ajoute également que la transmission de cette information épigénétique se fait au travers des générations. Aujourd'hui, on définit l'épigénétique comme "l'étude des changements dans la fonction des gènes qui sont mitotiquement et/ou méïotiquement héritables et qui n'entraînent pas de changements dans la séquence ADN sous-jacente" [Wu et Morris, 2001]. L'épigénétique peut donc expliquer pourquoi plusieurs cellules ayant le même contenu génomique (même ADN) peuvent avoir différentes fonctions, comme dans l'exemple d'une cellule souche qui donne plusieurs cellules différenciées impliquées dans diverses fonctions.

L'ADN des eucaryotes est organisé en unités fonctionnelles appelées nucléosomes, dont chacune est formée d'un coeur de protéines d'histones autour duquel s'enroule 146 (ou 147) nucléotides. Le nucléosome est répété tous les 200 nucléotides et s'associe avec divers facteurs nucléaires pour former une structure plus complexe, la chromatine [Luger, 2003]. La chromatine possède plusieurs niveaux de structure qui commencent par une structure primaire (la plus simple) représentée par l'arrangement linéaire des nucléosomes. Les nucléosomes s'associent entre eux pour former une structure secondaire qu'on appelle les fibres de 30 nm. Les structures secondaires peuvent à leur tour s'associer pour former des structures tertiaires encore plus complexes [Woodcock et Dimitrov, 2001]. Les différentes modifications épigénétiques (**Figure 5**) vont interagir avec des protéines pour définir différentes conformations de la chromatine régulant ainsi l'accès des facteurs de transcription à l'ADN. Les modifications épigénétiques incluent trois composants majeurs : la méthylation de l'ADN, les modifications d'histones et l'interférence par les petits ARN. On peut également y ajouter les protéines qui permettent de remodeler la structure de la chromatine [Sala et Corona, 2008] comme par exemple le complexe ATP-dépendant SWI/SNF, capable de catalyser le remodelage de la structure du nucléosome pour permettre aux facteurs de transcription de se fixer au niveau des régions promotrices [Côté *et al.*, 1994]. Les mécanismes épigénétiques sont impliqués dans plusieurs processus importants chez les mammifères [Biémont, 2010b].

Il a été démontré que l'environnement peut modifier les marques épigénétiques d'un génome [Bollati et Baccarelli, 2010]. Ainsi, une variété de produits toxiques (hydrocarbones, dioxine, pesticides, etc.) ainsi que d'autres facteurs de l'environnement peuvent agir épigénétiquement sur l'expression des gènes et des ET, pouvant aller jusqu'à provoquer l'apparition de maladies comme le cancer [Claes *et al.*, 2010]. En effet, une des caractéristiques du cancer est l'inactivation des gènes suppresseurs de tumeurs par des modifications épigénétiques (méthylation de l'ADN et modifications d'histones). Ces modifications étant réversibles, on peut utiliser des molécules chimiques qui les suppriment, rétablissant ainsi l'activité des gènes réprimés.

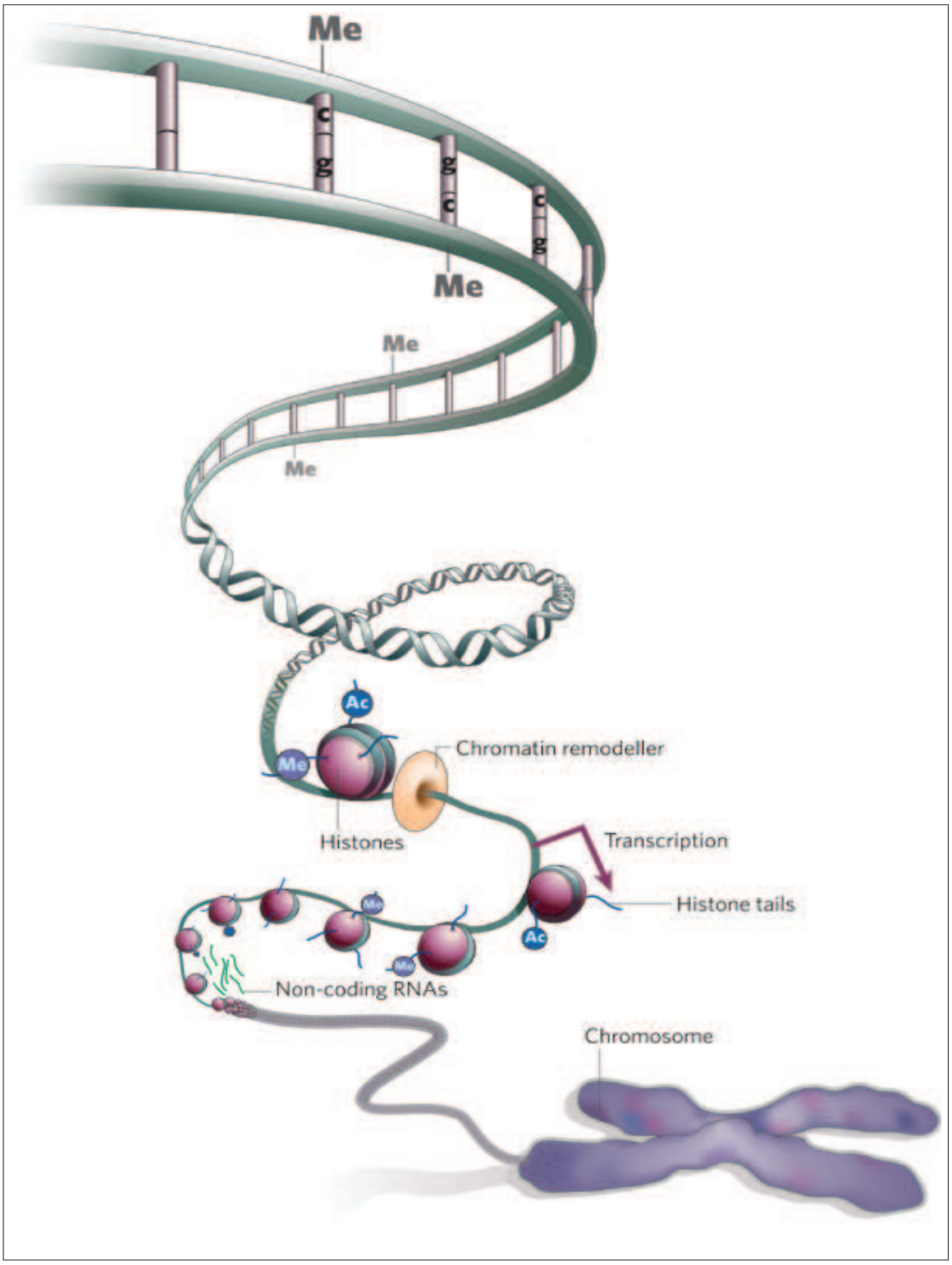

FIGURE 5: *Représentation schématique de la molécule d'ADN. Cette figure montre la structure de l'ADN eucaryotique ainsi que les trois modifications épigénétiques majeures : la méthylation de la molécule d'ADN, la modification des queues des molécules d'histones et les molécules d'ARN non codantes. Histone tails = queues d'histones ; Chromatin remodeller = remodeleur de la chromatine. Figure extraite de [American Association for Cancer Research Human Epigenome Task Force, 2008].*

Méthylation de l'ADN

La méthylation de l'ADN est l'opération qui consiste à ajouter un groupement méthyl (CH3) au niveau du Carbone 5 de la cytosine. Il existe deux types de méthylation : la méthylation symétrique qui a lieu sur les résidus CpG (Cytosine-phosphate-Guanine) et la méthylation asymétrique qui peut avoir lieu dans différents contextes nucléotidiques (CNG ou CNN, N correspondant à A, C ou T). Pendant longtemps, on pensait que la méthylation chez les vertébrés était exclusivement symétrique et que la méthylation asymétrique existait uniquement chez les plantes [Bernstein *et al.*, 2007]. Cependant l'établissement des méthylomes de deux types de cellules humaines a permis de mettre en évidence que ∼25% de la méthylation globale des cellules souches embryonnaires est asymétrique et que ce type de méthylation disparaît dans les cellules différenciées tels les fibroblastes [Lister *et al.*, 2009]. Dans le génome humain, la plupart des CpG sont méthylées sauf dans les régions riches en CG, appelées "ilôts CpG" (1% du génome humain), que l'on retrouve dans 60% des promoteurs chez l'homme et la souris [Robertson et Wolffe, 2000]. La méthylation est asymétrique à l'intérieur de la séquence des gènes mais pas au niveau des sites de fixation de l'ADN et des enhancers. Les défauts du système de méthylation sont associées à plusieurs maladies (syndromes d'ICF, de Rett et du X fragile) [Robertson et Wolffe, 2000], ce qui démontre que la méthylation de l'ADN est essentielle pour le développement chez les mammifères.

La méthylation de l'ADN dans le génome humain se trouve majoritairement au niveau des régions codantes des gènes, des régions intergéniques et sur les séquences répétées (ET inclus) ([Eckhardt *et al.*, 2006], [Weber *et al.*, 2005]). Cette méthylation a toujours été une marque répressive par défaut associée avec l'inactivité transcriptionnelle. Cependant une comparaison de l'activité des promoteurs des fibroblastes avec leur taux de méthylation a permis de voir que la plupart des promoteurs inactifs n'étaient pas méthylés [Weber *et al.*, 2007]. De plus, la comparaison avec les taux de méthylation des cellules germinales a permis d'identifier seulement 5% des ilôts CpG comme méthylés dans les fibroblastes. Un grand nombre de ces ilôts correspond à des gènes spécifiques des cellules germinales, suggérant ainsi qu'une des fonctions de la méthylation de l'ADN est de réprimer le programme germinal dans les cellules somatiques. Ces résultats montrent l'importance de la méthylation dans la stabilisation de l'identité cellulaire

contribuant ainsi à la stabilité du génome. Elle est également associée à la protection de l'intégrité du génome en inhibant toute activité des ET [Slotkin et Martienssen, 2007], ainsi qu'à d'autres fonctions importantes ([Weber et Schübeler, 2007] pour revue) comme l'inactivation du chromosome X, l'empreinte génomique, la régulation des gènes, la différentiation cellulaire, etc.

Les perturbations des profils de méthylation de l'ADN jouent un rôle important dans l'apparition des tumeurs ([Deltour *et al.*, 2005], [Linhart *et al.*, 2007]). Les changements de la méthylation à l'état tumoral surviennent dans des régions spécifiques du génome avec une hyperméthylation au niveau des gènes suppresseurs de tumeurs, ce qui entraîne leur inactivation, et une hypométhylation globale observée surtout au niveau des ET [Weber, 2008]. Cette hypométhylation entraîne une réactivation des ET, ce qui peut perturber le fonctionnement de la cellule. Plusieurs exemples dans la littérature montrent en effet une expression de différents ET dans différents types de cancer ([Szpakowski *et al.*, 2009], [Belancio *et al.*, 2010]).

Modifications d'histones

Les histones sont de petites protéines basiques formées d'un domaine globulaire et d'une extrémité aminoterminale (NH2-terminale) plus flexible qu'on appelle la "queue d'histone". Il existe cinq protéines d'histones, H1, H2A, H2B, H3 et H4 [Isenberg, 1979] qui s'organisent pour constituer le nucléosome, l'unité basique répétitive de la chromatine. Dans chaque nucléosome, l'ADN s'enroule autour d'un coeur d'histones formé par un tetramère H3-H4 et deux dimères H2A-H2B, sur lesquels vient se fixer l'histone H1 [Luger *et al.*, 1997]. Les queues d'histones émergent hors du nucléosome et sont sujettes à des modifications chimiques nombreuses (on en compte aujourd'hui plus d'une centaine) et variables, telles que l'acétylation, la méthylation ou la phosphorylation. Ces modifications correspondent à des changements post-traductionnels qui interviennent sur des résidus particuliers de l'extrémité N-terminale des histones. Même si nos connaissances sur ces modifications restent globalement limitées, des progrès considérables ont été faits quant à la compréhension des acétylations et des méthylations des résidus lysines, spécialement au niveau des histones H3 et H4 [Kouzarides, 2007]. Il faut noter ici que les

modifications d'histones et la méthylation d'ADN n'agissent pas séparément, mais qu'elles s'influencent mutuellement pour assurer une régulation plus stricte de la chromatine ([Espada *et al.*, 2004], [Tamaru et Selker, 2001], [Bachman *et al.*, 2003]). Tandis que la méthylation d'ADN est globalement associée à la répression de l'expression génique, les modifications d'histones peuvent avoir divers effets qui dépendent non seulement de la nature de la modification mais également du résidu sur lequel elle se situe. À titre d'exemple, l'acétylation et la méthylation des lysines sont associées à une chromatine accessible et une expression génique ([Kuo et Allis, 1998], [Wang *et al.*, 2008]) ; c'est le cas pour la méthylation de la lysine 4 sur l'histone H3 (H3K4me) et de la lysine 36 sur la même histone (H3K36me) ([Bannister et Kouzarides, 2005], [Barski *et al.*, 2007]). À l'inverse, la di- ou triméthylation de la lysine 9 (H3K9me2 et me3) ou de la lysine 27 sur l'histone H3 (H3K27me2 et me3) ainsi que la triméthylation de la lysine 20 sur l'histone H4 (H4K20me3) provoquent une condensation de la chromatine responsable de la répression des gènes ([Margueron *et al.*, 2005], [Schotta *et al.*, 2004], [Kourmouli *et al.*, 2004]).

Les modifications d'histones constituent un moyen complémentaire de la méthylation d'ADN pour limiter l'activité des ET. En effet, la fixation des modifications répressives au niveau des régions riches en ET provoque une condensation de la chromatine de cette région empêchant ainsi toute activité des ET. L'étude des modifications d'histones enrichies au niveau des ET, a montré que les nucléosomes associés aux ET, chez la souris, sont enrichis en H3K9 monométhylé (H3K9me, marque répressive) [Martens *et al.*, 2005] et que cette modification, associée avec la triméthylation de H4K20 (H4K20me3), jouent un rôle important dans le "silencing" des éléments répétés globalement et des ET plus spécifiquement [Mikkelsen *et al.*, 2007]. Des études similaires menées sur le génome humain ont permis de montrer un enrichissiment en H3K9me2 au niveau des éléments *Alu* [Kondo et Issa, 2003]. Plus globalement, dans le génome de la souris, les SINE montrent un enrichissement de la marque répressive H3K27me3 [Pauler *et al.*, 2009]. Ces résultats ont également été trouvés chez les plantes chez lesquelles les ET sont enrichis en H3k9me2 et H3K27me1 [Rigal et Mathieu, 2011]. Enfin, il a été montré que la méthylation de H3K9 est impliquée dans la régulation des gènes, dans la structure des chromosomes et dans le contrôle des ET [Peng et Karpen, 2007]. Tous ces résultats montrent ainsi une réelle répression de l'activité des ET par les modifications d'histones. Certains ET ne semblent

cependant pas être soumis à cette répression. En effet, une étude récente a montré que les ET impliqués dans des fonctions régulatrices chez l'homme (promoteurs humains qui dérivent des ET) sont enrichis en modifications d'histones activatrices [Huda *et al.*, 2011]. Ainsi, la régulation épigénétique des ET *via* les modifications d'histones dépend de leur activité dans le génome hôte.

Interférence par les ARN

Depuis la découverte des petits ARN en 1993, un grand nombre de classes a été identifié ([Ghildiyal et Zamore, 2009], pour revue). Ces classes diffèrent dans leur biogenèse, leur mode de régulation de la cible et les voies biologiques qu'ils régulent. Les caractéristiques qui définissent les ARN impliqués dans le "silencing" sont leur petite taille (∼21-30 nucléotides) et leur association avec les membres de la famille de protéines Argonaute (Ago), qui les guident vers leurs cibles. Les petits ARN jouent un rôle très important dans la dégradation et l'inactivation de l'expression des ARN aberrants (qui proviennent de la transcription d'un virus, transposon, etc.), assurant ainsi une stabilité génomique [Shalgi *et al.*, 2010], d'où leur appelation de système immunitaire du génome [Plasterk, 2002]. L'activité de répression peut se faire avant la transcription de l'ADN parasite, on parle dans ce cas d'une répression transcriptionnelle (TGS pour "Transcriptional Gene Silencing"), ou bien après la transcription, on parle dans ce cas d'une répression post-transcriptionnelle (PTGS pour "Post Transcriptional Gene Silencing"). On connaît aujourd'hui trois classes majeures de petits ARN :

- **Les small interfering RNA (siRNA)** : Ils existent dans les trois règnes des eucaryotes (plantes, métazoaires et champignons) et fournissent une réponse antivirale chez les plantes et les métazoaires au moins [Ghildiyal et Zamore, 2009]. L'ADN parasite peut être exogène ou provenir du génome hôte. Si les petits ARN sont produits à partir d'un parasite endogène, on parle d'endo-siRNA. Dans cette voie, l'identification d'un ARN aberrant induit la synthèse d'une molécule ARN double brin (dsRNA pour "double stranded RNA") par l'intermédiaire d'une ARN polymérase ARN dépendante (RdRp pour "RNA directed RNA polymerase"). Ce dsRNA est ensuite dégradé par l'enzyme DICER en petits

morceaux d'ARN de 20-21 nucléotides, qu'on appelle les ARN guides (siRNA pour "small interfering RNA"). Les siRNA sont ensuite incorporés au sein d'un complexe ribonucléoprotéique, appelé RISC ("RNA-Induced Silencing Complex"). Cette incorporation permet la localisation du complexe RISC au niveau de l'ARNm cible induisant ainsi sa dégradation *via* la nucléase Tudor-SN présente au sein du complexe [Robert et Bucheton, 2004].

- **Les microRNA (miRNA)** : Ils sont présents chez les plantes et les métazoaires mais pas chez les champignons. Les miRNA (20-25 nucléotides) reconnaîssent l'ARN aberrant grâce à des séquences complémentaires situées dans les régions 3'UTR des ARNm et se fixent dessus, empêchant ainsi leur traduction. Les miRNA peuvent également s'incorporer dans le complexe RISC agissant ainsi comme les siRNA [Kawasaki *et al.*, 2005].

- **Les piwi-interacting RNA (piRNA)** : Les piRNA sont les derniers petits ARN à avoir été découverts. Ils ont été trouvés chez les métazoaires seulement et possèdent la plus grande taille parmi toutes les classes des petits ARN (25-30 nucléotides). Ils diffèrent des deux classes précédentes par leur action spécifique au niveau de la voie germinale et par la structure de l'ARN précurseur. En effet, ce dernier est simple brin (ssRNA pour "single-stranded RNA") dans le cas des piRNA alors qu'il est double brin pour les siRNA et miRNA. L'ARN précurseur, qui est en antisens par rapport à l'ARNm cible, est dégradé par la protéine Argonaute en plusieurs piRNA [Obbard *et al.*, 2009], on parle ainsi d'une voie DICER-indépendante. Les piRNA peuvent être incorporés également dans un complexe protéique formé par les protéines Aubergine (Aub) et Piwi.

Les trois voies majeures des petits ARN ont été considérées initialement comme indépendantes et distinctes mais elles semblent interagir entre elles [Ghildiyal et Zamore, 2009]. Bien que leur cible primaire soit la molécule ARNm, ces trois voies sont également capables de réaliser des TGS puisque les protéines impliquées dans ces différentes voies peuvent interagir avec d'autres protéines induisant la méthylation de l'ADN des séquences cibles (RdDM, pour "RNA-directed DNA methylation") ainsi que la fixation de modifications d'histones répressives, ce qui provoque

une hétérochromatinisation locale (condensation de la chromatine) et par conséquent leur inaccessibilité.

Les piRNA sont directement impliqués dans le "silencing" des ET dans la lignée germinale des métazoaires assurant ainsi leur stabilité [Aravin *et al.*, 2001]. La plupart des études sur les piRNA sont réalisées chez la drosophile, mais l'identification chez les mammifères de certains piRNA qui ne proviennent pas d'ET a permis de diviser les piRNA en deux sous-groupes : les pré-pachytènes, qui sont responsables du "silencing" des ET (*L1* chez l'homme et IAP chez la souris) et les pachytènes, qui ne proviennent pas des ET et dont la fonction reste inconnue [Aravin *et al.*, 2007]. Au niveau somatique, le "silencing" des ET se fait essentiellement par les endo-siRNA et les miRNA. Les endo-siRNA qui dérivent des ET sont impliqués dans l'hétérochromatinisation locale chez la drosophile [Fagegaltier *et al.*, 2009] et dans la méthylation de l'ADN chez les plantes [Matzke *et al.*, 2009]. De plus, plusieurs études faites sur l'origine des miRNA ont montré que plus que 20% des miRNA humains proviennent des ET ([Piriyapongsa et Jordan, 2007],[Smalheiser et Torvik, 2005]), ce qui suppose que ces miRNA sont responsables de la répression des ET dans les cellules somatiques. Plusieurs études supportent cette hypothèse puisqu'on a trouvé par exemple que les éléments *Alu* sont ciblés par plusieurs miRNA dont un large cluster de 46 miRNA flanqués par des *Alu* situés sur le chromosome 19 [Lehnert *et al.*, 2009].

1.6.2 Autres voies de régulation

Édition des ARN

Un transcrit ARN subit plusieurs processus de maturation ("capping" en 5', polyadénylation en 3', épissage, etc.) après sa transcription à partir de la séquence génique. Ces processus de maturation sont indispensables pour obtenir une molécule ARNm mature et prête à être traduite en protéine. L'édition des ARN est un mécanisme de modification post-transcriptionnelle qui résulte en une séquence ARN différente de celle codée par l'ADN génomique, contribuant ainsi à la diversification des protéines produites par les gènes ([Nishikura, 2006], pour revue). L'édition de l'ARN se fait au niveau des sites d'édition qui peuvent être tissus-spécifiques [He *et al.*, 2011] et nécessitent la formation

d'une molécule dsRNA qui est éditée par l'enzyme ADAR ("Adenosine Deaminase Acting on RNA"). Cette enzyme désamine l'Adénosine (A) qui devient une Inosine (I). Cette Inosine est reconnue par les différents systèmes de traduction et RT comme une Guanosine (G). L'édition au niveau des séquences codantes peut altérer/modifier les fonctions des protéines qui en dérivent de diverses façons. En effet, elle peut modifier le site d'initiation de la transcription (ATG $\rightarrow$ GTG), les sites donneurs (AT $\rightarrow$ GT) et accepteurs (AA $\rightarrow$ AG) d'épissage et les codons stop, tout ceci résultant en une protéine très probablement tronquée et non fonctionnelle. Une étude faite sur les ARN humains a montré qu'environ 1,4% des ARN sont sujets à cette modification dont la plupart a lieu au niveau des introns (85-90%), le reste ayant lieu au niveau des UTR et des parties codantes [Athanasiadis *et al.*, 2004]. Cette même étude a montré également que les ET et plus spécifiquement, les *Alu*, sont la cible principale du mécanisme d'édition. Le fait que les éléments *Alu* soient les plus visés par cette modification vient de leur grande présence dans les introns (par rapport aux autres familles d'ET) sans oublier leur pourcentage élevé dans le génome humain (10%) et leur petite taille [Lander *et al.*, 2001]. L'édition des ET est un moyen important de régulation dans le sens où un ET qui apporte un certain avantage pour le génome hôte sera intégré dans le transcrit, *via* la modification des sites d'épissage et exclu dans le cas inverse.

Et aussi...

On peut enfin citer le "Repeat-Induced Point mutation" ou RIP qu'on a découvert chez *Neurospora crassa*, un champignon filamenteux [Galagan et Selker, 2004]. Dans ce cas, le RIP détecte toutes les séquences dupliquées qui ont une identité de séquence de $\sim$80% entre elles et qui ont une taille supérieure à 400 nucléotides, puis induit un grand nombre de mutations dans ces séquences, suivies d'une méthylation de la séquence d'ADN. Ceci explique qu'il n'y ait aucune activité transpositionnelle des ET dans ce génome.

1.7 Conclusion

Il est clair que plus de 60 ans après leur découverte le débat sur les éléments transposables reste totalement d'actualité avec des arguments pour chacune des théories qui expliquent leur présence et leur maintien dans les génomes hôtes. Les ET ont un large impact, à la fois positif et négatif, sur le génome, et ce dernier tend à réguler l'activité de ces éléments. Il y a donc une interaction permanente entre les ET et leur hôte que certains auteurs ont comparé à une interaction entre un parasite et un hôte, interaction dans laquelle ils ont considéré le génome comme un écosystème composé de plusieurs niches écologiques dans lesquelles un ensemble d'espèces (les ET) interagissent ([Brookfield, 2005], [Venner *et al.*, 2009]). Afin d'améliorer la compréhension de cette relation il est nécessaire d'examiner de manière plus approfondie les moyens utilisés par le génome pour contrôler les ET mais également de comprendre les modifications qui sont à l'origine de la perte de ce contrôle. Dans ce travail de thèse, je présenterai un élément de réponse concernant la régulation de la distribution des ET et la dérégulation de leur expression. Ainsi, dans la première partie de ma thèse, je me suis penché sur la distribution des ET dans le génome humain en analysant la variation de leur densité selon la fonction des gènes qui sont dans leur voisinage ; dans la deuxième partie j'ai analysé l'impact de l'altération des modifications d'histones associées aux différents composants géniques, dont les ET, sur la variation de l'expression des gènes, en condition tumorale.

Chapitre 2

Matériel et Méthodes

2.1 Les gènes

Les gènes identifiés dans le génome humain constituent le matériel de base utilisé pour l'ensemble des études réalisées au cours de ma thèse. Plusieurs bases de données comme **ENSEMBL** [Flicek *et al.*, 2008] (`www.ensembl.org`) ou **UCSC** [Sanborn *et al.*, 2011] (`http://genome.ucsc.edu/`), regroupent toutes les informations concernant les gènes des génomes séquencés. Étant donné le séquençage constant de nouvelles espèces ainsi que l'amélioration de l'annotation des génomes, les bases de données sont fréquemment mises à jour. Ces mises à jours sont accompagnées par une mise à jour de l'annotation des génomes. À titre d'exemple, j'ai comparé la version d'**ENSEMBL** utilisée pour ma thèse (version 50, Juillet 2008, version NCBI36 du génome humain) à la dernière mise à jour (version 62, Avril 2011, version GRCh37 du génome humain) et à la première version d'**ENSEMBL** (version 35, Novembre 2005, version NCBI35 du génome humain) (**Tableau 1**). Cet exemple montre les conséquences de l'amélioration de l'annotation génomique, ce qui explique la variation observée des nombres de gènes et de pseudogènes. Enfin plusieurs systèmes de requêtes sur les bases de données ont été développés parmi lesquels j'ai choisi d'utiliser **BioMart** [Haider *et al.*, 2009] (`www.biomart.org`) et plus spécifiquement l'outil **Martview** pour leur connection directe avec **ENSEMBL** .

Tableau 1: *Comparaison de trois versions d'ENSEMBL. Les nombres de gènes et pseudogènes ainsi que le nombre total de nucléotides pour chacune des versions d'ENSEMBL est indiqué.*

Version ENSEMBL	35(2005)	50(2008)	62(2011)
Nombre de gènes	22218	20435	20067
Nombre de pseudogènes	1976	6282	12542
Nombre total de nucléotides	3272187692	3253037807	3265996791

J'ai utilisé **Martview** pour récupérer l'ensemble des gènes humains. Pour chaque gène, j'ai récupéré ses "coordonnées", c'est-à-dire le chromosome sur lequel est situé le gène, son début, sa fin et l'ensemble des exons. Certains gènes possèdent plusieurs ARN messagers en raison de l'épissage alternatif. J'ai donc utilisé les données des exons de tous les messagers pour produire un ensemble d'exons uniques par gène (**Figure 6**), à partir desquels j'ai pu déduire les coordonnées des introns. La définition de la région promotrice

située en amont du gène varie en fonction de la taille. En effet, Kim *et al.* [Kim *et al.*, 2005] ont montré que 87% des facteurs de transcription sont situés dans la région 2,5 kilobases (kb) en amont du début du gène, la grande majorité se situant dans la région 1 kb en amont ; certains facteurs se trouvent à 9,8 kb du début du gène. Ainsi, j'ai considéré deux régions promotrices à des distances de 2 ou 10 kb par rapport au début du gène et les séquences correspondantes à ces régions ont été téléchargées *via* **Martview**. Pour certains gènes, j'ai récupéré leurs orthologues présents chez le chimpanzé, le macaque, l'orang-outan et la souris. Deux gènes dans deux espèces sont dits orthologues quand ils proviennent par héritage d'un gène ancestral commun. Il existe ainsi au plus un orthologue par génome pour chaque gène, qui est obtenu par la méthode du "Best Bidirectionnal Hit" ou BBH. Dans cette méthode, deux gènes protéiques a et b, appartenant respectivement aux génomes des organismes A et B, sont dits en BBH (orthologues) si : (1) les protéines codées par ces gènes se ressemblent (taille et identité de séquence), (2) le gène de A le plus ressemblant (ou meilleur hit) à b est a, (3) le gène de B le plus ressemblant à a est b. Dans les bases de données, certains gènes humains possèdent plusieurs gènes orthologues dans un génome donné, et un gène orthologue peut également correspondre à plusieurs gènes humains, ces deux cas étant probablement la conséquence d'une duplication. J'ai décidé de ne pas prendre en compte ce type de gènes et de ne garder que les gènes qui possèdent un orthologue et un seul dans chacune des autres espèces.

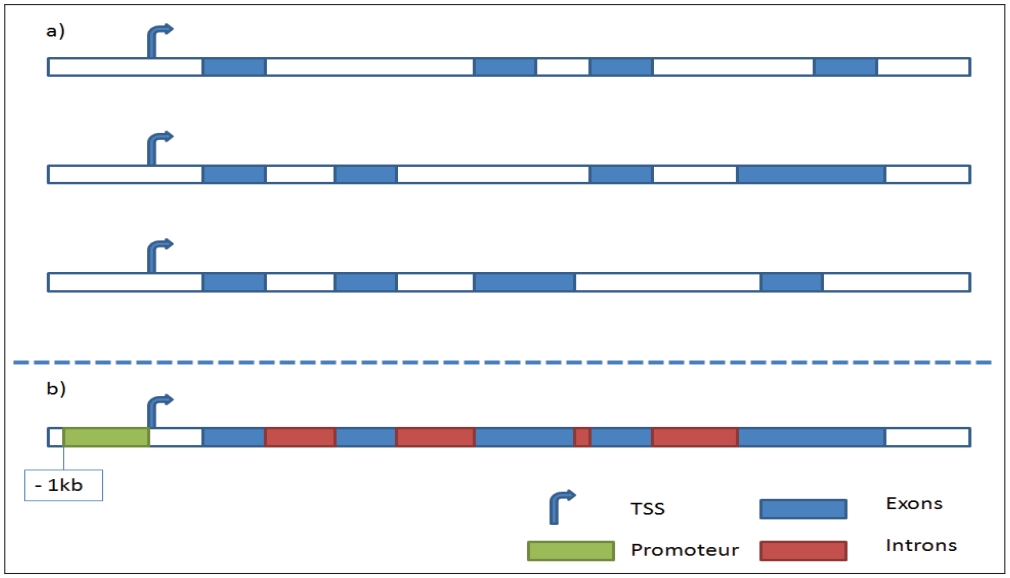

Figure 6: *Exemple d'exons uniques. (a) Exemple de différents ARNm du gène. (b) L'ensemble des ARNm permet d'établir un gène possèdant tous les exons à partir desquels on peut déduire la position des introns.*

2.2 Les Éléments Transposables

2.2.1 Recherche des Éléments Transposables dans les génomes

On peut déterminer le contenu en ET à l'intérieur et dans le voisinage de chaque gène par la méthode traditionnelle, qui consiste à utiliser le logiciel **Repeat Masker** [Smit *et al.*, 2010] (`www.repeatmasker.org`) qui détermine les positions des ET inclus dans une séquence donnée. Les coordonnées des ET peuvent être facilement accessibles en utilisant la version prémasquée du génome humain à condition que la version utilisée pour masquer le génome soit la même que celle qui a été utilisée pour la recherche des gènes. En effet, plusieurs génomes ont déjà été prémasqués et les différents fichiers de sortie sont disponibles et téléchargeables sur le site de **Repeat Masker** (`www.repeatmasker.org/PreMaskedGenomes.html`). Chaque génome prémasqué possède deux fichiers résultats : le fichier ayant pour extension ".align" qui contient l'alignement entre la séquence de la requête et les ET trouvés, le fichier ".out" qui contient la liste de tous les ET trouvés. Le fichier ".out" contient plusieurs informations sur chaque ET trouvé comme son début et sa fin dans la séquence requête, le pourcentage de divergence entre l'ET trouvé dans la séquence requête et la séquence consensus de l'ET en question. Ainsi j'ai téléchargé le fichier ".out" de la version prémasquée du génome humain contenant toutes les informations des ET inclus dans ce génome.

2.2.2 Calcul de la densité en Éléments Transposables

Le nombre d'ET inclus dans chaque gène s'obtient en croisant le fichier des coordonnées des gènes humains avec celui des ET, en se basant sur le fait que si le début et la fin de l'ET sont supérieurs et inférieurs, respectivement, au début et à la fin du gène, l'ET sera considéré comme inclus dans le gène en question. Je ne me suis pas limité à la séquence du gène puisque j'ai également pris en considération les régions flanquantes situées à 2 et 10 kb en amont et en aval du gène. Je peux ainsi obtenir le nombre d'ET inclus dans chaque gène et ses régions flanquantes. Dans le but de pouvoir classifier les gènes selon leur richesse en ET, le nombre d'ET par gène ne peut pas être utilisé parce qu'il présente un biais qui dépend de la taille du gène, dans le sens où un grand gène a plus de chance de contenir un grand nombre d'ET et être classé comme riche en ET par rapport à un

gène de petite taille. Afin de s'affranchir de ce biais, j'ai calculé la densité en ET pour chaque gène en divisant le nombre d'ET inclus dans le gène en question par la différence de taille du gène et de celle de tous les ET qui y sont inclus. Cette densité a été ensuite multipliée par 10^4 pour permettre une représentation plus commode des données.

$$\text{Densité ET} = \frac{\text{Nombre ET}}{\text{Taille gène} - \sum \text{Taille ET}} * 10^4$$

La **Figure 7** représente un exemple du calcul de la densité en ET alors que la **Figure 8** récapitule la démarche utilisée pour le calcul de la densité en ET. Les ET ont ensuite été considérés selon quatre types : les transposons à ADN, les rétrotransposons à LTR, les LINE et les SINE.

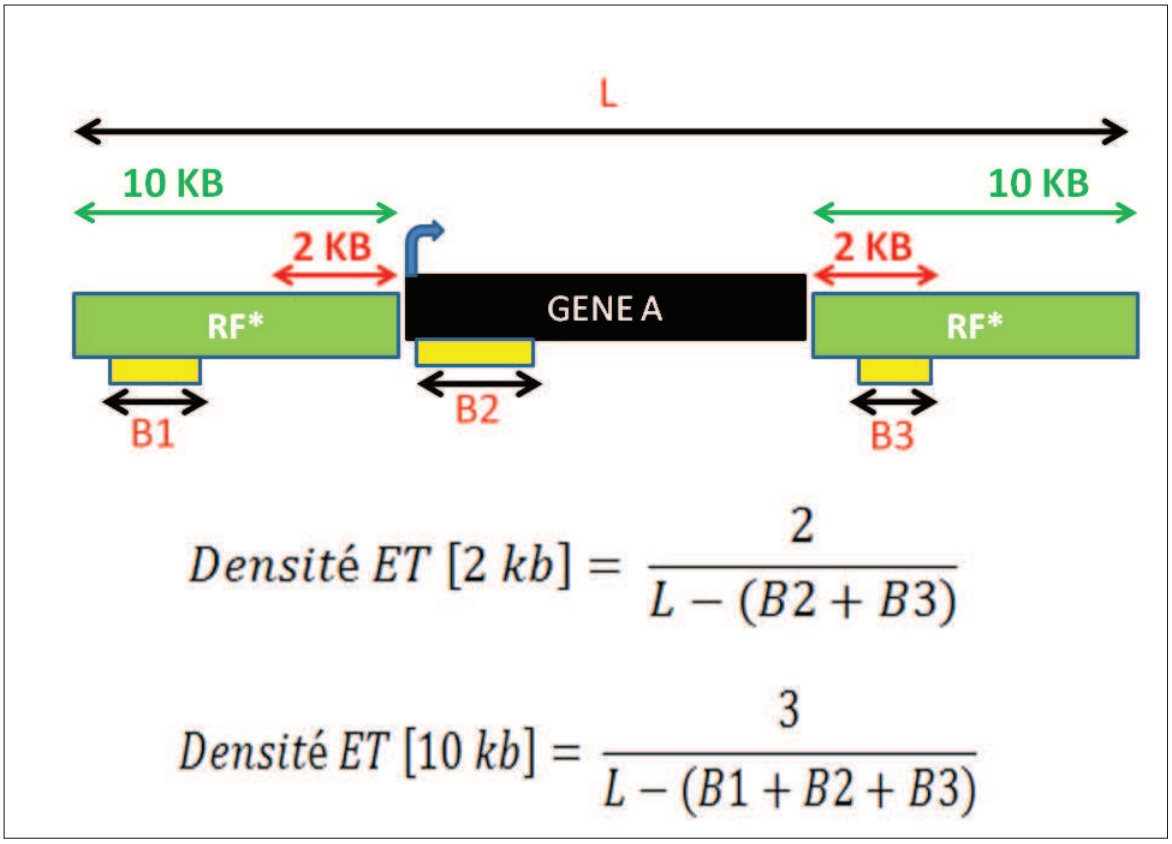

FIGURE 7: *Exemple de calcul de la densité en ET pour le gène A. Deux types de régions flanquantes sont définis (2 et 10 kb). Les ET sont représentés en jaune et l'indice en dessous représente leur taille. L est la taille totale du gène A (régions flanquantes comprises). La formule permettant de calculer la densité en ET pour chaque type de région flanquante est indiquée. RF* : Région Flanquante.*

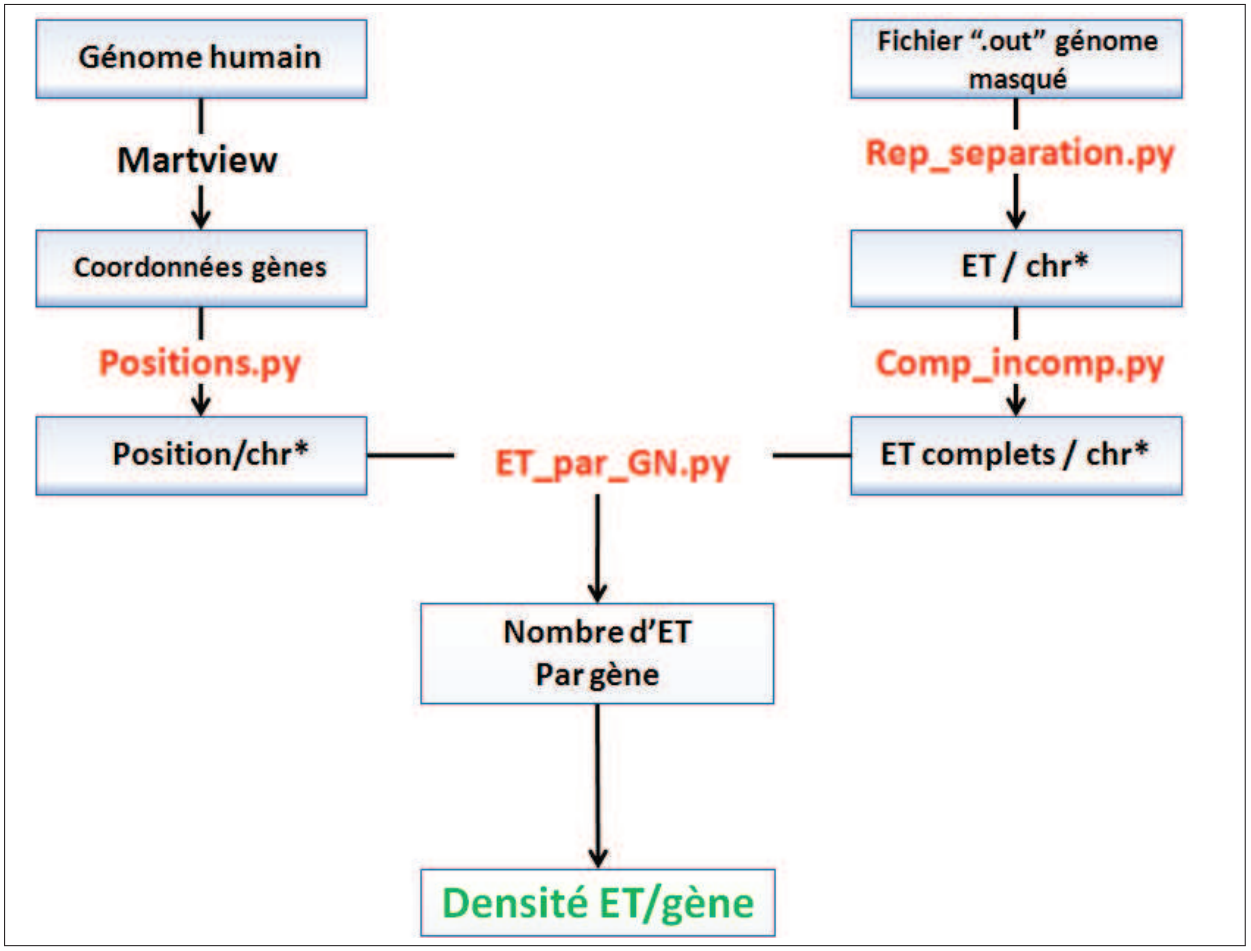

FIGURE 8: *Récapitulatif de la démarche utilisée pour le calcul de la densité en ET. D'un côté on récupère les coordonnées géniques qu'on sépare par chromosome, et de l'autre on sépare les coordonnées des ET par chromosome. Les deux informations sont ensuites croisées pour chaque chromosome pour obtenir le nombre et calculer la densité en ET par gène. En rouge sont représentés les scripts python que j'ai développés. chr* : chromosome.*

2.3 Étude de la fonction des gènes

L'analyse comparative des fonctions des différentes classes de gènes (comparaison deux à deux) a été faite *via* le logiciel **FatiGO** ([Al-Shahrour *et al.*, 2004], [Al-Shahrour *et al.*, 2007]) de la suite **Babelomics** [Medina *et al.*, 2010] (`http://babelomics.bioinfo.cipf.es/functional.html`). Il s'agit d'une application web qui permet d'extraire les informations sur les fonctions moléculaires, les processus biologiques et les compartiments cellulaires des gènes en se basant sur une ontologie (GO, pour "Gene Ontology"). Le projet de GO (`www.geneontology.org`) est une initiative créée en 2000 [Ashburner *et al.*, 2000] dont le but est de standardiser la représentation des attributs des

gènes et de leurs produits à travers les espèces. Les identifiants GO sont classés dans trois domaines qui ne se recouvrent pas [Gene Ontology Consortium, 2001] :

- **Fonction Moléculaire ("Molecular Function")** : cette partie décrit les activités biochimiques des produits des gènes. La description se limite à exposer le type du produit de transcription sans préciser ni le lieu ni le moment de son action (exemples : enzyme, transporteur, ligand, etc.).

- **Processus Biologique ("Biological Process")** : permet de décrire les processus biologiques dans lesquels le produit du gène est impliqué (exemples : biosynthèse de l'AMP cyclique, croissance cellulaire, etc.).

- **Composant Cellulaire ("Cellular Component")** : fait référence aux localisations cellulaires des molécules issues des gènes (exemples : noyau, appareil de Golgi, etc.).

Le GO est structuré sous la forme d'un réseau où chaque terme est un enfant (descendant) qui peut avoir un ou plusieurs parents (termes plus généraux) et aucun, un ou plusieurs enfants, en allant d'un niveau général vers un niveau de plus en plus spécifique. Il s'agit d'une structure en graphe acyclique orienté (DAG, pour "Directed Acyclic Graph"). Par exemple, le processus biologique *"hexose biosynthetic process"*, qui est donc le terme enfant, possède deux termes parents : les processus *"hexose metabolic process"* et *"monosaccharide biosynthetic process"*. Initialement, deux types de relations entre les termes parents et les termes enfants ont été définis [Gene Ontology Consortium, 2001] : le type *"part of"* où l'enfant est un des composants des parents (exemple : le flagelle périplasmique est une partie (*part of*) de l'espace périplasmique) et le type *"is a"*, qui signifie que le terme enfant est une instance du terme parent (exemple : le chromosome nucléaire est (*is a*) un chromosome). Récemment, une nouvelle mise à jour de GO [Gene Ontology Consortium, 2010] a permis de rajouter d'autres types de relations comme le type *"has part"*, qui signifie que le terme parent possède en partie le terme enfant et que donc quand le terme parent est présent, le terme enfant est sûrement présent à son tour. Cependant la présence du terme enfant n'implique pas forcément la présence du

terme parent (exemple : l'enveloppe cellulaire est constituée en partie (*has part*) de la membrane plasmique).

Le logiciel **FatiGO** possède plusieurs options permettant de chercher les fonctions d'une liste donnée de gènes, de comparer les fonctions d'une liste de gènes au génome entier ou de comparer les fonctions de deux listes de gènes entre elles. L'option de comparaison entre deux listes de gènes a été utilisée pour comparer les fonctions des différentes classes de gènes définies (selon leur densité en ET) entre elles. Dans ce cas, chaque liste de gènes est convertie en termes GO, et le logiciel affiche les fonctions moléculaires, processus biologiques et localisations cellulaires qui sont surreprésentées dans l'une des listes par rapport à l'autre. La significativité de la surreprésentation d'un terme GO donné est donnée par un test de Fischer. Enfin, étant donné que plusieurs tests peuvent être réalisés à la fois, **FatiGO** propose également une correction de tests multiples en se basant sur la correction de Benjamini et Hochberg (dite également FDR pour "False Discovery Rate").

2.4 Analyses évolutives

2.4.1 Calcul de la pression de sélection

La valeur, ω, indique le type de sélection exercée sur un gène donné. Cette valeur est égale au rapport $\frac{K_A}{K_S}$, où K_A indique le taux des substitutions non synonymes (qui changent l'acide aminé sous-jacent) et K_S, le taux de substitutions synonymes. Selon la valeur de ω, on décide de la nature de la sélection :

- Si $\omega > 1$: La sélection est positive ou adaptative.
- Si $\omega < 1$: La sélection est négative ou purificatrice. Elle tend à garder l'identité de la séquence nucléotidique en éliminant tout type de mutation.
- Si $\omega = 1$: L'évolution est neutre sans aucune sélection. Ainsi la mutation peut se fixer par chance.

Le calcul du ratio ω pour un gène humain donné peut être divisé en trois étapes :

1) Récupération des séquences orthologues d'un gène humain chez une ou plusieurs espèces proches (par exemple chimpanzé).

2) Alignement des séquences codantes des gènes orthologues *via* **Muscle** [Edgar, 2004].

3) Calcul du ratio ω par le programme **codeml** du paquet **PAML** ([Yang, 2007], [Bielawski et Yang, 2001]).

Les séquences codantes humaines qui ne possèdent pas de séquences orthologues chez les espèces proches ainsi que celles qui possèdent un codon stop interne sont exclues de notre analyse. De plus, les gènes orthologues dont les séquences sont quasiment identiques (pourcentage d'identité $\geq 99\%$) ont été également éliminées pour éviter le phénomène de saturation et l'obtention de valeurs aberrantes du ratio ω (par exemple 999).

2.4.2 Modèles d'évolution

Pour savoir si un gène donné a subi une pression de sélection différente spécifiquement chez une espèce donnée ou s'il est soumis à la même pression de sélection dans toutes les espèces considérées, j'ai analysé différents modèles de sélection en tenant compte de la phylogénie. Les tests ont été réalisés grâce au programme **codeml** du paquet **PAML** (exemples d'analyses :[Yang, 2007] [Yang, 2002] [Yang, 1998]). Dans mes analyses je me suis intéressé à trois modèles : **M0**, **M1** et **M2**. La **Figure 9** représente un exemple de l'utilisation de ces trois modèles.

Le modèle **M0** assume que la pression de sélection est identique sur toutes les branches de l'arbre (**Figure 9**), c'est-à-dire que le ratio ω_0 est égal entre les quatre espèces. Ce modèle est l'opposé du modèle **M1** qui suppose une pression de sélection indépendante sur chacune des branches de l'arbre, impliquant un ratio ω différent pour chaque espèce. Le modèle **M2** teste une pression de sélection (ω_1) différente sur une branche spécifique de l'arbre (branche de l'homme dans la **Figure 9**) par rapport à une pression commune (ω_0) sur les autres branches. Pour un gène donné, on peut tester un modèle par rapport à un autre (par exemple **M1** *versus* **M2**) et la significativité statistique du test est obtenue en comparant le double de la différence de la valeur de vraisemblance de chacun des modèles (2δ) à une valeur seuil d'une distribution χ^2 dont le nombre de degrés de liberté (ddl) est la différence du nombre de paramètres de chacun des modèles ([Yang, 2007], [Yang, 2002], [Yang, 1998]).

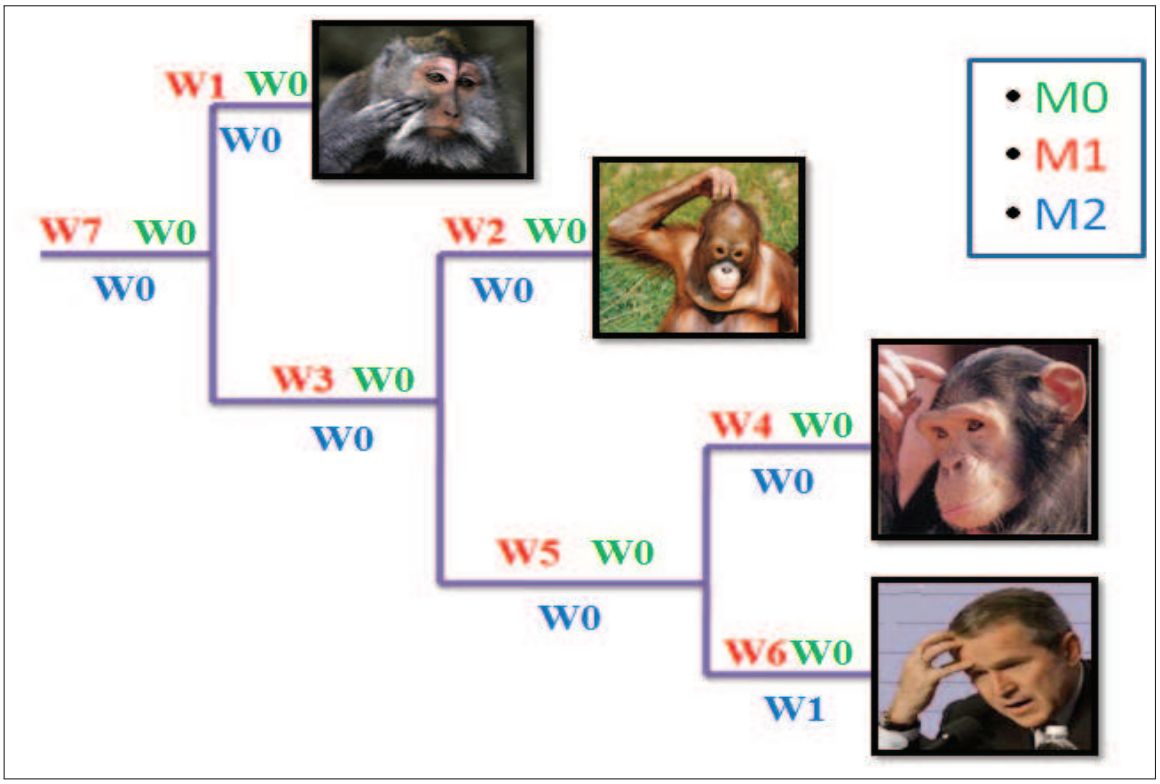

FIGURE 9: *Modèles d'évolution. Exemple d'un gène orthologue présent chez quatre espèces de primates : homme, chimpanzé, orang-outan et macaque. Trois modèles différents sont testés : le modèle M0 (vert), le modèle M1 (rouge) et le modèle M2 (bleu). W indique la valeur ω de la pression de sélection.*

2.4.3 Pourcentage d'identité

La valeur ω ne peut être calculée qu'au niveau des régions codantes. Ainsi l'estimation d'une pression de sélection qui s'exercerait au niveau des régions flanquantes des gènes doit être calculée par une autre approche. Le pourcentage d'identité entre deux séquences est un bon indicateur de cette pression dans le sens où une grande conservation d'une séquence donnée entre deux espèces (conservation supérieure au taux d'identité entre les deux génomes) suggère une pression de sélection qui empêche la fixation des mutations dans cette séquence. À l'opposé, une grande divergence suggère une absence de sélection, c'est-à-dire une évolution neutre. L'estimation du pourcentage d'identité est proche de la méthode du calcul de la valeur ω : après avoir récupéré (*via* **BioMart**) et aligné (*via* **Muscle**) les séquences flanquantes d'un gène orthologue donné, le pourcentage d'identité est calculé *via* le programme **dnadist** du paquet **phylip** [Felsenstein, 1989].

2.5 Enrichissement en marques d'histones

Les marques d'histones sont des modifications post-traductionnelles des queues des protéines d'histones. Ces modifications peuvent être quantifiées en terme d'enrichissement par le biais d'une immuno-précipitation de chromatine suivie, soit d'un séquençage des séquences ADN ("**Chip-seq**"), soit de leur fixation sur une puce à ADN ("**Chip-on-chip**"). L'immunoprécipitation d'une modification d'histone donnée (H3K4me3 par exemple) se fait en utilisant un anticorps spécifique de cette modification (anti H3K4me3). Ensuite l'ADN fixé à cette protéine est séquencé et localisé dans le génome. Il existe plusieurs logiciels qui permettent d'analyser les résultats de "**Chip-seq**" (localisation et normalisation) ([Taslim *et al.*, 2009], [Ji *et al.*, 2008], [Kharchenko *et al.*, 2008]).

Plusieurs données d'enrichissement en marques d'histones pour différents tissus sont disponibles et téléchargeables à partir du site d'**UCSC** dans la partie correspondante aux "**Chip-seq**" dans l'encyclopédie des éléments ADN (**ENCODE**, pour "Encyclopedia of DNA Elements") [Raney *et al.*, 2011] (`http://genome.ucsc.edu//cgi-bin/hgTrackUi?db=hg19&g=wgEncodeBroadHistone`). Pour chaque modification d'histone deux fichiers existent : Un fichier "peak" qui contient l'enrichissement normalisé pour chaque position nucléotidique du génome et un fichier "signal" qui contient l'enrichissement de chaque intervalle de 25 nucléotides. J'ai téléchargé le fichier "signal" correspondant à chaque modification d'histone. Ce fichier consiste en la division de chaque chromosome en plusieurs intervalles de 25 nucléotides et du calcul de l'enrichissement d'une modification donnée dans chaque intervalle. Ainsi, en croisant ces données avec les coordonnées géniques d'un gène d'intérêt, on peut obtenir l'enrichissement d'histone au niveau de ce gène.

2.6 Données d'expression

Si le génome est identique dans chacune des cellules d'un organisme donné, en revanche, les gènes peuvent avoir une expression spécifique différenciée dans le temps (propre à un stade du développement), dans l'espace (propre à un type cellulaire ou tissulaire) et/ou caractéristique d'un état donné (normal, tumoral, ou en réponse à un stimulus particulier). L'étude de l'expression génique consiste à caractériser et quantifier

les produits d'expression de l'ADN, c'est à dire les ARNm, de manière à identifier, dans un tissu, dans un état et à un moment donné du développement, les séquences actives. L'ensemble des ARNm d'une cellule donnée, constitue le transcriptome. Jusqu'à récemment, il existait trois méthodes dominantes d'estimation de l'expression génique : le séquençage d'étiquettes SAGE ("Serial Analysis of Gene Expression") et EST ("Expressed Sequence Tags") et l'hybridation (puces à ADN ou "Microarrays").

Les données d'expression que j'ai utilisées proviennent d'une étude faite chez l'homme et la souris dont le but était d'établir un "atlas" d'expression de tous les gènes protéiques dans différents tissus chez ces deux espèces [Su *et al.*, 2004]. Les auteurs ont utilisé des puces à ADN sur lesquelles ils ont fixé 44775 séquences humaines et 36812 séquences de la souris, qui sont les cibles potentielles des transcrits géniques. L'analyse qui a concerné 79 tissus chez l'homme et 61 tissus chez la souris, a été faite dans différents tissus (deux réplicats pour chaque tissu) pour pouvoir définir les gènes qui ont une expression tissu-spécifique et les gènes qui s'expriment dans tous les tissus. Parmi les 79 tissus humains étudiés, dix appartiennent au système immunitaire et six sont pathologiques (extraits d'une personne atteinte d'une tumeur). Ces données sont disponibles et téléchargeables à partir du site de **BioGPS** [Wu *et al.*, 2009], la base de données qui contient tous les résultats relatifs à cette analyse.

2.6.1 Divergence d'expression entre les états normal et tumoral

Cette étude a concerné 3350 gènes humains seulement. Ceci est du au fait que j'ai éliminé les gènes qui correspondent à plusieurs sondes ainsi que les gènes présents plusieurs fois dans le jeu de données du départ (cf. Chapitre 2, section 4.2, page 74). Le calcul de la divergence d'expression entre l'état normal et l'état tumoral a été fait selon deux méthodes différentes :

Distance euclidienne

Si on considère y le niveau d'expression d'un gène donné à l'état tumoral et x son niveau d'expression à l'état normal, la distance euclidienne entre les deux états pour le gène considéré est :

$$d = \sqrt{(x-y)^2}$$

Cette distance est supposée avoir des valeurs élevées quand la divergence d'expression des gènes entre les deux états est élevée, alors qu'on s'attend à des faibles valeurs quand la divergence est négligeable. Une fois la distance euclidienne calculée pour tous les gènes, l'histogramme de la variation de cette distance est représenté dans le but de diviser les données en quatre classes différentes selon leur divergence d'expression.

Distance moyenne

Cette distance correspond à la valeur absolue de la différence d'expression génique entre l'état normal (x) et l'état tumoral (y) divisée par la somme d'expression dans les deux états.

$$d_1 = \frac{|(x-y)|}{x+y}$$

Quand la divergence d'expression est grande entre les deux états, d_1 prend des valeurs proches de 1, tandis que ses valeurs seront proches de 0 quand la divergence d'expression est faible. Après avoir calculé les valeurs d_1 pour tous les gènes et afin de déterminer différentes classes de divergence d'expression, nous avons utilisé l'algorithme **K-means** [MacQueen, 1967] qui permet de regrouper les gènes en k clusters différents. Dans le cas d'un regroupement *via* **K-means**, le nombre de clusters est fixé *a priori* (k=4, dans notre analyse), l'algorithme va ensuite définir k centroïdes, un pour chaque cluster. Le regroupement est effectué en minimisant la somme des carrés des écarts entre les données et le centroïde du cluster correspondant. En d'autres termes, chaque gène sera classé dans le groupe dont le centroïde est le plus proche de son niveau de variation d'expression. L'utilisation de deux distances différentes pour le calcul de la divergence d'expression sera justifiée plus loin.

NB : Tous les tests statistiques réalisés ont été effectués par l'intermédiaire du logiciel **R** (`www.r-project.org`) [R Development Core Team, 2005]. En cas de tests multiples, la correction de Benjamini et Hochberg a été utilisée.

Chapitre 3

Les Éléments Transposables chez les primates : distribution, sélection, expression

3.1 Introduction

La plupart des études de la distribution des ET a été réalisée sur une partie des ET ou dans une portion spécifique du gène/génome. Ainsi la majorité des études s'est focalisée sur les rétrotransposons sans LTR étant donné leur nombre de copies très élevé alors que seulement un petit nombre d'études s'est intéressé aux transposons à ADN. La première étude menée sur l'ensemble du génome humain a été réalisée par le consortium du séquençage dans laquelle les auteurs ont noté une variation du nombre d'ET non seulement entre les chromosomes mais également à l'intérieur de ces derniers [Lander *et al.*, 2001]. Ainsi, certaines régions du génome sont très denses en ET alors que d'autres en sont quasiment dépourvues. Ces résultats ont permis de valider ce qu'on pensait avant, à savoir que la distribution des ET dans un génome n'est pas aléatoire mais que plusieurs facteurs y jouent un rôle important.

Simons *et al.* se sont intéressés aux régions dépourvues de tout type d'insertion d'ET, régions qu'ils ont appellé "Transposon Free Region" ou TFR ([Simons *et al.*, 2006], [Simons *et al.*, 2007]). Ils ont trouvé plusieurs centaines de TFR présents chez l'homme, la souris, et l'opossum [Simons *et al.*, 2006]. Ces TFR dont la taille peut aller jusqu'à 81 kb, ont été trouvées à proximité des gènes impliqués dans le développement et la transcription, ce qui a permis aux auteurs de supposer que la maintenance de ces régions est due principalement à une pression de sélection qui élimine l'insertion des ET. Ces TFR ont ensuite été trouvées chez les amphibiens et les poissons, ce qui indique une conservation de ces régions sans insertion d'ET depuis environ 500 millions d'années [Simons *et al.*, 2007]. Certains auteurs pensent que les gènes hautement exprimés ne tolèreraient aucune insertion d'ET dans leurs régions introniques, ce qui réduit leur taille et diminue le coût de la transcription [Castillo-Davis *et al.*, 2002]. D'autres auteurs avancent la possibilité que l'insertion d'un ET à l'intérieur d'un gène puisse interférer avec la transcription de ce dernier *via* les signaux de terminaison de transcription et les régions promotrices présentes au niveau des ET. La sélection agirait donc contre l'insertion des ET pour empêcher cette interférence transcriptionnelle [Mourier et Willerslev, 2008]. La sélection agit également sur l'orientation des ET, c'est-à-dire leur sens d'insertion par rapport à la région codante, avec par exemple une sélection négative contre l'insertion des LTR en sens par rapport à la région codante [Cutter *et al.*, 2005]. Cependant, d'autres études ont montré que la

sélection pourrait favoriser le maintien de certaines insertions d'ET, comme dans le cas des éléments *Alu* et *B1* (éléments majoritaires des SINE chez l'homme et chez la souris, respectivement) en amont et dans les régions introniques des gènes impliqués dans des catégories fonctionnelles spécifiques [Tsirigos et Rigoutsos, 2009]. Ainsi, la sélection va tendre à contre-sélectionner ou pas les ET dans le voisinage des gènes selon leur fonction.

L'implication de la fonction des gènes dans l'insertion des ET avait été proposée en 2003 par Grover *et al.* [Grover *et al.*, 2003]. Dans cette étude, les auteurs se sont intéressés aux éléments *Alu* au niveau des chromosomes humains 21 et 22, étude dans laquelle ils ont montré que les gènes associés à des fonctions importantes (développement, structure, etc.) sont pauvres en *Alu*, alors que les gènes riches en *Alu* sont associés à des fonctions de "second ordre" (métabolisme et transport). Il est cependant évident que la fonction des gènes n'est pas le seul facteur déterminant dans la distribution des *Alu* puisque le contenu en GC et la densité en gènes sont également impliqués [Grover *et al.*, 2004]. Plus tard, on a montré que l'expression des gènes associée à leur fonction influencent l'insertion des ET dans les régions introniques des mammifères dans le sens où des gènes hautement exprimés et impliqués dans des fonctions spécifiques sont dépourvus d'ET dans leurs introns [Sironi *et al.*, 2006]. Ainsi, l'impact des ET sur l'expression des gènes peut également expliquer leur contre sélection dans certaines régions génomiques même si cet impact n'est pas toujours mis en évidence puisqu'une comparaison de la divergence d'expression entre l'homme et le chimpanzé suite à des insertions spécifiques d'ET dans l'une des deux espèces n'a montré aucun effet significatif des ET [Warnefors *et al.*, 2010]. Néanmoins, l'analyse de l'expression des gènes humains dans les tissus normaux et tumoraux a montré que les SINE sont associés à la dérégulation des gènes en condition tumorale, l'effet étant d'autant plus fort que le nombre de SINE est élevé [Lerat et Sémon, 2007].

Au vu de tout ce qui précède, plusieurs questions subsistent. La fonction des gènes est-elle effectivement un facteur important dans la distribution des ET ? La pression de sélection explique-t-elle réellement l'absence totale des ET dans certains gènes ? Enfin, les ET ont-ils une influence sur l'expression des gènes voisins ? C'est à ces questions qu'on essaiera de répondre dans ce premier chapitre.

3.2 Gènes et voisinage en Éléments Transposables

3.2.1 Chez l'homme

Les coordonnées des gènes humains ont été extraits de la version 50 d'**ENSEMBL** (juillet 2008) et les coordonnées des ET présents dans le génome ont été extraits du fichier ".out" de la version masquée du génome humain. En croisant ces données, j'ai calculé la densité en ET pour chacun des gènes dans les deux régions étudiées (régions flanquantes à 2 et 10 kb). Il faut noter que parmi tous les ET je me suis focalisé sur les ET **complets** seulement parce qu'on considère qu'un ET complet a potentiellement plus de chance d'influencer les gènes voisins, *via* ses régions promotrices, qu'un ET tronqué. Un ET est considéré comme complet quand sa taille est supérieure ou égale à 95% de celle de l'élément de référence (présent dans **Repbase**) et quand la divergence par rapport à ce dernier est $\leq 20\%$. De la même façon, les éléments de type solo-LTR, qui contiennent des séquences régulatrices, ont été considérés comme complets et inclus dans mon analyse si leur taille était supérieure ou égale à 95% et leur divergence $\leq 20\%$ de la LTR de référence. La densité en ET a été ensuite utilisée pour définir quatre classes de gènes allant des gènes sans ET, appelés "TE-free", aux gènes denses en ET, appelés "TE-rich". La **Figure 10** représente le nombre de gènes humains dans chacune des quatre classes et pour les deux types de régions flanquantes.

On remarque que la distribution du nombre de gènes est identique quelle que soit la taille de la région flanquante avec un nombre de gènes qui diminue quand la densité en ET augmente. En effet, un petit nombre de gènes est complètement dépourvu d'ET complets puisque les gènes "TE-free" représentent 17,4% (6185 gènes) et 1,3% (462 gènes) du nombre total de gènes humains analysés (35641 gènes) dans les régions flanquantes de 2 et 10 kb, respectivement. Les gènes qui possèdent une densité en ET comprise entre 0 et 5 constituent 50% (17828 gènes) et 62,2% (22155 gènes) du nombre total de gènes en considérant les régions flanquantes de 2 et 10 kb, respectivement, alors que les gènes qui ont une densité comprise entre 5 et 10 représentent 26,1% (9307 gènes) et 30% (10693 gènes), respectivement. Enfin, les gènes classés comme "TE-rich" (densité ET $>$ 10) représentent 6,5% (2321 gènes) et 6,5% (2331 gènes) du nombre total de gènes avec les régions flanquantes de 2 et 10 kb, respectivement.

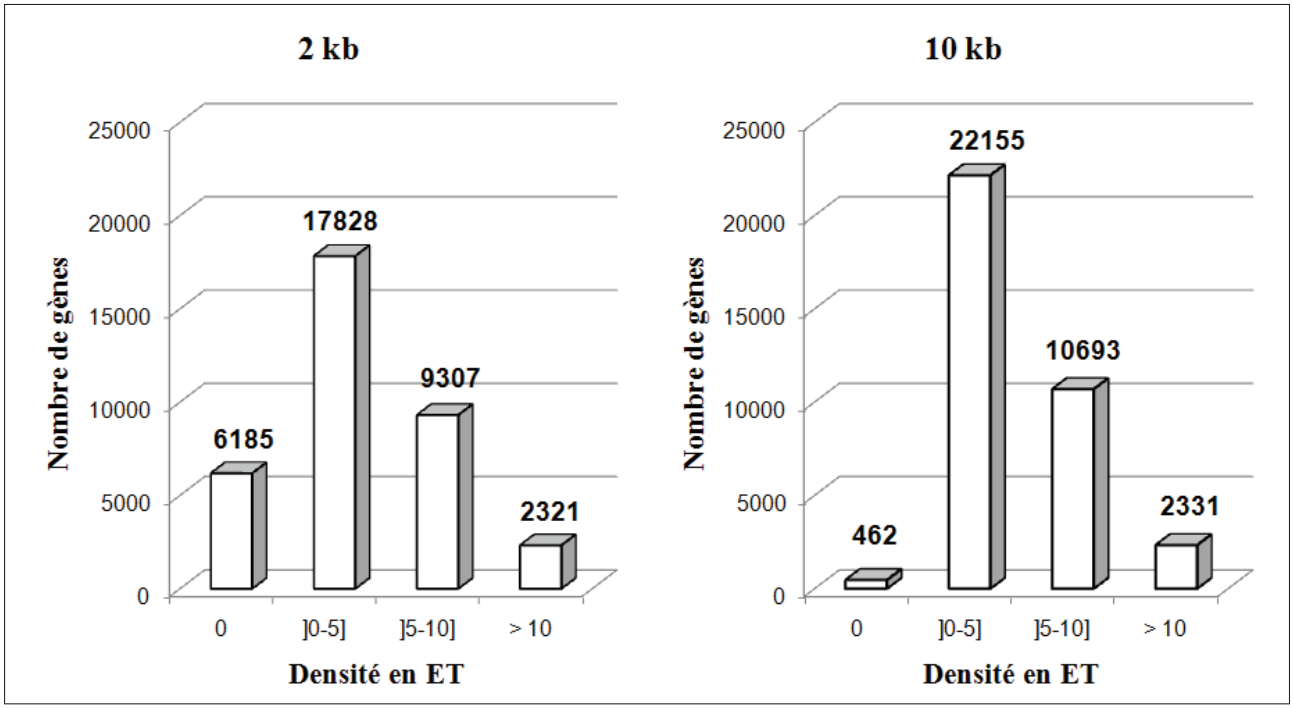

FIGURE 10: *Distribution du nombre de gènes humains en fonction de la densité en ET. La distribution est donnée pour les régions flanquantes de 2 et 10 kb. Le nombre de gènes appartenant à chaque classe est indiqué au-dessus de la barre correspondante.*

J'ai estimé la proportion de chacune des quatre familles d'ET (transposons à ADN, rétrotransposons à LTR, LINE et SINE) parmi les ET complets. Le **Tableau 2** représente le nombre ainsi que la proportion d'ET de chacune des quatre familles. On trouve que les SINE sont les éléments les plus fréquents avec une proportion de 74,6 et 75,8% quand on s'intéresse aux régions flanquantes de 2 et 10 kb respectivement. Les rétrotransposons à LTR constituent 13,7 et 14% des copies d'ET complets, ces proportions étant quasiment égales à celles des transposons à ADN qui constituent 11,2 et 9,8%, respectivement. Enfin, les éléments LINE sont les moins fréquents puisqu'ils représentent seulement 0,5 et 0,4% des copies d'ET complets dans les régions flanquantes de 2 et 10 kb, respectivement. Les proportions des rétrotransposons à LTR et des transposons à ADN sont globalement identiques avec le pourcentage relatif de chacune de ces classes parmi le nombre total d'ET observé dans le génome humain (14% pour les rétrotransposons à LTR et et 9% pour les transposons à ADN). Les SINE possèdent une proportion plus élevée dans mon jeu de données par rapport à leur proportion dans le génome humain (50%) mais ceci est

du à la proportion plus faible des LINE (0,5% dans mon jeu de données contre 27% dans le génome humain).

Tableau 2: *Nombre et proportion d'ET à l'intérieur et dans le voisinage des gènes pour les régions flanquantes de 2 et 10 kb.*

	Régions flanquantes 2 kb		Régions flanquantes 10 kb	
Classe d'ET	Nombre d'ET	Proportion (%)	Nombre d'ET	Proportion (%)
Rétrotransposons à LTR	65076	13,7	100279	14,0
LINE	2344	0,5	2926	0,4
SINE	353857	74,6	541017	75,8
Transposons à ADN	53040	11,2	69783	9,8
Total	474307	100	714005	100

La **Figure 11** donne la distribution du nombre de gènes selon la densité en ET pour chacune des quatre familles considérées. Le nombre de gènes par classe pour les transposons à ADN et les LINE montre la même distribution que celle observée quand tous les ET sont considérés ensemble (**Figure 10**) dans le sens où le nombre de gènes diminue quand la densité de la famille d'ET augmente. Pour les SINE, le nombre des gènes inclus dans la classe $]0 : 0,5]$ est supérieur à celui des gènes totalement dépourvus de SINE, quelle que soit la taille des régions flanquantes. Ceci s'explique par le grand nombre d'éléments SINE (supérieur à 1 million) présents dans le génome humain. Pour les rétrotransposons à LTR, la classe moyenne $]0,5 : 2]$ possède un nombre de gènes plus élevé que celui de la classe $]0 : 0,5]$ et ceci est net pour les régions flanquantes de 10 kb. Cependant, la classe des gènes les plus riches en rétrotransposons à LTR (>2) possède tout de même le plus faible effectif, comme c'est le cas pour les trois autres classes. Pris ensemble, ces résultats montrent que toutes les familles d'ET participent à la tendance de la distribution du nombre de gènes observée pour l'ensemble des ET.

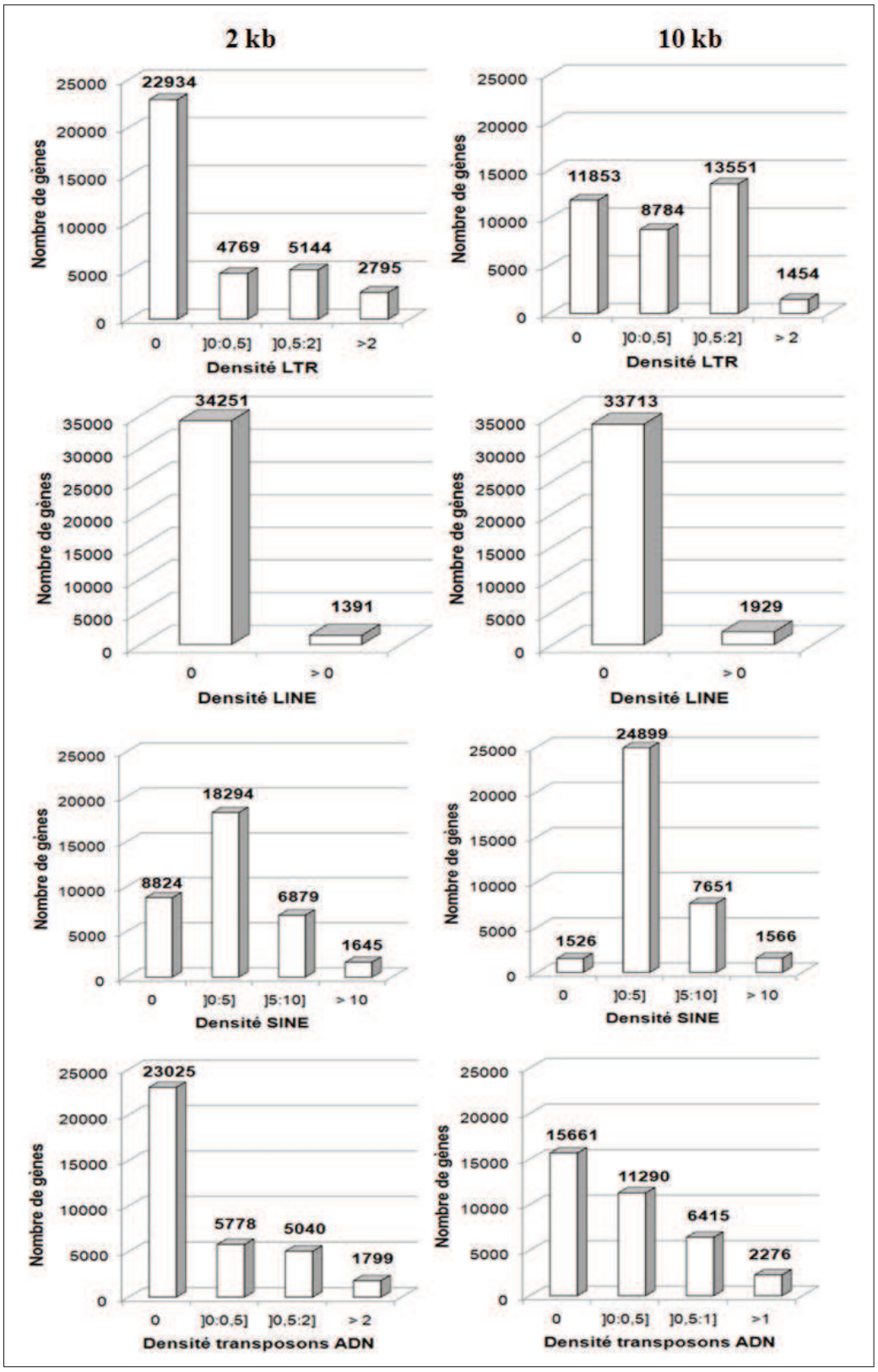

FIGURE 11: *Distribution du nombre de gènes humains pour chacune des quatre familles d'ET (Rétrotransposons à LTR, LINE, SINE et Transposons à ADN) pour les régions flanquantes de 2 et 10 kb. Le nombre de gènes appartenant à chaque classe est indiqué au-dessus de la barre correspondante.*

J'ai alors recherché les trois sous-familles d'éléments majoritaires dans chacune des quatre familles d'ET (**Tableau 3**). Les LINE complets correspondent à un seul type d'élément, *L1*, qui représente 99,57 et 99,66% des LINE quand on s'intéresse aux régions flanquantes de 2 et 10 kb, respectivement. Les SINE sont également constitués majoritairement des trois sous-familles d'éléments *Alu*, les jeunes (*Alu*Y), les intermédiaires (*Alu*S) et les vieux (*Alu*J), qui représentent en tout 90,93 et 91,33% des éléments SINE dans les régions flanquantes de 2 et 10 kb, respectivement. Les rétrotransposons à LTR sont constitués principalement des éléments MALR (62,91 et 59,88%), des ERV1 (21,29 et 23,56%) et ERVL (12,29% dans les régions flanquantes de 2 kb) et ERV2 (12,22% dans les régions flanquantes de 10 kb). Enfin les éléments de type MER (MER1 et MER2) sont majoritaires parmi les transposons à ADN complets puisqu'ils y représentent à eux seuls 81,85 et 81,5% dans les régions flanquantes de 2 et 10 kb, respectivement, tandis que les éléments *Mariner* ne participent qu'à hauteur de 4,49 et 4,42%.

Tableau 3: *Nombre et proportion des sous-familles d'ET à l'intérieur et dans le voisinage des gènes pour les régions flanquantes de 2 et 10 kb.*

	Gènes et régions flanquantes 2 kb			Gènes et régions flanquantes 10 kb		
	% parmi tous les ET	Sous-familles	% dans la classe	% parmi tous les ET	Sous-familles	% dans la classe
Rétrotransposons à LTR	13,72	MALR	62,91	14,05	MALR	59,88
		ERV1	21,29		ERV1	23,56
		ERVL	12,29		ERV2	12,22
		Autres	3,51		Autres	4,34
LINE	0,49	*L1*	99,57	0,41	*L1*	99,66
		Autres	0,43		Autres	0,34
SINE	74,61	*Alu*S	59,71	75,77	*Alu*S	60,55
		*Alu*J	16,31		*Alu*J	15,81
		*Alu*Y	14,91		*Alu*Y	14,97
		Autres	9,07		Autres	8,67
Transposons à ADN	11,18	MER1	58,79	9,77	MER1	58,79
		MER2	23,06		MER2	22,71
		Mariner	4,49		*Mariner*	4,42
		Autres	13,66		Autres	14,08

3.2.2 Chez les primates

Les gènes humains "TE-free" et "TE-rich" conservent-ils la même densité en ET dans les espèces proches ? Afin de répondre à cette question je me suis intéressé à trois espèces, le chimpanzé, le macaque et l'orang-outan, dont les génomes sont complètement séquencés et pour lesquels les coordonnées des ET sont également disponibles (fichiers ".out" des versions prémasquées des trois génomes). J'ai tout d'abord sélectionné les gènes humains qui ont seulement un gène orthologue et un seul dans chacune des trois autres espèces. Ceci m'a permis d'identifier 14744 gènes orthologues, parmi l'ensemble des gènes humains, pour lesquels j'ai déterminé la densité en ET présente à l'intérieur des gènes et dans les régions flanquantes de 2 et 10 kb pour chacune des trois espèces de primates. Sur les 14744 gènes orthologues, 877 sont "TE-free" dans le génome humain quand on considère 2 kb de région flanquante, parmi lesquels 606 gènes (67%) le sont également dans les génomes du chimpanzé, du macaque et de l'orang-outan, alors que 1496 gènes sont "TE-rich" dont seulement 17% (263 gènes) le sont également chez les trois autres espèces. Lorsque je m'intéresse à la région flanquante de 10 kb, sur les 89 gènes humains trouvés comme "TE-free", seuls 50 gènes (56%) le sont également chez les autres primates. Par contre sur les 1849 gènes humains identifiés comme "TE-rich", 289 gènes (15%) gardent leur classification de "TE-rich" sur les trois autres génomes. Globalement, ces résultats indiquent que les gènes "TE-free" ont plus tendance à conserver leur absence d'ET à travers l'évolution par rapport aux gènes "TE-rich". Tout ceci est résumé dans la **Figure 12**.

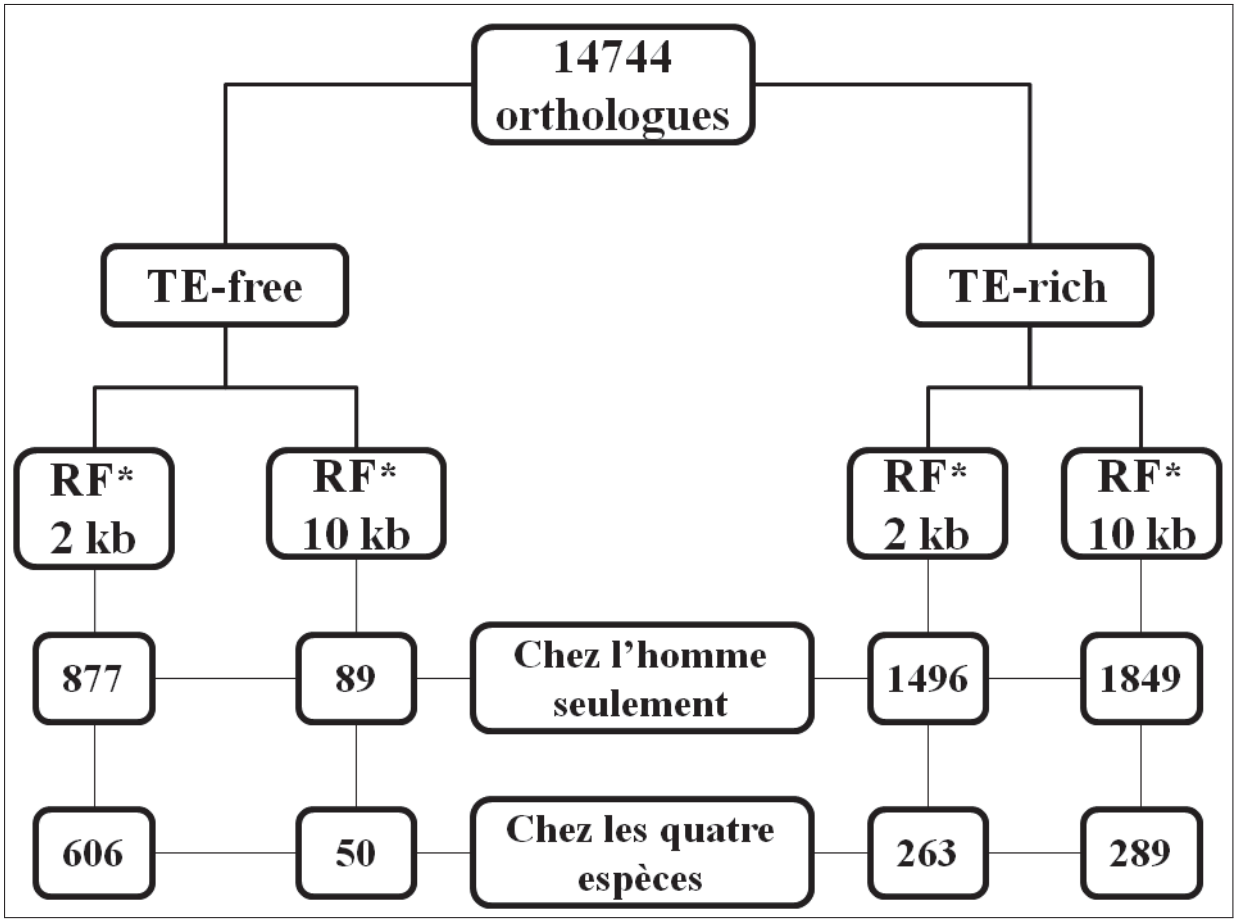

FIGURE 12: *Nombre de gènes TE-free et TE-rich chez l'homme et celui des quatre espèces de primates, dans les régions flanquantes de 2 et 10 kb. RF* : Régions flanquantes.*

3.3 La fonction des gènes selon leur voisinage en Éléments Transposables

Afin de vérifier si la fonction des gènes peut expliquer leur différence en termes de contenu en ET, j'ai tout d'abord comparé les fonctions des gènes appartenant aux deux classes extrêmes, c'est-à-dire les gènes "TE-free" et les gènes "TE-rich". Avec la région flanquante de 2 kb, cette comparaison a été faite entre 6185 gènes "TE-free" et 2321 gènes "TE-rich", alors que pour la région flanquante de 10 kb j'ai comparé 462 gènes "TE-free" à 2331 gènes "TE-rich".

La comparaison des fonctions a été faite *via* le logiciel **FatiGO** qui, après conversion des identifiants géniques en identifiants GO, indique les identifiants GO qui sont significativement surreprésentés dans une liste de gènes par rapport à une autre liste. **FatiGO** classe les résultats selon sept niveaux différents, allant de 3 à 9, le niveau 3 étant le plus général et 9 le plus spécifique. Afin d'être le plus clair possible, j'ai restreint dans la **Figure 13** ma présentation aux niveaux pour lesquels les différences sont nettes et statistiquement significatives entre les gènes "TE-free" et "TE-rich" pour les deux types de régions flanquantes (2 et 10 kb) tout en sachant que l'ensemble des résultats de cette comparaison sont présentés dans les Annexes (section B, page 113). Quand on considère la région flanquante de 2 kb, on remarque que, pour les trois niveaux choisis, les gènes "TE-free" sont plus souvent impliqués que les gènes "TE-rich" dans les fonctions de développement (développement des organismes multicellulaires, de la structure anatomique et du système nerveux), de régulation (processus biologiques et transcription) et de la transcription. Les gènes "TE-free" sont moins fréquemment impliqués que les gènes "TE-rich" dans les différents processus biosynthétiques et métaboliques ainsi que dans le transport de protéines. On observe également la même tendance quand on considère la région flanquante de 10 kb. Cependant, étant donné que notre analyse s'intéresse uniquement aux ET complets, on ne peut pas savoir si la différence fonctionnelle qu'on observe est due à la présence ou l'absence des ET complets dans le sens où les gènes "TE-free" peuvent contenir des ET partiels.

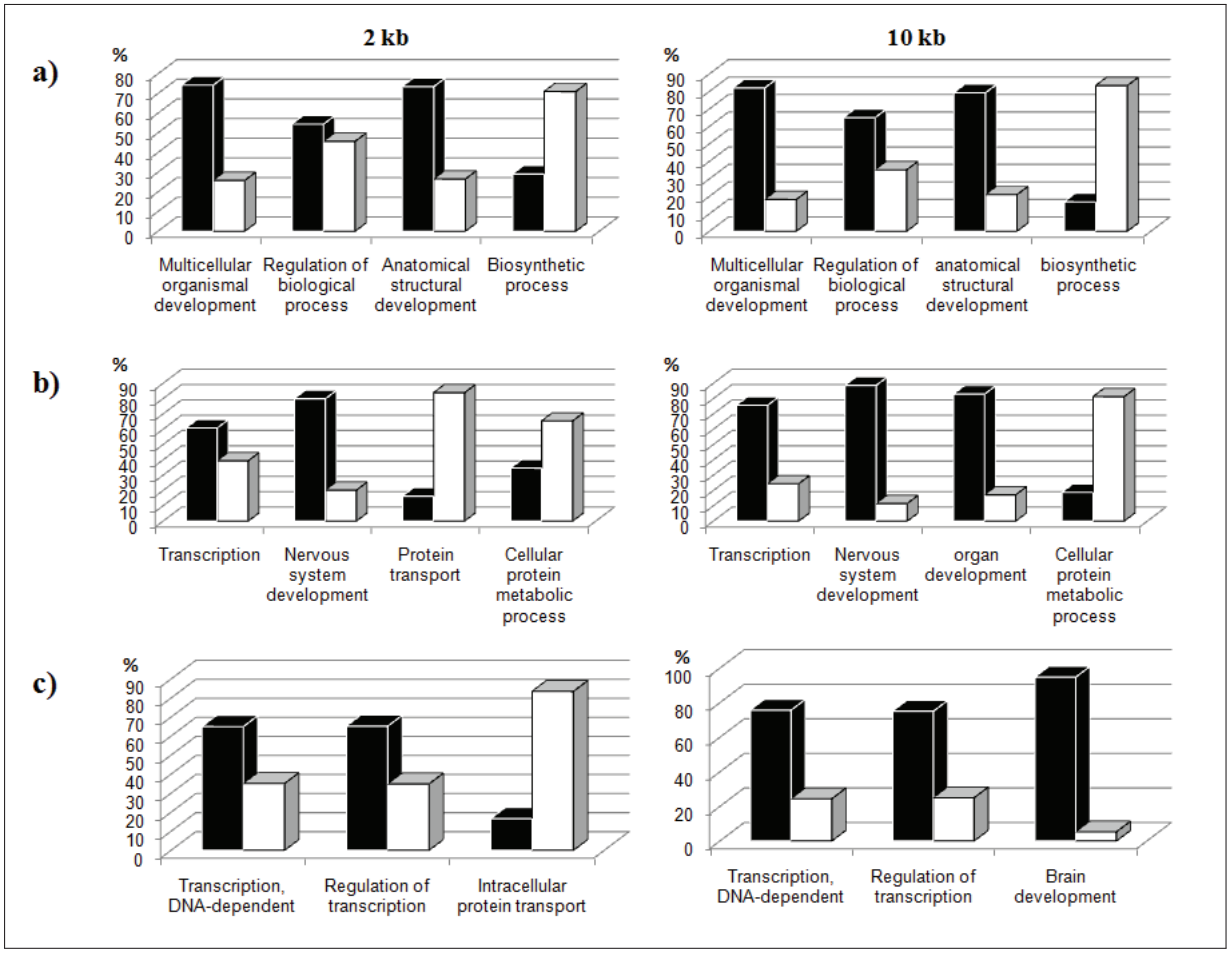

FIGURE 13: *Fonctions des gènes TE-free et TE-rich. Distribution du pourcentage des gènes humains TE-free (barre noire) et TE-rich (barre blanche) impliqués dans les processus biologiques correspondant aux niveaux 3 (a), 5 (b), et 7 (c) dans les régions flanquantes de 2 et 10 kb.*

Afin de tenir compte de ce problème, j'ai subdivisé les gènes "TE-free" en deux sous-classes distinctes selon qu'ils possèdent ou pas des ET partiels. Les gènes dépourvus de toute insertion d'ET (complète ou partielle) sont obtenus en croisant les coordonnées des gènes "TE-free" avec celles des TFR (obtenues à partir de l'étude correspondante [Simons *et al.*, 2006]). J'ai ainsi obtenu 971 gènes que j'ai appellé "TFR-genes". Pour les gènes restants, j'ai regardé le nombre d'ET partiels inclus pour calculer une densité en ET partiels au niveau de ces gènes et définir ainsi une classe de gènes dépourvus d'ET complets mais riches en ET partiels, classe que j'ai nommé "TE-Complete-Free-genes" ou "TCF-genes" (356 gènes). Enfin, parmi les gènes "TE-rich" (ET complets) je me suis intéressé à ceux qui ne possèdent pas d'ET partiels dans leurs séquences pour former la classe

"TE-Partial-Free-genes" ou "TPF-genes" (386 gènes). J'ai alors comparé les fonctions des gènes TCF et TPF à celles des TFR. Aucune différence fonctionnelle significative n'a été remarquée entre les gènes TFR et les gènes TCF. Par contre, la comparaison des fonctions des gènes TFR avec les gènes TPF a montré la même tendance que celle observée entre les gènes "TE-rich" et "TE-free" (**Tableau 4**). Ces résultats confirment que la fonction des gènes varie suivant la présence/absence des ET complets dans leur séquence ainsi que leur voisinage.

Tableau 4: *Différences fonctionnelles des gènes selon leur voisinage en ET complets ou partiels par rapport aux gènes TFR. * : p-value significative (< 0,05).*

Termes GO	P-value ajustée gènes TFR versus gènes TCF	P-value ajustée gènes TFR versus gènes TPF
Développement organismes multicellulaires	$0,284$	$3,25E-07^*$
Développement structure anatomique	$0,619$	$3,25E-07^*$
Régulation de processus biologiques	$0,686$	$6,71E-06^*$
Transcription	$0,395$	$2,52E-04^*$
Développement du système nerveux	$1,000$	$0,014^*$
Communication cellulaire	$1,000$	$0,043^*$
Processus neurologique	$0,619$	$0,049^*$
Transduction du signal	$1,000$	$0,026^*$
Perception sensorielle du stimulus chimique	$0,649$	$0,014^*$
Perception sensorielle de l'odeur	$0,888$	$0,043^*$

La comparaison des fonctions entre les deux classes extrêmes "TE-rich" et "TE-free" étant significative, j'ai comparé entre elles les fonctions des quatre classes prises deux à deux. Les résultats obtenus montrent la même tendance que ci-dessus lorsque les gènes de la classe "TE-free" sont comparés avec ceux des trois autres classes. Par contre la comparaison des fonctions entre les gènes "TE-rich" et ceux des classes intermédiaires (1 et 2) ne montre pas de différences de fonctions significatives. Ainsi, pour les analyses suivantes je me suis focalisé uniquement sur les deux classes géniques extrêmes ("TE-free" et "TE-rich").

Une question importante était de savoir si cette différence fonctionnelle entre les gènes "TE-free" et "TE-rich" était conservée chez les trois autres espèces de primates. Cependant, cette analyse n'étant pas réalisable *via* **FatiGO** car le chimpanzé, le macaque, et l'orang-outan ne sont pas disponibles dans la liste des espèces de **FatiGO**, j'ai comparé les fonctions des 606 gènes humains qui sont "TE-free" chez les quatres espèces de primates avec les 263 gènes qui sont "TE-rich" dans ces mêmes espèces. On peut ensuite extrapoler les résultats de cette comparaison aux espèces proches parce que les gènes orthologues sont supposés être impliqués dans les mêmes fonctions. Les résultats obtenus montrent clairement que les gènes "TE-free" sont impliqués dans des fonctions de développement et de régulation alors que les gènes "TE-rich" sont impliqués dans les fonctions de transport et de processus métaboliques ceci, dans les quatre espèces analysées.

3.4 Pressions de sélection

3.4.1 Au niveau des régions codantes

La différence de densité en ET au niveau des gènes selon leur fonction peut être expliquée par une pression de sélection qui s'exerce sur les régions codantes de ces gènes. Selon cette hypothèse, les gènes "TE-free", impliqués dans les fonctions de régulation et de structure, seraient soumis à une sélection purificatrice qui s'opposerait à la fixation des ET dans leur séquence. Cette sélection serait plus relâchée au niveau des gènes "TE-rich", impliqués dans les fonctions de métabolisme et de transport permettant ainsi la fixation de certaines insertions d'ET, ce qui augmente leur densité dans ces gènes. Dans le but d'éprouver cette hypothèse, j'ai calculé la valeur ω pour les gènes "TE-free" et "TE-rich" orthologues de l'homme et du chimpanzé. Cette analyse a été faite sur 1377 gènes "TE-free" et 824 gènes "TE-rich" quand on considère la région flanquante de 2 kb et sur 121 gènes "TE-free" et 982 gènes "TE-rich" quand 10 kb de régions flanquantes sont considérés. Les résultats de cette analyse indiquent une moyenne de ω plus élevée pour les gènes "TE-free" avec les deux types de régions flanquantes (0,55 *versus* 0,43 à 2 kb et 0,52 *versus* 0,40 à 10 kb) ce qui suggère une pression purificatrice plus forte au niveau des gènes "TE-rich". Cette différence de pression de sélection n'est cependant significative que lorsque les régions flanquantes de 2 kb sont considérées (Test de Wilcoxon : W =

522711,5, p-value = 0,001928 à 2 kb ; W = 54096,5, p-value = 0,1065 à 10 kb). Ces résultats indiquent que la pression de sélection sur les régions codantes des gènes n'est pas suffisante à elle seule pour expliquer leur variation en densité d'ET. Cependant, la séparation récente entre l'homme et le chimpanzé (estimée à 6 millions d'années) peut être à l'origine de ces résultats dans le sens où la proximité évolutive des deux espèces ne permettrait pas de détecter d'éventuelles traces de la pression de sélection.

Dans le but d'augmenter la puissance de détection, je me suis intéressé aux quatre espèces de primates : homme, chimpanzé, macaque et orang-outan. J'ai considéré dix scénarios différents selon la variation d'enrichissement en ET entre les gènes orthologues de ces quatre espèces (**Figure 14**). Dans les scénarios 1 et 2, les gènes orthologues appartiennent tous à la même classe ("TE-rich" pour le scénario 1 et "TE-free" pour le scénario 2) et deux modèles ont été testés : le modèle M0 qui assume une pression de sélection identique sur toutes les branches de l'arbre (**Figure 14**) et le modèle M1 qui suppose une pression de sélection différente sur chaque branche de l'arbre. Le ratio de vraisemblance calculé indique que le modèle M0 est celui qui correspond le mieux aux données des scénarios 1 et 2 (∼92% des gènes, **Tableau 5**). Ceci indique que toutes les branches de l'arbre possèdent la même valeur ω et que les gènes qui appartiennent à la même classe de densité en ET sont soumis à la même pression de sélection.

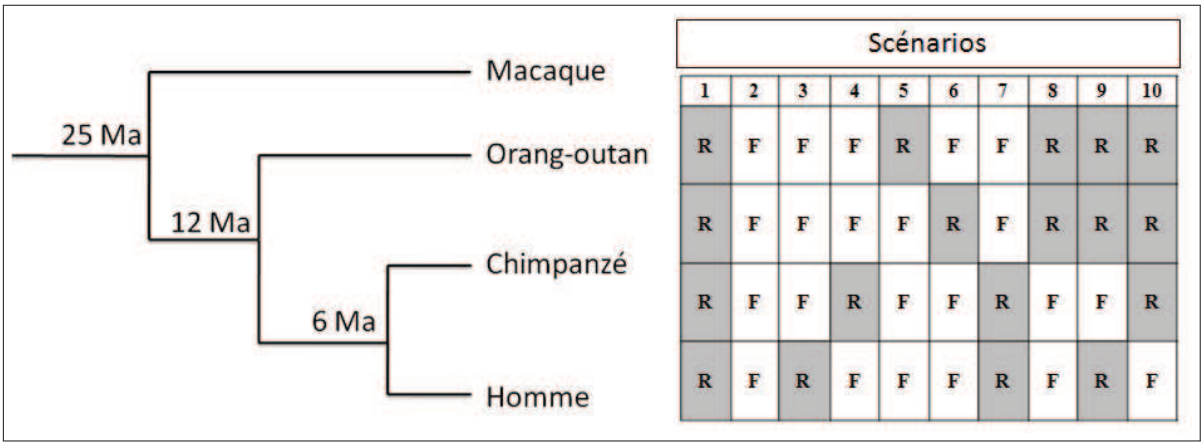

Figure 14: *L'arbre phylogénétique des quatre espèces de primates étudiées dans cette analyse est représenté à gauche avec sur chaque branche une estimation de la date de divergence (en millions d'années). À droite sont représentés les différents scénarios étudiés. R correspond à un gène riche en ET et F à un gène dépourvu d'ET.*

Dans les scénarios 3 à 10, les gènes orthologues peuvent appartenir à différentes classes de densité en ET, pour une ou plusieurs espèces. Ainsi, j'ai testé pour ces différents scénarios si la pression de sélection explique la variation de la densité en ET entre les gènes orthologues. Si on considère le scénario 3 à titre d'exemple, on remarque que les gènes humains sont classés "TE-rich" alors que leurs orthologues dans les trois autres espèces de primates sont "TE-free". Ceci peut être la conséquence d'une pression de sélection plus relâchée au niveau des gènes humains par rapport à leurs orthologues. Pour tester cette hypothèse, j'ai considéré le modèle M2, qui assume une valeur de pression de sélection différente ω_1 sur une branche spécifique de l'arbre (branche du gène humain dans le cas du scénario 3) par rapport à la valeur de la pression de sélection ω_0 qui agit sur les branches restantes de l'arbre. J'ai ensuite comparé le modèle M2 au modèle M0. Le **Tableau 5** montre que le modèle M0 est celui qui correspond le mieux aux données quel que soit le scénario considéré (3 à 10), indiquant que la pression de sélection est identique sur toutes les branches de l'arbre. Ainsi, malgré la différence de densité en ET entre les gènes orthologues, leur pression de sélection ne montre pas de différence significative.

Tableau 5: *Pour chaque scénario considéré, le nombre de gènes analysés est indiqué ainsi que le nombre de gènes qui correspond à chacun des modèles testés. Le pourcentage de gènes dans chaque modèle est indiqué entre parenthèses.*

Scénario	Nombre total de gènes analysés	Nombre de gènes correspondants au Modèle 0 (%)	Nombre de gènes correspondants au Modèle 1 (%)	Nombre de gènes correspondants au Modèle 2 (%)
1	1593	1456 (91,4)	126 (7,9)	-
2	335	309 (92,2)	24 (7,2)	-
3	174	167 (95,9)	-	7 (4,1)
4	37	34 (91,9)	-	3 (8,1)
5	194	167 (86,1)	-	24 (12,4)
6	33	27 (81,8)	-	4 (12,1)
7	215	198 (92,1)	-	17 (7,9)
8	34	32 (94,2)	-	1 (2,9)
9	95	83 (87,4)	-	11 (11,6)
10	24	22 (91,7)	-	2 (8,3)

Comme pour la comparaison entre l'homme et le chimpanzé, on peut se demander si la distance phylogénétique proche entre les quatre espèces de primates considérées dans notre analyse (le dernier ancêtre commun de ces quatre espèces est estimé à ∼25 millions d'années) est insuffisante pour permettre une détection d'un effet possible de la pression de sélection. En effet le taux d'identité important entre les séquences codantes des gènes orthologues peut fausser le calcul de la valeur ω à cause du phénomène de saturation. Afin de s'affranchir de cette saturation éventuelle, j'ai analysé les gènes orthologues de la souris dont la séparation avec l'homme est datée à ∼75 millions d'années. L'analyse a porté sur 991 gènes orthologues "TE-free" et 745 gènes "TE-rich" pour la région flanquante de 2 kb, et sur 91 gènes "TE-free" et 874 gènes "TE-rich" pour 10 kb de région flanquante. La moyenne de ω est significativement plus faible pour les gènes "TE-free" (ω=0,15 à 2 kb de région flanquante et ω=0,09 à 10 kb) par rapport aux gènes "TE-rich" (ω=0,17 à 2 kb de région flanquante et ω=0,16 à 10 kb) quelle que soit la taille des régions flanquantes considérées (test de wilcoxon : W = 334503 ; p-value = 0,0007982 pour la région flanquante de 2 kb ; W = 25977 ; p-value = 4,942e-08 à 10 kb de région flanquante). Ces résultats indiquent que la pression de sélection sur les régions codantes des gènes explique leurs différences en terme de densité en ET, avec une pression de sélection négative (ou purificatrice) qui s'exerce plus fortement sur les gènes "TE-free" que sur les gènes "TE-rich".

3.4.2 Au niveau des régions flanquantes

Étant donné que notre analyse des effets de la sélection n'a concerné que les régions codantes des gènes et que le calcul de la densité en ET pour un gène donné avait inclus ses régions flanquantes, on peut se demander si la pression de sélection agissant sur les régions flanquantes peut, elle aussi, être impliquée dans la variation de densité en ET. Puisque, la valeur de ω ne peut être calculée qu'au niveau des régions codantes, j'ai utilisé le taux d'identité de séquences comme un révélateur de la pression de sélection sur ces régions flanquantes. On considère que deux séquences d'ADN ayant un taux d'identité élevé sont très probablement soumises à une pression de sélection purificatrice plus forte par rapport à deux séquences ayant un taux d'identité faible. J'ai ainsi analysé le taux d'identité pour les régions flanquantes situées à 2 et 10 kb, en amont et en aval, des gènes

orthologues. Pour chaque gène humain, j'ai récupéré ses séquences flanquantes à 2 et 10 kb en amont et en aval, ainsi que celles de ses orthologues chez le chimpanzé, le macaque et l'orang-outan. La **Figure 15** représente les distributions du taux d'identité pour les différentes régions flanquantes considérées (en amont et en aval des gènes à 2 et 10 kb) des deux classes de gènes ("TE-rich" ou "TE-free").

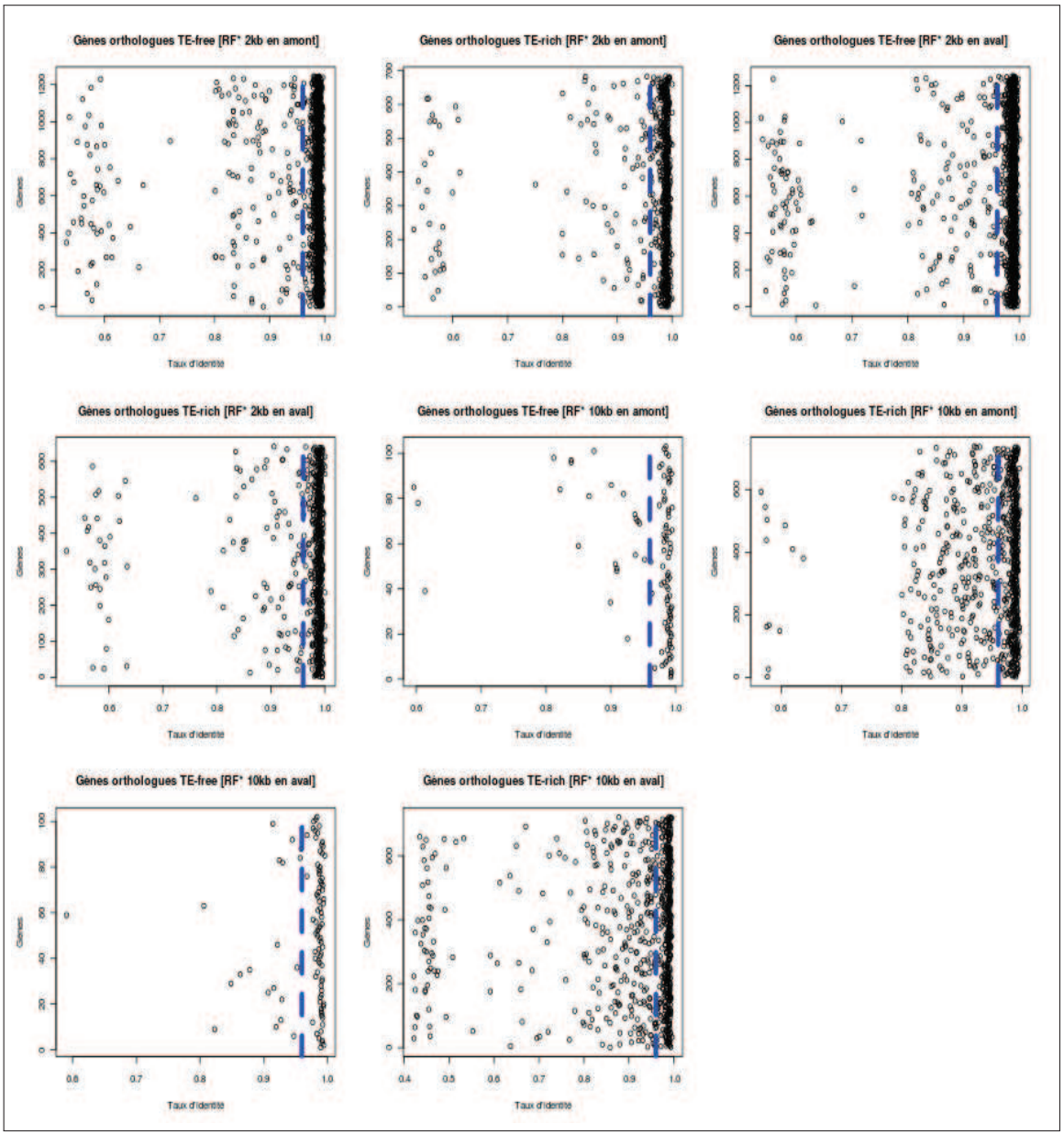

FIGURE 15: *Distributions du taux d'identité entre les séquences flanquantes de gènes humains et celles de leurs orthologues chez le chimpanzé pour les gènes TE-free et TE-rich. Le trait bleu correspond au pourcentage d'identité 97%. Le nombre de gènes est représenté sur l'axe des ordonnées et le taux d'identité sur l'axe des abscisses. RF* : Régions Flanquantes.*

Les résultats de l'analyse sont représentés dans le **Tableau 6**. Dans la comparaison homme-chimpanzé, le taux d'identité moyen des séquences flanquantes des gènes "TE-free" est significativement supérieur à celui des séquences encadrant les gènes "TE-rich" uniquement lorsqu'on considère les régions flanquantes de 10 kb (en amont et en aval des gènes). Cependant, la différence n'est pas significative pour les régions flanquantes de 2 kb. Pour la plupart des séquences flanquantes, les pourcentages d'identité sont globalement en accord avec le taux d'identité moyen entre l'homme et le chimpanzé (le taux de divergence nucléotidique entre les deux espèces varie entre 0,5 et 3% [Chimpanzee Sequencing and Analysis Consortium, 2005]). Par contre les plus faibles valeurs de pourcentage moyen d'identité s'expliquent par une minorité de séquences orthologues qui possèdent une forte divergence de séquences (**Figure 15**). Quand on compare les séquences flanquantes des gènes humains avec celles de leurs orthologues chez l'orang-outan, les taux d'identité de séquences flanquantes des gènes "TE-free" en amont pour les régions flanquantes de 2 kb et en aval pour les régions flanquantes de 2 et 10 kb sont significativement supérieurs à ceux des séquences flanquantes des gènes "TE-rich" (**Tableau 6**). Enfin, dans la comparaison homme/macaque, les séquences en amont et en aval des gènes "TE-free" pour les deux types de régions flanquantes (2 et 10 kb) montrent des taux d'identités significativement supérieurs à ceux des séquences en amont et en aval des gènes "TE-rich". Ces résultats indiquent que les régions flanquantes des gènes "TE-free" sont plus conservées à travers l'évolution que celles des gènes "TE-rich" ce qui peut s'expliquer par une pression de sélection purificatrice plus forte au niveau des séquences flanquantes des gènes "TE-free".

Tableau 6: *Taux d'identité des régions flanquantes des gènes orthologues TE-free et TE-rich dans les génomes de l'homme, du chimpanzé, de l'orang-outan et du macaque. * : p-value significative (< 0,05).*

Espèces comparées	Taille de la région flanquante (kb)	Nombre comparé de gènes TE-free	Nombre comparé de gènes TE-rich	Position de la région flanquante par rapport au gène	Moyenne du % d'identité des régions flanquantes des gènes TE-free	Moyenne du % d'identité des régions flanquantes des gènes TE-rich	P-value
Homme *vs* chimpanzé	2	1246	684	En amont En aval	96,3 95,7	96,1 96,0	0,60 0,29
	10	103	738	En amont En aval	95,8 96,9	94,7 90,9	$1,6E-04^*$ $9,2E-08^*$
Homme *vs* orang-outan	2	1178	532	En amont En aval	93,5 94,6	92,4 94,3	$2,2E-04^*$ 0,67
	10	95	445	En amont En aval	94,5 95,2	93,1 94,4	$1,8E-04^*$ $0,02^*$
Homme *vs* macaque	2	1080	445	En amont En aval	90,9 91,8	87,3 90,6	$8,9E-16^*$ $3,7E-02^*$
	10	89	335	En amont En aval	92,5 92,9	89,1 90,5	$3,1E-06^*$ $3,9E-09^*$

3.5 Expression des gènes

Afin de tester l'influence des ET sur l'expression des gènes voisins (cf. Introduction, section 1.3.2, page 12), j'ai comparé l'expression des gènes "TE-free" et "TE-rich". Les données d'expression utilisées proviennent de l'analyse de Su *et al.* [Su *et al.*, 2004] qui ont quantifié l'expression des gènes protéiques dans 79 tissus humains. Dans ces données, l'identifiant principal est celui de la sonde qui peut être associée à aucun, un ou plusieurs gènes. Afin de pouvoir associer de façon non ambigüe un gène à une seule sonde, j'ai éliminé de mon analyse tous les gènes associés à plusieurs sondes ainsi que les sondes présentes plusieurs fois. Ensuite j'ai croisé les données des gènes "TE-free" et "TE-rich" avec celles des données d'expression, ce qui m'a permis d'obtenir les niveaux d'expression pour 239 gènes "TE-free" et 158 gènes "TE-rich". La **Figure 16** donne les niveaux d'expression des gènes "TE-free" et "TE-rich" pour 19 tissus humains parmi les 79. Ces 19 tissus ont été choisis afin de pouvoir représenter tous les tissus qui montrent des différences significatives d'enrichissement tout en sachant que les niveaux d'expression des 60 tissus restants sont représentés dans les Annexes (**Figure 32**, page 112). Globalement, le niveau d'expression des gènes "TE-rich" est supérieur à celui des gènes "TE-free" et cette différence est significative pour 13 tissus (indiqués par un * sur la **Figure 16**). Cette différence d'expression a été observée dans trois tissus tumoraux parmi les six et dans les dix tissus immunitaires analysés.

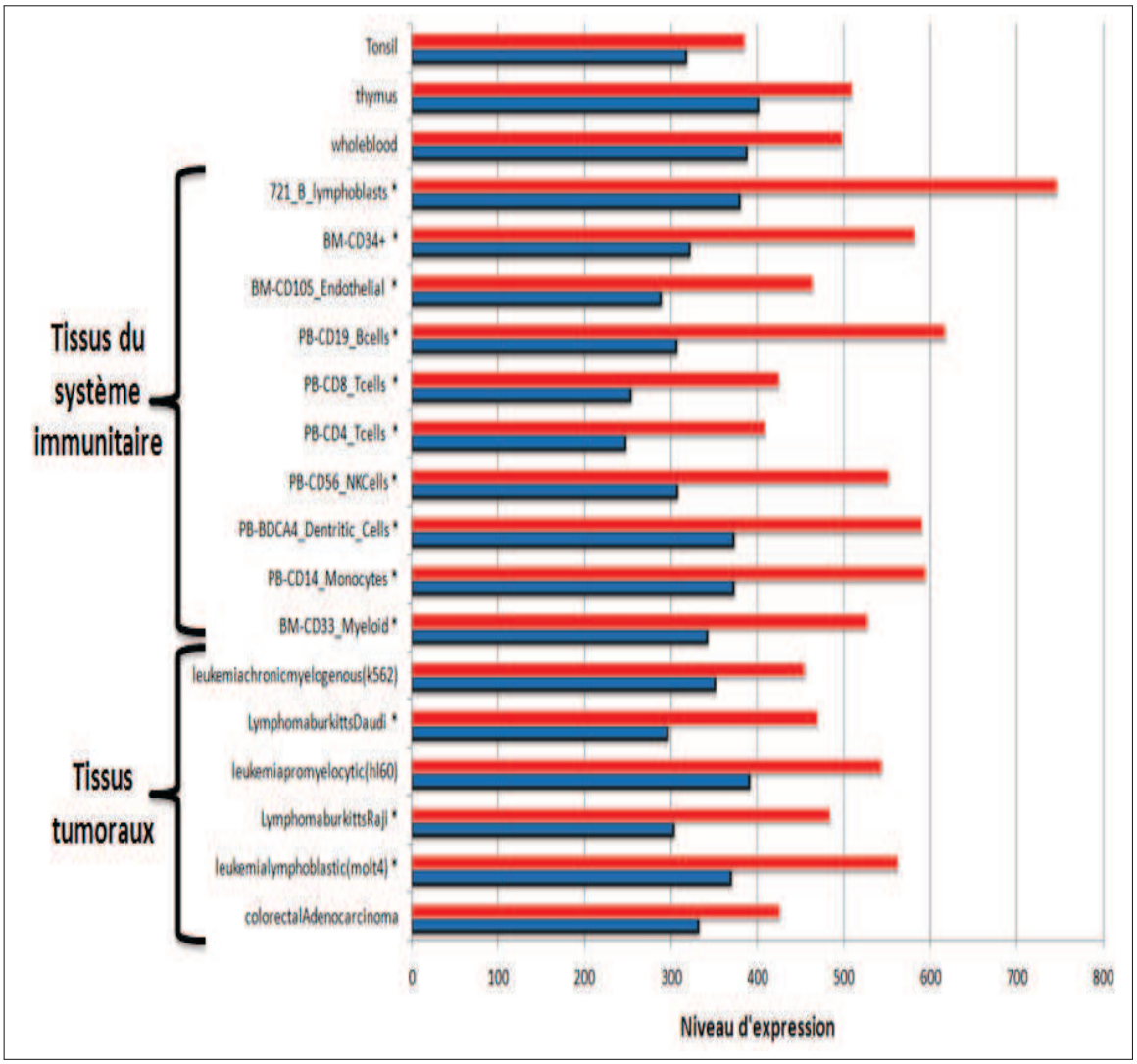

FIGURE 16: *Niveaux d'expression des gènes TE-free et TE-rich pour 19 tissus humains. Seule la région flanquante de 2 kb a été considérée. les barres rouges indiquent les gènes TE-rich et les barres bleues correspondent aux gènes TE-free. Les astérisques indiquent des différences significatives entre les deux catégories géniques (p-value$< 0,05$).*

3.6 Conclusion

Dans ce chapitre, je me suis penché sur les facteurs qui peuvent influer sur la distribution des ET. Je me suis intéressé à un facteur particulier, qui est la fonction des gènes dans le voisinage des insertions d'ET. J'ai identifié les ET complets et calculé la densité en ET complets pour tous les gènes en considérant les régions flanquantes de 2 et 10 kb. Cette densité m'a permis de diviser les gènes en plusieurs classes allant des gènes dépourvus de toute insertion d'ET complets ("TE-free") aux gènes riches en ET ("TE-rich"). La comparaison des fonctions de ces deux types de gènes a montré que les gènes impliqués dans des fonctions de régulation et de développement possèdent peu, voire aucune insertion d'ET complets, alors que les gènes impliqués dans les processus métaboliques et dans le transport possèdent une forte densité en ET. Cette tendance est conservée chez les primates. J'ai montré que la variation de la densité en ET entre les gènes s'explique par une pression de sélection purificatrice plus forte chez les gènes "TE-free", qui s'exerce non seulement au niveau des régions codantes mais également au niveau des régions flanquantes des gènes. Enfin, la comparaison du niveau d'expression des gènes "TE-free" et "TE-rich" a permis de voir que ces derniers possèdent un niveau d'expression plus élevé, particulièrement dans les tissus du système immunitaire et les tissus tumoraux, supposant ainsi un rôle important des ET dans la régulation des gènes impliqués dans ces processus.

Chapitre 4

Modifications d'histones, expression génique et cancer

4.1 Introduction

Le cancer est aujourd'hui la 3ème cause de décès dans le monde [(WHO), 2008] et est en phase de devenir la première dans le futur proche comme c'est déjà le cas dans certains pays comme la France [Aouba *et al.*, 2007]. Ainsi, il est très important de comprendre toutes les dérégulations génomiques dans les cellules cancéreuses. En effet, une meilleure compréhension des dysfonctionnements survenant pendant un cancer permettra de trouver certainement des moyens plus efficaces pour lutter contre sa progression. Les travaux de recherche sur le cancer se sont concentrés au départ sur les bases génétiques de cette maladie en cherchant tous les changements impliquant des gènes (mutations, réarrangements chromosomiques) qui peuvent aboutir au cancer [Hanahan et Weinberg, 2000]. On a pu ainsi identifier un certain nombre de gènes impliqués plus ou moins directement comme par exemple les gènes suppresseurs de tumeurs et les oncogènes. Néanmois, il existe d'autres mécanismes tout aussi importants comme par exemple les modifications épigénétiques [Barbacid, 1987] qui font actuellement l'objet de recherches intensives. Ainsi, un consortium [American Association for Cancer Research Human Epigenome Task Force, 2008] a été créé dans le but d'établir un épigénome humain complet ainsi que plusieurs épigénomes de référence, et de les rendre accessibles à la communauté scientifique *via* des bases de données.

Les processus épigénétiques incluent trois composants majeurs : la méthylation de l'ADN, les modifications d'histones et l'interférence par l'ARN [Bernstein *et al.*, 2007] (voir chapitre Introduction, section 1.6, page 21). La dérégulation d'une ou plusieurs composantes épigénétiques peut dans certains cas aboutir à un état cancéreux [Esteller, 2007]. Ainsi, la méthylation de l'ADN représente la composante épigénétique la plus dérégulée dans les cellules cancéreuses. À l'état normal, 3 à 6% des cytosines sont méthylées dans les cellules humaines mais cette méthylation n'est pas distribuée aléatoirement puisqu'elle est localisée préférentiellement sur les dinucléotides CpG (Cytosine-phosphate-Guanine) sauf dans les régions riches en CpG appelées "ilôts CpG", qui ne sont pas méthylées [Esteller, 2005]. Dans les cellules cancéreuses, les ilôts CpG situés dans les régions promotrices des gènes suppresseurs de tumeurs sont hyperméthylés, ce qui résulte en leur répression transcriptionnelle ([Esteller *et al.*, 2001], [Esteller, 2002]),

tandis qu'à un niveau plus général le génome subit une hypométhylation globale ([Feinberg et Vogelstein, 1983], [Goelz *et al.*, 1985]). Les miRNA et les modifications d'histones sont impliqués également dans le développement et la progression du cancer ([Silahtaroglu et Stenvang, 2010], [Ryan *et al.*, 2010], [Muntean et Hess, 2009]).

La dérégulation des différents processus épigénétiques à l'état tumoral peut engendrer une réactivation potentielle des ET. En effet, plusieurs études rapportent une expression de certains ET à l'état tumoral comme celle des LINE-1 (*L1*) dans certains cancers [Smith *et al.*, 2007] et des HERV dans différentes cellules cancéreuses comme le cancer du sein [Wang-Johanning *et al.*, 2001], le cancer de l'ovaire ([Wang-Johanning *et al.*, 2007], [Menendez *et al.*, 2004]) et les lignées de cellules leucémiques [Patzke *et al.*, 2002]. La majorité des dérégulations de gènes associés aux ET rapportées jusqu'à aujourd'hui concernent la méthylation de l'ADN, alors que peu d'études se sont intéressées aux modifications d'histones. La seule étude des modifications d'histones associées aux ET dans un état cancéreux a montré que les altérations touchent principalement l'histone H4 qui connaît une diminution de l'enrichissement de la monoacétylation au niveau de la lysine 16 (H4K16ac) et de la triméthylation au niveau de la lysine 20 (H4K20me3) [Fraga *et al.*, 2005].

Dans ce chapitre, j'ai comparé les variations d'enrichissement de plusieurs modifications d'histones entre une lignée de cellules extraites des fibroblastes du poumon d'une personne saine ("Normal Human Fibroblast Lung" ou NHFL) et une lignée de cellules cancéreuses extraites des poumons d'une personne atteinte d'une leucémie myéloide chronique (K562). Ces données ont permis d'aborder les questions suivantes :(i) quelles sont les modifications d'histones qui subissent des variations entre les deux états normaux et cancéreux ; (ii) ces variations sont-elles localisées dans une région spécifique des gènes ; (iii) la présence/absence des ET au niveau des gènes est-elle associée à des modifications spécifiques ; (iv) les variations des modifications d'histones sont-elles corrélées avec les variations de l'expression des gènes ?

4.2 Préparation des données

Les données d'expression utilisées dans cette analyse sont les mêmes que celles qui ont été utilisées dans le chapitre 1 [Su *et al.*, 2004]. Parmi ces données, j'ai extrait celles des deux tissus : le poumon normal et la lignée de cellules cancéreuses K562. Le niveau d'expression est disponible pour 6957 gènes **ENSEMBL**. Parmi ces gènes, j'ai éliminé les gènes qui correspondent à plusieurs sondes (107 gènes) et les gènes présents plusieurs fois dans le jeu de données (3233 gènes), ce qui a réduit mon jeu de données de départ à 3617 gènes uniques correspondant à des sondes uniques, pour lesquels j'ai récupéré les coordonnées géniques (début gène, fin gène et chromosome) *via* **BioMart**. J'ai ensuite éliminé de ce jeu de données les gènes situés sur les chromosomes Y et mitochondriaux puisqu'ils ne sont pas considérés dans les données des modifications d'histones. Ceci ajouté au fait que certains gènes n'étaient plus présents dans la version 50 d'**ENSEMBL** a réduit le jeu de données à 3350 gènes pour lesquels j'ai calculé la divergence d'expression entre les états normal et tumoral (cf. Matériel et Méthodes, section 2.6.1, page 44).

J'ai téléchargé le fichier "signal" d'enrichissement correspondant à chacune des huit modifications d'histones suivantes : H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K27ac, H3K27me3, H3K36me3 et H4K20me1. J'ai ensuite extrait l'enrichissement correspondant à chaque gène pour chaque modification d'histone (cf. Matériels et Méthodes, section 2.5, page 43) et ceci dans les deux états. Pour chaque gène du jeu de données de départ, j'ai considéré dix régions distinctes. La région "gène entier" va de 1 kb en amont du gène jusqu'à sa fin. Cette région a été divisée en trois sous-régions : la région "promoteur", qui va du début du gène jusqu'à 1 kb en amont, la région "exons", qui correspond à l'ensemble des exons uniques du gène, la région "introns", qui est déduite à partir des coordonnées de la région "exons". Chacune des trois sous-régions ("promoteur", "exons" et "introns") a ensuite été séparée suivant la présence ou l'absence des séquences d'ET. J'ai enfin récupéré l'enrichissement pour chaque région ("promoteur", "exons" et "introns", avec ou sans ET) de la même façon que ca avait été fait pour le gène entier.

4.3 Divergence d'expression génique entre les états normal et tumoral

La divergence d'expression génique entre les états normal et tumoral a été calculée par deux méthodes différentes : la distance euclidienne et la distance moyenne (cf. Matériel et Méthodes, section 2.6.1, page 44). La méthode basée sur la distance de corrélation [Glazko et Mushegian, 2010], qui permet de calculer, pour chaque gène, un coefficient de corrélation entre les niveaux d'expression géniques des deux états, ne pouvait pas être utilisée puisqu'on a seulement deux mesures pour chaque tissu dans chacun des états. Ensuite pour chaque méthode, quatre classes de gènes ont été définies suivant leur divergence d'expression, allant de la classe 1 (peu ou pas de divergence) à la classe 4 (grande divergence d'expression). La définition des différentes classes s'est faite en se basant sur la distribution des gènes pour la distance euclidienne (**Figure 17**) et en utilisant l'algorithme K-means pour la distance moyenne (**Figure 18**).

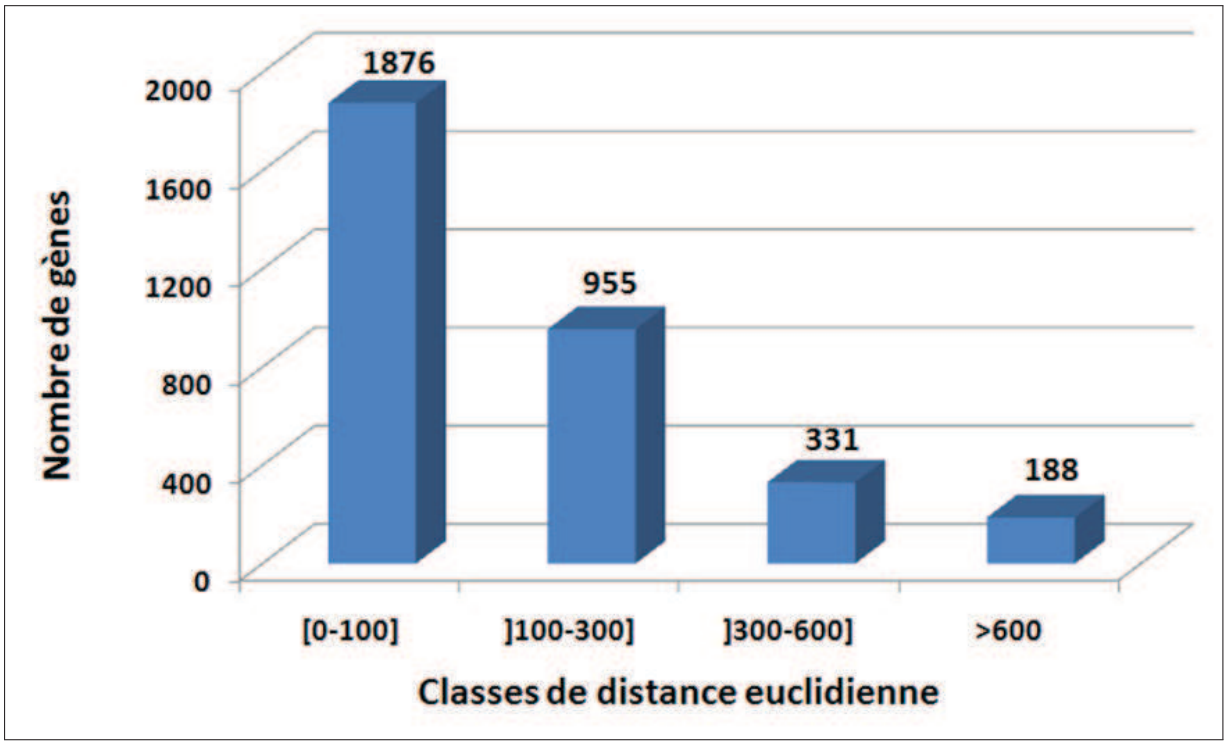

FIGURE 17: *Distribution du nombre de gènes humains selon leur classe de densité euclidienne. Le nombre de gènes appartenant à chaque classe est indiqué au dessus de la barre correspondante.*

La distribution du nombre de gènes est identique pour les deux méthodes avec un nombre de gènes plus faible pour une divergence d'expression plus grande. Pour la distance euclidienne, on compte 1876 gènes dans la classe 1, 955 gènes dans la classe 2, 331 gènes dans la classe 3 et 188 gènes dans la classe 4. Pour la distance moyenne, on compte 1444 gènes dans la classe 1, 928 gènes dans la classe 2, 635 gènes dans la classe 3 et 343 gènes dans la classe 4. Enfin, étant donné que certains gènes ne possèdent aucun intron et que tous les gènes ne possèdent pas forcément des séquences d'ET dans leurs différentes régions, le nombre de gènes analysés varie selon la région génique considérée. Le nombre de gènes analysés pour chacune des régions définies et pour chacune des méthodes est présenté dans le **Tableau 7**.

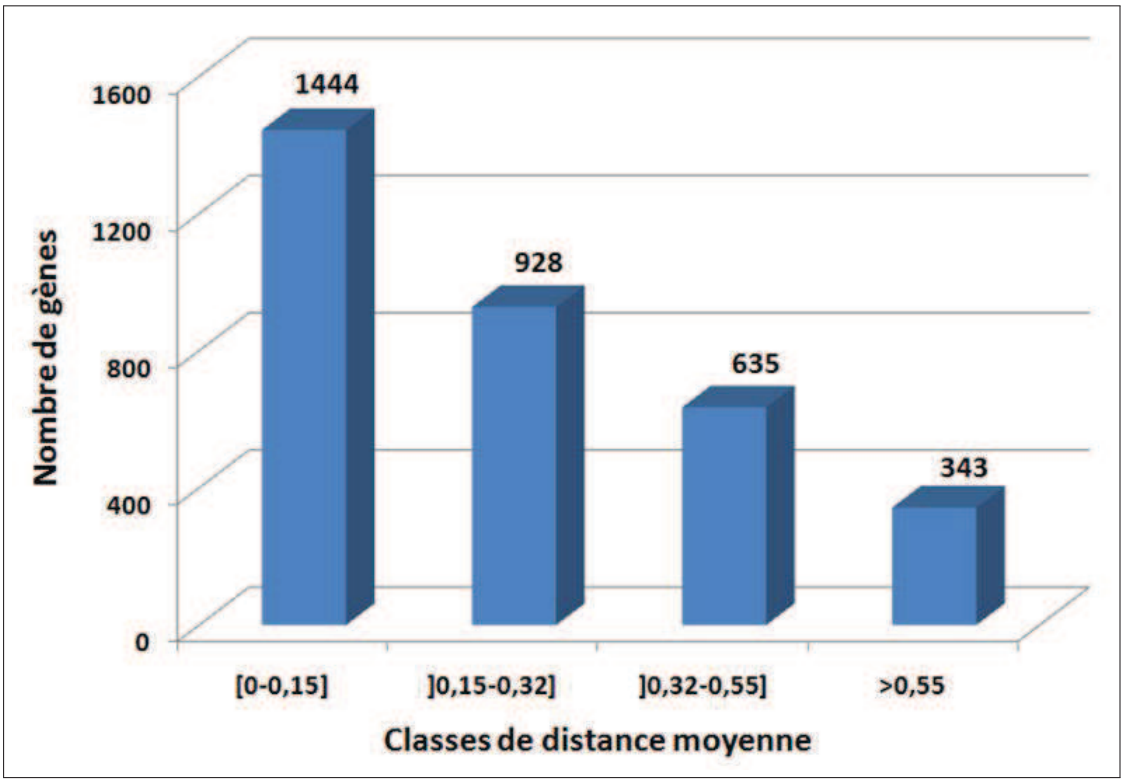

FIGURE 18: *Distribution du nombre de gènes humains selon leur classe de densité moyenne. Le nombre de gènes appartenant à chaque classe est indiqué au dessus de la barre correspondante.*

Tableau 7: *Nombre de gènes analysés par région génique et pour les distances euclidienne et moyenne.*

		Classe 1		Classe 2		Classe 3		Classe 4	
Région		Distance Euclidienne	Distance Moyenne	Distance Euclidienne	Distance Moyenne	Distance Euclidienne	Distance Moyenne	Distance Euclidienne	Distance Moyenne
Gène entier		1876	1444	955	928	331	635	188	343
Promoteur		1876	1444	955	928	331	635	188	343
Exons		1876	1444	955	928	331	635	188	343
Introns		1686	1289	876	850	311	590	181	325
Promoteur	Avec ET	1109	841	580	578	201	377	121	215
	Sans ET	1109	841	580	578	201	377	121	215
Exons	Avec ET	632	466	314	299	86	210	43	100
	Sans ET	632	466	314	299	86	210	43	100
Introns	Avec ET	1570	1196	824	797	291	546	163	309
	Sans ET	1570	1196	824	797	291	546	163	309

Bien que la méthode euclidienne soit souvent utilisée pour l'analyse de distances génétiques, elle présente un biais important dans le calcul de la divergence d'expression. Un exemple numérique de ce biais est représenté dans le **Tableau 8**. On voit que cette méthode permet de classer de façon identique les gènes B et C alors que la variation d'expression entre les états normal et tumoral est beaucoup plus importante, biologiquement parlant, pour le gène B que pour le gène C. En effet, le gène B a doublé son expression dans l'état tumoral par rapport à l'état normal alors que le gène C n'a subi qu'une faible augmentation de son expression à l'état tumoral. C'est également le cas pour les gènes A et B, qui doublent tous les deux leur niveau d'expression à l'état tumoral par rapport à l'état normal, ce qui veut dire que ces deux gènes (A et B) devraient être groupés dans la même classe de divergence d'expression. Ceci n'est pas le cas en utilisant la distance euclidienne puisque le gène A possède une distance très faible (distance=1) par rapport à celle du gène B (distance=100) ce qui indique que ces deux gènes ne seront certainement pas intégrés dans la même classe de divergence d'expression. Ce type de biais n'apparaît pas quand on utilise la méthode de la distance moyenne puisque les gènes A et B seront groupés dans la même classe de divergence d'expression. De plus, les gènes A et B présentent une divergence d'expression plus forte que celle du gène C.

Tableau 8: *Exemple numérique de comparaison d'expression par les distances euclidienne et moyenne pour trois gènes A, B et C.*

Gènes	Expression normale	Expression tumorale	Distance Euclidienne	Distance Moyenne
A	1	2	1	$\frac{1}{3}$
B	100	200	100	$\frac{1}{3}$
C	1100	1200	100	$\frac{1}{33}$

Le **Tableau 9** donne le nombre de gènes communs détectés selon les deux méthodes de calcul de la divergence d'expression. Pour chacune des méthodes, les gènes ont été classés en quatre classe de divergence d'expression. Le test de corrélation est positif et significatif entre les classifications des deux méthodes avec un coefficient de *Pearson* égal à 0,524, ce qui montre que la plupart des gènes conservent la même classe de divergence d'expression quelle que soit la méthode considérée. De plus, quand les gènes ne conservent pas la même classe entre les deux méthodes, la majorité des changements se fait vers la classe la plus proche (supérieure et inférieure).

Tableau 9: *Nombre de gènes communs détectés selon la méthode de classification. Le nombre de gènes par classe est indiqué en colonnes pour la distance moyenne et en ligne pour la distance euclidienne.*

	Distance Moyenne				
Distance Euclidienne	Classe 1	Classe 2	Classe 3	Classe 4	Total
Classe 1	1188	443	186	59	1876
Classe 2	229	338	260	128	955
Classe 3	21	112	119	79	331
Classe 4	6	35	70	77	188
Total	1444	928	635	343	3350

Cependant, ces résultats montrent qu'un grand nombre des gènes analysés ($\sim$48,6%, 1628 sur 3350 gènes) changent de classe entre les deux méthodes. Ainsi, pour les analyses restantes on se basera sur la classification des gènes à partir des résultats de la méthode de la distance moyenne tout en sachant que les résultats obtenus quand la distance euclidienne est utilisée pour classifier les gènes sont donnés dans les Annexes.

4.4 Comparaison des enrichissements en modifications d'histones entre un état normal et un état tumoral

Les enrichissements d'histones pour chaque gène et dans chacune des régions considérées ont été récupérés en croisant les coordonnées des gènes avec celles des enrichissements des modifications d'histones (section 4.2, page 74). La somme d'enrichissement pour chaque gène sur l'ensemble des régions a été normalisée par la taille de la région considérée. Ceci m'a permis d'obtenir pour chaque gène (dans chaque région génique) et pour chaque modification d'histone une seule valeur par état. Les valeurs d'enrichissement normalisées ont été comparées par un test de Wilcoxon sur données appariées. Cette démarche est résumée dans la **Figure 19**.

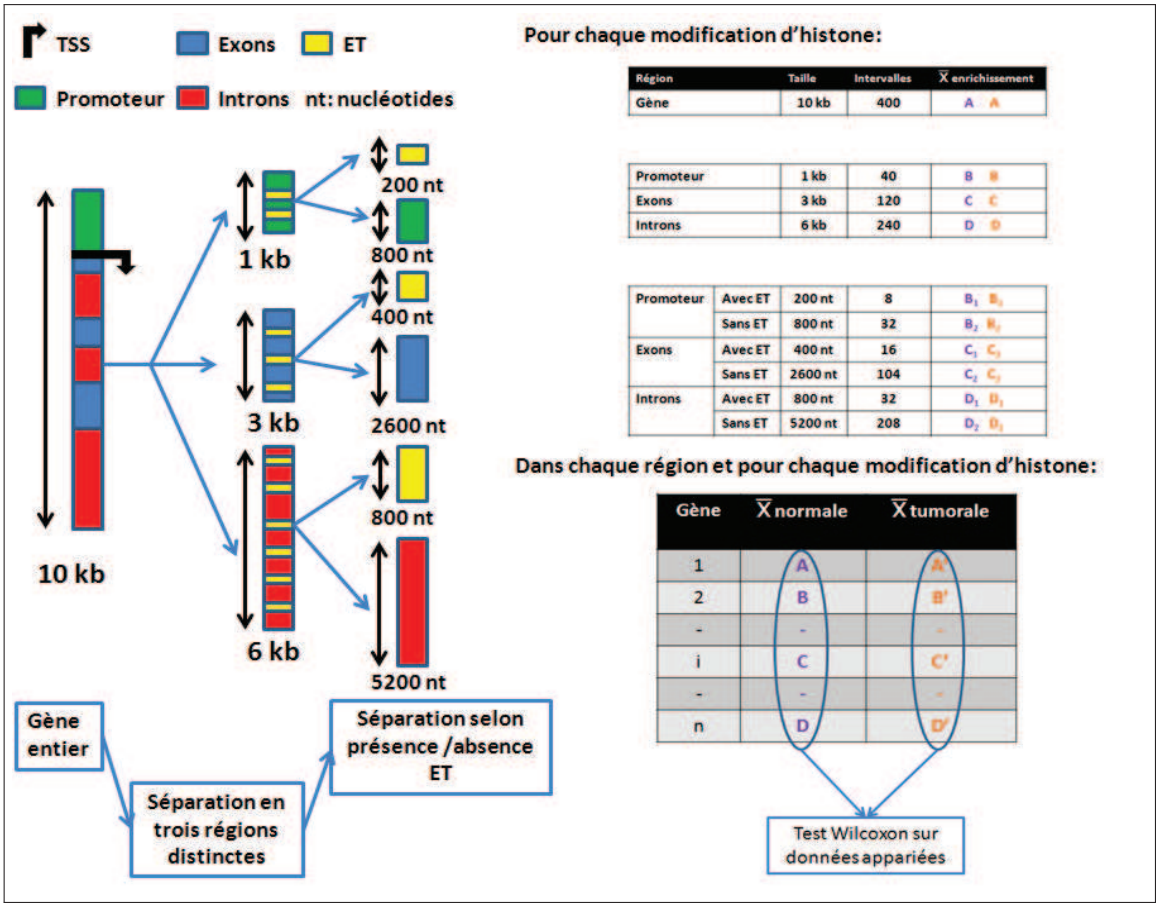

FIGURE 19: *Récapitulatif de la comparaison d'enrichissement. La première étape consiste à séparer chaque gène en différentes régions. Pour chaque région et chaque modification d'histone on calcule la moyenne d'enrichissement à l'état normal (violet) et tumoral (orange). Les moyennes d'enrichissement sont ensuite comparées par un test de wilcoxon sur données appariées pour détecter d'éventuelles variations statistiquement significatives.*

Lorsque tous les gènes (3350) sont considérés dans leur ensemble, les enrichissements sont significativement plus élevés à l'état tumoral pour les modifications d'histones suivantes : H3K4me1, H3K4me2, H3K9ac, H3K27ac, H3K27me3 et H4K20me1 (**Figure 20**). À l'opposé, les gènes dans les cellules normales montrent un enrichissement significativement plus élevé pour les modifications d'histones H3K4me3 et H3K36me3.

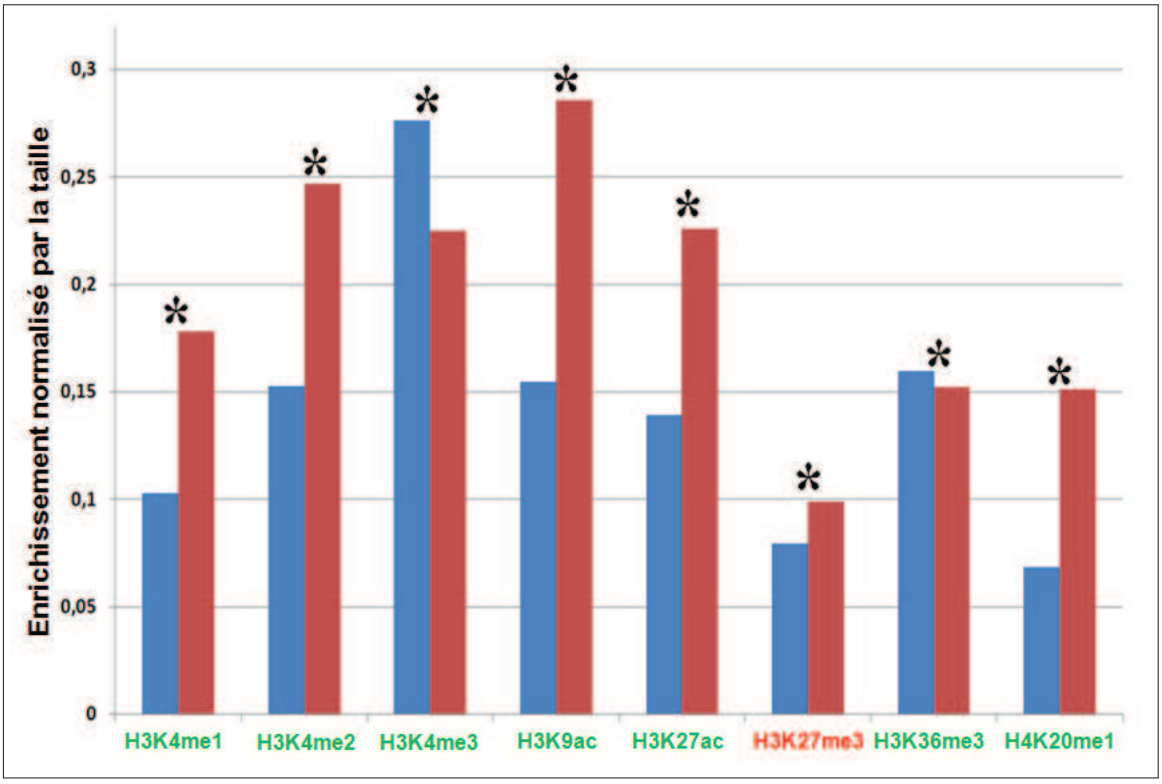

FIGURE 20: *Variations d'enrichissement des modifications d'histones dans les gènes. Barres bleues : état normal; barres rouges : état tumoral. Les modifications d'histones actives sont représentées en vert, les répressives en rouge. * : Différence d'enrichissement statistiquement significative (p-value $< 0,05$). Les P-values sont indiquées dans le Tableau 30 (cf. Annexes, section 30, page 130).*

Chaque gène a été divisé en trois sous-régions distinctes ("promoteur", "exons" et "introns") afin de voir si un comportement spécifique existe pour une région donnée. Comme le montre la **Figure 21**, on observe globalement les mêmes résultats que ceux obtenus sur les "gènes entiers" : les modifications H3K4me3 et H3K36me3 sont significativement plus enrichies à l'état normal, alors que les six modifications restantes montrent un enrichissement significativement plus élevé à l'état tumoral. Néanmois, H3K36me3 montre un enrichissement quasiment identique entre les états normal et tumoral dans les "introns" alors que dans les "exons" et le "promoteur" l'enrichissement

est significativement plus élevé dans les cellules normales. Ceci peut indiquer une augmentation d'enrichissement pour cette modification à l'état tumoral accompagnée d'une diminution à l'état normal. On note également des variations d'enrichissement entre les différentes modifications d'histones à l'intérieur de la même région. Si on s'intéresse, par exemple, à la région "promoteur", on remarque que la modification H3K27me3 montre un enrichissement plus faible par rapport aux autres modifications, c'est également le cas pour H3K36me3 et H4K20me1.

Chaque sous-région génique a été ensuite séparée selon la présence ou l'absence des séquences d'ET afin d'identifier un quelconque effet des ET sur la variation d'enrichissement des modifications d'histones entre les deux états. Les résultats (**Figure 22**) rejoignent ce qui a été observé pour le "gène entier" avec un enrichissement significativement plus fort dans les cellules tumorales pour les modifications d'histones H3K4me1, H3K9ac, H3K27ac, H3K27me3 et H4K20me1, et ceci indépendamment de la présence ou de l'absence d'ET dans les différentes régions. De plus, dans les "introns" et le "promoteur", l'enrichissement quasi identique entre les deux états pour la modification H3K36me3 est conservé en présence ou en absence des ET. Cependant, l'enrichissement de H3K4me3 dans le "promoteur" n'est pas significativement différent en présence des ET alors qu'en absence d'ET les cellules normales montrent un enrichissement significativement plus fort à l'état normal. Ceci peut indiquer que la présence des ET dans les "promoteurs" modifie l'enrichissement de H3K4me3 selon l'état de la cellule. Les ET semblent avoir également un effet sur l'enrichissement de H3K4me2 dans les "exons" puisque la variation d'enrichissement de H3K4me2 n'est pas significative en présence d'ET alors qu'en leur absence les cellules tumorales montrent un enrichissement significativement plus fort.

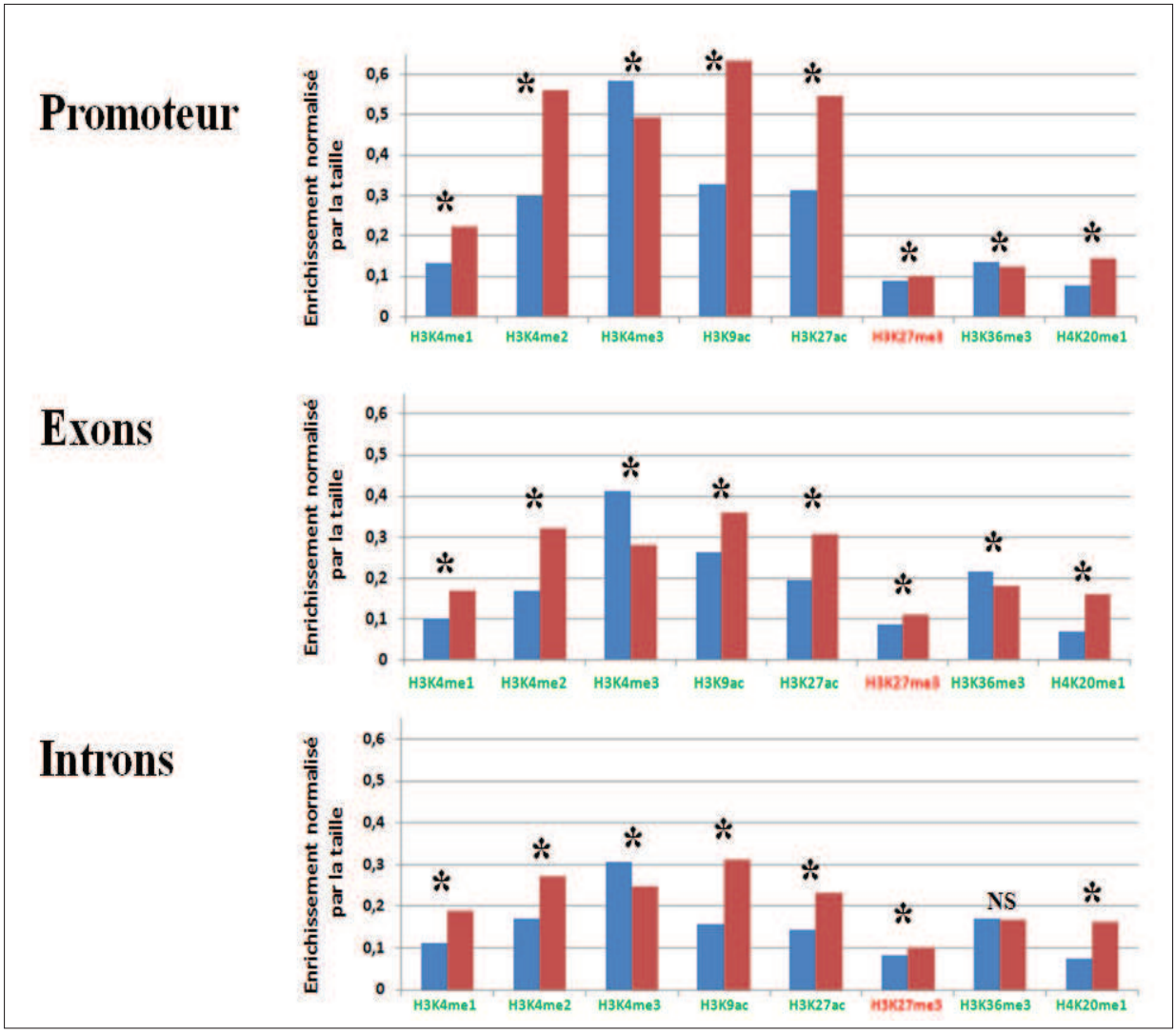

FIGURE 21: *Variations d'enrichissement des modifications d'histones dans les "promoteur", "exons" et "introns". Barres bleues : état normal ; barres rouges : état tumoral. Les modifications d'histones actives sont représentées en vert, les répressives en rouge. * : Différence d'enrichissement statistiquement significative (p-value < 0,05). NS : Différence d'enrichissement non significative. Les P-values sont indiquées dans le Tableau 30 (cf. Annexes, section 30, page 130).*

Il est important de noter que cette dernière analyse considère deux facteurs à la fois qui sont la présence/absence d'ET et l'état normal/tumoral de la cellule. Afin de s'affranchir de ce problème (analyse de deux facteurs à la fois) j'ai décidé de diviser cette analyse en deux analyses distinctes qui considèrent chacun des facteurs à part. Dans le but d'identifier un effet des ET sur les variations d'enrichissement des modifications d'histones, j'ai comparé, pour chacun des états considérés, les variations d'enrichissement au niveau des séquences d'ET inclus dans chacune des trois sous-régions géniques à celles

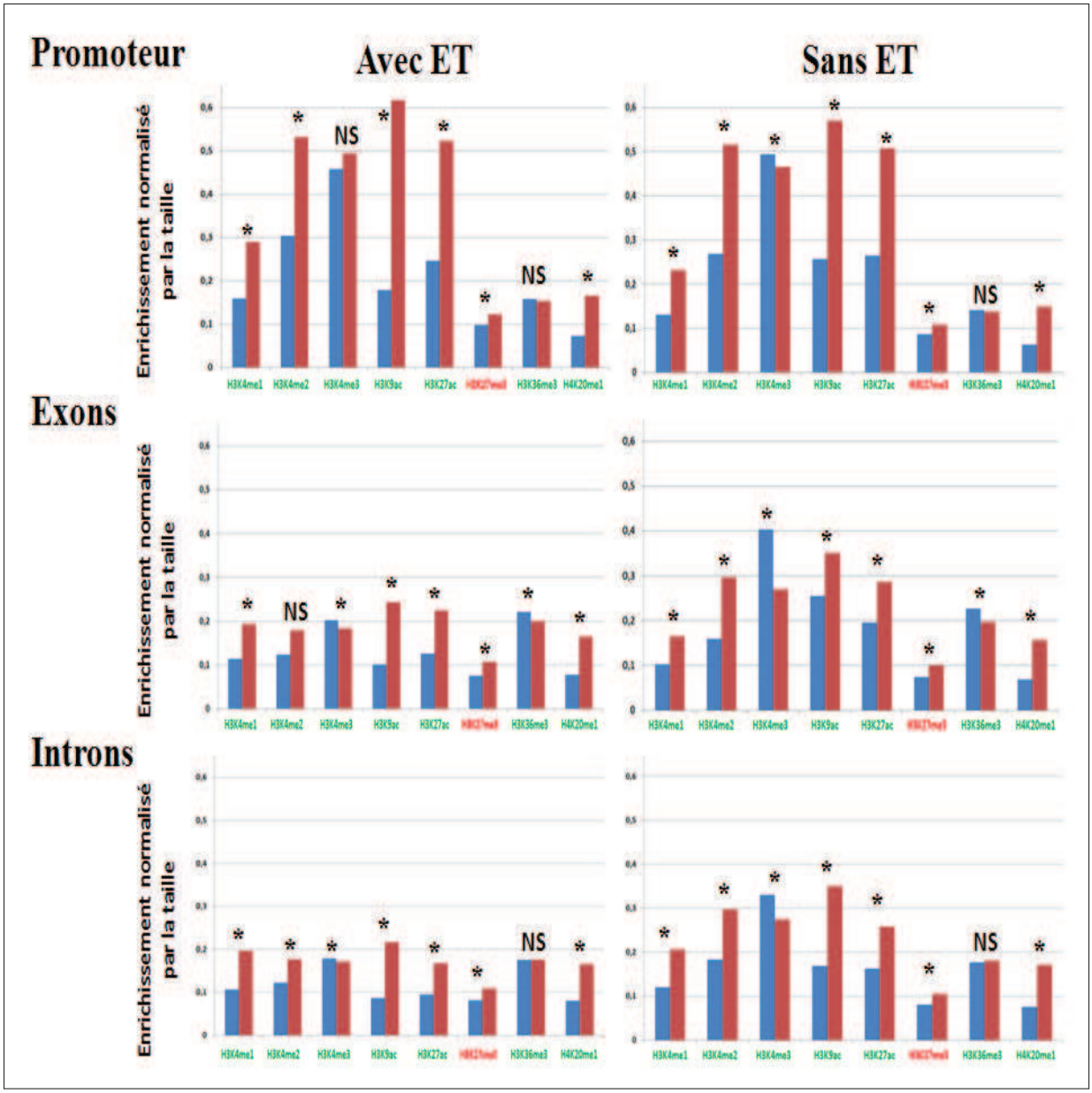

FIGURE 22: *Variations d'enrichissement des modifications d'histones dans les "promoteur", "exons" et "introns" selon la présence/absence d'ET. Barres bleues : état normal ; barres rouges : état tumoral. Les modifications d'histones actives sont représentées en vert, les répressives en rouge. * : Différence d'enrichissement statistiquement significative (p-value $< 0,05$). NS : Différence d'enrichissement non significative. Les P-values sont indiquées dans le Tableau 30 (cf. Annexes, section 30, page 130).*

de la sous-région considérée sans les séquences d'ET. La taille étant variable entre les ET et la sous-région sans les ET, la variation d'enrichissement a été normalisée par la taille.

Quand on considère les régions promotrices à l'état normal (**Figure 23**), les ET sont significativement plus enrichis que la région promotrice dépourvue d'ET pour les modifications d'histones H3K4me1, H3K4me3, H3K27ac, H3K27me3, H3K36me3 et H4K20me1. On observe la tendance inverse pour H3K9ac avec un enrichissement significativement plus important au niveau de la région promotrice dépourvue d'ET alors qu'il n'y a pas de différence d'enrichissement pour H3K4me2 entre les deux régions considérées. À l'état tumoral, les modifications d'histones H3K4me1, H3K27me3, H3K36me3 et H4K20me1 montrent la même différence d'enrichissement que celle observée à l'état normal où les séquences d'ET sont plus enrichies que les promoteurs dépourvus d'ET alors qu'il n'y a pas de différence d'enrichissement pour H3K4me2 entre les deux régions considérées. Cependant les modifications d'histones H3K4me3, H3K9ac et H3K27ac ne montrent plus de différence d'enrichissement significative entre les deux régions considérées. Il faut noter que pour la modification H3K9ac, l'enrichissement au niveau des séquences d'ET augmente de façon considérable à l'état tumoral pour atteindre un niveau comparable à celui des promoteurs sans les séquences d'ET. Ces résultats indiquent que les ET s'associent avec les modifications H3K4me1, H3K27me3, H3K36me3 et H4K20me1 quel que soit l'état de la cellule alors que les modifications H3K4me3 et H3K27ac sont associées aux ET à l'état normal uniquement.

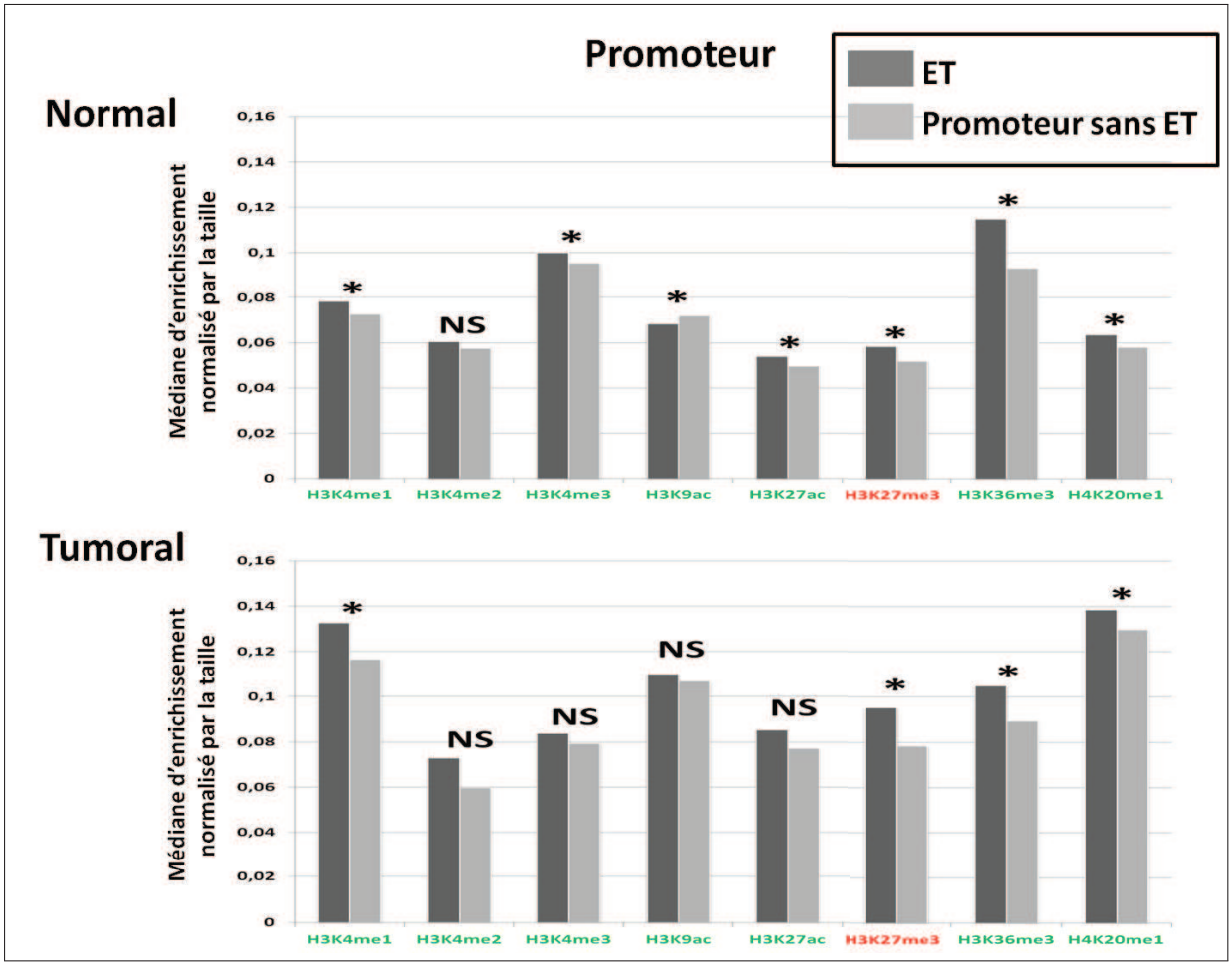

FIGURE 23: *Variations d'enrichissement des modifications d'histones dans le "promoteur" selon la présence/absence d'ET et l'état de la lignée cellulaire. Les modifications d'histones actives sont représentées en vert, les répressives en rouge. L'état de la lignée cellulaire est indiqué à côté de chaque graphe. * : Différence d'enrichissement statistiquement significative (p-value $< 0,05$). NS : Différence d'enrichissement non significative. Les P-values sont indiquées dans le Tableau 30 (cf. Annexes, section 30, page 130).*

Les régions "exons" (**Figure 24**) et "introns" (**Figure 25**) montrent des résultats semblables entre elles à l'état normal. En effet, les régions exoniques et introniques dépourvues d'ET montrent un enrichissement significativement plus important que les ET insérés dans ces régions pour les modifications d'histones H3K4me1, H3K4me2, H3K4me3, H3K9ac et H3K27ac. La tendance inverse est observée pour les modifications H4K20me1 ("exons" et "introns") et H3K27me3 ("introns" uniquement) qui montrent un enrichissement plus important au niveau des ET. Enfin il n'y a pas de différence d'enrichissement pour H3K36me3 selon la présence ou l'absence d'ET dans les régions

exonique et intronique. À l'état tumoral, on observe les mêmes tendances à l'exception des modifications H3K4me1 ("exons" et "introns") et H4K20me1 ("exons" seulement) pour lesquelles il n'y a plus de différence d'enrichissement significative entre les ET et les séquences dépourvues d'ET. Ainsi, les régions "exons" et "introns" montrent peu d'associations entre les ET et les modifications d'histones par rapport à celles observées au niveau de la région promotrice.

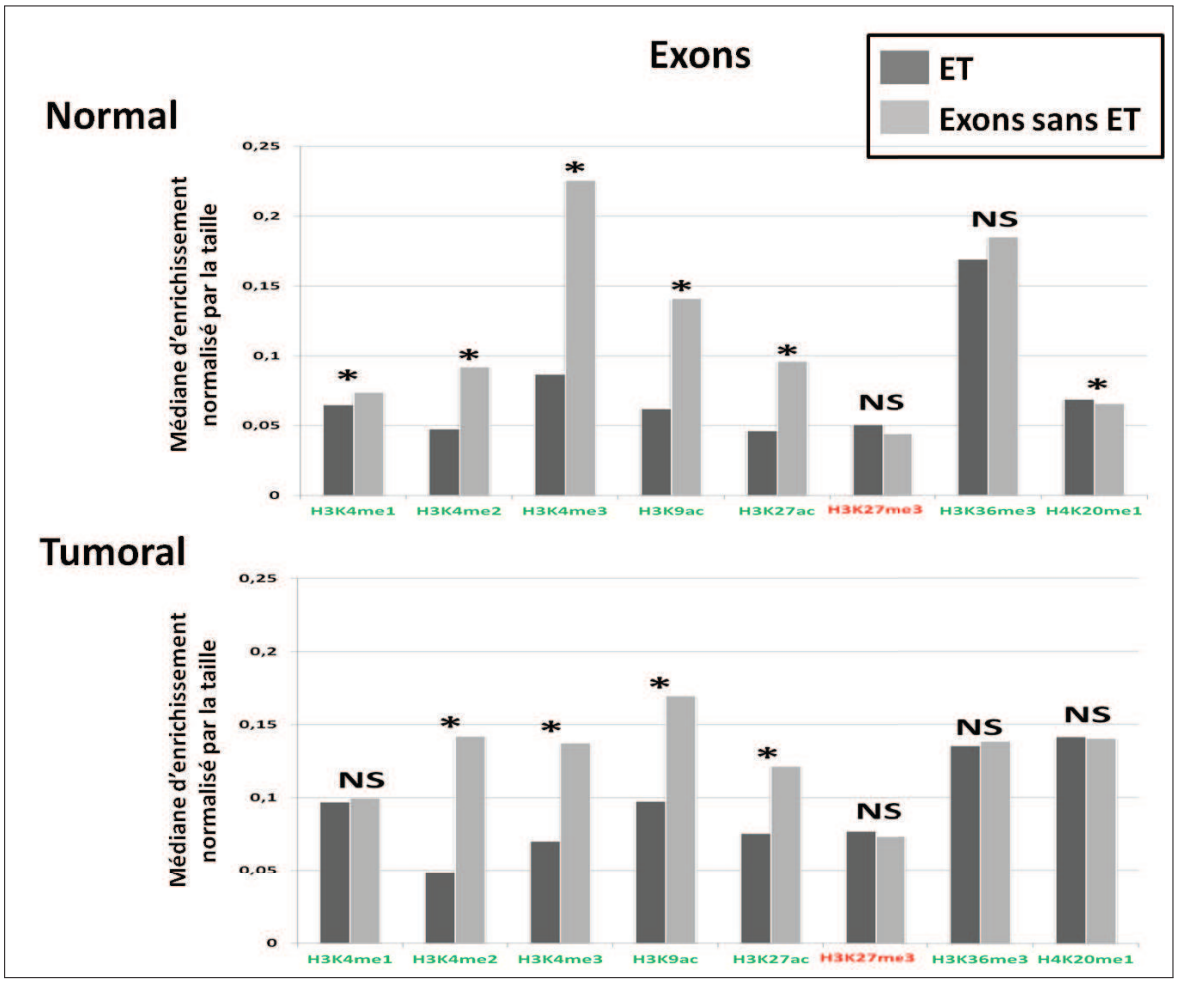

FIGURE 24: *Variations d'enrichissement des modifications d'histones dans les "exons" selon la présence/absence d'ET et l'état de la lignée cellulaire. Les modifications d'histones actives sont représentées en vert, les répressives en rouge. L'état de la lignée cellulaire est indiqué à côté de chaque graphe. * : Différence d'enrichissement statistiquement significative (p-value < 0,05). NS : Différence d'enrichissement non significative. Les P-values sont indiquées dans le Tableau 30 (cf. Annexes, section 30, page 130).*

Afin de valider les observations qui montrent des associations entre les ET et certaines modifications d'histones au niveau des promoteurs, j'ai comparé les variations d'enrichissement entre les promoteurs qui possèdent des ET dans leur séquence et

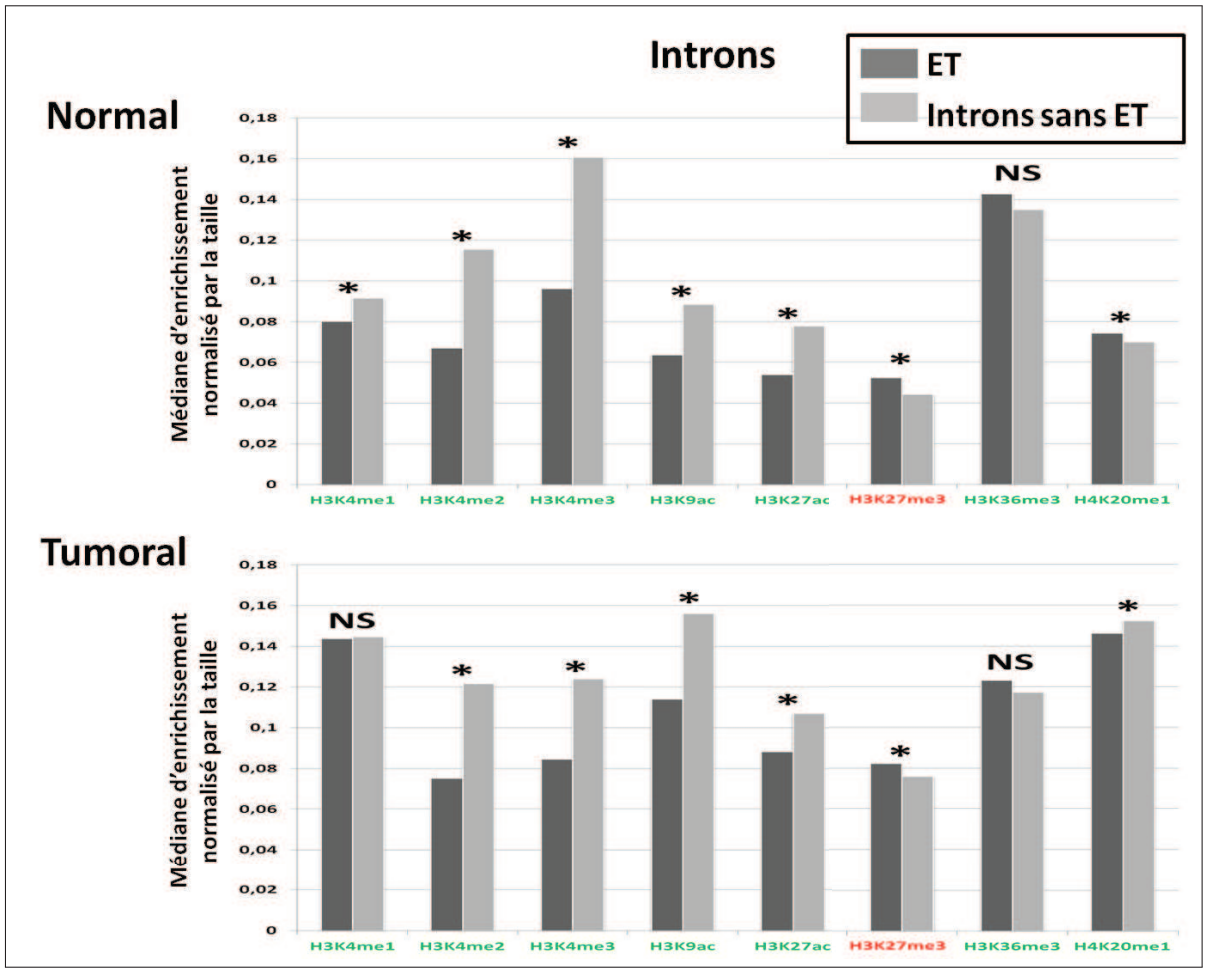

FIGURE 25: *Variations d'enrichissement des modifications d'histones dans les "introns" selon la présence/absence d'ET et l'état de la lignée cellulaire. Les modifications d'histones actives sont représentées en vert, les répressives en rouge. L'état de la lignée cellulaire est indiqué à côté de chaque graphe. * : Différence d'enrichissement statistiquement significative (p-value $<$ 0,05). NS : Différence d'enrichissement non significative. Les P-values sont indiquées dans le Tableau 30 (cf. Annexes, section 30, page 130).*

ceux qui sont complètement dépourvus d'ET. Dans ce cas on considère à la fois l'enrichissement des séquences d'ET et du promoteur pour les promoteurs qui possèdent des ET dans leur séquence. À l'état normal (**Figure 26**), on observe une association entre les promoteurs avec ET et les modifications d'histones H3K36me3 et H4K20me1 significativement plus importante qu'au niveau des promoteurs sans ET. La tendance inverse est observée pour les modifications H3K4me1, H3K4me2, H3K4me3, H3K9ac et H3K27ac avec un enrichissement plus important au niveau des promoteurs qui n'ont aucun ET. Seule la modification H3K27me3 ne montre pas une différence d'enrichissement entre les deux types de régions considérées. À l'état tumoral, on observe une association entre les promoteurs avec ET et les modifications d'histones H3K4me1, H3K27me3, et

H3K36me3 significativement plus importante qu'au niveau des promoteurs sans ET. Les modifications H3K4me2, H3K9ac, H3K27ac et H4K20me1 ne montrent plus une différence d'enrichissement significative entre les deux types de promoteurs avec une augmentation de l'enrichissement pour les trois premières modifications au niveau des promoteurs avec ET. Seule la modification H3K4me3 montre un enrichissement plus important au niveau des promoteurs sans ET. Ces résultats confirment les observations qui montrent une association entre les modifications d'histones et les ET qui dépendent à la fois de la région d'insertion considérée et de l'état de la cellule.

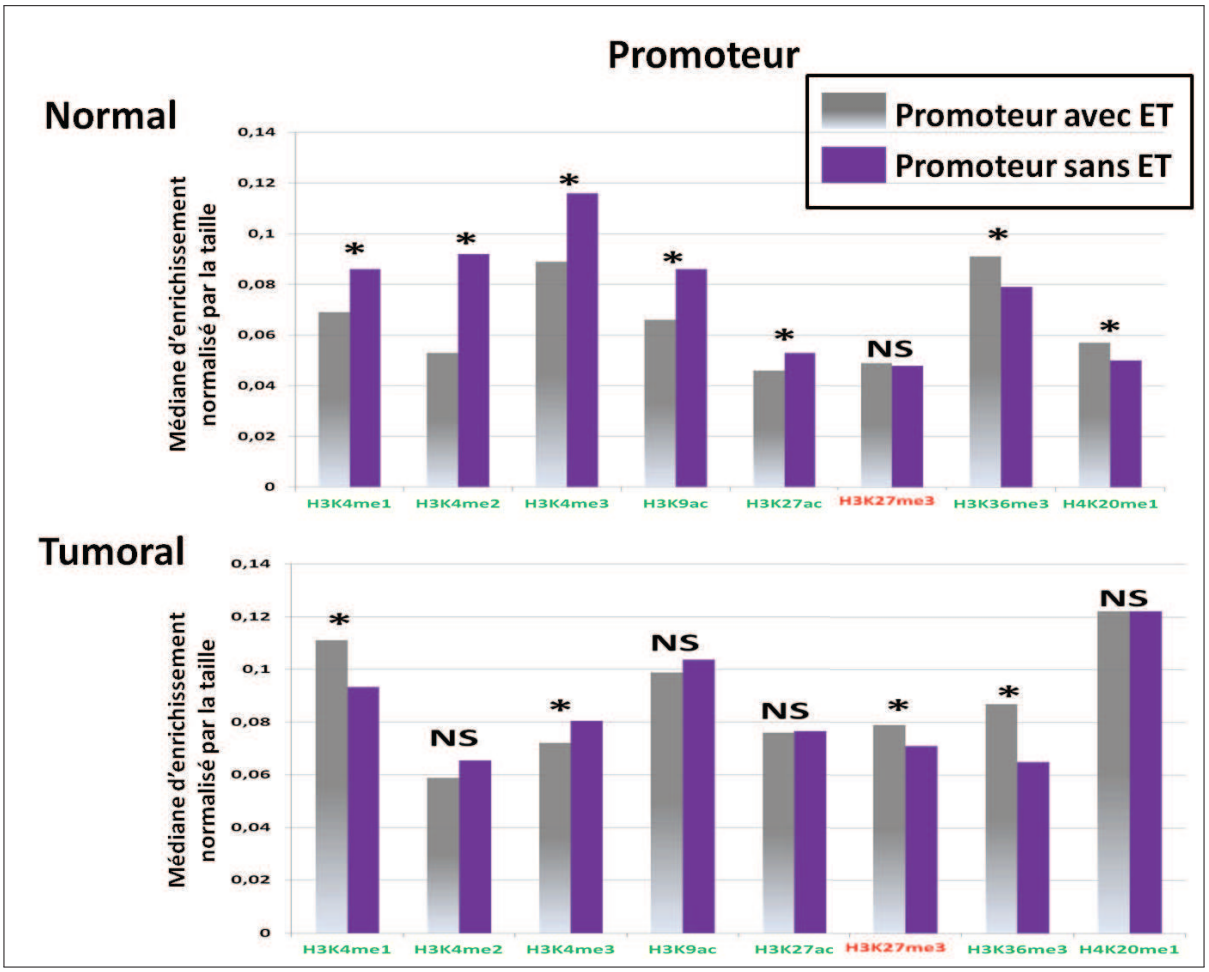

FIGURE 26: *Variations d'enrichissement des modifications d'histones dans le "promoteur" selon la présence/absence d'ET et l'état de la lignée cellulaire. Les modifications d'histones actives sont représentées en vert, les répressives en rouge. L'état de la lignée cellulaire est indiqué à côté de chaque graphe. * : Différence d'enrichissement statistiquement significative (p-value < 0,05). NS : Différence d'enrichissement non significative. Les P-values sont indiquées dans le Tableau 30 (cf. Annexes, section 30, page 130).*

4.5 Les modifications d'histones et la variation de l'expression génique entre les états normal et tumoral

Les modifications d'histones ont un effet important sur l'expression des gènes qui dépend à la fois de la modification d'histone et du résidu lysine sur lequel la modification a lieu. Afin de voir si les variations de l'expression des gènes entre les états normal et tumoral pourraient être liées aux variations d'enrichissement des modifications d'histones, j'ai divisé les gènes analysés précédemment en quatre classes de divergence d'expression (allant de 1 à 4, avec 1 qui correspond à la classe qui possède la divergence d'expression la plus faible alors que la classe 4 possède la plus forte divergence d'expression) et j'ai calculé ensuite les enrichissements des gènes de ces quatre classes. Enfin, un test de wilcoxon sur données appariées a été utilisé pour comparer les valeurs d'enrichissement normalisées.

La **Figure 27** montre la même variation d'enrichissement des modifications d'histones quelle que soit la classe de divergence d'expression lorsque les "gènes entiers" sont considérés. Une différence existe cependant pour la modification H3K36me3, qui est significativement plus enrichie dans les gènes des classes 1 et 2 à l'état normal par rapport à l'état tumoral, alors que les gènes des classes 3 et 4 montrent un enrichissement identique entre les deux états. On peut noter également que certaines modifications d'histones, comme H3K27ac, H3K4me3 et H3K9ac, montrent une augmentation de la variation d'enrichissement entre les deux états avec l'augmentation de la divergence d'expression alors que pour d'autres modifications, comme H3K27me3 et H4K20me1, la variation d'enrichissement entre les deux états est quasiment identique pour les gènes quelle que soit leur classe de divergence d'expression.

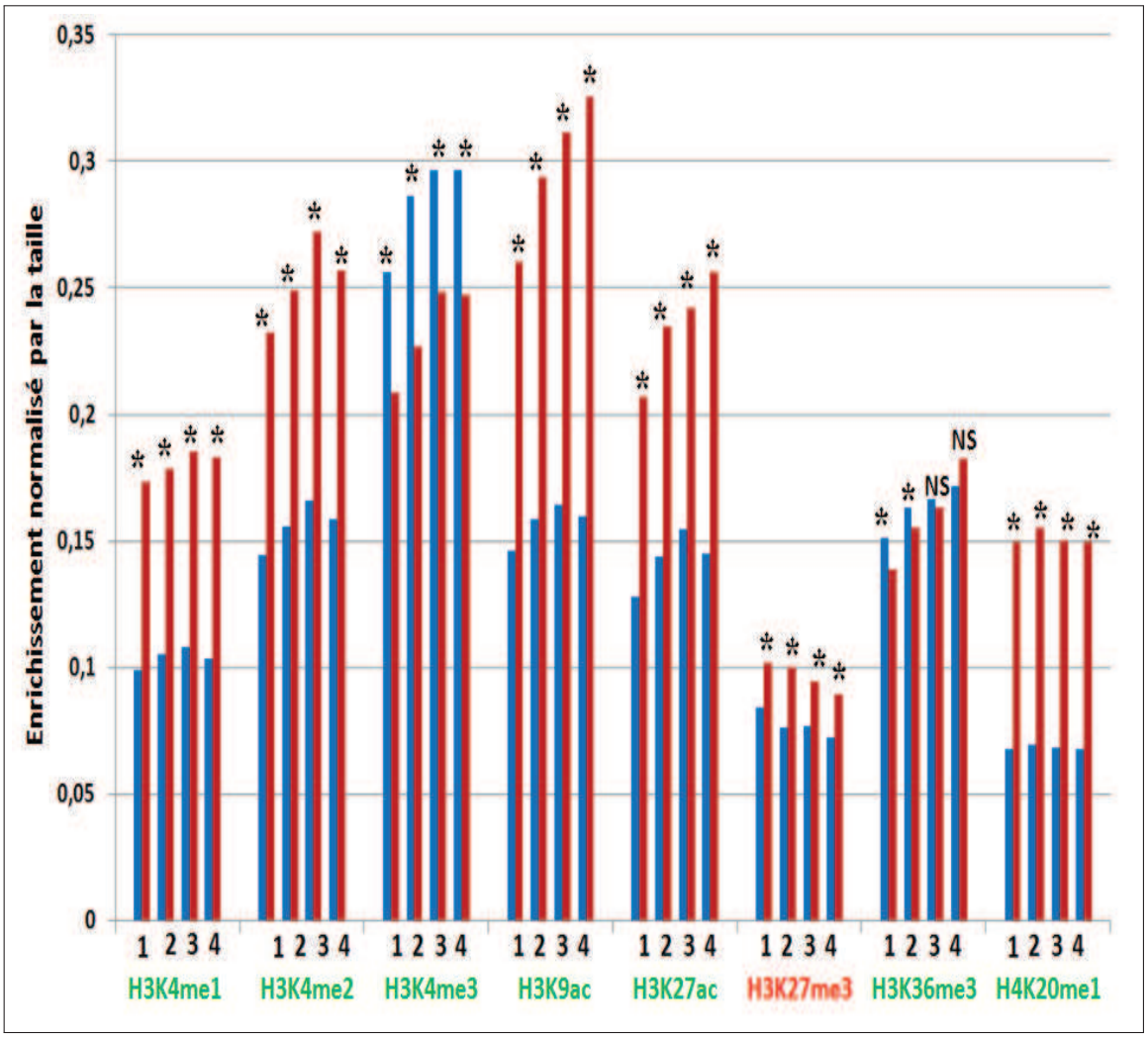

FIGURE 27: *Variations d'enrichissement des modifications d'histones dans les gènes pour quatre classes de divergence d'expression. Barres bleues : état normal ; barres rouges : état tumoral. Les modifications d'histones actives sont représentées en vert, les répressives en rouge. * : Différence d'enrichissement statistiquement significative (p-value < 0,05). NS : Différence d'enrichissement non significative. 1, 2, 3 et 4 : classes de divergence d'expression.*

J'ai ensuite divisé chaque gène en trois sous-régions ("promoteur", "exons" et "introns") pour déterminer si, pour chacune des régions, une corrélation existe entre les variations d'enrichissement des modifications d'histones et la divergence d'expression (**Figure 28**). Pour le "promoteur", les modifications d'histones H3K4me3 (classe 4) et H3K36me3 (classes 3 et 4) ne montrent pas de différence d'enrichissement significative tandis que les gènes des classes restantes montrent un enrichissement significativement plus élevé à l'état normal qu'à l'état tumoral. Pour les "exons", seule la modification H3K36me3 montre une différence entre les gènes des quatre classes puisque les gènes de

la classe 4 ne montrent pas un enrichissement significativement plus fort à l'état normal alors que les gènes des classes 1, 2 et 3 montrent une différence significative. Pour les "introns", aucune différence n'existe entre les gènes quelle que soit leur classe de divergence d'expression.

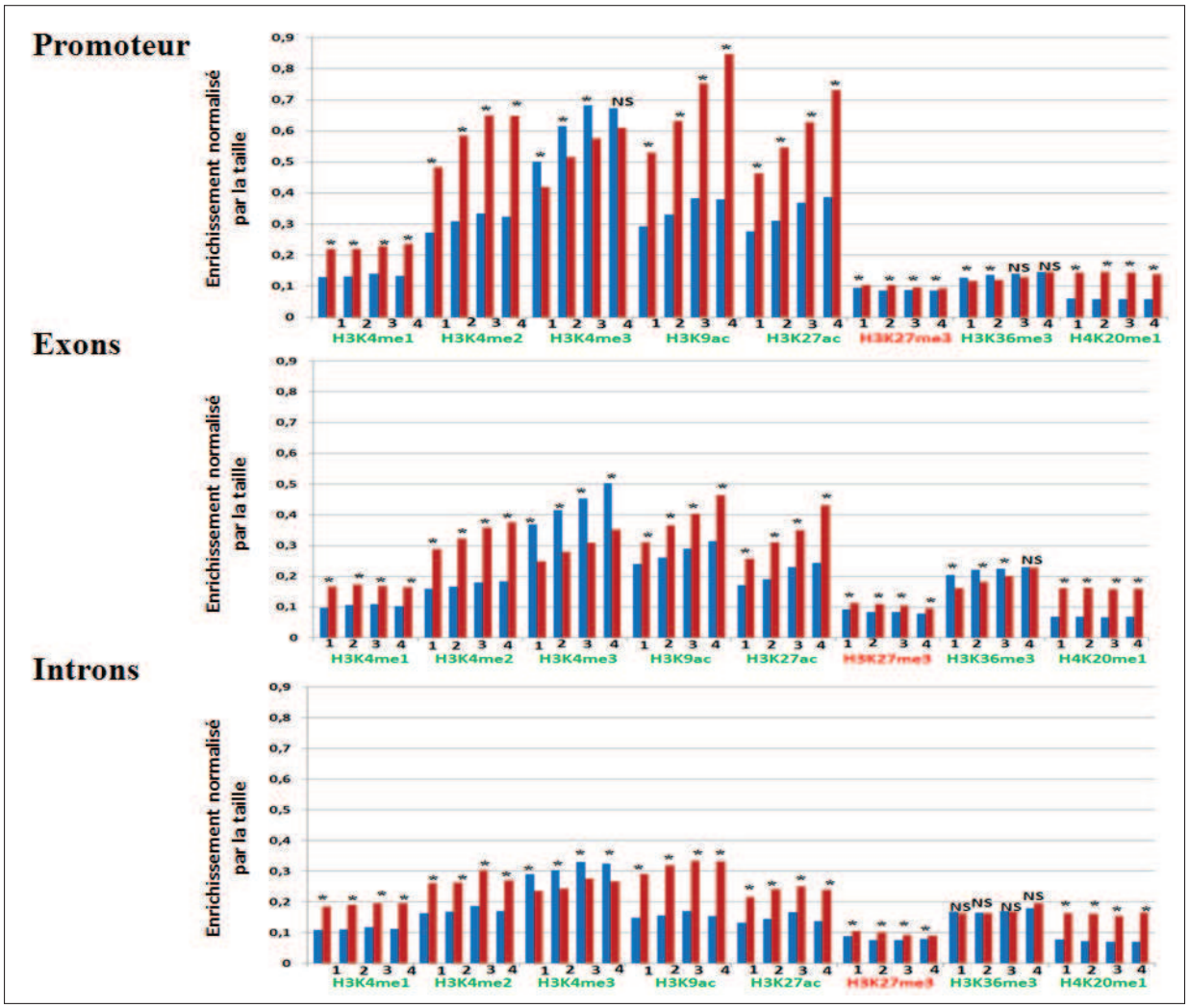

FIGURE 28: *Variations d'enrichissement des modifications d'histones dans les "promoteur", "exons" et "introns" pour les différentes classes de divergence d'expression. Barres bleues : état normal ; barres rouges : état tumoral. Les modifications d'histones actives sont représentées en vert, les répressives en rouge. * : Différence d'enrichissement statistiquement significative (p-value < 0,05). NS : Différence d'enrichissement non significative. 1, 2, 3 et 4 : classes de divergence d'expression.*

La séparation de chacune des trois sous-régions en deux régions distinctes en fonction de la présence/absence des séquences d'ET m'a permis de noter quelques différences

entre les gènes suivant leur classe de divergence d'expression (**Figure 29**). Ainsi, dans le "promoteur", la présence des ET est toujours associée avec un enrichissement de H3K4me3 plus fort à l'état tumoral pour tous les gènes quelle que soit leur classe même si la différence n'est pas significative. Cette même modification montre également une différence d'enrichissement non significative entre les deux états pour les gènes des classes 2, 3 et 4 des "exons" en présence des ET. Enfin, dans les "introns", l'enrichissement de H3K36me3 est identique entre les deux états en absence et en présence des ET et ceci quelle que soit la classe de divergence d'expression.

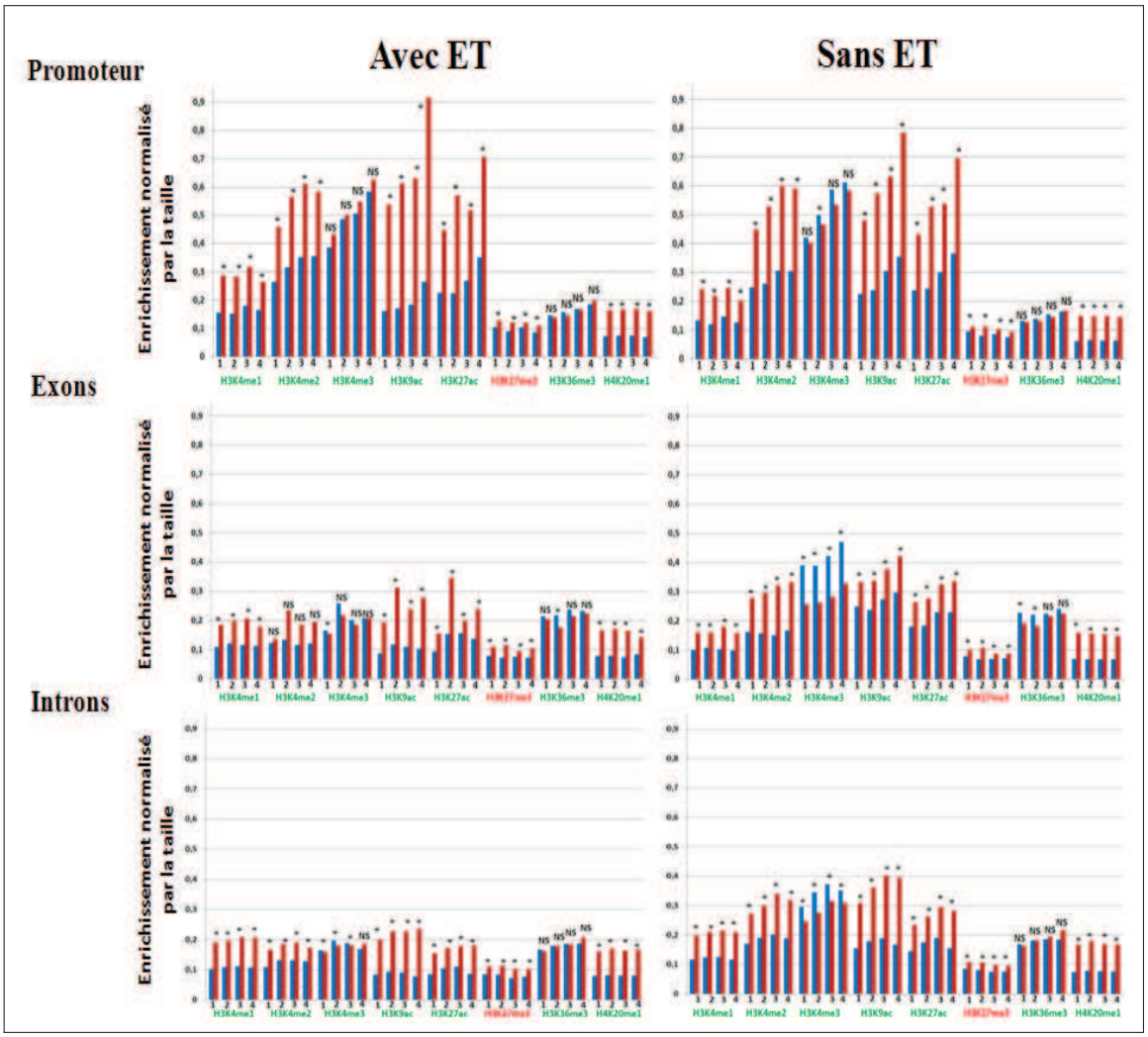

FIGURE 29: *Variations d'enrichissement des modifications d'histones dans les "promoteur", "exons" et "introns" selon la présence/absence d'ET pour les différentes classes de divergence d'expression. Barres bleues : état normal; barres rouges : état tumoral. Les modifications d'histones actives sont représentées en vert, les répressives en rouge. * : Différence d'enrichissement statistiquement significative (p-value < 0,05). NS : Différence d'enrichissement non significative. 1, 2, 3 et 4 : classes de divergence d'expression.*

Les variations d'enrichissement des modifications d'histones, entre les états normal et tumoral, augmentent avec la divergence d'expression, ce qui suppose une influence du taux d'enrichissement sur cette divergence d'expression. Afin de pouvoir estimer la part de la divergence d'expression due à la variation d'enrichissement des modifications d'histones, j'ai choisi d'utiliser une analyse en composantes prinicpales ou ACP. L'ACP a été faite sur les variations d'enrichissement des modifications d'histones pour les différentes classes géniques de divergence d'expression. La variation d'enrichissement pour une modification donnée a été calculée en divisant la différence d'enrichissement de cette modification entre les deux états par la taille de la région considérée :

$$\text{Variation enrichissement} = \frac{\text{Enrichissement tumoral - Enrichissement normal}}{\text{Taille région génique}}$$

J'ai réalisé une ACP "inter groupes" dont l'objectif est de révéler d'éventuelles différences entre les groupes considérés (classes de divergence d'expression) selon un facteur donné (variation d'enrichissement d'histones). Le ratio de l'ACP "inter groupes" nous permet d'avoir une indication sur la part de variabilité expliquée par le facteur. Quand on considère la région "gène entier" (**Figure 30**), le cercle de corrélations indique que les modifications H3K27ac, H3K9ac et H3K4me2 sont celles qui expliquent la plus grande part de variation sur l'axe 1, alors que H3K27me3 et H4K20me1 expliquent la variation de l'axe 2. De plus, cette analyse montre pas ou très peu de variation entre les gènes selon leur classe de divergence d'expression . En effet, les centroïdes des quatre classes sont quasiment identiques avec une divergence intra-classe un peu plus importante dans la classe 4 (pour laquelle on a la plus forte divergence d'expression).

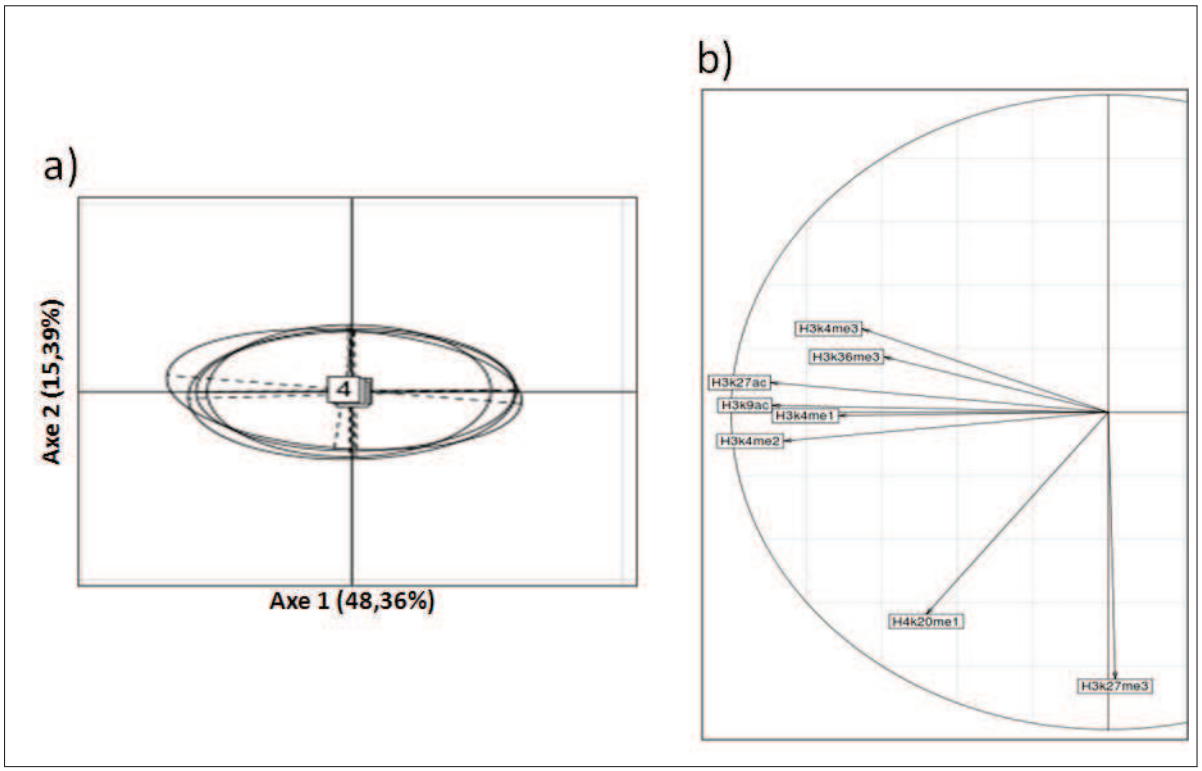

FIGURE 30: *ACP sur la variation d'enrichissement dans la région "gène entier". (a) La présentation graphique de la variation d'enrichissement des modifications d'histones entre les quatre classes de divergence d'expression. (b) Le cercle des corrélations indiquant les huit modifications d'histones.*

Les mêmes résultats sont observés quand je sépare le gène en trois sous-régions ("promoteur", "exons" et "introns") (**Figure 31**), il est important de noter également que la région "introns" montre une divergence inter-classes faible. Enfin, la séparation des régions suivant la présence/absence des ET montre également les mêmes résultats (Annexes, **Figure 36** page 127, **Figure 37** page 128, **Figure 38** page 129). Ces résultats supposent que la divergence d'expression ne peut pas être expliquée par les variations d'enrichissement des modifications d'histones quand toutes les modifications sont considérées ensemble. Néanmoins, la divergence d'expression pourrait être expliquée par une modification d'histone spécifique.

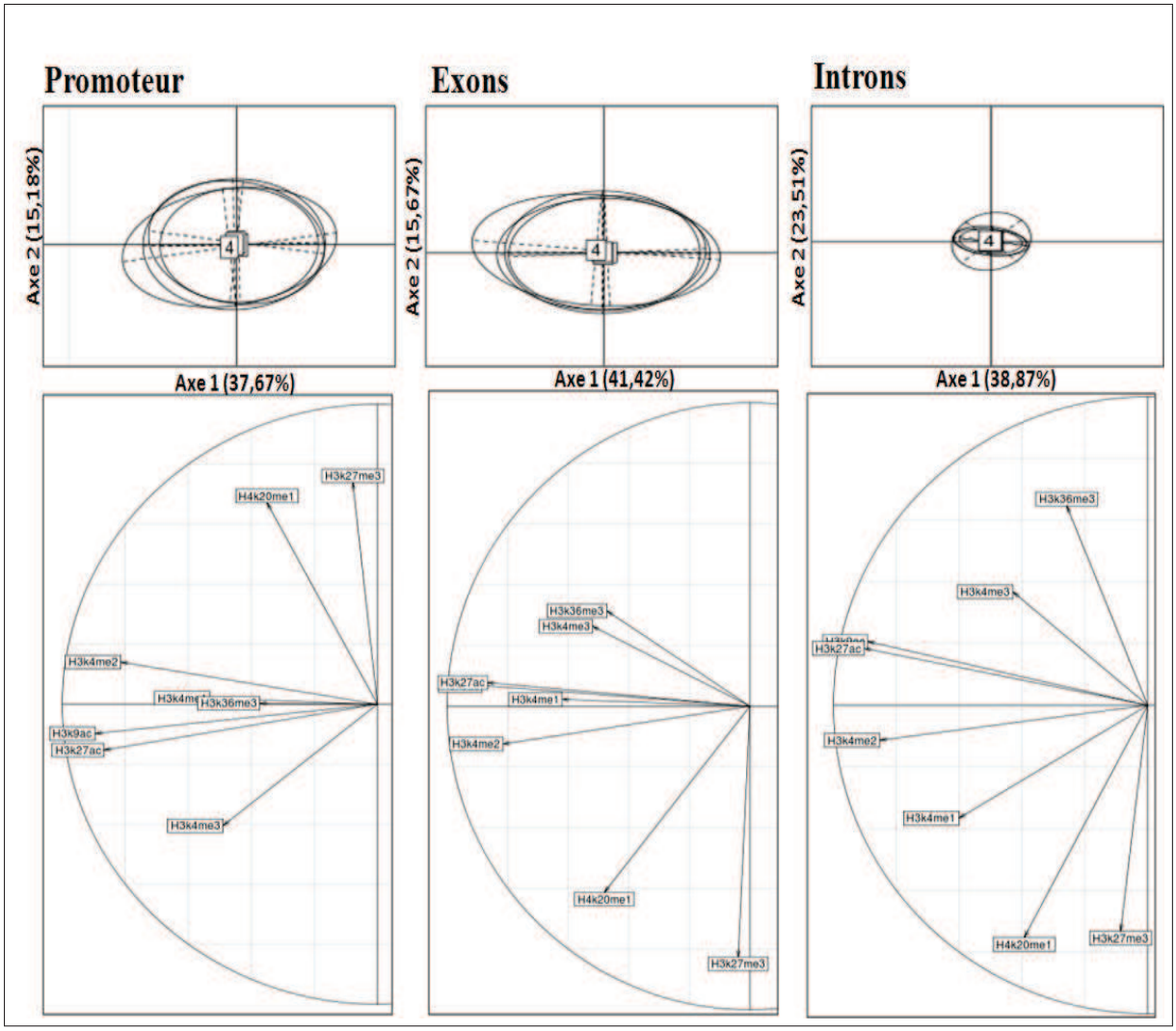

Figure 31: *ACP sur la variation d'enrichissement dans les "promoteur", "exons" et "introns". La présentation graphique de la variation d'enrichissement des modifications d'histones entre les quatre classes de divergence d'expression. Le cercle de corrélation indiquant les huit modifications d'histones est représenté en bas. La région considérée est indiquée au dessus de chaque graphe.*

Pour tester ceci, j'ai réalisé des tests de corrélation dans chaque région et pour chaque modification d'histone, entre la variation d'enrichissement et la divergence d'expression génique. Si la variation d'enrichissement d'une modification d'histone spécifique explique effectivement la divergence d'expression on s'attend à avoir une valeur de corrélation (coefficient de *Pearson*) significativement élevée. Les résultats obtenus (**Tableaux 10**, **11**, **12**) montrent que, pour la plupart des régions géniques et des modifications d'histones, les corrélations entre la variation d'enrichissement des modifications d'histones et la divergence d'expression génique ne sont pas significatives.

Tableau 10: *Les coefficients de Pearson entre les variations d'enrichissement des modifications d'histones et la divergence d'expression génique dans le "gène entier". * : corrélation significative (p-value < 0,05).*

Région génique	H3K4me1	H3K4me2	H3K4me3	H3K9ac	H3K27ac	H3K27me3	H3K36me3	H4K20me1
Gène entier	0,06*	0,10*	0,06*	0,18*	0,11*	0,02	0,19*	0,02

Tableau 11: *Les coefficients de Pearson entre les variations d'enrichissement des modifications d'histones et la divergence d'expression génique dans les "promoteur", "exons" et "introns". * : corrélation significative (p-value < 0,05).*

Région génique	H3K4me1	H3K4me2	H3K4me3	H3K9ac	H3K27ac	H3K27me3	H3K36me3	H4K20me1
Promoteur	0,02	0,03	0,01	0,05*	0,04*	0,01	0,03	0,01
Exons	<0,01	0,02	0,03	0,01	0,02	0,01	0,01	<0,01
Introns	0,01	0,02	0,03	0,03	0,01	0,03	0,03	<0,01

Tableau 12: *Les coefficients de Pearson entre les variations d'enrichissement des modifications d'histones et la divergence d'expression génique dans les "promoteur", "exons" et "introns" selon la présence/absence d'ET. * : corrélation significative (p-value < 0,05).*

Région génique	H3K4me1	H3K4me2	H3K4me3	H3K9ac	H3K27ac	H3K27me3	H3K36me3	H4K20me1
Avec ET promoteur	0,05*	0,07*	0,02	0,07*	0,07*	0,02	0,05*	0,02
Sans ET promoteur	0,04	0,05*	0,03	0,07*	0,04	0,02	0,03	0,02
Avec ET exons	0,01	0,05	0,03	<0,01	0,01	0,06	0,07*	0,01
Sans ET exons	0,01	0,02	<0,01	0,03	0,04	0,01	0,03	0,03
Avec ET introns	0,02	0,01	0,01	0,04*	<0,01	0,04*	0,03	0,02
Sans ET introns	0,02	0,05*	0,02	0,03	0,02	0,04	0,04*	0,02

Il existe néanmoins quelques corrélations significatives mais le coefficient de *Pearson* est très faible pour avoir une significativité biologique. En effet, la plus forte corrélation est observée pour la modification H3K36me3 dans la région "gène entier" avec un coefficient de *Pearson* de 0,194 alors que la moyenne de toutes les corrélations significatives est égale à 0,074. Ces résultats sont confirmés par l'analyse "inter groupes" (**Tableaux 13**, **14**, **15**) dont le ratio varie entre 0,07% et 0,42% pour les différentes régions géniques considérées avec une valeur moyenne égale à 0,21%. Ceci veut dire que les variations d'enrichissement des modifications d'histones expliquent seulement 0,21% de la divergence d'expression des gènes.

Tableau 13: *Les ratios de l'analyse "inter groupes" dans la région "gène entier".*

Région génique	Ratio "inter groupes"
Gène entier	0,13

Tableau 14: *Les ratios de l'analyse "inter groupes" dans les régions "promoteur", "exons" et "introns".*

Région génique	Ratio "inter groupes"
Promoteur	0,17
Exons	0,25
Introns	0,07

Tableau 15: *Les ratios de l'analyse "inter groupes" dans les régions "promoteur", "exons" et "introns" selon la présence/absence d'ET.*

Régions	Région génique	Ratio "inter groupes"
Promoteurs	Avec ET	0,14
	Sans ET	0,19
Exons	Avec ET	0,42
	Sans ET	0,32
Introns	Avec ET	0,13
	Sans ET	0,23

4.6 Conclusion

Dans cette analyse, j'ai comparé les enrichissements de huit modifications d'histones entre les états normal et tumoral, et déterminé leur impact sur la divergence d'expression des gènes. Les résultats obtenus montrent que l'enrichissement en H3K4me1, H3K4me2, H3K9ac, H3K27ac, H3K27me3 et H4K20me1 augmente à l'état tumoral par rapport à l'état normal, alors que H3K4me3 et H3K36me3 montrent la tendance inverse. Ces résultats bien que globalement identiques montrent certaines différences selon la région génique considérée ("promoteur", "exons" et "introns"). Je peux citer à titre d'exemple la modification H3K36me3 dont la variation d'enrichissement est significative dans les "exons" mais ne l'est pas dans les "introns" et le "promoteur". De plus, les ET semblent être associés spécifiquement avec certaines modifications d'histones étudiées comme par exemple H3K4me1, H3K4me3 et H3K36me3 dans le "promoteur" et semblent jouer un rôle important dans la variation d'enrichissement de ces marques à l'état tumoral. Même si on remarque, pour certaines modifications, une augmentation de la variation d'enrichissement entre les états normal et tumoral, cette augmentation n'explique pas la divergence d'expression entre les deux états puisque seul $\sim$0,21% de la divergence d'expression est expliquée par les variations d'enrichissement entre les deux états. Enfin, les corrélations entre les variations d'enrichissement de chaque modification d'histone et la divergence d'expression sont pour la plupart statistiquement non significatives ou très faibles pour avoir une significativité biologique.

Chapitre 5

Discussion et perspectives

5.1 Distribution des Éléments Transposables

Dans le premier chapitre, j'ai analysé le contenu des gènes humains, du chimpanzé, de l'orang-outan et du macaque en terme de copies complètes d'ET présentes dans les séquences géniques et leur voisinage (2 et 10 kb de régions flanquantes). J'ai montré que les gènes "TE-free" et "TE-rich" possèdent des fonctions différentes. Les gènes dépourvus d'ET ont un rôle dans le développement et dans la régulation de la transcription, tandis que les gènes qui possèdent une forte proportion d'ET dans leurs séquences et leur voisinage sont impliqués dans les fonctions de métabolisme et de transport. Cette tendance a été observée pour chacune des quatre familles d'ET (transposons à ADN, rétransposons à LTR, LINE et SINE). Par conséquent, on peut extrapoler les résultats de l'étude réalisée sur les éléments *Alu* des chromosomes humains 21 et 22 [Grover *et al.*, 2003], qui a montré que les gènes "*Alu*-rich" et "*Alu*-poor" possèdent des fonctions différentes, à l'ensemble des ET ainsi qu'à la totalité du génome humain. Des fonctions similaires ont été trouvées également pour les gènes qui ne possèdent pas d'ET dans leurs introns ; ces gènes sont impliqués dans la morphogenèse et dans la transcription et possèdent des régions introniques extrêmement conservées [Sironi *et al.*, 2006]. La même tendance a été observée dans notre analyse dans laquelle seuls les ET complets ont été considérés. Ceci suppose fortement que les différences de fonctions observées entre les gènes expliquent la présence ou l'absence des séquences d'ET complètes.

Malgré la grande proportion d'ET présents dans les génomes de mammifères, certaines régions, dont la taille peut aller jusqu'à 100 kb, sont complètement dépourvues de toute insertion d'ET [Simons *et al.*, 2006]. Ces régions, connues sous le nom de TFR (pour "Transposon Free Region"), ont été également identifiées dans les génomes des amphibiens ainsi que ceux des poissons [Simons *et al.*, 2007]. La plupart de ces régions sont associées à des gènes impliqués dans des fonctions importantes telles que la régulation de la transcription et le développement. L'existence de telles régions est probablement due à une forte pression de sélection empêchant leur interruption par des séquences d'ET. Ainsi, le maintien des régions dépourvues de rétrotransposons a été montré comme une conséquence de l'action de la pression de sélection contre l'interférence transcriptionnelle des ET, c'est-à-dire contre l'activité transcriptionnelle des ET qui pourrait interférer avec celle des gènes voisins [Mourier et Willerslev, 2008]. La sélection agirait ainsi en empêchant la fixation de

toute insertion d'ET dans le voisinage des gènes qui possèdent des fonctions importantes. Dans notre étude, seuls 14% (875 gènes parmi 6185) des gènes "TE-free" sont dans les régions TFR. Ce faible pourcentage peut être expliqué par le fait qu'on ait considéré seulement les ET complets. Il est toutefois important de noter que les gènes "TE-free" possèdent tous le même type de fonctions que les gènes localisés dans les TFR.

La différence de contenu en ET entre les gènes "TE-free" et "TE-rich" peut être due à une pression de sélection purificatrice plus forte qui agit sur les gènes "TE-free" par rapport aux gènes "TE-rich". Cette hypothèse est soutenue par le taux ω des gènes "TE-free", qui est significativement plus faible que celui des gènes "TE-rich" quand les gènes orthologues de l'homme et de la souris sont comparés. Ceci suggère que cette pression de sélection purificatrice tend à éliminer toute insertion d'ET complets à l'intérieur ou dans le voisinage des gènes "TE-free". Néanmois, cette pression de sélection est difficile à mettre en évidence quand deux espèces proches, comme l'homme et le chimpanzé, sont comparées. Deux facteurs peuvent expliquer cette difficulté : 1) le temps, relativement court (6 Millions d'années), écoulé depuis la séparation entre l'homme et le chimpanzé n'est pas suffisant pour que la sélection puisse agir sur tous les gènes qui ont subi des insertions d'ET, 2) l'identité forte entre les deux génomes ne permet pas de détecter de telles contraintes sélectives. Ce même raisonnement peut expliquer les valeurs du taux ω dans la comparaison multi-espèces. Dans cette comparaison, on s'attendait à obtenir des taux ω différents si la pression de sélection était le facteur qui explique la différence de densité d'ET. Cependant, les résultats obtenus montrent que le modèle qui assume une pression de sélection ω identique (modèle M0) sur toutes les branches de l'arbre, est toujours le modèle qui correspond le mieux à nos données, et ceci même quand les gènes orthologues possèdent des densités en ET différentes. Ceci suggère que les espèces de primates sont trop proches pour pouvoir détecter d'éventuelles différences de pression de sélection, différences qui sont détectées si on compare les gènes orthologues de l'homme et de la souris.

Les régions non codantes peuvent être la cible de pression de sélection [Lowe *et al.*, 2007]. Ainsi la comparaison de la conservation de séquences des régions situées en amont et en aval des gènes "TE-free" et "TE-rich" entre homme et chimpanzé,

homme et orang-outan et homme et macaque, a montré que le pourcentage moyen d'identité de séquences des régions entourant les gènes "TE-free" était significativement supérieur à celui des régions encadrant les gènes "TE-rich", quand 10 kb de régions flanquantes sont considérées. Ceci implique que les régions encadrant les gènes "TE-free" sont plus conservées que celles qui encadrent les gènes "TE-rich". Cependant l'analyse limitée aux régions flanquantes de 2 kb, ne montre pas une conservation de séquences plus forte pour les gènes "TE-free" dans les comparaisons homme et chimpanzé, et homme et orang-outan. On peut supposer que les promoteurs des gènes, qui sont probablement localisés dans les régions flanquantes de 2 kb, sont déjà sujets à de fortes pressions de sélection purificatrice étant donné leur importance fonctionnelle, et ceci indépendamment de la présence ou l'absence d'ET. On peut alors penser que la significativité de l'analyse (comparaison taux d'identité à 2 kb de régions flanquantes) dans la comparaison homme-macaque peut être due au taux d'identité global plus faible entre les deux espèces ($\sim 93,5\%$) [Rhesus Macaque Genome Sequencing and Analysis Consortium , 2007]. Ceci a permis de montrer que quand le taux global d'identité de séquences est élevé, il est difficile de détecter une différence de conservation de séquences qui dépend d'un facteur donné (par exemple, la densité en ET). On peut ainsi conclure que les régions flanquantes des gènes "TE-free" sont plus fortement conservées suite, probablement, à une exposition à une pression de sélection purificatrice plus forte que celle qui agit sur les régions flanquantes des gènes "TE-rich". Ces résultats confirment d'anciennes observations qui avaient suggéré que la différence de densité en ET entre les gènes était due à une action de la pression de sélection, sans toutefois quantifier cette dernière. On a ainsi montré que les pressions de sélection n'agissent pas seulement au niveau des régions codantes, comme il est souvent supposé, mais également au niveau des régions non codantes.

La régulation de l'expression génique peut se faire au niveau transcriptionnel par l'intermédiaire des régions promotrices et des régions régulatrices situées en amont des gènes (ou "*cis*-regulatory sequences"), au niveau post-transcriptionnel par le biais des régions non traduites (UTR, pour "UnTranslated Regions") des ARNm, et à un ordre-supérieur au niveau de la chromatine. Il a été montré que quasiment 25% des promoteurs humains contiennent des séquences qui dérivent des ET et que ces ET participent à hauteur de 2,5% aux "*cis*-regulatory sequences" [Jordan *et al.*, 2003]. Ainsi les éléments

SINE, qui possèdent des promoteurs de l'ARN polymérase III, peuvent promouvoir la transcription des gènes dépendants de l'ARN polymérase II [Oliviero et Monaci, 1988] et possèdent dans leurs séquences des sites de fixation des facteurs de transcription [Polak et Domany, 2006], leur permettant ainsi de contrôler l'activité des gènes localisés dans leur voisinage. Néanmois, il a été noté que l'enrichissement en éléments *Alu*, à l'intérieur et aux alentours des gènes largement exprimés, peut être seulement dû à l'insertion préférentielle de ces éléments à côté des gènes de ménage [Urrutia *et al.*, 2008]. La présence des sites de fixation des facteurs de transcription a été montrée comme étant plus abondante au niveau des vieilles sous-familles des éléments *Alu* par rapport aux jeunes éléments, situés sur le chromosome humain 22 [Shankar *et al.*, 2004]. Ces résultats sont en accord avec nos observations puisque la sous-famille la plus abondante des SINE dans notre analyse correspond aux *Alu*S (sous-famille d'âge intermédiaire). De plus, les analyses de la région de fixation de sept facteurs de transcription chez les mammifères a montré que cinq de ces facteurs sont associés avec des familles distinctes d'ET, comme par exemple les ERV1 qui sont associés avec les TP153, indiquant ainsi que les ET jouent un rôle important dans l'expansion du répertoire des sites de fixation des facteurs de transcription chez les mammifères [Bourque *et al.*, 2008].

Les éléments *L1*, les rétrotransposons à LTR et les transposons à ADN montrent une affinité de fixation au nucléosome plus forte que celle des éléments *Alu*, ce qui peut aboutir à une différence de la conformation de la chromatine et par conséquent une différence dans l'expression des gènes [Huda *et al.*, 2009]. Les différentes familles d'ET peuvent ainsi avoir des influences variables sur l'expression des gènes. Quand toutes les familles d'ET sont considérées, les gènes humains dont les promoteurs sont enrichis en ET possèdent en moyenne un niveau d'expression plus fort et plus large que celui des gènes dont les promoteurs sont dépourvus d'ET [Huda *et al.*, 2009]. Cet effet des ET sur l'expression des gènes a été également trouvé chez les rongeurs, chez qui une corrélation significative a été observée entre l'expression des gènes et les ET insérés récemment, ce qui indique que ces insertions altèrent significativement l'expression des gènes [Pereira *et al.*, 2009]. La comparaison des voisinages génomiques des gènes de l'homme et du chimpanzé a également montré que l'expression des gènes qui possèdent un voisinage génomique conservé est différente de celle des gènes qui possèdent des insertions d'ET dans leur

voisinage [De *et al.*, 2009]. Étant donné l'influence des ET sur l'expression des gènes situés dans leur voisinage, on peut s'attendre à une pression de sélection purificatrice plus forte au niveau des gènes impliqués dans des fonctions essentielles, comme la régulation et le développement, leur permettant ainsi de devenir très probablement des gènes "TE-free". Cette pression est supposée être plus relachée au niveau des gènes impliquées dans d'autres fonctions moins vitales autorisant ainsi l'insertion et la fixation des ET dans leur voisinage.

La comparaison du niveau global d'expression des gènes "TE-free" et "TE-rich", dans 79 tissus humains, a montré un niveau d'expression moyen plus élevé des gènes "TE-rich" par rapport aux gènes "TE-free". Il est frappant que cette différence d'expression soit trouvé dans trois des six tissus tumoraux et dans tous les tissus du système immunitaire analysés. Ces résultats sont en accord avec l'étude de Lerat et Sémon [Lerat et Sémon, 2007], selon laquelle les niveaux d'expression des gènes diffèrent entre les tissus normaux et tumoraux suivant le nombre des SINE dans le voisinage. Les auteurs suggèrent une implication potentielle des SINE dans la cascade de dérégulation des gènes à l'état tumoral, alors que ces éléments sont réprimés dans les tissus normaux. Les modifications d'histones ou d'autres mécanismes épigénétiques telle que la méthylation d'ADN, qui sont connus pour leur répression de l'activité des ET, peuvent être associés à cette cascade de dérégulation dans les tissus tumoraux. C'est pour tester cette hypothèse que nous avons étudié les variations d'enrichissement des modifications d'histones associées aux différentes régions géniques, et à la présence d'ET, dans un tissu tumoral par rapport à un tissu normal.

5.2 Les modifications d'histones

Après la fin du projet de séquençage du génome humain en 2001, de nouveaux efforts ont été entrepris pour la complétion de l'épigénome humain [American Association for Cancer Research Human Epigenome Task Force, 2008]. Cet épigénome est supposé contenir toutes les modifications capables d'influencer l'expression des gènes sans modifier la séquence d'ADN sous-jacente [Bernstein *et al.*, 2007]. Parmi ces modifications, l'attention se porte aujourd'hui plus particulièrement sur deux catégories : la méthylation d'ADN et les modifications d'histones. Il a été largement montré au cours des 20 dernières années que ces deux modifications subissent des changements importants à l'état tumoral, cancéreux plus spécifiquement, par rapport à l'état normal, provoquant ainsi des modifications de l'expression génique ([Feinberg et Tycko, 2004], [Jones et Baylin, 2007]).

J'ai comparé les enrichissements des modifications d'histones entre deux lignées cellulaires du poumon, dont une extraite d'une personne saine et la deuxième d'une personne atteinte d'une leucémie (lignée de cellules cancéreuses). Je me suis intéressé à huit modifications d'histones, qui ont déjà fait l'objet de plusieurs études. Ces études ont montré que ces différentes modifications ont des localisations préférentielles différentes dans le gène. Ainsi, on sait maintenant que les marques H3K4me1, H3K4me2 et H3K4me3 sont localisées dans la région promotrice des gènes avec un enrichissement plus fort quand on s'approche du site d'initiation de la transcription (TSS, pour "Transcription Start Site") [Barski *et al.*, 2007]. C'est également le cas pour H3K9ac et H3K27ac pour lesquelles on a déjà montré qu'elles font partie d'un ensemble de 17 modifications d'histones qui possèdent une localisation préférentielle au niveau de la région promotrice et assurent un rôle important dans l'initiation de la transcription [Wang *et al.*, 2008]. À l'opposé, H3K36me3 et H4K20me1 montrent une localisation plus importante en aval du TSS [Barski *et al.*, 2007]. Nos résultats sont en accord avec ces observations puisqu'on observe un enrichissement plus important pour les marques H3K4me1, H3K4me2, H3K4me3, H3K9ac et H3K27ac dans la région "promoteur", alors que H3K36me3 et H4K20me1 montrent un enrichissement plus élevé dans les régions "exons" et "introns", respectivement. Enfin, la marque H3K27me3 montre un enrichissement faible quelle que soit la région considérée, ce qui peut s'expliquer par le fait que cette marque soit répressive

donc associée à des gènes non exprimés [Barski *et al.*, 2007], qui ne présentent qu'une petite partie de l'ensemble des gènes inclus dans notre analyse.

L'ensemble des modifications d'histones analysées subissent des variations d'enrichissement entre les deux états normal et tumoral. Ainsi, les marques H3K4me1, H3K4me2, H3K9ac, H3K27ac, H3K27me3 et H4K20me1 montrent un enrichissement significativement plus élevé à l'état tumoral par rapport à celui de l'état normal, ceci quelle que soit la région génique considérée. À l'inverse, les deux modifications d'histones H3K4me3 et H3K36me3 montrent une diminution significative de leur enrichissement à l'état tumoral sauf pour les "introns" dans lesquels H3K36me3 ne montre pas de différence significative d'enrichissement entre les deux états. Certaines études proposent que les variations globales des modifications d'histones peuvent servir de marqueur pour un type de cancer spécifique. Ainsi un enrichissement plus faible des deux marques H3K4me2 et H3K18ac est un prédicteur d'un risque élevé d'un cancer de la prostate, et une diminution globale de l'enrichissement des modifications d'histones indique un phénotype cancéreux plus agressif [Seligson *et al.*, 2009]. Au vue des résultats obtenus dans notre analyse on peut supposer que les deux modifications d'histones H3K4me3 et H3K36me3 pourraient être associées à une leucémie myéloide. Cette hypothèse nécessitera certainement de plus amples vérifications en étudiant les variations d'enrichissement d'autres modifications d'histones.

La séparation des gènes en quatre classes suivant leur divergence d'expression génique, a montré les mêmes variations d'enrichissement entre les deux états normal et tumoral quelle que soit la classe considérée. Il existe cependant une exception majeure concernant H3K36me3 qui ne montre pas de différence d'enrichissement significative entre les états normal et tumoral pour les gènes des classes 3 et 4 de divergence d'expression. Ceci a été observé quand le gène entier a été analysé et semble être essentiellement dû aux régions "introns" et "promoteur" pour lesquelles on observe le même résultat. Plusieurs études récentes ont montré que H3K36me3 montre un enrichissement important spécifiquement an niveau des régions exoniques des gènes dans lequelles elle est supposée jouer un rôle primordial dans l'épissage alternatif ([Kolasinska-Zwierz *et al.*, 2009], [Schwartz *et al.*, 2009], [Muers, 2009], [Luco *et al.*, 2010], [Huff *et al.*, 2010]). Ainsi, on

peut supposer que pour modifier l'expression d'un gène, les variations d'enrichissement de H3K36me3 auront très probablement lieu au niveau des régions exoniques et à moindre niveau dans le reste du gène, ce qui peut expliquer la non significativité de la variation d'enrichissement de cette marque au niveau du promoteur et des introns.

De nombreuses études ont montré qu'à l'état tumoral, les ET sont sujets à des altérations de leur régulation épigénétique qui peut entraîner leur réactivation ([Muntean et Hess, 2009], [Belancio *et al.*, 2010]). Ainsi, l'étude de Fraga *et al.* [Fraga *et al.*, 2005] a montré une diminution d'enrichissement de H4K20me3 et H4K16ac spécifique au niveau des ET dans le cancer du colon et des tumeurs hématologiques malignes. De plus, Szpakowski *et al.* ont montré chez les primates une perte de la répression épigénétique des ET en condition tumorale qui affecte principalement les rétroéléments [Szpakowski *et al.*, 2009]. Les résultats de mon étude montrent une variation de l'enrichissement des marques d'histones associées aux ET entre les deux états normal et tumoral. Cette variation pourrait être responsable de leur dérégulation. De plus, ces variations d'enrichissement sont différentes selon la région à laquelle on s'intéresse. Ainsi, à titre d'exemple, cette variation d'enrichissement entre les deux états concerne la modification H3K4me1 dans les régions exoniques et introniques alors qu'elle concerne la modification H3K9ac dans la région promotrice. Ces résultats indiquent que l'association entre les ET et les modifications d'histones varient selon la région génique considérée. D'un autre côté, H3K4me1 et H3K9ac n'ont jamais été associées aux ET auparavant, puisqu'on considérait que ce sont les modifications d'histones H3K9me3 et H4K20me3 qui sont responsables de la répression des ET [Mikkelsen *et al.*, 2007]. Nos résultats sont les premiers à supposer cette association entre les ET et différentes modifications d'histones qui dépend à la fois de la région génique et de l'état de la cellule.

Les modifications d'histones ont une influence importante sur l'expression des gènes qui dépend à la fois de la modification en question mais également du résidu lysine sur lequel se situe cette modification (cf. Introduction, section 1.6, page 21). Pour vérifier ceci, on a divisé les gènes en quatre classes distinctes selon leur divergence d'expression entre les états normal et tumoral. Les variations d'enrichissement entre les quatre classes de divergence d'expression semblent vérifier l'impact des modifications d'histones

sur l'expression puisqu'on observe une augmentation de la variation d'enrichissement quand la divergence d'expression génique augmente. Néanmoins, les analyses ACP "inter groupes" et les études de corrélation entre les variations d'enrichissement des huit modifications d'histones et les divergences d'expression géniques ont montré qu'un très faible pourcentage ($\sim$0,21%) de la divergence d'expression entre les deux états peut être expliqué par les variations d'enrichissement des modifications d'histones. Plusieurs hypothèses peuvent expliquer ces résultats. Tout d'abord notre étude n'a porté que sur huit modifications d'histones alors que le génome en contient plus d'une centaine. On peut donc supposer que d'autres modifications d'histones pourraient expliquer cette divergence d'expression. De plus, il est possible que les altérations de la méthylation d'ADN soient majoritairement responsables de la divergence d'expression. On est malheureusement incapables d'éprouver cette hypothèse puisqu'on ne possède pas les données des taux de méthylation d'ADN de la lignée tumorale analysée.

5.3 Perspectives

Dans la première partie de ce travail, je me suis intéressé exclusivement aux ET complets, Il serait très important de voir si la fonction des gènes peut expliquer également la densité en ET incomplets dans leur voisinage. En effet, un ET incomplet, qui a gardé sa région promotrice, peut lui aussi influencer l'expression des gènes voisins. De plus mon analyse n'a porté que sur les primates et plus spécifiquement sur l'homme. Il est donc indispensable de vérifier les résultats obtenus chez d'autres espèces. Il faut noter ici que j'ai essayé de réaliser cette analyse chez diverses espèces (vache, poulet, poisson zèbre, drosophile, rat, souris et moustique) mais les annotations GO encore limitées pour ces différentes espèces ne m'ont pas permis d'avoir des résultats fiables. Enfin, la comparaison d'expression entre les gènes "TE-rich" et "TE-free" suppose une dérégulation des ET dans les tissus tumoraux. Cette dérégulation peut-elle également expliquer la différence d'expression des gènes selon leur densité en ET dans les tissus du système immunitaire ? La réponse à cette question nécessitera certainement des analyses plus approfondies mais il a été montré récemment que le transcriptome est comparable entre les cellules du tissu immunitaire et les tissus tumoraux [Yang *et al.*, 2008]. Il sera donc très important de décortiquer de façon plus approfondie le comportement des ET dans les tissus du système immunitaire.

Dans ma deuxième étude, je me suis intéressé à huit modifications d'histones, alors que l'homme en possède plus d'une centaine. Il serait donc très important de vérifier si les tendances observées pour les huit modifications d'histones seront identiques pour d'autres modifications. Il est clair qu'il est difficile d'analyser toutes les modifications d'histones simultanément mais certaines modifications semblent plus primordiales que d'autres selon la question que l'on se pose. Je peux citer par exemple les modifications H3K9me3 et H4K20me3 qui sont connues pour leur importance dans la répression des ET et qui seront donc les premières à analyser pour toute question relative aux ET. Les résultats de cette analyse indiquent également un effet important des ET sur la variation d'enrichissement des modifications d'histones. Cet effet des ET doit être décortiqué de façon plus détaillée en fonction de la famille de l'ET et de sa séquence (complète ou pas). De plus, cette analyse a montré un impact faible des variations d'enrichissement des huit modifications d'histones sur la divergence d'expression génique, mais le nombre de modifications est

faible pour pouvoir tirer des conclusions générales d'où l'utilité, encore une fois, de considérer plus de modifications d'histones. On sait aujourd'hui que les mécanismes épigénétiques comportent trois composantes majeures (méthylation d'ADN, modifications d'histones et les petits ARN). Il est donc important de considérer ces trois voies ensemble pour déterminer la part de chacune de ces composantes dans la dérégulation des cellules tumorales. Enfin, les données épigénétiques disponibles ne cessant d'augmenter dans les bases des données, la répétition de la deuxième analyse avec un nombre plus important de tissus (normaux et tumoraux à la fois) me semble indispensable pour éprouver les hypothèses émises lors de cette analyse.

Annexes

A Variations d'expression et richesse en ET

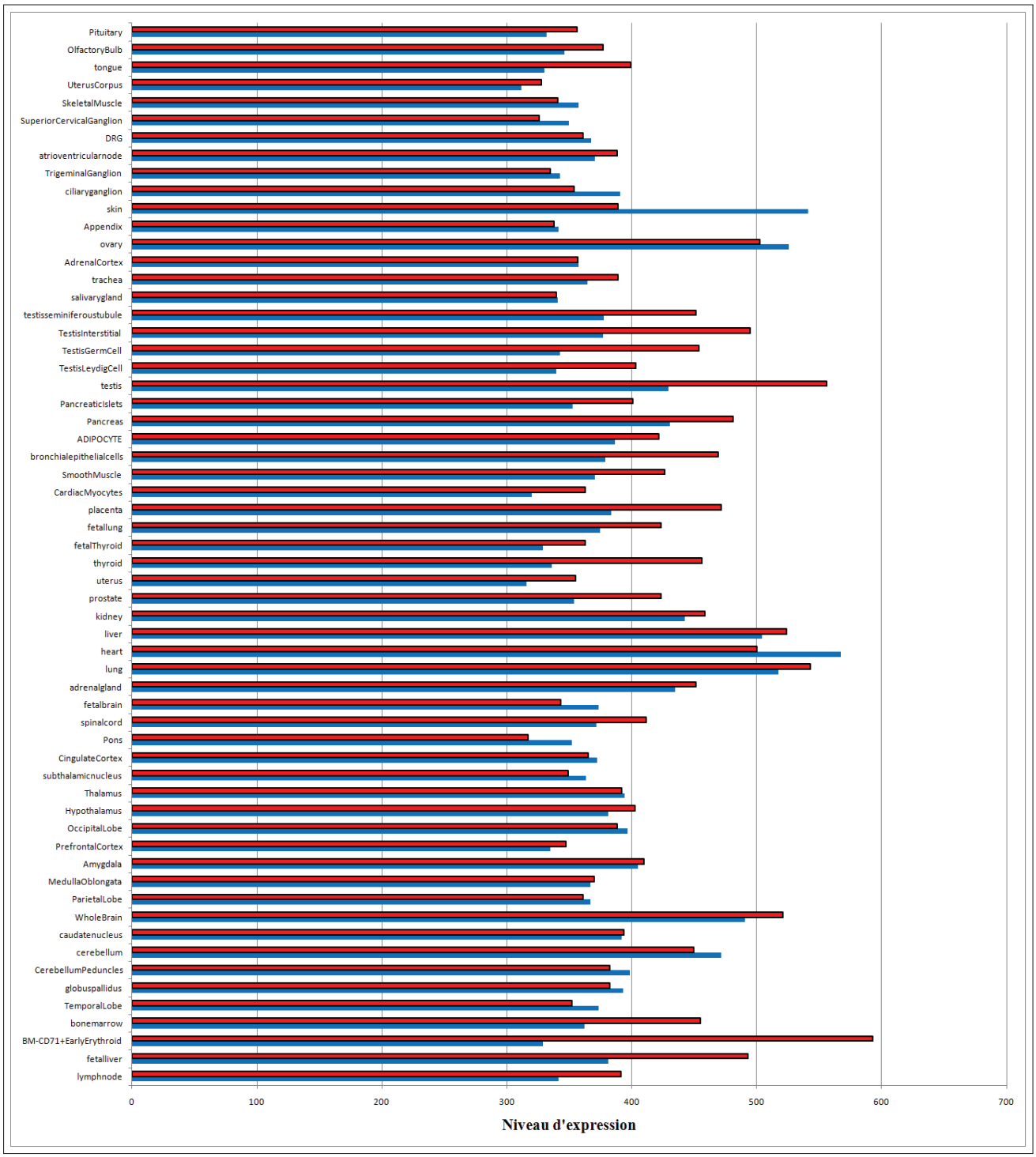

FIGURE 32: *Niveaux d'expression des gènes TE-free et TE-rich pour 60 tissus humains. Seule la région flanquante de 2 kb a été considérée. les barres rouges indiquent les gènes TE-rich et les barres bleues correspondent aux gènes TE-free.*

B Résultats de la comparaison des fonctions des gènes suivant leur densité en ET

B.1 Régions flanquantes de 2 kb

Tableau 16: *Fonction gènes et densité en ET - niveau 3. Le pourcentage de gènes "TE-free" et "TE-rich" impliqués dans les différents processus biologiques qui ont montré une différence significative entre les deux types de gènes. Les résultats correspondent au niveau 3 des processus biologiques.*

Processus biologique	Gènes "TE-free"	Gènes "TE-rich"	P-value	P-value corrigée
Multicellular organismal development	74,21%	25,79%	5.68784e-20	3.86773e-18
Neurological process	78,76%	21,24%	2.66223e-18	9.05157e-17
Anatomical structure development	73,49%	26,51%	8.24867e-17	1.8697e-15
Cell communication	62,51%	37,49%	3.97166e-11	6.75182e-10
Cellular metabolic process	43,44%	56,56%	4.35423e-09	5.92175e-08
Biosynthetic process	29,03%	70,97%	3.38035e-08	3.28377e-07
Establishment of localization	32,52%	67,48%	3.19691e-08	3.28377e-07
Protein localization	19,45%	80,55%	1.34186e-07	1.14059e-06
Primary metabolic process	44,06%	55,94%	3.71337e-07	2.80565e-06
Response to endogenous stimulus	22,64%	77,36%	0.000408092	0.00277502
Catabolic process	33,59%	66,41%	0.00423334	0.0205619
Response to biotic stimulus	70,92%	29,08%	0.00394885	0.0205619
Macromolecule metabolic process	46,09%	53,91%	0.00347692	0.0205619
Cellular developmental process	57,99%	42,01%	0.00402306	0.0205619
Cell adhesion	70,50%	29,50%	0.00552801	0.0250603
Reproductive process	74,69%	25,31%	0.00994027	0.0422461
Regulation of biological process	54,39%	45,61%	0.012372	0.0494881

Tableau 17: *Fonction gènes et densité en ET - niveau 4. Le pourcentage de gènes "TE-free" et "TE-rich" impliqués dans les différents processus biologiques qui ont montré une différence significative entre les deux types de gènes. Les résultats correspondent au niveau 4 des processus biologiques.*

Processus biologique	Gènes "TE-free"	Gènes "TE-rich"	P-value	P-value corrigée
System development	77,23%	22,77%	2.68341e-18	4.34713e-16
Sensory perception	80,65%	19,35%	8.10445e-18	6.5646e-16
Signal transduction	62,73%	37,27%	9.60894e-11	3.99902e-09
Anatomical structure morphogenesis	75,26%	24,74%	9.87412e-11	3.99902e-09
Protein metabolic process	34,38%	65,62%	2.10071e-10	6.80629e-09
Cellular macromolecule metabolic process	35,40%	64,60%	1.02964e-08	2.78003e-07
Transport	31,82%	68,18%	1.61985e-08	3.7488e-07
Establishment of protein localization	18,65%	81,35%	1.1741e-07	2.37755e-06
Cellular biosynthetic process	29,09%	70,91%	2.09924e-07	3.77864e-06
Pattern specification process	92,60%	7,40%	3.09028e-07	5.00625e-06
Embryonic development	81,85%	18,15%	7.63259e-06	0.000112407
Regulation of metabolic process	59,78%	40,22%	2.97447e-05	0.000376474
Cellular localization	27,10%	72,90%	3.02108e-05	0.000376474
Cell-cell signaling	73,65%	26,35%	4.4883e-05	0.000519361
Generation of precursor metabolites and energy	28,19%	71,81%	5.06243e-05	0.000546743
Cofactor metabolic process	15,36%	84,64%	0.00010146	0.00102729
Organic acid metabolic process	25,02%	74,98%	0.000118993	0.00113393
Phosphorus metabolic process	29,61%	70,39%	0.000132221	0.00118999
Response to DNA damage stimulus	17,88%	82,12%	0.000167953	0.00143202
Lipid metabolic process	28,43%	71,57%	0.000245982	0.00199246
Regulation of cellular process	55,87%	44,13%	0.00135538	0.0104558
Membrane organization and biogenesis	18,92%	81,08%	0.00170544	0.0125582
Cell differentiation	58,35%	41,65%	0.00259512	0.0182786
Amino acid and derivative metabolic process	27,47%	72,53%	0.00331331	0.0223649

Tableau 18: *Fonction gènes et densité en ET - niveau 5. Le pourcentage de gènes "TE-free" et "TE-rich" impliqués dans les différents processus biologiques qui ont montré une différence significative entre les deux types de gènes. Les résultats correspondent au niveau 5 des processus biologiques.*

Processus biologique	Gènes "TE-free"	Gènes "TE-rich"	P-value	P-value corrigée
Cell surface receptor linked signal transduction	74,31%	25,69%	1.6351e-20	4.64368e-18
Sensory perception of chemical stimulus	84,58%	15,42%	1.96438e-18	2.78942e-16
Organ development	78,61%	21,39%	2.33003e-15	2.20576e-13
Cellular protein metabolic process	34,59%	65,41%	1.355e-09	9.62053e-08
Nervous system development	79,86%	20,14%	2.22415e-09	1.26332e-07
Protein transport	16,22%	83,78%	3.93743e-08	1.86372e-06
Biopolymer modification	30,88%	69,12%	1.73459e-07	7.03746e-06
Embryonic morphogenesis	94,80%	5,20%	6.76012e-06	0.000239984
Macromolecule biosynthetic process	27,69%	72,31%	9.13964e-06	0.000288406
Transcription	60,67%	39,33%	1.50829e-05	0.000428356
Coenzyme metabolic process	11,20%	88,80%	1.80814e-05	0.000466829
Electron transport	19,08%	80,92%	2.38074e-05	0.000563443
Establishment of cellular localization	26,57%	73,43%	3.34499e-05	0.000730752
Regulation of cellular metabolic process	59,66%	40,34%	3.9558e-05	0.000769713
Cell fate commitment	100%	0%	4.06539e-05	0.000769713
Phosphate metabolic process	29,29%	70,71%	7.77444e-05	0.00137996
Carboxylic acid metabolic process	24,73%	75,27%	0.000109832	0.00183484
Vesicle-mediated transport	23,09%	76,91%	0.000234361	0.00369769
Negative regulation of developmental process	100%	0%	0.000371985	0.0055602
Regionalization	88,60%	11,40%	0.000398266	0.00565538
Cellular lipid metabolic process	28,22%	71,78%	0.000926022	0.0125233
Cell migration	79,42%	20,58%	0.001372	0.0177112
Regulation of hydrolase activity	17,09%	82,91%	0.00173163	0.0213819
Cell cycle phase	23,97%	76,03%	0.00350867	0.0415193

Tableau 19: *Fonction gènes et densité en ET - niveau 6. Le pourcentage de gènes "TE-free" et "TE-rich" impliqués dans les différents processus biologiques qui ont montré une différence significative entre les deux types de gènes. Les résultats correspondent au niveau 6 des processus biologiques.*

Processus biologique	Gènes "TE-free"	Gènes "TE-rich"	P-value	P-value corrigée
G-protein coupled receptor protein signaling pathway	80,09%	19,91%	1.22396e-23	4.39402e-21
Sensory perception of smell	84,92%	15,08%	3.63675e-17	6.52797e-15
Protein modification	30,49%	69,51%	1.09537e-07	1.3108e-05
Tissue development	86,97%	13,03%	1.56423e-07	1.40389e-05
Central nervous system development	86,92%	13,08%	3.5052e-06	0.000251673
Neurogenesis	84,17%	15,83%	1.1083e-05	0.000663133
Organ morphogenesis	78,25%	21,75%	1.38931e-05	0.000712519
Intracellular transport	22,23%	77,77%	1.66634e-05	0.00074777
DNA replication	14,14%	85,86%	2.09473e-05	0.000767111
RNA processing	21,76%	78,24%	2.1368e-05	0.000767111
RNA biosynthetic process	60,56%	39,44%	4.24912e-05	0.00138676
Skeletal development	87,14%	12,86%	4.8566e-05	0.00145293
Phosphorylation	27,24%	72,76%	5.56115e-05	0.00153573
Regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	59,69%	40,31%	8.2112e-05	0.00210559
Translation	27,35%	72,65%	0.000101373	0.00242618
DNA repair	15,54%	84,46%	0.00016645	0.00373472
M phase	17,67%	82,33%	0.000291224	0.00614997
Negative regulation of cell differentiation	100%	0%	0.000683203	0.0136261
Monocarboxylic acid metabolic process	19,57%	80,43%	0.00121651	0.0229856
mRNA metabolic process	23,19%	76,81%	0.00138581	0.0248753
Coenzyme biosynthetic process	13,53%	86,47%	0.00201057	0.0343711
Cell projection morphogenesis	91,63%	8,37%	0.00227398	0.0371072

Tableau 20: *Fonction gènes et densité en ET - niveau 7. Le pourcentage de gènes "TE-free" et "TE-rich" impliqués dans les différents processus biologiques qui ont montré une différence significative entre les deux types de gènes. Les résultats correspondent au niveau 7 des processus biologiques.*

Processus biologique	Gènes "TE-free"	Gènes "TE-rich"	P-value	P-value corrigée
Regulation of transcription	64,95%	35,05%	3.25274e-11	1.1905e-08
Transcription, DNA-dependent	65,18%	34,82%	7.73806e-11	1.41607e-08
Ectoderm development	92,96%	7,04%	2.33509e-07	2.74037e-05
Generation of neurons	88,51%	11,49%	2.99494e-07	2.74037e-05
Brain development	92,27%	7,73%	1.16444e-06	8.52371e-05
Intracellular protein transport	16,56%	83,44%	0.000209377	0.012772
Tissue morphogenesis	85,87%	14,13%	0.000313803	0.0164074
RNA splicing	18,81%	81,19%	0.000372714	0.0170517
Cell projection organization and biogenesis	92,96%	7,04%	0.000515431	0.0209609
mRNA processing	19,65%	80,35%	0.000713436	0.0253305
Post-translational protein modification	35,94%	64,06%	0.000761298	0.0253305

Tableau 21: *Fonction gènes et densité en ET - niveau 8. Le pourcentage de gènes "TE-free" et "TE-rich" impliqués dans les différents processus biologiques qui ont montré une différence significative entre les deux types de gènes. Les résultats correspondent au niveau 8 des processus biologiques.*

Processus biologique	Gènes "TE-free"	Gènes "TE-rich"	P-value	P-value corrigée
Regulation of transcription, DNA-dependent	64,12%	35,88%	1.27409e-09	4.63769e-07
Neuron differentiation	96,13%	3,87%	4.48659e-08	8.16559e-06
Epidermis development	92,20%	7,80%	8.7913e-07	0.000106668
Chromatin assembly or disassembly	89,87%	10,13%	0.000107519	0.00978422
Protein-DNA complex assembly	85,53%	14,47%	0.000357108	0.0259975
RNA splicing, via transesterification reactions	9,71%	90,29%	0.000585096	0.0339618
Transcription from RNA polymerase II promoter	67,42%	32,58%	0.000653112	0.0339618

Tableau 22: *Fonction gènes et densité en ET - niveau 9. Le pourcentage de gènes "TE-free" et "TE-rich" impliqués dans les différents processus biologiques qui ont montré une différence significative entre les deux types de gènes. Les résultats correspondent au niveau 9 des processus biologiques.*

Processus biologique	Gènes "TE-free"	Gènes "TE-rich"	P-value	P-value corrigée
Chromatin assembly	94,19%	5,81%	4.84438e-05	0.0120621
Neuron development	93,74%	6,26%	8.40563e-05	0.0120621
Regulation of transcription from RNA polymerase II promoter	69,98%	30,02%	0.000193169	0.0152934
Epidermis morphogenesis	89,02%	10,98%	0.000213149	0.0152934
RNA splicing, via transesterification reactions with bulged adenosine as nucleophile	10,19%	89,81%	0.000768606	0.044118

B.2 Régions flanquantes de 10 kb

Tableau 23: *Fonction gènes et densité en ET - niveau 3. Le pourcentage de gènes "TE-free" et "TE-rich" impliqués dans les différents processus biologiques qui ont montré une différence significative entre les deux types de gènes. Les résultats correspondent au niveau 3 des processus biologiques.*

Processus biologique	Gènes "TE-free"	Gènes "TE-rich"	P-value	P-value corrigée
Multicellular organismal development	81,91%	18,09%	1.08306e-19	6.49838e-18
Anatomical structure development	79,13%	20,87%	1.21998e-11	3.65994e-10
Regulation of biological process	64,94%	35,06%	1.77806e-08	3.55612e-07
Neurological process	78,85%	21,15%	9.06399e-05	0.0013596
Biosynthetic process	16,75%	83,25%	0.000423314	0.00507977
Macromolecule metabolic process	56,60%	43,40%	0.00302607	0.0302607

Tableau 24: *Fonction gènes et densité en ET - niveau 4. Le pourcentage de gènes "TE-free" et "TE-rich" impliqués dans les différents processus biologiques qui ont montré une différence significative entre les deux types de gènes. Les résultats correspondent au niveau 4 des processus biologiques.*

Processus biologique	Gènes "TE-free"	Gènes "TE-rich"	P-value	P-value corrigée
Regulation of metabolic process	73,49%	26,51%	8.74103e-15	1.11011e-12
System development	83,14%	16,86%	1.22145e-13	7.7562e-12
Nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	67,32%	32,68%	5.47386e-10	2.31727e-08
Regulation of cellular process	66,11%	33,89%	3.90227e-09	1.23897e-07
Pattern specification process	94,54%	5,46%	7.62612e-09	1.93703e-07
Biopolymer metabolic process	63,90%	36,10%	3.79522e-08	8.03322e-07
Protein metabolic process	18,51%	81,49%	1.05887e-06	1.92108e-05
Cellular macromolecule metabolic process	19,04%	80,96%	2.84829e-06	4.52167e-05
Embryonic development	89,28%	10,72%	1.67485e-05	0.000236341
Sensory perception	81,63%	18,37%	3.86859e-05	0.000491311
Anatomical structure morphogenesis	75,46%	24,54%	0.000116437	0.00134432
Cellular biosynthetic process	17,85%	82,15%	0.000966736	0.0102313
Generation of precursor metabolites and energy	10,63%	89,37%	0.00388738	0.0379767

Tableau 25: *Fonction gènes et densité en ET - niveau 5. Le pourcentage de gènes "TE-free" et "TE-rich" impliqués dans les différents processus biologiques qui ont montré une différence significative entre les deux types de gènes. Les résultats correspondent au niveau 5 des processus biologiques.*

Processus biologique	Gènes "TE-free"	Gènes "TE-rich"	P-value	P-value corrigée
Transcription	75,65%	24,35%	5.73128e-17	1.2093e-14
Regulation of cellular metabolic process	73,70%	26,30%	4.77301e-15	5.03552e-13
RNA metabolic process	72,18%	27,82%	8.24219e-11	4.34776e-09
Organ development	82,90%	17,10%	8.69638e-09	3.66987e-07
Nervous system development	88,58%	11,42%	8.69638e-09	3.66987e-07
Cellular protein metabolic process	18,76%	81,24%	1.55303e-06	5.46148e-05
Embryonic morphogenesis	97,84%	2,16%	4.56666e-06	0.000137652
Regionalization	92,82%	7,18%	2.49682e-05	0.000658536
Sensory perception of chemical stimulus	86,60%	13,40%	2.92097e-05	0.000684806
Cell migration	91,88%	8,12%	0.000132889	0.00280397
Embryonic pattern specification	100%	0%	0.000305392	0.00585798
Urogenital system development	96,28%	3,72%	0.00136907	0.0240728
Cell fate commitment	91,51%	8,49%	0.00157689	0.0255942
Biopolymer modification	19,13%	80,87%	0.00217844	0.0328322

Tableau 26: *Fonction gènes et densité en ET - niveau 6. Le pourcentage de gènes "TE-free" et "TE-rich" impliqués dans les différents processus biologiques qui ont montré une différence significative entre les deux types de gènes. Les résultats correspondent au niveau 6 des processus biologiques.*

Processus biologique	Gènes "TE-free"	Gènes "TE-rich"	P-value	P-value corrigé
RNA biosynthetic process	76,43%	23,57%	3.74406e-18	1.04085e-15
Regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	75,25%	24,75%	5.65793e-17	7.86452e-15
Central nervous system development	95,44%	4,56%	3.11737e-07	2.88876e-05
Neurogenesis	92%	8%	1.17317e-06	8.15353e-05
Skeletal development	92,62%	7,38%	2.57297e-06	0.000143057
Sensory perception of smell	87,35%	12,65%	2.06601e-05	0.000957253
Organ morphogenesis	83,14%	16,86%	0.000161238	0.00640347
G-protein coupled receptor protein signaling pathway	76,41%	23,59%	0.000192793	0.00669956
Sensory organ development	96,91%	3,09%	0.000238607	0.00737029
Kidney development	96,17%	3,83%	0.00150798	0.0419219

Tableau 27: *Fonction gènes et densité en ET - niveau 7. Le pourcentage de gènes "TE-free" et "TE-rich" impliqués dans les différents processus biologiques qui ont montré une différence significative entre les deux types de gènes. Les résultats correspondent au niveau 7 des processus biologiques.*

Processus biologique	Gènes "TE-free"	Gènes "TE-rich"	P-value	P-value corrigé
Transcription, DNA-dependent	75,65%	24,35%	1.89687e-19	5.10257e-17
Regulation of transcription	75,04%	24,96%	8.12551e-19	1.09288e-16
Generation of neurons	91,68%	8,32%	1.57099e-06	0.000140866
Brain development	94,75%	5,25%	2.46793e-06	0.000165968
Ear development	100%	0%	5.27714e-05	0.0028391

Tableau 28: *Fonction gènes et densité en ET - niveau 8. Le pourcentage de gènes "TE-free" et "TE-rich" impliqués dans les différents processus biologiques qui ont montré une différence significative entre les deux types de gènes. Les résultats correspondent au niveau 8 des processus biologiques.*

Processus biologique	Gènes "TE-free"	Gènes "TE-rich"	P-value	P-value corrigé
Regulation of transcription, DNA-dependent	74,72%	25,28%	3.20133e-19	9.31588e-17
Neuron differentiation	92,68%	7,32%	1.12682e-05	0.00163953
Ear morphogenesis	100%	0%	6.96931e-05	0.00676023
Inner ear development	100%	0%	0.000483925	0.0352055

Tableau 29: *Fonction gènes et densité en ET - niveau 9. Le pourcentage de gènes "TE-free" et "TE-rich" impliqués dans les différents processus biologiques qui ont montré une différence significative entre les deux types de gènes. Les résultats correspondent au niveau 9 des processus biologiques.*

Processus biologique	Gènes "TE-free"	Gènes "TE-rich"	P-value	P-value corrigé
Inner ear morphogenesis	100%	0%	0.000101936	0.0136948
Neuron development	92,70%	7,30%	0.00012859	0.0136948
Regulation of transcription from RNA polymerase II promoter	79,51%	20,49%	0.000201182	0.0142839

C Résultats de la méthode euclidienne

C.1 Gène entier

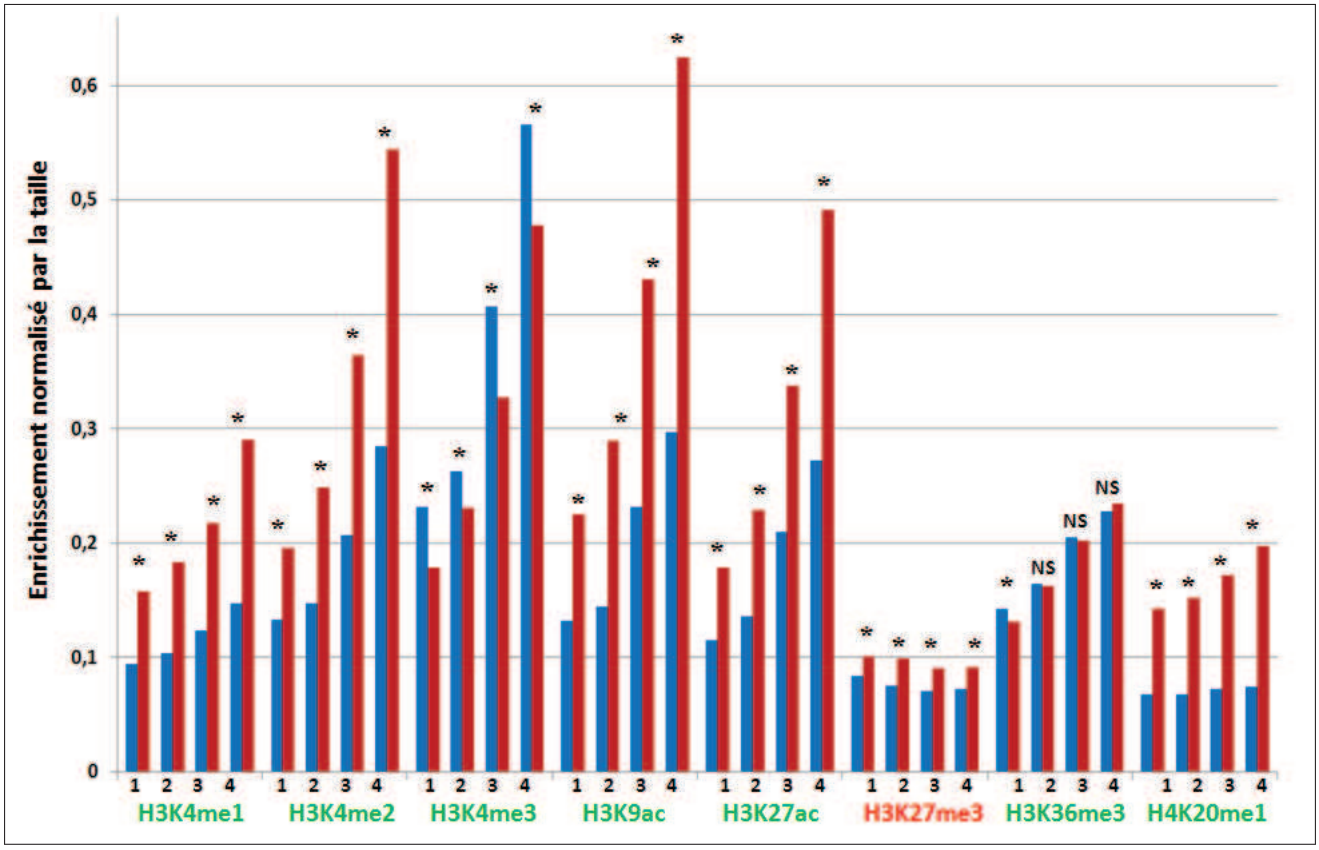

FIGURE 33: *Variations d'enrichissement des modifications d'histones dans les gènes pour quatre classes de divergence d'expression. Barres bleues : état normal ; barres rouges : état tumoral. Les modifications d'histones actives sont représentées en vert, les répressives en rouge. * : Différence d'enrichissement statistiquement significative (p-value < 0,05). NS : Différence d'enrichissement non significative. 1, 2, 3 et 4 : classes de divergence d'expression.*

C.2 Promoteur, exons et introns

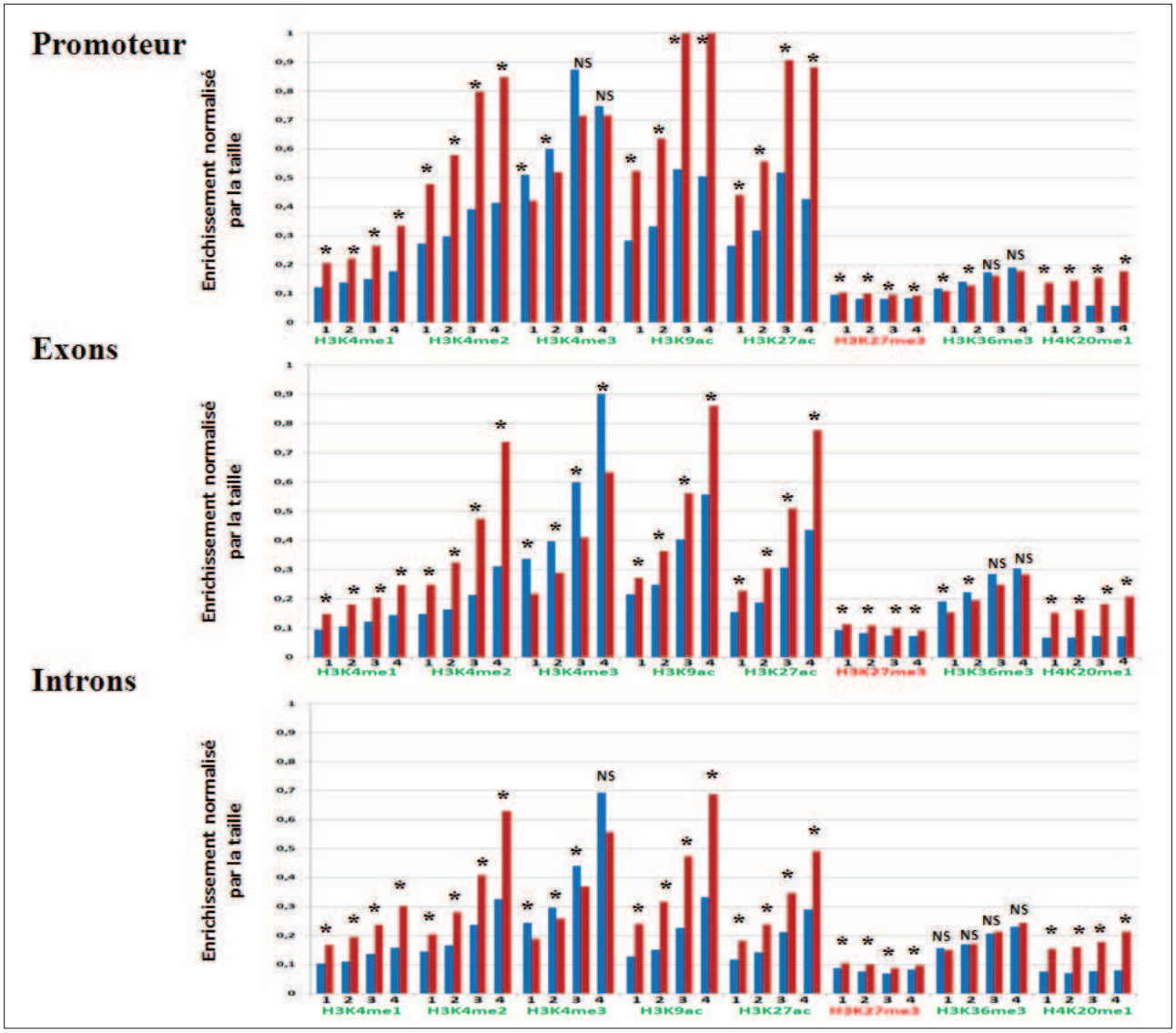

FIGURE 34: *Variations d'enrichissement des modifications d'histones dans les "promoteur", "exons" et "introns" pour les différentes classes de divergence d'expression. Barres bleues : état normal ; barres rouges : état tumoral. Les modifications d'histones actives sont représentées en vert, les répressives en rouge. * : Différence d'enrichissement statistiquement significative (p-value $< 0,05$). NS : Différence d'enrichissement non significative. 1, 2, 3 et 4 : classes de divergence d'expression.*

C.3 Absence/présence des Éléments Transposables

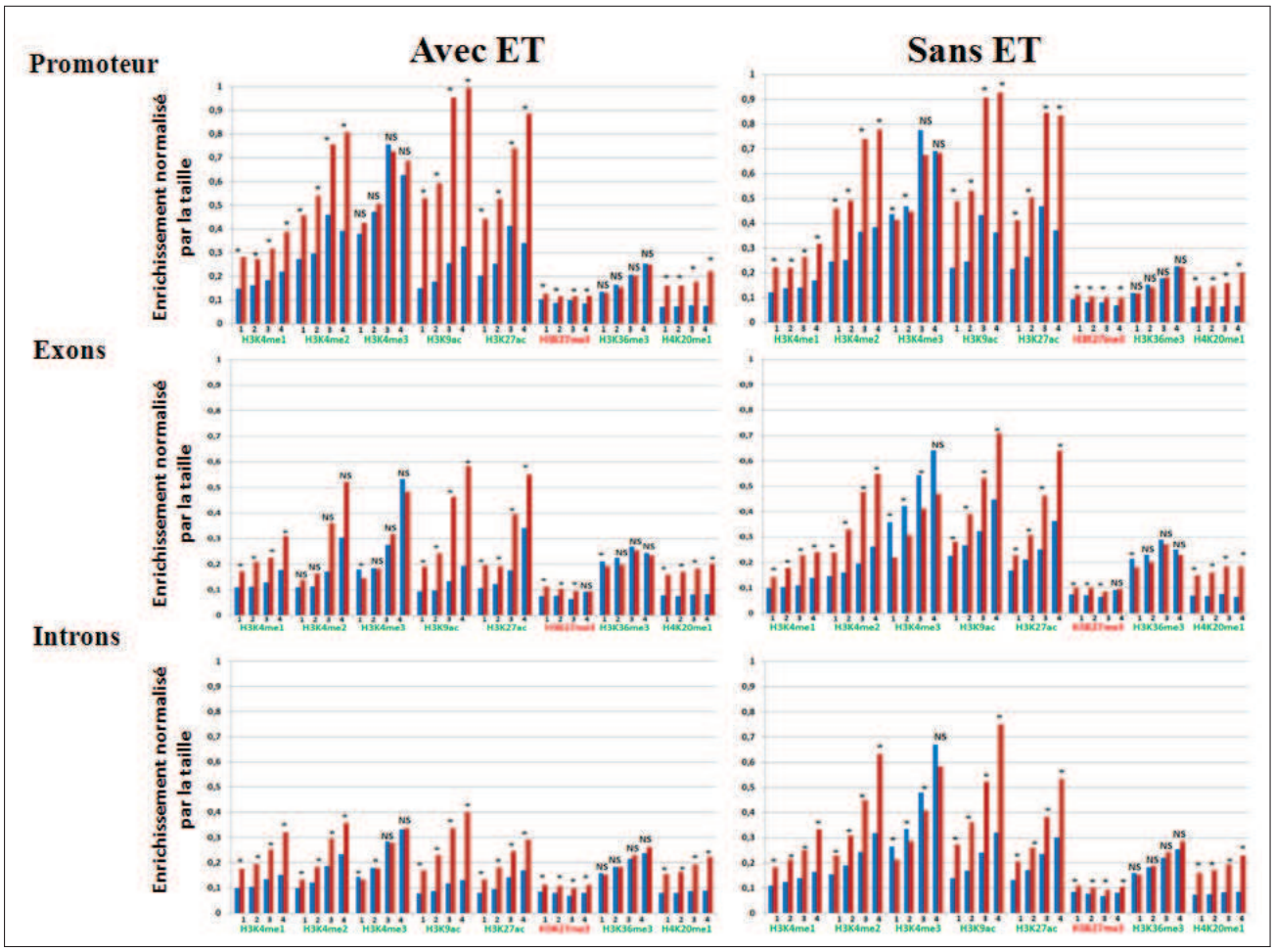

FIGURE 35: *Variations d'enrichissement des modifications d'histones dans les "promoteur", "exons" et "introns" selon la présence/absence d'ET pour les différentes classes de divergence d'expression. Barres bleues : état normal; barres rouges : état tumoral. Les modifications d'histones actives sont représentées en vert, les répressives en rouge. * : Différence d'enrichissement statistiquement significative (p-value < 0,05). NS : Différence d'enrichissement non significative. 1, 2, 3 et 4 : classes de divergence d'expression.*

D ACP et Éléments Transposables

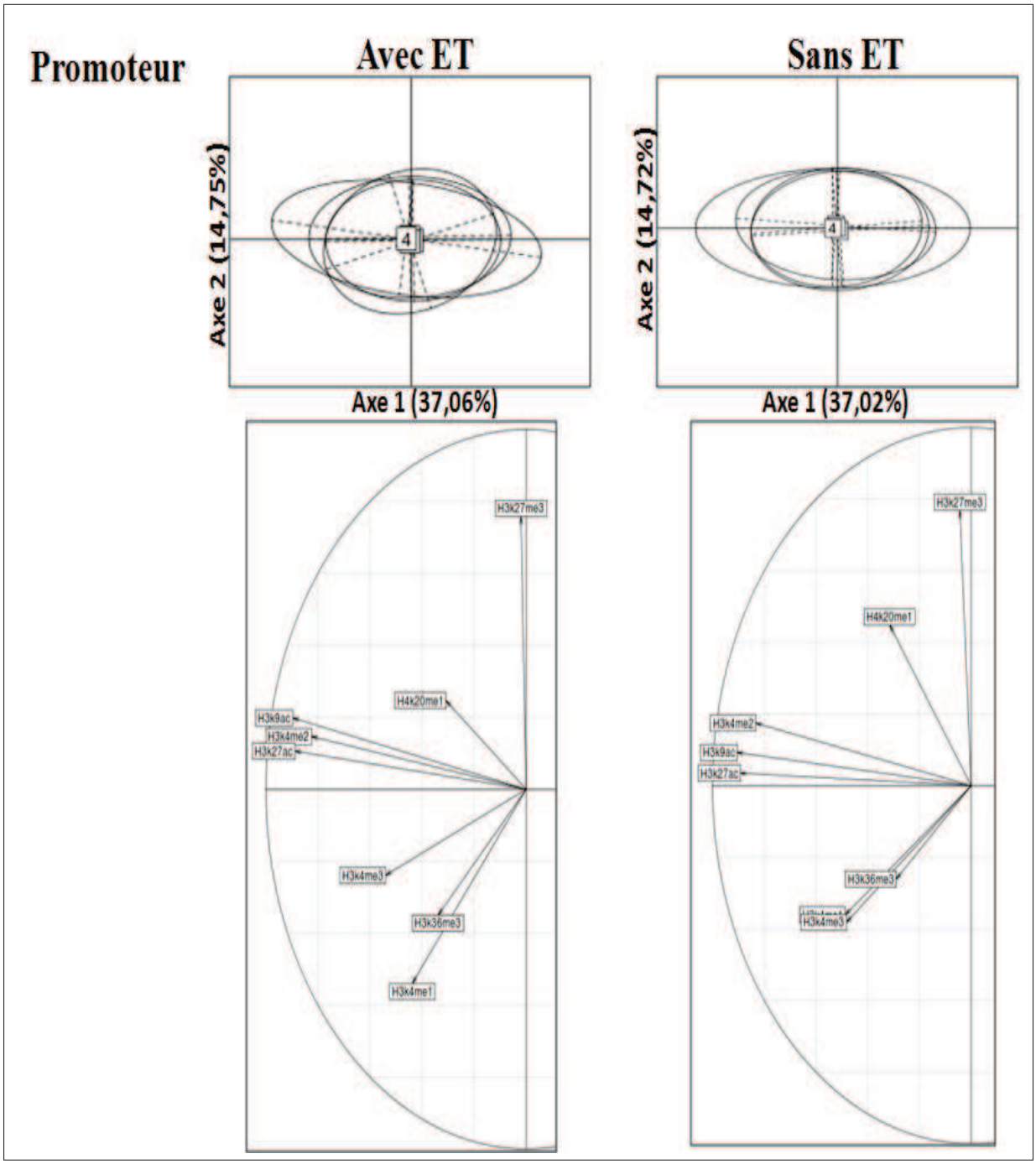

Figure 36: *Graphiques ACP "promoteur" et ET. La présentation graphique de la variation d'enrichissement des modifications d'histones entre les quatre classes de divergence d'expression en présence/absence des ET. 1 représente la classe de gènes avec la plus faible divergence d'expression et 4 la classe ayant la divergence d'expression la plus forte. En bas, le cercle des corrélations indiquant les huit modifications d'histones.*

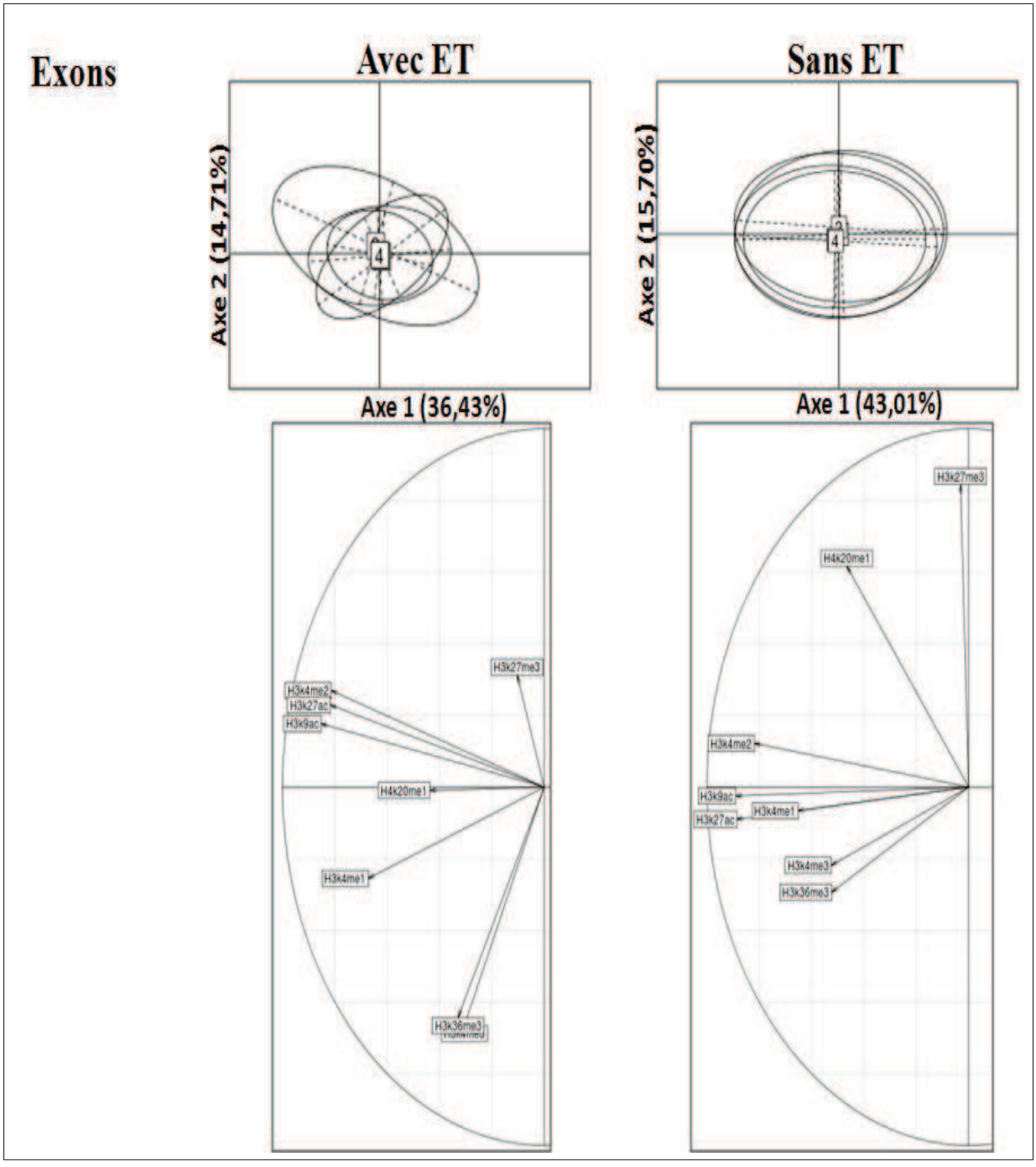

FIGURE 37: *Graphiques ACP "exons" et ET. La présentation graphique de la variation d'enrichissement des modifications d'histones entre les quatre classes de divergence d'expression en présence/absence des ET. 1 représente la classe de gènes avec la plus faible divergence d'expression et 4 la classe ayant la divergence d'expression la plus forte. En bas, le cercle des corrélations indiquant les huit modifications d'histones.*

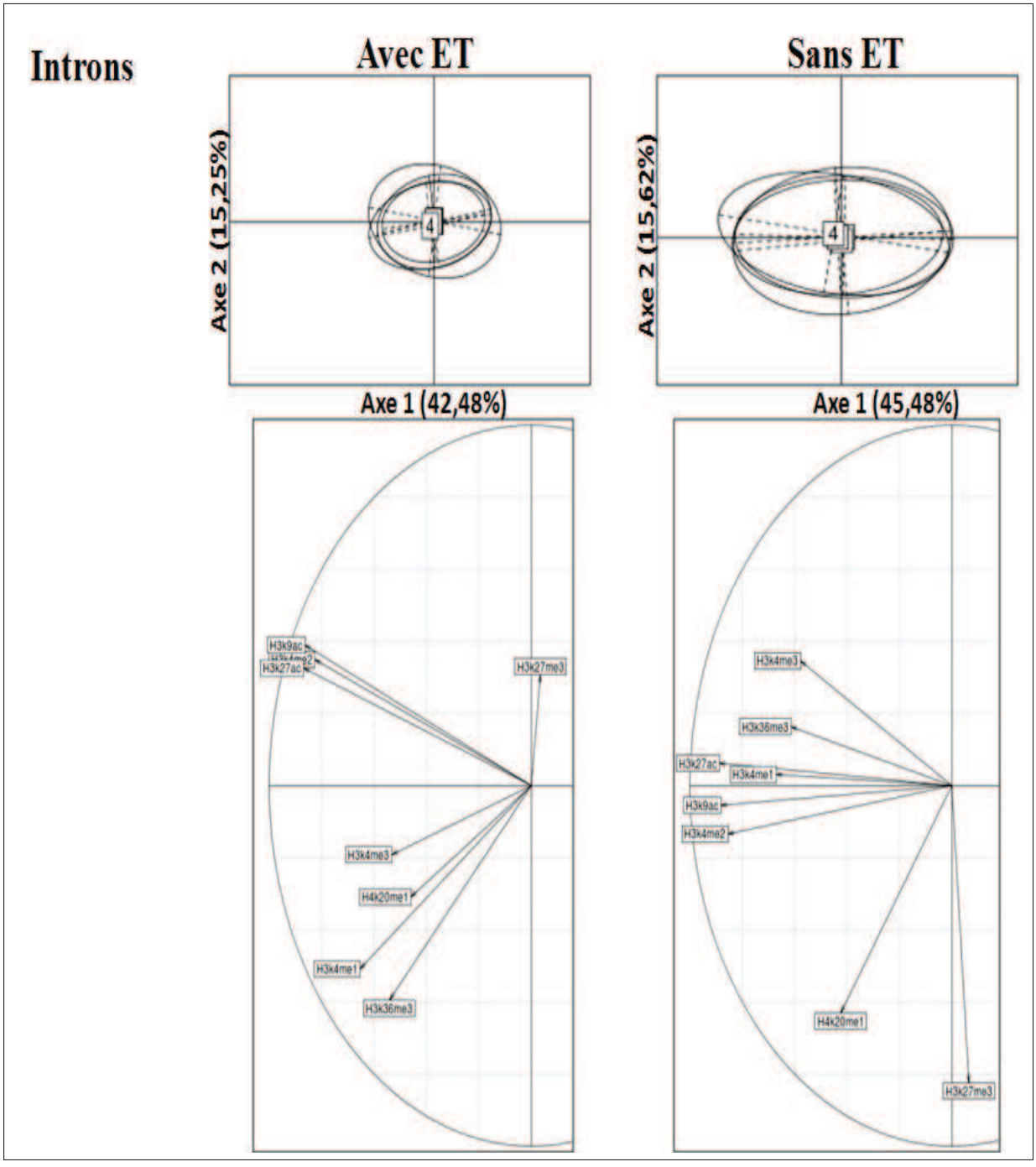

FIGURE 38: *Graphiques ACP "introns" et ET. La présentation graphique de la variation d'enrichissement des modifications d'histones entre les quatre classes de divergence d'expression en présence/absence des ET. 1 représente la classe de gènes avec la plus faible divergence d'expression et 4 la classe ayant la divergence d'expression la plus forte. En bas,le cercle des corrélations indiquant les huit modifications d'histones.*

E P-value des analyses statistiques

Tableau 30: *Les P-value des comparaisons statistiques après la correction de tests multiples (Benjamini et Hochberg) sont données pour chacune des régions étudiées ainsi que pour chacune des modifications d'histones.*

Figure	Region	H3K4me1	H3K4me2	H3K4me3	H3K9ac	H3K27ac	H3K27me3	H3K36me3	H4K20me1
20	Gène entier	8.8e-14	8.8e-14	8.8e-14	8.8e-14	8.8e-14	8.8e-14	0.002956	8.8e-14
21	Promoteur	8.8e-14	8.8e-14	8.8e-14	8.8e-14	8.8e-14	8.8e-14	2.38e-08	8.8e-14
21	Exons	8.8e-14	8.8e-14	8.8e-14	8.8e-14	8.8e-14	8.8e-14	8.8e-14	8.8e-14
21	Introns	8.8e-14	8.8e-14	8.8e-14	8.8e-14	8.8e-14	8.8e-14	1	8.8e-14
22	Promoteur [Avec ET]	8.8e-14	8.8e-14	1	8.8e-14	8.8e-14	8.8e-14	1	8.8e-14
22	Promoteur [Sans ET]	8.8e-14	8.8e-14	8.8e-14	8.8e-14	8.8e-14	8.8e-14	1	8.8e-14
22	Exons [Avec ET]	8.8e-14	1	6.24e-06	8.8e-14	8.8e-14	8.8e-14	3.62e-07	8.8e-14
22	Exons [Sans ET]	8.8e-14	8.8e-14	8.8e-14	8.8e-14	8.8e-14	8.8e-14	1	8.8e-14
22	Introns [Avec ET]	8.8e-14	8.8e-14	8.8e-14	8.8e-14	8.8e-14	8.8e-14	1	8.8e-14
22	Introns [Sans ET]	8.8e-14	8.8e-14	8.8e-14	8.8e-14	8.8e-14	8.8e-14	1	8.8e-14
23	Promoteur [Normal]	0.0003167	0.3598	0.003818	8.73e-08	0.0205	0.01432	3.68e-06	2.16e-08
23	Promoteur [Tumoral]	1.76e-08	0.8141	0.3968	0.8306	0.7152	1.98e-06	9.12e-06	5.85e-06
24	Exons [Normal]	0.0243	2.2e-16	2.2e-16	2.2e-16	2.2e-16	0.3573	0.2242	4.39e-05
24	Exons [Tumoral]	0.7731	2.2e-16	2.2e-16	2.2e-16	2.2e-16	0.7355	0.9192	0.7514
25	Introns [Normal]	2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.2e-16	7.24e-13	0.093	2.2e-16
25	Introns [Tumoral]	0.3271	2.2e-16	2.2e-16	2.2e-16	2.2e-16	8.94e-07	0.2204	1.44e-04
26	Promoteur [Normal]	0.003567	5.95e-10	1.07e-11	4.24e-14	0.0132	0.247	6.35e-06	2.27e-07
26	Promoteur [Tumoral]	2.95e-05	0.05859	0.01332	0.1695	0.9301	0.0011	2.52e-16	0.9293

Publication

J Mol Evol (2010) 71:180–191
DOI 10.1007/s00239-010-9376-5

Genes Devoid of Full-Length Transposable Element Insertions are Involved in Development and in the Regulation of Transcription in Human and Closely Related Species

Hussein Mortada · Cristina Vieira · Emmanuelle Lerat

Received: 1 March 2010 / Accepted: 26 July 2010 / Published online: 27 August 2010

Abstract Transposable elements (TEs) are major components of mammalian genomes, and their impact on genome evolution is now well established. In recent years several findings have shown that they are associated with the expression level and function of genes. In this study, we analyze the relationships between human genes and full-length TE copies in terms of three factors (gene function, expression level, and selective pressure). We classified human genes according to their TE density, and found that TE-free genes are involved in important functions such as development, transcription, and the regulation of transcription, whereas TE-rich genes are involved in functions such as transport and metabolism. This trend is conserved through evolution. We show that this could be explained by a stronger selection pressure acting on both the coding and non-coding regions of TE-free genes than on those of TE-rich genes. The higher level of expression found for TE-rich genes in tumor and immune system tissues suggests that TEs play an important role in gene regulation.

Keywords Transposable element · Gene function · Genome evolution · Primates

Electronic supplementary material The online version of this article (doi:10.1007/s00239-010-9376-5) contains supplementary material, which is available to authorized users.

H. Mortada · C. Vieira · E. Lerat
Université de Lyon, F-69000 Lyon, France

H. Mortada · C. Vieira · E. Lerat (✉)
Laboratoire de Biométrie et Biologie Evolutive,
Université Claude Bernard—Lyon 1, CNRS, UMR 5558,
F-69622 Villeurbanne, France
e-mail: emmanuelle.lerat@univ-lyon1.fr

Introduction

Transposable elements (TEs) are genomic sequences able to replicate themselves, and to move from one site to another within genomes. They are present in almost all the organisms in which they have been looked for, and can make up a large proportion of a genome: ~45% of primate genomes (The International Human Genome Sequencing and Analysis Consortium 2001; The Chimpanzee Genome Sequencing and Analysis Consortium 2005; The Rhesus Macaque Genome Sequencing and Analysis Consortium 2007), 38.5% of the mouse genome (The Mouse Genome Sequencing and Analysis Consortium 2002), 5% of euchromatin in *Drosophila melanogaster* (The Drosophila 12 Genomes Consortium 2007), and in some plants, such as maize, they can account for 80% of the genome (Schnable et al. 2009). Several classes of TE can be distinguished on the basis of their sequence and structure (Wicker et al. 2007), however, they are usually divided into two main classes based on their replication system (Finnegan 1989). The class II elements, or DNA transposons, use a DNA intermediate and move by a "cut and paste" mechanism; the class I elements, or retrotransposons, use an RNA intermediate and move by a "copy and paste" mechanism. The retrotransposons are further divided into two subclasses, those that have "long terminal repeats" (LTRs) at their extremities (LTR retrotransposons), and those that do not (non-LTR retrotransposons). There are also two subfamilies of non-LTR retrotransposons: the long interspersed nuclear elements (LINEs) and the short interspersed nuclear elements (SINEs).

TEs can insert into the *cis*-regulatory regions of genes, thus acting as regulatory sequences that control the expression of these genes (Mariño-Ramírez et al. 2005). For example, *Alu* elements, a particular family of SINEs,

contain binding sites for transcription factors, which allow these elements to control the activity of nearby genes (Polak and Domany 2006). The presence of SINEs in the gene neighborhood has been found to be associated with the deregulation of genes in tumoral conditions (Lerat and Sémon 2007). TE insertion and recombination between TE copies, either on the same or on different chromosomes, can also lead to duplication/deletion events, chromosomal translocations, and more complicated chromosomal rearrangements. These modifications are responsible for some human diseases (Kazazian and Moran 1998; Kazazian 2004; Deininger and Batzer 1999), and *Alu* insertions have been linked to 16 diseases (Deininger and Batzer 1999), and *L1* insertions to 15 diseases (Chen et al. 2005). For instance, one type of breast cancer is caused by an *Alu* insertion in exon 22 of the BRCA2 gene, and hemophilia A by the insertion of an *L1* in exon 14 of coagulation factor VIII (Kazazian et al. 1988). Forty-nine diseases, including Tay-Sachs disease (Myerowitz and Hogikyan 1987), are known to be the consequence of *Alu/Alu* recombination (Deininger and Batzer 1999). *L1/L1* recombination is known to be responsible for two diseases, glycogen storage disease (Burwinkel and Kilimann 1998) and duplication of the β-globin gene (Fitch et al. 1991). Despite their deleterious impact, TEs can also have positive effects on the host genome (Biemont and Vieira 2006), and some copies have been domesticated by the genome. The human protein SETMAR, which is a fusion between an H3 methylase and a transposase belonging to DNA transposons of the *mariner* family (Liu et al. 2007), has a role in both DNA methylation and DNA repair. The RAG1 and RAG2 proteins, which initiate the V(D)J recombination, a site-specific somatic recombination necessary for the assembly of the variable region of B-cell receptor/immunoglobulin and T-cell receptor genes from different gene segments, have been shown to be derived from an ancient transposase (Roth and Craig 1998).

The distribution of *Alu*s and LINEs in the human genome has been studied with regard to the characteristics of the genes into which or near to which these elements are inserted. Two categories of gene function have been distinguished according to their *Alu* content: *Alu*-poor genes are implicated in important functions such as regulation and transcription, whereas *Alu*-rich genes tend to be associated with transport or metabolism functions (Grover et al. 2003). This difference in *Alu* distribution could be explained by the involvement of *Alus* in regulatory processes. Human *L1* elements tend to be inserted into gene-poor regions, not only because they can affect gene expression since they have their own promoter in their 5′-UTR region, but also because of their capacity to produce truncated transcripts by introducing more transcription termination signals (Perepelitsa-Belancio and Deininger 2003) or by reducing the amount of transcript produced, thus, reducing protein expression (Han et al. 2004). Moreover, the function and expression level of genes have been found to be associated with TE insertion/fixation in mammalian introns (Sironi et al. 2006).

Most of the works referred to above, which analyze the relationships between the TEs and the host genes, concern particular types of TE families, such as *Alu* and *L1* elements, which constitute the largest proportion of the TEs in the human genome. Only a few studies have investigated the impact of DNA transposons or retrotransposons, including the inserted forms of retroviruses. In our study, we focused on the relationships between overall TE insertions and human genes. We have considered only the complete TE copies, compared to their reference, that could still have an intact and potentially active promoter region that could allow them to have a particular influence on neighboring genes. We show here that the functions of TE-rich genes differ from those of TE-free genes for all the TE families studied. This difference seems to be associated with the different selection pressures acting on the two categories of gene, with selection acting more strongly on TE-poor genes and their flanking regions. In addition, using microarray data, we show that the expression of TE-free and TE-rich genes differs in tumor and immune system tissues. This suggests that TE-rich genes are deregulated in tumor tissues, and that they are subject to specific regulation in immune system tissues.

Materials and Methods

The number of genes used in the different analyses are indicated in two summary figures proposed as supplementary material.

The Data

We used the BioMart tool (http://www.biomart.org/; Smedley et al. 2009) to extract all human gene coordinates (gene start and end, chromosome location) based on the Ensembl database (release 50, July 2008), which constitutes 35,641 genes. We then downloaded the list of TE coordinates obtained from the corresponding human genome version (hg18) available on the Repeat Masker website (http://www.repeatmasker.org/PreMaskedGenomes.html). For each gene, we considered the 2- and 10-kb flanking regions located upstream and downstream of the gene. We chose these values because it has been demonstrated that the promoter regions of human genes can be located up to the 10 kb of the gene start, with the majority of the promoters being within a 2-kb region (Kim et al. 2005).

We calculated the number of complete copies of TEs inserted into each gene and into its 2- and 10-kb flanking regions. We considered a TE copy to be complete if its length was at least 95% of that of the complete reference element, and if its divergence to the complete reference element was <20% (Lerat and Sémon 2007). We also considered all solo-LTRs of more than 95% of the length of full-length reference LTRs, and with a divergence to the reference LTR to be <20%, to be complete, because such sequences harbor regulatory regions. We then calculated the overall TE density for each gene as the fraction of the number of complete TEs within the gene and its flanking regions, and the difference between the length of the gene plus the length of the flanking regions, and the length of the TEs. The value found was multiplied by 10^4 to allow a more convenient representation of the data.

$$\text{TE density} = \frac{\text{Number of TEs}}{\text{Gene length} - \text{TE length}} \times 10^4$$

Similarly, for each gene we calculated the density in DNA transposons, LTR retrotransposons, LINEs, and SINEs. We used the TE density to define four classes of genes, extending from TE-free genes to TE-rich genes. We defined all genes with a TE density >10 as TE-rich genes.

We used the web-based application FatiGO (http://babelomics.bioinfo.cipf.es/EntryPoint?loadForm=fatigo; Al-Shahrour et al. 2004, 2005) to compare the functions of the two sets of genes (TE-free and TE-rich) for the two sizes of flanking region. FatiGO was used to extract Gene Ontology (GO) terms that were significantly over- or under-represented in one of the two sets of genes. GO terms are classified into three non-overlapping domains (The Gene Ontology Consortium 2000, 2001). The molecular function section of FatiGO describes the biochemical activities of gene products. The description is limited to describing the type of transcription product, without specifying either when or where it is produced (e.g., enzyme, transporter, ligand, etc.), the biological processes in which the gene product is involved (e.g., biosynthesis of cyclic AMP, cell growth, etc.), and the cellular component within which the products derived from the genes are located (e.g., nucleus, golgi apparatus, etc.).

Evolutionary Analyses

We used BioMart to extract the orthologous gene sequences of the chimpanzee, and then computed the selection pressure ratio, ω (or K_a/K_s, where K_a represents the rate of non-synonymous substitutions and K_s the rate of synonymous substitutions), of TE-free and TE-rich genes. We eliminated any human genes without orthologous genes in chimpanzee and the genes with internal stop codons. To avoid saturation in the computation of selective ratio and bias in the statistical analyses, we also removed the genes with almost identical sequences between human and chimpanzee (with a % identity between 99 and 100%). We performed pairwise alignments of each pair of human and chimpanzee genes using muscle (Edgar 2004), and we used the codeml program from the PAML package (Yang 2007) to calculate the ω ratio. The same method was used to compute the ω ratio between human genes and their orthologous genes in mouse.

We compared the ω ratios of orthologous genes along the phylogenetic tree of four primates: human, chimpanzee, orangutan, and rhesus macaque. Using BioMart, we searched for the "one-to-one" orthologous protein coding genes present just once in each of the four genomes. The orthologous genes obtained were aligned using muscle (Edgar 2004). Ten different scenarios were thus defined depending on the neighboring TE status of the genes (see Fig. 1). For example, scenario 1 indicates that the genes in all four species are TE-rich, whereas scenario 10 indicates that human genes are TE-free while the other primate genes are TE-rich. In order to evaluate the contribution of selection pressure to the difference in TE density observed between

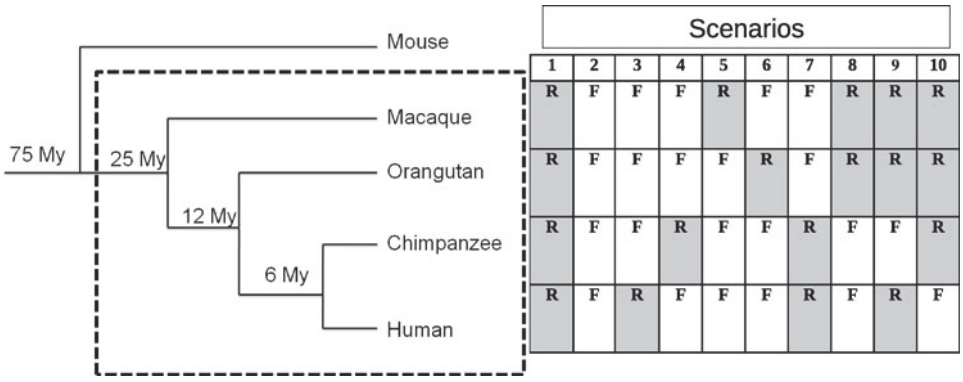

Fig. 1 Phylogeny of the four primates studied (branch length ignored). The tree is rooted by the mouse. The date of divergence in million years (My) is indicated on each branch. The table on the right shows the 10 gene evolution scenarios associated with the ω comparative analysis (see "Materials and Methods" section). F in white cells: TE-free genes; R in gray cells: TE-rich genes

the four primate species, different models were studied for the different scenarios considered. The M0 model assumes the same ω ratio for all branches in the tree; the M1 model supposes an independent ω ratio on each branch of the phylogeny; the M2 model assumes a $\omega 1$ ratio for a specific branch, which differs from the background $\omega 0$ ratio of the tree. These computations were done using the codeml program from the PAML package (Yang 2007). Two models were studied for each scenario: M0 versus M1 when the orthologous genes in the four primate genomes had the same TE density, and M0 versus M2 when the TE density of orthologous genes was different in at least one species. The statistical significance was assessed by comparing twice the difference in the likelihood scores (2ω), with a ω^2 distribution with a number of degrees of freedom equal to the difference in the number of parameters between the models.

The percentage identities of the flanking regions upstream and downstream of each orthologous gene pair of human and chimpanzee, human and orangutan, and human and macaque, were calculated using the dnadist program of the phylip package (Felsenstein 1989). To align the flanking sequences of each orthologous gene pair, the 2- and 10-kb flanking region sequences were submitted to RepeatMasker in order to mask the incomplete TEs present in these sequences. Sequences containing more than 60% of TE sequences were not taken into consideration in our analysis, because once the TEs had been masked, the DNA sequences were too short to be aligned.

Expression Data

The gene expression levels of human genes were retrieved from the data of Su et al. (2004), which were obtained by high-density oligonucleotide arrays. In this study, Su et al. (2004) surveyed the expression levels of almost all protein-encoding genes in 79 human tissues including six tumor tissues. Two determinations of the expression level were available for each tissue, and we used the average value as the expression level per tissue for each gene. In this way, we obtained the expression level for 158 TE-rich genes and 239 TE-free genes.

Results

Distribution of TEs in the Neighborhood of Genes in Four Primate Species

The number of human genes in the four classes of full-length TE density, for the 2- and 10-kb flanking regions is shown in Fig. 2. The distribution of the number of genes is the same whatever the size of the flanking region, with the number of genes decreasing as the TE density increases. Only a few genes were entirely devoid of complete TEs, with TE-free genes accounting for 17.4% (6,185 genes) and 1.3% (462 genes) of the total number of genes (35,641 human genes) for the 2- and 10-kb flanking regions, respectively. Genes with a TE density comprised between 0 and 5 constituted 50% (17,828 genes) and 62.2% (22,155 genes) of the total number of genes for the 2- and 10-kb flanking regions, respectively, while genes with a TE density comprised between 5 and 10 constituted 26.1% (9,307 genes) and 30% (10,693 genes), respectively. Finally, genes with a TE density of more than 10 (TE-rich genes) corresponded to 6.5% (2,321 genes) and 6.5% (2,331 genes) of the total number of genes for the 2- and 10-kb flanking regions, respectively.

We determined the proportions of LTR retrotransposons, LINEs, SINEs, and DNA transposons among the complete TEs located in and near human genes for both flanking region sizes (Table 1 and Supplementary Table 1). We found that SINEs were the most frequent elements,

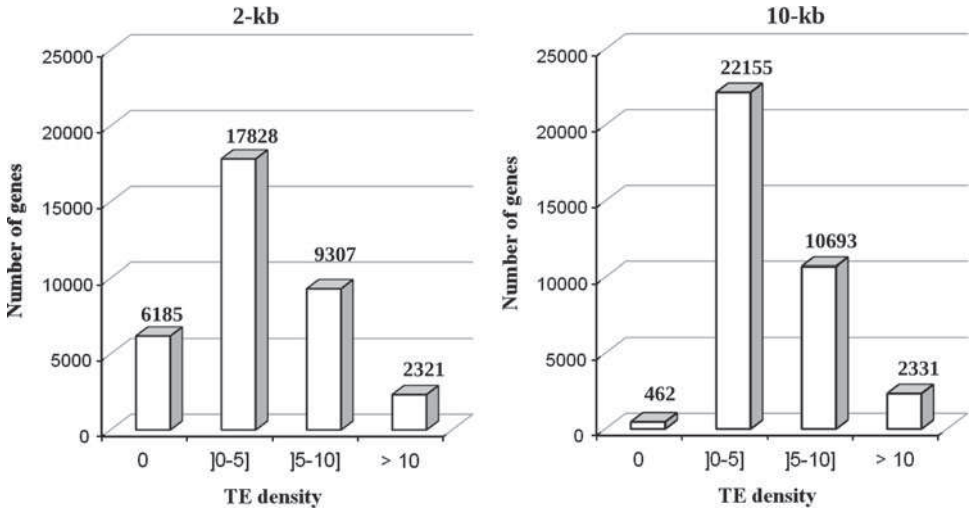

Fig. 2 Distribution of the number of human genes according to their TE density, for the 2- and 10-kb flanking regions. The numbers of genes are indicated at the top of each bar

Table 1 Proportion of each TE classes and subfamilies represented inside and in the neighborhood of genes

	Gene + 2-kb of flanking region			Gene + 10-kb of flanking region		
	% Of all TEs	Subfamilies	% Inside class	% Of all TEs	Subfamilies	% Inside class
DNA transposons	11.18	MER1	58.79	9.77	MER1	58.79
		MER2	23.06		MER2	22.71
		Mariner	4.49		Mariner	4.42
		Others	13.66		Others	14.08
LINEs	0.49	L1	99.57	0.41	L1	99.66
		Others	0.43		Others	0.34
SINEs	74.61	AluS	59.71	75.77	AluS	60.55
		AluJ	16.31		AluJ	15.81
		AluY	14.91		AluY	14.97
		Others	9.07		Others	8.67
LTR retrotransposons	13.72	MALR	62.91	14.05	MALR	59.88
		ERV1	21.29		ERV1	23.56
		ERVL	12.29		ERV2	12.22
		Others	3.51		Others	4.34

corresponding to 74.6 and 75.8% of complete TE copies considering the 2- and 10-kb flanking regions, respectively. The LTR retrotransposons constituted 13.7 and 14% of complete TE copies considering the 2- and 10-kb flanking regions, respectively, which was quite similar to the proportion of DNA transposons, which constituted 11.2 and 9.8% considering the 2- and 10-kb flanking regions, respectively. Finally, complete LINE elements were the least frequently occurring elements, accounting for only 0.5 and 0.4% considering the 2- and 10-kb flanking regions, respectively. These percentages are in agreement with the observations made globally for the complete human genome sequence (The International Human Genome Sequencing and Analysis Consortium 2001). We determined the distribution of the number of genes according to TE density for each TE family (see Supplementary Figs. 1, 2, 3, and 4). The numbers of genes per class of DNA transposon, SINE, or LINE showed the same distribution pattern as for all TEs, i.e., the number of genes decreased as the TE density increased. However, there were more DNA transposon-free and LINE-free genes than genes harboring these TEs in their vicinity. This tendency was less obvious for LTR retrotransposons, although there were fewer LTR retrotransposon-rich genes than genes with low LTR retrotransposon density. Overall, these results show that all TE families contribute to the trend observed for TEs as a whole.

In order to find out whether the density of complete TEs observed in human TE-free and TE-rich genes is conserved in closely related species, we looked at the orthologous genes in three species, chimpanzee, macaque, and orangutan, of which the genomes have been completely sequenced and TE lists are available. We selected human genes with only a single ortholog in each of the other three species, i.e., human genes that had only one orthologous gene in chimpanzee, one in macaque, and one in orangutan (see "Materials and Methods" section for details). We identified 14,744 human genes with one and only one ortholog in each of the other three species. We determined the density of TEs present in and near these genes for the 2- and 10-kb flanking regions in the other three primates. Among the 897 genes that were TE-free in the human genome, 67% (606 genes) were also TE-free in the chimpanzee, macaque, and orangutan, and among the 1,496 human TE-rich genes, 17% (263 genes) were also TE-rich in all of the other three species when we considered 2-kb flanking regions. When we considered the 10-kb flanking regions, we found that 56% of the 89 human TE-free genes (50 genes) were also TE-free in the other three primate species, and 15% (289 genes) of the 1,849 human TE-rich genes were also TE-rich in the other species. This indicates that TE-free genes seem to be more prone to conserve this characteristic throughout evolution than TE-rich genes to conserve the contrasting characteristic.

Gene Functions According to Their TE Neighborhood

We first compared the functions of the human genes displaying extreme TE densities, i.e., the TE-free genes and the TE-rich genes (TE density of more than 10). For the 2-kb flanking regions, we compared 6,185 TE-free genes to 2,321 TE-rich genes, and for the 10-kb flanking regions we compared 462 TE-free genes to 2,331 TE-rich genes.

Gene functions were assigned by the FatiGO software (Al-Shahrour et al. 2004, 2005) to seven levels, from level 3 to level 9, ranging from general to more specific functions. For clarity's sake, only the levels for which clear and

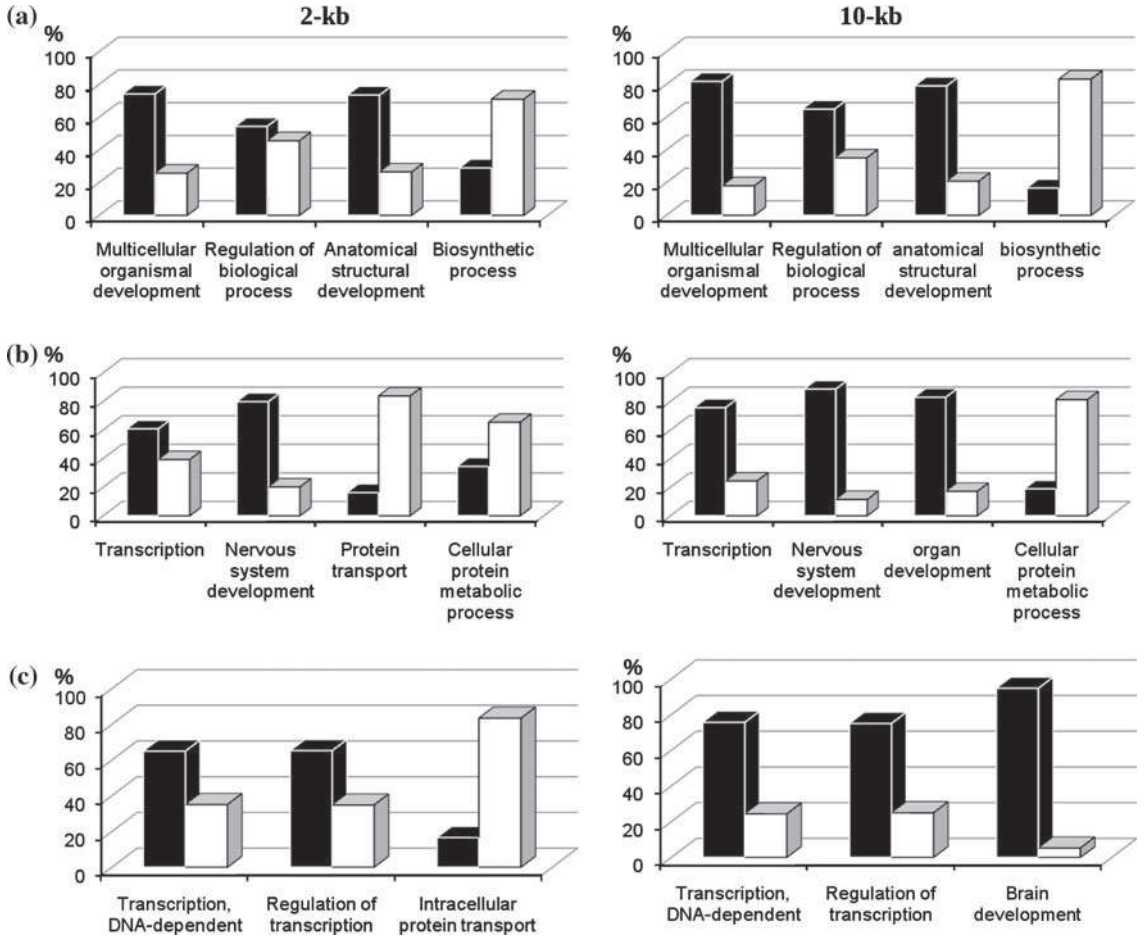

Fig. 3 Distribution of the percentage of TE-free (*black bars*) and TE-rich (*white bars*) genes involved in biological processes at gene ontology levels 3 (**a**), 5 (**b**), and 7 (**c**), considering 2-kb and 10-kb flanking regions

statistically significant differences between the TE-free and TE-rich genes were found for either the 2- or the 10-kb flanking regions are shown in Fig. 3. In the case of the 2-kb flanking regions, we observed for the three chosen levels that TE-free genes were more often involved than TE-rich genes in multicellular organismal development, nervous system development, and the regulation of transcription (Fig. 3). In contrast, TE-free genes were less often involved in biosynthetic processes, protein transport, and intracellular protein transport than TE-rich genes (Fig. 3). The same trends were observed for the 10-kb flanking regions. In order to verify if the effect we observed is related to the presence or absence of complete TEs, we compared the genes free of any TEs, both complete and incomplete (genes corresponding to those included in transposon free regions (TFRs) as defined by Simons et al. (2006, 2007)) (971 TFR genes), the genes that are complete TE-free and TE-partial rich (356 TCF genes), and the genes that are TE-partial free and TE-complete rich (386 TPF genes). No significant difference in the gene function was observed between the TFR genes and the TCF genes, whereas TFR genes and the TPF genes did display significant function difference in biological processes (see Supplementary Table 2). These results confirm that the function of these genes is mainly linked to the presence or absence of complete TE sequences inside the gene or in their flanking regions, with no influence of the presence or absence of partial TE sequences.

We then compared the functions of the genes belonging to the intermediate classes of TE density ($0 < \text{TE density} \leq 5$ and $5 < \text{TE density} \leq 10$) to those of TE-free and TE-rich genes, for both sizes of flanking region. Overall, we observed the same tendency as when the TE-free genes were compared to the TE-rich genes, although it was less marked, with the exception of the fact that the functions of the TE-rich genes did not differ significantly from those of the genes in the intermediate class (Chen et al. 2005; Finnegan 1989) (data not shown). We will therefore focus the rest of our analysis on the genes belonging to the extreme classes, i.e., the TE-free genes and the TE-rich genes.

We compared the functions of TE-free and TE-rich genes in orthologous genes in the chimpanzee, orangutan,

and macaque. Since orthologous genes are likely to be involved in the same functions, we looked for the ontologies of human genes and compared the functions of human TE-free genes that were also TE-free in the other three species (606 genes) to those of the human TE-rich genes that were also TE-rich in the other species (263 genes). We found the same degree of difference of function as when the functions of human TE-free genes were compared to those of human TE-rich genes, showing that this feature is conserved in primates.

Selective Pressures Acting on the Coding Regions of Genes Depending on Their TE Neighborhood

To test whether TE density could be linked to selective pressures acting on the genes, we computed the ω ratio of the human and chimpanzee TE-free and TE-rich orthologous genes. We analyzed 1,377 TE-free genes and 824 TE-rich genes for the 2-kb flanking regions, and 121 TE-free genes and 982 TE-rich genes for the 10-kb flanking regions. The median of the ω ratio appeared to be higher for TE-free genes than for TE-rich genes for both groups of flanking regions (0.33 vs. 0.28 for 2 kb, 0.37 vs. 0.26 for 10 kb), but this difference was statistically significant only for the 2-kb flanking region category (Wilcoxon rank sum test = 522,711.5, $P = 0.002$). This indicates that selection pressure on the coding part of the genes is not sufficient to explain the absence of TE insertions in and near the TE-free genes. This could be due to the fact that the two species are too closely related for any significant difference to be detected. To boost the power of the detection, we performed a comparative evolutionary analysis using the genomes of these four primates: human, chimpanzee, orangutan, and macaque. We considered 10 different scenarios (Fig. 1). In scenarios 1 and 2, all orthologous genes belonged to the same class (either TE-free or TE-rich), and two models were compared: the M0 model, which assumes the same selective ω ratio for all branches of the tree (Fig. 1), and the M1 model, which assumes an independent ω ratio for each branch of the phylogeny. The likelihood ratio test indicated that the M0 model fitted the data significantly better for both scenarios 1 and 2 (92% of genes fitted scenarios 1 and 2). This means that all branches of the tree have the same ω ratio, and so genes belonging to the same TE density class have been subjected to the same selection pressure.

In scenarios 3–10, orthologous genes can belong to different TE-density classes, in one or more species. We tested whether this could reflect the fact that different selection pressures were acting on these genes. For example, in scenario 3, the human genes were TE-rich, whereas the orthologous genes in the other three species were TE-free. This could result from exposure to more relaxed selection pressure in the human genome than in those of the other species. We then tested the M2 model, which assumes that the $\omega 1$ ratio for a specific branch (the human branch in scenario 3) differs from the background ratio $\omega 0$ of the tree, and compared it to the M0 model. Despite their differing TE densities, the selection pressures exerted on the orthologous genes showed no significant difference. In scenarios 3–10, the M0 model was always the one that fitted our data best, which indicates that the selection pressure is the same in all branches of the tree.

The genomic proximity of the human genome to that of the other three primates (the divergence between humans and macaques is dated to around 25 Million years (My) ago), which could explain the absence of obvious differences as a saturation phenomenon is possible. To overcome this difficulty, we compared the ω ratio of TE-free and TE-rich orthologous genes in the human and mouse genomes, species which diverged 75-My ago. We analyzed 991 TE-free orthologous genes and 745 TE-rich genes (2-kb flanking regions), and 91 TE-free genes and 874 TE-rich genes (10-kb flanking regions). The median of the ω ratio values was significantly lower for TE-free genes than for TE-rich genes for both sizes of flanking regions (0.15 vs. 0.17 for 2 kb, Wilcoxon rank sum test = 334,503, $P = 0.0008$; 0.09 vs. 0.16 for TE-rich genes for 10 kb, Wilcoxon rank sum test = 25,977, $P = 4.9e^{-8}$). This indicates that the selection pressure on the coding part of the gene can explain the differences in TE density between TE-free and TE-rich genes, with stronger negative selection pressure on the TE-free genes being evident when human and mouse genes were compared.

Selective Pressures Acting on the Flanking Regions of Genes Depending on Their TE Neighborhood

Since selection pressure can also act on non-coding regions, we next analyzed the sequence identity of the 2- and 10-kb regions downstream and upstream of the genes. For each human gene, we retrieved the sequences of the 2- and 10-kb regions upstream and downstream of its orthologous gene in chimpanzee, orangutan, and macaque. An example of the distribution of the percentage identity of the flanking regions for the different categories of genes in the human–chimpanzee comparison is shown in Supplementary Fig. 5. For the human–chimpanzee comparison, a Wilcoxon rank sum test showed that the mean percentage identity of the regions flanking TE-free genes is significantly higher than that of the regions flanking the TE-rich genes, for both the downstream and upstream regions, for 10-kb but not for 2-kb (Table 2). The percentage identity values are congruent with the global nucleotide divergence observed between the two genomes (from 0.5 to 3.0%, The Chimpanzee Genome Sequencing and Analysis Consortium 2005). The mean percentage identity is lowered by a minority of orthologous sequences

Table 2 Mean percentage identities of the regions flanking orthologous genes in the human, chimpanzee, orangutan, and macaque genomes

Species compared	Flanking region size (kb)	Number of TE-free genes	Number of TE-rich genes	Position of the flanking region relative to the gene	Mean % identity of the flanking regions of TE-free genes	Mean % identity of the flanking regions of TE-rich genes	*P* value
Human–chimpanzee	2	1,246	684	Upstream	96.3	96.1	0.60
				Downstream	95.7	96.0	0.29
	10	103	738	Upstream	95.8	94.7	1.6e−04*
				Downstream	96.9	90.9	9.2e−08*
Human–orangutan	2	1,178	532	Upstream	93.5	92.4	2.2e−04*
				Downstream	94.6	94.3	0.67
	10	95	445	Upstream	94.5	93.1	1.8e−04*
				Downstream	95.2	94.4	0.02*
Human–macaque	2	1,080	445	Upstream	90.9	87.3	8.9e−16*
				Downstream	91.8	90.6	3.7e−02*
	10	89	335	Upstream	92.5	89.1	3.1e−06*
				Downstream	92.9	90.5	3.9e−09*

* Significant *P* values

that have a high divergence (see Supplementary Fig. 8). When the human and orangutan genomes were compared, the mean percentage identities of the 2- and 10-kb upstream flanking regions, and the 10-kb downstream flanking regions, were significantly higher for regions flanking TE-free genes than for regions flanking TE-rich genes (Table 2). In the human/macaque comparison, the upstream and downstream 2- and 10-kb flanking regions of TE-free genes displayed significantly higher identity percentages than those flanking TE-rich genes (Table 2). These results indicate that the regions flanking TE-free genes seem to be better conserved through evolutionary time than those flanking TE-rich genes.

Gene Expression

Figure 4 shows the expression level variations of TE-free and TE-rich genes for 19 of the 79 human tissues analyzed by microarray (Su et al. 2004). The variation of the expression level of the remaining tissues is shown in Supplementary Fig. 9. A Wilcoxon rank sum test revealed that the level of expression of the TE-free genes was significantly lower than that of the TE-rich genes in 13 of the 79 tissues analyzed (indicated by the asterisks in Fig. 4). This difference was observed in three of the six tumor tissues, and in all 10 immune system tissues investigated.

Discussion

In this study, we analyzed the characteristics of human, chimpanzee, orangutan, and macaque genes in terms of the presence of complete TE copies either within their sequence or in their vicinity (within 2- or 10-kb flanking sequences). We showed that TE-free and TE-rich genes tend to have different types of function. Genes devoid of TEs have a role in development and in the regulation of transcription, while genes having a high proportion of TE insertions within their sequence and in their vicinity tend to be involved in metabolism and transport functions. This tendency was observed for all the TE families investigated (DNA transposons, LTR retrotransposons, LINEs, and SINEs). Consequently, we can extrapolate the findings of a previous study of *Alu* elements and human chromosomes 21 and 22, which had shown that *Alu*-rich genes and *Alu*-poor genes have different functions (Grover et al. 2003), to the whole set of TE families and indeed to the entire human genome. Similarly, genes with no TE in their introns were usually involved in morphogenesis and development as well as in transcription, and displayed extremely well-conserved intronic regions (Sironi et al. 2006). We show the same tendency in our present work in which only complete TEs are considered. This strongly suggests that the differences in gene function are likely to be related to the presence or absence of complete TE sequences.

Despite the high proportion of TEs in mammalian genomes, some regions, as much as 100 kb in length, are devoid of TEs (Simons et al. 2006). These regions, which are known as TFRs, have been also identified in amphibian and fish genomes (Simons et al. 2007). Most of these regions are associated with genes involved in important functions such as the regulation of transcription and development. The existence of such genomic regions could be due to strong evolutionary selection preventing their interruption by

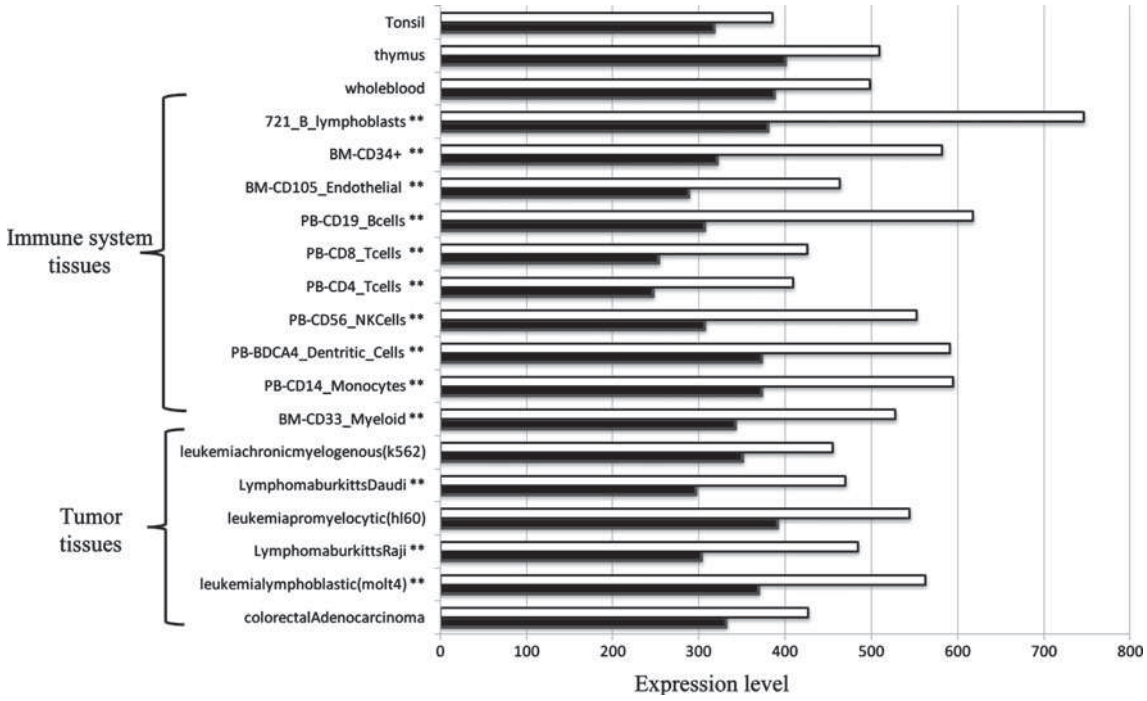

Fig. 4 Variation of the expression level of TE-free (*black bars*) and TE-rich genes (*white bars*) among 19 tissues considering 2-kb flanking regions. The *asterisks* indicate significantly differences between the two categories of genes

TE-derived sequences. The maintenance of retroelement-free regions in the human genome has indeed been shown to be due to selection acting against transcriptional interference from TEs, i.e., against the transcriptional activity of TEs that could interfere with the transcriptional activity of the genes in their vicinity (Mourier and Willerslev 2008). Selection would thus tend to eliminate any TEs inserted in the vicinity of genes with important functions in order to prevent potential interference. About 14% (875 of 6,185) of our TE-free genes in the 2-kb flanking regions were included in these TFR regions. This low proportion could be attributable to the fact that we only considered complete TEs, but these TE-free genes all had the same types of function as all the genes located in TFRs.

The difference in the amount of TEs in TE-free and TE-rich genes could be attributable to greater purifying selection acting on TE-free than on TE-rich genes. This hypothesis is supported by the significantly lower ω ratio of TE-free genes than of TE-rich genes, when human and mouse orthologs were compared. This suggests that stronger purifying selection tends to eliminate any complete TEs inserted within or in the vicinity of the TE-free genes. However, selection pressure is difficult to identify when orthologous genes of closely related species, such as human and chimpanzee, are compared. This could be due either to the fact that selection has not yet had time to act on all the genes that have undergone TE insertions because of the relatively recent divergence between human and chimpanzee (6 My), or that the species are too closely related to make it possible to identify such selective constraints. The same reasoning explains the values of the ω ratio in the multispecies comparisons. If selection pressure is the factor that explains the difference in TE density between genes, we would expect to find different ω ratios between these genes. However, the M0 model, which assumes that all branches of the tree have the same ω ratio, was always the model that fitted our data best, even when the orthologous genes had different TE densities. This suggests that these primate species are too closely related to make it possible to detect any difference in selection pressures, whereas this could be detected in the human/mouse comparison.

Non-coding regions can also be the target of selection pressure (Lowe et al. 2007). When we compared the sequence conservation of the flanking regions downstream and upstream of TE-free and TE-rich genes in human and chimpanzee, human and orangutan, and human and macaque, we observed that the mean percentage identity of regions surrounding TE-free genes was significantly higher than that of surrounding TE-rich genes for the 10-kb flanking regions. This finding implies that the regions surrounding TE-free genes are more highly conserved than those surrounding TE-rich genes. The same analysis for the 2-kb flanking regions did not always confirm higher sequence conservation around the TE-free genes. We can assume that promoters of genes, which are probably located in the 2-kb flanking regions, are subjected to strong purifying selection

due to their functional importance, regardless of the presence or absence of TEs. This difference became significant in the human and macaque comparison since the global sequence identity between the two genomes is lower [about 93.5% (The Rhesus Macaque Genome Sequencing and Analysis Consortium 2007)]. This showed that when the global sequence identity between genomes is high, as it is between human and chimpanzee, and between human and orangutan, it becomes difficult to detect any difference in sequence conservation depending on a factor such as TE density. We conclude that the regions flanking TE-free genes are better conserved, and thus likely to be exposed to a stronger purifying selection than the regions flanking TE-rich genes. These results confirmed previous observations suggesting that the difference in TE density between TE-free genes and TE-rich genes could be due to different selection pressures, but without quantifying them. We showed that these different selection pressures act not only on coding regions, as it is often assumed, but also on non-coding regions.

Gene expression can be regulated transcriptionally by promoters and *cis*-regulatory sequences, post-transcriptionally at the level of the untranslated regions (UTR) of mRNA, and at the higher-order level of the chromatin. It has been shown that almost 25% of human promoter regions contain TE-derived sequences, and that TEs provide about 2.5% of all human *cis*-regulatory sequences (Jordan et al. 2003). For example, SINE elements, which possess RNA polymerase III promoters, have been shown to promote the transcription of RNA polymerase II genes (Oliviero and Monaci 1988), and to present internal transcription factor binding sites within their sequences (Polak and Domany 2006) that can account for the control of the activity of genes located in their vicinity. However, it has been observed that the enrichment of a particular subfamily of SINEs, the *Alus*, in and around broadly expressed genes could only be a by-product of a preferential insertion bias of these elements near housekeeping genes (Urrutia et al. 2008). The presence of transcription factor sites has also been shown to be most abundant in ancient subfamilies of *Alus* than in young ones when tested on the elements from the chromosome 22 of human (Shankar et al. 2004). These observations can account for our observations as the most abundant subfamily of SINEs in our analysis correspond to *AluS* (Table 1). Analyses of the binding regions of seven mammalian transcription factors showed that five of them are associated with distinct families of TEs, like ERV1 being associated with TP153, indicating that TEs play an important role in expanding the repertoire of binding sites in mammals (Bourque et al. 2008). Moreover, *L1*, LTR retrotransposons, and DNA transposons displayed a higher affinity for nucleosome binding than *Alus*, which would result in a difference in the chromatin conformation and thus in expression (Huda et al. 2009). The different TE classes may thus have a different influence on genes. When all TE families are considered, human genes with TE-enriched promoters on average display greater and broader expression than gene promoters devoid of TEs (Huda et al. 2009). This effect of TEs on gene expression was also found in rodents, in which a correlation between gene expression and recent TE insertions was observed, indicating that these insertions do significantly alter gene expression patterns (Pereira et al. 2009). The comparison of genomic neighborhoods of human and chimpanzee genes has also shown that the expression of genes with a conserved genomic neighborhood was different from that of genes with TEs inserted in their vicinity (De et al. 2009). Since a TE sequence can affect gene regulation, selection pressure can be expected to be tighter for genes with crucial functions, such as development and regulation, making them more likely to be TE-free, while selection is expected to be more relaxed for genes with other, less vital functions, allowing TE insertions to be maintained in their vicinity. In our analysis, we found different levels of overall expression between TE-free and TE-rich genes among 79 human tissues, with the expression of TE-rich genes being greater than that of TE-free genes. It is striking that this difference in expression was found in three of the six tumor tissues, and in all 10 of the immune system tissues analyzed. These findings are consistent with the study of Lerat and Sémon (2007), in which levels of gene expression differed in the context of tumoral and normal conditions depending on the SINE neighborhood, suggesting that SINEs could be involved in gene deregulation under tumoral conditions, whereas they are silent in normal tissues. DNA methylation or other epigenetic mechanisms, which are known to regulate and even silence TE activity, could be associated with this gene deregulation in tumor tissues, because it is known that the loss of methylation in such tissues can affect TEs (Szpakowski et al. 2009). This is illustrated by the increase in the activity of retrotransposons that follows the loss of DNA methylation in tumoral conditions, which has been reported for human endogenous retroviruses in breast (Wang-Johanning et al. 2001) and ovarian cancers (Menendez et al. 2004; Wang-Johanning et al. 2007), and in leukemia cell lines (Patzke et al. 2002). Could this epigenetic gene deregulation postulated in tumor tissues be involved to explain the difference in gene expression according to the TE density observed in immune system tissues? This is a question that warrants to be investigated because it has been recently shown that the transcriptome is comparable between immune system cells and tumor tissues (Yang et al. 2008). We can thus hypothesize that the same kind of "deregulation" of the TEs via some epigenetic mechanisms could influence the silencing of the TEs near genes and change their expression both in tumor and immune tissues.

From this study, we can conclude that gene function is an important factor in determining the distribution of TEs in the human genome, and also in the genomes of the closely related species chimpanzee, orangutan, and macaque. Selection has the effect of eliminating TE insertions within and near genes with important functions. In addition, the insertion of TEs seems to be associated with a marked difference in gene expression in human tumor and immune system tissues. It would therefore be of great interest to analyze other primate species and indeed more widely phylogenetically divergent species, such as the mouse, to assess more precisely the effect of TEs on gene expression, especially under tumoral conditions.

Acknowledgments We would like to thank Christian Biémont for his comments and his critical reading of this manuscript, and Monika Ghosh for English correction.

References

Al-Shahrour F, Diaz-Uriarte R, Dopazo J (2004) FatiGO: a web tool for finding significant associations of gene ontology terms with group of genes. Bioinformatics 20:578–580

Al-Shahrour F, Minguez P, Vaquerizas JM, Conde L, Dopazo J (2005) BABELOMICS: a suite of web tools for functional annotation and analysis of groups of genes in high-throughput experiments. Nucleic Acid Res 33:460–464

Biemont C, Vieira C (2006) Junk DNA as an evolutionary force. Nature 443:521–524

Bourque G, Leong B, Vega VB, Chen X, Lee YL, Srinivasan KG, Chew JL, Ruan Y, Wei CL, Ng HH, Liu ET (2008) Evolution of the mammalian transcription factor binding repertoire via transposable elements. Genome Res 18:1752–1762

Burwinkel B, Kilimann MW (1998) Unequal homologous recombination between *LINE-1* elements as a mutational mechanism in human genetic disease. J Mol Biol 277:513–517

Chen JM, Stenson PD, Cooper DN, Ferec C (2005) A systematic analysis of LINE-1 endonuclease-dependent retranspositional events causing human genetic disease. Hum Genet 117:411–427

De S, Teichmann SA, Babu MM (2009) The impact of genomic neighborhood on the evolution of human and chimpanzee transcriptome. Genome Res 19:785–794

Deininger PL, Batzer MA (1999) *Alu* repeats and human disease. Mol Genet Metab 67:183–193

Edgar RC (2004) MUSCLE: multiple alignment with high accuracy and high throughput. Nucleic Acid Res 32:1792–1797

Felsenstein J (1989) PHYLIP—phylogeny interference package. Cladistics 5:164–166

Finnegan DJ (1989) Eukaryotic transposable elements and genome evolution. Trends Genet 5:103–107

Fitch DHA, Bailey WJ, Tagle DA, Goodman M, Sieu L, Slightom JL (1991) Duplication of the γ-globin gene mediated by *L1* long interspersed repetitive elements in a early ancestor of simian primates. Proc Nat Acad Sci USA 88:7396–7400

Grover D, Majumder PP, Rao CB, Brahmachari SK, Mukerji M (2003) Nonrandom distribution of *Alu* elements in genes of various functional categories: insight from analysis of human chromosomes 21 and 22. Mol Biol Evol 20:1420–1424

Han JS, Szak ST, Boeke JD (2004) Transcriptional disruption by the *L1* retrotransposon and implications for mammalian transcriptomes. Nature 429:268–274

Huda A, Marino-Ramirez L, Landsman D, Jordan IK (2009) Repetitive DNA elements, nucleosome binding and human gene expression. Gene 436:12–22

Jordan IK, Rogozin IB, Glazko GV, Koonin EV (2003) Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends Genet 19:68–72

Kazazian HH (2004) Mobile elements: drivers of genome evolution. Science 303:1626–1632

Kazazian HH, Moran JV (1998) The impact of *L1* retrotransposons on the human genome. Nat Genet 19:19–24

Kazazian HH, Wong C, Youssoufian H, Scott AF, Phillips DG, Antonarakis SE (1988) Haemophilia A resulting from *de novo* insertion of *L1* sequences represents a novel mechanism for mutation in man. Nature 332:164–166

Kim TH, Barrera LO, Zheng M, Qu C, Singer MA, Richmond TA, Wu Y, Green RD, Ren B (2005) A high-resolution map of active promoters in the human genome. Nature 436:876–880

Lerat E, Sémon M (2007) Influence of the transposable element neighborhood on human gene expression in normal and tumor tissues. Gene 396:303–311

Liu D, Bischerour J, Siddique A, Buisine N, Bigot Y, Chalmers R (2007) The human SETMAR protein preserves most of the activities of the ancestral *Hmar1* transposase. Mol Cell Biol 27:1125–1132

Lowe CB, Bejerano G, Haussler D (2007) Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. Proc Nat Acad Sci USA 104:8005–8010

Mariño-Ramírez L, Lewis KC, Landsman D, Jordan IK (2005) Transposable elements donate lineage-specific regulatory sequences to host genomes. Cytogenet Genome Res 110:333–341

Menendez L, Benigno BB, McDonald JF (2004) *L1* and *HERV-W* retrotransposons are hypomethylated in human ovarian carcinomas. Mol Cancer 3:12

Mourier T, Willerslev E (2008) Does selection against transcriptional interference shape retroelement-free regions in mammalian genomes? PLoS ONE 3:e3760

Myerowitz R, Hogikyan ND (1987) A deletion involving *Alu* sequences in the β-hexosaminidase α-chain gene of French Canadians with Tay-Sachs disease. J Biol Chem 262:15396–15399

Oliviero S, Monaci P (1988) RNA polymerase III promoter elements enhance transcription of RNA polymerase II genes. Nucleic Acids Res 16:1285–1293

Patzke S, Lindeskog M, Munthe M, Aasheim HC (2002) Characterization of a novel human endogenous retrovirus, *HERV-H/F*, expressed in human leukemia cell lines. Virology 303:164–173

Pereira V, Enard D, Eyre-Walker A (2009) The effect of transposable element insertions on gene expression evolution in rodents. PLoS ONE 4:e4321

Perepelitsa-Belancio V, Deininger PL (2003) RNA truncation by premature polyadenylation attenuates human mobile element activity. Nat Genet 35:363–366

Polak P, Domany E (2006) *Alu* elements may contain binding sites for transcription factors and may play a role in regulation and developmental processes. BMC Genomics 7:133–148

Roth DB, Craig NL (1998) VDJ recombination: a transposase goes to work. Cell 94:411–414

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, Minx P, Reily AD, Courtney L, Kruchowski SS, Tomlinson C, Strong C, Delehaunty K, Fronick C, Courtney B, Rock SM, Belter E, Du F, Kim K, Abbott RM, Cotton M, Levy A, Marchetto P, Ochoa K, Jackson SM, Gillam B, Chen W, Yan L, Higginbotham J, Cardenas M, Waligorski J, Applebaum E, Phelps L, Falcone J, Kanchi K,

Thane T, Scimone A, Thane N, Henke J, Wang T, Ruppert J, Shah N, Rotter K, Hodges J, Ingenthron E, Cordes M, Kohlberg S, Sgro J, Delgado B, Mead K, Chinwalla A, Leonard S, Crouse K, Collura K, Kudrna D, Currie J, He R, Angelova A, Rajasekar S, Mueller T, Lomeli R, Scara G, Ko A, Delaney K, Wissotski M, Lopez G, Campos D, Braidotti M, Ashley E, Golser W, Kim H, Lee S, Lin J, Dujmic Z, Kim W, Talag J, Zuccolo A, Fan C, Sebastian A, Kramer M, Spiegel L, Nascimento L, Zutavern T, Miller B, Ambroise C, Muller S, Spooner W, Narechania A, Ren L, Wei S, Kumari S, Faga B, Levy MJ, McMahan L, Van Buren P, Vaughn MW, Ying K, Yeh CT, Emrich SJ, Jia Y, Kalyanaraman A, Hsia AP, Barbazuk WB, Baucom RS, Brutnell TP, Carpita NC, Chaparro C, Chia JM, Deragon JM, Estill JC, Fu Y, Jeddeloh JA, Han Y, Lee H, Li P, Lisch DR, Liu S, Liu Z, Nagel DH, McCann MC, SanMiguel P, Myers AM, Nettleton D, Nguyen J, Penning BW, Ponnala L, Schneider KL, Schwartz DC, Sharma A, Soderlund C, Springer NM, Sun Q, Wang H, Waterman M, Westerman R, Wolfgruber TK, Yang L, Yu Y, Zhang L, Zhou S, Zhu Q, Bennetzen JL, Dawe RK, Jiang J, Jiang N, Presting GG, Wessler SR, Aluru S, Martienssen RA, Clifton SW, McCombie WR, Wing RA, Wilson RK (2009) The B73 maize genome: complexity, diversity, and dynamics. Science 326:1112–1115

Shankar R, Grover D, Brahmachari SK, Mukerji M (2004) Evolution and distribution of RNA polymerase II regulatory sites from RNA polymerase III dependant mobile Alu elements. BMC Evol Biol 4:37

Simons C, Pheasant M, Makunin IV, Mattick JS (2006) Transposon-free regions in mammalian genomes. Genome Res 16:164–172

Simons C, Pheasant M, Makunin IV, Mattick JS (2007) Maintenance of transposon-free regions throughout vertebrate evolution. BMC Genomics 8:470

Sironi M, Menozzi G, Comi GP, Cereda M, Cagliani R, Bresolin N, Pozzoli U (2006) Gene function and expression level influence the insertion/fixation dynamics of distinct transposon families in mammalian introns. Genome Biol 7:R120

Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, Kasprzyk A (2009) BioMart—biological queries made easy. BMC Genomics 14:10–22

Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Nat Acad Sci USA 101:6062–6067

Szpakowski S, Sun X, Lage JM, Dyer A, Rubinstein J, Kowalski D, Sasaki C, Costa J, Lizardi PM (2009) Loss of epigenetic silencing in tumors preferentially affects primate-specific retroelements. Gene 448:151–167

The Chimpanzee Genome Sequencing and Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437:69–87

The Drosophila 12 Genomes Consortium (2007) Evolution of genes and genomes on the drosophila phylogeny. Nature 450:203–218

The Gene Ontology Consortium (2000) Gene ontology: tool for the unification of biology. Nat Genet 25:25–29

The Gene Ontology Consortium (2001) Creating the gene ontology resource: design and implementation. Genome Res 11: 1425–1433

The International Human Genome Sequencing and Analysis Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921

The Mouse Genome Sequencing and Analysis Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562

The Rhesus Macaque Genome Sequencing and Analysis Consortium (2007) Evolutionary and biomedical insights from the rhesus macaque genome. Science 316:222–234

Urrutia AO, Ocaña LB, Hurst LD (2008) Do *Alu* repeats drive the evolution of the primate transcriptome? Genome Biol 9:R25

Wang-Johanning F, Frost AR, Johanning GL, Khazaeli MB, LoBuglio AF, Shaw DR, Strong TV (2001) Expression of human endogenous retrovirus k envelope transcripts in human breast cancer. Clin Cancer Res 7:1553–1560

Wang-Johanning F, Liu J, Rycaj K, Huang M, Tsai K, Rosen DG, Chen DT, Lu DW, Barnhart KF, Johanning GL (2007) Expression of multiple human endogenous retrovirus surface envelope proteins in ovarian cancer. Int J Cancer 120:81–90

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, Paux E, SanMiguel P, Schulman AH (2007) A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8:973–982

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586–1591

Yang X, Sun X, Xie J, Lu Z (2008) Comparability of gene expression in human blood, immune and carcinoma cells. Appl Math Comput 205:178–184

Bibliographie

[Abrahamsen *et al.*, 2004] Abrahamsen, M. S., Templeton, T. J., Enomoto, S. et *al.* (2004). Complete genome sequence of the apicomplexan, *Cryptosporidium parvum*. *Science*, 304:441–445.

[Al-Shahrour *et al.*, 2004] Al-Shahrour, F., Díaz-Uriarte, R. et Dopazo, J. (2004). FatiGO : a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, 20:578–580.

[Al-Shahrour *et al.*, 2007] Al-Shahrour, F., Minguez, P., Tárraga, J. et *al.* (2007). FatiGO + : a functional profiling tool for genomic data. Integration of functional annotation, regulatory motifs and interaction data with microarray experiments. *Nucleic Acids Res*, 35:W91–W96.

[American Association for Cancer Research Human Epigenome Task Force, 2008] American Association for Cancer Research Human Epigenome Task Force (2008). Moving AHEAD with an international human epigenome project. *Nature*, 454:711–715.

[Aouba *et al.*, 2007] Aouba, A., péquignot, F., Toullec, A. L. et *al.* (2007). Les causes médicales de décès en France en 2004 et leur évolution 1980-2004. *Bulletin Epidémiologique Hebdomadaire*, 35-36:308–314.

[Aravin *et al.*, 2001] Aravin, A. A., Naumova, N. M., Tulin, A. V. et *al.* (2001). Double-stranded RNA-mediated silencing of genomic tandem repeats and transposable elements in the *Drosophila melanogaster* germline. *Curr Biol*, 11:1017–1027.

[Aravin *et al.*, 2007] Aravin, A. A., Sachidanandam, R., Girard, A. et *al.* (2007). Developmentally regulated piRNA clusters implicate MILI in transposon control. *Science*, 316:744–747.

[Ashburner *et al.*, 2000] Ashburner, M., Ball, C. A., Blake, J. A. et *al.* (2000). Gene ontology : tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25:25–29.

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

[Bachman *et al.*, 2003] Bachman, K. E., Park, B. H., Rhee, I. et *al.* (2003). Histone modifications and silencing prior to DNA methylation of a tumor suppressor gene. *Cancer Cell*, 3:89–95.

[Bailey *et al.*, 2000] Bailey, J. A., Carrel, L., Chakravarti, A. et *al.* (2000). Molecular evidence for a relationship between LINE-1 elements and X chromosome inactivation : the Lyon repeat hypothesis. *Proc Natl Acad Sci U S A*, 97:6634–6639.

[Bailey *et al.*, 2003] Bailey, J. A., Liu, G. et Eichler, E. E. (2003). An *Alu* transposition model for the origin and expansion of human segmental duplications. *Am J Hum Genet*, 73:823–834.

[Bannister et Kouzarides, 2005] Bannister, A. J. et Kouzarides, T. (2005). Reversing histone methylation. *Nature*, 436:1103–1106.

[Barbacid, 1987] Barbacid, M. (1987). ras genes. *Annu Rev Biochem*, 56:779–827.

[Barski *et al.*, 2007] Barski, A., Cuddapah, S., Cui, K. et *al.* (2007). High-resolution profiling of histone methylations in the human genome. *Cell*, 129:823–837.

[Bartolomé *et al.*, 2002] Bartolomé, C., Maside, X. et Charlesworth, B. (2002). On the abundance and distribution of transposable elements in the genome of *Drosophila melanogaster*. *Mol Biol Evol*, 19:926–937.

[Bejerano *et al.*, 2006] Bejerano, G., Lowe, C. B., Ahituv, N. et *al.* (2006). A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature*, 441:87–90.

[Belancio *et al.*, 2008] Belancio, V. P., Hedges, D. J. et Deininger, P. (2008). Mammalian non-LTR retrotransposons : for better or worse, in sickness and in health. *Genome Res*, 18:343–358.

[Belancio *et al.*, 2010] Belancio, V. P., Roy-Engel, A. M. et Deininger, P. L. (2010). All y'all need to know 'bout retroelements in cancer. *Semin Cancer Biol*, 20:200–210.

[Bennetzen, 2000] Bennetzen, J. L. (2000). Transposable element contributions to plant gene and genome evolution. *Plant Mol Biol*, 42:251–269.

[Bernstein *et al.*, 2007] Bernstein, B. E., Meissner, A. et Lander, E. S. (2007). The mammalian epigenome. *Cell*, 128:669–681.

[Bielawski et Yang, 2001] Bielawski, J. P. et Yang, Z. (2001). Positive and negative selection in the DAZ gene family. *Mol Biol Evol*, 18:523–529.

[Bingham *et al.*, 1982] Bingham, P. M., Kidwell, M. G. et Rubin, G. M. (1982). The molecular basis of P-M hybrid dysgenesis : the role of the P element, a P-strain-specific transposon family. *Cell*, 29:995–1004.

[Biémont, 2010a] Biémont, C. (2010a). A brief history of the status of transposable elements : from junk DNA to major players in evolution. *Genetics*, 186:1085–1093.

[Biémont, 2010b] Biémont, C. (2010b). From genotype to phenotype. What do epigenetics and epigenomics tell us ? *Heredity*, 105:1–3.

[Biémont et Vieira, 2006] Biémont, C. et Vieira, C. (2006). Genetics : junk DNA as an evolutionary force. *Nature*, 443:521–524.

[Bollati et Baccarelli, 2010] Bollati, V. et Baccarelli, A. (2010). Environmental epigenetics. *Heredity*, 105:105–112.

[Bourque, 2009] Bourque, G. (2009). Transposable elements in gene regulation and in the evolution of vertebrate genomes. *Curr Opin Genet Dev*, 19:607–612.

[Bourque *et al.*, 2008] Bourque, G., Leong, B., Vega, V. B. et *al.* (2008). Evolution of the mammalian transcription factor binding repertoire *via* transposable elements. *Genome Res*, 18:1752–1762.

[Bowen et Jordan, 2002] Bowen, N. J. et Jordan, I. K. (2002). Transposable elements and the evolution of eukaryotic complexity. *Curr Issues Mol Biol*, 4:65–76.

[Brayton *et al.*, 2007] Brayton, K. A., Lau, A. O. T., Herndon, D. R. et *al.* (2007). Genome sequence of *Babesia bovis* and comparative analysis of apicomplexan hemoprotozoa. *PLoS Pathog*, 3:1401–1413.

[Britten, 2006] Britten, R. (2006). Transposable elements have contributed to thousands of human proteins. *Proc Natl Acad Sci U S A*, 103:1798–1803.

[Brookfield, 2005] Brookfield, J. F. Y. (2005). The ecology of the genome - mobile DNA elements and their hosts. *Nat Rev Genet*, 6:128–136.

[Burki et Kaessmann, 2004] Burki, F. et Kaessmann, H. (2004). Birth and adaptive evolution of a hominoid gene that supports high neurotransmitter flux. *Nat Genet*, 36:1061–1063.

[Burwinkel et Kilimann, 1998] Burwinkel, B. et Kilimann, M. W. (1998). Unequal homologous recombination between LINE-1 elements as a mutational mechanism in human genetic disease. *J Mol Biol*, 277:513–517.

[Böhne *et al.*, 2008] Böhne, A., Brunet, F., Galiana-Arnoux, D. et *al.* (2008). Transposable elements as drivers of genomic and biological diversity in vertebrates. *Chromosome Res*, 16:203–215.

[Capy *et al.*, 1997] Capy, P., Langin, T., Higuet, D. et *al.* (1997). Do the integrases of LTR-retrotransposons and class II element transposases have a common ancestor ? *Genetica*, 100:63–72.

[Carlton *et al.*, 2002] Carlton, J. M., Angiuoli, S. V., Suh, B. B. et *al.* (2002). Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*. *Nature*, 419:512–519.

[Carroll, 2005] Carroll, S. B. (2005). Evolution at two levels : on genes and form. *PLoS Biol*, 3:e245.

[Castillo-Davis *et al.*, 2002] Castillo-Davis, C. I., Mekhedov, S. L., Hartl, D. L. et *al.* (2002). Selection for short introns in highly expressed genes. *Nat Genet*, 31:415–418.

[Cavalier-Smith et Beaton, 1999] Cavalier-Smith, T. et Beaton, M. J. (1999). The skeletal function of non-genic nuclear DNA : new evidence from ancient cell chimaeras. *Genetica*, 106:3–13.

[Chimpanzee Sequencing and Analysis Consortium, 2005] Chimpanzee Sequencing and Analysis Consortium (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 437:69–87.

[Claes *et al.*, 2010] Claes, B., Buysschaert, I. et Lambrechts, D. (2010). Pharmaco-epigenomics : discovering therapeutic approaches and biomarkers for cancer therapy. *Heredity*, 105:152–160.

[Cutter *et al.*, 2005] Cutter, A. D., Good, J. M., Pappas, C. T. et *al.* (2005). Transposable element orientation bias in the *Drosophila melanogaster* genome. *J Mol Evol*, 61:733–741.

[Côté *et al.*, 1994] Côté, J., Quinn, J., Workman, J. L. et *al.* (1994). Stimulation of GAL4 derivative binding to nucleosomal DNA by the yeast SWI/SNF complex. *Science*, 265:53–60.

[Dasilva *et al.*, 2002] Dasilva, C., Hadji, H., Ozouf-Costaz, C. et *al.* (2002). Remarkable compartmentalization of transposable elements and pseudogenes in the heterochromatin of the *Tetraodon nigroviridis* genome. *Proc Natl Acad Sci U S A*, 99:13636–13641.

[Dawkins, 2006] Dawkins, R. (1976/2006). *The Selfish Gene*. New York City : Oxford University Press.

[De *et al.*, 2009] De, S., Teichmann, S. A. et Babu, M. M. (2009). The impact of genomic neighborhood on the evolution of human and chimpanzee transcriptome. *Genome Res*, 19:785–794.

[Deininger et Batzer, 1999] Deininger, P. L. et Batzer, M. A. (1999). *Alu* repeats and human disease. *Mol Genet Metab*, 67:183–193.

[Deininger *et al.*, 2003] Deininger, P. L., Moran, J. V., Batzer, M. A. et *al.* (2003). Mobile elements and mammalian genome evolution. *Curr Opin Genet Dev*, 13:651–658.

[Deltour *et al.*, 2005] Deltour, S., Chopin, V. et Leprince, D. (2005). Modifications épigénétiques et cancer. *Médecine sciences*, 21:405–411.

[Dimitri *et al.*, 2003] Dimitri, P., Junakovic, N. et Arcà, B. (2003). Colonization of heterochromatic genes by transposable elements in Drosophila. *Mol Biol Evol*, 20:503–512.

[Doolittle et Sapienza, 1980] Doolittle, W. F. et Sapienza, C. (1980). Selfish genes, the phenotype paradigm and genome evolution. *Nature*, 284:601–603.

[Drosophila 12 Genomes Consortium , 2007] Drosophila 12 Genomes Consortium (2007). Evolution of genes and genomes on the Drosophila phylogeny. *Nature*, 450:203–218.

[Duret *et al.*, 2000] Duret, L., Marais, G. et Biémont, C. (2000). Transposons but not retrotransposons are located preferentially in regions of high recombination rate in *Caenorhabditis elegans*. *Genetics*, 156:1661–1669.

[Eckhardt *et al.*, 2006] Eckhardt, F., Lewin, J., Cortese, R. et *al.* (2006). DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat Genet*, 38:1378–1385.

[Edgar, 2004] Edgar, R. C. (2004). MUSCLE : a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5:113.

[Espada *et al.*, 2004] Espada, J., Ballestar, E., Fraga, M. F. et *al.* (2004). Human DNA methyltransferase 1 is required for maintenance of the histone H3 modification pattern. *J Biol Chem*, 279:37175–37184.

[Esteller, 2002] Esteller, M. (2002). CpG island hypermethylation and tumor suppressor genes : a booming present, a brighter future. *Oncogene*, 21:5427–5440.

[Esteller, 2005] Esteller, M. (2005). Aberrant DNA methylation as a cancer-inducing mechanism. *Annu Rev Pharmacol Toxicol*, 45:629–656.

[Esteller, 2007] Esteller, M. (2007). Cancer epigenomics : DNA methylomes and histone-modification maps. *Nat Rev Genet*, 8:286–298.

[Esteller *et al.*, 2001] Esteller, M., Corn, P. G., Baylin, S. B. et *al.* (2001). A gene hypermethylation profile of human cancer. *Cancer Res*, 61:3225–3229.

[Evgen'ev *et al.*, 2000] Evgen'ev, M. B., Zelentsova, H., Poluectova, H. et *al.* (2000). Mobile elements and chromosomal evolution in the virilis group of Drosophila. *Proc Natl Acad Sci U S A*, 97:11337–11342.

[Fagegaltier *et al.*, 2009] Fagegaltier, D., Bougé, A.-L., Berry, B. et *al.* (2009). The endogenous siRNA pathway is involved in heterochromatin formation in Drosophila. *Proc Natl Acad Sci U S A*, 106:21258–21263.

[Fedoroff, 1999] Fedoroff, N. V. (1999). Transposable elements as a molecular evolutionary force. *Ann N Y Acad Sci*, 870:251–264.

[Feinberg et Tycko, 2004] Feinberg, A. P. et Tycko, B. (2004). The history of cancer epigenetics. *Nat Rev Cancer*, 4:143–153.

[Feinberg et Vogelstein, 1983]
Feinberg, A. P. et Vogelstein, B. (1983). Hypomethylation distinguishes genes of some human cancers from their normal counterparts. *Nature*, 301:89–92.

[Felsenstein, 1989] Felsenstein, J. (1989). PHYLIP-phylogeny inference package (version 3.2). *Cladistics*, 5:164–166.

[Feng *et al.*, 2002] Feng, Q., Zhang, Y., Hao, P. et *al.* (2002). Sequence and analysis of rice chromosome 4. *Nature*, 420:316–320.

[Feschotte et Pritham, 2007] Feschotte, C. et Pritham, E. J. (2007). DNA transposons and the evolution of eukaryotic genomes. *Annu Rev Genet*, 41:331–368.

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

[Flicek *et al.*, 2008] Flicek, P., Aken, B. L., Beal, K. et *al.* (2008). Ensembl 2008. *Nucleic Acids Res*, 36:D707–D714.

[Fraga *et al.*, 2005] Fraga, M. F., Ballestar, E., Villar-Garea, A. et *al.* (2005). Loss of acetylation at Lys16 and trimethylation at Lys20 of histone H4 is a common hallmark of human cancer. *Nat Genet*, 37(4):391–400.

[Fugmann, 2010] Fugmann, S. D. (2010). The origins of the Rag genes–from transposition to V(D)J recombination. *Semin Immunol*, 22:10–16.

[Galagan et Selker, 2004] Galagan, J. E. et Selker, E. U. (2004). RIP : the evolutionary cost of genome defense. *Trends Genet*, 20:417–423.

[Gardner *et al.*, 2005] Gardner, M. J., Bishop, R., Shah, T. et *al.* (2005). Genome sequence of *Theileria parva*, a bovine pathogen that transforms lymphocytes. *Science*, 309:134–137.

[Gardner *et al.*, 2002] Gardner, M. J., Shallom, S. J., Carlton, J. M. et *al.* (2002). Sequence of *Plasmodium falciparum* chromosomes 2, 10, 11 and 14. *Nature*, 419:531–534.

[Gene Ontology Consortium, 2001] Gene Ontology Consortium (2001). Creating the gene ontology resource : design and implementation. *Genome Res*, 11:1425–1433.

[Gene Ontology Consortium, 2010] Gene Ontology Consortium (2010). The Gene Ontology in 2010 : extensions and refinements. *Nucleic Acids Res*, 38:D331–D335.

[Ghildiyal et Zamore, 2009] Ghildiyal, M. et Zamore, P. D. (2009). Small silencing RNAs : an expanding universe. *Nat Rev Genet*, 10:94–108.

[Gibbs *et al.*, 2004] Gibbs, R. A., Weinstock, G. M., Metzker, M. L. et *al.* (2004). Genome sequence of the brown norway rat yields insights into mammalian evolution. *Nature*, 428:493–521.

[Glazko et Mushegian, 2010] Glazko, G. et Mushegian, A. (2010). Measuring gene expression divergence : the distance to keep. *Biol Direct*, 5:51.

[Goelz *et al.*, 1985] Goelz, S. E., Vogelstein, B., Hamilton, S. R. et *al.* (1985). Hypomethylation of DNA from benign and malignant human colon neoplasms. *Science*, 228:187–190.

[Gould et Vrba, 1982] Gould, S. et Vrba, E. (1982). Exaptation : a missing term in the science of form. *Paleobiology*, 8:4–15.

[Grover *et al.*, 2003] Grover, D., Majumder, P. P., Rao, C. B. et *al.* (2003). Nonrandom distribution of *Alu* elements in genes of various functional categories : insight from analysis of human chromosomes 21 and 22. *Mol Biol Evol*, 20:1420–1424.

[Grover *et al.*, 2004] Grover, D., Mukerji, M., Bhatnagar, P. et *al.* (2004). *Alu* repeat analysis in the complete human genome : trends and variations with respect to genomic composition. *Bioinformatics*, 20:813–817.

[Haider *et al.*, 2009] Haider, S., Ballester, B., Smedley, D. et *al.* (2009). BioMart Central Portal–unified access to biological data. *Nucleic Acids Res*, 37:W23–W27.

[Hanahan et Weinberg, 2000] Hanahan, D. et Weinberg, R. A. (2000). The hallmarks of cancer. *Cell*, 100:57–70.

[He *et al.*, 2011] He, T., Wang, Q., Feng, G. et *al.* (2011). Computational detection and functional analysis of human tissue-specific A-to-I RNA editing. *PLoS One*, 6:e18129.

[Hickey, 1982] Hickey, D. A. (1982). Selfish DNA : a sexually-transmitted nuclear parasite. *Genetics*, 101:519–531.

[Hill *et al.*, 2000] Hill, A. S., Foot, N. J., Chaplin, T. L. et *al.* (2000). The most frequent constitutional translocation in humans, the t(11 ;22)(q23 ;q11) is due to a highly specific *Alu*-mediated recombination. *Hum Mol Genet*, 9:1525–1532.

[Holliday, 1987] Holliday, R. (1987). The inheritance of epigenetic defects. *Science*, 238:163–170.

[Holliday, 1994] Holliday, R. (1994). Epigenetics : an overview. *Dev Genet*, 15:453–457.

[Holt *et al.*, 2002] Holt, R. A., Subramanian, G. M., Halpern, A. et *al.* (2002). The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science*, 298:129–149.

[Honeybee Genome Sequencing Consortium, 2006] Honeybee Genome Sequencing Consortium (2006). Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature*, 443:931–949.

[Hoskins *et al.*, 2002] Hoskins, R. A., Smith, C. D., Carlson, J. W. et *al.* (2002). Heterochromatic sequences in a Drosophila whole-genome shotgun assembly. *Genome Biol*, 3:RESEARCH0085.

[Hua-Van *et al.*, 2005] Hua-Van, A., Rouzic, A. L., Maisonhaute, C. et *al.* (2005). Abundance, distribution and dynamics of retrotransposable elements and transposons : similarities and differences. *Cytogenet Genome Res*, 110:426–440.

[Huda *et al.*, 2011] Huda, A., Bowen, N. J., Conley, A. B. et *al.* (2011). Epigenetic regulation of transposable element derived human gene promoters. *Gene*, 475:39–48.

[Huda *et al.*, 2009] Huda, A., Mariño-Ramírez, L., Landsman, D. et *al.* (2009). Repetitive DNA elements, nucleosome binding and human gene expression. *Gene*, 436:12–22.

[Huff *et al.*, 2010] Huff, J. T., Plocik, A. M., Guthrie, C. et *al.* (2010). Reciprocal intronic and exonic histone modification regions in humans. *Nat Struct Mol Biol*, 17:1495–1499.

[Isenberg, 1979] Isenberg, I. (1979). Histones. *Annu Rev Biochem*, 48:159–191.

[Jacob, 1977] Jacob, F. (1977). Evolution and tinkering. *Science*, 196:1161–1166.

[Ji *et al.*, 2008] Ji, H., Jiang, H., Ma, W. et *al.* (2008). An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat Biotechnol*, 26:1293–1300.

[Jones et Baylin, 2007] Jones, P. A. et Baylin, S. B. (2007). The epigenomics of cancer. *Cell*, 128:683–692.

[Jordan *et al.*, 2003] Jordan, I. K., Rogozin, I. B., Glazko, G. V. et *al.* (2003). Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends Genet*, 19:68–72.

[Jurka, 2004] Jurka, J. (2004). Evolutionary impact of human *Alu* repetitive elements. *Curr Opin Genet Dev*, 14:603–608.

[Jurka *et al.*, 2005] Jurka, J., Kapitonov, V. V., Pavlicek, A. et *al.* (2005). Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res*, 110:462–467.

[Kapitonov et Jurka, 2008] Kapitonov, V. V. et Jurka, J. (2008). A universal classification of eukaryotic transposable elements implemented in Repbase. *Nat Rev Genet*, 9:411–2 ; author reply 414.

[Katinka *et al.*, 2001] Katinka, M. D., Duprat, S., Cornillot, E. et *al.* (2001). Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. *Nature*, 414:450–453.

[Kawasaki *et al.*, 2005] Kawasaki, H., Taira, K. et Morris, K. V. (2005). siRNA induced transcriptional gene silencing in mammalian cells. *Cell Cycle*, 4:442–448.

[Kazazian, 2004] Kazazian, H. H. (2004). Mobile elements : drivers of genome evolution. *Science*, 303:1626–1632.

[Kharchenko *et al.*, 2008] Kharchenko, P. V., Tolstorukov, M. Y. et Park, P. J. (2008). Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotechnol*, 26:1351–1359.

[Kidwell, 2002] Kidwell, M. G. (2002). Transposable elements and the evolution of genome size in eukaryotes. *Genetica*, 115:49–63.

[Kim *et al.*, 2005] Kim, T. H., Barrera, L. O., Zheng, M. et *al.* (2005). A high-resolution map of active promoters in the human genome. *Nature*, 436:876–880.

[Kolasinska-Zwierz *et al.*, 2009] Kolasinska-Zwierz, P., Down, T., Latorre, I. et *al.* (2009). Differential chromatin marking of introns and expressed exons by H3K36me3. *Nat Genet*, 41:376–381.

[Kondo et Issa, 2003] Kondo, Y. et Issa, J.-P. J. (2003). Enrichment for histone H3 lysine 9 methylation at *Alu* repeats in human cells. *J Biol Chem*, 278:27658–27662.

[Kourmouli *et al.*, 2004] Kourmouli, N., Jeppesen, P., Mahadevhaiah, S. et *al.* (2004). Heterochromatin and tri-methylated lysine 20 of histone H4 in animals. *J Cell Sci*, 117:2491–2501.

[Kouzarides, 2007] Kouzarides, T. (2007). Chromatin modifications and their function. *Cell*, 128:693–705.

[Kuo et Allis, 1998] Kuo, M. H. et Allis, C. D. (1998). Roles of histone acetyltransferases and deacetylases in gene regulation. *Bioessays*, 20:615–626.

[Lander *et al.*, 2001] Lander, E. S., Linton, L. M., Birren, B. et *al.* (2001). Initial sequencing and analysis of the human genome. *Nature*, 409:860–921.

[Lehnert *et al.*, 2009] Lehnert, S., Loo, P. V., Thilakarathne, P. J. et *al.* (2009). Evidence for co-evolution between human microRNAs and *Alu*-repeats. *PLoS One*, 4:e4456.

[Lerat et Sémon, 2007] Lerat, E. et Sémon, M. (2007). Influence of the transposable element neighborhood on human gene expression in normal and tumor tissues. *Gene*, 396:303–311.

[Linhart *et al.*, 2007] Linhart, H. G., Lin, H., Yamada, Y. et *al.* (2007). Dnmt3b promotes tumorigenesis in vivo by gene-specific de novo methylation and transcriptional silencing. *Genes Dev*, 21:3110–3122.

[Lippman *et al.*, 2004] Lippman, Z., Gendrel, A.-V., Black, M. et *al.* (2004). Role of transposable elements in heterochromatin and epigenetic control. *Nature*, 430:471–476.

[Lister *et al.*, 2009] Lister, R., Pelizzola, M., Dowen, R. H. et *al.* (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462:315–322.

[Lowe *et al.*, 2007] Lowe, C. B., Bejerano, G. et Haussler, D. (2007). Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. *Proc Natl Acad Sci U S A*, 104:8005–8010.

[Luco *et al.*, 2010] Luco, R. F., Pan, Q., Tominaga, K. et *al.* (2010). Regulation of alternative splicing by histone modifications. *Science*, 327:996–1000.

[Luger, 2003] Luger, K. (2003). Structure and dynamic behavior of nucleosomes. *Curr Opin Genet Dev*, 13:127–135.

[Luger *et al.*, 1997] Luger, K., Mäder, A. W., Richmond, R. K. et *al.* (1997). Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature*, 389:251–260.

[MacQueen, 1967] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability.*

[Margueron *et al.*, 2005] Margueron, R., Trojer, P. et Reinberg, D. (2005). The key to development : interpreting the histone code ? *Curr Opin Genet Dev*, 15:163–176.

[Mariño-Ramírez et Jordan, 2006] Mariño-Ramírez, L. et Jordan, I. K. (2006). Transposable element derived DNaseI-hypersensitive sites in the human genome. *Biol Direct*, 1:20.

[Martens *et al.*, 2005] Martens, J. H. A., O'Sullivan, R. J., Braunschweig, U. et *al.* (2005). The profile of repeat-associated histone lysine methylation states in the mouse epigenome. *EMBO J*, 24:800–812.

[Matzke *et al.*, 2009] Matzke, M., Kanno, T., Daxinger, L. et *al.* (2009). RNA-mediated chromatin-based silencing in plants. *Curr Opin Cell Biol*, 21:367–376.

[McClintock, 1950] McClintock, B. (1950). The origin and behavior of mutable loci in maize. *Proc Natl Acad Sci U S A*, 36:344–355.

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

[Medina *et al.*, 2010] Medina, I., Carbonell, J., Pulido, L. et *al.* (2010). Babelomics : an integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling. *Nucleic Acids Res*, 38:W210–W213.

[Medstrand *et al.*, 2002] Medstrand, P., van de Lagemaat, L. N. et Mager, D. L. (2002). Retroelement distributions in the human genome : variations associated with age and proximity to genes. *Genome Res*, 12:1483–1495.

[Menendez *et al.*, 2004] Menendez, L., Benigno, B. B. et McDonald, J. F. (2004). L1 and HERV-W retrotransposons are hypomethylated in human ovarian carcinomas. *Mol Cancer*, 3:12.

[Miki *et al.*, 1996] Miki, Y., Katagiri, T., Kasumi, F., Yoshimoto, T. et *al.* (1996). Mutation analysis in the BRCA2 gene in primary breast cancers. *Nat Genet*, 13:245–247.

[Miki *et al.*, 1992] Miki, Y., Nishisho, I., Horii, A. et *al.* (1992). Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer. *Cancer Res*, 52:643–645.

[Mikkelsen *et al.*, 2007] Mikkelsen, T. S., Ku, M., Jaffe, D. B. et *al.* (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, 448:553–560.

[Mills *et al.*, 2007] Mills, R. E., Bennett, E. A., Iskow, R. C. et *al.* (2007). Which transposable elements are active in the human genome ? *Trends Genet*, 23:183–191.

[Misumi *et al.*, 2005] Misumi, O., Matsuzaki, M., Nozaki, H. et *al.* (2005). *Cyanidioschyzon merolae* genome. A tool for facilitating comparable studies on organelle biogenesis in photosynthetic eukaryotes. *Plant Physiol*, 137:567–585.

[Moran *et al.*, 1999] Moran, J. V., DeBerardinis, R. J. et Kazazian, H. H. (1999). Exon shuffling by L1 retrotransposition. *Science*, 283:1530–1534.

[Mourier et Willerslev, 2008] Mourier, T. et Willerslev, E. (2008). Does selection against transcriptional interference shape retroelement-free regions in mammalian genomes ? *PLoS One*, 3:e3760.

[Mouse Genome Sequencing Consortium, 2002] Mouse Genome Sequencing Consortium (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420:520–562.

[Muers, 2009] Muers, M. (2009). A peak for exons. *Nat Rev Genet*, 10:1038.

[Muntean et Hess, 2009] Muntean, A. G. et Hess, J. L. (2009). Epigenetic dysregulation in cancer. *Am J Pathol*, 175:1353–1361.

[Muotri *et al.*, 2007] Muotri, A. R., Marchetto, M. C. N., Coufal, N. G. et *al.* (2007). The necessary junk : new functions for transposable elements. *Hum Mol Genet*, 16:R159–R167.

[Nekrutenko et Li, 2001] Nekrutenko, A. et Li, W. H. (2001). Transposable elements are found in a large number of human protein-coding genes. *Trends Genet*, 17:619–621.

[Nene *et al.*, 2007] Nene, V., Wortman, J. R., Lawson, D. et *al.* (2007). Genome sequence of *Aedes aegypti*, a major arbovirus vector. *Science*, 316:1718–1723.

[Nishikura, 2006] Nishikura, K. (2006). Editor meets silencer : crosstalk between RNA editing and RNA interference. *Nat Rev Mol Cell Biol*, 7:919–931.

[Obbard *et al.*, 2009] Obbard, D. J., Gordon, K. H. J., Buck, A. H. et *al.* (2009). The evolution of RNAi as a defence against viruses and transposable elements. *Philos Trans R Soc Lond B Biol Sci*, 364:99–115.

[Oliver et Greene, 2009] Oliver, K. R. et Greene, W. K. (2009). Transposable elements : powerful facilitators of evolution. *Bioessays*, 31:703–714.

[Oliviero et Monaci, 1988] Oliviero, S. et Monaci, P. (1988). RNA polymerase III promoter elements enhance transcription of RNA polymerase II genes. *Nucleic Acids Res*, 16:1285–1293.

[Orgel et Crick, 1980] Orgel, L. E. et Crick, F. H. (1980). Selfish DNA : the ultimate parasite. *Nature*, 284:604–607.

[Orgel *et al.*, 1980] Orgel, L. E., Crick, F. H. et Sapienza, C. (1980). Selfish DNA. *Nature*, 288:645–646.

[Ostertag *et al.*, 2003] Ostertag, E. M., Goodier, J. L., Zhang, Y. et *al.* (2003). SVA elements are nonautonomous retrotransposons that cause disease in humans. *Am J Hum Genet*, 73:1444–1451.

[Pace et Feschotte, 2007] Pace, J. K. et Feschotte, C. (2007). The evolutionary history of human DNA transposons : evidence for intense activity in the primate lineage. *Genome Res*, 17:422–432.

[Pace *et al.*, 2008] Pace, J. K., Gilbert, C., Clark, M. S. et *al.* (2008). Repeated horizontal transfer of a DNA transposon in mammals and other tetrapods. *Proc Natl Acad Sci U S A*, 105:17023–17028.

[Pardue *et al.*, 2005] Pardue, M.-L., Rashkova, S., Casacuberta, E. et *al.* (2005). Two retrotransposons maintain telomeres in Drosophila. *Chromosome Res*, 13:443–453.

[Paslier et Bernot, 2001] Paslier, D. L. et Bernot, A. (2001). Le Projet Génome Humain : quinze ans d'efforts. *Médecine/sciences*, 17:294–8.

[Patzke *et al.*, 2002] Patzke, S., Lindeskog, M., Munthe, E. et *al.* (2002). Characterization of a novel human endogenous retrovirus, HERV-H/F, expressed in human leukemia cell lines. *Virology*, 303:164–173.

[Pauler *et al.*, 2009] Pauler, F. M., Sloane, M. A., Huang, R. et *al.* (2009). H3K27me3 forms BLOCs over silent genes and intergenic regions and specifies a histone banding pattern on a mouse autosomal chromosome. *Genome Res*, 19:221–233.

[Peng et Karpen, 2007] Peng, J. C. et Karpen, G. H. (2007). H3K9 methylation and RNA interference regulate nucleolar organization and repeated DNA stability. *Nat Cell Biol*, 9:25–35.

[Pereira *et al.*, 2009] Pereira, V., Enard, D. et Eyre-Walker, A. (2009). The effect of transposable element insertions on gene expression evolution in rodents. *PLoS One*, 4:e4321.

[Picard *et al.*, 1978] Picard, G., Bregliano, J. C., Bucheton, A. et *al.* (1978). Non-mendelian female sterility and hybrid dysgenesis in *Drosophila melanogaster*. *Genet Res*, 32:275–287.

[Piriyapongsa et Jordan, 2007] Piriyapongsa, J. et Jordan, I. K. (2007). A family of human microRNA genes from miniature inverted-repeat transposable elements. *PLoS One*, 2:e203.

[Piriyapongsa *et al.*, 2007a] Piriyapongsa, J., Mariño-Ramírez, L. et Jordan, I. K. (2007a). Origin and evolution of human microRNAs from transposable elements. *Genetics*, 176:1323–1337.

[Piriyapongsa *et al.*, 2007b] Piriyapongsa, J., Polavarapu, N., Borodovsky, M. et *al.* (2007b). Exonization of the LTR transposable elements in human genome. *BMC Genomics*, 8:291.

[Plasterk, 2002] Plasterk, R. H. A. (2002). RNA silencing : the genome's immune system. *Science*, 296:1263–1265.

[Polak et Domany, 2006] Polak, P. et Domany, E. (2006). *Alu* elements contain many binding sites for transcription factors and may play a role in regulation of developmental processes. *BMC Genomics*, 7:133.

[Pritham, 2009] Pritham, E. J. (2009). Transposable elements and factors influencing their success in eukaryotes. *J Hered*, 100:648–655.

[R Development Core Team, 2005] R Development Core Team (2005). *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[Raney *et al.*, 2011] Raney, B. J., Cline, M. S., Rosenbloom, K. R. et *al.* (2011). ENCODE whole-genome data in the UCSC genome browser (2011 update). *Nucleic Acids Res*, 39:D871–D875.

[Rhesus Macaque Genome Sequencing and Analysis Consortium , 2007] Rhesus Macaque Genome Sequencing and Analysis Consortium (2007). Evolutionary and biomedical insights from the rhesus macaque genome. *Science*, 316:222–234.

[Rice Chromosome 10 Sequencing Consortium, 2003] Rice Chromosome 10 Sequencing Consortium (2003). In-depth view of structure, activity, and evolution of rice chromosome 10. *Science*, 300:1566–1569.

[Rigal et Mathieu, 2011] Rigal, M. et Mathieu, O. (2011). A "mille-feuille" of silencing : Epigenetic control of transposable elements. *Biochim Biophys Acta*.

[Rizzon *et al.*, 2002] Rizzon, C., Marais, G., Gouy, M. et *al.* (2002). Recombination rate and the distribution of transposable elements in the *Drosophila melanogaster* genome. *Genome Res*, 12:400–407.

[Rizzon *et al.*, 2003] Rizzon, C., Martin, E., Marais, G. et *al.* (2003). Patterns of selection against transposons inferred from the distribution of Tc1, Tc3 and Tc5 insertions in the mut-7 line of the nematode *Caenorhabditis elegans*. *Genetics*, 165:1127–1135.

[Robert et Bucheton, 2004] Robert, V. et Bucheton, A. (2004). Régulation de l'expression des séquences répétées et interférence par l'ARN. *Médecine/Sciences*, 20:767–772.

[Robertson et Wolffe, 2000] Robertson, K. D. et Wolffe, A. P. (2000). DNA methylation in health and disease. *Nat Rev Genet*, 1:11–19.

[Rubin *et al.*, 1982] Rubin, G. M., Kidwell, M. G. et Bingham, P. M. (1982). The molecular basis of P-M hybrid dysgenesis : the nature of induced mutations. *Cell*, 29:987–994.

[Ryan *et al.*, 2010] Ryan, B. M., Robles, A. I. et Harris, C. C. (2010). Genetic variation in microRNA networks : the implications for cancer research. *Nat Rev Cancer*, 10:389–402.

[Sakano *et al.*, 1979] Sakano, H., Hüppi, K., Heinrich, G. et *al.* (1979). Sequences at the somatic recombination sites of immunoglobulin light-chain genes. *Nature*, 280:288–294.

[Sala et Corona, 2008] Sala, A. et Corona, D. F. V. (2008). Epigenetics : More than genetics. *Fly (Austin)*, 2:165–168.

[Sanborn *et al.*, 2011] Sanborn, J. Z., Benz, S. C., Craft, B. et *al.* (2011). The UCSC Cancer Genomics Browser : update 2011. *Nucleic Acids Res*, 39:D951–D959.

[SanMiguel *et al.*, 1996] SanMiguel, P., Tikhonov, A., Jin, Y. K. et *al.* (1996). Nested retrotransposons in the intergenic regions of the maize genome. *Science*, 274:765–768.

[Sasaki *et al.*, 2002] Sasaki, T., Matsumoto, T., Yamamoto, K. et *al.* (2002). The genome sequence and structure of rice chromosome 1. *Nature*, 420:312–316.

[Schnable *et al.*, 2009] Schnable, P. S., Ware, D., Fulton, R. S. et *al.* (2009). The B73 maize genome : complexity, diversity, and dynamics. *Science*, 326:1112–1115.

[Schotta *et al.*, 2004] Schotta, G., Lachner, M., Sarma, K. et *al.* (2004). A silencing pathway to induce H3-K9 and H4-K20 trimethylation at constitutive heterochromatin. *Genes Dev*, 18:1251–1262.

[Schwartz *et al.*, 1998] Schwartz, A., Chan, D. C., Brown, L. G. et *al.* (1998). Reconstructing hominid Y evolution : X-homologous block, created by X-Y transposition, was disrupted by Yp inversion through LINE-LINE recombination. *Hum Mol Genet*, 7:1–11.

[Schwartz *et al.*, 2009] Schwartz, S., Meshorer, E. et Ast, G. (2009). Chromatin organization marks exon-intron structure. *Nat Struct Mol Biol*, 16:990–995.

[Seligson *et al.*, 2009] Seligson, D. B., Horvath, S., McBrian, M. A. et *al.* (2009). Global levels of histone modifications predict prognosis in different cancers. *Am J Pathol*, 174:1619–1628.

[Shalgi *et al.*, 2010] Shalgi, R., Pilpel, Y. et Oren, M. (2010). Repression of transposable-elements - a microRNA anti-cancer defense mechanism ? *Trends Genet*, 26:253–259.

[Shankar *et al.*, 2004] Shankar, R., Grover, D., Brahmachari, S. K. et *al.* (2004). Evolution and distribution of RNA polymerase II regulatory sites from RNA polymerase III dependant mobile *Alu* elements. *BMC Evol Biol*, 4:37.

[Silahtaroglu et Stenvang, 2010] Silahtaroglu, A. et Stenvang, J. (2010). MicroRNAs, epigenetics and disease. *Essays Biochem*, 48:165–185.

[Simons *et al.*, 2007] Simons, C., Makunin, I. V., Pheasant, M. et *al.* (2007). Maintenance of transposon-free regions throughout vertebrate evolution. *BMC Genomics*, 8:470.

[Simons *et al.*, 2006] Simons, C., Pheasant, M., Makunin, I. V. et *al.* (2006). Transposon-free regions in mammalian genomes. *Genome Res*, 16:164–172.

[Sironi *et al.*, 2006] Sironi, M., Menozzi, G., Comi, G. P. et *al.* (2006). Gene function and expression level influence the insertion/fixation dynamics of distinct transposon families in mammalian introns. *Genome Biol*, 7:R120.

[Slotkin et Martienssen, 2007] Slotkin, R. K. et Martienssen, R. (2007). Transposable elements and the epigenetic regulation of the genome. *Nat Rev Genet*, 8:272–285.

[Smalheiser et Torvik, 2005] Smalheiser, N. R. et Torvik, V. I. (2005). Mammalian microRNAs derived from genomic repeats. *Trends Genet*, 21:322–326.

[Smit *et al.*, 2010] Smit, A. F. A., Hubley, R. et Green, P. (1996-2010). RepeatMasker Open3.0.

[Smith *et al.*, 2007] Smith, I. M., Mydlarz, W. K., Mithani, S. K. et *al.* (2007). DNA global hypomethylation in squamous cell head and neck cancer associated with smoking, alcohol consumption and stage. *Int J Cancer*, 121:1724–1728.

[Sorek *et al.*, 2002] Sorek, R., Ast, G. et Graur, D. (2002). *Alu*-containing exons are alternatively spliced. *Genome Res*, 12:1060–1067.

[Stenger *et al.*, 2001] Stenger, J. E., Lobachev, K. S., Gordenin, D. et *al.* (2001). Biased distribution of inverted and direct *Alus* in the human genome : implications for insertion, exclusion, and genome stability. *Genome Res*, 11:12–27.

[Su *et al.*, 2004] Su, A. I., Wiltshire, T., Batalov, S. et *al.* (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A*, 101:6062–6067.

[Surzycki et Belknap, 2000] Surzycki, S. A. et Belknap, W. R. (2000). Repetitive-DNA elements are similarly distributed on *Caenorhabditis elegans* autosomes. *Proc Natl Acad Sci U S A*, 97:245–249.

[Szpakowski *et al.*, 2009] Szpakowski, S., Sun, X., Lage, J. M. et *al.* (2009). Loss of epigenetic silencing in tumors preferentially affects primate-specific retroelements. *Gene*, 448:151–167.

[Tamaru et Selker, 2001] Tamaru, H. et Selker, E. U. (2001). A histone H3 methyltransferase controls DNA methylation in *Neurospora crassa*. *Nature*, 414:277–283.

[Taslim *et al.*, 2009] Taslim, C., Wu, J., Yan, P. et *al.* (2009). Comparative study on ChIP-seq data : normalization and binding pattern characterization. *Bioinformatics*, 25:2334–2340.

[Tenaillon *et al.*, 2010] Tenaillon, M. I., Hollister, J. D. et Gaut, B. S. (2010). A triptych of the evolution of plant transposable elements. *Trends Plant Sci*, 15:471–478.

[Tsirigos et Rigoutsos, 2009] Tsirigos, A. et Rigoutsos, I. (2009). *Alu* and *B1* repeats have been selectively retained in the upstream and intronic regions of genes of specific functional classes. *PLoS Comput Biol*, 5:e1000610.

[Urrutia *et al.*, 2008] Urrutia, A. O., Ocaña, L. B. et Hurst, L. D. (2008). Do *Alu* repeats drive the evolution of the primate transcriptome ? *Genome Biol*, 9:R25.

[Venner *et al.*, 2009] Venner, S., Feschotte, C. et Biémont, C. (2009). Dynamics of transposable elements : towards a community ecology of the genome. *Trends Genet*, 25:317–323.

[Volff, 2006] Volff, J.-N. (2006). Turning junk into gold : domestication of transposable elements and the creation of new genes in eukaryotes. *Bioessays*, 28:913–922.

[von Sternberg et Shapiro, 2005] von Sternberg, R. et Shapiro, J. A. (2005). How repeated retroelements format genome function. *Cytogenet Genome Res*, 110:108–116.

[Waddington, 1942] Waddington, C. (1942). The epigenotype. *Endeavour*, 1:18–20.

[Wang *et al.*, 2008] Wang, Z., Zang, C., Rosenfeld, J. A. et *al.* (2008). Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat Genet*, 40:897–903.

[Wang-Johanning *et al.*, 2001] Wang-Johanning, F., Frost, A. R., Johanning, G. L. et *al.* (2001). Expression of human endogenous retrovirus k envelope transcripts in human breast cancer. *Clin Cancer Res*, 7:1553–1560.

[Wang-Johanning *et al.*, 2007] Wang-Johanning, F., Liu, J., Rycaj, K. et *al.* (2007). Expression of multiple human endogenous retrovirus surface envelope proteins in ovarian cancer. *Int J Cancer*, 120:81–90.

[Warnefors *et al.*, 2010] Warnefors, M., Pereira, V. et Eyre-Walker, A. (2010). Transposable elements : insertion pattern and impact on gene expression evolution in hominids. *Mol Biol Evol*, 27:1955–1962.

[Weber, 2008] Weber, M. (2008). Profils de méthylation de l'ADN dans les cellules normales et cancéreuses. *Médecine/Sciences*, 24:731–734.

[Weber *et al.*, 2005] Weber, M., Davies, J. J., Wittig, D. et *al.* (2005). Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat Genet*, 37:853–862.

[Weber *et al.*, 2007] Weber, M., Hellmann, I., Stadler, M. B. et *al.* (2007). Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat Genet*, 39:457–466.

[Weber et Schübeler, 2007] Weber, M. et Schübeler, D. (2007). Genomic patterns of DNA methylation : targets and function of an epigenetic mark. *Curr Opin Cell Biol*, 19:273–280.

[(WHO), 2008] (WHO), W. H. O. (2008). the global burden of disease : 2004 update. *WHO*.

[Wicker *et al.*, 2007] Wicker, T., Sabot, F., Hua-Van, A. et *al.* (2007). A unified classification system for eukaryotic transposable elements. *Nat Rev Genet*, 8:973–982.

[Wong *et al.*, 2000] Wong, G. K., Passey, D. A., Huang, Y. et *al.* (2000). Is "junk" DNA mostly intron DNA ? *Genome Res*, 10:1672–1678.

[Woodcock et Dimitrov, 2001] Woodcock, C. L. et Dimitrov, S. (2001). Higher-order structure of chromatin and chromosomes. *Curr Opin Genet Dev*, 11:130–135.

[Wu et Morris, 2001] Wu, C. et Morris, J. R. (2001). Genes, genetics, and epigenetics : a correspondence. *Science*, 293:1103–1105.

[Wu *et al.*, 2009] Wu, C., Orozco, C., Boyer, J. et *al.* (2009). BioGPS : an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol*, 10:R130.

[Xu *et al.*, 2004] Xu, P., Widmer, G., Wang, Y. et *al.* (2004). The genome of *Cryptosporidium hominis*. *Nature*, 431(7012):1107–1112.

[Yang *et al.*, 2008] Yang, X., Suna, X., Xie, J. et *al.* (2008). Comparability of gene expression in human blood, immune and carcinoma cells. *Appl Math Comput*, 205:178–184.

[Yang, 1998] Yang, Z. (1998). Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol*, 15:568–573.

[Yang, 2002] Yang, Z. (2002). Inference of selection from multiple species alignments. *Curr Opin Genet Dev*, 12:688–694.

[Yang, 2007] YANG, Z. (2007). PAML 4 : phylogenetic analysis by maximum likelihood. *Mol Biol Evol*, 24:1586–1591.