



**HAL**  
open science

## **TE variation in natural populations of *Drosophila*: copy number, transcription and chromatin state**

Rita Rebollo

### ► **To cite this version:**

Rita Rebollo. TE variation in natural populations of *Drosophila*: copy number, transcription and chromatin state. Agricultural sciences. Université Claude Bernard - Lyon I, 2009. English. <NNT: 2009LYO10186>. <tel-00580831>

**HAL Id: tel-00580831**

**<https://theses.hal.science/tel-00580831v1>**

Submitted on 29 Mar 2011

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization

THESE DE L'UNIVERSITE DE LYON

Délivrée par

L'UNIVERSITE CLAUDE BERNARD LYON 1

EVOLUTION ECOSYSTEMES MICROBIOLOGIE MODELISATION

DIPLOME DE DOCTORAT

(arrêté du 7 août 2006)

soutenue publiquement le 26 octobre 2009

par

Mlle Rita REBOLLO

TITRE : TE variation in natural populations of *Drosophila* : copy number, transcription and chromatin state

Directeur de thèse : Cristina VIEIRA

JURY : Pr Gunter REUTER

Pr Maria del Pilar GARCIA GUERREIRO

Pr Claudia Marcia Aparecida CARARETO

Dr Benjamin LOPPIN

Pr Dominique MOUCHIROUD

Dr Cristina VIEIRA





Université Claude Bernard – Lyon 1

Laboratoire de Biométrie et Biologie Evolutive

CNRS UMR 5558

43 Boulevard du 11 novembre 1918

69622 Cedex Villeurbanne

Equipe TrEEP : Transposable elements, evolution and population

Rita REBOLLO

[rebollo@biomserv.univ-lyon1.fr](mailto:rebollo@biomserv.univ-lyon1.fr)

Mots Clés – Key words

Eléments transposables – Transposable elements

*Drosophila*

Populations naturelles – Natural populations

Délétion – Deletion

Epigénétique – Epigenetics

# UNIVERSITE CLAUDE BERNARD - LYON 1

## **Président de l'Université**

Vice-président du Conseil Scientifique

Vice-président du Conseil d'Administration

Vice-président du Conseil des Etudes et de la Vie  
Universitaire

Secrétaire Général

**M. le Professeur L. Collet**

M. le Professeur J-F. Mornex

M. le Professeur G. Annat

M. le Professeur D. Simon

M. G. Gay

## ***COMPOSANTES SANTE***

Faculté de Médecine Lyon Est – Claude Bernard

Faculté de Médecine Lyon Sud – Charles Mérieux

UFR d'Odontologie

Institut des Sciences Pharmaceutiques et Biologiques

Institut des Sciences et Techniques de Réadaptation

Département de Formation et Centre de Recherche en  
Biologie Humaine

Directeur : M. le Professeur J. Etienne

Directeur : M. le Professeur F-N. Gilly

Directeur : M. le Professeur D. Bourgeois

Directeur : M. le Professeur F. Locher

Directeur : M. le Professeur Y. Matillon

Directeur : M. le Professeur P. Farge

## ***COMPOSANTES SCIENCES ET TECHNOLOGIE***

Faculté des Sciences et Technologies

UFR Sciences et Techniques des Activités Physiques  
et Sportives

Observatoire de Lyon

Institut des Sciences et des Techniques de l'Ingénieur  
de Lyon

Institut Universitaire de Technologie A

Institut Universitaire de Technologie B

Institut de Science Financière et d'Assurance

Institut Universitaire de Formation des Maîtres

Directeur : M. Le Professeur F. Gieres

Directeur : M. C. Collignon

Directeur : M. B. Guiderdoni

Directeur : M. le Professeur J. Lieto

Directeur : M. le Professeur C. Coulet

Directeur : M. le Professeur R. Lamartine

Directeur : M. le Professeur J-C. Augros

Directeur : M R. Bernard

VARIATION DES ELEMENTS TRANSPOSABLES DANS LES POPULATIONS NATURELLES DE  
*DROSOPHILA* : NOMBRE DE COPIES, TRANSCRIPTION ET ETAT DE LA CHROMATINE.

Les éléments transposables (ET) sont une source majeure de variation génétique, ce qui leur confère un rôle essentiel dans l'évolution des génomes. Certes présents dans tous les génomes analysés à ce jour, leurs proportions sont fortement variables entre espèces et aussi entre populations, suggérant une relation unique entre génome hôte et ET. Grâce à un système modèle composé de populations naturelles de deux espèces proches (*Drosophila melanogaster* et *D. simulans*) avec des quantités différentes en ET, nous avons pu comparer les relations génome hôte/ET. Nous nous sommes particulièrement intéressés à l'élément *helen* qui, chez *D. simulans*, montre une activité faible, malgré un nombre de copies élevé. Cette activité moindre est associée à de nombreuses délétions internes des copies, suggérant un mécanisme de régulation d'ET par des délétions de l'ADN. Un autre système de régulation de l'activité des ET utilise le contrôle épigénétique, ce qui permet le maintien des copies d'ET dans le génome mais un blocage de leur activité. Le remodelage de la chromatine est un système épigénétique bien décrit chez la drosophile. Les régions chromatiniennes des génomes sont associées à différents types de modifications d'histone. Nous avons mis en évidence, dans des populations de *D. melanogaster* et *D. simulans*, une variation conséquente de modifications d'histones de type hétérochromatique, H3K27me3 et H3K9me2, associées à des copies de différents ET. De plus, nous avons décrit des populations chez *D. simulans* dites déréprimées, chez lesquelles certains éléments sont surexprimés et présentent des localisations probablement hétéchromatiques. Les ET sont donc contrôlés par le génome hôte par des délétions internes et probablement par un système épigénétique variable. De plus, dans certaines populations, des copies peuvent échapper à ce contrôle et envahir le génome. Les ET sont donc des grands créateurs de variabilité génétique mais permettent aussi une territorialisation chromatinienne du génome car ils portent des modifications épigénétiques précises et sont capables de les étendre à leurs environnements génomiques. Ceci leur confère la fonction "d'épigénétique mobile".

TE VARIATION IN NATURAL POPULATIONS OF *DROSOPHILA* : COPY NUMBER, TRANSCRIPTION  
AND CHROMATIN STATE

Transposable elements (TEs) are one major force of genome evolution thanks to their ability to create genetic variation. TEs are ubiquitous and their proportion is variable between species and also populations, suggesting that a tight relationship exists between genomes and TEs. The model system composed of the natural populations of the twin sisters *Drosophila melanogaster* and *D. simulans* is interesting to compare host/TE relationship, since both species harbour different amounts of TE copies. The *helena* element is nearly silenced in *D. simulans* natural populations despite a very high copy number. Such repression is associated to abundant internally deleted copies suggesting a regulatory mechanism of TEs based on DNA deletion. Another pathway of TE regulation is through epigenetics where the host genome is able to keep intact the DNA sequences of TEs and still silence their activities. Chromatin remodelling is well known in *drosophila* and specific histone modifications can be associated to specific chromatin domains. We observed an important variation on H3K27me3 and H3K9me2, two heterochromatic marks, on TE copies in *D. melanogaster* and *D. simulans* natural populations. Also, we show that derepressed lines of *D. simulans* exist for specific elements, have high TE transcription rates and are highly associated to non constitutive heterochromatic marks. TEs are therefore controlled by the host genome through DNA deletion and a possible chromatin remodelling mechanism. Not only genetic variability is enhanced by TEs but also epigenetic variability, allowing the host genome to be partitioned into chromatin domains. TEs are therefore mandatory to gene network regulation through their ability of “jumping epigenetics”.

La partie codante des génomes ne constitue pas, dans beaucoup d'espèces, la majorité des séquences d'ADN présentes (Biemont and Vieira, 2006). En réalité, ces régions codantes sont souvent noyées dans une abondance de séquences répétées, qui pendant longtemps n'étaient pas associées à des « fonctions » précises et ont par conséquent été nommées « Junk DNA » (Ohno, 1972). Aujourd'hui, multiples sont les effets décrits de ces répétitions : 1) effets directs sur le génome, telle la domestication des parties codantes des répétitions (au sein de centromères et télomères par exemple (Sinzelle et al., 2009) ou leur utilisation dans des systèmes de régulation du génome hôte (séquences régulatrices d'une répétition permettant le contrôle des gènes (Marino-Ramirez et al., 2005), ou dérivés ARN permettant la régulation de gènes par ARN interférence (Hasler et al., 2007) – 2) effets indirects, tel l'apport de variabilité génétique (multiplication et recombinaisons (Hedges and Deininger, 2007) et l'augmentation de la taille des génomes hôtes (Boulesteix et al., 2006), deux phénomènes qui participent à l'évolution des espèces.

Les éléments transposables (ET), partie majeure du « Junk DNA », sont des séquences d'ADN capables de se multiplier à l'intérieur d'un génome, engendrant des mutations. Selon leur mode de déplacement, les ET peuvent être classés en deux grands groupes : les rétrotransposons, qui se mobilisent *via* une molécule d'ARN par un mécanisme de « copier/coller », et les transposons, qui se mobilisent *via* une molécule d'ADN par un « couper/coller » (cf. Figure 1 page 103). Les mécanismes de transposition ainsi que l'utilisation d'ET pour les réparations de cassures double brin, induisent l'augmentation du nombre de copies de ces séquences dans le génome. Les processus de délétions et recombinaisons induisent à leurs tours une diminution du nombre de copies dans le génome. La proportion d'ET est très variable dans les espèces analysées, allant de 10% pour *Arabidopsis thaliana* à 80% pour le maïs (Biemont and Vieira, 2006). De plus, des variations dans la proportion d'ET sont aussi observées entre des populations d'une même espèce (Vieira et al., 1999; Biemont, 2008), suggérant l'existence d'une relation très étroite entre les ET et le génome. C'est exactement cette spécificité relationnelle que nous avons essayé de comprendre en étudiant un système modèle composé de deux espèces proches, *D. melanogaster* et *D. simulans*, qui ne possèdent pas la même quantité d'ET (15% et 5% environ, respectivement). Ces deux espèces sont cosmopolites, étroitement liées à l'Homme, et ont une aire de distribution assez proche, avec cependant des endroits du globe qui n'ont pas encore été colonisés par *D. simulans*. De plus, *D. simulans*, contrairement à *D.*

*melanogaster*, ne possède pas un nombre constant de copies d'ET, c'est-à-dire, une faible variance du nombre de copies d'ET entre les populations naturelles ; ce qui permet une étude intraspécifique des relations ET/hôte. Aussi bien par des analyses bibliographiques que par des travaux de recherche, nous avons essayé de comprendre la relation ET/hôte chez la drosophile, avec une description des facteurs qui interviennent dans cette relation.

À travers une analyse bibliographique, nous avons constaté que les promoteurs des ET constituent un premier système de régulation car l'invasion d'un génome est dépendante des facteurs de transcription fournis par l'hôte (Fablet et al., 2007). Sachant qu'il existe plusieurs types de séquences régulatrices, spécifiques à chaque type d'ET, et qu'à ceci s'ajoute la variation de facteurs de transcription entre les espèces, nous suggérons que la relation ET/hôte est en partie une conséquence de la relation « promoteur ET / facteur de transcription de l'hôte » (Fablet et al., 2007).

Parmi les systèmes de régulation qui contrent les effets mutagènes des ET et empêchent l'invasion des génomes, nous avons étudié le système de régulation par délétion avec l'exemple de l'élément *helena* chez la drosophile (Rebollo et al., 2008). *Helena* est un rétrotransposon de type LINE (long interspersed nuclear element), qui possède une dynamique inverse chez *D. melanogaster* par rapport aux autres ET car très peu de copies y sont observées (Vieira et al., 1999). L'analyse par Southern blot de cet élément dans les populations naturelles de *D. melanogaster* ainsi que l'analyse *in silico* du génome séquencé de cette espèce, montrent qu'*helena* est certes présent dans ces génomes mais à très faible fréquence, très détérioré et de façon fixée entre les populations, suggérant une absence d'activité de cet élément. En effet, aucune copie complète d'*helena* n'est détectée, ce qui nous permet d'émettre l'hypothèse d'une éventuelle extinction de cet élément chez *D. melanogaster*. Au contraire, dans les populations naturelles de *D. simulans* et dans le génome séquencé, les copies d'*helena* sont observées en abondance. De plus, le grand polymorphisme d'insertion et l'observation de copies complètes suggèrent que l'élément est actif. Néanmoins, l'analyse exhaustive de différentes copies d'*helena* dans plusieurs populations de *D. simulans* et dans le génome séquencé montre un mécanisme de délétions internes important, déjà décrit chez *D. melanogaster* (Petrov and Hartl, 1998). L'absence de transcrits observée par northern blot accentue l'idée que les copies d'*helena* dans le génome de *D. simulans* ne soient pas actives et cela probablement dû aux délétions abondantes (Rebollo et al., 2008). Cette analyse montre, premièrement, que l'observation d'un polymorphisme d'insertion au sein des populations de *D. simulans*, n'est pas corrélée à une activité forte de cet élément, mais probablement à une activité récente. Deuxièmement, nous observons beaucoup de séquences

ayant des délétions internes mais qui sont très similaires entre elles (séquences nucléotidiques), suggérant la rapidité de l'apparition des délétions. L'état de l'élément *helena* aussi bien chez *D. melanogaster* que chez *D. simulans* semble inactif mais la présence d'une copie complète dans le génome séquencé de *D. simulans* suggère une possibilité d'invasion future du génome. En effet, des vagues de transpositions ont souvent été observées chez les mammifères pour des éléments de type LINE (Adey et al., 1994; Khan et al., 2006). Les mécanismes d'élimination des séquences d'ET ont déjà été décrits dans d'autres espèces chez lesquelles les recombinaisons homologues privent les éléments de type LTR (long terminal repeats) de leurs parties codantes (certaines recombinaison LTR-LTR) ou peuvent causer la délétion complète d'un élément (Vitte and Panaud, 2003; van de Lagemaat et al., 2005). Or, le mécanisme de délétion interne décrit ici ne semble pas commun à toutes les espèces ni aux éléments de type LINE. La poursuite de ce travail par A. Granzotto et collaborateurs (Granzotto et al., 2009) et par E. Lerat (communication personnelle) montre que le système de délétions internes est commun à d'autres *drosophilidae* et agit sur tous les rétrotransposons analysés chez *D. simulans*. De plus, *D. melanogaster* semble être la seule à échapper à ce système même si l'élément *helena* s'y trouve très dégradé. On observe donc : 1) une relation spécifique entre l'élément *helena* et le génome de *D. melanogaster* par rapport aux autres éléments de cette espèce, 2) un comportement différent entre les ET et le génome de *D. melanogaster* par rapport aux autres espèces de drosophiles. Les systèmes d'élimination des séquences d'ET sont donc variables entre les espèces et entre les éléments et participent à la spécificité de la relation ET/hôte.

Le deuxième système de régulation des ET que nous avons étudié est le contrôle épigénétique qui rend possible le maintien des ET dans le génome hôte tout en empêchant leur activité. Trois grandes voies épigénétiques permettent la régulation des ET et du reste du génome : la conformation de la chromatine, la méthylation de l'ADN et les processus impliquant des petits ARN (Lisch, 2009). Nous nous sommes premièrement concentrés sur l'analyse des structures chromatiniennes des ET dans les populations naturelles de *D. melanogaster* et *D. simulans*. L'ADN s'enroule autour d'un grand complexe protéique appelé nucléosome qui est formé de dimères d'histones. Les queues N-terminales des histones se trouvent libres du nucléosome et subissent des modifications post-traductionnelles leur conférant des états physiques précis. Les modifications de type acétylation, souvent sur les lysines N-terminales, sont observées surtout au niveau de la chromatine propice à la transcription, c'est à dire, de l'euchromatine. Au contraire, la méthylation des lysines (à l'exception de la lysine 4 sur l'histone 3 – H3K4) est plus souvent observée dans des zones où

la chromatine est fermée, c'est à dire, l'hétérochromatine. Sachant qu'il existe un nombre variable de copies d'ET entre les deux espèces de drosophile analysées et aussi entre les populations naturelles de *D. simulans*, nous nous sommes demandé s'il existait une variation dans l'association aux marques d'histones avec ces éléments.

Notre étude porte sur quatre ET (trois éléments qui possèdent des LTR : *412*, *tirant* et *roo*, et un élément de type LINE : *F*) dans deux populations de *D. melanogaster* et cinq populations de *D. simulans*. Premièrement, la comparaison du nombre de copies obtenu par hybridation *in situ* avec une sonde d'ET complète faite par Vieira et collaborateurs (Vieira et al., 1999) avec les données d'amplification par PCR quantitative d'un petit fragment d'ET (environ 400pb) suggèrent que chez *D. simulans* les copies d'ET sont probablement délétées, ce qui renforce l'hypothèse de mécanisme de délétion interne décrite ci-dessus. De plus, nous avons recherché les associations entre les marques d'histones permissives (H3K4me2) ou répressives (H3K27me3 et H3K9me2) et les ET, puis quantifié les transcrits de chaque ET dans les sept populations de drosophile étudiées. Il existe effectivement une variation des associations épigénétiques sur les ET non seulement entre les espèces mais aussi entre les populations naturelles. En effet, deux systèmes différents agissent chez l'espèce *D. melanogaster*, pour laquelle deux populations analysées présentent environ le même nombre de copies d'ET, les mêmes associations aux marques d'histones, mais une de ces populations est réprimée pour les quatre ET étudiés. Il existe probablement d'autres systèmes, tels la méthylation de l'ADN ou la RNAi, qui pourraient être différents entre ces deux populations et qui expliqueraient le comportement des ET analysés. Chez *D. simulans*, nous avons mis en évidence des populations dites « dérprimées » qui possèdent un taux important de transcrits d'un ET donné, beaucoup de représentations de cet élément dans le génome et des copies fortement associées à H3K27me3. Il pourrait donc exister un système de régulation épigénétique qui serait variable entre les espèces de drosophile et au sein d'une même espèce, et qui pourrait contribuer aux différences dans les quantités d'ET observées.

Il est important de comprendre que les deux systèmes étudiés (délétions et régulation épigénétique) sont complémentaires car la suppression épigénétique n'empêche pas les recombinaisons souvent délétères dues à la présence de répétitions. De plus, la suppression totale des ET d'un génome peut aussi être délétère à long terme car les ET apportent une variabilité considérée comme importante pour l'évolution des espèces (McClintock, 1984). En effet, une analyse critique de la bibliographie concernant l'impact des ET dans l'évolution des espèces ainsi que les travaux de recherche de l'équipe, nous amènent à proposer l'hypothèse selon laquelle les variations environnementales ayant un impact sur la régulation épigénétique

des ET pourraient moduler la réponse du génome hôte à un quelconque stress. Un pic d'activité des ET pourrait promouvoir une augmentation rapide de la variabilité génétique et donc participer à une réponse adaptative au stress. Une répression *de novo* des copies récemment transposés pourrait se mettre en place dans le génome réarrangé et on observerait une stabilisation de l'activité des éléments et des systèmes de régulation. Une variation dans la taille du génome réarrangé reste possible mais pas nécessaire, permettant d'inclure les espèces à « petit génome » dans cette hypothèse d'adaptation par des bursts de transposition. La régulation épigénétique des ET est donc, aussi bien que les effets mutagènes des ET, importante pour l'évolution des espèces. De plus, les ET créent, modulent et contrôlent des réseaux de gènes, car comme décrit plus haut, ils possèdent des séquences régulatrices capables de contrôler l'expression des gènes (Feschotte, 2008). A ceci s'ajoute leur caractère multi copies et ubiquitaires dans le génome, qui ne fait qu'augmenter leurs effets potentiels. Nous proposons qu'un nouveau niveau de régulation de gènes existe, où non seulement les séquences régulatrices des ET sont nécessaires mais aussi les marques épigénétiques sur les copies d'ET. Les copies permettraient la compartimentation du génome en zones chromatiniennes définies et influenceraient aussi les gènes avoisinants par leurs marques épigénétiques.

## SOMMAIRE

|   |     |
|---|-----|
| INTRODUCTION.....   | 13  |
| THE EVOLUTION OF RETROTRANSPOSON REGULATORY REGIONS AND ITS CONSEQUENCES ON THE <i>DROSOPHILA MELANOGASTER</i> AND <i>HOMO SAPIENS</i> HOST GENOMES ..... | 16  |
| Conclusion.....   | 24  |
| LOSING <i>HELENA</i> : THE EXTINCTION OF A <i>DROSOPHILA</i> LINE-LIKE ELEMENT .....  | 25  |
| Conclusion.....   | 50  |
| VARIATION OF HISTONE MODIFICATIONS ASSOCIATED TO TRANSPOSABLE ELEMENTS IN NATURAL POPULATIONS OF <i>D. MELANOGASTER</i> AND <i>D. SIMULANS</i> .....      | 52  |
| JUMPING GENES AND EPIGENETICS : TOWARDS NEW SPECIES.....  | 89  |
| GENERAL CONCLUSIONS.....  | 109 |
| COMPLEMENTARY WORK.....   | 111 |
| REFERENCES.....   | 127 |

## INTRODUCTION

Apart from Greeks, with Aristotle who once proposed that human beings had “heritable characteristics” in their blood, Mendel was the first who showed the existence of “heritable factors”, i.e. genes, in 1865 (Griffiths, 1999; Griffiths, 2000). Since that time, important discoveries were made: genes are on chromosomes (Morgan, 1910), genes code for proteins (Beadle and Tatum, 1941), DNA is organized in a double helix (Watson and Crick, 1953) etc. In the 70’s, the study of whole genomes suggested a discrepancy between genome size and “complex species”: the C-value paradox (reviewed in (Gregory, 2005). In parallel, the discovery of jumping genes by B. McClintock in the 50’s (Mc, 1950; McClintock, 1953) and non coding DNA repeats during the 70’s, induced several hypotheses on the quality of the information handled by genomes. Jumping genes (or transposable elements – TEs) are DNA molecules capable of moving from one chromosomal location to another. Such observation was hardly accepted by the scientific community who just got used to the idea of genes fixed on chromosomes. Because of such scepticism and the lack of function for all repeats, these DNA sequences such as TEs were called “junk DNA” (Ohno, 1972). The most complete answer from the scientific community to the C-value paradox arose from whole genome sequencing projects that began in the end of the 70’s. Genome size (and gene amount) was not correlated to “species complexity” because the “junk DNA” proportion of genomes can be extremely high as it can be variable (Gregory, 2005). We know today that repeats can vary from 80% in some species (maize) to 10% in others (Arabidopsis) (Biemont and Vieira, 2006). It has also been shown that these repeats are essential for centromeres and telomeres integrity, they may be co-opted by the genome (Sinzelle et al., 2009) and a high number of regulatory sequences in humans are derived from TEs (Marino-Ramirez et al., 2005). Furthermore, repeats are important recombinogenic substrates allowing for rapid genome remodelling (Hedges and Deininger, 2007). We indeed know today that “junk DNA” is nothing but informative (Muotri et al., 2007).

We will solely be interested in TEs because they represent the largest part of “junk DNA”, are present in both euchromatin and dense heterochromatin, and have the ability to influence all the genome. TEs can be separated into two major classes given their transposition mechanisms (for a figure, please refer to “Figure 1” page 103). Class I moves through an RNA molecule and are called retrotransposons. Class II jumps through a DNA molecule, and are called DNA transposons. Both classes can be further separated into subclasses regarding TE structure and evolutionary aspects. Finally, TE families can be distinguished regarding

DNA sequence homology. TEs can be considered as genetic units bringing positive, neutral and negative effects to the genome. In order to understand the relationship that exists between TEs and genomes we studied natural regulation of TEs in *Drosophila*. Indeed, fruit flies are one of the most classical models for genetic studies as showed by TH Morgan researches in white-eye flies and his hypotheses on genes and chromosomes (Morgan, 1910). *Drosophila* offers important experimental qualities as they are easy to keep in laboratories, breed extremely fast and can be maintained in important amounts of individuals. The existence of natural populations of *Drosophila* from close species, allows us to study different combinations of TEs and genomes.

The goal of my work was to understand the connexion between TE natural variation between and inside species, and genome response to TEs. A bibliographic analysis along with other members of the group, on TE regulatory regions suggests that common features exist between different types of TEs and different species (Fablet et al., 2007). However, the arm race of TEs and genomes necessarily brings specificity to each TE/genome complex, suggesting that regulatory mechanisms of TEs may vary between species and TEs analysed (Fablet et al., 2007). In order to experimentally test such suggestion, we used the system model composed of the sibling species, *D. melanogaster* and *D. simulans* along with several natural populations of both species. *D. melanogaster* was previously described as harbouring tree times more copies than *D. simulans* [16]. Such data was based on the study of natural populations allowing the observation of variation not only between species but most important, inside each species. Interestingly, all TE families are over represented in *D. melanogaster* natural populations compared to *D. simulans* ones. Few elements, as the LINE-like *helena* element, harboured the opposite dynamic, being outnumbered in *D. simulans* natural populations. As a result of such observation, the first question addressed experimentally during my PhD was to understand *helena* opposite occurrence in our system model. Natural population variability of *helena* between *D. simulans* and *D. melanogaster* confirmed the specificity of TE regulation between species (Rebollo et al., 2008). We hypothesize on the existence of a rapid internal deletion mechanism for *D. simulans* TEs. Such hypothesis was confirmed by E. Lerat *in silico* TE analysis and broaden to other species of *Drosophila* (personal communication).

The differences between *D. melanogaster* and *D. simulans* in controlling TEs and the different chromatin regions occupied by TEs in both species (Vieira et al., 1999; Capy and Gibert, 2004) lead us to our second question : could epigenetic regulatory mechanisms also be

variable between both species. We concentrated our analysis on chromatin remodelling processes by describing histone marks associated to TE copies. Also, we characterized all TE families analysed for copy number and expression patterns in every natural population studied. Our data suggests that *D. melanogaster* has more heterochromatic elements than *D. simulans*. Also, important variation of histone marks between all element analysed regarding natural populations or species suggest that chromatin conformation might be specific of the context “TE/locus/genome”. In *D. simulans* natural populations, TEs seem to be internally deleted confirming the existence of a deletion mechanism suggested for the *helena* element and the *in silico* analysis of E. Lerat. No correlation between epigenetic marks (histone post translational modifications) could however be made with TE expression or copy number. We can nevertheless propose that since expression patterns of TEs are strain-specific and often species-specific, different regulatory mechanisms might exist.

The evolutionary importance of genetic and epigenetic lineage specific regulations of TEs suggested by these experiments leads us to reflect on the impact of transposition in speciation. We know that epigenetics is influenced by environment and is the most important regulatory mechanisms of TEs. We propose therefore, in a bibliographic survey, that changes in environment can induce changes in TE epigenetic regulation, inducing bursts of transposition. The consequences of such bursts can reach speciation if important karyotypic modifications are observed.

# The evolution of retrotransposon regulatory regions and its consequences on the *Drosophila melanogaster* and *Homo sapiens* host genomes

Marie Fablet, Rita Rebollo, Christian Biémont, Cristina Vieira \*

UMR CNRS 5558, Biométrie et Biologie Evolutive, Université Claude Bernard Lyon 1, 69622 Villeurbanne Cedex, France

Received 14 June 2006; received in revised form 11 August 2006; accepted 15 August 2006

Available online 24 August 2006

Received by M. Batzer

## Abstract

It has now been established that transposable elements (TEs) make up a variable, but significant proportion of the genomes of all organisms, from Bacteria to Vertebrates. However, in addition to their quantitative importance, there is increasing evidence that TEs also play a functional role within the genome. In particular, TE regulatory regions can be viewed as a large pool of potential promoter sequences for host genes. Studying the evolution of regulatory region of TEs in different genomic contexts is therefore a fundamental aspect of understanding how a genome works. In this paper, we first briefly describe what is currently known about the regulation of TE copy number and activity in genomes, and then focus on TE regulatory regions and their evolution. We restrict ourselves to retrotransposons, which are the most abundant class of eukaryotic TEs, and analyze their evolution and the subsequent consequences for host genomes. Particular attention is paid to much-studied representatives of the Vertebrates and Invertebrates, *Homo sapiens* and *Drosophila melanogaster*, respectively, for which high quality sequenced genomes are available.

© 2006 Elsevier B.V. All rights reserved.

**Keywords:** LTR; UTR; Sequence variability; LINES

## 1. Introduction

It has now been established that transposable elements (TEs) make up a significant proportion of the genomes of all organisms, ranging from more than 50% of the human genome (International Human Genome Sequencing Consortium, 2001), to about 28% of the *Drosophila melanogaster* genome (Biémont and Vieira, 2005), and up to 90% of the genomes of some plants (Bennetzen, 2000). The increasing data obtained from genome-wide sequencing projects indicate that TEs have major functions in the organisms in which they are located. In particular, it has been shown that almost 25% of promoter regions in the human genome contain TE-derived sequences (Jordan et al., 2003), and that TEs have donated transcriptional regulatory signals to many mammalian genes (van de Lagemaat et al., 2003). Because TEs

are abundant in most genomes, they can be considered to constitute a large pool of potential promoter regions for new host regulatory sequences. Finding out how TEs are regulated is therefore important if we are to understand how the genome works.

TE regulatory regions are known to be very rapidly evolving sequences (Arkhipova et al., 1995), a characteristic of eukaryotic regulatory regions attributed to having to cope with changing genomic environments (Ludwig et al., 2000). In this article, we propose to focus on the evolution of the TE regulatory regions, and the subsequent consequences for their host genomes. We only deal with retrotransposons, which are generally the most abundant class of TEs in terms of their proportion in the genomes, and are widely distributed amongst eukaryotic genomes (Hua-Van et al., 2005). Of course, there are some exceptions, as for instance the non-autonomous Class II elements MITEs, which contributed to more than 70% of TEs in *Oriza sativa*, and about 50% in *Arabidopsis thaliana* (Hua-Van et al., 2005; Turcotte et al., 2001), or the Class II *mariner* element, which is very abundant in some Arthropods, such as *Ceratitidis capitata* (Torti et al., 2000). Retrotransposons comprise

**Abbreviations:** TE, Transposable Element; UTR, UnTranslated Region; LTR, Long Terminal Repeat; ORF, Open Reading Frame; Inr, Initiator; DPE, Downstream Promoter Element; ASP, AntiSense Promoter.

\* Corresponding author. Tel.: +33 4 72 43 29 18; fax: +33 4 72 43 13 88.

E-mail address: [vieira@biomserv.univ-lyon1.fr](mailto:vieira@biomserv.univ-lyon1.fr) (C. Vieira).

two classes of elements, the elements bordered by long terminal repeats, known as LTR retrotransposons, and the long interspersed nuclear elements (non-LTR retrotransposons), also known as LINES. Both these classes are mobilized *via* an RNA intermediate, and insert themselves into the genome by means of the reverse transcriptase that they encode in their second open reading frame (ORF). Both of these elements also encode a GAG-related, RNA binding protein in ORF1. The most important difference between LINES and LTR elements resides in their regulatory sequences. Here, we want to show that the modalities and outcomes of the rapid evolution of the regulatory region vary according to the class of retrotransposons (LTR retrotransposons/LINES) and the genome in which they reside. We focus particularly on two much-studied representatives of the Vertebrates and Invertebrates, *Homo sapiens* and *D. melanogaster*, respectively, for which the sequenced genomes are available.

## 2. Transposable elements in the genomes of *H. sapiens* and *D. melanogaster*

### 2.1. The major class of TEs is not the same in these two genomes

*Drosophila* and human genomes display quite different TE proportions (Fig. 1). *LINE-1* (*L1*) elements are the most abundant family of autonomously-replicating retroelements in mammals, accounting for 17% of the human genome (Bannert and Kurth, 2004). Nearly 8% of human DNA consists of sequences assigned to HERVs (Human Endogenous Retroviruses) (International Human Genome Sequencing Consortium, 2001), which will be considered here to be LTR retroelements. HERVs are probably genomic traces of numerous germ-line retroviral infections (Belshaw et al., 2004), and waves of invasion are known to have occurred repeatedly during primate evolution (Sverdlov, 2000). On the contrary, LINES and LTR retrotransposons correspond to 0.9% and 2.7% of the euchromatin of *D. melanogaster*, respectively (Kaminker et al., 2003). Considering that the heterochromatin represents 30–40% of the DNA of the genome of *D. melanogaster*, that 50–60% of the heterochromatin exhibits sequence similarities to known TEs (Hoskins et al., 2002; Kapitonov and Jurka, 2003), and assessing that LINES and LTR retrotransposons display the same relative proportions in euchromatin and heterochromatin, then LINES and LTR retrotransposons represent roughly about 3% and 10%, respectively, of the whole genome of *D. melanogaster*. The proportion of LINES is significantly lower than in the human genome. Another major difference between both genomes is that the genome of *D. melanogaster*

does not display any SINEs (Short Interspersed Nuclear Elements, usually associated with LINES), whereas these are prevalent in the human genome (International Human Genome Sequencing Consortium, 2001).

LINES and LTR retrotransposons do not only constitute different proportions of the genomes of *D. melanogaster* and *H. sapiens*, they also display contrasting activity patterns in both genomes. In *H. sapiens*, retrotransposons account for a large fraction of the genome, but have little activity (less than 0.2% of spontaneous mutations in humans are caused by *L1* insertions (Kazazian, 1998)), while in *D. melanogaster*, in spite of their low copy number, TEs are responsible for more than 50% of naturally-occurring mutations with major morphological effects (Eikbush and Furano, 2002). In insects, sixty or so retrotransposon families are active, whereas just one LINE family, the *L1* family, has been the major source of retrotranspositional activity in primates (International Human Genome Sequencing Consortium, 2001). Retrotransposons in the *D. melanogaster* genome seem to result from recent transposition events (Lerat et al., 2003), whereas most HERV families are now apparently extinct, even though the activity of some, such as the *HERV-L* family, appears to have persisted for a long time (Benit et al., 1999), and *HERV-K* remained active for some time after the chimpanzee/human split (Mayer et al., 1999). These HERV families consist mainly of solo LTRs, and transpositionally-deficient and transcriptionally-silent members. These differences in proportions and activity strongly suggest that TEs must be regulated differently in human and *Drosophila* genomes.

### 2.2. Overall host regulation of retrotransposon activity and copy number in the genomes of *H. sapiens* and *D. melanogaster*

Various mechanisms are involved in regulating retrotransposon activity and copy number, and they are quite different between *D. melanogaster* and *H. sapiens*. Indeed, in human cells, TEs are transcriptionally silenced due to hypermethylated CpG dinucleotides (Lavie et al., 2005), which is the major way of controlling their activity. In *H. sapiens*, abnormalities in retrotransposon hypermethylation are frequently linked to cancers, as shown by Menendez et al. (2004), who have demonstrated that *L1* and *HERV-W* are hypomethylated in ovarian carcinomas, and that relative levels of expression of these retrotransposons are significantly higher in malignant ovarian tissues. On the contrary, until very recently, the genome of *Drosophila* was considered to be non-methylated (Patel and Gopinathan, 1987; Tweedie et al., 1997), in contrast to most vertebrates and other insects. However, it is now realized that *Drosophila* does in fact methylate, mostly in young embryos (Lyko et al., 2000; Lyko, 2001), and mainly on the cytosines that are associated with the A and T nucleotides, which contrasts with the other species, in which methylation takes place on the CpG. However, the impact of methylation on the *Drosophila* genome remains to be clarified, since individuals with DNA methylation mutations do not display any specific dysfunction (Kunert et al., 2003).

In contrast to TE regulation mechanisms in the human genome, which mainly act at the transcriptional level, the TE

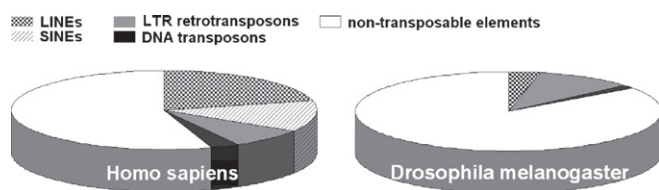


Fig. 1. Relative proportions of the various TE classes in the sequenced genomes of *Homo sapiens* and *Drosophila melanogaster*.

copy numbers are decreased by selection against insertions in the *D. melanogaster* genome. Selection can indeed act either directly against the deleterious effects of TE insertions or indirectly against the chromosomal rearrangements resulting from ectopic recombination between copies of elements belonging to the same family (Charlesworth et al., 1997; Biémont et al., 1997). Eikbush and Furano (2002) have suggested that one way to explain the difference in the total amount of TEs between the *D. melanogaster* and *H. sapiens* genomes could be that mammals have a much lower rate of ectopic homologous recombination than *D. melanogaster* (Cooper et al., 1998; van de Lagemaat et al., 2005). The explanation for the differences in TE amount and composition in different genomes is not straightforward because recombination, presence or absence of heterochromatine, host effective population size and breeding system may be involved. It should thus be borne in mind that the amount of TEs may be very different in closely related species, as has been shown for the twin species *D. melanogaster* and *D. simulans*. Indeed, the total amount of TE insertion sites is three times higher in *D. melanogaster* than in *D. simulans* (Vieira et al., 1999; Vieira and Biémont, 2004). Several hypotheses have been proposed to explain this difference, involving the genome properties of the species (Kidwell and Lisch, 2000) or the species ecology (Vieira et al., 2002), but no irrefutable evidence has been provided, and we can envisage that epigenetic mechanisms may also be responsible for controlling TEs in *Drosophila*.

While retrotransposon activity depends on host control mechanisms, such as methylation, recombination rate and epigenetic phenomena, they are also subject, as in the ‘arms race’ scenario, to their own regulatory signals, most of which are located in their 5′ regulatory regions.

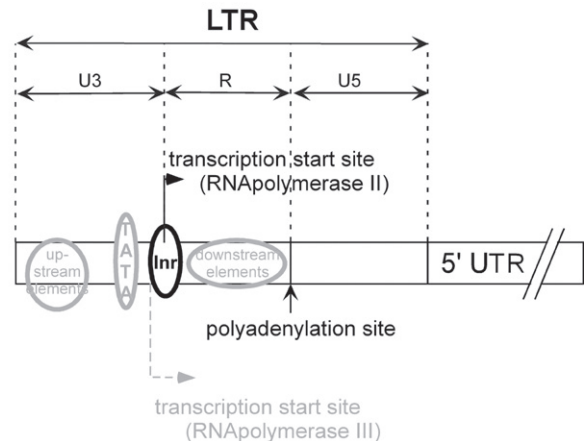
### 3. Structure and variability of retrotransposon regulatory regions

Retrotransposons are transcribed *via* RNA polymerase II, and display a tripartite canonical eukaryotic promoter structure consisting of the TATA box, the initiator element (Inr), and the downstream promoter element (DPE) (Arkhipova et al., 1995) (Fig. 2). The Inr and DPE motifs can act in conjunction to provide a binding site for TFIID in the absence of a TATA box, and a strict requirement for spacing between them has been demonstrated in both humans and *Drosophila* (Burke and Kadonaga, 1996). In some TEs, an alternative RNA start site for RNA polymerase III can be observed, as for instance 10′bp upstream of the regular RNA polymerase II start site in the *mdg1* LTR retrotransposon of *Drosophila* (Arkhipova, 1995), or in the human *L1* promoter (Kurose et al., 1995).

#### 3.1. Structure of LINE element regulatory regions

Due to their retrotransposition mechanism, the regulation signals of LINES must be located downstream from the 5′ RNA start site. Therefore, instead of TATA boxes, some internal promoters are found that consist of the Inr and the DPE (Arkhipova et al., 1995). The consensus 5′ UTR region of *L1* copies contains such an internal promoter plus a YY1 binding

#### A. LTR retrotransposons



#### B. LINES

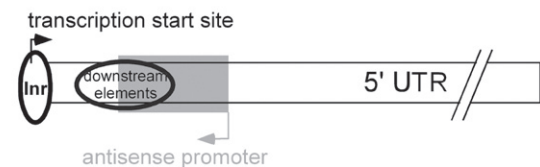


Fig. 2. Structure of typical LTR retrotransposon (A) and LINE (B) regulatory regions. Only the most common features are shown. The black elements are mandatory and found in every retrotransposon, whereas the gray elements are not essential. However, for LTR retrotransposons, the presence of either the TATA box or the downstream promoter element (DPE) is required. TATA: TATA box, Inr: Initiator.

site, which is necessary for the transcription to be initiated (Khan et al., 2006). Moreover, it has recently been shown that the *L1* transcription start site is not always located exactly at nucleotide +1, but can range from nt −9 to +4, so that the Inr extends into upstream sequences (Lavie et al., 2006). In *Drosophila*, even though the regulatory regions of non-LTR elements are characteristic of each element, some 5′ UTRs have the same overall structure. For the *I*, *doc* and *F* elements in *Drosophila* for example, the 5′ UTR promoter has two regions (A and B), which boost the role of the initiator. The B region can be divided into 3 regions containing DNA binding sites. The RNA machinery is recruited synergistically by all of these regions (Minchiotti et al., 1997). In mice, the *L1* 5′ UTR is composed of multiple direct repeats of individual promoter elements, or monomers of about 210 bp, located upstream from a single-copy, non-monomeric sequence (Mears and Hutchison, 2001; Goodier et al., 2001; Kazanian, 1998). The monomers act synergistically to increase the promoter activity.

In addition to the internal promoter necessary for their own transcription, LINES can also display antisense promoters (ASP). In *Drosophila*, the *F* element possesses two convergently-oriented promoters within its 5′ UTR, which overlap by approximately 100 bp (Contursi et al., 1993). The full-length human *L1* 5′ UTR also displays an antisense promoter, which can drive the transcription of adjacent cellular genes in the opposite direction (Speck, 2001). Transcription-factor binding sites for both promoters overlap in the 5′ UTR region: two

binding sites for SOX factors involved in the *L1* sense transcription are located in the region from nt 400 to 600, as is an enhancer element containing binding sites for Ets and Sp1 factors, capable of functioning in either orientation (Speek, 2001). Sequences critical for ASP activity are located in a small region extending from nt 399 to 467, and a 15-bp region located from nt 544 to 558, containing two Sp1 binding sites, may also contribute to the overall *L1* ASP activity (Speek, 2001). Chimeric transcripts derived from the *L1* ASP are highly represented in expressed sequence tag (EST) datasets, from both normal and tumor cells, and the ASP can be located between 1 kb and more than 60 kb from the protein coding sequences (Speek, 2001; Nigumann et al., 2002). It is therefore believed that hypermethylation of *L1* is a major defense mechanism, repressing the oncogenes that use this ASP as an alternative promoter, and which could be very damaging for the organism (Roman-Gomez et al., 2005).

### 3.2. Structure of LTR retrotransposon regulatory regions

The regulatory region of an LTR retrotransposon consists of the 5' LTR and frequently also of a 5' UTR, also known as the ULR (Untranslated Leader Region). The LTRs contain the promoter necessary for transcription and specify the terminator and polyadenylation signals needed for RNA processing. The LTRs consist of three discrete regions: U3 (the nearest to 5', up to the transcription start site), R (for 'repeat', extending from the transcription start site to the polyadenylation site), and U5 (the nearest to 3', downstream from the polyadenylation site). During reverse transcription, the R and U5 regions at the 5' end of the RNA template are reverse transcribed first, followed by a template switch to the 3' end of the RNA, where the reverse-transcribed R region of the DNA binds to the 3' R region of the RNA template (Varmus, 1988). The transcription start signal is located in U3, and the U5 region can have a silencing effect on transcription (Prudhomme et al., 2004). Some LTR retrotransposons have two TATA boxes, such as the *BARE-1* element in barley (Manninen and Schulman, 1993), both of which are able to direct RNA transcription, but which do so under different conditions (Suoniemi et al., 1996). In contrast, a substantial proportion of elements can be TATA-less, such as *mdg1* and *gypsy* in *Drosophila* (Arkhipova et al., 1995). Bidirectional transcription is also possible for LTR retrotransposons. The *Drosophila mdg1* element displays multiple start sites for the antisense direction (Arkhipova and Ilyin, 1991), and the LTRs of some families of HERVs have also been shown to be able to activate transcription in both directions, a characteristic probably linked to the presence of Sp1 binding sites (Dunn et al., 2006; Feuchter and Mager, 1990). Kovalskaya et al. (2006) found transcripts promoted by the human *HERV-K (HML 2)* provirus and solitary LTRs, the transcriptional start point of which is located at the extreme 3' end of the R region of the LTR. The R region is therefore excluded from transcripts initiated on LTRs, whereas a classical retroviral life-cycle model implies that the transcription is driven from between the LTR U3 and R elements. Kovalskaya et al. (2006) propose that a shift from the start site could be explained by the presence of at least

two alternative promoters, one of which is used for viral gene expression, and the other for the transcription of retrotransposition-competent copies of the integrated provirus.

The 5' UTR regions of LTR retrotransposons are multifunctional, and are involved in the transcriptional enhancement of the elements (Wilson et al., 1998; Smith and Corces, 1995), in the isolation of transcriptional units from the neighboring regulatory elements by means of an insulator, such as in the *Idefix* element of *Drosophila* (Conte et al., 2002a), and in the interactions with the nuclear matrix via a scaffold-attachment region (Nabirochkin et al., 1998, for the *Drosophila gypsy* element). The 5' UTR of *Idefix* in *D. melanogaster* has also been shown to exhibit internal ribosome entry site (IRES) activity that enables it to promote the translation of a downstream cistron in a cap-independent manner (Meignin et al., 2003). Recently, an IRES upstream of each ORF has been found in the mouse *L1* element (Li et al., 2006).

### 3.3. Variability of the TE regulatory regions

The regulatory regions of retrotransposons are functionally very important. For instance, 10 point mutations in the U3 region of the human *ERVWE1* element are enough to make 5' and 3' LTRs display significant different promoter activities (Prudhomme et al., 2004). Despite this, regulatory regions of retrotransposons are often reported to be some of the most rapidly evolving regions in the sequence of the elements (Jordan and McDonald, 1998). In particular, the U3 sequence has been shown to be the most variable region of LTR retrotransposons, as for example in the tobacco *Tnt1* element (Casacuberta et al., 1995). The *Retrolyc1* LTR retrotransposon from the plant *Lycopersicon* displays high levels of diversity in its U3 region, with a 53-bp motif repeated up to four times in tandem, the number of repeats having a significant impact on the transcription potential of the element (Araujo et al., 2001).

In the 5' UTR, the regulatory region of the *copla* LTR retrotransposon of *Drosophila* displays multiple copies of an 8-bp motif (TTGTGAAA), with similarity to the core sequence of the SV40 enhancer (McDonald et al., 1997). Naturally-occurring variations in the number of these motifs are correlated with the enhancer strength of the 5' UTR (Matyunina et al., 1996). Costas et al. (2001) PCR-amplified the 5' LTR–UTR region of the *blood* LTR retrotransposon in some species of the *D. melanogaster* subgroup, and revealed length variability corresponding to a 49-bp gap at the end of the 5' LTR, and two gaps of 13 and 49 bp, respectively, in the UTR. Two classes of variants were subsequently identified, one of them being mainly represented by heterochromatic elements (L class), and the other corresponding to active, euchromatic elements (S class). In natural populations of *D. simulans*, the *412* LTR retrotransposon displays variability in sequences of its regulatory region, resulting from nucleotide substitutions, indels, and duplications of motifs, and also, in some populations, some chimerical elements, showing similarity to *412* in some regions and to *mdg1* in others, were found (Mugnier et al., 2005).

All mammalian *L1* LINES have very similar coding regions, but unrelated 5' UTRs and promoters. Five subfamilies have

therefore been classified in the human genome, in which the most active copies are the most recent. In the human genome, even though *L1* copies of different subfamilies all belong to the same lineage, there are striking differences in the transcription factor binding sites found in the 5' UTRs. The active subfamily may have new binding sites that the pre-existing ones did not, such as the SRY-related transcription factor binding sites and the RUNX3 binding sites, which are essential for transcription activation and therefore allow the younger *L1* copies to be active (Khan et al., 2006).

It is therefore obvious that diversity exists in the retrotransposon regulatory region, and many examples are found in the literature. We will now discuss the mechanisms that may have led to this enormous variability in regulatory region sequences, which are quite different between LTR retrotransposons and LINES.

#### 4. Evolution of the regulatory region

##### 4.1. Retrotransposition, an error-prone mechanism

The lack of any proofreading/repair activity of RNA polymerase and reverse transcriptase means that the replication of retrotransposons (LTR and non-LTR) is a very error-prone mechanism. As a consequence, the replication of a single retrotransposon can generate a population of closely related, but not identical sequences resembling the 'quasi-species' populations of RNA viruses (Domingo et al., 1985; Casacuberta et al., 1995). In particular, secondary structures and direct repeats, as well as runs of identical nucleotides, can considerably increase the mutation rate (Pathak and Temin, 1990, 1992). Repeated motifs are often bordered by runs of T's (McDonald et al., 1997), which are known to facilitate template slippages during reverse transcription (Burns and Temin, 1994). McDonald et al. (1997) have proposed that, since reverse transcription is prone to generating short regional duplications, which are characteristic of eukaryotic enhancers, LTR retrotransposons may play a role in the evolution of these enhancers.

##### 4.2. Recombination in LTR retrotransposons

Recombination is a major mechanism generating variability in LTR retrotransposons. Frequent recombination events are known to occur between the two genomic RNA strands packaged within the LTR retroelement capsids (Zhang and Temin, 1994). Jordan and McDonald (1998) found evidence for recombination by template switching between *Ty1* and *Ty2* elements of *Saccharomyces cerevisiae*. Template switching is thought to occur within the R region of the LTR, and these authors did indeed find elements containing hybrid LTRs, in which U3 regions displayed phylogenetic patterns different from those of the R and U5 regions. In natural populations of *D. simulans*, the regulatory region of some *412* elements has been found to result from a recombination of a *412*-LTR and a *mgd1*-5' UTR (Mugnier et al., 2005). Recombination has also been found in the *BARE* LTR retrotransposon in barley, involving *BARE* and *Wis-2* elements (Vicent et al., 2005). In mice, recombination events

between different ancient *L1* copies have engendered new 5' UTR active *L1* copies. For instance, the *G<sub>F</sub>* subfamily possesses different monomers from different sequences of the *F* subfamily, and a new, non-monomeric region (Goodier et al., 2001). Modular evolution (Lerat et al., 1999), which results from recombination events occurring between the regulatory regions of two elements, could be a quick and efficient way to change expression patterns and the regulatory interacting system, thus allowing new variants to evade regulation by the host (Mugnier et al., 2005).

##### 4.3. Recruitment of novel regulatory regions in LINES

Due to their different retrotransposition mechanisms, LINES are not affected by recombination the way LTR retrotransposons are. On the contrary, in the human *L1* family, novel regulatory regions (5' UTR) have frequently been recruited during the evolution of the family, whereas the two open reading frames have remained relatively conserved (Khan et al., 2006). The 5' UTR of the ancestral *L1* primate subfamily has been replaced at least eight times over the last 70 My. *L1* families with different 5' UTRs can coexist for a period of time without competing, because they do not rely on the same host-encoded factors for their transcription. The most powerful 5' UTR would eventually generate a mobilization burst and would therefore give rise to a new *L1* subfamily (Khan et al., 2006). Over the course of evolution, each of these subfamilies has been replaced by a new active one, and consequently all active *L1* copies belong to a single lineage (Boissinot and Furano, 2005). The same life-cycle pattern has been identified in mice *L1* elements, but several recombination events have complicated the phylogenetic tree (Mears and Hutchison, 2001). In contrast to the situation in humans, active *Drosophila* LINE elements belonging to different lineages can be found coexisting at the same time. In the zebrafish genome, *L1* elements also behave totally differently from their counterparts in the human genome: the 5' UTRs have different sequences, and therefore the various *L1* elements do not compete with each other for host transcriptional factors. The zebrafish genome thus contains more than 30 distinct *L1* lineages (Furano et al., 2004).

Some authors have also proposed that, since mammalian *L1* elements have unrelated 5' UTRs or promoters, it is possible that *L1* promoters were originally derived from host genes located close to *L1* insertion sites by a promoter-capturing mechanism (Speck, 2001).

#### 5. Impacts on the host genome

##### 5.1. Interference with host genome performance

Since the retrotransposon regulatory region displays promoter behavior, any insertion of a retrotransposon into the genome could affect the transcription of nearby host genes. In particular, the regulatory region could promote ectopic expression or simply modify the expression level of the host genes (Borie et al., 2000). For instance, it has been demonstrated that an LTR can bidirectionally promote the transcription of two human

genes (Dunn et al., 2006). Ectopic transcription by *L1* ASP can also be responsible for transcriptional interference and posttranscriptional silencing of cellular genes (Nigumann et al., 2002). van de Lagemaat et al. (2003) have illustrated that a major impact of TEs is their ability to induce changes in gene regulation without destroying existing gene functions. In particular, these authors showed that TE regulatory regions act as alternative promoters of many genes, therefore modifying expression patterns or tissue-specificity. LTR elements, which carry more transcription regulating signals than LINES, are also less often found in gene promoter regions, probably because a high number of regulatory signals are more likely to alter gene expression in a greater extent and have deleterious effects (Thornburg et al., 2006). Most of the genes exhibiting TE regulatory regions in their promoters are involved in functions such as the stress response, immunity, and response to external stimuli, whereas genes playing a role in development and metabolism are less likely to have TE promoters inserted within their UTRs. In addition, Mariño-Ramírez et al. (2005) have shown that the regulatory sequences contributed by TEs are exceptionally lineage specific, suggesting that TEs can drive the diversification of gene regulation between evolutionary lineages. TEs can interfere with host genes, not only by promoting their transcription, but also by introducing into their 5' UTRs novel regulatory sequences, which may act as enhancers or insulators, as in the case of the *D. melanogaster* elements *ZAM* and *Idefix*, respectively, which modify the proper regulation of the *white* gene in the eyes (Conte et al., 2002a). Promoter competition between host genes and retrotransposons has also been shown to be a mechanism of transcriptional interference, disrupting endogenous enhancer–promoter communications (Conte et al., 2002b). The rapid evolution of the regulatory region of TEs can therefore have a wide panel of significant effects on the way the genome works.

## 5.2. Inter-element competition and new waves of retrotransposition

Generation of new variants of retrotransposon regulatory regions is a mechanism by which competition between elements can occur, the genome being invaded by the ‘best-adapted’ element. There can also be competition between different groups of copies (highly active, less active, non-autonomous) without elimination of the less active or non-autonomous elements (Fig. 3) (Deceliere et al., 2005; Le Rouzic and Capy, 2005). Variants of the *1731* retrotransposon in *D. melanogaster*, which display both extended transcriptional profiles due to changes in the LTR sequence, and altered translational strategy due to loss of frameshifting, have supplanted the more ancient forms (Kalmykova et al., 2004). Costas et al. (2001) found two classes of variants of the regulatory region in the *blood* LTR retrotransposon, which they suggest may compete with each other. The active *blood* subfamily (S) transposes at a faster rate, and therefore survives and propagates, whereas the other subfamily (L) is condemned to disappear or to remain confined within the heterochromatin. Inter-element competition has also been postulated for the two subfamilies of the *tirant* LTR retrotransposon in *D. melanogaster* and *D. simulans*: one subfamily, the C-type, is active and has invaded the chromosome arms, whereas transcription is impeded in another subfamily, the S-type, which is confined to the heterochromatin and has a very low copy number (Fablet et al., 2006). Host genomic constraints also interfere with inter-element competition, since, unlike *D. melanogaster* and *D. simulans*, *D. mauritiana* and *D. sechellia* do not display any expansion of the S subfamily of *blood* in their genomes (Costas et al., 2001). To generate new characters in this inter-element competition, inter-element recombination may, for instance, be an effective strategy by which retroelements can rapidly evolve novel regulatory sequence combinations (Jordan

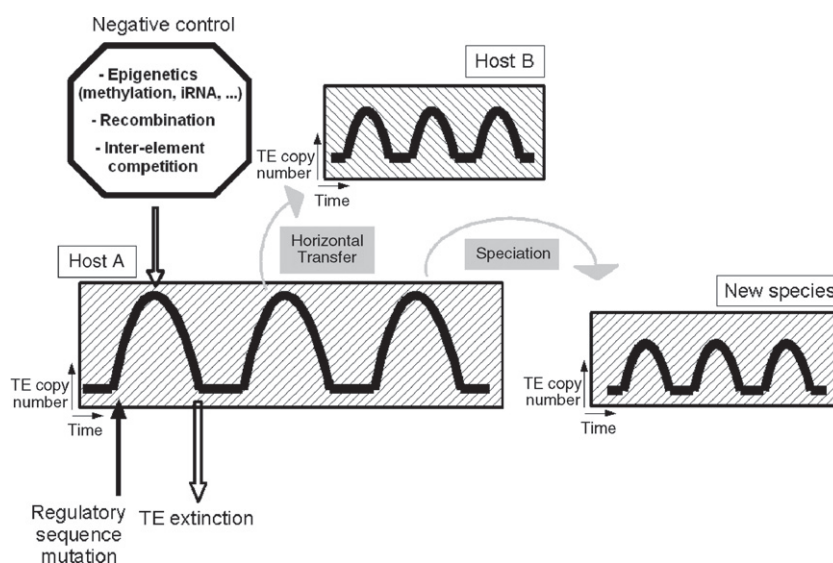


Fig. 3. The influence of regulatory sequences in the TE life cycle. The black curve represents the evolution of TE copy number in a given host genome. TE regulatory sequences can be subject to negative control by the host genome, leading to a decrease in TE copy number and even TE extinction. Depending on its regulatory mechanisms and the host response, a TE can display waves of increasing and decreasing copy number. The evolution of the regulatory region may allow the TE to invade a new host, in which the TE will once again be subjected to successive waves of increasing and decreasing copy number.

and McDonald, 1998). What has happened to the *L1* element in the human genome is an extreme form of this inter-element competition, since virtually only one lineage is active at a time (Khan et al., 2006).

The rules governing the inter-element competition are mainly determined by the host. A high degree of variability within transcriptional regulatory regions could therefore allow retrotransposons to explore transcriptional patterns in order to facilitate their coexistence with their host genome. Consistent with this hypothesis, *Tnt1* elements within various plant host species tend to develop different U3 regulatory sequences that are adapted to each host genome (Araujo et al., 2001). Cooperation between cellular and retroviral elements is another way of creating variability in the regulation of a retrotransposon. One unusual example is that of the *ERVWE1* element, the expression of which is regulated by a bipartite element consisting of a cyclic AMP-inducible LTR retroviral promoter adjacent to a cellular enhancer conferring a high level of expression and placental tropism (Prudhomme et al., 2004).

The innovation race of the regulatory region engenders a close relationship between the host factors and its transposable elements. This can result either in an adaptation between the transposable elements and their host genome, or a new burst of transposition and genome invasion. But according to the genome and type of retrotransposon considered (LTR/non-LTR), the modalities and outcomes of these evolutionary mechanisms are quite different.

## Acknowledgements

We thank Monika Ghosh for reviewing the English text. This work was funded by the Centre National de la Recherche Scientifique (UMR 5558 and GDR 2157 on Transposable Elements).

## References

- Araujo, P.G., Casacuberta, J.M., Costa, A.P.P., Hashimoto, R.Y., Grandbastien, M.A., Van Sluys, M.A., 2001. Retrolyc1 subfamilies defined by different U3 LTR regulatory regions in the *Lycopersicon* genus. *Mol. Genet. Genomics* 266, 35–41.
- Arkipova, I.R., 1995. Complex patterns of transcription of a *Drosophila* retrotransposon *in vivo* and *in vitro* by RNA polymerases II and III. *Nucleic Acids Res.* 23, 4480–4487.
- Arkipova, I.R., Ilyin, Y.V., 1991. Properties of promoter regions of *mdg1* *Drosophila* retrotransposons indicate that it belongs to a specific class of promoters. *EMBO J.* 10, 1169–1177.
- Arkipova, I.R., Lyubomirskaya, N.V., Ilyin, Y.V., 1995. *Drosophila* Retrotransposons. Springer, Berlin. pp 8–9, 41–46.
- Bannert, N., Kurth, R., 2004. Retroelements and the human genome: new perspectives on an old relation. *Proc. Natl. Acad. Sci. U. S. A.* 5, 14572–14579.
- Belshaw, R., et al., 2004. Long-term reinfection of the human genome by endogenous retroviruses. *Proc. Natl. Acad. Sci. U. S. A.* 101, 4894–4899.
- Benit, L., Lallemand, J.B., Casella, J.F., Philippe, H., Heidmann, T., 1999. ERV-L elements: a family of endogenous retrovirus-like elements active throughout the evolution of mammals. *J. Virol.* 73, 3301–3308.
- Bennetzen, J.L., 2000. Transposable element contributions to plant gene and genome evolution. *Plant Mol. Biol.* 42, 251–269.
- Biémont, C., Vieira, C., 2005. What transposable elements tell us about genome organization and evolution? The case of *Drosophila*. *Cytogenet. Genome Res.* 110, 25–34.
- Biémont, C., Tsitrone, A., Vieira, C., Hoogland, C., 1997. Transposable element distribution in *Drosophila*. *Genetics* 147, 1997–1999.
- Boissinot, S., Furano, A.V., 2005. The recent evolution of human *L1* retrotransposons. *Cytogenet. Genome Res.* 110, 402–406.
- Borie, N., Loevenbruck, C., Biémont, C., 2000. Developmental expression of the 412 retrotransposon in natural populations of *D. melanogaster* and *D. simulans*. *Genet. Res.* 76, 217–226.
- Burke, T.W., Kadonaga, J.T., 1996. *Drosophila* TFIID binds to a conserved downstream basal promoter element that is present in many TATA-box-deficient promoters. *Genes Dev.* 10, 711–724.
- Burns, D.P., Temin, H.M., 1994. High rates of frameshift mutations within homo-oligomeric runs during a single cycle of retroviral replication. *J. Virol.* 68, 4196–4203.
- Casacuberta, J.M., Vernhettes, S., Grandbastien, M.A., 1995. Sequence variability within the tobacco retrotransposon *Tnt1* population. *EMBO J.* 14, 2670–2678.
- Charlesworth, B., Langley, C.H., Sniegowski, P.D., 1997. Transposable element distribution in *Drosophila*. *Genetics* 147, 1993–1995.
- Conte, C., Dastugue, B., Vauzy, C., 2002a. Coupling of enhancer and insulator properties identified in two retrotransposons modulate their mutagenic impact on nearby genes. *Mol. Cell. Biol.* 22, 1767–1777.
- Conte, C., Dastugue, B., Vauzy, C., 2002b. Promoter competition as a mechanism of transcriptional interference mediated by retrotransposons. *EMBO J.* 21, 3908–3916.
- Contursi, C., Minchiotti, G., DiNocera, P.P., 1993. Functional dissection of two promoters that control sense and antisense transcription of *Drosophila melanogaster* *F* elements. *J. Mol. Biol.* 234, 988–997.
- Cooper, D.M., Schimenti, K.J., Schimenti, J.C., 1998. Factors affecting ectopic gene conversion in mice. *Mamm. Genome* 277, 513–517.
- Costas, J., Valadé, E., Naveira, H., 2001. Amplification and phylogenetic relationships of a subfamily of *blood*, a retrotransposable element of *Drosophila*. *J. Mol. Evol.* 52, 342–350.
- Deceliere, G., Charles, S., Biémont, C., 2005. The dynamics of transposable elements in structured populations. *Genetics* 169, 467–474.
- Domingo, E., et al., 1985. The quasispecies (extremely heterogeneous) nature of viral RNA genome populations: biological relevance — a review. *Gene* 40, 1–8.
- Dunn, C.A., Romanish, M.T., Gutierrez, L.E., van de Lagemaat, L.N., Mager, D.L., 2006. Transcription of two human genes from a bidirectional endogenous retrovirus promoter. *Gene* 366, 335–342.
- Eikbush, T.H., Furano, A.V., 2002. Fruit flies and humans respond differently to retrotransposons. *Curr. Opin. Genet. Dev.* 12, 669–674.
- Fablet, M., McDonald, J.F., Biémont, C., Vieira, C., 2006. Ongoing loss of the tirant transposable element in natural populations of *Drosophila simulans*. *Gene* 375, 54–62.
- Feuchter, A., Mager, D.L., 1990. Functional heterogeneity of a large family of human LTR-like promoters and enhancers. *Nucleic Acids Res.* 18, 1261–1270.
- Furano, A.V., Duvernell, D.D., Boissinot, S., 2004. L1 (LINE-1) retrotransposon diversity differs dramatically between mammals and fish. *Trends Genet.* 20, 9–14.
- Goodier, J.L., Ostertag, E.M., Du, K., Kazazian Jr., H.H., 2001. A novel active L1 retrotransposon subfamily in the mouse. *Genome Res.* 11, 1677–1685.
- Hoskins, R.A., et al., 2002. Heterochromatic sequences in *Drosophila* whole-genome shotgun assembly. *Genome Biol.* 3, 0085.1–0085.16.
- Hua-Van, A., Le Rouzic, A., Maisonhaute, C., Capy, C., 2005. Abundance, distribution and dynamics of retrotransposable elements and transposons: similarities and differences. *Cytogenet. Genome Res.* 110, 426–440.
- International Human Genome Sequencing Consortium, 2001. Initial sequencing and analysis of the human genome. *Nature* 409, 860–921.
- Jordan, I.K., McDonald, J.F., 1998. Evidence for the role of recombination in the regulatory evolution of *Saccharomyces cerevisiae* *Ty* elements. *J. Mol. Evol.* 47, 14–20.
- Jordan, I.K., Rogoniz, I.B., Glazko, G.V., Koonin, E.V., 2003. Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends Genet.* 19, 68–72.
- Kalmykova, A.I., et al., 2004. Selective expansion of newly evolved genomic variants of retrotransposon *1731* in the *Drosophila* genomes. *Mol. Biol. Evol.* 21, 2281–2289.
- Kaminker, J.S., et al., 2003. The transposable elements of the *Drosophila melanogaster* euchromatin: a genomic perspective. *Genome Biol.* 3, 0084.1–0084.2.

- Kapitonov, V.V., Jurka, J., 2003. Molecular paleontology of transposable elements in the *Drosophila melanogaster* genome. Proc. Natl. Acad. Sci. U. S. A. 100, 6569–6574.
- Kazazian Jr., H.H., 1998. Mobile elements and disease. Curr. Opin. Genet. Dev. 8, 343–350.
- Khan, H., Smit, A., Boissinot, S., 2006. Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. Genome Res. 16, 78–87.
- Kidwell, M.G., Lisch, D.R., 2000. Transposable elements and host genome evolution. Trends Ecol. Evol. 15, 95–99.
- Kovalskaya, E., Buzdin, A., Gogvadze, E., Vinogradova, T., Sverdlov, E., 2006. Functional human endogenous retroviral LTR transcription start sites are located between the R and U5 regions. Virology 346, 373–378.
- Kunert, N., Marhold, J., Stanke, J., Stach, D., Lyko, F., 2003. A Dnmt2-like protein mediates DNA methylation in *Drosophila*. Development 130, 5083–5090.
- Kurose, K., Hata, K., Hattori, M., Sakaki, Y., 1995. RNA polymerase III dependence of the human L1 promoter and possible participation of the RNA polymerase II factor YY1 in the RNA polymerase III transcription system. Nucleic Acids Res. 23, 3704–3709.
- Lavie, L., Kitova, M., Maldener, E., Meese, E., Mayer, J., 2005. CpG methylation directly regulates transcriptional activity of the human endogenous retrovirus family HERV-K (HML-2). J. Virol. 79, 876–883.
- Lavie, L., Maldener, E., Brouha, B., Meese, E.U., Mayer, J., 2006. The human L1 promoter: variable transcription initiation sites and a major impact of upstream flanking sequence on promoter activity. Genome Res. 14, 2253–2260.
- Lerat, E., Brunet, F., Bazin, C., Capy, P., 1999. Is the evolution of transposable elements modular? Genetica 107, 15–25.
- Lerat, E., Rizzon, C., Biémont, C., 2003. Sequence divergence within transposable element families in the *Drosophila melanogaster* genome. Genome Res. 13, 1889–1896.
- Le Rouzic, A., Capy, P., 2005. The first steps of transposable elements invasion: parasitic strategy vs. genetic drift. Genetics 169, 1033–1043.
- Li, P.W., Li, J., Timmerman, S.L., Krushel, L.A., Martin, S.L., 2006. The dicistronic RNA from the mouse LINE-1 retrotransposon contains an internal ribosome entry site upstream of each ORF: implications for retrotransposition. Nucleic Acids Res. 34, 853–864.
- Ludwig, M.Z., Bergman, C., Patel, N.H., Kreitman, M., 2000. Evidence for stabilizing selection in a eukaryotic enhancer element. Nature 403, 564–567.
- Lyko, F., 2001. DNA methylation learns to fly. Trends Genet. 17, 169–172.
- Lyko, F., Ramsahoye, B.H., Jaenisch, R., 2000. DNA methylation in *Drosophila melanogaster*. Nature 408, 538–540.
- Manninen, I., Schulman, A.H., 1993. *BARE-1*, a copia-like retroelement in barley (*Hordeum vulgare* L.). Plant Mol. Biol. 22, 829–846.
- Mariño-Ramírez, L., Lewis, K.C., Landsman, D., Jordan, I.K., 2005. Transposable elements donate lineage-specific regulatory sequences to host genomes. Cytogenet. Genome Res. 110, 333–341.
- Matyunina, L.V., Jordan, I.K., McDonald, J.F., 1996. Naturally occurring variation in *copia* expression is due to both element (cis) and host (trans) regulatory variation. Proc. Natl. Acad. Sci. U. S. A. 93, 7097–7102.
- Mayer, J., Sauter, M., Racz, A., Scherer, D., Mueller-Lantzsch, N., Meese, E., 1999. An almost-intact human endogenous retrovirus K on human chromosome 7. Nat. Genet. 21, 257–258.
- McDonald, J.F., Matyunina, L.V., Wilson, S., Jordan, I.K., Bowen, N.J., Miller, W.J., 1997. LTR retrotransposons and the evolution of eukaryotic enhancers. Genetica 100, 3–13.
- Mears, M.L., Hutchison, C.A., 2001. The evolution of modern lineages of mouse L1 elements. J. Mol. Evol. 52, 51–62.
- Meignin, C., Bailly, J.L., Arnaud, F., Dastugue, B., Vaury, C., 2003. The 5' untranslated region and Gag product of *Idefix*, a long terminal repeat-retrotransposon from *Drosophila melanogaster*, act together to initiate a switch between translated and untranslated states of the genomic mRNA. Mol. Cell. Biol. 23, 8246–8254.
- Menendez, L., Benigno, B.B., McDonald, J.F., 2004. L1 and HERV-W retrotransposons are hypomethylated in human ovarian carcinomas. Mol. Cancer 3, 12–16.
- Minchiotti, G., Contursi, C., Di Nocera, P.P., 1997. Multiple downstream promoter modules regulate the transcription of the *Drosophila melanogaster* *I*, *doc* and *F* elements. J. Mol. Biol. 267, 37–46.
- Mugnier, N., Biémont, C., Vieira, C., 2005. New regulatory regions of *Drosophila* 412 retrotransposable element generated by recombination. Mol. Biol. Evol. 22, 747–757.
- Nabirochkin, S., Ossokina, M., Heidmann, T., 1998. A nuclear matrix/scaffold attachment region co-localizes with the *gypsy* retrotransposon insulator sequence. J. Biol. Chem. 273, 11899–11906.
- Nigumann, P., Redik, K., Mätlik, K., Speck, M., 2002. Many human genes are transcribed from the antisense promoter of L1 retrotransposon. Genomics 79, 628–634.
- Patel, C.V., Gopinathan, K.P., 1987. Determination of trace amounts of 5-methylcytosine in DNA by HPLC. Anal. Biochem. 164, 164–169.
- Pathak, V.K., Temin, H.M., 1990. Broad spectrum of *in vivo* forward mutations, hypermutations, and mutational hotspots in a retroviral shuttle vector after a single replication cycle: substitutions, frameshifts, and hypermutations. Proc. Natl. Acad. Sci. U. S. A. 87, 6019–6023.
- Pathak, V.K., Temin, H.M., 1992. 5-azacytidine and RNA secondary structure increase the retrovirus mutation rate. J. Virol. 66, 3093–3100.
- Prudhomme, S., Oriol, G., Mallet, F., 2004. A retroviral promoter and a cellular enhancer define a bipartite element which controls env ERVWE1 placental expression. J. Virol. 78, 12157–12168.
- Roman-Gomez, J., et al., 2005. Promoter hypomethylation of the LINE-1 retrotransposable elements activates sense/antisense transcription and marks the progression of chronic myeloid leukemia. Oncogene 24, 7213–7223.
- Smith, P.A., Corces, V.G., 1995. The suppressor of Hairy-wing protein regulates the tissue-specific expression of the *Drosophila gypsy* retrotransposons. Genetics 139, 215–228.
- Speck, M., 2001. Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. Mol. Cell. Biol. 21, 1973–1985.
- Suoniemi, A., Narvanto, A., Schulman, A.H., 1996. The *BARE-1* retrotransposon is transcribed in barley from an LTR promoter active in transient assays. Plant Mol. Biol. 31, 295–306.
- Sverdlov, E.D., 2000. Retroviruses and primate evolution. Bioessays 22, 161–171.
- Thornburg, B.G., Gotea, V., Makalowski, W., 2006. Transposable elements as a significant source of transcription regulating signals. Gene 365, 104–110.
- Torti, C., et al., 2000. Evolution of different subfamilies of *mariner* elements within the medfly genome inferred from abundance and chromosomal distribution. Chromosoma 108, 523–532.
- Turcotte, K., Srinivasan, S., Bureau, T., 2001. Survey of transposable elements from rice genomic sequences. Plant J. 25, 169–179.
- Tweedie, S., Charlton, J., Clark, V., Bird, A., 1997. Methylation of genomes and genes at the invertebrate-vertebrate boundary. Mol. Cell. Biol. 17, 1469–1475.
- van de Lagemaat, L., Landry, J.R., Mager, D., Medstrand, P., 2003. Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. Trends Genet. 19, 530–536.
- van de Lagemaat, L.N., Gagnier, L., Medstrand, P., Mager, D.L., 2005. Genomic deletions and precise removal of transposable elements mediated by short identical DNA segments in primates. Genome Res. 15, 1243–1249.
- Varmus, H., 1988. Retroviruses. Science 240, 1427–1435.
- Vicient, C.M., Kalendar, R., Schulman, A.H., 2005. Variability, recombination, and mosaic evolution of the barley *BARE-1* retrotransposon. J. Mol. Evol. 61, 275–291.
- Vieira, C., Biémont, C., 2004. Transposable element dynamics in two sibling species: *Drosophila melanogaster* and *Drosophila simulans*. Genetica 120, 115–123.
- Vieira, C., Lepetit, D., Dumont, S., Biémont, C., 1999. Wake up of transposable elements following *Drosophila simulans* worldwide colonization. Mol. Biol. Evol. 16, 1251–1255.
- Vieira, C., Nardon, C., Arpin, C., Lepetit, D., Biémont, C., 2002. Evolution of genome size in *Drosophila*. Is the invader's genome being invaded by transposable elements? Mol. Biol. Evol. 19, 1154–1161.
- Wilson, S., Matyunina, L.V., McDonald, J.F., 1998. An enhancer region within the *copia* untranslated leader contains binding sites for *Drosophila* regulatory proteins. Gene 209, 239–246.
- Zhang, J., Temin, H.M., 1994. Retrovirus recombination depends on the length of sequence identity and is not error prone. J. Virol. 68, 2409–2414.

## CONCLUSION

In this bibliographic work, we have shown that retrotransposons differ in their regulatory sequences (from LTR elements to LINE-like elements) and, as a consequence, impact differently the genomes in which they reside. Also, we have shown that similar elements behave differently according to the species analysed. All these data suggest that a specific relationship exists between TEs and each invaded genome. Several parameters might influence the genome/TE relationship : TE families, species genomes, environment and ecological traits are described to influence TE/genome relationship (Waterland and Jirtle, 2003; 2004; Clark et al., 2007; Dramard et al., 2007; Carbone et al., 2009). Therefore, one way to understand such specificity is to decrease the number of different parameters involved in the equation TE/genome. The study of natural populations belonging to the same species decreases the disparities between genomic environments, since higher genetic homologies can be observed intra-species than between species, and allows for a simpler view of TE/genome interactions. *D. melanogaster* and *D. simulans*, as described above, have different amounts of repeats and TEs (~15% and ~5% of TEs respectively) (Dowsett and Young, 1982; Vieira et al., 1999). Such pattern can be observed among TE families that are often overrepresented in *D. melanogaster* natural populations (Vieira et al., 1999). There are a few elements that are outnumbered in natural populations of *D. simulans* and therefore considered as exceptions. One of them is the LINE-like element *helena*. We decided to make a survey of this uncommon *helena* element in natural populations of both *D. melanogaster* and *D. simulans*. The goal of this study is firstly to understand two different behaviours of one same element in two close species. Secondly, the use of natural populations allows us to test intra-specific natural variation in TE regulatory mechanisms. Surprisingly, the supposed active *helena* element in *D. simulans* natural populations is the target of an important internal deletion mechanism. Such system seems to be absent from *D. melanogaster* natural populations, and unpublished data by E. Lerat generalize it to most of the other *D. simulans* elements.

Research article

Open Access

## Losing *helena*: The extinction of a drosophila line-like element

Rita Rebollo<sup>1</sup>, Emmanuelle Lerat<sup>1</sup>, Liliana Lopez Kleine<sup>1,2</sup>,  
Christian Biémont<sup>1</sup> and Cristina Vieira\*<sup>1</sup>

Address: <sup>1</sup>Université de Lyon; Université Lyon 1; CNRS; UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, Villeurbanne F-69622, France and <sup>2</sup>UR477 de Biochimie Bactérienne, UR341 de Mathématiques et informatiques Appliquées, INRA. 78352 Jouy en Josas, France

Email: Rita Rebollo - rebollo@biomserv.univ-lyon1.fr; Emmanuelle Lerat - lerat@biomserv.univ-lyon1.fr; Liliana Lopez Kleine - Liliana.Lopez@jouy.inra.fr; Christian Biémont - biemont@biomserv.univ-lyon1.fr; Cristina Vieira\* - vieira@biomserv.univ-lyon1.fr

\* Corresponding author

Published: 31 March 2008

Received: 31 October 2007

BMC Genomics 2008, 9:149 doi:10.1186/1471-2164-9-149

Accepted: 31 March 2008

This article is available from: <http://www.biomedcentral.com/1471-2164/9/149>

© 2008 Rebollo et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

### Abstract

**Background:** Transposable elements (TEs) are major players in evolution. We know that they play an essential role in genome size determination, but we still have an incomplete understanding of the processes involved in their amplification and elimination from genomes and populations. Taking advantage of differences in the amount and distribution of the Long Interspersed Nuclear Element (LINE), *helena* in *Drosophila melanogaster* and *D. simulans*, we analyzed the DNA sequences of copies of this element in samples of various natural populations of these two species.

**Results:** *In situ* hybridization experiments revealed that *helena* is absent from the chromosome arms of *D. melanogaster*, while it is present in the chromosome arms of *D. simulans*, which is an unusual feature for a TE in these species. Molecular analyses showed that the *helena* sequences detected in *D. melanogaster* were all deleted copies, which diverged from the canonical element. Natural populations of *D. simulans* have several copies, a few of them full-length, but most of them internally deleted.

**Conclusion:** Overall, our data suggest that a mechanism that induces internal deletions in the *helena* sequences is active in the *D. simulans* genome.

### Background

Genome evolution occurs by several processes, including global genome duplications, segmental duplications and the amplification/deletion of repetitive sequences. Among the repeated sequences, transposable elements (TEs), which constitute a high proportion in many genomes, play an important role in genome evolution [1]. The transposition rates of these TEs depend on the amount and type of the TEs present in the genome; they are not constant over time, but are subject to amplification bursts in certain species and populations [2]. As a

result, genomes contain widely differing amounts of TEs that are not directly correlated to their activity levels. For instance, the human genome is composed of at least 50% of TEs, but only very few are active, and they are responsible for less than 1% of mutations [3]. In contrast, in *Drosophila melanogaster*, only 18% of the genome is composed of TEs, but a high proportion of mutations (more than 50%) is attributable to their transposition [4].

A TE life cycle can be viewed as successive waves of transposition/loss: invasion of the host genome by TEs being

followed by their progressive elimination [5,6]. For example, the *LINE-1* (*L1*) element has colonized the entire human genome by successful waves of transposition [2,7], and today it is the most abundant TE family in this genome. However, in humans, most of the elements have been inactivated either by structural changes or by epigenetic control, such as DNA methylation [7]. In *D. melanogaster*, the *I* factor has recently reinvaded this genome after being lost from the chromosome arms [8]. TE elimination from genomes is therefore a commonly observed phenomenon, although no real-time observation of a TE extinction has ever been reported. As TEs have a considerable influence on remodeling the genome structure [9], we need to understand the dynamics of changes in their copy numbers. One way to investigate these dynamics is to analyze closely related species with differing TE amounts, such as *D. simulans* and *D. melanogaster*. These species diverged 2 to 3 million years ago [10] and have differing proportions of TEs: *D. melanogaster* contains more than 18% of TEs, whereas *D. simulans* contains only 5% [11].

*D. simulans* has fewer copies of most TEs [11], but there are a few exceptions. The DNA-transposon *hobo* is more abundant in *D. simulans*, the retrovirus-like *gypsy* and *ZAM* elements have the same low number of copies in both species, and the LINE-like element *helena* is present in the *D. simulans* genome (10 insertion sites as determined by *in situ* hybridization), but has not been detected in the chromosome arms of *D. melanogaster* [11,12]. The striking distribution of *helena* in natural populations of these two species, and the fact that degenerated copies are found in the sequenced *D. melanogaster* genome, make this LINE-like element an ideal model system to study the real time TE life cycle.

Petrov and colleagues [13] proposed that deletions are common events in *Drosophila*, and based this suggestion on the analysis of partial *helena* sequences from different *Drosophila* species. However, it is difficult to extrapolate this to other TEs, if we take into account the fact that *helena* is one of the few degenerate TEs in the *D. melanogaster* sequenced genome [14]. Comparing closely related species with differing TE amounts, could be used to test the importance of this deletion process in regulating TE genome invasions.

We analyzed the structure and activity of *helena* using the sequenced genomes and of 41 natural populations of *D. melanogaster* and *D. simulans*. We show that the elimination of *helena* from its host genomes is a very quick process, and that it is mediated by massive internal deletions in the element [15]. We conclude that the process of elimination of *helena* is far advanced in *D. melanogaster*. but is still in progress in *D. simulans*.

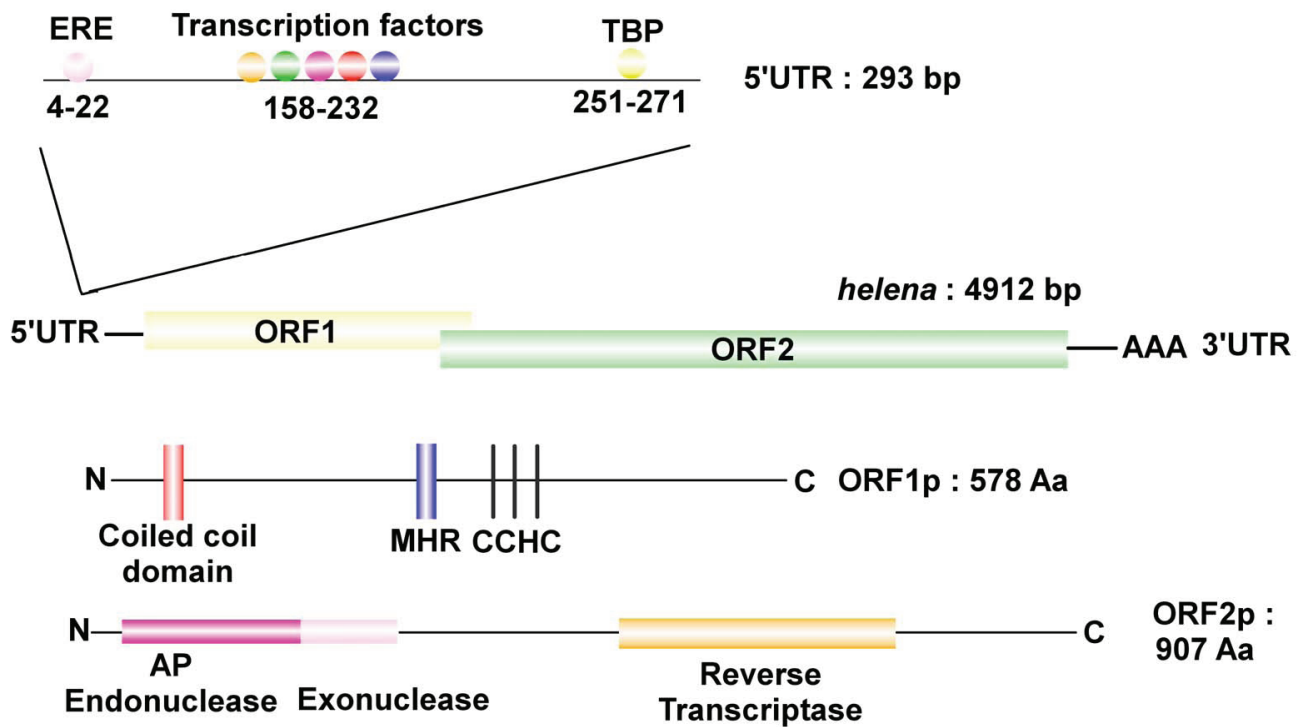
## Results

### In silico identification of a complete copy of *helena* in the *D. simulans* genome

Because no full-length copy of *helena* had previously been described, we performed a bioinformatic search for such a copy in the draft sequence of the *D. simulans* genome [16]. We found a 4912-bp copy of *helena* on the chromosome arm 3R (at position 1506433 – 1511368 on the minus strand) (Figure 1). *Helena* belongs to the *jockey* clade [17], has a 25-bp poly A tail, and two overlapping open reading frames (ORF1 and ORF2). The first ORF is 1737-bp and codes a 579-amino acid (aa) protein that has high similarities to the *gag* protein of other LINE-like elements, such as *X*, *jockey* and *HeT-A* [18,19]. The *gag*-like protein contains the major homology region (MHR), followed by a cysteine-rich domain (CX<sub>2</sub>CX<sub>4</sub>HX<sub>4</sub>C, CX<sub>2</sub>CX<sub>3</sub>HX<sub>4</sub>C, CX<sub>2</sub>CX<sub>3</sub>HX<sub>6</sub>C). This region is common to all *gag*-like proteins, and confers an RNA or DNA single-strain binding property on these elements, as well as being essential for *gag* oligomerization. *Helena* has a coiled-coil domain located in the 5' region of the *gag* protein, something that had previously only been seen in L1 elements from mammals [20] and in some LTR retrotransposons from *Drosophila* [21]. The second ORF, which starts on the last base of ORF1, is 2721-bp, and codes a 907-aa protein corresponding to the *pol* gene, which is very similar to the protein of the *BS* and *jockey* elements. The *pol*-like protein contains all the domains necessary for its function: an apyrimidic endonuclease and an exonuclease (from amino acid 4 to 221), plus a reverse transcriptase domain (from amino acid 493 to 746). Both ORFs are intact, could produce transcripts, and are surrounded by two untranslated regions (5'UTR and 3'UTR respectively). Because the regulatory region is often defined in the 5'UTR [22], we performed a bioinformatic search for transcription factors binding sites in this region. A single region was detected containing several transcription factor binding sites, such as SP1 and upstream stimulating factor-like (USF) binding domains. This region also displays a binding site for a TATA-binding protein (TBP), and an estrogen response element (ERE).

### The copies of *helena* in the sequenced genomes of *D. melanogaster* and *D. simulans*

Using the complete sequence of *helena* as a query, we found 62 *helena* sequences in the *D. simulans* genome (see Additional File 1 for details). Twenty-eight of these copies were located on the chromosome arms, and the remaining 34 were in the U part of the genome, that may correspond to heterochromatin. The copies ranged in size from 107 bp to 5098 bp. However, it is difficult to determine the exact size of some copies due to the presence of numerous undetermined bases. Two copies (chr2R\_13305831 and chrX\_16602314) were longer than the reference sequence due to insertions. In some other



**Figure 1**  
**helena structure.** Full-length copy of *helena* in the *D. simulans* genome (3R: 1506433 – 15011368). DNA sequence: UTR, untranslated region; ORF, open reading frame; AAA, polyA tail. 5'UTR: ERE, estrogen response element; GATA, SPI, stimulating protein I; USF-like, upstream stimulating factor-like; Protein sequences: MHR, major homology region; CCHC, cysteine rich domain; AP, apyrimidic. See Materials and Methods for prediction information.









cases, we may be looking at fragments of the same copy; however, the distances that separate them are too large to allow us to find out with certainty whether they come from the same copy. The estimated number of 62 copies in *D. simulans* may therefore be an overestimation. The average percentage identity is 96.1% for all copies, with an average of 97.4% for the copies in the euchromatin, and of 94.9% for the copies in the U part.

In the sequenced genome of *D. melanogaster*, we found 26 copies of *helena* (see Additional File 2 for details), which ranged in size from 91 bp to 4805 bp. The average percentage identity was 80.4% for all copies, with an average of 78.7% for the copies in the chromosome arms, and of 83.7% for the copies located in the U part. Most of the copies in this genome have therefore been degraded, with numerous internal deletions or insertions. All copies are truncated on the 5' side, and are DOA (Dead on arrival) copies, apart from the 3L\_23487977 copy, although even this displays some internal deletions.

We analyzed in greater detail any copies that could correspond to the most recent insertions in both *D. melanogaster* and *D. simulans*. We used specific blast criteria to identify these copies: we selected matches with at least 90% identity, and a length at least 50% of that of the complete copy, with e-values of less than 10e-10 (Figure 2). In *D. melanogaster*, only the 3L\_23487977 copy described above met all the blast criteria. It is obviously an inactive copy, since more than four deletions were detected within its sequence. In *D. simulans*, six copies were found that matched the blast criteria, including the complete copy on the 3R chromosome (position 1506433 – 1511368 on the minus strand); the other five copies had internal deletions, and insertions were detected in four of them.

**Chromatin localization of helena copies in natural populations**

We used *in situ* hybridization to estimate the number of *helena* insertion sites located on the arms (euchromatin) of the polytene chromosomes from salivary glands of both *D. melanogaster* and *D. simulans*. Both species had centromeric staining, but only *D. simulans* from natural

| Species                | Chromosome/<br>Copy number | Copies of <i>helena</i> more than 50% as long as the reference copy<br> |
|------------------------|----------------------------|---|
| <i>D. melanogaster</i> | 3L / 2                     |   |
|                        |                            |   |
|                        | 3R / 0                     |   |
|                        | 2L / 0                     |   |
|                        | 2R / 0                     |   |
|                        | X / 0                      |   |
|                        | U / 0                      |   |
| Total 2                |                            |   |
| <i>D. simulans</i>     | 3L / 1                     |   |
|                        | 3R / 1                     |   |
|                        | 2L / 0                     |   |
|                        | 2R / 0                     |   |
|                        | X / 1                      |   |
|                        | U / 2                      |   |
|                        |                            |   |
| Total 5                |                            |   |

**Figure 2**  
**Scheme of *helena* copies.** Representation of *helena* copies in the *D. melanogaster* and *D. simulans* genomes with at least 90% identity and 50% of the length of the complete copy, and with e-values of less than  $10e^{-10}$ ; Triangles = insertions; Spaces = deletions.

populations presented euchromatic bands (mean copy number  $10.7 \pm 2.2$ ) with no fixed sites (see Additional File 3 for details on insertion sites per population). With this experiment we did not detect any insertions of *helena* in the chromosome arms, which could be explained by the short size of the elements and the divergence to the probe used.

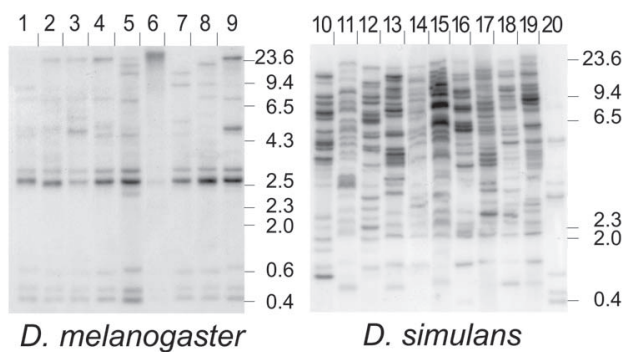
*Inter-population polymorphism*

We analyzed the inter-population *helena* copy number polymorphism by Southern blot, using a restriction enzyme that does not cut inside the element. This method detects both heterochromatic and euchromatic sequences. As shown in Figure 3, *D. melanogaster* had 8 to 11 bands per population, and several bands were shared by different populations. These copies could correspond to ancient and fixed heterochromatic copies in the *D. melanogaster* genome. *D. simulans* populations contained

numerous *helena* copies (19 to 30 copies per population) with a high level of insertion polymorphism. Since the enzyme used for the Southern blot did not cut inside the element, all bands over 4.5 kb could correspond to a complete element. Because both species harbored bands over 4.5 kb, they could have full-length *helena* copies.

*PCR screening*

Three sets of primers were used to amplify the whole ORF1 and two fragments of the ORF2. No bands corresponding to the ORF1 were observed in any of the *D. melanogaster* natural populations in agreement with the absence of this ORF in the sequenced genome. There was a high level of size polymorphism for the *D. melanogaster* ORF2, corresponding to the different internal deletions already analyzed by Petrov et al. [13,23]. In contrast, *D. simulans* displayed low size polymorphism for *helena* ORF1 and ORF2. Indeed, only one population out of



**Figure 3**  
**Southern blot analysis of *D. melanogaster* and *D. simulans* populations.** Lanes 1 to 9 are *D. melanogaster* populations (Bolivia, Brazzaville, Canton, Chicharo, Reunion Island, Arabia, Virasoro, Vietnam, and ISO for the 9<sup>th</sup> and 20<sup>th</sup> lane). Lanes 1 to 19 are *D. simulans* populations (Amieu, Eden, Valence, Canberra, Papeete, Moscow, Makindu, Zimbabwe, Cann River and Reunion Island). For both Southern blots, the DNA size is estimated in base pairs.

twenty had two sets of ORF1 (Papeete). All *D. simulans* populations had two to three sets of ORF2.

#### Analysis of *helena* copies in *D. simulans* populations

PCR fragments obtained from the *D. simulans* population screening mentioned above were cloned and sequenced. Surprisingly several common indels were detected at the same positions in different sequences from all the populations analyzed. Phylogenetic reconstructions based on the ORF1 and ORF2 (Figures 4 and 5, alignments from Additional Files 4 and 5) showed that the sequences that displayed the same indels are grouped in the tree. This suggests that the deletion or insertion events were produced before the amplification of these sequences. Some copies had only a few insertions, and might be inactive since their reading frames were not preserved. However, the amplification of some of these copies could have been promoted in trans. We did not find any population that had both complete ORFs. Nevertheless, some ORFs had no internal stop codons in their sequence, suggesting that copies bearing them could be active despite the deletions.

The percentage identity between the reference copy and ORF1 ranges between 96% and 99%, meaning that these copies have not diverged much. No relationship was detected between the size and location of the deletions, and the percentage identity. For the ORF2, we found copies with a percentage identity of more than 93% that reached 100% for some copies. A common 401-bp deletion was found in copies with 93% identity with the com-

plete *helena* copy, but no correlation was observed with the percentage identity for the other deletions or insertions. Based on the age estimation of each copy, we found that most of the oldest ORF2 fragments had the 401-bp deletion. Several young copies of both ORF1 and ORF2 displayed major internal deletions, showing that the mechanism leading to these deletions is much more powerful than copy divergence in inactivating them.

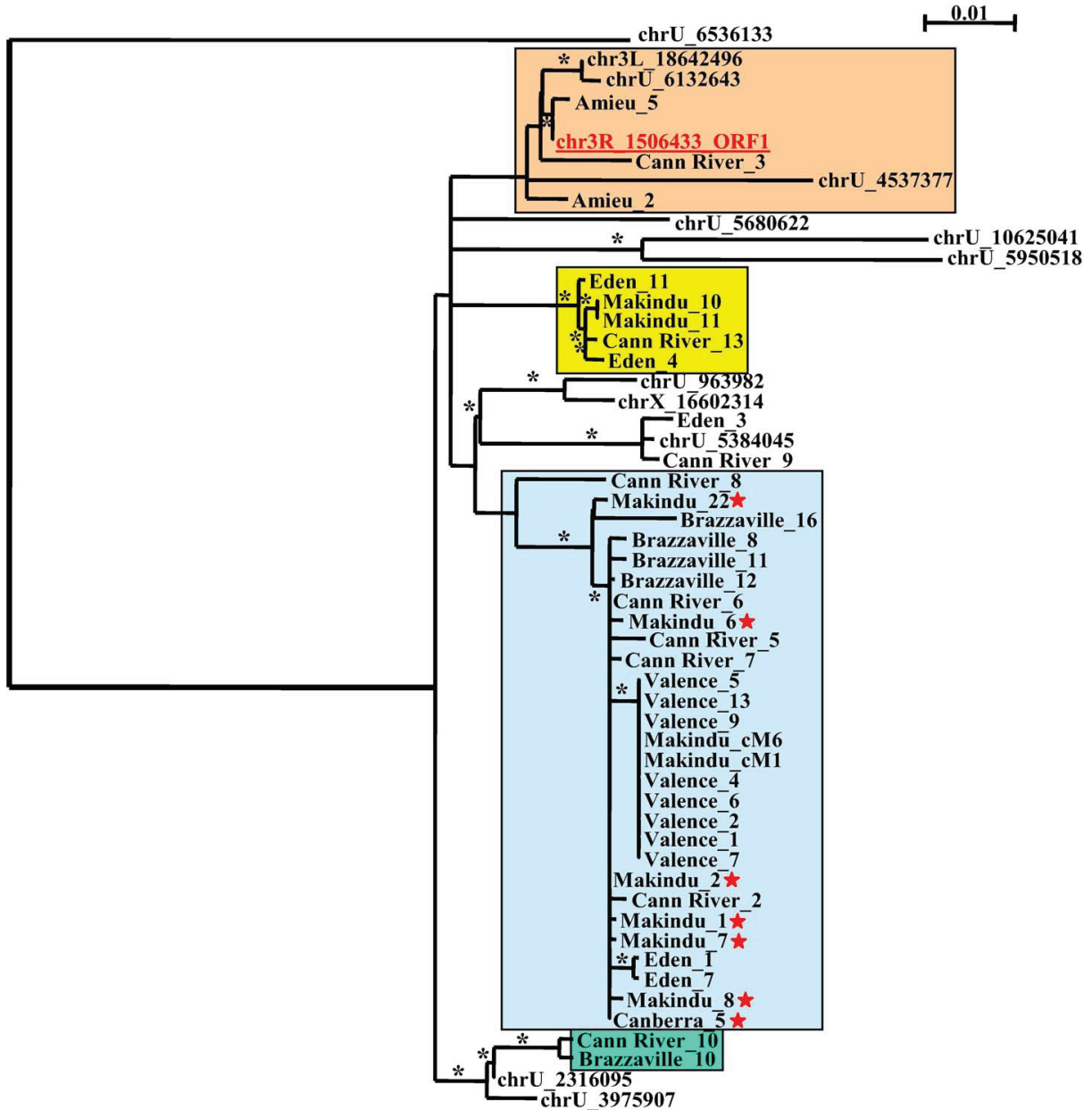
#### Transcript analysis in *D. simulans*

To test the transcriptional potential of *helena* in *D. simulans*, we performed northern blot and RT-PCR in various populations (see Materials and Methods). Since the sequenced genome was obtained from strains from North America, we added three populations from this continent (San Antonio, SW3-S2 and San Diego). No transcripts were found by Northern blot in any of the populations. However, the RT-PCR method detected transcripts of both ORF1 and ORF2 in the Valence population, an extremely low signal for ORF1 and ORF2 in three of the American populations, and ORF2 transcripts in the Amieu population. None of the other populations had any transcript for *helena*, implying that this element is extinct in these populations. Since the Northern blot technique is less sensitive than RT-PCR, these findings suggest that *helena* is transcribed at extremely low levels in the populations in which some transcripts were detected by RT-PCR.

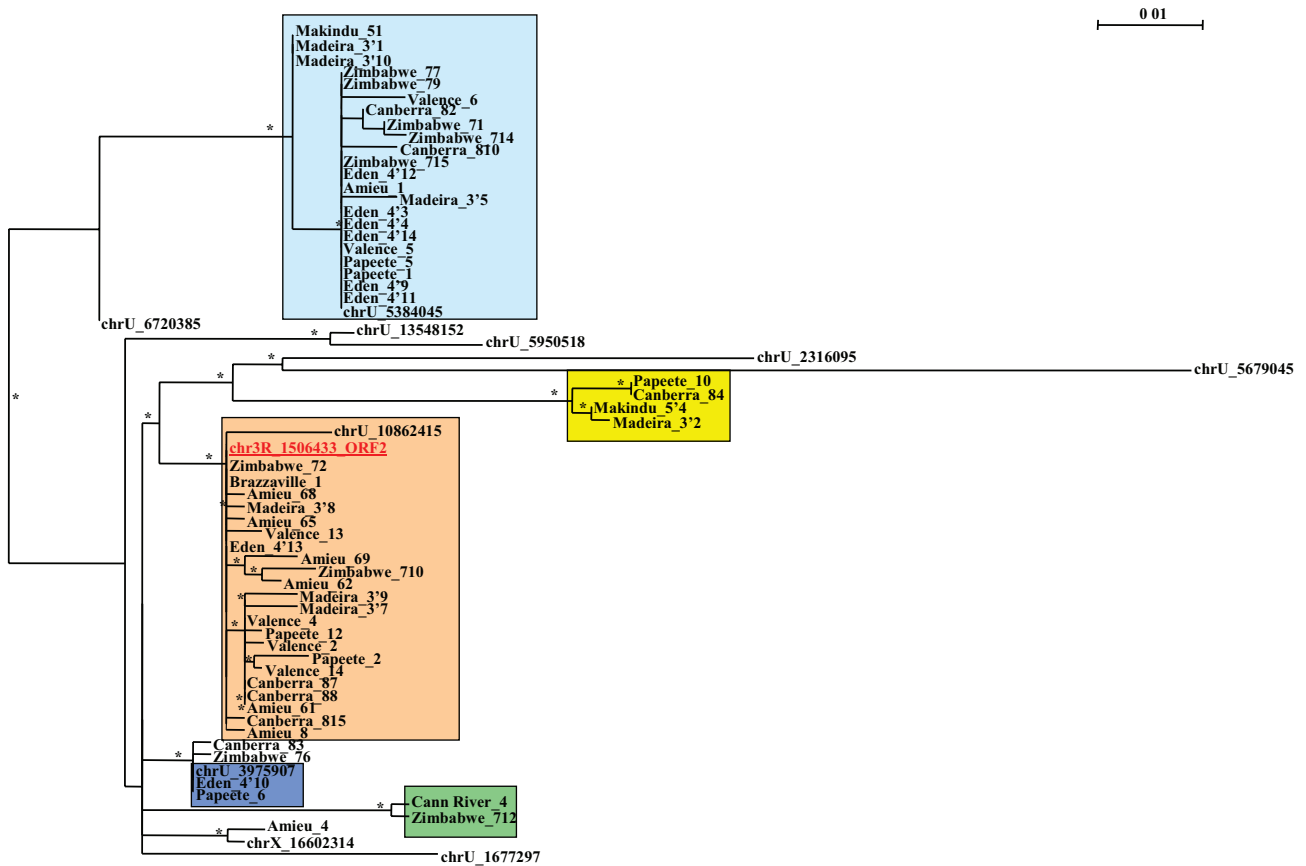
#### Discussion

Our *in silico* and experimental analyses of the *D. melanogaster* genome show that *helena* copies are mostly DOA, devoid of ORF1, and therefore unable to transpose autonomously. All these features have been associated with elements that are in the process of inactivation [20]. The scenario for the *helena* copies in *D. simulans* is quite different. Analysis of the sequenced genome of this species allowed us to identify a full-length copy of *helena* with the structures required for an active element: two intact ORFs, a poly-A tail, and regulatory regions. The high level of insertion polymorphism detected in the natural populations suggests that *helena* is an active element or has been active recently. However, sequence analysis of the two ORFs of *helena* in the natural populations revealed two main points: **first**, both ORFs are intact in only very few populations; **second**, even though the sequences of *helena* are very similar at the nucleotide level, their deletion features differ.

The first point was strengthened by the almost total absence of *helena* transcripts in all natural populations of *D. simulans*, which means that very few copies were involved in transcription. Because a single master copy is enough to maintain TE transposition [24], the putative activity of *helena* in this species could reside in the full-length copy probably present in some populations such as



**Figure 4**  
**phylogenetic tree of the DNA sequences from the ORF1 region.** The reconstruction was performed on the cloned DNA sequences of the ORF1 region from the different populations of *D. simulans*, and from some sequences detected in the sequenced genome (we eliminated sequences that were too short relative to the global length of the alignment). Colored boxes identify sequences harboring common patterns of deletions and/or insertions. All the positions are given by reference to the complete copy. Green box: sequences display the same deletions of 118 bp (at position 45), 3 bp (at position 839), 6 bp (at position 846), 1 bp (at position 854) and a 1-bp insertion (at position 415). Blue box: sequences display the same 28-bp deletion (at position 1092) – those with a red star also have a 77-bp deletion (at position 508), and a 91-bp deletion (at position 593). Yellow box: sequences display the same deletions of 1 bp (at position 160), 28 bp (at position 165), 19 bp (at position 954), 2 bp (at position 989), and 37 bp (at position 1006), and an insertion of 1 bp (at position 322). Orange box: sequences with no deletion or insertion, very closely related to the complete copy chr3R\_1506433. Black asterisks correspond to bootstrap values greater than 50%.



**Figure 5**  
**phylogenetic tree of the DNA sequences from the ORF2 region.** The reconstruction was performed on the cloned DNA sequences of the ORF2 region from the different populations of *D. simulans* and from some sequences detected in the sequenced genome (we eliminated sequences that were too short relative to the global length of the alignment). Colored boxes represent sequences harboring common patterns of deletions and/or insertions. All the positions are given relative to the complete copy. Green box: sequences display the same 4-bp deletion (at position 547) and the same 2-bp insertion (at position 112). Dark blue box: sequences display the same 1-bp deletion (at position 521). Yellow box: sequences display the same deletions of 3 bp (at position 143), 8 bp (at position 335), 4 bp (at position 345), and an insertion of 2 bp (at position 460). Light blue box: sequences display the same 401-bp deletion (at position 179). Orange box: sequences with no deletion or insertion, very closely related to the complete copy chr3R\_1506433. Black asterisks correspond to bootstrap values greater than 50%.

Valence, where we did observe transcripts. Hence, we would expect sequences that are still similar at nucleotide level to differ only in the 5' end truncation size, as usually observed for LINE-like elements. However, as mentioned in the second point, we actually observed many other kinds of internal deletions that occurred throughout the length of the element. Intriguingly this deletion-promoting process appears to be quite powerful in inactivating the elements, and could be even more powerful than other mutation processes such as point mutations. This means that a real-time loss of *helena* is ongoing in all *D. simulans* populations.

The nature of the mechanisms leading to internally deleted copies is still unknown. In humans, LINE elements can be spliced [25], a process that creates internal deletions. We used bioinformatic analyses, but were unable to find splice sites in the full-length *helena* sequence. Also, although recombination between mRNAs can produce internal deletions [26], *helena* sequences are not sufficiently divergent to allow us to infer the origin of a single copy.

*Helena* appears to be extinct in *D. melanogaster*, and this recalls the *I* element, which also disappeared from the *D. melanogaster* genome in the past, and reinvaded it only recently [27]. The *I* and *helena* elements are both LINE-like

elements, leading us to wonder whether amplification/loss of copies could be a characteristic of this type of element. Waves of amplification/loss have been observed in humans, where only the youngest L1 subfamily is active, perhaps as a result of competition between different L1 subfamilies [2,7].

Our data show that *helena* has been almost entirely removed from the *D. melanogaster* genome, and was not subjected to the recent wave of transpositions reported for other elements [14]. In *D. simulans*, we did observe an insertion site polymorphism of *helena*, but this corresponds to copies that are being internally deleted by an efficient mechanism. This may be generalized on the basis of data on the LTR retrotransposons *412* and *tirant*, which have also internal deletions [28,29]. We still do not clearly understand which mechanisms lead to a low copy number in the *D. simulans* genome, but a mechanism promoting internal deletions could be a major force at work [30].

### Conclusion

TEs are major players in genome evolution, and the way they are controlled by the host genome is one of the most fundamental questions in evolutionary genomics. Here we show that two closely related species of drosophila have a TE family at different stages in its life cycle. The mechanism by which this is achieved in *D. simulans* implies that a very efficient internal deletion mechanism is acting on TEs, which is more powerful than the simple neutral evolution of non-active elements. The difference in the amount of TEs between *D. melanogaster* and *D. simulans* could be explained by such a process, that doesn't seem to be very active in *D. melanogaster* present populations.

### Methods

#### Natural populations

We worked on fly samples collected from several geographically-distinct, natural populations (confer each Method for natural populations investigated). These populations were maintained in the laboratory as isofemale lines or small mass cultures with around 50 pairs in each generation.

#### In situ hybridization

Polytene chromosomes from salivary glands of third-instar female larvae were prepared and treated with nick-translated, biotinylated DNA probes, as previously described [31]. Insertion sites were visible as brown bands resulting from a dye-coupled reaction with peroxidase substrate and diaminobenzidine. The insertion site numbers of the TE(s) were determined on all the long chromosomes arms (X, 2L, 2R, 3L, 3R), and were summed to give the total number of labeled sites per diploid genome. We

did not take into account the insertions located in pericentromeric regions 20, 40, 41, 80, and 81, because TE site number estimations in these regions are difficult and not reliable for all chromosomes or all squashes. We used a probe (1278 bp) from *helena* of *D. sechellia* (AF012044). The following populations of *D. melanogaster* were investigated: Portugal (Chicharo), Saudi Arabia, Congo (Brazzaville), Reunion Island, Argentina (Virasoro), Bolivia, China (Canton), Vietnam, and Iso line. The *D. simulans* populations analyzed were from France (Valence), Russia (Moscow), Kenya (Makindu), Zimbabwe, Reunion Island, Australia (Eden, Cann River and Canberra), French Polynesia (Papeete), New Caledonia (Amieu).

#### Southern blot

DNA was extracted from one or five adult females by a standard phenol-chlorophorm-salt method with proteinase K digestion. The *D. melanogaster* populations analyzed were from Bolivia, Congo (Brazzaville), China (Canton), Portugal (Chicharo), Reunion Island, Saudi Arabia, Argentina (Virasoro), Vietnam, and ISO line. The *D. simulans* populations analyzed were from New Caledonia (Amieu), Australia (Eden, Cann River and Canberra), France (Valence), French Polynesia (Papeete), Russia (Moscow), Kenya (Makindu), Zimbabwe, and Reunion Island. The DNA was cut using the *HindIII* enzyme, which has no restriction site within the *helena* sequence, and therefore allowed us to estimate the number of complete *helena* copies. Electrophoresis of a 0.8% agarose gel containing digested DNA was carried out for 17 h. The DNA was denatured (NaOH 0.5 M), and then transferred overnight to a Hybond-N+ nylon membrane. Pre-hybridization and hybridization were carried out at 67°C using a Denhardt 5× solution. The probe used for hybridization (AF012044) was radiolabeled with <sup>32</sup>P, using a random procedure from Amersham.

#### Amplification of ORF1 and ORF2

DNA was extracted from single flies by a standard phenol-chlorophorm method. The following populations of *D. melanogaster* were investigated: France (Valence and Saint Cyprien), Portugal (Chicharo), Saudi Arabia, Senegal, Congo (Brazzaville), Reunion Island, Guadeloupe, Argentina (Virasoro), Bolivia, and China (Canton). The *D. simulans* populations analyzed were from France (Valence), Russia (Moscow), Egypt (Tanta), Congo (Brazzaville), Kenya (Meru, Kwalé and Makindu), Zimbabwe, Tanzania (Arusha), Puerto Rico, Japan, Australia (Eden, Cann River and Canberra), French Polynesia (Papeete), Saint Martin, Hawaii, New Caledonia (Amieu), and Portugal (Madeira).

PCR was run using 1 µg DNA with the two following primers – ORF1: H1for (285 5' AAC TGT AAA ATG GAT ACG AAC A 3' 306), H1rev (1808 5' GCC ACT TCA TAA

ATT GTT CC 3' 1827). – ORF2: Hel2F (2325 5' CCG GGC TGG GCG ATA TGG 3' 2342), Hel2R (4548 CGT ACA TAC CAG GGG CAG TTG G 3' 4569). PCR was run in 30 cycles with annealing temperatures of 57°C (ORF1) and 56°C (ORF2). We used Euroblue taq from Eurobio. DNA amplified fragments were purified and cloned on competent bacteria (Qiagen kits). Four primers were used for sequencing: M13 forward and reverse; Seq1 (5' CTC TTC CTT CAT TTG GTA CG 3') and Seq2 (5' AAG GGG AAA CAG TGA GAA TA 3') for the complete ORF1; Seq3F (5' TTA GAC CAT GCT CTC GGT TA 3') and Seq3R (5' TGT CAA TTC CTG GAG CTT TA 3') for a fragment of ORF2. Sequencing was performed by Genome Express. Accession numbers (Genbank) from [EU168807](#) to [EU168844](#) correspond to ORF1 fragments. Accession numbers (Genbank) from [EU170377](#) to [EU170431](#) correspond to ORF2 fragments.

#### RT-PCR

Total RNA was extracted from four adult females, four adult males, 10 ovaries and 10 testes from *D. simulans* populations (France (Valence), Congo (Brazzaville), Kenya (Makindu), Zimbabwe, Australia (Canberra), New Caledonia (Amieu), Portugal (Madeira), United States (San Antonio, San Diego, Arena, SW3)) with the RNeasy protect mini kit from Qiagen. RNA extracts were treated with the Ambion's DNA-free kit. ThermoScript RT-PCR system from Invitrogen was used to synthesize four different cDNA pools (55°C for 90 min and 85°C for 5 min): a control reaction with no reverse transcriptase to test DNA contamination, a pool of total cDNA synthesized with oligo-dt primers, two specific cDNA pools obtained with H1R (ORF1) and Hel2R (ORF2), respectively, corresponding to *helena* transcripts. All four cDNA pools were tested for the presence of *actin* cDNA (house keeping gene) by PCR with Act5cfw (5'ATGTGACGAAGAAGTTG3') and Act5cRv (5'TTAGAAGCACTTGCGGTGCA3') primers. Oligo-dt and specific *helena* cDNA pools were analyzed by PCR using ORF1 and ORF2 specific primers (H1R/H1F and Hel2R/Hel2F).

#### Northern blot

Total RNA was extracted from adult females or embryos from several *D. simulans* populations (Valence, Makindu, Amieu, Brazzaville) with the RNeasy protect mini kit from Qiagen. Total RNA extracts were treated with the Ambion's DNA-free kit. Electrophoresis (MOPS, formaldehyde gel) was run for 3 h after RNA denaturing. After washing (water and NaOH, 75 mM) RNA was passively transferred to a nylon membrane, and cross linked for 2 hours at 64°C. Blots were pre-hybridized in hybridization buffer, then hybridized overnight at 42°C in hybridization buffer containing a <sup>32</sup>P-labeled *helena* cDNA probe. The radiolabeled cDNA probe was prepared using a Meg-

aprime DNA Labeling Kit according to the manufacturer's protocol (Cat # RPN 1607; Amersham Biosciences, Little Chalfont, Buckinghamshire, England). Following hybridization, blots were washed in 2 × SSC/0.1% SDS at 42°C and then exposed to X-ray film (KODAK).

#### Identification of *helena* copies in the complete genomes

We retrieved the sequences of the chromosome arms 2L, 2R, 3L, 3R, 4, X and the unassigned part (named U) corresponding to the first release of the mosaic assembly of the genome of *D. simulans* available at the ftp site of the Genome Sequencing Center at the Washington University Medical School [32]. This mosaic assembly corresponds to different strains of *D. simulans*. We also used the sequenced genome of *D. melanogaster* [33]. We will refer to the *helena* copies found in the genomes according to the chromosome name and the start position of the copy (for example chr2L\_133500 corresponds to a copy found on the 2L chromosome, and it starts at position 133500).

The *helena* element was only found in the databases as fragments of the reverse transcriptase (RT). We retrieved the longest sequences from *D. yakuba* (accession number in Genbank [AF012049](#)), *D. melanogaster* ([AF012030](#)) and *D. virilis* ([U26847](#)) to build a chimeric, 1532-bp sequence. Using this chimeric element we searched for copies in the *D. simulans* sequenced genome using blastn [34]. The reconstructed *helena* sequence (ID Helena\_DS) is available in Repbase [35]. Only matches with an e-value of less than 10e<sup>-10</sup> have been retained, and any separated by distances of less than 300 bp have been merged. As the query used corresponds to a small portion of the ORF2, in order to search for longer sequences of *helena*, we retrieved the matches after adding 5000 bp around their positions. We then performed multiple alignments of these sequences using clustalw [36] in order to detect the longest copies. By this procedure, we identified a sequence on the chromosome 3R that was the longest of the matches detected. The prediction of potentially coding parts was made using the ORF finder program available on the NCBI web site [37], and this allowed us to identify two ORFs. It was not possible to use the presence of target site duplication to determine the exact position of the beginning of the sequence, because the copy was surrounded by unidentified bases, and so we performed a blast search in the draft sequence of *D. sechellia*, the closest relative to *D. simulans*. This allowed us to find a homologous copy, and to identify the beginning of the complete copy of *helena*. Once this copy had been identified, it was used as a query to perform blast searches in the *D. simulans* and *D. melanogaster* genomes to determine the *helena* copy populations.

### Sequence analysis

The computation of the percentage identity was performed using the DNADIST program in the PHYLIP package [38]. We used the sequence editor Seaview [39] to visualize the sequences and the alignments. Splice sites and transcription binding sites were predicted by the Softberry tools [40] and Genomatix [41]; PEPcoil ([42] allowed us to find the coiled coil domain in ORF 1. Conserved domains in both ORF1 and ORF2 were predicted with the "Conserved domain search" tool from NCBI. Sequenced copies were aligned with T\_coffee [43]. Phylogenetic analysis were made using maximum likelihood with HKY substitution model implemented in PhyML [44]. The reconstruction was performed on the cloned DNA sequences of the ORF1 and ORF2 region from the different populations of *D. simulans*, and from some sequences detected in the sequenced genome (we eliminated sequences that were too short relative to the global length of the alignment).

Age was estimated using the Bowen and McDonald method [45] with the formula  $Age = K/(2r)$ , where K is the divergence between the two copies calculated from the Kimura two-parameter distance via DNAdist, and r is the synonymous substitution rate per site per million years in *D. melanogaster* ( $r = 0.016$  from Li [46]). It is important to note that the age of *helena* copies is underestimated due to the lack of knowledge about conversion and substitution rates in *D. melanogaster* genome, and is also unreliable when applied to old and highly diverged copies.

### Abbreviations

DOA, dead on arrival; LINE, long interspersed nuclear element; L1, LINE 1; LTR, long terminal repeat; ORF, open reading frame; TE, transposable element

### Authors' contributions

RR and LL carried out the molecular and genetic studies, RR and EL did the bioinformatic analysis, CV and CB contributed to the design of the study, CV designed and coordinated the study. All the authors contributed to writing of the paper. All the authors have read and approved the final manuscript.

### Additional material

#### Additional File 1

*helena* copies in the *Drosophila simulans* sequenced genome. The data provided is a list of the *D. simulans* copies  
Click here for file  
[<http://www.biomedcentral.com/content/supplementary/1471-2164-9-149-S1.doc>]

#### Additional File 2

*helena* copies in the *Drosophila melanogaster* sequenced genome. The data provided is a list of the *D. melanogaster* copies.  
Click here for file  
[<http://www.biomedcentral.com/content/supplementary/1471-2164-9-149-S2.doc>]

#### Additional File 3

*Helena* copy number in *D. melanogaster* and *D. simulans* populations by *in situ* hybridization in polytene chromosomes. The data provided de *in situ* hybridization results.  
Click here for file  
[<http://www.biomedcentral.com/content/supplementary/1471-2164-9-149-S3.xls>]

#### Additional File 4

Alignment of *helena* ORF1. The data provided the alignment used to construct the tree on Figure 4  
Click here for file  
[<http://www.biomedcentral.com/content/supplementary/1471-2164-9-149-S4.pdf>]

#### Additional File 5

Alignment of *helena* ORF2. The data provided the alignment used to construct the tree on Figure 5  
Click here for file  
[<http://www.biomedcentral.com/content/supplementary/1471-2164-9-149-S5.pdf>]

### Acknowledgements

We thank Monika Ghosh for reviewing the English text. This work was funded by the Centre National de la Recherche Scientifique (UMR 5558) and the IMPBio program.

### References

1. Biemont C, Vieira C: **Genetics: junk DNA as an evolutionary force.** *Nature* 2006, **443**(7111):521-524.
2. Khan H, Smit A, Boissinot S: **Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates.** *Genome research* 2006, **16**(1):78-87.
3. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al.: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409**(6822):860-921.
4. Eickbush TH, Furano AV: **Fruit flies and humans respond differently to retrotransposons.** *Curr Opin Genet Dev* 2002, **12**(6):669-674.
5. Bennetzen JL: **Transposable elements, gene creation and genome rearrangement in flowering plants.** *Current opinion in genetics & development* 2005, **15**(6):621-627.
6. Vitte C, Bennetzen JL: **Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution.** *Proc Natl Acad Sci USA* 2006, **103**(47):17638-17643.
7. Boissinot S, Furano AV: **The recent evolution of human L1 retrotransposons.** *Cytogenet Genome Res* 2005, **110**(1-4):402-406.
8. Chambeyron S, Bucheton A: **I elements in *Drosophila*: *in vivo* retrotransposition and regulation.** *Cytogenetic and genome research* 2005, **110**(1-4):215-222.
9. SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL: **The paleontology of intergene retrotransposons of maize.** *Nature genetics* 1998, **20**(1):43-45.

10. Li YJ, Satta Y, Takahata N: **Paleo-demography of the *Drosophila melanogaster* subgroup: application of the maximum likelihood method.** *Genes & genetic systems* 1999, **74(4)**:117-127.
11. Vieira C, Lepetit D, Dumont S, Biemont C: **Wake up of transposable elements following *Drosophila simulans* worldwide colonization.** *Mol Biol Evol* 1999, **16(9)**:1251-1255.
12. Biemont C, Nardon C, Deceliere G, Lepetit D, Loevenbruck C, Vieira C: **Worldwide distribution of transposable element copy number in natural populations of *Drosophila simulans*.** *Evolution Int J Org Evolution* 2003, **57(1)**:159-167.
13. Petrov DA, Hartl DL: **High rate of DNA loss in the *Drosophila melanogaster* and *Drosophila virilis* species groups.** *Molecular Biology and Evolution* 1998, **15(3)**:293-302.
14. Lerat E, Rizzon C, Biemont C: **Sequence divergence within transposable element families in the *Drosophila melanogaster* genome.** *Genome Res* 2003, **13(8)**:1889-1896.
15. Arkhipova IR, Lyubomirskaya NV, Ilyn YV: ***Drosophila* Retrotransposons.** Berlin: Springer; 1995.
16. **Assembly/Alignment/Annotation of 12 related *Drosophila* species** [<http://rana.lbl.gov/drosophila/index.html>]
17. Malik HS, Burke WD, Eickbush TH: **The age and evolution of non-LTR retrotransposable elements.** *Molecular biology and evolution* 1999, **16(6)**:793-805.
18. Hohjoh H, Singer MF: **Cytoplasmic ribonucleoprotein complexes containing human LINE-I protein and RNA.** *Embo J* 1996, **15(3)**:630-639.
19. Pardue ML, Danilevskaya ON, Lowenhaupt K, Wong J, Erby K: **The gag coding region of the *Drosophila* telomeric retrotransposon, HeT-A, has an internal frame shift and a length polymorphic region.** *J Mol Evol* 1996, **43(6)**:572-583.
20. Casavant NC, Scott L, Cantrell MA, Wiggins LE, Baker RJ, Wichman HA: **The end of the LINE?: lack of recent LI activity in a group of South American rodents.** *Genetics* 2000, **154(4)**:1809-1817.
21. Rohrmann GF, Karplus PA: **Relatedness of baculovirus and gypsy retrotransposon envelope proteins.** *BMC Evol Biol* 2001, **1**:1.
22. Fablet M, Rebollo R, Biemont C, Vieira C: **The evolution of retrotransposon regulatory regions and its consequences on the *Drosophila melanogaster* and *Homo sapiens* host genomes.** *Gene* 2007, **390(1-2)**:84-91.
23. Lozovskaya ER, Nurminsky DI, Petrov DA, Hartl DL: **Genome size as a mutation-selection-drift process.** *Genes Genet Syst* 1999, **74(5)**:201-207.
24. Brookfield JF: **Evolutionary forces generating sequence homogeneity and heterogeneity within retrotransposon families.** *Cytogenet Genome Res* 2005, **110(1-4)**:383-391.
25. Belancio VP, Hedges DJ, Deininger P: **LINE-I RNA splicing and influences on mammalian gene expression.** *Nucleic Acids Res* 2006, **34(5)**:1512-1521.
26. Roy SW, Penny D: **Widespread intron loss suggests retrotransposon activity in ancient apicomplexans.** *Molecular biology and evolution* 2007, **24(9)**:1926-1933.
27. Bucheton A, Vaury C, Chaboissier MC, Abad P, Pelisson A, Simonelig M: **I elements and the *Drosophila* genome.** *Genetica* 1992, **86(1-3)**:175-190.
28. Fablet M, McDonald JF, Biemont C, Vieira C: **Ongoing loss of the tirant transposable element in natural populations of *Drosophila simulans*.** *Gene* 2006, **375**:54-62.
29. Mugnier N, Biemont C, Vieira C: **New regulatory regions of *Drosophila* 412 retrotransposable element generated by recombination.** *Molecular biology and evolution* 2005, **22(3)**:747-757.
30. Vieira C, Biemont C: **Transposable element dynamics in two sibling species: *Drosophila melanogaster* and *Drosophila simulans*.** *Genetica* 2004, **120(1-3)**:115-123.
31. Biemont C, Monti-Dedieu L, Lemeunier F: **Detection of transposable elements in *Drosophila* salivary gland polytene chromosomes by in situ hybridization.** *Methods Mol Biol* 2004, **260**:21-28.
32. **Genome Sequencing Center at the Washington University Medical School** [<http://hgdownload.cse.ucsc.edu/downloads.html#droSim>]
33. **Ensembl Genome Browser** [<http://www.ensembl.org/index.html>]
34. Karlin S, Altschul SF: **Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes.** *Proc Natl Acad Sci USA* 1990, **87(6)**:2264-2268.
35. **Repbase** [<http://www.girinst.org/>]
36. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22(22)**:4673-4680.
37. **ORF Finder** [<http://www.ncbi.nlm.nih.gov/gorf/gorf.html>]
38. Felsenstein J: **An alternating least squares approach to inferring phylogenies from pairwise distances.** *Syst Biol* 1997, **46(1)**:101-111.
39. Galtier N, Gouy M, Gautier C: **SEAVIEW and PHYLO WIN: two graphic tools for sequence alignment and molecular phylogeny.** *Comput Appl Biosci* 1996, **12(6)**:543-548.
40. **Softberry tools** [<http://www.softberry.com/berry.phtml>]
41. **Genomatix** [<http://www.genomatix.de>]
42. Lupas A, Van Dyke M, Stock J: **Predicting coiled coils from protein sequences.** *Science* 1991, **252(5010)**:1162-1164.
43. Notredame C, Higgins DG, Heringa J: **T-Coffee: A novel method for fast and accurate multiple sequence alignment.** *J Mol Biol* 2000, **302(1)**:205-217.
44. Guindon S, Gascuel O: **A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood.** *Systematic biology* 2003, **52(5)**:696-704.
45. Bowen NJ, McDonald JF: ***Drosophila* euchromatic LTR retrotransposons are much younger than the host species in which they reside.** *Genome research* 2001, **11(9)**:1527-1540.
46. Li W: **Molecular Evolution.** Sunderland, MA: Sinauer Associates; 1997.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:  
[http://www.biomedcentral.com/info/publishing\\_adv.asp](http://www.biomedcentral.com/info/publishing_adv.asp)



Supplementary Table 1: *helena* copies in the *Drosophila simulans* sequenced genome

| chromosome            | strand   | start          | stop           | length (bp) | %identity with the complete <i>helena</i> |
|-----------------------|----------|----------------|----------------|-------------|---|
| 2L <sup>£</sup>       | +        | 1592410        | 1592944        | 534         | 99.25                                     |
| 2L <sup>¥</sup>       | +        | 1674490        | 1674844        | 255         | 99.21                                     |
| 2L <sup>*</sup>       | -        | 15248615       | 15252309       | 3680        | 99.75                                     |
| 2L <sup>£</sup>       | +        | 17606797       | 17606986       | 190         | 100.00                                    |
| 2L <sup>£¥</sup>      | +        | 21324887       | 21324994       | 108         | 90.74                                     |
| 2L <sup>£</sup>       | -        | 21614564       | 21615205       | 642         | 99.22                                     |
| 2L <sup>£</sup>       | +        | 21940141       | 21940640       | 500         | 96.59                                     |
| 2R <sup>£</sup>       | -        | 1707808        | 1707920        | 113         | 97.32                                     |
| 2R <sup>£</sup>       | -        | 11963548       | 11964592       | 1045        | 99.81                                     |
| 2R <sup>£</sup>       | -        | 13305831       | 13310157       | 4327        | 99.22                                     |
| 2R <sup>£</sup>       | -        | 14875310       | 14876458       | 1149        | 99.91                                     |
| 2R <sup>¥</sup>       | -        | 14892514       | 14892896       | 383         | 98.35                                     |
| 2R <sup>£</sup>       | -        | 15489919       | 15490045       | 128         | 96.06                                     |
| 3L <sup>£</sup>       | +        | 6812601        | 6816422        | 3821        | 99.63                                     |
| 3L <sup>£</sup>       | -        | 13781457       | 13782139       | 683         | 99.56                                     |
| 3L <sup>£</sup>       | -        | 14741536       | 14741739       | 204         | 99.51                                     |
| 3L <sup>£</sup>       | -        | 18635059       | 18635811       | 753         | 99.07                                     |
| 3L <sup>¥</sup>       | -        | 18642496       | 18643290       | 795         | 100.00                                    |
| 3L <sup>£</sup>       | +        | 22376004       | 22376483       | 479         | 100.00                                    |
| 3R <sup>£¥</sup>      | -        | 30161          | 30267          | 107         | 88.79                                     |
| 3R <sup>£</sup>       | +        | 5983652        | 5983777        | 126         | 95.20                                     |
| 3R <sup>£</sup>       | -        | 1504547        | 1505432        | 885         | 99.77                                     |
| <b>3R<sup>§</sup></b> | <b>-</b> | <b>1506433</b> | <b>1511368</b> | <b>4912</b> | <b>/</b>                                  |
| 3R <sup>£</sup>       | +        | 17102251       | 17102891       | 641         | 99.53                                     |
| 3R <sup>£</sup>       | +        | 18466755       | 18467294       | 539         | 99.81                                     |
| X <sup>£</sup>        | +        | 4098908        | 4099039        | 131         | 96.92                                     |
| X <sup>£¥</sup>       | -        | 16594484       | 16595282       | 799         | 93.59                                     |
| X <sup>£*</sup>       | +        | 16602314       | 16610995       | 8682        | 96.94                                     |
| X <sup>£</sup>        | -        | 16622045       | 16623141       | 1096        | 95.62                                     |
| U <sup>£</sup>        | +        | 233264         | 234099         | 835         | 95.08                                     |
| U <sup>¥§</sup>       | -        | 963982         | 966622         | 2641        | 97.01                                     |
| U <sup>£</sup>        | +        | 1677297        | 1679692        | 2396        | 96.16                                     |
| U <sup>¥</sup>        | -        | 1817942        | 1818183        | 242         | 96.28                                     |
| U <sup>£§</sup>       | +        | 2316095        | 2319355        | 3260        | 94.32                                     |
| U <sup>£¥</sup>       | +        | 3850078        | 3850295        | 218         | 98.17                                     |
| U <sup>§</sup>        | +        | 3975907        | 3980988        | 5098        | 97.12                                     |

|                   |   |          |          |      |       |
|-------------------|---|----------|----------|------|-------|
| U <sup>£</sup>    | - | 4355937  | 4357004  | 1068 | 94.10 |
| U <sup>£¥</sup>   | + | 4537377  | 4539308  | 1932 | 96.83 |
| U <sup>£</sup>    | - | 5367908  | 5368352  | 444  | 99.55 |
| U <sup>£¥*</sup>  | + | 5384045  | 5387314  | 3270 | 96.16 |
| U <sup>£¥</sup>   | + | 5494730  | 5495329  | 600  | 95.67 |
| U <sup>£</sup>    | - | 5537882  | 5538331  | 449  | 95.54 |
| U <sup>£¥\$</sup> | - | 5679045  | 5679704  | 660  | 91.92 |
| U <sup>£¥</sup>   | - | 5680622  | 5681500  | 879  | 95.56 |
| U <sup>£¥\$</sup> | + | 5950518  | 5951775  | 1258 | 93.13 |
| U <sup>£¥</sup>   | - | 6132643  | 6133267  | 625  | 99.36 |
| U <sup>£¥\$</sup> | - | 6536133  | 6538085  | 1953 | 87.11 |
| U <sup>£¥\$</sup> | + | 6720385  | 6721210  | 826  | 96.73 |
| U <sup>£</sup>    | + | 6958505  | 6960048  | 1543 | 95.46 |
| U <sup>£¥</sup>   | + | 7409968  | 7410136  | 169  | 95.83 |
| U <sup>£¥</sup>   | - | 8905908  | 8906162  | 255  | 98.43 |
| U <sup>£</sup>    | - | 9126245  | 9126804  | 559  | 95.40 |
| U <sup>£¥</sup>   | + | 9852927  | 9853034  | 108  | 89.81 |
| U <sup>£</sup>    | + | 10162660 | 10163444 | 784  | 96.30 |
| U <sup>£¥</sup>   | + | 10317347 | 10317513 | 167  | 82.04 |
| U <sup>£¥</sup>   | + | 10625041 | 10625415 | 375  | 94.31 |
| U <sup>£¥</sup>   | + | 10862415 | 10864252 | 1838 | 97.82 |
| U <sup>£</sup>    | - | 11024602 | 11025244 | 643  | 91.19 |
| U <sup>£\$</sup>  | - | 11023792 | 11024075 | 284  | 94.72 |
| U <sup>£</sup>    | - | 12899234 | 12899861 | 627  | 93.95 |
| U <sup>£¥\$</sup> | - | 13548152 | 13549068 | 917  | 95.85 |
| U <sup>¥</sup>    | - | 13613535 | 13613737 | 203  | 94.58 |
| U <sup>¥</sup>    | - | 15302493 | 15302886 | 394  | 97.21 |

§ the complete *helena* sequence

\* sequences with internal deletions and insertions

\$ sequences with internal deletions

# sequence with insertions

£ sequences truncated in 5'

¥ sequences truncated in 3'

Supplementary Table 2: *helena* copies in the *Drosophila melanogaster* sequenced genome

| chromosome       | strand | start    | stop     | length (bp) | %identity with the complete <i>helena</i> |
|------------------|--------|----------|----------|-------------|---|
| 2L <sup>\$</sup> | +      | 20553860 | 20554967 | 1108        | 94.00                                     |
| 2L               | +      | 20502277 | 20502573 | 297         | 76.00                                     |
| 2L*              | -      | 21790993 | 21793986 | 2994        | 61.00                                     |
| 2R*              | -      | 483468   | 488273   | 4805        | 83.00                                     |
| 2R*              | +      | 1630552  | 1632697  | 2146        | 64.66                                     |
| 2R*              | -      | 1841627  | 1843766  | 1907        | 68.81                                     |
| 3L <sup>#</sup>  | +      | 16603717 | 16607074 | 3358        | 65.00                                     |
| 3L               | +      | 18688097 | 18688680 | 583         | 70.05                                     |
| 3L <sup>\$</sup> | -      | 23487977 | 23490595 | 2619        | 93.00                                     |
| 3R               | -      | 484629   | 484719   | 91          | 79.00                                     |
| 3R               | -      | 401148   | 401432   | 285         | 92.00                                     |
| 3R*              | +      | 3313732  | 3314009  | 278         | 95.00                                     |
| 3R*              | -      | 3884273  | 3886085  | 1813        | 70.21                                     |
| 3R*              | -      | 7777247  | 7779392  | 2146        | 64.65                                     |
| X                | -      | 18697277 | 18697732 | 458         | 84.00                                     |
| X*               | -      | 22519248 | 22521013 | 1765        | 90.00                                     |
| X <sup>\$</sup>  | -      | 2255945  | 2256468  | 524         | 88.00                                     |
| U                | +      | 1548034  | 1548124  | 91          | 86.00                                     |
| U                | +      | 1572147  | 1572700  | 554         | 90.00                                     |
| U                | -      | 3163038  | 3163543  | 506         | 72.00                                     |
| U                | -      | 3164132  | 3164716  | 585         | 70.00                                     |
| U <sup>\$</sup>  | -      | 390489   | 392808   | 2320        | 87.00                                     |
| U <sup>\$</sup>  | +      | 4397920  | 4400343  | 2423        | 94.00                                     |
| U*               | +      | 5254588  | 5256562  | 1975        | 67.34                                     |
| U                | +      | 5813553  | 5813690  | 138         | 92.00                                     |
| U                | +      | 5814272  | 5814990  | 719         | 95.00                                     |

\* sequences with internal deletions and insertions

\$ sequences with internal deletions

# sequence with insertions

Supplementary Table 3: *Helena* copy number in *D. melanogaster* and *D. simulans* populations by in situ hybridization in polytene chromosomes

|                        |                            | <i>helena</i> copy number in polytene chromosomes | Additional Centromeric and Peri-centromeric staining |
|------------------------|----------------------------|---|--|
| <i>D. melanogaster</i> | Saudi Arabia               | 0   | P / C  |
|                        | Bolivia                    | 0   |  |
|                        | Brazzaville (Congo)        | 0   | P / C  |
|                        | Canton (China)             | 0   | P  |
|                        | Chicharo (Portugal)        | 0   | P  |
|                        | Reunion                    | 0   | C  |
|                        | Virasoro (Argentina)       | 1   | P  |
|                        | Vietnam                    | 0   | P  |
|                        | ISO                        | 0   | P/C  |
| <i>D. simulans</i>     | Valence (France)           | 13  |  |
|                        | Makindu (Kenya)            | 9   | C  |
|                        | Moscou (Russia)            | 10  |  |
|                        | Canberra (Australia)       | 11  |  |
|                        | Zimbawe                    | 13  |  |
|                        | Reunion                    | 10  |  |
|                        | Cann River (Australia)     | 12  |  |
|                        | Amieu (New Caledonia)      | 9   |  |
|                        | Eden (Australia)           | 10  | C  |
|                        | Papeete (French Polynesia) | 10  |  |



241  
chr3R\_1506433  
Brazzaville\_16  
Eden 3  
Cann-River 9  
Amieu 2  
Makindu 2  
Makindu 6  
Makindu 7  
Valence 6  
Makindu 8  
Makindu 22  
Canberra 5  
Cann-River 2  
Cann-River 6  
Cann-River 7  
Cann-River 5  
Cann-River 8  
Valence 5  
Valence 9  
Eden 1  
Valence\_13  
Eden 7  
Valence 4  
Valence 2  
Valence 1  
Valence 7  
Makindu cM1  
Makindu cM6  
Makindu 1  
Amieu 5  
Eden 11  
Cann-River 10  
Brazzaville 10  
Cann-River 3  
Cann-River 13  
Eden 4  
Makindu 10  
Makindu 11  
Brazzaville 11  
Brazzaville 12  
Brazzaville 8  
chrU 963982  
chr3L 18642496  
chrU 3975907  
chrU 10625041  
chrU 2316095  
chrU 4537377  
chrU 5384045  
chrU 5680622  
chrU 5950518  
chrU 6132643  
chrX 16602314  
chrU 6536133

301  
chr3R\_1506433  
Brazzaville\_16  
Eden 3  
Cann-River 9  
Amieu 2  
Makindu 2  
Makindu 6  
Makindu 7  
Valence 6  
Makindu 8  
Makindu 22  
Canberra 5  
Cann-River 2  
Cann-River 6  
Cann-River 7  
Cann-River 5  
Cann-River 8  
Valence 5  
Valence 9  
Eden 1  
Valence\_13  
Eden 7  
Valence 4  
Valence 2  
Valence 1  
Valence 7  
Makindu cM1  
Makindu cM6  
Makindu 1  
Amieu 5  
Eden 11  
Cann-River 10  
Brazzaville 10  
Cann-River 3  
Cann-River 13  
Eden 4  
Makindu 10  
Makindu 11  
Brazzaville 11  
Brazzaville 12  
Brazzaville 8  
chrU 963982  
chr3L 18642496  
chrU 3975907  
chrU 10625041  
chrU 2316095  
chrU 4537377  
chrU 5384045  
chrU 5680622  
chrU 5950518  
chrU 6132643  
chrX 16602314  
chrU 6536133

361  
chr3R\_1506433  
Brazzaville\_16  
Eden 3  
Cann-River 9  
Amieu 2  
Makindu 2  
Makindu 6  
Makindu 7  
Valence 6  
Makindu 8  
Makindu 22  
Canberra 5  
Cann-River 2  
Cann-River 6  
Cann-River 7  
Cann-River 5  
Cann-River 8  
Valence 5  
Valence 9  
Eden 1  
Valence\_13  
Eden 7  
Valence 4  
Valence 2  
Valence 1  
Valence 7  
Makindu cM1  
Makindu cM5  
Makindu 1  
Amieu 5  
Eden 11  
Cann-River 10  
Brazzaville 10  
Cann-River 3  
Cann-River 13  
Eden 4  
Makindu 10  
Makindu 11  
Brazzaville 11  
Brazzaville 12  
Brazzaville 8  
chrU 963982  
chr3L 18642496  
chrU 3975907  
chrU 10625041  
chrU 2316095  
chrU 4537377  
chrU 5384045  
chrU 5680622  
chrU 5950518  
chrU 6132643  
chrX 16602314  
chrU 6536133

421  
chr3R\_1506433  
Brazzaville\_16  
Eden 3  
Cann-River 9  
Amieu 2  
Makindu 2  
Makindu 6  
Makindu 7  
Valence 6  
Makindu 8  
Makindu 22  
Canberra 5  
Cann-River 2  
Cann-River 6  
Cann-River 7  
Cann-River 5  
Cann-River 8  
Valence 5  
Valence 9  
Eden 1  
Valence\_13  
Eden 7  
Valence 4  
Valence 2  
Valence 1  
Valence 7  
Makindu cM1  
Makindu cM6  
Makindu 1  
Amieu 5  
Eden 11  
Cann-River 10  
Brazzaville 10  
Cann-River 3  
Cann-River 13  
Eden 4  
Makindu 10  
Makindu 11  
Brazzaville 11  
Brazzaville 12  
Brazzaville 8  
chrU 963982  
chr3L 18642496  
chrU 3975907  
chrU 10625041  
chrU 2316095  
chrU 4537377  
chrU 5384045  
chrU 5680622  
chrU 5950518  
chrU 6132643  
chrX 16602314  
chrU 6536133

481  
chr3R\_1506433  
Brazzaville\_16  
Eden 3  
Cann-River 9  
Amieu 2  
Makindu 2  
Makindu 6  
Makindu 7  
Valence 6  
Makindu 8  
Makindu 22  
Canberra 5  
Cann-River 2  
Cann-River 6  
Cann-River 7  
Cann-River 5  
Cann-River 8  
Valence 5  
Valence 9  
Eden 1  
Valence\_13  
Eden 7  
Valence 4  
Valence 2  
Valence 1  
Valence 7  
Makindu cM1  
Makindu cM6  
Makindu 1  
Amieu 5  
Eden 11  
Cann-River 10  
Brazzaville 10  
Cann-River 3  
Eden 4  
Makindu 10  
Makindu 11  
Brazzaville 11  
Brazzaville 12  
Brazzaville 8  
chrU\_963982  
chr3L\_18642496  
chrU\_3975907  
chrU\_10625041  
chrU\_2316095  
chrU\_4537377  
chrU\_5384045  
chrU\_5680622  
chrU\_5950518  
chrU\_6132643  
chrX\_16602314  
chrU\_6536133

541  
chr3R\_1506433  
Brazzaville\_16  
Eden 3  
Cann-River 9  
Amieu 2  
Makindu 2  
Makindu 6  
Makindu 7  
Valence 6  
Makindu 8  
Makindu 22  
Canberra 5  
Cann-River 2  
Cann-River 6  
Cann-River 7  
Cann-River 5  
Cann-River 8  
Valence 5  
Valence 9  
Eden 1  
Valence\_13  
Eden 7  
Valence 4  
Valence 2  
Valence 1  
Valence 7  
Makindu cM1  
Makindu cM6  
Makindu 1  
Amieu 5  
Eden 11  
Cann-River 10  
Brazzaville 10  
Cann-River 3  
Eden 4  
Makindu 10  
Makindu 11  
Brazzaville 11  
Brazzaville 12  
Brazzaville 8  
chrU\_963982  
chr3L\_18642496  
chrU\_3975907  
chrU\_10625041  
chrU\_2316095  
chrU\_4537377  
chrU\_5384045  
chrU\_5680622  
chrU\_5950518  
chrU\_6132643  
chrX\_16602314  
chrU\_6536133

601  
chr3R\_1506433  
Brazzaville\_16  
Eden 3  
Cann-River 9  
Amieu 2  
Makindu 2  
Makindu 6  
Makindu 7  
Valence 6  
Makindu 8  
Makindu 22  
Canberra 5  
Cann-River 2  
Cann-River 6  
Cann-River 7  
Cann-River 5  
Cann-River 8  
Valence 5  
Valence 9  
Eden 1  
Valence\_13  
Eden 7  
Valence 4  
Valence 2  
Valence 1  
Valence 7  
Makindu cM1  
Makindu cM5  
Makindu 1  
Amieu 5  
Eden 11  
Cann-River 10  
Brazzaville 10  
Cann-River 3  
Eden 4  
Makindu 10  
Makindu 11  
Brazzaville 11  
Brazzaville 12  
Brazzaville 8  
chrU\_963982  
chr3L\_18642496  
chrU\_3975907  
chrU\_10625041  
chrU\_2316095  
chrU\_4537377  
chrU\_5384045  
chrU\_5680622  
chrU\_5950518  
chrU\_6132643  
chrX\_16602314  
chrU\_6536133

661  
chr3R\_1506433  
Brazzaville\_16  
Eden 3  
Cann-River 9  
Amieu 2  
Makindu 2  
Makindu 6  
Makindu 7  
Valence 6  
Makindu 8  
Makindu 22  
Canberra 5  
Cann-River 2  
Cann-River 6  
Cann-River 7  
Cann-River 5  
Cann-River 8  
Valence 5  
Valence 9  
Eden 1  
Valence\_13  
Eden 7  
Valence 4  
Valence 2  
Valence 1  
Valence 7  
Makindu cM1  
Makindu cM6  
Makindu 1  
Amieu 5  
Eden 11  
Cann-River 10  
Brazzaville 10  
Cann-River 3  
Eden 4  
Makindu 10  
Makindu 11  
Brazzaville 11  
Brazzaville 12  
Brazzaville 8  
chrU\_963982  
chr3L\_18642496  
chrU\_3975907  
chrU\_10625041  
chrU\_2316095  
chrU\_4537377  
chrU\_5384045  
chrU\_5680622  
chrU\_5950518  
chrU\_6132643  
chrX\_16602314  
chrU\_6536133

721  
chr3R\_1506433  
Brazzaville\_16  
Eden 3  
Cann-River 9  
Amieu 2  
Makindu 2  
Makindu 6  
Makindu 7  
Valence 6  
Makindu 8  
Makindu 22  
Cann-River 5  
Cann-River 2  
Cann-River 6  
Cann-River 7  
Cann-River 8  
Valence 5  
Valence 9  
Eden 1  
Valence\_13  
Eden 7  
Valence 4  
Valence 2  
Valence 1  
Valence 7  
Makindu cM1  
Makindu cM6  
Makindu 1  
Amieu 5  
Eden 11  
Cann-River 10  
Brazzaville 10  
Cann-River 3  
Cann-River 13  
Eden 4  
Makindu 10  
Makindu 11  
Brazzaville 11  
Brazzaville 12  
Brazzaville 8  
chrU 963982  
chr3L 18642496  
chrU 3975907  
chrU 10625041  
chrU 2316095  
chrU 4537377  
chrU 5384045  
chrU 5680622  
chrU 5950518  
chrU 6132643  
chrX 16602314  
chrU 6536133

781  
chr3R\_1506433  
Brazzaville\_16  
Eden 3  
Cann-River 9  
Amieu 2  
Makindu 22  
Makindu 6  
Makindu 7  
Valence 6  
Makindu 8  
Makindu 22  
Cannbera 5  
Cann-River 2  
Cann-River 6  
Cann-River 7  
Cann-River 8  
Valence 5  
Valence 9  
Eden 1  
Valence\_13  
Eden 7  
Valence 4  
Valence 2  
Valence 1  
Valence 7  
Makindu cM1  
Makindu cM6  
Makindu 1  
Amieu 5  
Eden 11  
Cann-River 10  
Brazzaville 10  
Cann-River 3  
Cann-River 13  
Eden 4  
Makindu 10  
Makindu 11  
Brazzaville 11  
Brazzaville 12  
Brazzaville 8  
chrU 963982  
chr3L 18642496  
chrU 3975907  
chrU 10625041  
chrU 2316095  
chrU 4537377  
chrU 5384045  
chrU 5680622  
chrU 5950518  
chrU 6132643  
chrX 16602314  
chrU 6536133

841  
chr3R\_1506433  
Brazzaville\_16  
Eden 3  
Cann-River 9  
Amieu 2  
Makindu 2  
Makindu 6  
Makindu 7  
Valence 6  
Makindu 8  
Makindu 22  
Cannbera 5  
Cann-River 2  
Cann-River 6  
Cann-River 7  
Cann-River 8  
Valence 5  
Valence 9  
Eden 1  
Valence\_13  
Eden 7  
Valence 4  
Valence 2  
Valence 1  
Valence 7  
Makindu cM1  
Makindu cM5  
Makindu 1  
Amieu 5  
Eden 11  
Cann-River 10  
Brazzaville 10  
Cann-River 3  
Cann-River 13  
Eden 4  
Makindu 10  
Makindu 11  
Brazzaville 11  
Brazzaville 12  
Brazzaville 8  
chrU 963982  
chr3L 18642496  
chrU 3975907  
chrU 10625041  
chrU 2316095  
chrU 4537377  
chrU 5384045  
chrU 5680622  
chrU 5950518  
chrU 6132643  
chrX 16602314  
chrU 6536133

901  
chr3R\_1506433  
Brazzaville\_16  
Eden 3  
Cann-River 9  
Amieu 2  
Makindu 2  
Makindu 6  
Makindu 7  
Valence 6  
Makindu 8  
Makindu 22  
Cannbera 5  
Cann-River 2  
Cann-River 6  
Cann-River 7  
Cann-River 8  
Valence 5  
Valence 9  
Eden 1  
Valence\_13  
Eden 7  
Valence 4  
Valence 2  
Valence 1  
Valence 7  
Makindu cM1  
Makindu cM6  
Makindu 1  
Amieu 5  
Eden 11  
Cann-River 10  
Brazzaville 10  
Cann-River 3  
Cann-River 13  
Eden 4  
Makindu 10  
Makindu 11  
Brazzaville 11  
Brazzaville 12  
Brazzaville 8  
chrU 963982  
chr3L 18642496  
chrU 3975907  
chrU 10625041  
chrU 2316095  
chrU 4537377  
chrU 5384045  
chrU 5680622  
chrU 5950518  
chrU 6132643  
chrX 16602314  
chrU 6536133

961

chr3R\_1506433  
Brazzaville\_16  
Eden 3  
Cann-River 9  
Amieu 2  
Makindu 2  
Makindu 6  
Makindu 7  
Valence 6  
Makindu 8  
Makindu 22  
Canberra 5  
Cann-River 2  
Cann-River 6  
Cann-River 7  
Cann-River 5  
Cann-River 8  
Valence 5  
Valence 9  
Eden 1  
Valence\_13  
Eden 7  
Valence 4  
Valence 2  
Valence 1  
Valence 7  
Makindu cM1  
Makindu cM6  
Makindu 1  
Amieu 5  
Eden 11  
Cann-River 10  
Brazzaville 10  
Cann-River 3  
Cann-River 13  
Eden 4  
Makindu 10  
Makindu 11  
Brazzaville 11  
Brazzaville 12  
Brazzaville 8  
chrU 963982  
chr3L 18642496  
chrU 3975907  
chrU 10625041  
chrU 2316095  
chrU 4537377  
chrU 5384045  
chrU 5680622  
chrU 5950518  
chrU 6132643  
chrX 16602314  
chrU 6536133

1021

chr3R\_1506433  
Brazzaville\_16  
Eden 3  
Cann-River 9  
Amieu 2  
Makindu 2  
Makindu 6  
Makindu 7  
Valence 6  
Makindu 8  
Makindu 22  
Canberra 5  
Cann-River 2  
Cann-River 6  
Cann-River 7  
Cann-River 5  
Cann-River 8  
Valence 5  
Valence 9  
Eden 1  
Valence\_13  
Eden 7  
Valence 4  
Valence 2  
Valence 1  
Valence 7  
Makindu cM1  
Makindu cM6  
Makindu 1  
Amieu 5  
Eden 11  
Cann-River 10  
Brazzaville 10  
Cann-River 3  
Cann-River 13  
Eden 4  
Makindu 10  
Makindu 11  
Brazzaville 11  
Brazzaville 12  
Brazzaville 8  
chrU 963982  
chr3L 18642496  
chrU 3975907  
chrU 10625041  
chrU 2316095  
chrU 4537377  
chrU 5384045  
chrU 5680622  
chrU 5950518  
chrU 6132643  
chrX 16602314  
chrU 6536133

1081

chr3R\_1506433  
Brazzaville\_16  
Eden 3  
Cann-River 9  
Amieu 2  
Makindu 2  
Makindu 6  
Makindu 7  
Valence 6  
Makindu 8  
Makindu 22  
Canberra 5  
Cann-River 2  
Cann-River 6  
Cann-River 7  
Cann-River 5  
Cann-River 8  
Valence 5  
Valence 9  
Eden 1  
Valence\_13  
Eden 7  
Valence 4  
Valence 2  
Valence 1  
Valence 7  
Makindu cM1  
Makindu cM5  
Makindu 1  
Amieu 5  
Eden 11  
Cann-River 10  
Brazzaville 10  
Cann-River 3  
Cann-River 13  
Eden 4  
Makindu 10  
Makindu 11  
Brazzaville 11  
Brazzaville 12  
Brazzaville 8  
chrU 963982  
chr3L 18642496  
chrU 3975907  
chrU 10625041  
chrU 2316095  
chrU 4537377  
chrU 5384045  
chrU 5680622  
chrU 5950518  
chrU 6132643  
chrX 16602314  
chrU 6536133

1141

chr3R\_1506433  
Brazzaville\_16  
Eden 3  
Cann-River 9  
Amieu 2  
Makindu 2  
Makindu 6  
Makindu 7  
Valence 6  
Makindu 8  
Makindu 22  
Canberra 5  
Cann-River 2  
Cann-River 6  
Cann-River 7  
Cann-River 5  
Cann-River 8  
Valence 5  
Valence 9  
Eden 1  
Valence\_13  
Eden 7  
Valence 4  
Valence 2  
Valence 1  
Valence 7  
Makindu cM1  
Makindu cM6  
Makindu 1  
Amieu 5  
Eden 11  
Cann-River 10  
Brazzaville 10  
Cann-River 3  
Cann-River 13  
Eden 4  
Makindu 10  
Makindu 11  
Brazzaville 11  
Brazzaville 12  
Brazzaville 8  
chrU 963982  
chr3L 18642496  
chrU 3975907  
chrU 10625041  
chrU 2316095  
chrU 4537377  
chrU 5384045  
chrU 5680622  
chrU 5950518  
chrU 6132643  
chrX 16602314  
chrU 6536133

1201  
chr3R\_1506433  
Brazzaville\_16  
Eden 3  
Cann-River 9  
Amieu 2  
Makindu 2  
Makindu 6  
Makindu 7  
Valence 6  
Makindu 8  
Makindu 22  
Canberra 5  
Cann-River 2  
Cann-River 6  
Cann-River 7  
Cann-River 8  
Valence 5  
Valence 9  
Eden 1  
Valence\_13  
Eden 7  
Valence 4  
Valence 2  
Valence 1  
Valence 7  
Makindu\_cM1  
Makindu\_cM6  
Makindu 1  
Amieu 5  
Eden 11  
Cann-River 10  
Brazzaville 10  
Cann-River 3  
Cann-River 13  
Eden 4  
Makindu 10  
Makindu 11  
Brazzaville 11  
Brazzaville 12  
Brazzaville 8  
chrU\_963982  
chr3L\_18642496  
chrU\_3975907  
chrU\_10625041  
chrU\_2316095  
chrU\_4537377  
chrU\_5384045  
chrU\_5680622  
chrU\_5950518  
chrU\_6132643  
chrX\_16602314  
chrU\_6536133

1261  
chr3R\_1506433  
Brazzaville\_16  
Eden 3  
Cann-River 9  
Amieu 2  
Makindu 2  
Makindu 6  
Makindu 7  
Valence 6  
Makindu 8  
Makindu 22  
Canberra 5  
Cann-River 2  
Cann-River 6  
Cann-River 7  
Cann-River 8  
Valence 5  
Valence 9  
Eden 1  
Valence\_13  
Eden 7  
Valence 4  
Valence 2  
Valence 1  
Valence 7  
Makindu\_cM1  
Makindu\_cM6  
Makindu 1  
Amieu 5  
Eden 11  
Cann-River 10  
Brazzaville 10  
Cann-River 3  
Cann-River 13  
Eden 4  
Makindu 10  
Makindu 11  
Brazzaville 11  
Brazzaville 12  
Brazzaville 8  
chrU\_963982  
chr3L\_18642496  
chrU\_3975907  
chrU\_10625041  
chrU\_2316095  
chrU\_4537377  
chrU\_5384045  
chrU\_5680622  
chrU\_5950518  
chrU\_6132643  
chrX\_16602314  
chrU\_6536133

1321  
chr3R\_1506433  
Brazzaville\_16  
Eden 3  
Cann-River 9  
Amieu 2  
Makindu 2  
Makindu 6  
Makindu 7  
Valence 6  
Makindu 8  
Makindu 22  
Canberra 5  
Cann-River 2  
Cann-River 6  
Cann-River 7  
Cann-River 8  
Valence 5  
Valence 9  
Eden 1  
Valence\_13  
Eden 7  
Valence 4  
Valence 2  
Valence 1  
Valence 7  
Makindu\_cM1  
Makindu\_cM6  
Makindu 1  
Amieu 5  
Eden 11  
Cann-River 10  
Brazzaville 10  
Cann-River 3  
Cann-River 13  
Eden 4  
Makindu 10  
Makindu 11  
Brazzaville 11  
Brazzaville 12  
Brazzaville 8  
chrU\_963982  
chr3L\_18642496  
chrU\_3975907  
chrU\_10625041  
chrU\_2316095  
chrU\_4537377  
chrU\_5384045  
chrU\_5680622  
chrU\_5950518  
chrU\_6132643  
chrX\_16602314  
chrU\_6536133

1381  
chr3R\_1506433  
Brazzaville\_16  
Eden 3  
Cann-River 9  
Amieu 2  
Makindu 2  
Makindu 6  
Makindu 7  
Valence 6  
Makindu 8  
Makindu 22  
Canberra 5  
Cann-River 2  
Cann-River 6  
Cann-River 7  
Cann-River 8  
Valence 5  
Valence 9  
Eden 1  
Valence\_13  
Eden 7  
Valence 4  
Valence 2  
Valence 1  
Valence 7  
Makindu\_cM1  
Makindu\_cM6  
Makindu 1  
Amieu 5  
Eden 11  
Cann-River 10  
Brazzaville 10  
Cann-River 3  
Cann-River 13  
Eden 4  
Makindu 10  
Makindu 11  
Brazzaville 11  
Brazzaville 12  
Brazzaville 8  
chrU\_963982  
chr3L\_18642496  
chrU\_3975907  
chrU\_10625041  
chrU\_2316095  
chrU\_4537377  
chrU\_5384045  
chrU\_5680622  
chrU\_5950518  
chrU\_6132643  
chrX\_16602314  
chrU\_6536133

1441

chr3R 1506433  
Brazzaville\_16  
Eden 3  
Cann-River 9  
Amieu 2  
Makindu 2  
Makindu 6  
Makindu 7  
Valence 6  
Makindu 8  
Makindu\_22  
Canberra 5  
Cann-River 2  
Cann-River 6  
Cann-River 7  
Cann-River 5  
Cann-River 8  
Valence 5  
Valence 9  
Eden 1  
Valence\_13  
Eden 7  
Valence 4  
Valence 2  
Valence 1  
Valence 7  
Makindu\_cM1  
Makindu\_cM6  
Makindu 1  
Amieu 5  
Eden 11  
Cann-River 10  
Brazzaville 10  
Cann-River 3  
Cann-River 13  
Eden 4  
Makindu 10  
Makindu 11  
Brazzaville 11  
Brazzaville 12  
Brazzaville 8  
chrU 963982  
chr3L 18642496  
chrU 3975907  
chrU 10625041  
chrU 2316095  
chrU 4537377  
chrU 5384045  
chrU 5680622  
chrU 5950518  
chrU 6132643  
chrX 16602314  
chrU 6536133

1501

chr3R 1506433  
Brazzaville\_16  
Eden 3  
Cann-River 9  
Amieu 2  
Makindu 2  
Makindu 6  
Makindu 7  
Valence 6  
Makindu 8  
Makindu\_22  
Canberra 5  
Cann-River 2  
Cann-River 6  
Cann-River 7  
Cann-River 5  
Cann-River 8  
Valence 5  
Valence 9  
Eden 1  
Valence\_13  
Eden 7  
Valence 4  
Valence 2  
Valence 1  
Valence 7  
Makindu\_cM1  
Makindu\_cM6  
Makindu 1  
Amieu 5  
Eden 11  
Cann-River 10  
Brazzaville 10  
Cann-River 3  
Cann-River 13  
Eden 4  
Makindu 10  
Makindu 11  
Brazzaville 11  
Brazzaville 12  
Brazzaville 8  
chrU 963982  
chr3L 18642496  
chrU 3975907  
chrU 10625041  
chrU 2316095  
chrU 4537377  
chrU 5384045  
chrU 5680622  
chrU 5950518  
chrU 6132643  
chrX 16602314  
chrU 6536133







## CONCLUSION

The *helena* element is common in *D. melanogaster* and *D. simulans* natural populations. However, *helena* is extinct in *D. melanogaster* since it is no longer able to transpose due to the lack of complete open reading frames. In *D. simulans*, few natural populations harbour complete copies of *helena* but all of these strains have numerous internally deleted copies of this element. Given the repetitive nature of TEs, recombination is often the cause of copy internal deletion (Bennetzen, 2002). More specifically, solo LTR formation allows the genome to keep the regulatory parts of the elements and throw the coding parts away. In barley, TE conversion into solo LTRs is faster than element insertion, and therefore counteracts with TE invasion (Shirasu et al., 2000). In rice, 190Mb of LTR elements were deleted in the last 8 million years (Ma et al., 2004). Important deletions can also induce TE extinction and have severe consequences on the host genome as described in rodents where massive radiation seems to concord in time with TE extinction (Casavant et al., 2000). Thanks to comparative genomics, we know that precise TE deletion can also occur when target sites duplications (created upon retrotransposon insertions) are the objects of recombination (van de Lagemaat et al., 2005). The host genome is therefore able to eliminate TE sequences. Nevertheless, one should not forget that recombination events can also be harmful for the host genome when unequal homologous recombination occurs (Hedges and Deininger, 2007). Indeed 0.3% of human genetic diseases are probably related to Alu unequal homologous recombination (Deininger and Batzer, 1999). As a conclusion, our data show that deletion occurs inside the TE sequence, which is not a common described phenomenon especially in the case of LINE elements, such as *helena*, in which only 5' truncations are expected from the transposition mechanism.

Sequence deletion is therefore a very effective system for suppressing the transposition activity of TEs, but the consequences can also be deleterious for the host. Genome regulation system is also capable of avoiding transposition by controlling the epigenetic environment of TE copies while keeping their sequences unmodified. Full-length TEs might therefore be protected from deletion by epigenetic means but also might be harmless for the host genome. Epigenetic regulation of TEs involves chromatin remodeling factors, DNA methylation and non coding small RNAs (Lippman and Martienssen, 2004). In several species, specific conserved repressive histone marks (N-terminal tail post translational modifications) were observed in nucleosomes wrapping TEs. Usually, histone methylation in lysine residues (H3K9m, H3K27m, H4K20m) are typical from a closed chromatin conformation in contrast

to acetylation of histones and methylation in H3K4, which are often observed in open chromatin structures.

Epigenetic natural variation is a common feature among species but is hardly ever studied (Richards, 2008). *D. melanogaster* and *D. simulans* natural populations harbor an important TE copy number variation as described above (15% and 5% respectively). Interestingly, only 5,4% and 2,7% of the euchromatin are composed of TEs in *D. melanogaster* and *D. simulans* sequenced genomes (Clark et al., 2007). TEs seem therefore rather heterochromatic but what strikes us the most is that from the 15% of TEs present in *D. melanogaster*, the majority could be heterochromatic while *D. simulans* might harbor as much TEs in the euchromatin as in the heterochromatin. Thus, we wonder if natural variation in the histone association with TEs could explain such scenarios. We analyzed four elements in both species natural populations for TE copy number, TE expression, and TE association to histone marks. Also, we further analysed the impact of such natural variation in genome evolution. Hence, a bibliographic assay follows the histone project.

## For submission to plos Genetics

Variation of histone modifications associated to transposable elements in natural populations of *D. melanogaster* and *D. simulans*

**Running head** : Variation of TE / histone association in *Drosophila*

Rita Rebollo<sup>1</sup>, Flora Begeot<sup>2</sup>, Béatrice Horard<sup>3</sup>, Marion Delattre<sup>2</sup>, Eric Gilson<sup>3</sup> and Cristina Vieira<sup>1\*</sup>

1. CNRS, UMR558, Laboratoire de Biométrie et Biologie Evolutive, Université Lyon 1, Villeurbanne, France

2. NCCR “Frontiers in Genetics”, Department of Zoology and Animal Biology, University of Geneva, Geneva, Switzerland

3. Laboratoire de Biologie Moléculaire de la Cellule-UMR 5239 CNRS/ENS LYON/ Université LYON 1/ HCL. 165, Chemin du Grand Revoyet, 69495 Pierre Bénite or 46, allée d’Italie, 69364 Lyon cedex 07, France

\* Address for correspondence: [vieira@biomserv.univ-lyon1.fr](mailto:vieira@biomserv.univ-lyon1.fr)

## Abstract

The presence of transposable elements (TEs) in genomes is certainly one source of genetic variability creating genetic novelty and affecting genome evolution. Examination of sequenced genomes revealed an important diversity in TE families, copy number and localization between several species. For instance, the twin species, *D. melanogaster* and *D. simulans*, display different amount of TEs despite sharing the same TE composition. Furthermore, previous analyses of natural populations of *D. simulans* have shown high polymorphism regarding TE copy number inside the species. Several factors might influence TE diversity and abundance in a genome. Among molecular mechanisms that could be a source of variation in TE success, is epigenetics. In this report, we present the first analysis of the epigenetic status of various TE families in several populations of *D. melanogaster* and *D. simulans*. Our data demonstrate that although TEs are mainly decorated with repressive

histone marks we observed intra and inter specific variations suggesting that association between TEs and epigenetic marks is dependent on the TE family/Genome analyzed.

### **Author Summary**

Transposable elements (TEs) are DNA sequences able to move inside the genome. Transposition and recombination events of TEs may induce mutations and spread copies in the genome. Activity of TE copies is dependent on chromatin localization and since the major part of transposition events is lineage specific, TE copy localization is variable between species, and inside species. We use the model system composed of the fruit flies, *D. melanogaster* and *D. simulans*, along with natural populations from both species in order to compare histone post translational modifications and TE dynamics. Our data suggests that different associations of histone marks exists between these very close species. Also, natural populations of both *D. melanogaster* and *D. simulans* do not behave equally to TE presence since TE expression varies between populations analyzed. This research shows how population studies in epigenetics are mandatory to understand how TEs are contributing to genome evolution.

## Introduction

DNA repeats increase species genetic variability and are therefore an important motor of genome evolution (Biemont and Vieira, 2006). We concentrate this study on transposable elements (TEs), which are an important component of DNA repeats. TEs are mobile sequences, that may induce mutations through their mobilisation and through copy recombination often in an individual specific manner (Hedges and Deininger, 2007). TEs may offer new genetic features (Muotri et al., 2007; Sinzelle et al., 2009) and regulatory sequences (Marino-Ramirez et al., 2005; Muotri et al., 2007; Polavarapu et al., 2008) participating on the formation and remodelling of host gene networks (McClintock, 1984; Feschotte, 2008). Factors that govern intra- and inter-species TE diversity are complex. They are likely a combination of intrinsic properties of TEs themselves (transposition mechanism, infectivity...), properties of host ecology (species effective size and structure...), and properties of the genome (TE regulation, gene density, genome size...). Among molecular mechanisms that could be a source of variation of TE expression/activity, is epigenetics. Indeed, it is now well documented among the whole eukaryote taxa that regulation of TE activity involves epigenetic pathways as chromatin remodelling processes, DNA methylation and non coding small RNAs (Lisch, 2009; Obbard et al., 2009). TEs are maintained in the genome but are usually silenced through epigenetics, protecting the genome from negative effects of transposition/recombination, but keeping the potential for creating variability. When DNA sequences (such as genes) are positioned near or within heterochromatin, no/low expression can be detected as described for position effect variegation (PEV) models (Girton and Johansen, 2008). The localization of TEs within particular chromatin environment might therefore be correlated to TE expression and further activity.

The basic unit of chromatin is the nucleosome, composed of an octamer of histones (H2A, H2B, H3 and H4, twice each) (Ito, 2007). Double stranded DNA wraps nucleosomes every 147bp and with other non histone proteins, forms the chromatin (Ito, 2007). Histones have a free N terminal tail which is the target of many post-translational modifications, as for example methylation and acetylation of lysine residues. Each histone modification is the target of chromatin remodelling factors and allow for chromatin conformation to change. Generally, acetylation of lysines are associated to an active chromatin (euchromatin) while methylation, depending on the lysine modified, can correlate to activation (lysine 4 of histone 3) or repression (lysine 9 and 27 of histone 3) of gene expression (Kim et al., 2009). TEs are

considered as repressed elements, i.e., packaged in heterochromatic domains harbouring repressive histone marks (Kondo and Issa, 2003; Ding et al., 2007; Bernatavichute et al., 2008). *Drosophila* constitutive heterochromatin is enriched in H3K9me2 while facultative heterochromatin is labelled with H3K27me3 (Ebert et al., 2006). Furthermore heterochromatin is often enriched in TEs suggesting a role for histone modifications and chromatin related proteins in TE regulation. (Ebert et al., 2006; Schulze and Wallrath, 2007). In this regard, several reports indicate that 1. TEs are associated with histone modifications and proteins typical of heterochromatic domains (de Wit et al., 2005; Klenov et al., 2007; Matyunina et al., 2008; Fablet et al., 2009; Phalke et al., 2009) and 2. Distinct chromatin patterns might be observed between different TE families but also inside a TE family (Mito et al., 2005; Fablet et al., 2009). The epigenetic regulation system, even if still rarely studied in natural populations, has been shown to be flexible (Richards, 2008). For instance, the epigenetic polymorphism observed between twin brothers (Ballestar, 2009), or the presence of different patterns of gene methylation between *Arabidopsis thaliana* ecotypes (Tanurdzic et al., 2008) exemplifies epigenetic natural variation. No study on natural population has ever been made on TE histone modifications restricting our knowledge of the impact, and the dynamic of chromatin remodelling on TE diversity and stabilization in genomes. However, the need of population studies in the field of epigenetics is already claimed by many, and the importance of such experiments is no longer questionable (Richards, 2006; Bossdorf et al., 2008; Johannes et al., 2008). *Drosophila* natural populations are an excellent model to analyse such questions. Despite the fact that *Drosophila* has fewer copies of TEs when compared to other organisms such as humans (15% for *Drosophila* and 50% for humans), it has a high percentage of spontaneous mutations due to TE activity (Biemont and Vieira, 2006). Hence, *Drosophila* has putatively active elements (2004) and constitute an interesting model to the study of the impact of TEs on genetic variability and genome evolution. Studies on natural populations are approaches allowing to apprehend the genetic novelty associated with TE diversity and abundance (Fablet et al., 2006; Rebollo et al., 2008). *D. melanogaster* and *D. simulans*, for instance, contain the same TE families, with in most of the cases more than 90% of sequence identity (Bartolome et al., 2009). However, an overrepresentation of almost all TEs is observed in *D. melanogaster* (Vieira et al., 1999). Natural populations of *D. melanogaster* are relatively homogeneous for TEs, since copies are present at high number in all strains analyzed (Vieira et al., 1999). In contrast, *D. simulans* natural populations are highly polymorphic since for a given element high copy number can be observed in one strain while another one is completely devoided (Vieira et al., 1999). These observations were

established from the estimation of TE copy number through *in situ* hybridization experiments by which centromeric, telomeric and dense heterochromatic regions can not be individually counted (Vieira et al., 1999). Therefore, variation in copy number between *D. melanogaster* and *D. simulans* natural populations reflects only euchromatic copies. Also, euchromatic TE representation is different between both species sequenced genomes (~5% and 2% for *D. melanogaster* and *D. simulans* respectively (Clark et al., 2007)) suggesting a role for TE chromatin localization in host response.

We propose to investigate somatic TE association with a subset of histone marks in regard of TE copy number variation and TE transcription in natural populations of both *D. melanogaster* and *D. simulans*. The TEs chosen for this assay were copy number polymorphic between natural populations of *D. simulans* (Vieira et al., 1999) and chosen as being LTR elements (*412*, *tirant* and *roo*) and LINE-like elements (*F*). No DNA transposons are copy number polymorphic between natural populations of *D. simulans* and were therefore excluded from our analysis. For each TE family, we observed transcriptionally derepressed populations in both species of *Drosophila*. Transcription profiles are specific for each natural population and also each TE analysed. Histone patterns are polymorphic between natural populations and distinct between TE families. We confirm that different associations to previously described regulatory marks exist between both species. Also, and more importantly, single natural populations can act as transcriptionally active lines for specific TEs, showing that mobile elements can counteract global host regulation system, and that natural populations may be used as natural mutants. These findings enlighten the nuanced world of natural variation of TE epigenetic marks and show that even with a highly described epigenetic regulatory system, genomes can be invaded and TEs can colonize populations.

## Results and discussion

### 1. TE occurrence in natural populations of *D. melanogaster* and *D. simulans*

All four elements analysed are retrotransposons that either have long terminal repeats (LTR) (two *copia*-like elements, *412* and *roo*, a *gypsy*-like element, *tirant*) or not (LINE-like element *F*) (Figure 1). They harbour open reading frames (ORFs) that may code for retrovirus like proteins as GAG, POL and ENV (Meyerowitz and Hogness, 1982; Yuki et al., 1986; Fablet et al., 2006; Tchurikov and Kretova, 2007). Initial *in situ* hybridization experiments revealed that these elements are over-represented in *D. melanogaster* strains compared to *D. simulans* (Vieira et al., 1999). For example, Vieira and collaborators observed 18 times more *F* copies in *D. melanogaster* than in *D. simulans* natural populations (mean copy number between natural populations analysed). *412*, *tirant* and *F* were also described as being copy number polymorphic between natural populations of *D. simulans* (Vieira et al., 1999; Borie et al., 2000; Fablet et al., 2006). For example, the *F* element varied between 37 copies in a few *D. simulans* natural populations to being completely absent from others (Vieira et al., 1999). In contrast, *Roo* is one exception, since it was detected in high copy number in both species and all natural populations analysed (Vieira et al., 1999). All data collected by Vieira et al. was obtained through *in situ* hybridization of polytene chromosomes of *Drosophila* salivary glands using large probes corresponding to almost full length elements (Vieira et al., 1999). *In situ* hybridization only allowed the authors to count for copies present in the arms of polytene chromosomes (so called euchromatic). However, all four elements are present on chromosome arms and in putatively constitutive heterochromatin regions (pericentromeric staining on polytene chromosomes). In order to apprehend genome wide variation in TE copy number, we used quantitative PCR to amplify an internal region of *412*, *tirant*, *roo* and *F* elements (Figure 2). As expected, *D. melanogaster* presents more TE copies than *D. simulans* (~4 times more, supplementary Figure 1). Interestingly, *roo* element, which was as highly represented in *D. melanogaster* as in *D. simulans* natural populations, has a significantly lower copy number in *D. simulans* (~9 fold). In addition, copy number polymorphism previously observed between natural populations of *D. simulans* (Vieira et al., 1999) were not reproduced. Instead, either no copy variation between *D. simulans* populations is observed (*F* Kruskal-Wallis chi p value ~0,14), or a different pattern of copy number distribution is noted among *D. simulans* natural populations (*412* and *tirant*). For example, the Canberra *D.*

*simulans* strain was previously described as harbouring more than 60 copies of *412* (Vieira et al., 1999) and we observe an almost complete lack of this element in this population. The same can be described for Grand Ferrade *D. simulans* and the *F* element. We are quantifying the presence of a unique small internal region of each element (150 bp maximum) and any deleted copy in that region will not be amplified by this technique. In contrast, using a complete copy as a probe, Vieira et al were able to count every copy presented in the euchromatic regions, even importantly deleted copies. The comparison of both datasets strongly suggests that *412*, *tirant*, *roo* and *F* are internally deleted in *D. simulans* natural populations. The analysis of both sequenced genomes confirmed that a large proportion of these elements present internal deletions in *D. simulans* genome (will be described in a separate report). These observations are consistent with previous hypothesis of an active degradation process in natural populations of *D. simulans*, effective on various elements such as *412* and *tirant* but also on other elements as the LINE-like element *helena* (Mugnier et al., 2005; Fablet et al., 2006; Rebollo et al., 2008). Furthermore, some lines, previously described as devoid of copies of a given TE family, present copies through real time PCR. Indeed, *F* and *tirant* elements are amplified in *D. simulans* populations such as Grand Ferrade, Makindu or Zimbabwe initially described as empty. We interpret these differences as heterochromatic copies located in centro and telomeric domains and hence ignored in the previous *in situ* hybridization experiments.

We confirmed previous data showing that *D. melanogaster* has generally more copies of all TE families analyzed than *D. simulans*, and we show that intact copies in *D. simulans* are probably in low copy number. Are these patterns the result of differential transcriptional activity of TEs in the two *Drosophila* species? Full length copies of all TEs analysed can be observed in the sequenced genomes of both species (Bartolome et al., 2009) and copy number variation is observed for some elements in natural populations of both species. Are these elements still producing transcripts (full length or deleted ones) and are they accounting for copy number variation of each TE?

## **2. Expression of *412*, *tirant*, *roo* and *F* in natural populations of *D. melanogaster* and *D. simulans***

In order to test if some of the copies present in natural populations are transcriptionally active, we quantified TE expression. We extracted total RNAs from embryos and amplified total cDNA by quantitative PCR (Figure 3). The amplified region in all elements is the same

as the region targeted for copy number analysis described above (Figure 1). Each TE family presented a particular expression pattern in *D. melanogaster* and *D. simulans* natural populations but globally, *D. melanogaster* Senegal strain harboured more transcripts than all other strains.

*D. melanogaster* populations, Senegal and Chicharo, harboured constant patterns between all elements: Chicharo displays always lower expression level than Senegal (-3 fold, Kruskal Wallis p value  $1,6 \cdot 10^{-3}$ ). Such difference can not be explained by copy number variation for all elements in *D. melanogaster*, except for the 412 family (Figure 2 supplementary materials).

The majority of natural populations of *D. simulans* harbour transcripts for at least one of the four elements. Nevertheless, no correlation between copy number and transcription level can be visualised. However, it is interesting to note that the expression of 412 in *D. simulans* natural populations has a tendency to follow Vieira et al. *in situ* hybridization quantification data [12] (Figure 3 supplementary materials). One hypothesis to explain such tendency is that 412 would be transcribed at high levels from few full-length copies and that transcription may be associated to high transposition rate. In this context, each newly transposed copy would be internally deleted by the *D. simulans* genome, while the master copies remain intact. The second hypothesis is to think that fragments of 412 observed by *in situ* hybridization are producing transcripts regardless of copy quality (full-length, truncated and internally deleted copies). Such transcripts might be originated from host genes and external promoters. Further analyses on transposition rates are being held along with RACE-PCRs in order to identify external promoters. *Tirant* has a specific expression pattern in *D. simulans*, where a few natural populations, harbouring almost the same number of *tirant* copies (Makindu, Canberra and Zimbabwe), are either silenced or extremely transcriptionally active (700 fold between Makindu and Canberra for instance). The *roo* element has almost the same dynamics as *tirant*, since several natural populations of *D. simulans* harbour more transcripts than both natural populations of *D. melanogaster*. Both data suggests that *D. simulans* even though having fewer copies than *D. melanogaster*, does present transcripts for TEs, comforting the idea that few copies may be responsible for high transcription rate. The *F* element has no copy number variation for *D. simulans* and no transcript variation (p values  $> 0,1$  for Kruskal Wallis tests).

As a conclusion, expression of a TE family in a given population does not correlate with copy number and we hypothesize that different regulation systems may exist inside the same species. Moreover, populations of *D. simulans* harbouring the same range of TE copies can either be transcriptionally active (for example, *tirant* in *D. simulans* Makindu strain) or

silenced (*tirant* in Canberra *D. simulans* strain). *412*, *tirant*, *roo* and *F*, do not have a preferential insertion site in the host genome, they are observed in both heterochromatic and euchromatic regions. Epigenetic regulation of TEs is well known, and chromatin remodelling processes are thought to control gene expression. In *D. simulans*, *412* correlation between fragments present in the chromosome arms and expression suggests that the more *412* copies are present in the euchromatin regions, the higher the expression. Also, we previously showed that in *D. simulans* Makindu, *tirant* copies embedded in H3K9me2 heterochromatic domains are transcriptionally inert while copies associated with transcription are decorated with H3K27me3 (Fablet et al., 2009). Do TE families present distinct chromatin status between species and between populations? Do different TE families harbour different chromatin patterns inside a population?

### **3. Histone modifications within *412*, *tirant*, *roo* and *F* elements in natural populations of *D. melanogaster* and *D. simulans***

#### **a. Chromatin immunoprecipitation of histone marks associated to TEs**

ChIP analysis coupled with qPCR allowed us to observe permissive (H3K4me2) and repressive marks (H3K9me2 and H3K27me3) of histones on TE copies during embryogenesis. It is important to note that the region amplified by qPCR is near or overlapping the region already described for copy number and expression assays (Figure 1). Also, quantification of the immunoprecipitated material is relative to each TE copy number in each natural population (qPCR in the input samples). Plus, for all ChIP reactions, antibodies recognizing the histone 3 allowed us to check for nucleosome presence and lack of variation (Figure 4 supplementary materials). All elements analyzed were not associated with the histone modification typical of active regions, consistent with the common hypothesis that TEs present in the genome are globally in a repressed state (Figure 5 supplementary materials). Heterochromatin can be either constitutive (centromeres, telomeres, repeat-rich regions) or facultative (for instance, cell cycle and developmental stage dependent) (Trojer and Reinberg, 2007). H3K27me3 is typical from facultative heterochromatin in *Drosophila* (supplementary figure 6) (Czermin et al., 2002). Patterns obtained for H3K27me3 enrichment are TE family- and population-specific (Figure 4). *412* and *tirant* are globally more enriched in H3K27me3 in *D. melanogaster* natural populations than in *D. simulans* ones while no difference is observed between both species for *roo* and *F*. H3K27me3 enrichment does not

correlate with transcript variation of the four elements analysed. Indeed, *D. melanogaster* populations, although displaying different amount of transcripts, are both associated to H3K27me3. *D. simulans* natural populations harbour distinct patterns of H3K27me3 on different TE families including populations as enriched as *D. melanogaster*, and others completely devoid of such histone mark. Strikingly, the most transcriptionally active *412* population of *D. simulans*, Canberra, has the highest association of H3K27me3 among all natural populations. Similarly, *tirant* transcriptionally active line, Makindu, also harbors the highest H3K27me3 association. It is important to note that, as described by Fablet et al, bivalent marks as H3K27me3 and small association to H3K4me2, along with low association with H3K9me2, are observed in *tirant* expressed copies (Fablet et al., 2009).

H3K9me2 is typical from repressive constitutive heterochromatic regions (supplementary figure 6), as telomeres, centromeres, the 4<sup>th</sup> chromosome and the Y chromosome in *Drosophila* (Riddle et al., 2009). Similarly to H3K27me3, population- and TE-specific profiles are observed for H3K9me2 in embryos (Figure 4). Enrichment for H3K9me2 for TEs are weaker than compared to constitutive heterochromatic regions but are significant (supplementary figure 6). H3K9me2 patterns observed are different from the H3K27me3 ones. *Roo* presents a very low association with H3K9me2 in both species suggesting that *roo* elements are mainly outside of constitutive heterochromatin, regardless of the species analyzed and that these elements are more decorated with H3K27me3 than H3K9me2. *412* and *F* are more represented in the constitutive heterochromatin of *D. melanogaster* than *D. simulans* but this does not explain both transcription profiles. *Tirant* is equally enriched in H3K9me2, and small variations can be observed between natural populations.

Our ChIP coupled with qPCR approach allowed us to apprehend the mean status of all individual elements within one specific TE family. This cluster analysis does not allow the detection of differences that may be present at one particular chromosomal locus (TE single copy) in each natural population. In order to have an individual picture, we observed, on polytene chromosomes, through ImmunoFish, the associations between the different copies present on the chromosome arms, and histone post-translational modifications. We analyzed H3K27me3 pattern in both natural populations of *D. melanogaster* and in polymorphic natural populations of *D. simulans*. Colocalization between elements and H3K27me3 was only observed for a few copies of the *412* element in Senegal (*D. melanogaster*) and Canberra (*D. simulans*) lines (Figure 5). All other elements present on the chromosomes arms were devoid of H3K27me3 modification. These observations suggest that if H3K27me3 is acting as a repressive mark, then expression of the elements analyzed can originate from copies on the

chromosome arms devoid of any H3K27me3 modification and probably H3K9me2. One could wonder which histone marks are indeed associated to transcriptionally active copies. TEs analysed in *D. melanogaster* were globally more enriched in heterochromatic marks than *D. simulans* TEs. Are the different histone profiles between populations a result from differential expression of enzymes in charge of chromatin modification?

#### **b. Variation in key players of chromatin remodelling and other epigenetic pathways**

In order to apprehend variation in the deposition of histone modifications, we assayed for transcriptional activity of genes involved in post-translational modification of histones and other epigenetic pathways (Figure 6). Enhancer of zeste (E(z)), responsible for methylation of H3K27me3, Su(var)3-9, responsible for H3K9me2 (Ebert et al., 2006) were significantly overexpressed in *D. melanogaster* compared to *D. simulans* natural populations. One could ask if lower expression of both enzymes could be correlated to lower association to H3K27me3 and H3K9me2 in *D. simulans*. We assayed for other epigenetic players, in order to check for variation between both species (supplementary figure 7). RPD3, a histone deacetylase (HDAC1 homologous), ASH1 (a multi histone methyltransferase that acts in enhancing active patterns), Su(var)4-20 (H4K20me3) and HP1 (heterochromatin protein 1) expression do not present different profiles for *D. melanogaster* or *D. simulans*. It is interesting to note that H3K9me2 is often described in association with HP1 and our data show that despite having different expression patterns for Suvar3(9), HP1 is constant between natural populations and both species. Ash1 and Su(var)4-20 are extremely down regulated in a few *D. simulans* natural populations. MBD23 (methyl binding DNA protein) and DNMT2 (DNA methyl transferase, results not shown) do not present any variation between both species and natural populations. These data shows therefore that there are significant differences between the two species analyzed, *D. melanogaster* and *D. simulans* and that again, intra specific variation can be observed.

## Conclusion

Our global analysis of TEs in *Drosophila* natural populations as illustrated in Figure 7 points to several facts. The first general conclusion is that inter and intra species variations are observed at transcriptional and epigenetic levels. TE families present distinct patterns of regulation, and variation does exist between copies of a given TE family. Although some TE expression coincides with reduced levels of repressive marks (*roo* in *D. simulans* for example), such state is not the general rule. We demonstrate that some copies of TE families are mainly devoid of repressive histone marks in the genome. One may suggest that transcripts emerge preferentially from these “nude” copies. Indeed the second conclusion is about the analysis of two natural populations of *D. melanogaster* (Chicharo and Senegal) showing that each TE family presents distinct expression profiles in different populations, with transcription levels of all elements analyzed, highly different between the two populations of *D. melanogaster* species. Two hypotheses can be made to explain this observation: first, high transcription level in Senegal emerge from copies devoid of repressive histone modifications, while low transcription level in Chicharo results from packaging in repressive chromatin of mostly all TE copies present in the genome. However, we observed that the average enrichments for the histone methylation marks are largely similar between these two *D. melanogaster* populations. This suggests that an alteration of the histone methylation pattern is not the major cause of differential transcription between the two populations. If, histone post-translational modification pathways are equally effective in both *D. melanogaster* natural populations, repression of TEs in Chicharo might involve additional regulatory pathway. In this regard, a particular attention should be paid to the production of small interfering RNAs (siRNA) since several recent studies reveal the central role played by small RNAs in the genome, as TE silencers. The *F*, *412*, *tirant* and *roo* elements are present in flamenco and/or 42AB regions from which emerge a high number of germinal and somatic small RNAs (Brennecke et al., 2008; Malone et al., 2009). Chicharo line could thus be strongly silenced through small RNAs produced from 42AB and the *flamenco* locus. It should be noticed that no analysis has ever been performed on natural populations, and that we have no information on the variability of the insertions that can be found in the flamenco locus. Further analysis of siRNA production will thus be helpful to enlighten silencing of TEs in natural populations. Third, some *D. simulans* populations present elevated levels of *412* and *tirant* transcripts. Interestingly, in these populations the studied TEs present more prominent association with H3K27me3 than H3K9me2. Nevertheless, since amplification of

immunoprecipitated fragments of ChIP is not copy specific, one could suggest that the expressed TE copies are not associated to H3K27me3. This hypothesis is moreover in agreement with our results from ImmunoFish experiments. Copy localization of TEs in the two transcriptionally active genomes of *D. simulans* is mandatory to make ChIP copy specific and therefore realize the *in vivo* distribution of histone marks with TE copies. For example, we previously reported that an expressed *tirant* copy is preferentially decorated with H3K27me3 and H3K4me2 but not at all H3K9me2 compared to other copies (Fablet et al., 2009). Unfortunately, TE copies are very well conserved and transcript analysis does not allow us to distinguish between copies.

The comparative analysis of TE copy number by *in situ* hybridization published before (Vieira et al., 1999) and the quantitative PCR assays done in this project have shown that *D. simulans* strains harbour rearranged or deleted copies. Indeed, the *in situ* hybridization experiments performed with full length probes of the elements do not reflect the integrity of the copies detected in one genome. This has already been suggested by other works (Mugnier et al., 2005; Rebollo et al., 2008) in which deleted copies have been described, and definitely raises the question of TE regulation in *D. simulans*. Is this enigmatic TE specific deletion mechanism random or is it targeting copies present preferentially in chromosome arms (euchromatic) or in centromeric regions (heterochromatic)? We noticed that strains devoid of any copies on chromosomal arms (Vieira et al., 1999) display qPCR amplification products indicating that copies are present elsewhere in the genome. In contrast, strains harboring high number of euchromatic copies were almost not amplified by qPCR. From these observations it is tempting to hypothesize that the deletion mechanism is targeting essentially euchromatic copies.

Given the diversity of TEs and the variable degree they populate genomes, several pathways have emerged to restrict the deleterious effects of “uncontrolled” activity. From our results we propose two non-exclusive pathways. Our observations and a previous report (Phalke et al., 2009), suggest that local chromatin environment defined by histone modifications might control TE expression in *D. melanogaster* and *D. simulans*. Although the general idea is that TEs are embedded in repressive chromatin, no general rule appears from our analysis. Some copies might be packaged in constitutive repressed chromatin (H3K9me2, H3K27me3, H4K20me2) although others might be more permissive to transcriptional elongation (absence of H3K27me3 and/or H3K9me2). This pathway relies on chromatin remodellers activity such as histone methyltransferase activity and on the use of small RNAs emerging from the transposons to guide silencing. In addition, as illustrated in *D. simulans*, a deletion

mechanism targeting all TE families tend to restrict the deleterious activity inherent to full length copies. In contrast to *D. melanogaster*, *D. simulans* populations appear as a panel of several extinction profiles of TEs with a lot of truncated copies. In a few strains we still observe high transcription but low supposed full length copy number, and in others, no transcription and only copies embedded in heterochromatic regions. Based on our observations, we hypothesize that in *D. simulans*, copies present in the chromosome arms are potentially active and also the preferential targets of the deletion mechanism, while copies that are conserved in heterochromatin regions, are the target of repressive histone marks (H3K9me2 and H3K27me3), and are protected from the deletion processes. Further investigations are required to determine how the deletion mechanism is set up in *D. simulans* natural populations. How heterochromatic copies are maintained in the genome? In the modern view of epigenetics, small RNAs have the most important role to play: they can bring specificity to the system, start the system and end the system. If, as showed above, TEs are present in the heterochromatin, are still transcribing at a low rate, they could target the deletion mechanism to copies in the chromosome arms. It is thus essential to verify if the deleted copies are indeed exclusive from euchromatin and if the heterochromatic copies are able to produce small RNAs. Natural population studies offer therefore the perfect model to study such a system.

Genetic diversity is increased by TE in natural populations. For the first time, we identified broad epigenetic variability associated to TE variation. Therefore, TEs will not only provide important sources of genetic variability for the host but also allow a second level of organization and regulation through histone modifications. It is important to note that such conclusions are only possible thanks to a populational approach where diversity can be observed directly in nature. Since TEs play a role in regulation of gene networks (Feschotte, 2008), because they are multicopies and genome widespread, one could easily suggest that histone modifications associated to TEs are highly important to genome compartmentalization and differential regulation between populations. *D. melanogaster* originated in East Africa and has colonized the world in a recent past (Lachaise and Silvain, 2004). The dynamics of TEs in this species is starting to be well known and it is well described that TEs are very recent and have been submitted to recent bursts of transposition and horizontal transfers (Lerat et al., 2003; Bartolome et al., 2009). *D. simulans* is also an East African species but the colonization of the world started later and there are still regions where no *D. simulans* are found. TE dynamics in *D. simulans* also seems to be very different from *D. melanogaster*, since very few full length copies are described. Former population genetics models have

proposed that the differences in the amount of TEs between the two species could be explained by differences in species effective size, with *D. simulans* having a more important size and a more efficient selection against TEs. Even if this statement is true, which is far from being confirmed (Nolte and Schlotterer, 2008), we shouldn't expect differences in the quality and structure of the elements. Our data show that the analysis of TE dynamics will also depend on the fine mechanism of regulation such as the epigenetic pathways, that were never included in the population genetic models, and that will allow variability in the expression levels of specific elements in specific populations. Once the system breaks down and that a TE copy is transcribed, one might observe increase in copy number. The understanding of TE impacts on genome evolution and their implications on gene networks should associate ecology, epigenetics and natural population (Vieira et al., 2009).

## Materials and methods

### 1. Natural populations

We worked on fly samples collected from several geographically distinct regions. *D. melanogaster* natural populations were collected from Portugal (Chicharo) and Africa (Senegal). *D. simulans* natural populations are from Kenya (Makindu), Zimbabwe, Australia (Canberra), French Polynesia (Papeete) and France (Grand Ferrade). These populations were maintained in the laboratory as isofemale lines or small mass cultures with around 50 pairs in each generation. Flies stocks were maintained at 24°C.

### 2. TE copy number estimation

DNA was extracted from 14h to 16h embryos using “DNAeasy blood and tissue kit” from Qiagen, three times for each natural population. For each sample we used DNA diluted at 0.16ng/μl. Linear real time PCR was performed using Power SYBR Green Master Mix (Applied Biosystems), on a SDS 7900 HT instrument (Applied Biosystems) with the following parameters: 50°C for two minutes, 95°C for ten minutes, and 45 cycles of 95°C 15 seconds–60°C one minute. Each genomic region was tested with specific primers designed using the program Primer Express v 2.0 (Applied Biosystems) with default parameters. Primers sequences are available upon request. Triplicates of samples and triplicates of PCR were performed and the results obtained for each tested genes were normalized with three or four control genes treated in parallel (RP49, RNA pol II, EFG1). Raw Ct values obtained with SDS 2.2 (Applied Biosystems) were imported in Excel and normalization factor and fold changes were calculated using the GeNorm method (Vandesompele et al., , 2002). Real-time PCR and data analysis were performed at the Genomics Platform, NCCR “Frontiers in Genetics” (<http://www.frontiers-in-genetics.org/genomics.htm>).

### 3. TE expression quantification

RNA was isolated from 14h to 16h embryos. Extraction of RNA was done using TRIzol<sup>®</sup> reagent (Invitrogen) according to the manufacturer’s recommendation followed by chloroform /Isoamyl-alcohol (24:1) purification. After DNase treatment (Ambion DNA free<sup>™</sup>), the

OD260/280 (interval 1.9-2.1) of the RNA samples were determined by spectrophotometry. The integrity of the RNA was assessed by Agilent 2100 Bioanalyser (Agilent Technologies Inc, Palo Alto CA). For each sample, one  $\mu\text{g}$  of total RNA was used to make cDNA using random hexamers and the Superscript II reverse transcriptase (Invitrogen). Further analysis of transcripts was done by quantitative PCR as described above (including primers).

#### 4. Transcription of epigenetic factors

Total RNA was extracted from 14h to 16h embryos from all natural populations analysed (two biological replicates for each strain) with the RNeasy protect mini kit from Qiagen.  $1\mu\text{g}$  of total RNA extracts were treated with the Ambion's DNA-free kit. ThermoScript RT-PCR system from Invitrogen was used to synthesize two different cDNA (55°C for 90 min and 85°C for 5 min): a control reaction with no retrotranscriptase to test DNA contamination, a pool of total cDNA synthesized with a mix of oligo-dt/random primers 1:1. The cDNA samples were diluted 80 fold, and PCR was carried out using QuantiTect SYBR Green PCR kit (Roche) on the LightCycler (Roche) using primers specific from each enzyme analysed. Primers were chosen surrounding introns in order to amplify 150-250bp fragments of cDNA (primers available upon request). Genes analysed were chose as being part of chromatin remodelling processes, as Su(var)3-9, Ash1, E(z), HP1, HDAC3, Suvar420 and other epigenetic pathways as DNA methylation with MBD23 and dDNMT2. Quantitative PCR cycling conditions were 5 min at 95°C (1 cycle), 15 s at 95°C, followed by 10 s at 60°C and 20 s at 72°C (50 cycles). Reactions were done in duplicate, and standard curves were calculated from serial dilutions of specific amplified PCR fragments. The quantity of the transcripts was estimated relative to the RP49 and 18sRNA expression. Relative quantification was calculated as described above for TE expression.

#### 5. Histone modification within TEs

Extraction of chromatin from 14h to 16h *D. simulans* embryos and immunoprecipitation were adapted from (Sandmann et al., 2006). Cell lysis buffer was changed to 5mM PIPES pH 8, 85 mM KCl, 0.5% Nonidet P-40 supplemented with protease inhibitors. Chromatin was sheared with a Bioruptor sonicator water bath (Diagenode, Liège, Belgium) for 6 X (30-s on/30-s off) cycles at high power in order to have random fragments from 1 kb to 500 bp. Sheared chromatin was incubated overnight at 4°C with antibodies recognizing H3K9me2

(Millipore, Billerica, MA, USA; 07441), H3K27me3 (Millipore 07449), H3K4me2 (Millipore 07030), H3 (Abcam, Cambridge, UK; ab1791), and rabbit IgG (Sigma-Aldrich, St. Louis, MO, USA; I5006). The antigen-antibody complexes were washed as described before (Sandmann et al., 2006), but a second washing solution was modified as followed: TE 2X, 500 mM NaCl, 1% Triton, 0.1% SDS. To quantify each IP, real-time PCR was performed using QuantiTect SYBR Green PCR kit (Qiagen) on a MXP3000P PCR system (Stratagene, La Jolla, CA, USA). Reactions were done in duplicates, and standard curves were calculated on serial of input chromatin. To evaluate the relative enrichment of each TE after IP, we calculated the difference in cycles between the IP-enriched sample and the input DNA for TE copies and for a control (*actin*-CG4027). Primers were chosen in the coding region for each TE (Figure 1) and are available upon request.

## 6. Colocalisation of facultative heterochromatin and TE copies

Immunofish was adapted from (Lavrov et al., 2004). Fly stocks were maintained at 18°C during this experiment. Salivary glands were extracted from third-instar larvae in 1%NP40 and 0,6% NaCl. Cross linked salivary glands (2% Formaldehyde for 2 minutes) were crushed and fixated on slides for 3 minutes. Slides were conserved less than one week at 4°C in PBS. Dried Slides were incubated in 2X SSC for 45 min at 65 °C. After being dehydrated in cold ethanol baths, polytenes were denatured for 10 min in 0,07N NaOH at room temperature. After being washed with 2X SSC, slides were dehydrated a second time. Each TE probe was labelled during a PCR with DIG DNA labelling mix from Roche. Primers used for probe amplification target the internal region of all elements (Figure 1) and are available upon request. Slides were incubated over night at 37°C in a humid dark chamber with a denaturated mix of hybridization buffer (Formamide + 50% Dextran sulfate sodium salt + 20xSSC), hareng sperm and DIG labelled probe. The next morning, slides were washed in 2xSSC, four times at 42°C and 15min at room temperature in PBS. Slides were than incubated for one hour with blocking solution (PBS, 5% BSA, 0,1% Triton ) at room temperature. Histone modification H3K27me3 was targeted using Millipore 07449 antibody over night at 4°C, diluted 100 times in blocking solution. PBS washing steps the next morning were followed by deep washing (300mM of NaCl, 0,2% NP40, 0,2% Tween and PBS, second wash gas 400mM NaCl) in a shaking table for 15min each. Polytenes were incubated sequentially with Anti Dig monoclonal from Roche (1/25), AlexaFluor 555 Donkey Anti mouse IgG (1/100) and AlexaFluor 488 Donkey anti rabbit IgG (1/100) for 2h each. Deep washing of slides once again was carried and followed

by PBS washing. DAPI staining was carried with Vectashield Hard Set, AbCys mounting medium for fluorescence from Vector laboratories (1:3 with DAPI Vectashield Hard Set, AbCys). The slides were observed with AxioImager Z1 fluorescent microscope (Zeiss). The image obtained was treated with ImageJ software.

### **Acknowledgements**

We thank Nelly Bulet, Olaf Nickel, Dr Sameer Phalke, Professor Dr Gunter Reuter, Dr Jany Peng and Dr Gary Karpen for their valuable help and Dr Marie Fablet for all the comments on previous versions of this manuscript. This work was funded by CNRS ANR, Rhone Alpes region project, FINOVI and Fapesp. Pr Eric Gilson is an associate member of European Union 6<sup>th</sup> framework program *The Epigenome Network of Excellence* (NoE) and his laboratory was supported by European Union 6<sup>th</sup> framework program grant RISCRAAD and ARECA framework program from Canceropole Lyon Auvergne Rhône Alpes.

## Figure legends

### Figure 1

#### TE schematic representation

TEs analysed are LTR elements (*412*, *tirant* and *roo*) or LINE-like elements (*F*). Boxes represent open reading frames (ORF), a part from the LTR (long terminal repeats) box. ORF may code for retrovirus-like proteins GAG, POL and ENV. Primers used for quantitative PCR amplification (estimation of copy number, TE expression and enrichment in specific histone marks) are showed in black arrows. The probe used to observe colocalization of histone marks and TE single copy is represented as a green line.

### Figure 2

Copy number estimation of *412*, *tirant*, *roo* and *F* from *in situ* hybridization in polytene chromosomes (adapted from (Vieira et al., 1999)) and quantitative PCR.

Right graphs are adapted from (Vieira et al., 1999) and show chromosomal arm copies (euchromatic) for the four elements analyzed. Note that *D. melanogaster* has more elements than *D. simulans* and intra specific variation is observed for *D. simulans*. Left graphs show quantification of copy internal regions through quantitative PCR. Absence or different patterns of variation are observed for *D. simulans* natural populations. Blue bars represent *D. melanogaster* natural populations and red bars *D. simulans*. Please refer to materials and methods for qPCR relative quantification and, note that reference genes are equally represented between both species and natural populations allowing us to compare all strains together (Supplementary Figure 1).

### Figure 3

#### Transcription of TEs in natural populations of *D. melanogaster* and *D. simulans*

*D. melanogaster* (bleu bars) harbors transcription for the four elements analyzed. *D. simulans* (red bars) has very active and silenced populations for all LTR elements. Expression of all TEs is variable between both species and between all natural populations analyzed. No common pattern is observed for a specific natural population or a specific TE. Please refer to materials and methods for qPCR relative quantification and, note that reference genes are equally expressed between both species and natural populations allowing us to compare all strains together (Supplementary Figure 1).

#### Figure 4

H3K27me3 and H3K9me2 enrichment on TEs in natural populations of *D. melanogaster* and *D. simulans*.

Variation of H3K27me3 and H3K9me2 is observed between *D. melanogaster* (blue) and *D. simulans* (red) and between all natural populations analysed. *412* and *tirant* are polymorphic for H3K27me3 but *roo* and *F* harbour lower fold changes. The *roo* element has the highest enrichment rate for H3K9me2 while the *F* element has the lowest. Please, note that reference genes are equally enriched between both species and natural populations allowing us to compare all strains together (supplementary figure 1).

#### Figure 5

Colocalisation of H3K27me3 with *412* element in natural populations of *D. simulans* and *D. melanogaster*

Only Senegal from *D. melanogaster* and Canberra from *D. simulans* have shown colocalisation of facultative heterochromatic mark H3K27me3 and *412* copies.

#### Figure 6

Quantification of transcripts of genes involved in H3K27me3 and H3K9me2 chromatin post-translational modifications. Please, note that reference genes are equally expressed between both species and natural populations allowing us to compare all strains together (supplementary Figure 1).

#### Figure 7

##### Table of results

Blue bars represents *D. melanogaster* natural population and red ones for *D. simulans*'. For each analysis (Copy number, transcripts, H3K27me3 or H3K9me2 results) and each element (*412*, *tirant*, *roo* and *F*), the highest strain enrichment is chosen as being 100% and used for normalisation of other strains.

## Figures

Figure 1

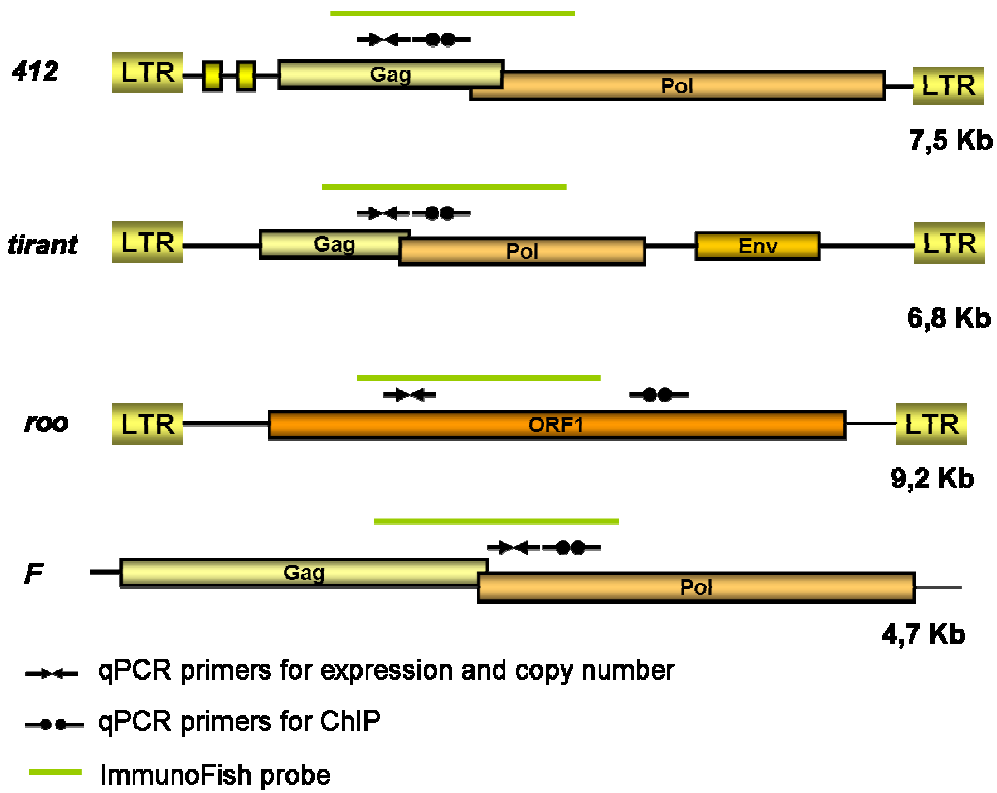


Figure 2

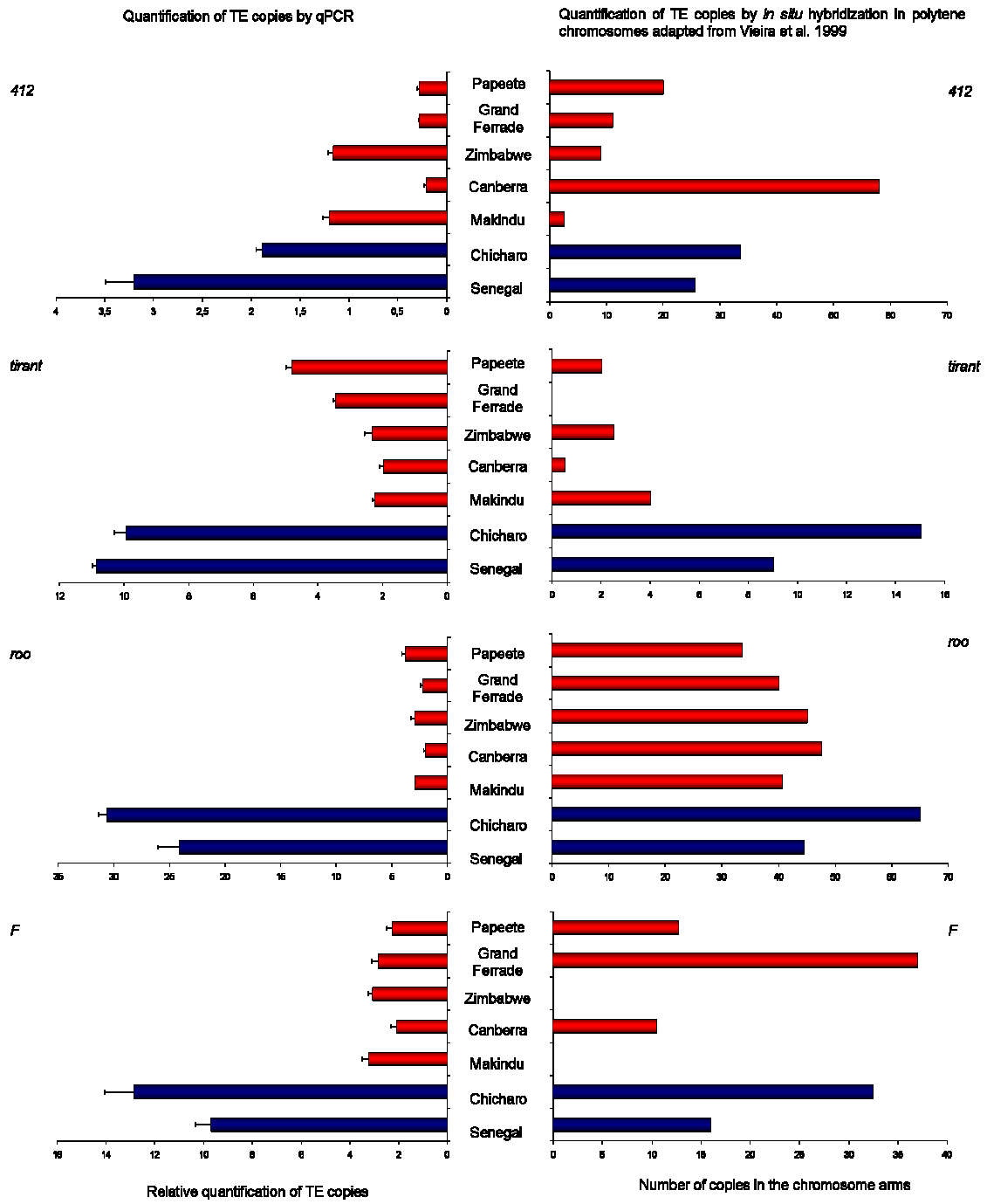


Figure 3

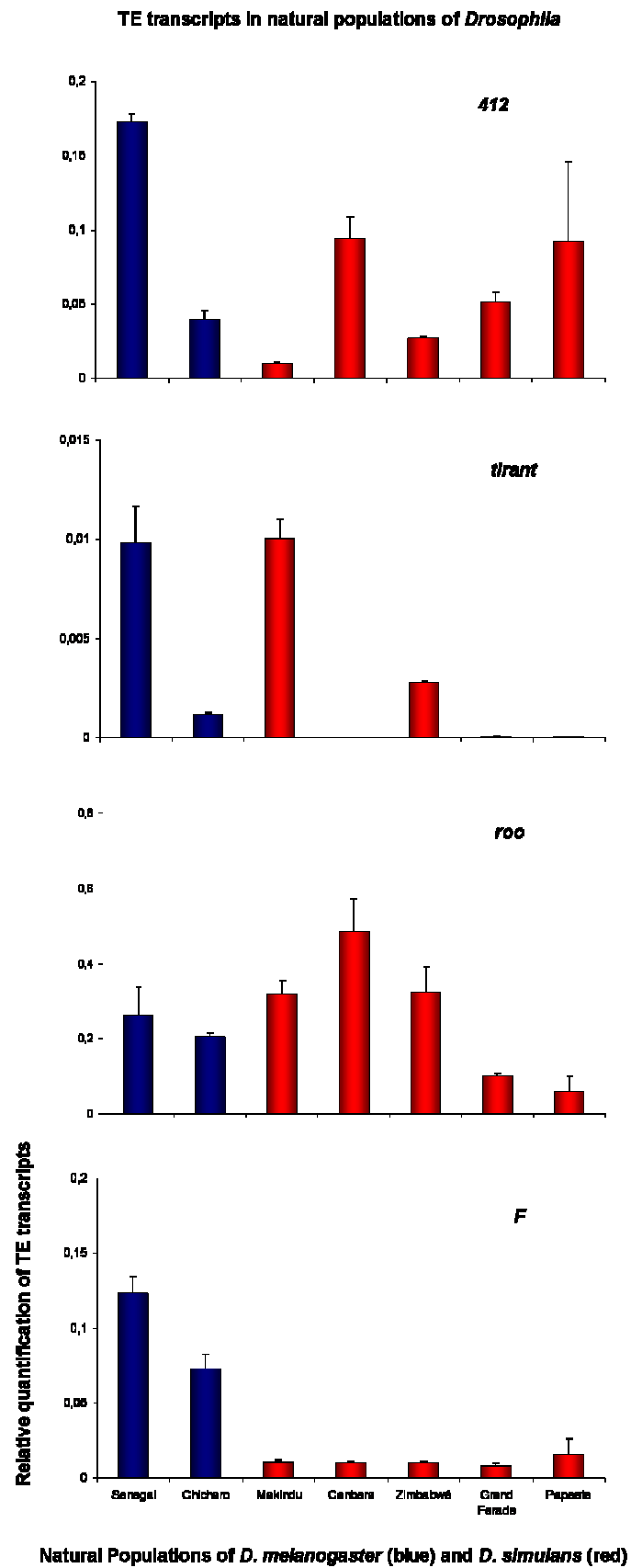
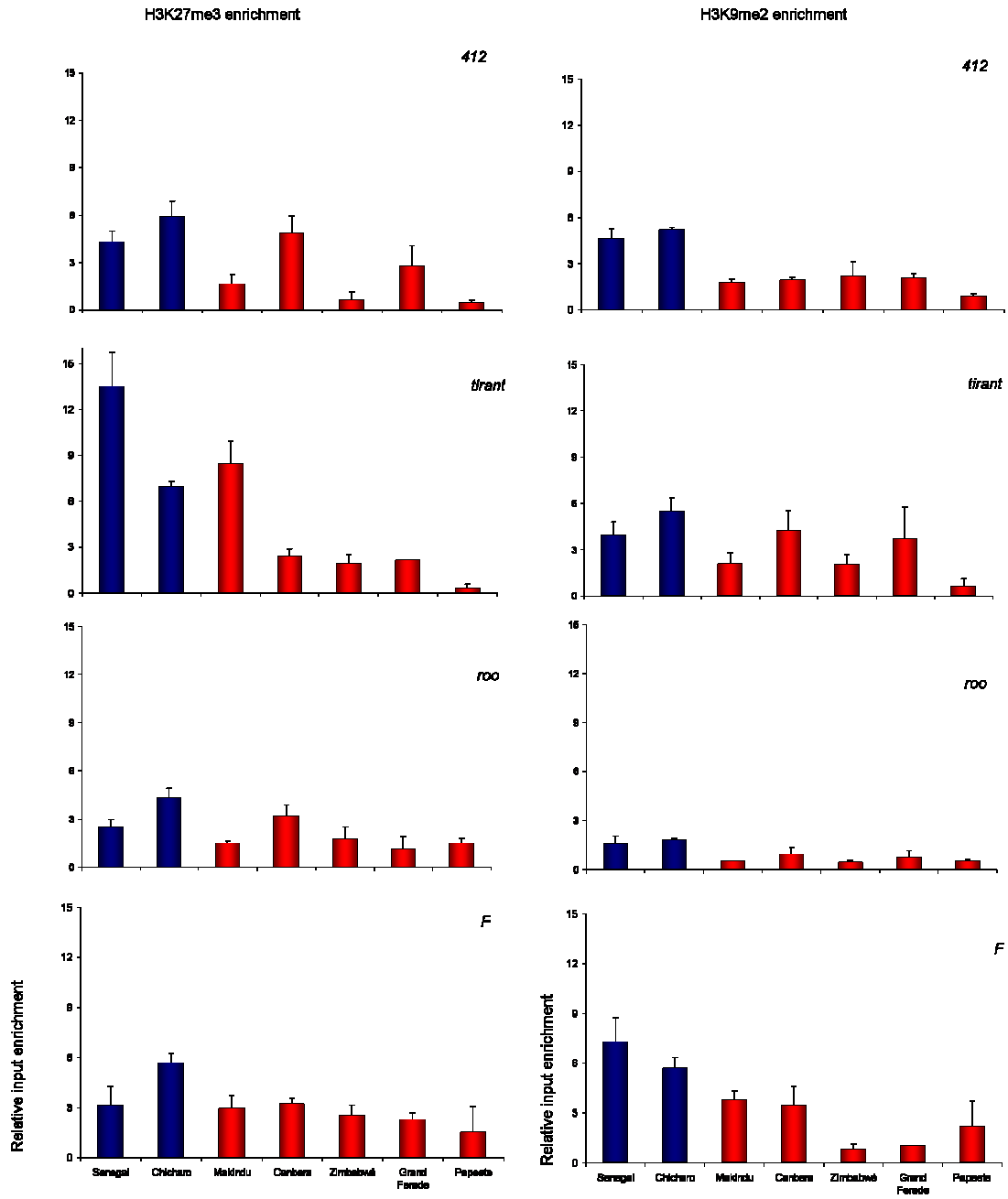


Figure 4



Natural Populations of *D. melanogaster* (blue) and *D. simulans* (red)

Figure 5

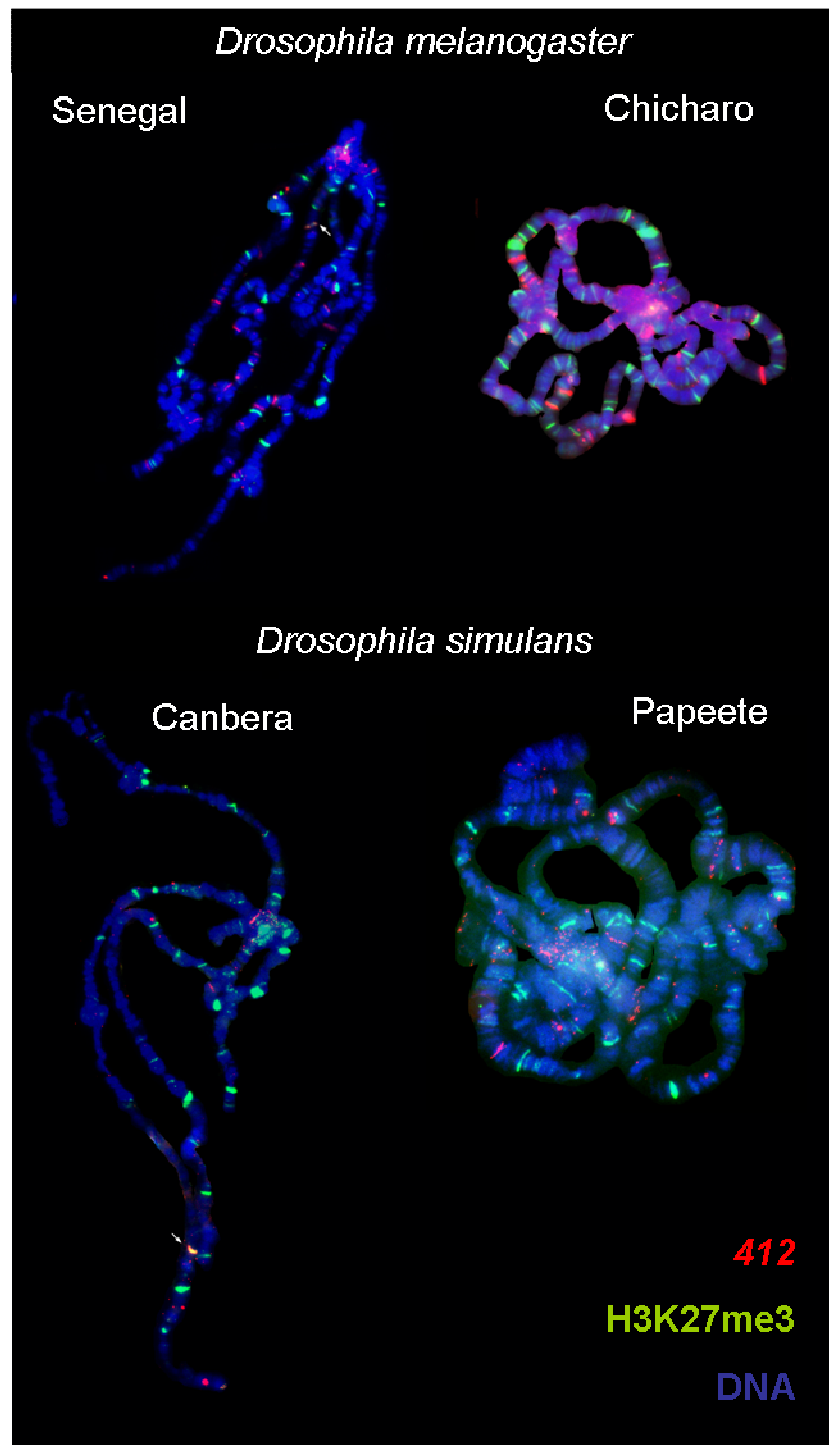


Figure 6

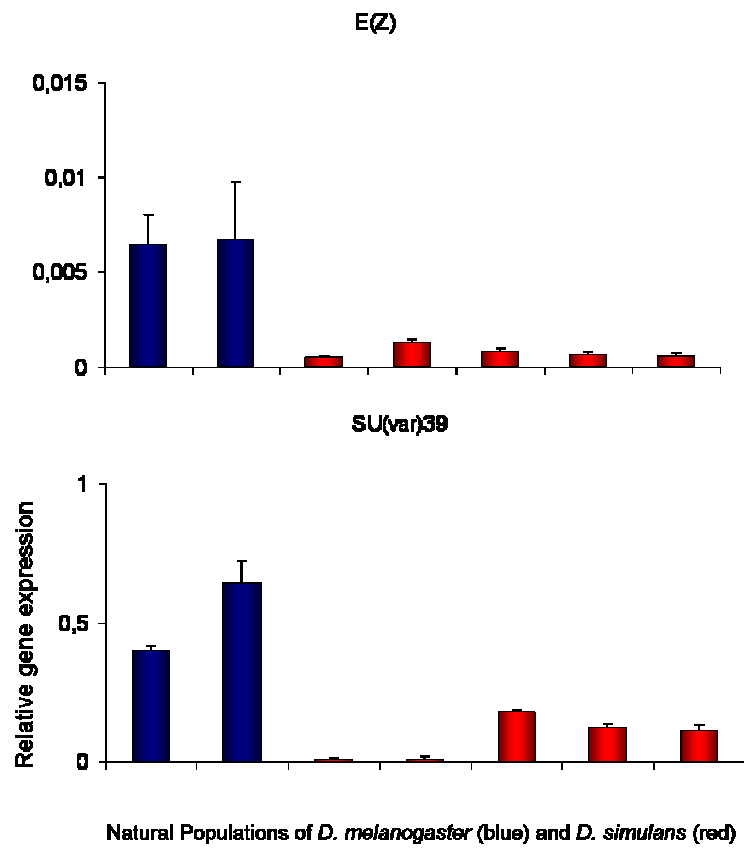
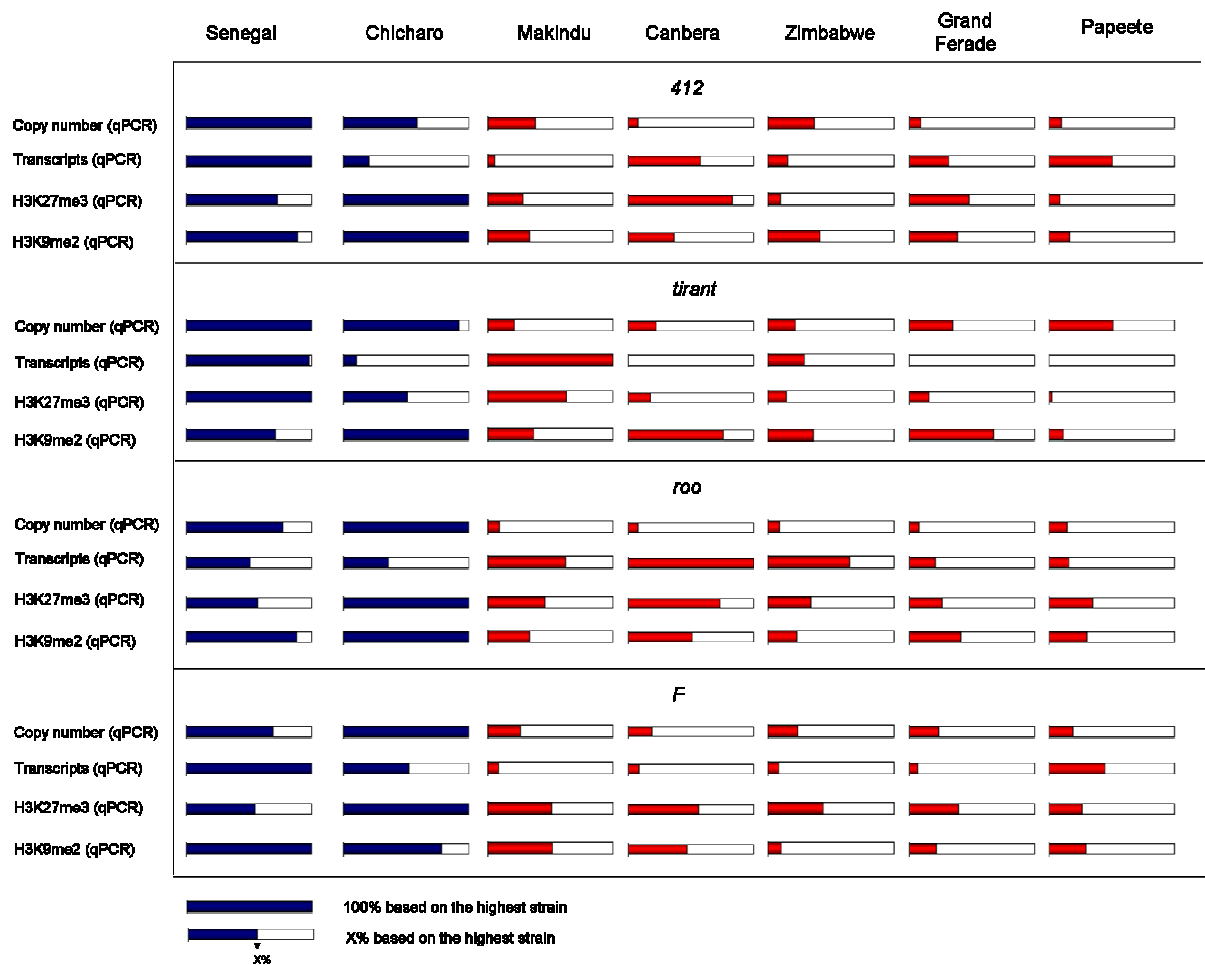


Figure 7



## Supplementary materials legends

### Figure 1

Ct comparison between *D. melanogaster* and *D. simulans* for reference genes.

For all experiments (copy number, transcription estimation and ChIP) reference genes used for both species, *D. melanogaster* and *D. simulans* are either equally represented, equally transcribed or equally associated to histone post translational modifications. Such data allow us to compare natural populations of *D. simulans* with populations of *D. melanogaster*. Also, *D. melanogaster* TE copy number present earlier Cts than *D. simulans* natural populations showing that both natural populations of *D. melanogaster* have more TE copies than *D. simulans*.

### Figure 2

Correlation between copy number and TE transcription for *D. melanogaster* natural populations.

### Figure 3

Tendency of *412* transcription to be higher in high copy number natural populations of *D. simulans*.

### Figure 4

Histone 3 enrichment for TEs in natural populations of *D. melanogaster* and *D. simulans*. Please, note that reference genes are equally enriched between both species and natural populations allowing us to compare all strains together (supplementary figure 1).

### Figure 5

H3K4me2 ChIP data

A. H3K4me2 lack of enrichment for TEs in natural populations of *D. melanogaster* and *D. simulans*. B. Enrichment of H3K4me2 was nevertheless observed in absolute quantification of the reference gene (actin) compared to TEs. C. H3K4me2 was also relative enriched to actin when *rp49* was analyzed and absent in a already described X region. Please, note that reference genes are equally enriched between both species and natural populations allowing us to compare all strains together (supplementary figure 1).

Figure 6

Specific H3K9me2 association to constitutive heterochromatin and comparison with TEs

A. *Satellite 1688* located in constitutive heterochromatin of *D. melanogaster* Chicharo strain is enriched in H3K9me2 but not H3K27me3. B. Higher enrichment of *satellite 1688* compared to TEs in *D. melanogaster* Chicharo strain.

Figure 7

Quantification of transcripts of genes involved chromatin post-translational modifications. Please, note that reference genes are equally expressed between both species and natural populations allowing us to compare all strains together (supplementary Figure 1).

Supplementary materials

Figure 1

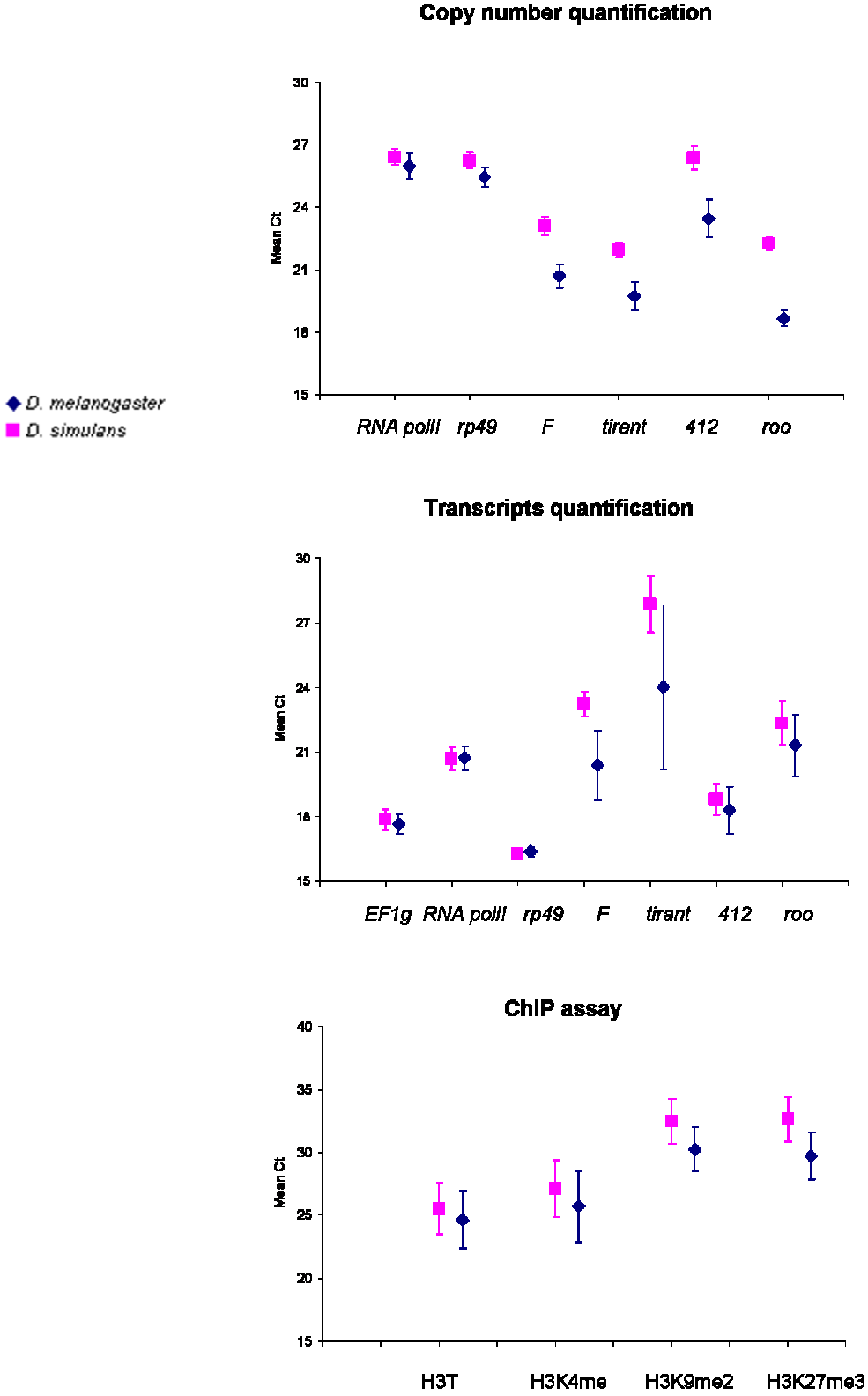


Figure 2

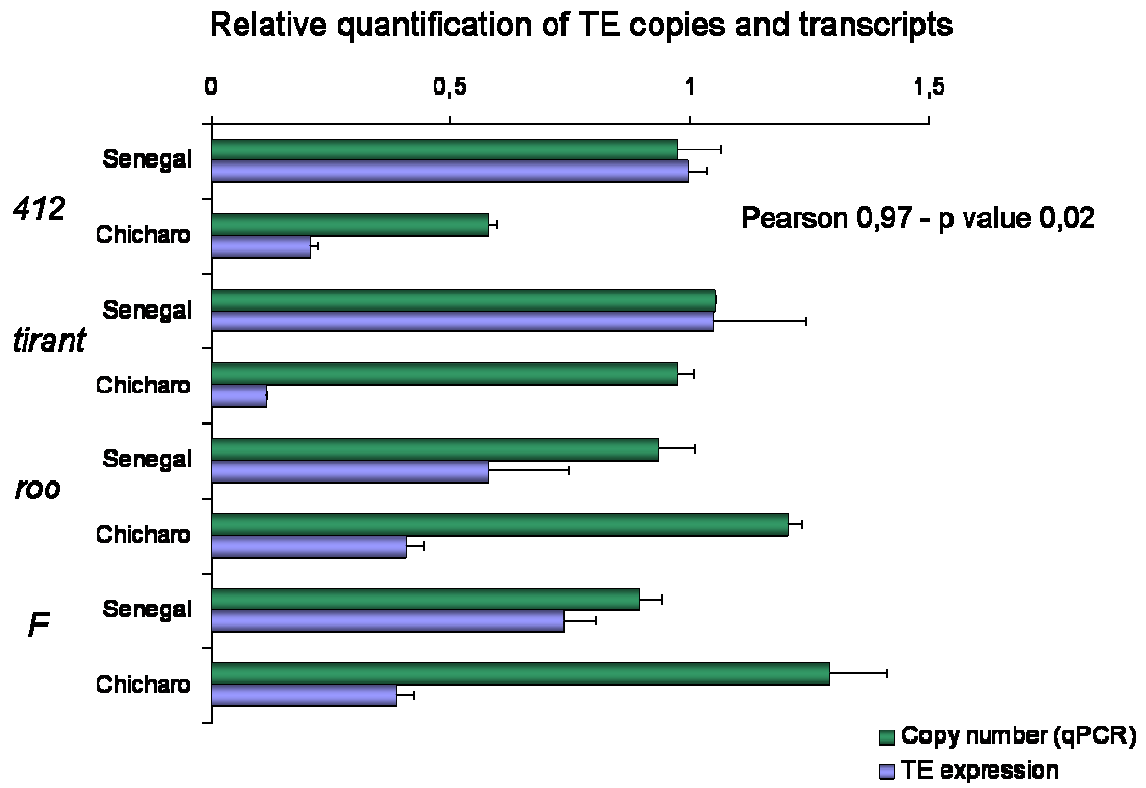


Figure 3

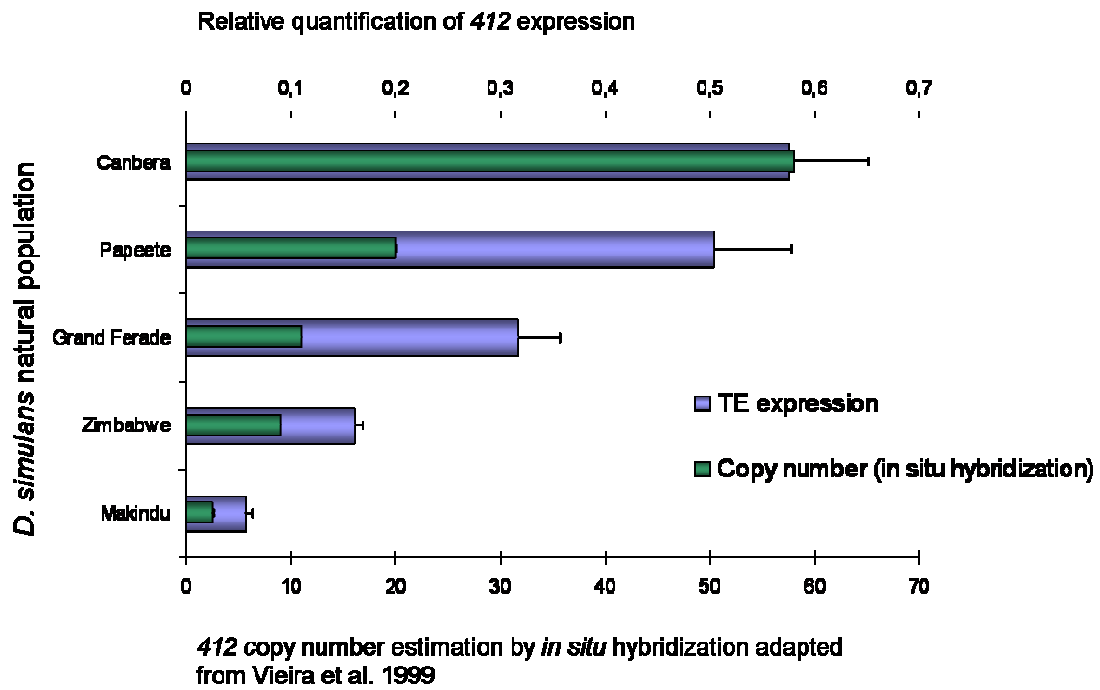


Figure 4

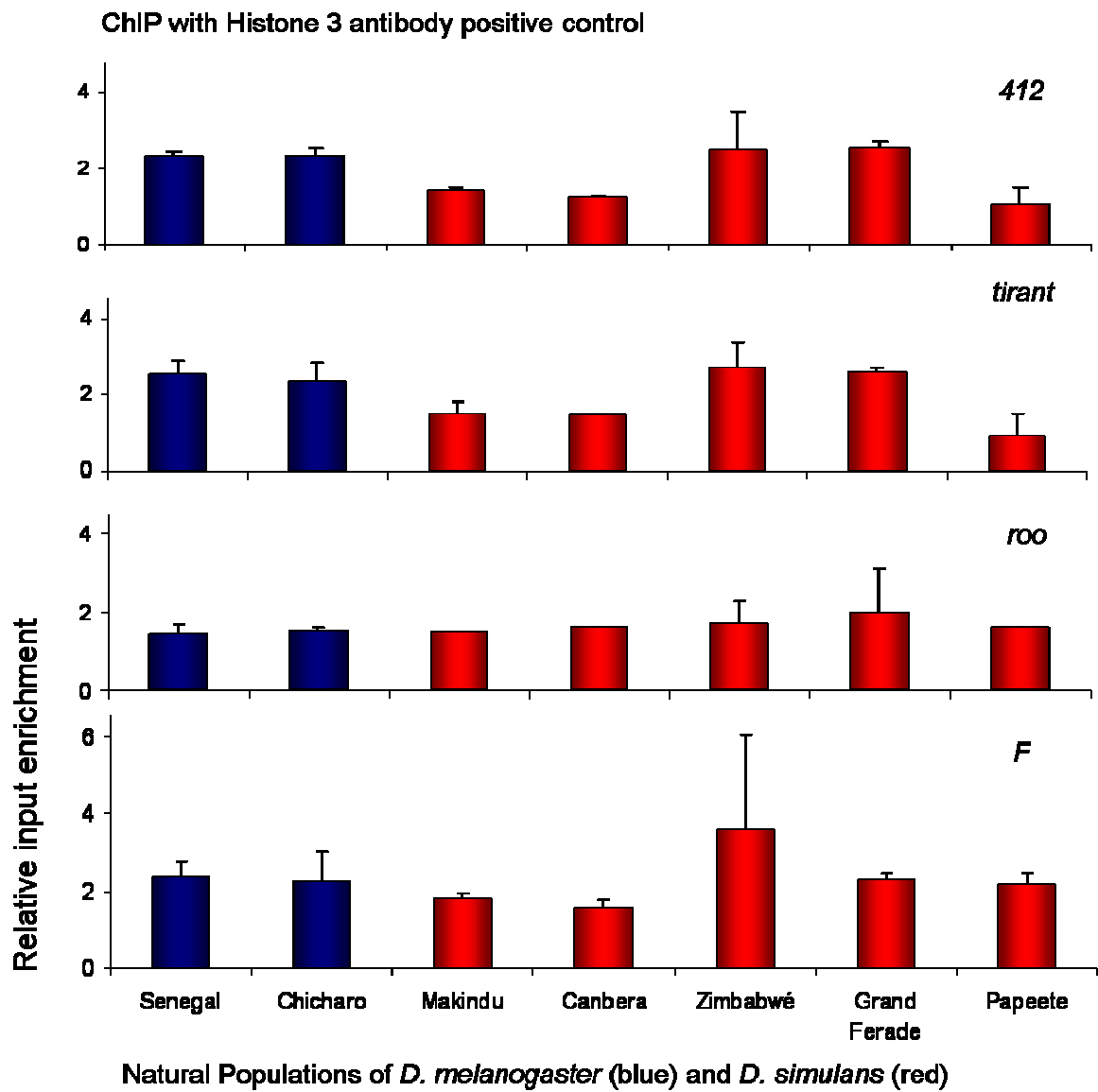


Figure 5

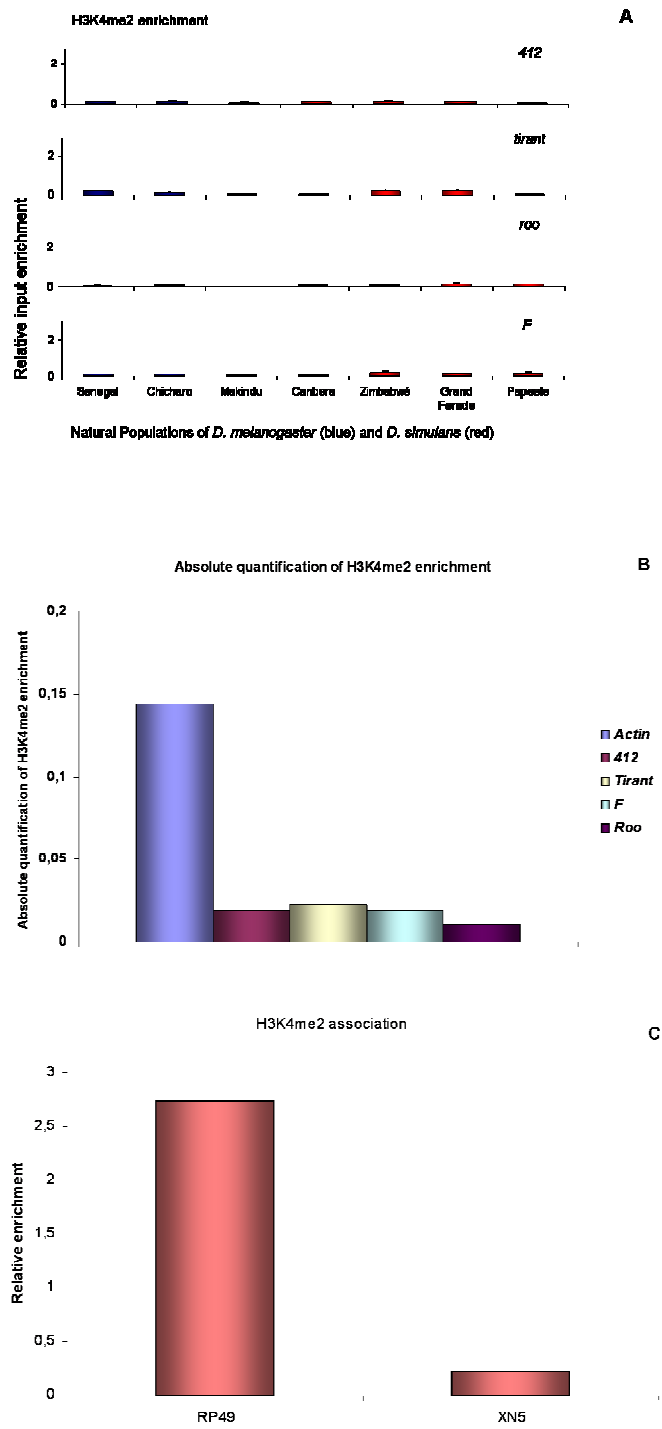


Figure 6

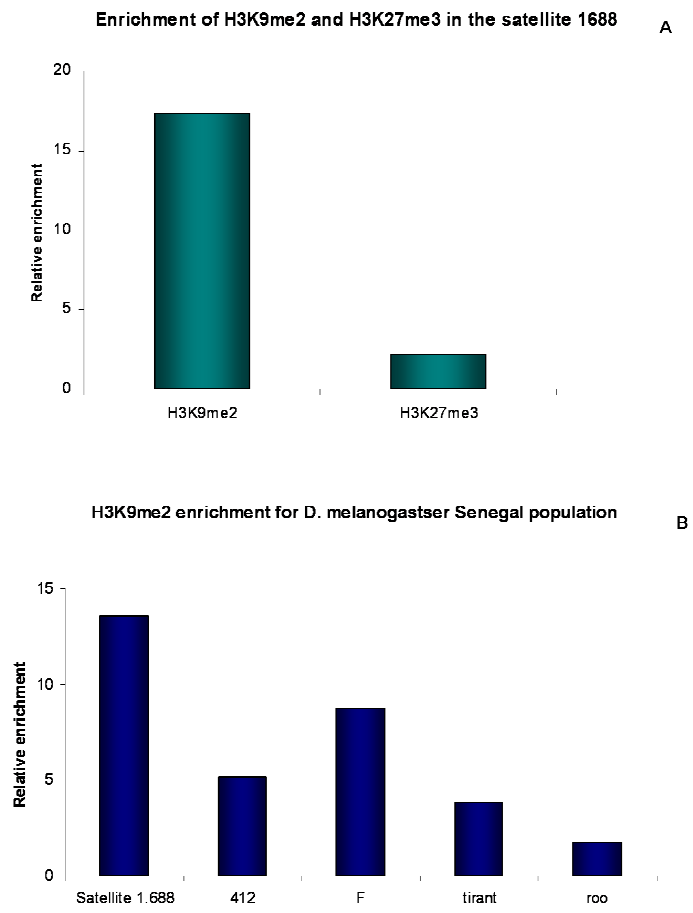
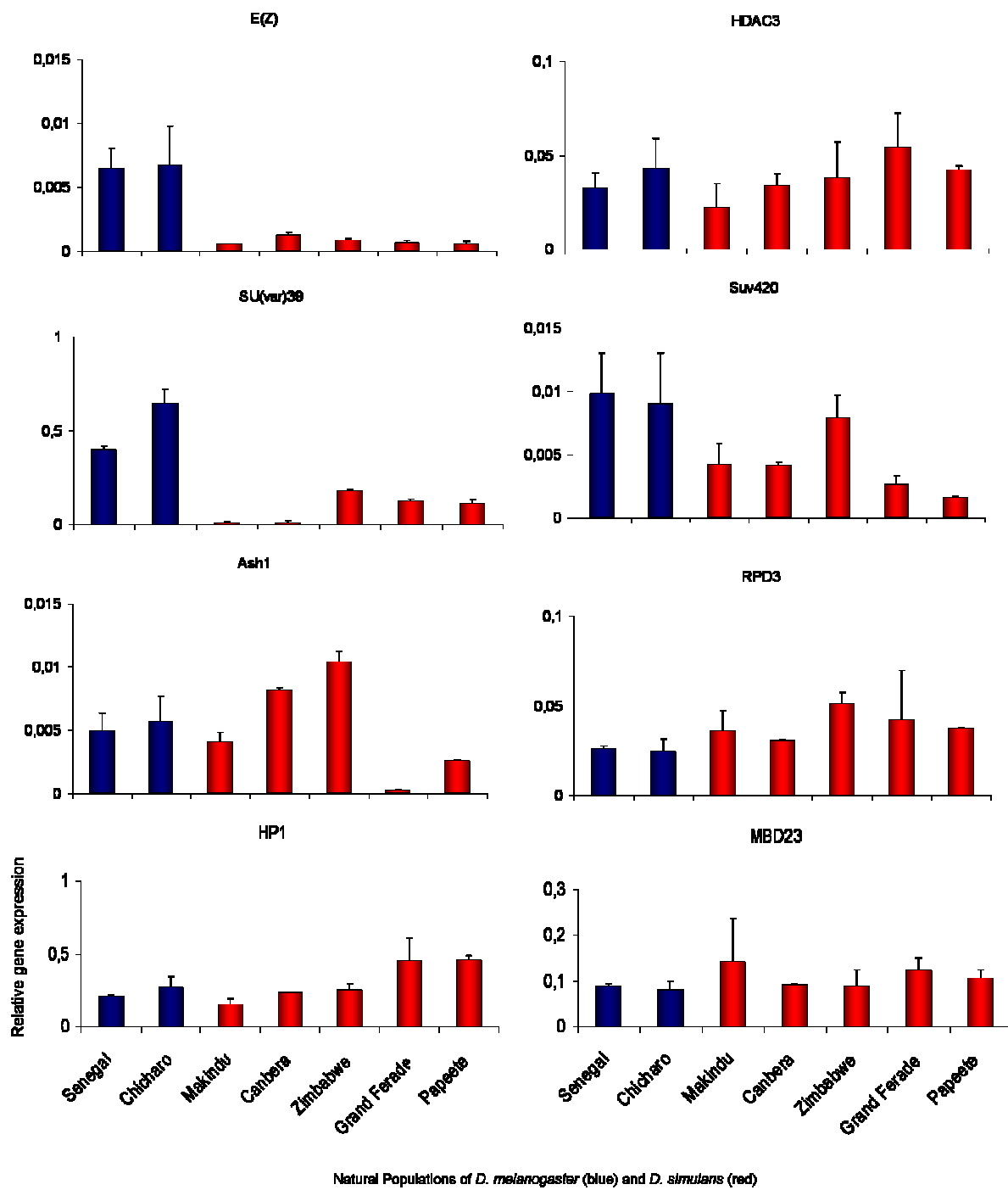


Figure 7



**Revised version for re submission to Gene**

## **Jumping genes and epigenetics : towards new species**

**Rita Rebollo<sup>1</sup>, Benjamin Hubert<sup>1</sup>, Béatrice Horard<sup>2</sup> and Cristina Vieira<sup>1\*</sup>**

1. CNRS, UMR558, Laboratoire de Biométrie et Biologie Evolutive, Université Lyon 1, Villeurbanne, France

2. Laboratoire de Biologie Moléculaire de la Cellule-UMR 5239 CNRS ;ENS LYON ; Université LYON 1 ; HCL. 165, Chemin du Grand Revoyet, 69495 Pierre Bénite or 46, allée d'Italie, 69364 Lyon cedex 07, France

**\* Corresponding author:**

**Cristina Vieira**

**[vieira@biomserv.univ-lyon1.fr](mailto:vieira@biomserv.univ-lyon1.fr)**

**Tel.: +33 4 72 43 29 18; fax: +33 4 72 43 13 88.**

## **Abstract**

Transposable elements (TE) are responsible for rapid genome remodelling through the creation of new regulatory gene networks and chromosome restructuring. TEs are often regulated by the host through epigenetic systems but environmental changes can induce physiological and, therefore, epigenetic stresses, disrupting TE tight control. The consequent TE mobilisation drives, in turn, genome restructuring that may offer a genetic innovative escape to the host. We suggest, therefore, that macroevolution and speciation might originate from an unleashing of the TEs from epigenetic host control. To understand the impact of TEs and their importance in host genome evolution, it is essential to study TE epigenetic variation in natural populations. We propose to focus on recent data demonstrating the correlation between variations in the epigenetic control of TEs in species/populations and genome evolution.

## **Keywords:**

**Transposable elements; evolution; speciation; natural populations; epigenetic control**

## **Abbreviations:**

**TE : transposable element**

## 1. Introduction

Genome sequencing programs provided new clues in understanding the lack of correlation between phenotypic complexity and genome size – the so called “C value” paradox – by pointing out that genome size differs between species mostly in the non coding parts (reviewed in (Biemont and Vieira, 2006)). For instance, the human genome is composed of ~98% of non coding DNA (2004) while the fruit fly, *Drosophila melanogaster*, has a very compact genome with only ~12% of this type of sequence (Dowsett and Young, 1982). This variable, but seemingly useless, part of the genome was named “junk DNA” and is mostly represented by repetitive sequences, such as satellite DNAs and transposable elements (TEs), which will be the focus of this short review (Figure 1 for a eukaryotic TE classification). Our work in understanding TE dynamics in natural populations led us to question the influence of TE mobilisation in speciation (Vieira et al., 1999; Rebollo et al., 2008; Fablet et al., 2009). Speciation may be a slow process involving the fixation of genetic differences between individuals, either by changing their fitness and/or ecological specialization and/or simply inducing genic incompatibility, all followed by micro population isolation. Speciation is also thought to happen fast in the case of non-genic speciation, i.e., when important karyotypic differences are observed between individuals of the same species causing sexual isolation. We are interested in rapid evolution induced by chromosomal incompatibilities inside one species. Chromosomal incompatibility, i.e., the impossibility of pairing for both parental homologous chromosomes, can happen with inter-specific crosses but also between individuals of the same species when chromosome rearrangements have occurred. Such incompatibility can lead to progeny lethality or sterility causing isolation of fertile survivors into micro populations. TEs are inducers of genome remodelling either through transposition, as described for lines of

maize where Ac/Ds alternative transposition (from the ends of two elements) is directly responsible for major chromosomal rearrangements (translocations, duplications, inversions ...) (Zhang et al., 2009), or for all recombination events due to TE copies (Hedges and Deininger, 2007). A burst of transposition might, therefore, be essential for rapid karyotypic isolation and, hence, macroevolution. TEs are often kept silenced in genomes and in order for such karyotypic changes to happen it is necessary for TEs to escape the regulation systems, which are mainly composed of epigenetic mechanisms. We propose that macroevolution and speciation could be concomitant with rapid genome remodelling induced by TE awakening from tight epigenetic repression, as indicated by many observations.

## **2. Junk and Genome: a partnership in progression**

The developmental and evolutionary importance of TEs is no longer questionable. In general, all families of DNA repeats could potentially have an impact on genome organisation, either as generators of genome instability, since multicopy elements are powerful recombinogenic substrates (Hedges and Deininger, 2007), or as components of essential chromosomal domains, like centromeres and telomeres in many species (Wong and Choo, 2004; Lamb et al., 2007). It is clear that TE replication might induce genetic mutations via transposition. For instance, in *Drosophila* a high rate of spontaneous mutations are related to transposition events (Biemont and Vieira, 2006). The immediate effects of transposition might be harmful, as illustrated by the occurrence of several diseases described in humans (reviewed in (Callinan and Batzer, 2006)). Despite the damaging effects, TEs are maintained in almost all genomes either as full length or truncated copies. Full length copies have kept their ability to mutate the genome through transposition while the truncated copies have often lost the capacity to jump. However, the truncated versions might putatively be recruited by

the genome. Indeed, recent reports have proposed a co-evolution of TEs and hosts, leading to the participation of TEs (often truncated copies) in complex genomic processes (as co-opted open reading frames, post transcriptional gene regulation, gene protein translation enhancement etc. as reviewed in (Muotri et al., 2007; Sinzelle et al., 2009)). Moreover, truncated copies may also serve as recombinogenic substrates to other truncated or full-length copies, inducing genome rearrangements. TE copies have been the source of new regulatory sequences, alternative splice sites, polyadenylation signals (Marino-Ramirez et al., 2005), and new transcription factor binding sites (Polavarapu et al., 2008). Also, TEs enhance genome regulation as, for example, when they are the origin of microRNAs, which are able to regulate gene expression (Hasler et al., 2007; Piriyaopongsa et al., 2007). Since TEs are widespread in the genome and have so many influences on gene regulation, several authors have suggested a vital need for TEs in creating, remodelling and regulating gene networks (McClintock, 1984; Feschotte, 2008). Therefore, as we hypothesize, a strong impact on genome structure and regulation could be observed after TE bursts of transposition.

### **3. Driving speciation through TE bursts of transposition**

TEs have been observed in all sequenced genomes analysed to date and comparative genomics allows a broad insight into the variability in TE content between genomes of different species. For instance, TEs represent 60% of the maize genome (Messing et al., 2004) and only 14% of *Arabidopsis thaliana*'s (2000). Variations in genome size between closely related species can also be correlated to differences in the amount of TEs. For instance, in the *D. melanogaster* species subgroup, bigger genome sizes are in part attributed to high TE-like sequences, as estimated from the amount of reverse transcriptase related sequences by dotblot (Boulesteix et al., 2006). In cotton, a significant variation in the total amount of TEs is

observed between *gossypium* species (40% to 65%) (Hawkins et al., 2006). Such disparities depend on particular TE subfamilies, such as a massive amplification of the *gypsy*-like *Gorge3* element in increased genome size cotton species as shown by whole genome shotgun sequencing (Hawkins et al., 2006). Variations in the relative proportion of TEs can also be intra-specific, as noted for the euchromatic copies of several TE families when counting on polytene chromosomes in natural populations of *D. simulans* (Vieira et al., 1999). Variation in the proportion of TEs illustrates not only the influence of a given species/population on the host/TE relationship but also suggests that TEs might play a significant role in speciation through their individual specific influence on the host genome.

TE abundance, TE-derived genomic features and chromosomal rearrangements involving TE sequences are frequently lineage specific and, therefore, suggest that variations might have either stabilized after speciation or contributed to the species evolutionary process (Marino-Ramirez et al., 2005; Bohne et al., 2008). One should note that correlating TE transposition consequences and speciation is rather tricky, since an exact timing of TE bursts and natural species diversification is difficult to measure. Several researchers have tried to obtain examples of timing concordance between bursts of transposition or massive TE extinction with speciation (Table 1). Significant TE activity is observed in several species often during a period of radiation, suggesting that massive speciation and massive TE activity may be associated. The genetic distance between two organisms is calculated as a function of their genetic divergence, so every episode creating divergence, such as lineage specific transposition events, might contribute to the reproductive isolation of those organisms. Either as a cause or as a consequence of genetic differentiation, TE patterns that are different between individuals of the same species may serve not only as genetic markers for researchers but also as evidence of the speciation process occurring within the species (Esnault et al., 2008). For instance, significant TE insertion site polymorphism can be observed in rice for the

japonica and indica cultivars, accounting for 14% of their genetic differences (Huang et al., 2008). Since the exact evolutionary history of a species is difficult to determine, it is useful to study interspecies crosses as a model for macroevolution. Indeed, interspecies hybrids are classical examples of bursts of transposition causing important dysfunctions and potentially acting as inducers of rapid speciation (reviewed in (Michalak, 2009)) (Table 2). Three hybrids of sunflower species have a 50% larger genome than the parental lines because of a massive TE transposition and they are thought to have undergone rapid speciation (in fewer than 60 generations for one of the hybrids) (Ungerer et al., 1998; Ungerer et al., 2006). Also, in dosage dependent crosses between *A. thaliana* and *A. arenosa* (crosses where the amount of maternal and paternal genome is variable and may be different), high expression of the paternal *A. arenosa athila* element in the hybrid is correlated to seed lethality (Josefsson et al., 2006). Such observations are essential for understanding hybrid “compatibility” since *A. arenosa* and *A. thaliana* hybridization has succeeded at least once in nature (Jakobsson et al., 2006). In insects, *Drosophila buzzatii* and *D. koepferae* are still able to interbreed and have common TE families that are maintained in each genome. Crosses between these two species highly induce the transposition of *osvaldo* in hybrids whilst it is repressed in both parental genomes (Labrador et al., 1999). In wallabies, interspecific hybrids present variable centromeres composed of satellite repeats and newly replicated TE copies (Metcalf et al., 2007). All these examples suggest that TE bursts of transposition occurring during hybrid speciation may induce important karyotypic changes because of the capacity of TEs to induce major chromosomal rearrangements and ectopic recombination (Hedges and Deininger, 2007; Weil, 2009). In this way, TE mobilization will thus support novel phenotypes followed by micro population ecological isolation, the necessary components for rapid and divergent evolution. Based on hybrid crosses, we hypothesise that bursts of transposition in individuals of one species followed by intraspecific crosses could generate new phenotypes.

Persistence of TEs in most genomes, despite the initial deleterious effects, suggests a strict regulation system keeping TEs silenced. Expression of TEs is dependent on transcription factors presence as illustrated by the evolution of L1 lineage in humans. Indeed, recruitment of regulatory regions in new L1 subfamilies harbouring new transcription factor binding sites are essential for L1 expression (Khan et al., 2006). Also, cellular inhibitors may influence TE transposition post transcriptionally, as observed for some members of the APOBEC family, capable of reducing HERV-K infectivity (50 fold) (Lee and Bieniasz, 2007) and block *alu* transposition in an ORF1p and ORF2p L1 independent manner (Hulme et al., 2007). Moreover, transposition of mobile elements induces DNA breaks suggesting an interaction between the host DNA repair machinery and TEs. ERCC1-XPF heterodimers are implicated in DNA repair processes and limit L1 insertion (Gasior et al., 2008). Apart from cellular inhibitors and transcription factor dependency, TEs are also transcriptionally and post-transcriptionally regulated through epigenetic means (Lisch, 2009). Although regulation of TEs by the host genome might be tight it also needs to remain versatile in order to keep a “rapid genome remodelling source”. Such flexibility is insured in many eukaryotic organisms via epigenetic regulatory mechanisms.

#### **4. TE Epigenetic reprogramming**

Epigenetic regulation of TEs involves interdependent pathways, such as chromatin remodelling factors, DNA methylation and non-coding small RNAs (Lisch, 2009; Obbard et al., 2009) (Table 3 for a general view on TE epigenetic regulation). In rice, for instance, mutants of histone methyltransferase, specific for repressive chromatin marks, induce DNA demethylation of *Tos17* (*copia*-like retrotransposon) and, consequently, transposition (Ding et al., 2007). Also in plants, it is possible to observe RNA dependent DNA methylation

(RDDM) of TEs and genes, which is reversible since it is dependent on the presence of small interfering RNAs (Matzke et al., 2007). Recent investigations have highlighted the central role of RNA in controlling TE activity: such system was probably present in a common eukaryote ancestor; it is well conserved between species; and it may act as an immunological system against non self RNAs (Obbard et al., 2009). Also, small RNAs allow for target specificity of DNA methylation or histone modification in a given sequence. For instance, epigenetic instability in long-term cultured cells of *A. thaliana* evolves into hypomethylation of specific TEs and subsequent activation (Tanurdzic et al., 2008). Indeed, *athila* or *copia* elements are hypomethylated regardless of their location, but no change is observed for *gypsy* class elements (Tanurdzic et al., 2008). Such specificity is possibly due to siRNAs being produced differently between TE families under stress such as this, varying from 21nt and 24nt for hypomethylated activated TEs and only 24nt for silenced *gypsy* class elements (Tanurdzic et al., 2008). The epigenetic regulation system is, indeed, rather efficient; it is general because the TE families capable of invasion are multiple and divergent but, at the same time, it also appears to be specific, targeting single TE families through sequence specific small RNAs. Each pathway of the epigenetic regulation of TEs seems, therefore, to be essential and also extremely rigorous. Naturally the question arises as to how TEs could possibly invade a genome if they are being held prisoners in a perfect prison. It is well known that TEs do transpose often at a very low rate suggesting that the perfect prison is, after all, flexible. Indeed, it has recently been suggested that small RNAs can be linked with total or partial silence of elements, as observed in *Drosophila* hybrid dysgenesis. Indeed, intraspecific *Drosophila* crosses may cause hybrid dysgenesis of P and I elements showing very important deleterious effects, such as female sterility or chromosomal abnormalities (Bucheton et al., 1984; Castro and Carareto, 2004). In these crosses individuals of the same species have different amounts of TEs since one of the parents has an empty genome. A deficit in small

interfering RNA (piRNA) in the maternal gamete allows originally silenced TEs to transpose in the hybrids (Brennecke et al., 2008; Chambeyron et al., 2008). However, despite the associated female sterility or embryonic lethality, it should not be forgotten that a few hybrids do survive and are fertile, thus propagating newly mobilised copies and possible chromosomal changes to the population.

The study of natural populations and the observation of the natural variability that exists in epigenetic host control can lead to an understanding of TE induced macroevolution. Epigenetic variation in hybrids, in allopolyploid species, and in single individuals could awake TEs, induce a burst of transposition and, as described above, increase karyotypic changes followed by ecological isolation. TE epigenetic regulation can be observed in somatic tissues (Barbot et al., 2002; Malone et al., 2009) and in germline tissues (Malone et al., 2009). A model has been proposed by Slotkin and co-researchers (Slotkin et al., 2009) in *A. thaliana* where the naturally occurring hypomethylation of TEs in the vegetative nucleus (somatic) influences TE activity in the neighbouring sperm cell. In *Drosophila*, the vertical propagation of some retroviral elements (proviral amplification) depends on the inhibition, via small RNA interference, of full length retroviruses in the follicular cells adjacent to the oocyte (Brennecke et al., 2007; Lau et al., 2009; Malone et al., 2009). Both types of regulation (somatic and germinal), if variable, can influence the population behaviour by creating phenotypical variations that might be inherited. Variation in TE epigenetic regulation can be observed such as in *A. thaliana* where the LINE-like element *sadhu* varies between three ecotypes showing different degrees of DNA methylation and different silencing states (Rangwala et al., 2006). Other epialleles, variation in the epigenetic regulation of a given sequence between tissues and/or individuals of the same population, were most described in plants and mice. However, improvement in population epigenetics is still necessary along with ecological epigenetic studies in order to fully understand natural population variation in

epigenetic regulation (Bossdorf et al., 2008; Johannes et al., 2008; Richards, 2008). TE epigenetic regulation is, therefore, a variable and flexible mechanism that may induce massive TE transposition in the germline and consequent chromosomal rearrangements.

The gibbon species has rapidly accumulated chromosomal rearrangements and, hence, offers an interesting model for karyotypic evolution and speciation. Carbone and collaborators recently described an example of the different epigenetic regulation of *Alu* elements between humans and gibbons associated with breakpoints between both species (Carbone et al., 2009). They observed a CpG content higher in the gibbon *alu* elements near the breakpoints (typical from active elements) and such elements are undermethylated compared with humans. *Alu* elements present in the breakpoints are probably active and responsible, in part, for the rapid chromosomal remodelling in the gibbon. The authors propose that “the association between undermethylation and chromosomal rearrangement in gibbons suggests a correlation between epigenetic state and structural genome variation in evolution”.

Conjugating two different genomes in the same organism, as is the case for hybrids or in allopolyploidization, may require significant adaptations of all the regulatory mechanisms, including TE epigenetic regulation (reviewed in (Michalak, 2009)) (Table 2). In wallabies, interspecies crosses cause a burst of transposition of a retrotransposon, together with a genome-wide hypomethylation (O'Neill et al., 1998). Such a burst of transposition targeted only one parental genome and results in extended centromeres, suggesting a rapid karyotype differentiation from the parents (O'Neill et al., 1998). Other natural crosses were analysed by the authors and hypomethylation of the hybrids was always observed as *de novo* chromosomal changes. In allopolyploidization (“condition whereby evolution is *accelerated* and fitness is enhanced” as suggested by Liu and Wendel), TE transposition may also be concomitant with genome-wide epigenetic changes (Liu and Wendel, 2003). These examples show how genome remodelling could occur after epigenetic variation in TE copies. However,

we need to establish what are the causes of genome wide epigenetic modifications and subsequent TE activation. Interspecific crosses induce genomic stress, *i.e.* changes in genomic stability (chromatin changes, density in repeats...) and organization (DNA recombination, TE replication, retroposed or duplicated genes...), that could indeed have an impact on epialleles and provoke TE activation. Genome-wide epigenetic changes might play a role in genome adaptation to environmental changes. One could easily hypothesize that TE awakening is due to epigenetic changes, and that those variations start at the level of one individual as a response to specific environmental changes. The ecological outcomes of TE mobilisation due to environmental changes may be numerous, such as survival of the host, increase of host fitness, micro population isolation etc. The consequent spread of these factors into populations can lead to sexual isolation and speciation.

## **5. Environment induces epigenetic reprogramming**

Several investigations have demonstrated that modifications in the environment induce epigenetic modifications and, therefore, transcription state changes (Jaenisch and Bird, 2003; Han and Boeke, 2005). Such transcriptional changes are a source of phenotypic variability that may be explored by organisms through increasing the host “adaptative potential”. Indeed, diet changes, temperature variation, stress etc. all have an impact on gene regulation (Waterland and Jirtle, 2004; Cropley et al., 2006; Gibert et al., 2007; Chinnusamy and Zhu, 2009). Similarly to epigenetic regulation, diet changes, temperature variations, stress etc. might affect TE transposition (El-Sawy et al., 2005; Hashida et al., 2006; Ebina and Levin, 2007; Cho et al., 2008). Consequently, activation of TEs could be the result of relaxed epigenetic control induced by environmental changes. There is a huge amount of literature

relating to the activation of TEs by environmental stresses, but only a few examples illustrate the link existing between environmental epigenetic instability and transposition.

Early nutrition has an impact on the epigenetic regulation of TEs, especially DNA methylation, as reviewed by (Waterland and Jirtle, 2004). The *aguti* gene controls hair color in mice (brown in wild type). An insertion of an IAP retrotransposon in the first exon induces ectopic and variable expression of *aguti*. LTR from IAP elements is regulated by DNA methylation but varies between individuals. Dietary supplementation (methyl donors) shifts the phenotype to wild type brown colour, concomitant to a higher DNA methylation in the IAP element (Waterland and Jirtle, 2003). Heat treatment and aging induce the transcription of older heterochromatic I copies and, hence, the production of small interfering RNAs (siRNA) repressing active I elements in the germline (Dramard et al., 2007). Individuals exposed to the pollutant benzene show decreased DNA methylation of L1 and *alul* elements (Bollati et al., 2007) and, similarly, benzo(*a*)pyrene increases retrotransposition of L1 elements in HeLa cells (Stribinskis and Ramos, 2006). In rice, spaceflight induces transposition of several TEs, sometimes associated with hypomethylation within the element (Long et al., 2009). In mice, a long-term peroxisome proliferator diet induces hypomethylation of satellites, IAP and L1/L2 elements (Pogribny et al., 2007).

## **6. Conclusion**

Epigenetic regulation of TE copies has two main consequences: 1) the environment can have a direct influence on TE activity through epigenetic instability; 2) the presence of TE sequences in the host genome in a “harmless” state. Since bursts of transposition have been observed in several species it is tempting to suggest that the defense system has, at least temporarily, broken down. However, this failure is transient and the host may rapidly silence

“de novo” TE copies. Although the benefit is not immediate, transposition might have a long term advantage. Indeed, the consequences of transposition bursts will be numerous, resulting in a renewal of genetic diversity, which is the major condition for genome evolution and the action of selection. This genetic diversity is, thus, fundamental for renewing gene networks and inducing the emergence of new species. Each environmental change indirectly creates an increase in host genetic variability and selection can finally act within a bigger repertoire of genetic information. Epigenetic instability of TEs would cause significant genetic variability, thus leading to selection of the best adapted organism.

Figures

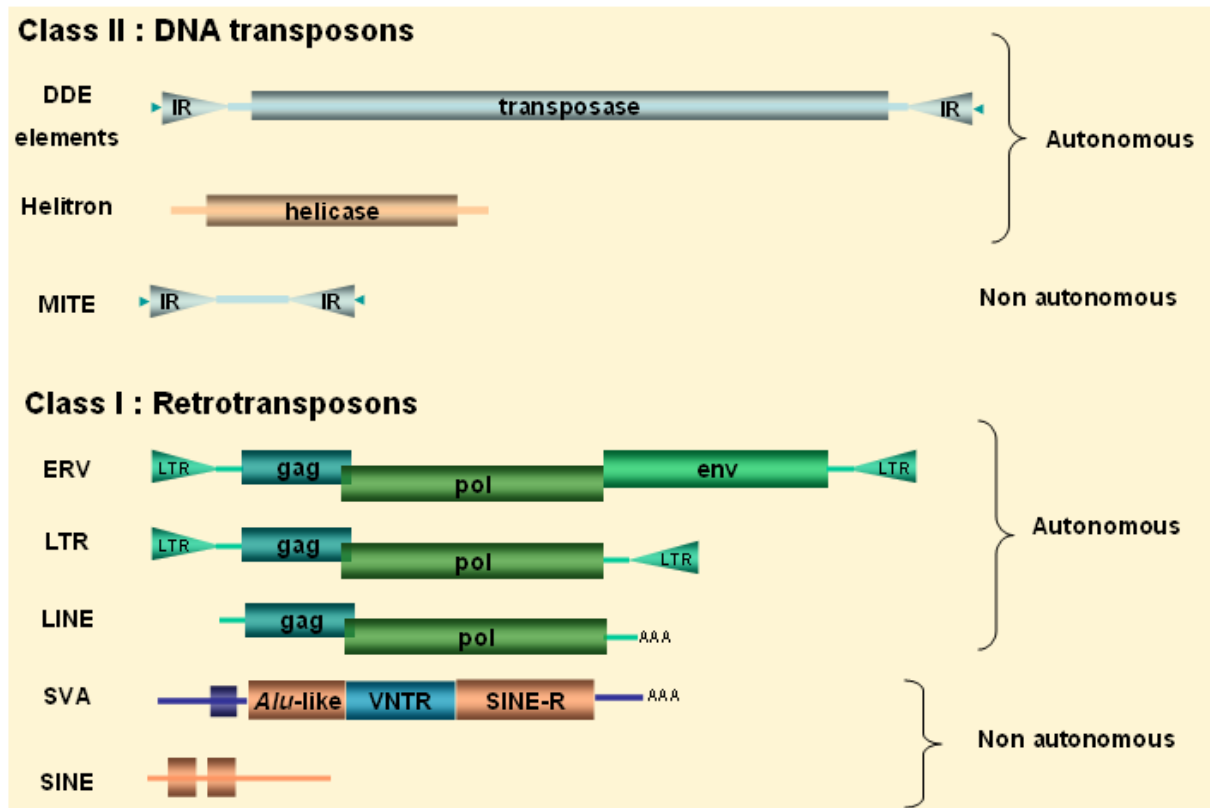


Figure 1

**Eucaryote transposable element general classification**

TEs are abundant and ubiquitous mobile sequences capable of jumping inside the genome. Differences in the transposition mechanisms allow a classification of TEs into two major classes: Class I Retrotransposons “copy&paste” through an RNA intermediate and Class II DNA transposons just “cut&paste” their own molecule. Autonomous retrotransposons can harbour long terminal repeats in their ends (LTR) or not (LINE-like) and might be infectious agents (endogenous retrovirus). Non autonomous retrotransposons, such as SINEs, are dependent on autonomous elements to be “copied&pasted” in trans. The same dependency is observed among DNA transposons where MITEs need full length transposase coded by autonomous DNA transposons to be “cut&pasted” in trans. Full length helitrons, newly classified II DNA transposons, play an important role in exon shuffling thanks

to their “rolling circle” replication mechanism. For a recent classification of eukaryote TEs, please refer to (Wicker et al., 2007). Boxes represent open reading frames, triangles are either inverted repeats (IR), in blue, or long terminal repeats (LTR), in green, and small blue arrows are duplicated insertion site representations. DDE elements : transposases carrying aspartate (D), aspartate (D,) glutamate (E) motif. MITE : miniature inverted repeated elements ; ERV : endogenous retrovirus ; LINE : long interspersed nuclear element ; SVA : composite element composed of parts of SINE, VNTR (variable number of tandem repeats) and *alu* repeats – first box represents CCCTCT hexamere repeats ; SINE : short interspersed nuclear element – red box represents diagnostic features ; Gag, Pol, Env : retroviral-like proteins coded by TE open reading frames.

## Tables

| TE burst of transposition concomitant with radiation periods |   |   |
|--|---|---|
| Models   | TE events /Species history  | References  |
| Eutheriens   | <ul style="list-style-type: none"> <li>• Decreased L1 and SINE accumulation during emergence of African apes (14-15 Mya).</li> <li>• Generation of L1 subfamilies in less than 0.3My concomitant with intense speciation in <i>Rattus sensu stricto</i>;</li> <li>• Timing of Lx family (L1 ancestral family) amplification is close to the murine radiation ;</li> <li>• Rapid speciation in the genus <i>Taterillus</i> (gerbil) has occurred and massive transposition of TEs in new lineages was observed.</li> <li>• DNA elements were extremely active during <i>Myotis</i> radiation.</li> </ul> | (Pascale et al., 1990; Verneau et al., 1998; Lander et al., 2001; Dobigny et al., 2004; Ray et al., 2008) |
| Fish   | <ul style="list-style-type: none"> <li>• DNA transposon bursts of transposition are concomitant with speciation events in pseudotetraploid salmonids and after genome duplication ;</li> </ul>  | (Volf et al., 2001; de Boer et al., 2007)   |
| Unicellular eukaryotes                                       | <ul style="list-style-type: none"> <li>• Acquisition and consequent transposition of an endogenous retrovirus element in <i>Entamoeba histolytica</i> and lineage specific enrichment in TEs might affect speciation and pathogenicity .</li> </ul>   | (Lorenzi et al., 2008)  |

Table 1 : TE burst of transposition concomitant with radiation periods

| Hybrid analysis : epigenetic remodelling and TE activation |  |   |
|--|--|---|
| Models   | Experiment conclusions   | Reference   |
| Eutheriens   | <ul style="list-style-type: none"> <li>• <i>Mus musculus</i> &amp; <i>M. caroli</i> crosses induce retroelements hypomethylation on chromosome 10, the substrate of double minute chromosome formation in interspecies hybrids.</li> </ul>   | (Brown et al., 2008)  |
| Insects  | <ul style="list-style-type: none"> <li>• <i>D. melanogaster</i> intraspecific crosses can result in hybrid dysgenesis, associated with P or I elements mobilisation, dependent on rasiRNA production in the germinal cell line and causing several abnormalities (such as female sterility);</li> <li>• Crosses between <i>D. buzzatti</i> &amp; <i>D. koepferae</i> awake <i>osvaldo</i> copies in the hybrid.</li> </ul>   | (Labrador et al., 1999; Brennecke et al., 2008)   |
| Marsupials   | <ul style="list-style-type: none"> <li>• Interspecific macropodid hybrids (<i>Macropus rufogriseus</i> &amp; <i>M. agilis</i>) present centromeric instability due to TEs and satellite replication, probably inducing karyotypic isolation relative to the parental species;</li> <li>• Genome-wide hypomethylation and centromeric expansion, due to TE activation, are observed in <i>M. eugenii</i> or <i>Wallabia bicolor</i> hybrids.</li> </ul>   | (O'Neill et al., 1998; Metcalfe et al., 2007)   |
| Plants   | <ul style="list-style-type: none"> <li>• In <i>A. thaliana</i> &amp; <i>A. arenosa</i> dosage-dependent crosses, the usually silenced paternal <i>athila</i> elements are activated concomitantly with the deregulation of polycomb complex-dependent gene regulation;</li> <li>• Wheat allotetraploid formation is accompanied by TE activation, DNA methylation and gene expression alterations;</li> <li>• <i>Helianthus annuus</i> &amp; <i>H. petiolaris</i> hybrids have a 50% larger genome than parental individuals due to TE amplification ;</li> <li>• DNA introgression in <i>Zizania latifolia</i> causes TE activation through modifications in DNA methylation and morphological novelties compared to the primordial line. Note that <i>de novo</i> stable silencing of TEs is observed in the introgression lines.</li> </ul> | (Ungerer et al., 1998; Kashkush et al., 2002; Liu and Wendel, 2003; Josefsson et al., 2006; Ungerer et al., 2006) |

Table 2 : Hybrid analysis : epigenetic remodelling and TE activation

| General view of TE epigenetic regulation |  |
|--|--|
| Histone modifications                    | <p>Position effect variegation (PEV) is the mechanism describing transcription variation of a given gene correlated to its chromatin localization. Mutations in Su(var) genes responsible for such variegation are often accompanied by TE amplification. The major function of this gene family is to post-translationally modify histone N terminal ends. Usually histone methylation in lysine residues (H3K9m, H3K27m, H4K20m) is typically from a closed chromatin conformation, in contrast to acetylation of histones and methylation in H3K4 which are often observed in open chromatin structures. TEs are highly associated with repressive marks as in general H3K9me3 in humans, H4K20me3 in drosophila and H3K9me2 in plants. Regarding chromatin malleability, histone variants and non histone proteins (as HP1) are also involved in TE regulation.</p>  |
| DNA methylation                          | <p>In plants and mammals, DNA methylation plays an important role in silencing TEs. In insects, DNA methylation is observed as a silencing process in genes and TEs.</p>   |
| Non coding RNAs                          | <p>Post translational gene silencing (PTGS) through small interfering RNAs (siRNA) processed by the AGO/DICER/RISC complex is also a mechanism that can be used to silence TEs. Indeed siRNAs derived from TE copy transcripts can target full length and putatively active TE transcripts, thus preventing TE transposition. Piwi related RNAs (piRNAs or rasiRNAs for repeated associated small interfering RNAs) in Drosophila are processed by the piwi/Aub/AGO3 pathway, are 24-30 nt and are known to silence TEs in the germline whereas endo-siRNA (endogenous small interfering RNAs) processed by DICER2/AGO2 are 21 nt and are capable of somatic silencing TEs. Germinal and somatic silencing are therefore possible thanks to non coding RNAs. However, the presence and the transcription of a TE copy in the genome is essential to engage PTES (post translational transposable element silencing). The idea of the immune system is, hence, appropriate since having non coding RNAs of a given TE family will preserve the genome from further invasions.</p> |

Table 3 : General view of TE epigenetic variation

## Acknowledgments

Owing to space limitations, we sincerely regret the omission of many outstanding publications in this field. We thank Dr. Marie Fablet and Dr. Daphne Reiss for their valuable advice. We thank Valerie James for the English corrections. This work was supported by ANR Genemobile, FINOVI and CIBLE2008 (Région Rhône Alpes).

## GENERAL CONCLUSIONS

TEs harbor an intimate relationship with the genomes in which they are present. As a consequence of the immediate deleterious results of some transpositions, several mechanisms have been selected that allow genome to counteract negative effects. Elimination of TE sequences as observed with *helena* is also observed with LTR elements, SINEs, and almost all elements in *D. simulans*, *D. sechelia* and *D. yakuba* ((Rebollo et al., 2008; Granzotto et al., 2009), E. Lerat personal communication). Interestingly, the *D. simulans* twin sister, *D. melanogaster*, does not harbour the same deletion mechanism, showing again how specific the arm race for invasion/protection might be. TE expression can also be suppressed without elimination of their sequences. Indeed, through epigenetic regulation, silencing of TEs is possible while they are still maintained in the genome. Chromatin data obtained in this work shows the existence of epigenetic variability in TE families between natural populations and species of *Drosophila*. Our data do not show a natural variation in the epigenetic regulation of TEs since no correlation can be observed between TE expression and local chromatin structure. It is important to note that further data is needed on TE copy specific analysis in order to correlate histone marks and TE dynamics. Such kind of project needs not only deep sequencing programs in order to locate each copy and to make ChIP copy specific, but also bioinformatic analysis.

Apart from the chromatin remodelling factors analysed in this work, DNA methylation and small interfering RNAs are also part of the epigenetic TE regulation. It has been proposed that the original role of epigenetic defence, especially small RNAs, is to protect the genome against foreign DNA, as TEs, and act as an extrapolated immunological system (Obbard et al., 2009). Therefore, TEs and epigenetic regulation factors have coexisted and coevolved for a long time. One could wonder if epigenetic regulation system was not adapted for genes while specific for parasite DNAs. The ability of TEs to insert near genes and to be genome widespread would perhaps generalise the flexible epigenetic mechanisms to its neighbourhood, as already observed for heterochromatin spreading. Genes would thus be epigenetically regulated thanks to TEs. Chromatin territories would be created, and specific regulatory elements would be targeted to each specific region. Also, as suggested by our data, if epigenetic regulation is actually element-specific, invasion of genomes by TEs become possible if they are unknown for the system. However, de novo silencing of TEs through epigenetic learning would avoid TE activity to be maintained. Hence, TEs that are new for a

genome may have the opportunity to engage invasion until the moment epigenetic mechanisms start to act and then suppress TE activity. In that case, epigenetic flexibility allows a renouveau of genetic material to the first contact with TEs. If any of these hypotheses is true, it means that efficiency of the epigenetic system to control the first invasion is essential. If no “epigenetic immunological system” is engaged then the genome will be invaded and the TEs will cause important chromosomal changes, and sometimes even macroevolution.

## COMPLEMENTARY WORK

As a member of the « Transposable element, evolution and populations » team, I had the opportunity to participate to other members' projects. I will succinctly discuss them below, outlining when I conceived the experiment.

### **The *tirant* element regulation**

*Tirant*, as described above, is an LTR retrotransposon present in both species of *D. melanogaster* and *D. simulans* (Fablet et al., 2007). Copy number polymorphism of this element is observed in *D. simulans* in which most natural populations are empty (polytene chromosomal arms) of *tirant* copies (Vieira et al., 1999). However, through Southern blot, Fablet et al. (2000) have showed that these strains of *D. simulans* presented *tirant* copies, henceforth located in heterochromatic regions, that are not polytenized (Fablet et al., 2009). In addition, characterization of *tirant* copies in both species revealed two different subfamilies: a common type present in both species (C-type) and a *D. simulans* type, specific to this species (S-type). Interestingly the S-type was specific to heterochromatic regions as showed by genome walking and Southern blot deductions (Fablet et al., 2009). The C-type on the other hand, could be observed in both heterochromatic and euchromatic regions (Fablet et al., 2006). In order to verify that both types of elements were indeed present in a specific local chromatin conformation, I ran a ChIP assay targeting euchromatic (H3K4me2) and heterochromatic (H3K9me2, H3K27me3) histone marks associated with locus-specific insertions of *tirant*. Indeed, *tirant* S-type was highly enriched in H3K9me2, suggesting its presence in dense and constitutive heterochromatin, whereas the C-type presents a weaker enrichment in H3K9me2 and H3K27me3, but a higher one in H3K4me2. The S-type is indeed present in a heterochromatic region as observed through genome walking. Since only C-type transcripts were found (Fablet et al., 2006), Fablet et al wanted to verify that the lack of expression of *tirant* S-type was indeed due to a heterochromatic silencing. In order to test this hypothesis, we transfected promoters specific to the S-type or the C-type along with a reporter gene in S2 cells of *Drosophila* (Fablet et al., 2009). Both promoters were able to induce expression of the reporter gene suggesting that it is only the heterochromatic localisation of the S-type that is responsible for its silencing. Here follows the scientific article published on these data.

## Genomic environment influences the dynamics of the *tirant* LTR retrotransposon in *Drosophila*

Marie Fablet,\* Emmanuelle Lerat,\* Rita Rebollo,\* Béatrice Horard,<sup>†</sup> Nelly Burlet,\*  
Sonia Martinez,\* Émilie Brasset,<sup>‡</sup> Eric Gilson,<sup>†</sup> Chantal Vauray,<sup>‡</sup> and Cristina Vieira\*<sup>1</sup>

\*CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, Université Lyon 1, Villeurbanne, France; <sup>†</sup>CNRS UMR 5161/ENS Lyon, Faculté de médecine Lyon Sud, Laboratoire de Biologie Moléculaire de la Cellule, Université Lyon 1/HCL, Pierre Bénite, France; and <sup>‡</sup>UMR/CNRS 6247, INSERM, U931, Faculté de Médecine, BP38, Clermont Université, Clermont-Ferrand, France

**ABSTRACT** Combining genome sequence analysis and functional analysis, we show that some full-length copies of *tirant* are present in heterochromatic regions in *Drosophila simulans* and that when tested *in vitro*, these copies have a functional promoter. However, when inserted in heterochromatic regions, *tirant* copies are inactive *in vivo*, and only transcription of euchromatic copies can be detected. Thus, our data indicate that the localization of the element is a hallmark of its activity *in vivo* and raise the question of genomic invasions by transposable elements and the importance of their genomic integration sites.—Fablet, M., Lerat, E., Rebollo, R., Horard, B., Burlet, N., Martinez, S., Brasset, E., Gilson, E., Vauray, C., Vieira, C. Genomic environment influences the dynamics of the *tirant* LTR retrotransposon in *Drosophila*. *FASEB J.* 23, 1482–1489 (2009)

**Key Words:** chromatin • endogenous retrovirus • natural populations

TRANSPOSABLE ELEMENTS (TEs), which are DNA sequences that can move and multiply along the chromosomes, are now considered to be full-fledged components of genomes, able to play various and numerous functional roles (1, 2). However, their dynamics within a given genome and natural populations is far from being fully understood (3, 4), since TE amount and genomic distribution vary considerably between different species and populations. In most organisms, a high proportion of the heterochromatic genomic compartment is composed of TEs, which are usually organized in clusters (5, 6) and mainly correspond to deleted elements. However, an increasing amount of data shows that heterochromatin also harbors complete copies of TEs, which may constitute a “reservoir” of new elements that may be reactivated (7–9) and thus are potentially able to invade gene-rich regions like the euchromatin, which may have a considerable evolutionary impact. This idea is supported by data showing that the distribution of TEs between euchromatin and heterochromatin is variable and depends on the populations or strains analyzed (10–13). However, we still do not have a refined analysis of the structure and activity of TEs located in different chromatin conformations within a given genome and between genomes. From this perspective, *tirant*, an endogenous retrovirus from *Dro-*

*sophila* belonging to the *gypsy*-like long terminal repeat (LTR) retrotransposon subclass, is an interesting model, since its genomic copy number varies considerably in different natural populations of *Drosophila simulans*, ranging from 0 euchromatic insertion in most worldwide populations, to 2 to 5 in East African populations (10, 13). A previous study of the *tirant* regulatory region in natural populations of *Drosophila melanogaster* and *D. simulans* revealed two subfamilies of *tirant* regarding the 5′ LTR-untranslated region (UTR) (14). One subfamily, called C type, corresponds to the euchromatic insertions in African populations, and is found in the heterochromatin of all populations worldwide. This C type has been shown to be expressed when located in euchromatin. The other subfamily, called S type, is found in all populations, at very low copy numbers (14), and despite its high levels of sequence conservation between populations, is not found to be actively transcribed.

In an attempt to understand the dynamics of *tirant* and the influence of genomic localization on its activity, we relate data on the *in vitro* expression of *tirant* and on genome walking and sequence analysis. We show that the chromatin environment is a hallmark for expression and that heterochromatin harbors full-length elements that have invasive features. We tested whether the different 5′ LTR-UTR variants were able to promote expression using the reporter gene technique and correlated this to the genomic location of each *tirant* insertion and local chromatin structure. We then identified the genomic context of the insertions in the *D. simulans* sequenced genome, and described the copies of *tirant*.

### MATERIALS AND METHODS

#### *Drosophila* natural populations

We worked on fly samples collected from several geographically distinct natural populations of *D. melanogaster* and *D. simulans*. A list of the populations analyzed is provided in Supplemental

<sup>1</sup> Correspondence: Université de Lyon, F-69000, Lyon, Université Lyon 1, CNRS, UMR5558, Laboratoire de Biométrie et Biologie Evolutive, F-69622, Villeurbanne, France. E-mail: [vieira@biomserv.univ-lyon1.fr](mailto:vieira@biomserv.univ-lyon1.fr)  
doi: 10.1096/fj.08-123513

Table 1. These populations were maintained in the laboratory at 17°C as isofemale lines or small mass cultures with around 50 pairs in each generation.

## Reporter gene assays

### Constructions

We used 3 variants of the *tirant* 5' LTR-UTR region previously amplified by polymerase chain reaction (PCR) from natural population samples (14) (GenBank accession numbers: AY756122, AY756118, and AY756121; see Fig. 1 for a detailed diagram). We refer to these sequences as S/C-*i,j*, where S or C is the type, *i* is the number of repeats of a 19-bp motif in the LTR, and *j* is the number of repeats of a 102-bp motif in the 5' UTR: S-1,5 (AY756122), C-2,2 (AY756121), and C-2,4 (AY756118) (Fig. 1). The 5' LTR-UTR regions were cloned separately upstream of a *lacZ* reporter gene into the *SphI-PstI* (for S-1,5) and *SphI-XbaI* (for C-2,2 and C-2,4) sites of the *pPelican* plasmid (15). Plasmids were amplified in TOP 10 competent cells from Invitrogen (Carlsbad, CA, USA) and purified with Plasmid Mini/Midi kits from Qiagen (Courtaboeuf, France) (final elution with water). The presence of the 5' LTR-UTR regions in the plasmids were confirmed by PCR and restriction profiles.

### Transfection into S2 cells

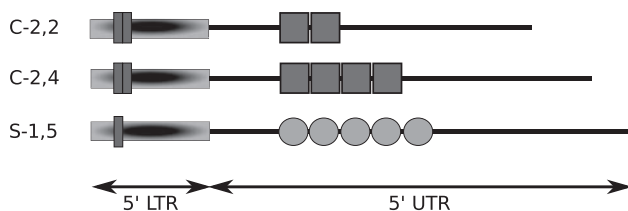
Transfections were done using Celfectin (Invitrogen). Cells were then incubated for 72 h at 22°C. S2 cells were cotransfected with *tirant* constructs and the *pGL3-Control* vector (Promega, Madison, WI, USA), consisting of a luciferase gene downstream of an SV40 promoter.  $\beta$ -Galactosidase activity was normalized *vs.* luciferase activity.

The ratios of the reported  $\beta$ -galactosidase activity to the luciferase activity were compared for the various constructions. After cell lysis, the  $\beta$ -galactosidase and luciferase activities were measured using the  $\beta$ -gal reporter gene assay, chemiluminescence kit (Roche Diagnostics, Mannheim, Germany), and the Luciferase Assay System kit (Promega), respectively.

Each transfection was repeated six times. Three wells were transfected with *pPelican* alone, as a negative control, and three others were transfected with a vector carrying a *lacZ* gene downstream of an SV40 promoter, as a positive control.

### Genome-walking analysis

We determined the insertion site sequences for each copy of *tirant* with the Universal Genome Walker kit (Clontech, Mountain View, CA, USA) in three genomes from natural populations



**Figure 1.** Structure of the variants of the 5' LTR-UTR region tested. Rectangles and ovals stand for the different tandem repeats. Variants are named as follows: S/C-*i,j*, where S or C is the subfamily, *i* is the number of 19-bp tandem repeats in the LTR, and *j* is the number of 102-bp repeats in the 5' UTR.

of *D. simulans*: the Makindu population from Kenya (5 euchromatic copies), the Chicharo population from Portugal (0 euchromatic copies), and the Zimbabwe population (2 euchromatic copies). Genomic DNA was extracted from 20 female adults from each population. For each population, 4 genomic libraries were obtained using the restriction enzymes supplied by the provider: *DraI*, *EcoRV*, *PvuII*, and *StuI*, with 9, 5, 0, and 2 restriction sites, respectively, along the *tirant* reference sequence. Adaptors were ligated according to the supplier's instructions, and two nested PCRs were subsequently done, with primers specific to *tirant* LTR and to the adaptors. The external *tirant*-specific primer can amplify both C and S types: 5' GTT TAG AGG CGT GGG GGT TTA GAA TC 3'. The internal *tirant*-specific primers specifically used for the C and S types are 5' TGT AAG CAT AAT GAA CAT GCC GAC TC 3' and 5' TGT AAA CAT AAT TTC CAT GCC ACT TC 3', respectively. PCRs were done in 2 steps using the Advantage Genomic PCR kit (Clontech).

### PCR amplification of *tirant* entire copies

To amplify each copy of *tirant*, we designed specific primers from the flanking sequences of each insertion (Supplemental Table 2). We realized 3-step PCRs with the Expand Long Template PCR System (Roche) with an annealing temperature of 60°C and the System 1 buffer, provided by the supplier.

### Populational screening of *tirant* insertions

For some of the insertion sites determined by genome walking in the Makindu population that corresponded to potentially full-length copies of *tirant*, we designed a pair of primers with the forward primer in the *env* gene, and the reverse primer in the 3' flanking region. We looked for the presence of PCR products in a pool of 3 flies for each natural population, which would indicate that the corresponding insertion of *tirant* was shared by several populations. The Makindu population was systematically used as a positive control for the PCR assays. The screened populations and the list of the primers are presented in Supplemental Table 3. We used the EuroBlueTaq enzyme from Eurobio (Les Ulis, France). The PCR was a 3-step reaction run in 30 cycles with an annealing temperature of 57°C. Amplified products were migrated on a 1% agarose gel.

### Local chromatin structure analysis

#### Chromatin immunoprecipitation (ChIP)

Extraction of chromatin from 16-h *D. simulans* Makindu embryos and immunoprecipitation (IP) were adapted from Sandmann *et al.* (16). Cell lysis buffer was changed to 5 mM PIPES pH 8, 85 mM KCl, 0.5% Nonidet P-40 supplemented with protease inhibitors. Chromatin was sheared with a Bioruptor sonicator water bath (Diagenode, Liège, Belgium) for 6 × 30-s on/30-s off cycles at high power in order to have random fragments from 1 kb to 500 bp. Shared chromatin was incubated overnight at 4°C with antibodies recognizing H3K9me2 (Millipore, Billerica, MA, USA; 07441), H3K27me3 (Millipore 07449), H3K4me2 (Millipore 07030), H3 (Abcam, Cambridge, UK; ab1791), and rabbit IgG (Sigma-Aldrich, St. Louis, MO, USA; I5006). The antigen-antibody complexes were washed as described before (16), but a second washing solution was modified as followed: TE 2×, 500 mM NaCl, 1% Triton, 0.1% SDS.

## Real-time PCR

To quantify each IP, real-time PCR was performed using QuantiTect SYBR Green PCR kit (Qiagen) on a MXP3000P PCR system (Stratagene, La Jolla, CA, USA). Reactions were done in duplicates, and standard curves were calculated on serial of input chromatin. To evaluate the relative enrichment of C and S copies after IP, we calculated the difference in cycles between the IP-enriched sample and the input DNA for *tirant* copies and for a control (*actin*-CG4027). C 2 and S 5 copies (see **Table 1**, Makindu population) were amplified using primers in the flanking region and in the *tirant* LTR (C 2 specific primers: forward, 5' GTG TTC CAG TTG CCG TCT TC 3'; reverse, 5' TTT TCT GGG GTG TTG TAC GC 3'; S 5 specific primers: forward, 5' TGC TCT CAA CTG CGC GCG AGT TAC 3'; reverse, 5' GTA AA ATA ATT TCC ATG CCA CTT C 3').

## Analysis of *tirant* copies from the *D. simulans* sequenced genome

We retrieved the sequences of the chromosome arms 2L, 2R, 3L, 3R, 4, X, and the unassigned part (named U), of the first release of the mosaic assembly of the genome of *D. simulans* available at the ftp site of the Genome Sequencing Center at the Washington University Medical School (<http://hgdownload.cse.ucsc.edu/downloads.html#droSim>). This mosaic assembly corresponds to different *D. simulans* strains. In the following, we refer to the *tirant* copies using the chromosome name and the start position of the copy (*e.g.*, chr2L\_21040234 corresponds to a copy found on chromosome 2L and that starts at position 21040234).

To search for any kind of copy of *tirant* of the two variant types in the *D. simulans* sequenced genome, we used the sequences of the S type (GenBank AY756122) and the C type (GenBank AY756123) cloned by Fablet *et al.* (14). As these

sequences corresponded only to the 5' LTR-UTR and 5' region of the first open reading frame (ORF) *gag*, we also used the complete sequenced *tirant* from *D. melanogaster* (GenBank AY928610) (17), which is a C type, to identify sequences corresponding to internal parts and to potentially complete copies of *tirant* in *D. simulans*. With all these sequences, we searched for *tirant* copies in the *D. simulans* sequenced genome using Blastn (18). Matches with an  $e\_value < 10^{-10}$  were retained, and those with distances  $< 300$  bp were merged. The detected copies were compared pairwise with the three query sequences using the program matcher (EMBOSS website; <http://emboss.bioinformatics.nl/cgi-bin/emboss/matcher>), and the corresponding percentage identity was computed with the dnadist module from the PHYLIP package (19). The detection of the ORFs was performed using the ORF Finder program at the National Center for Biotechnology Information website ([www.ncbi.nlm.nih.gov/projects/gorf/](http://www.ncbi.nlm.nih.gov/projects/gorf/)).

## RESULTS

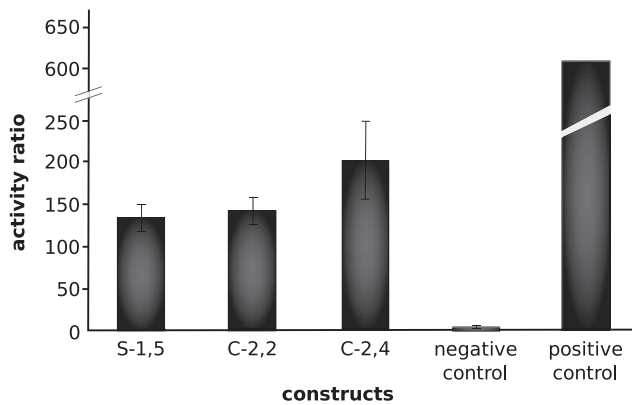
### Reporter gene assays

Three types of *tirant* 5' LTR-UTR regions were tested for the promotion of *lacZ* expression in S2 cells: S-1,5, which is an S type, and C-2,2 and C-2,4, which are C types with two and four 102-bp motifs in the 5' UTR, respectively (see Fig. 1). Note that the number of repeats of the 102-bp motif is always 5 for all S types previously identified in natural populations (14). As shown in **Fig. 2**, the activity ratios for S-1,5, C-2,2, and C-2,4 were significantly different from both the positive and negative controls (see Materials and Methods for

TABLE 1. *Tirant* insertion sites determined by genome walking

| Insertion           | Type | Location        | Comments                            | Length         |
|---------------------|------|-----------------|-------------------------------------|----------------|
| Makindu population  |      |                 |                                     |                |
| 1                   | C    | X               | Unannotated DNA                     | Full length    |
| 2                   | C    | 2L              | <i>tkv</i> gene intron              | Full length    |
| 3                   | C    | 3R              | Unannotated DNA                     | Full length    |
| 4                   | C    | Centromeric DNA | Maupiti islands                     | Full length    |
| 5                   | S    |                 | <i>R1</i> non-LTR retrotransposon   | Full length    |
| 6                   | C    | 2L              | <i>hobo</i> transposon              | Short fragment |
| 7                   | C    | 2L              | <i>frogger</i> LTR retrotransposon  | Short fragment |
| 8                   | S    | 2R              | Heterochromatin                     | Short fragment |
| 9                   | C    |                 | <i>MAX</i> LTR retrotransposon      | Short fragment |
| 10                  | C    | 3R              | Unannotated DNA                     | NA             |
| Zimbabwe population |      |                 |                                     |                |
| 1                   | S    |                 | <i>R1</i> non-LTR retrotransposon   | Full length    |
| 2                   | C    | 3               | Heterochromatin                     | Long fragment  |
| 3                   | C    | 2L              | 3' from CG13786 gene                | Short fragment |
| 4                   | C    | 2L              | Heterochromatin                     | Short fragment |
| 5                   | C    | 3R              | Unannotated DNA                     | Short fragment |
| 6                   | C    | Centromeric DNA | Maupiti islands                     | Short fragment |
| 7                   | C    | 3               | Heterochromatin                     | NA             |
| Chicharo population |      |                 |                                     |                |
| 1                   | C    | Centromeric DNA | Maupiti islands                     | Full length    |
| 2                   | C    |                 | Heterochromatin                     | Long fragment  |
| 3                   | C    |                 | <i>diver2</i> LTR retrotransposon   | Short fragment |
| 4                   | S    |                 | <i>Rt1b</i> non-LTR retrotransposon | NA             |

It is not possible to determine the position on the chromosomes of some sites, especially those corresponding to insertions into other transposable elements. NA, electrophoresis of the PCR product shows multiple bands, possibly indicating embedded insertions of *tirant*.



**Figure 2.** Reporter gene assays, transfection into S2 cells results. Activity ratio: ratio of the  $\beta$ -galactosidase and luciferase activities measured. Negative control corresponds to transfection of an empty *pPelican* vector, and positive control to transfection of a vector carrying the *lacZ* gene downstream of an SV40 promoter. No SD could be calculated for the positive control because 2 of 3 measurements were out of range for the machine used.

details). One-tailed pairwise *t* tests revealed a significant difference between C-2,2 and C-2,4 ( $P=0.012$ ), and between S-1,5 and C-2,4 ( $P=0.006$ ), whereas C-2,2 and S-1,5 were not significantly different ( $P=0.216$ ). This shows that both the C and S types are able to promote the expression of the reporter gene and are therefore theoretically able to promote *tirant* expression in an endogenous context.

#### Analysis of the genomic context of each copy of *tirant*

The reporter gene assay results cannot be directly extrapolated to the endogenous copies of *tirant* without taking into account the genomic localization of each copy (*i.e.*, in euchromatin or heterochromatin), which could influence transcription. We therefore determined the genomic localization of each insertion of *tirant* by the genome-walking technique.

The genome-walking protocol allowed us to identify 10, 7, and 4 *tirant* insertion sites in the genomes of the Makindu, Zimbabwe, and Chicharo populations, respectively. The results are presented in Table 1. The number of insertions was slightly higher than that obtained by *in situ* hybridization (10). This is due to the following facts: 1) only euchromatic copies on the chromosome arms were detected by *in situ* hybridization, 2) full-length copies as well as solo LTRs could be detected by genome walking, whereas the detection size threshold for *in situ* hybridization excluded that for solo LTRs, and 3) the genome walking was done on DNA extracted from a pool of 20 individuals, thus revealing the sum of the insertions resulting from individual variability, whereas insertion sites were estimated for single individuals by the *in situ* hybridization technique.

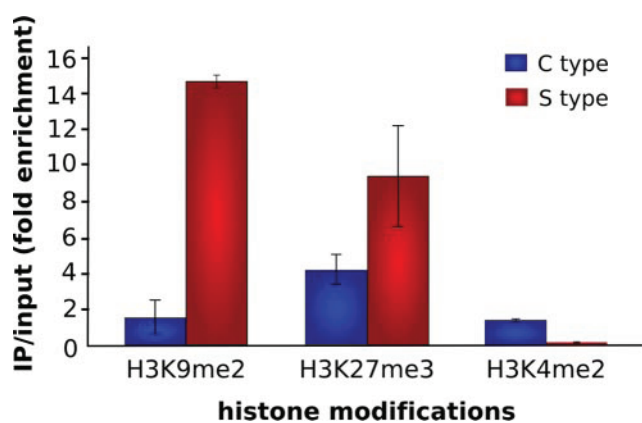
Since the sequenced genome of *D. simulans* was very poorly annotated at the time of the study, we used the annotations of the corresponding regions in the *D.*

*melanogaster* sequenced genome. The only shared site among the genomes of the 3 natural populations analyzed in the present study corresponded to an insertion into the Maupiti Islands centromeric DNA, which is an A + T-rich region displaying many insertions of TEs, mainly non-LTR retrotransposons, which showed a high degree of similarity with the *F* and *doc* elements (9). Of the other 18 insertions of *tirant* identified in the 3 genomes, 7 were found within other TEs, such as *hobo* (20), *frogger* (21), *MAX* (21), *diver2* (22), *RI* (23), and *Rt1b* (24). Five of the 18 insertions were heterochromatic, and the remaining 6 were found in unannotated DNA which probably corresponds to noncoding DNA. With the data of the exact insertion site for each copy of *tirant*, we were able to design primers specific to the flanking region of each individual copy of *tirant*, in order to amplify it by long PCR in the different populations. As expected from previous Southern blot results (14), we obtained full-length amplicons, as well as shorter ones, for each of the 3 populations tested. The detailed results and the correspondence with insertion sites are shown in Table 1. Five of the insertions in the Makindu population were of the expected size for a full-length *tirant*, *i.e.*, 8.5 kb. In the other cases, the amplicons corresponding to short fragments were assumed to be solo LTRs (the amplicons were sequenced for insertions Makindu 7 and Zimbabwe 4). We used a PCR test to check whether the potentially full-length insertions were present in other populations than Makindu, and we found that these insertions were specific to the Makindu population and were not found in the other populations.

To determine whether local chromatin environment could vary between C and S copies, we identified histone modification marks on two different *tirant* full-length copies: C 2 and S 5 (see Table 1). ChIP followed by copy-specific real-time PCR analysis was performed using antisera against permissive (H3K4me2) and repressive (H3K9me2 and H3K27me3) histone marks (25). We observed that the S-type copy of *tirant* is strongly associated with the repressive histone marks H3K9me2 and H3K27me3. In contrast, the C-type copy of *tirant* is characterized by the coenrichment of the repressive mark H3K27me3 and the permissive H3K4me2 (Fig. 3). These observations suggest that the only full-length S-type copy present in the Makindu genome is embedded in a repressive chromatin environment, while the C-type copy can be in a bivalent chromatin domain.

#### D. *simulans* sequenced genome analysis

Among the 20 copies of *tirant* identified in the *D. simulans* sequenced genome, 5 are of the S type and 15 of the C type (Table 2 and Fig. 4A, B). For each type, we found one almost-complete sequence. The full-length S-type copy U\_8880559 is located in the U part of the genome and corresponds to a sequence of 8508 bp. It is flanked by two LTRs of 427 bp sharing 99.30% identity, which indicates that this is a recent insertion. Two ORFs correspond to potentially complete *gag* (positions 2028 to 2978) and *env* (positions 6847 to 8118) genes. The *pol* ORF (positions 3218 to 6455) displays some frameshifts that result in inframe stop



**Figure 3.** Histone marks association with full-length C- and S-type copies of *tirant*. ChIP analysis of C- and S-type copies of *tirant* in *D. simulans* embryos from the population of Makindu. See Materials and Methods for fold enrichment computation. H3K9me2, H3K27me3, and H3K4me2 are epigenetic marks characteristic of constitutive heterochromatin, nonconstitutive heterochromatin, and euchromatin, respectively.

codons. The full-length C-type copy U\_10384153 is also located in the U part of the genome, and corresponds to a sequence of 6803 bp, surrounded by undetermined bases, which made it impossible for us to determine the presence of LTRs at the extremities. Two ORFs correspond to potentially complete *gag* (positions 911 to 2044) and *env* (positions 5693 to 6801) genes; the *pol* ORF (positions 2284 to 5515) displays a large internal deletion that creates a frameshift.

Comparison of the cloned flanking regions with those of the copies identified *in silico* showed that the 2L-random-887610 copy (Table 1; Fig. 4B) is present in both the Makindu and Zimbabwe populations (insertions Ma-

kindu 7 and Zimbabwe 4). The sequencing of PCR products from populations shows that this insertion is a solo LTR. However, it is not possible to conclude about the corresponding insertion found in the sequenced genome, since it is flanked downstream by undetermined bases. The other insertions detected in the 3 populations correspond to unique insertions that are not present in the sequenced genome of *D. simulans*.

## DISCUSSION

The S and C types of *tirant*, previously characterized in *D. melanogaster* and *D. simulans* (14), appear to behave quite differently in the genome of *D. simulans*. It had previously been shown that the C type was polymorphic, presenting varying numbers of a 102-bp motif in its UTR, whereas the S type was monomorphic, with always 5 repeats of this motif in its UTR. In addition, the S-type copy number was low and homogeneous among *D. simulans* natural populations, while the C-type copy number was high in East African populations, but quite low in the surrounding populations (14).

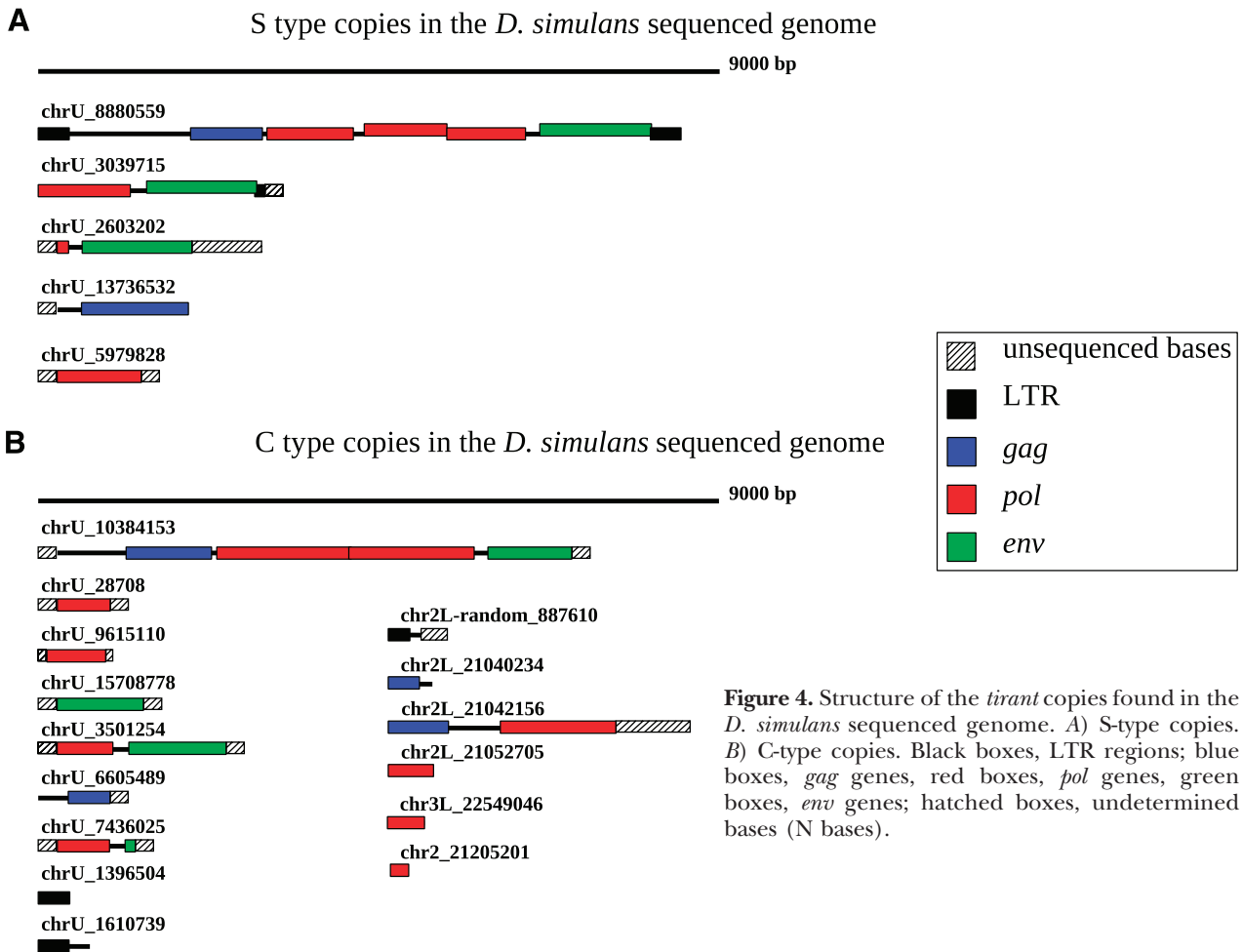
### *Tirant* activity

Reporter gene assays in S2 cells suggest that both the S and C types are able to promote the expression of the reporter gene, with significant differences in the expression strength between the C-type sequences tested. However, as revealed by previous RT-PCR experiments on gonads, only the C-type sequences are expressed when located in the euchromatin (14). The S type, which was localized in the heterochromatin, was not

TABLE 2. Positions of the *tirant* copies in the *D. simulans* sequenced genome

| Chromosome | Strand | Start    | Stop     | Type | Length (bp) | Percentage identity          |                    |                    |              |
|------------|--------|----------|----------|------|-------------|------------------------------|--------------------|--------------------|--------------|
|            |        |          |          |      |             | <i>tirant</i> Dm<br>AY928610 | C type<br>AY756123 | S type<br>AY756122 | chrU_8880559 |
| U          | -      | 13736532 | 13737472 | S    | 942         | 79.10                        | 72.97              | 88.03              | 94.58        |
| U          | -      | 5979828  | 5980937  | S    | 1110        | 85.57                        | NA                 | NA                 | 99.28        |
| U          | -      | 8880559  | 8889066  | S    | 8508        | 80.06                        | 73.35              | 99.47              | /            |
| U          | +      | 3039715  | 3042707  | S    | 2993        | 78.43                        | NA                 | NA                 | 98.47        |
| U          | -      | 2603202  | 2604990  | S    | 1791        | 76.16                        | NA                 | NA                 | 99.44        |
| U          | +      | 1610739  | 1611425  | C    | 687         | 89.70                        | 89.10              | 58.48              | 60.00        |
| 3L         | -      | 22549046 | 22549538 | C    | 493         | 85.39                        | NA                 | NA                 | 82.62        |
| U          | -      | 1396504  | 1396898  | C    | 422         | 96.68                        | 99.26              | 89.01              | 88.99        |
| U          | +      | 28708    | 29409    | C    | 702         | 94.14                        | NA                 | NA                 | 86.55        |
| U          | +      | 9615110  | 9615888  | C    | 780         | 93.93                        | NA                 | NA                 | 85.19        |
| U          | +      | 10384153 | 10390953 | C    | 6803        | 98.63                        | 99.26              | 78.16              | 80.81        |
| U          | -      | 15708778 | 15709918 | C    | 1142        | 95.97                        | NA                 | NA                 | 74.82        |
| U          | +      | 3501254  | 3503485  | C    | 2232        | 97.17                        | NA                 | NA                 | 75.21        |
| U          | +      | 6605489  | 6606440  | C    | 952         | 92.30                        | 94.11              | 79.87              | 79.38        |
| U          | +      | 7436025  | 7437064  | C    | 1040        | 95.76                        | NA                 | NA                 | 82.19        |
| 2L-random  | +      | 887610   | 888048   | C    | 440         | 90.38                        | 90.76              | 86.11              | 87.78        |
| 2L         | +      | 21040234 | 21040766 | C    | 533         | 83.86                        | NA                 | NA                 | 74.56        |
| 2L         | +      | 21042156 | 21045168 | C    | 2845        | 82.96                        | 67.84              | 65.90              | 80.52        |
| 2L         | +      | 21052705 | 21053256 | C    | 552         | 86.79                        | NA                 | NA                 | 84.04        |
| 2L         | +      | 21205201 | 21205440 | C    | 240         | 91.25                        | NA                 | NA                 | 86.67        |

NA, value cannot be computed.



**Figure 4.** Structure of the *tirant* copies found in the *D. simulans* sequenced genome. A) S-type copies. B) C-type copies. Black boxes, LTR regions; blue boxes, *gag* genes, red boxes, *pol* genes, green boxes, *env* genes; hatched boxes, undetermined bases (N bases).

expressed in any of the natural populations analyzed (14), but we show in this study, by the *in vitro* experiments, that it is potentially able to drive expression. Two hypotheses can be proposed to explain this discrepancy: either the S type is inhibited in its endogenous natural context (e.g., by a repressive genomic environment, such as heterochromatin), or it is able to promote expression in a *D. melanogaster* genetic context (e.g., S2 cells), but is repressed in a *D. simulans* genetic context. An alternative hypothesis would be that *tirant* expression is allowed in S2 cells, which are derived from embryonic cells, whereas it could be specifically repressed in some tissues, such as ovaries.

The observed situation for *tirant* in *D. simulans* is comparable to cases reported in *D. melanogaster*, in which one line displays genetic instability—also known as the permissive line—with an elevated copy number and a high rate of activity for a particular TE, such as the endogenous retroviruses *gypsy* (26), *ZAM*, and *Idefix* (27). Stable lines—also known as restrictive lines—are not totally devoid of these elements but have few or no copies in the chromosome arms (28, 29). The Makindu population, which has full-length copies of *tirant* in the euchromatin, behaves like a permissive line in which the control of *tirant* is partly or totally impaired, which would explain its activity and relative high copy number in this population. In contrast, the other populations

analyzed in this study only have a few inactive heterochromatic copies of *tirant*.

#### *Tirant* copies genomic contexts

##### *S* type

The genome-walking protocol showed the presence of one or two S-type insertions per genome (i.e., per population), which is consistent with what was previously found by Southern blot analysis (14). One of these insertions is heterochromatic (Makindu 8), and the other three are located in copies of either the *R1* non-LTR retrotransposon (Makindu 5 and Zimbabwe 1) or the *Rt1b* non-LTR retrotransposon (Chicharo 4). *Rt1b*, also known as *Waldo-A*, belongs to the *R1* clade (24). We cannot precisely locate the insertions of these elements, but we know that a high proportion of *R1* copies is located in centromeric heterochromatin (23), or in ribosomal genes that are often inactivated by local heterochromatin formation (30). In this regard, we found that the unique full-length S-type copy of *tirant* in the Makindu genome is associated with H3K9me2 and H3K27me3, two specific repressive histone marks (25). This confirms the assumption that S-type copies of *tirant* are inserted into inactivated genomic regions, mainly heterochromatin, which could explain the ab-

sence of S-type transcripts in endogenous conditions revealed by RT-PCR, even though the S-type 5' LTR-UTR region was shown to drive expression in the S2 cells.

In the sequenced genome of *D. simulans*, 5 S-type copies were detected, only one of which corresponded to a full-length copy. The shorter ones displayed high-percentage identities with the full-length S-type copy, except the U\_13736243 copy (94.58% identity, Table 2), which is the shortest sequence for this type. Moreover, the shorter copies were surrounded by undetermined bases, indicating that they could be longer, and that we are probably underestimating the copy number of the S type. We were unable to identify the exact location of the S-type insertions, since the *D. simulans* sequenced genome was not annotated yet.

### C type

Unlike the S type, which appears to be exclusively heterochromatic, insertions of the C type are both euchromatic and heterochromatic. In the Chicharo genome, in which no *tirant* copy can be detected using *in situ* hybridization, 3 C-type copies were found when analyzing the genomic sequence, one in the heterochromatin, another in centromeric DNA (Maupiti Islands), which is also heterochromatic, and the last one in the LTR retrotransposon *diver2*. *Diver2* insertions in the *D. melanogaster* genome are either centromeric or telomeric (22) and so correspond to heterochromatin. Therefore, all *tirant* insertions found by genome walking in the Chicharo genome can be assumed to be heterochromatic.

The *tirant* C-type insertion 7 from Zimbabwe is of particular interest since its flanking region displays blast matches with the heterochromatic gene *parp* (31). Tulin *et al.* (31) showed that the genomic region where *parp* is located in *D. melanogaster* is rich in TEs, especially *gypsy* elements that have lost their LTRs and insulators. These authors also showed that a mutation in the regulatory region of the gene *parp* deregulates the LTR-retrotransposon *copia*, the transcription level of which increases 50-fold. The idea that the insertion of *tirant* near *parp* in the Zimbabwe genome could regulate other TEs, or promote changes in chromatin structure warrants further investigation. While it was previously assumed from Southern blot data that the Zimbabwe genome had no full-length copy of *tirant*, the long PCR experiments suggest that insertion 1 is potentially full length. This discrepancy could be explained by the presence of restriction polymorphism in this sequence.

In the Makindu genome, the size of PCR products obtained for half of the detected insertions indicates that the copies are (potentially) full length. Insertion sites are varied for these copies: intron of a gene, Maupiti Islands, *RI* transposable element, and unannotated DNA. Interestingly, we observed that one of these full-length C-type *tirant* copies, insertion C 2, harbors simultaneously repressive H3K27me3 and permissive H3K4me2 histone marks. In the past few years, several studies in mouse and human ES cells reported the occurrence of such "bivalent domains" predominantly on key developmental regulators (reviewed in Pietersen and van Lohuizen; ref. 32). It was proposed that this unusual combination of marks keeps genes repressed or expressed at very low level but poised

for later activation or complete repression (32). In this regard, we can propose that the Makindu genome contains C-type copies of *tirant* susceptible to be expressed in response to appropriate developmental or cellular cues. Full-length insertions are thus not associated specifically with euchromatin but are also present in heterochromatic regions.

### Internal dynamics of *tirant* in the *D. simulans* genome: the cohabitation of two subfamilies

Subfamilies with differing regulatory regions have been reported for various TEs, such as *copia* (33), *blood* (34), and *412* (35) in *Drosophila*, and *Tnt1* in plants (36). In some cases, such as *copia*, subfamilies are associated with different levels of expression (4, 32). The situation observed for *blood* in the *D. melanogaster* subgroup of species (34), is very similar to what we found for *tirant*. One of the *blood* subfamilies, long (L), is mainly heterochromatic, and its insertion sites are shared by most populations. The other subfamily, short (S), is potentially active and euchromatic. One particular population of *D. simulans* has more *blood* insertions on the chromosome arms than the other populations of the species (13), which is interpreted as an invasion of euchromatin by the S subfamily (34). The authors, therefore, assume the existence of competition between the subfamilies, leading to the elimination from euchromatin of the L subfamily, which is replaced by the S subfamily. This situation is quite similar to what we observe for *tirant*, with the S type restricted to the heterochromatin and untranscribed, and the C type that has invaded euchromatin in the East African populations of *D. simulans*.

The dynamics of the *tirant* types is not incompatible with an overall loss of the element from the genome of most *D. simulans* populations worldwide, as was previously proposed (14) but indicates that one population (Makindu), which can be considered as permissive for *tirant*, has been subjected to some deregulation, leading to transpositions, and an increase in copy number. Our results also suggest that the S and C types of *tirant* are subjected to several different mechanisms of regulation, which include chromatin conformation and therefore the epigenetic regulation machinery.

For a long time, heterochromatin has been considered to be the graveyard of TEs, since the absence of recombination would lead to the accumulation of these sequences (37). This model is consistent with data showing that TEs are organized in clusters in the heterochromatin and that most of them are highly rearranged and deleted (6, 38). However, analyses of the *Drosophila* heterochromatic sequences (8, 39) and several experimental studies have shown that full-length elements persist in this part of the genome, and in some cases could be active (7, 9). The identification of full-length S-type copies of *tirant* in the heterochromatin of the *D. simulans* natural populations and sequenced genome, associated with the data obtained for the *in vitro* expression of these type of elements, suggest that the heterochromatin harbors potentially invasive elements, which may be reactivated under particular

conditions that have not yet been determined and that warrant further studies. FJ

We thank Christian Biémont for useful comments, and Monika Ghosh for reviewing the English text. The work in C. Vieira's laboratory was funded by the Centre National de la Recherche Scientifique (UMR 5558 and GDR 2157 on TEs). The work in E.G.'s laboratory was funded by ARC (ARECA program on epigenetic profiling).

## REFERENCES

1. Biémont, C., and Vieira, C. (2005) What transposable elements tell us about genome organization and evolution: the case of *Drosophila*. *Cytogenet. Genome Res.* **110**, 25–34
2. Biémont, C., and Vieira, C. (2006) Genetics: junk DNA as an evolutionary force. *Nature* **443**, 521–524
3. Biémont, C., Vieira, C., Hoogland, C., Cizeron, G., Loevenbruck, C., Arnault, C., and Carante, J. P. (1997) Maintenance of transposable element copy number in natural populations of *Drosophila melanogaster* and *D. simulans*. *Genetica* **100**, 161–166
4. Vieira, C., and Biémont, C. (2004) Transposable element dynamics in two sibling species: *Drosophila melanogaster* and *Drosophila simulans*. *Genetica* **120**, 115–123
5. Bergman, C. M., Quesneville, H., Anxolabehere, D., and Ashburner, M. (2006) Recurrent insertion and duplication generate networks of transposable element sequences in the *Drosophila melanogaster* genome. *Genome Biol.* **7**, R112
6. Vauray, C., Bucheton, A., and Pelisson, A. (1989) The beta heterochromatic sequences flanking the I elements are themselves defective transposable elements. *Chromosoma* **98**, 215–224
7. Kogan, G. L., Tulin, A. V., Aravin, A. A., Abramov, Y. A., Kalmykova, A. I., Maisonhaute, C., and Gvozdev, V. A. (2003) The GATE retrotransposon in *Drosophila melanogaster*: mobility in heterochromatin and aspects of its expression in germline tissues. *Mol. Genet. Genomics.* **269**, 234–242
8. Mugnier, N., Gueguen, L., Vieira, C., and Biémont, C. (2008) The heterochromatic copies of the LTR retrotransposons as a record of the genomic events that have shaped the *Drosophila melanogaster* genome. *Gene* **411**, 87–93
9. Sun, X., Le, H. D., Wahlstrom, J. M., and Karpen, G. H. (2003) Sequence analysis of a functional *Drosophila* centromere. *Genome Res.* **13**, 182–194
10. Biémont, C., Nardon, C., Deceliere, G., Lepetit, D., Loevenbruck, C., and Vieira, C. (2003) Worldwide distribution of transposable element copy number in natural populations of *Drosophila simulans*. *Evolution Int. J. Org. Evolution* **57**, 159–167
11. Charlesworth, B., Jarne, P., and Assimacopoulos, S. (1994) The distribution of transposable elements within and between chromosomes in a population of *Drosophila melanogaster*. III. Element abundances in heterochromatin. *Genet. Res.* **64**, 183–197
12. Junakovic, N., Terrinoni, A., Di Franco, C., Vieira, C., and Loevenbruck, C. (1998) Accumulation of transposable elements in the heterochromatin and on the Y chromosome of *Drosophila simulans* and *Drosophila melanogaster*. *J. Mol. Evol.* **46**, 661–668
13. Vieira, C., Lepetit, D., Dumont, S., and Biémont, C. (1999) Wake up of transposable elements following *Drosophila simulans* worldwide colonization. *Mol. Biol. Evol.* **16**, 1251–1255
14. Fablet, M., McDonald, J. F., Biémont, C., and Vieira, C. (2006) Ongoing loss of the tirant transposable element in natural populations of *Drosophila simulans*. *Gene* **375**, 54–62
15. Barolo, S., Carver, L. A., and Posakony, J. W. (2000) GFP and beta-galactosidase transformation vectors for promoter/enhancer analysis in *Drosophila*. *BioTechniques* **29**, 726, 728, 730, 732
16. Sandmann, T., Jakobsen, J. S., and Furlong, E. E. (2006) ChIP-on-chip protocol for genome-wide analysis of transcription factor binding in *Drosophila melanogaster* embryos. *Nat. Protocols* **1**, 2839–2855
17. Lerat, E., Rizzon, C., and Biémont, C. (2003) Sequence divergence within transposable element families in the *Drosophila melanogaster* genome. *Genome Res.* **13**, 1889–1896
18. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402
19. Felsenstein, J. (1989) PHYLIP-Phylogeny Inference Package (Version 3.2). *Cladistics* **5**, 3
20. Bonnard, E., Higuier, D., and Bazin, C. (1997) Characterization of natural populations of *Drosophila melanogaster* with regard to the hobo system: a new hypothesis on the invasion. *Genet. Res.* **69**, 197–208
21. Kaminker, J. S., Bergman, C. M., Kronmiller, B., Carlson, J., Svirskas, R., Patel, S., Frise, E., Wheeler, D. A., Lewis, S. E., Rubin, G. M., Ashburner, M., and Celniker, S. E. (2002) The transposable elements of the *Drosophila melanogaster* euchromatin: a genomics perspective. *Genome Biol.* **3**, RESEARCH0084
22. Kapitonov, V. V., and Jurka, J. (2003) Molecular paleontology of transposable elements in the *Drosophila melanogaster* genome. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 6569–6574
23. Eickbush, D. G., and Eickbush, T. H. (1995) Vertical transmission of the retrotransposable elements R1 and R2 during the evolution of the *Drosophila melanogaster* species subgroup. *Genetics* **139**, 671–684
24. Busseau, I., Berezikov, E., and Bucheton, A. (2001) Identification of Waldo-A and Waldo-B, two closely related non-LTR retrotransposons in *Drosophila*. *Mol. Biol. Evol.* **18**, 196–205
25. Ebert, A., Lein, S., Schotta, G., and Reuter, G. (2006) Histone modification and the control of heterochromatic gene silencing in *Drosophila*. *Chromosome Res.* **14**, 377–392
26. Kim, A., Terzian, C., Santamaria, P., Pelisson, A., Purd'homme, N., and Bucheton, A. (1994) Retroviruses in invertebrates: the gypsy retrotransposon is apparently an infectious retrovirus of *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. U. S. A.* **91**, 1285–1289
27. Desset, S., Conte, C., Dimitri, P., Calco, V., Dastugue, B., and Vauray, C. (1999) Mobilization of two retroelements, ZAM and Idefix, in a novel unstable line of *Drosophila melanogaster*. *Mol. Biol. Evol.* **16**, 54–66
28. Desset, S., Meignin, C., Dastugue, B., and Vauray, C. (2003) COM, a heterochromatic locus governing the control of independent endogenous retroviruses from *Drosophila melanogaster*. *Genetics* **164**, 501–509
29. Prud'homme, N., Gans, M., Masson, M., Terzian, C., and Bucheton, A. (1995) Flamenco, a gene controlling the gypsy retrovirus of *Drosophila melanogaster*. *Genetics* **139**, 697–711
30. Strohnner, R., Nemeth, A., Nightingale, K. P., Grummt, I., Becker, P. B., and Langst, G. (2004) Recruitment of the nucleolar remodeling complex NoRC establishes ribosomal DNA silencing in chromatin. *Mol. Cell. Biol.* **24**, 1791–1798
31. Tulin, A., Stewart, D., and Spradling, A. C. (2002) The *Drosophila* heterochromatic gene encoding poly(ADP-ribose) polymerase (PARP) is required to modulate chromatin structure during development. *Genes Dev.* **16**, 2108–2119
32. Pietersen, A. M., and van Lohuizen, M. (2008) Stem cell regulation by polycomb repressors: postponing commitment. *Curr. Opin. Cell. Biol.* **20**, 201–207
33. Matyunina, L. V., Jordan, I. K., and McDonald, J. F. (1996) Naturally occurring variation in copia expression is due to both element (cis) and host (trans) regulatory variation. *Proc. Natl. Acad. Sci. U. S. A.* **93**, 7097–7102
34. Costas, J., Valade, E., and Naveira, H. (2001) Amplification and phylogenetic relationships of a subfamily of blood, a retrotransposable element of *Drosophila*. *J. Mol. Evol.* **52**, 342–350
35. Mugnier, N., Biémont, C., and Vieira, C. (2005) New regulatory regions of *Drosophila* 412 retrotransposable element generated by recombination. *Mol. Biol. Evol.* **22**, 747–757
36. Vernhettes, S., Grandbastien, M. A., and Casacuberta, J. M. (1998) The evolutionary analysis of the Tnt1 retrotransposon in *Nicotiana* species reveals the high variability of its regulatory sequences. *Mol. Biol. Evol.* **15**, 827–836
37. Charlesworth, B., and Langley, C. H. (1989) The population genetics of *Drosophila* transposable elements. *Annu. Rev. Genet.* **23**, 251–287
38. Daniels, S. B., and Strausbaugh, L. D. (1986) The distribution of P-element sequences in *Drosophila*: the willistoni and saltans species groups. *J. Mol. Evol.* **23**, 138–148
39. Hoskins, R. A., Carlson, J. W., Kennedy, C., Acevedo, D., Evans-Holm, M., Frise, E., Wan, K. H., Park, S., Mendez-Lago, M., Rossi, F., Villasante, A., Dimitri, P., Karpen, G. H., and Celniker, S. E. (2007) Sequence finishing and mapping of *Drosophila melanogaster* heterochromatin. *Science* **316**, 1625–1628

Received for publication October 7, 2008.  
Accepted for publication December 18, 2008.

## **SUPPLEMENTARY TABLE S1.**

*Populations tested for the presence of Makindu-specific insertions of tirant.*

In brackets are the number of isofemale lines tested for each population.

|                                 |                            |
|---------------------------------|----------------------------|
| Antibes (France) (3)            | Kwale (Mayotte) (1)        |
| Arusha (Tanzania) (1)           | Madeira (3)                |
| Austria (1)                     | Malta (1)                  |
| Brazzaville (Congo) (1)         | Moscow (Russia) (3)        |
| Canaries (1)                    | Nasrallah (Tunisia) (3)    |
| Capri (Italy) (1)               | Privas (France) (2)        |
| Chicharo (Portugal) (2)         | Rome (Italy) (1)           |
| Cordoba (Spain) (1)             | Saint Cyprien (France) (1) |
| Czech Republic (1)              | Sotchi (Russia) (2)        |
| Djerba (Tunisia) (1)            | Tanta (Egypt) (2)          |
| Grand Ferrade (France) (4)      | Valence (France) (3)       |
| Israel (2)                      | Yaounde (Cameroon) (2)     |
| Johannesburg (South Africa) (1) | Zimbabwe (2)               |

## **SUPPLEMENTARY TABLE S2.**

*Primers used to amplify each individual insertion of tirant in the genome of natural populations.*

| population | insertion # | primer sequence  |
|------------|-------------|--|
| Makindu    | 1           | Fw: 5' CAG TGG GCA CTA GCT AAG ATG GCA GTC 3'<br>Rv: 5' C ATT ACA CCC ACG ATC CAC ATG ACA TCT ACC 3'     |
|            | 2           | Fw: 5' GC GTT GCT CTA AGG GGA TGA AAA GGG 3'<br>Rv: 5' CT GAG CAC TTG ATT TGG GCT TAG ACA GGC 3'         |
|            | 3           | Fw: 5' CC TTT GGC AAC TGC TCG AGT GTT GTT TTG 3'<br>Rv: 5' CCA TTA ATT GGC TGC AAG CGC GAG TTA CC 3'     |
|            | 4           | Fw: 5' GGT GGT CTG CAC GCG AGT TAC CAC 3'<br>Rv: 5' GTT CAC TAG TGC ACT TGA GCT TTT TTC GCG 3'           |
|            | 5           | Fw: 5' TGC TCT CAA CTG CGC GCG AGT TAC CAC 3'<br>Rv: 5' GCG GGG AAG AAG ACG TGG GTA TAG C 3'             |
|            | 6           | Fw: 5' CGC TCA CCA AGT TAG CCG CTG TAG ACC 3'<br>Rv: 5' CGA TCG GCC GCA GGT CAC TTT TTG TAC 3'           |
|            | 7           | Fw: 5' ATT TTC AAT TGA CGC GCG AGT TAC CAC CCC 3'<br>Rv: 5' CA CAA AGT GAG AAC CGC GAC GCC AAT TCC 3'    |
|            | 8           | Fw: 5' CAT TTA CCG TTC ATT TGC TCG GCG TCC 3'<br>Rv: 5' CAT ACA GTT GCC GAC GTT TTC CAT TGG 3'           |
|            | 9           | Fw: 5' G TAC TCA AAG CTC CTC CAC GGC TCG 3'<br>Rv: 5' G GAG AGA AGG AGA GCG AGT TAC CAC 3'               |
|            | 10          | Fw: 5' AGC GGA GTG CGC ACG AGT TAC CAC 3'<br>Rv: 5' GTC ATC AGT TGA CGA ACG GCT ATG GAA TC 3'            |
| Zimbabwe   | 1           | Fw: 5' CTC TCA ACT GCG CGC GAG TTA CCA C 3'<br>Rv: 5' GAG AGT GGA CGA GTG GGA ATT GAG TGA C 3'           |
|            | 2           | Fw: 5' CAG ATG ACC GTA TCG TTC GTG CC 3'<br>Rv: 5' GGT GGT AAC TCG CGC GTC AAT TG 3'                     |
|            | 3           | Fw: 5' CAT TCG CTG GCC AAA AGG CAC GAG TTA C 3'<br>Rv: 5' CAT GGC TGG CTC ATG GGT TTG TC 3'              |
|            | 4           | Fw: 5' TTA TTT TCA ATT GAC GCG CGA GTT ACC ACC CCA C 3'<br>Rv: 5' GTG AGA ACC GCG ACG CCA ATT CCA ATG 3' |
|            | 5           | Fw: 5' CAT TTT TTT ATT ATG CGC CGC GCG AGT TAC 3'<br>Rv: 5' TAG CAG GGC TTC CTG CCG ATT TGC 3'           |
|            | 6           | Fw: 5' TAT CCG GTG GTC TGC ACG CGA GTT AC 3'<br>Rv: 5' GTG GCT ATT TTG TTC ACT AAT GCA CTT GAG C 3'      |
| Chicharo   | 1           | Fw: 5' TCC TCC AAT ATA AGT TAA CAC AAG CAG TTC T 3'<br>Rv: 5' GGC TAT TTT GTT CAC TAG TGC ACT TGA GC 3'  |
|            | 2           | Fw: 5' ATC CGG TGG TCT GCA CGC GAG TTA C 3'<br>Rv: 5' GTT TGA CCT CAC ACA TAG AGA GTG GCT 3'             |
|            | 3           | Fw: 5' TGT GTC TGG AGG CTG CGC GAG TTA C 3'<br>Rv: 5' AGC TAT TAA GGC GCT ACC GCA AGG 3'                 |
|            | 4           | Fw: 5' GAG GAG TTT CTG GAG GCG CTA GTT ACC AC 3'<br>Rv: 5' CTT CGG GCT TCG CCC TCT GCG AT 3'             |

### SUPPLEMENTARY TABLE S3.

*Primers used to test for the presence of Makindu-specific insertions of tirant in natural populations of D. simulans.*

For each of the following PCR reactions, the forward primer is located within the ORF envelope, and the reverse primer is designed in the 3' flanking region of *tirant*.

| Insert ion | Forward primer                   | Reverse primer                    | hybrid temp |
|------------|----------------------------------|-----------------------------------|-------------|
| #1         | 5' AAC GCC CCT ATA GCC AAG AT 3' | 5' CAC GAT CCA CAT GAC ATC TAC 3' | 54°C        |
| #2         |                                  |                                   |             |
| #3         |                                  | 5' GCG TTG CTC TAA GGG GAT G 3'   | 54°C        |
|            |                                  | 5' GCA ACT GCT CGA GTG TTG TT 3'  | 55°C        |

The data obtained in Fablet et al comes from somatic tissues analysis since embryos were used in all experiments (Fablet et al., 2009). Transposition events in somatic tissues, a already discussed above, may have deleterious effects for the genome but also may have an impact in germline regulation. Germline regulation of TEs is necessary for genome integrity so inheritance of proper genetic material can be done. Difference between somatic and germinal regulation of TEs has already been described but again, no variation in natural populations has ever been researched. In order to understand TE regulation in germinal tissues of natural populations, I assayed for *tirant* transcription (*gag*, *pol* and *env*) in ovaries of both *D. melanogaster* and *D. simulans* natural populations (Figure 1 hereafter). Again, different expression levels are observed among natural populations of *D. simulans* where only one strain is active. Moreover, contrarily to the expression test made in embryos and presented above, *D. melanogaster* natural populations are silenced for *tirant*. Therefore, a silencing mechanism is present in the germline but is absent in the somatic tissues. An alternative hypothesis is that enhancers of *tirant* expression are present in the somatic tissues but absent in the germinal line. Further experiments are being held in the laboratory in order to understand *tirant* regulation in germinal tissues.

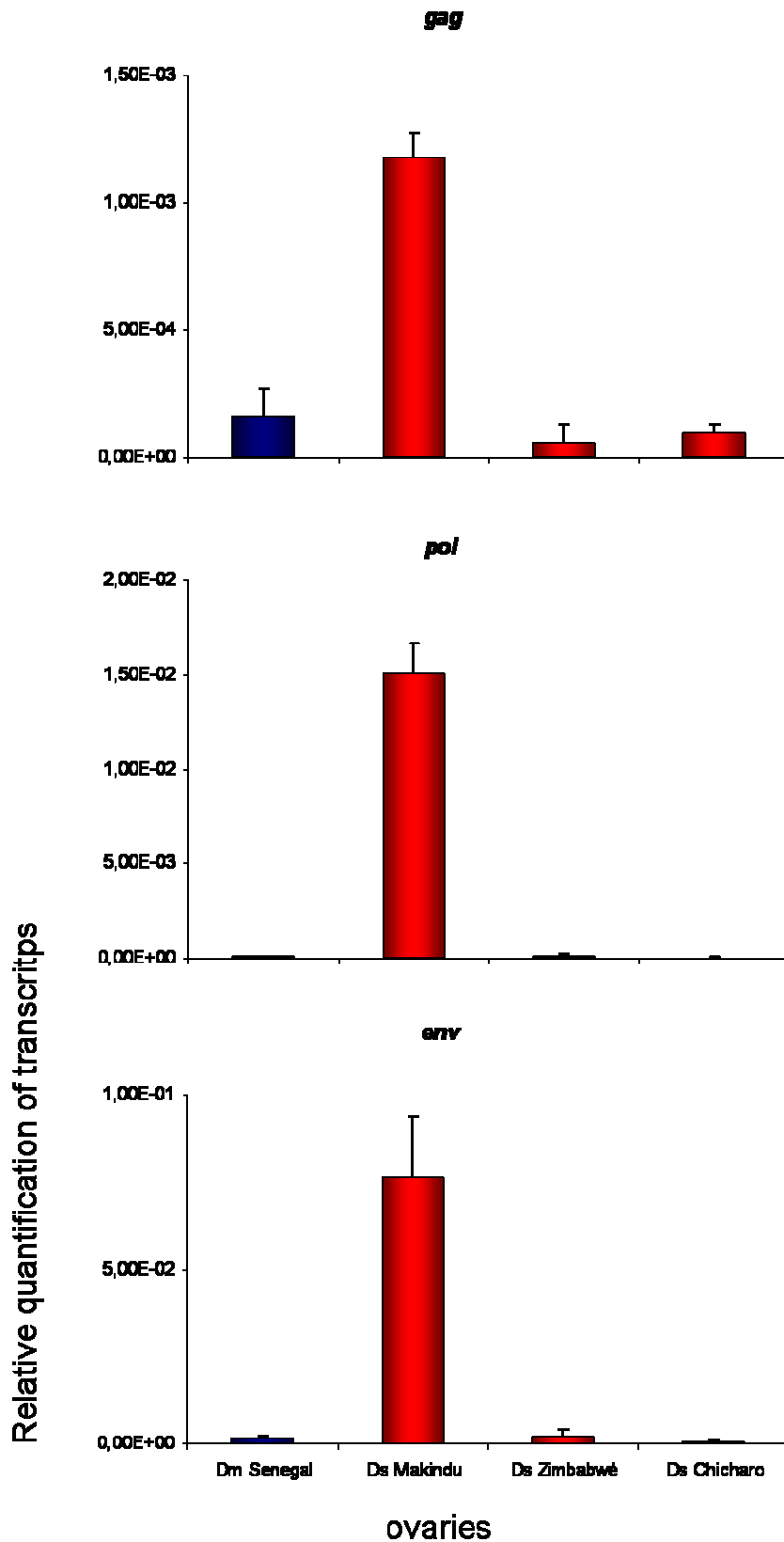


Figure 1. Relative quantification of *tirant gag*, *pol* and *env* genes in natural populations of *D. melanogaster* (blue) and *D. simulans* (red). High expression of *tirant* in ovaries of *D. simulans* Makindu strain is observed. For materials and methods please refer to article page 88. These results are part of Hubert et al. in preparation.

### **The *tirant* element is an endogenous retrovirus**

The infectious quality of *tirant* was analysed through the characterization of the *env* gene by Marie Fablet, Nelly Burlet, Cristina Vieira and I. The *env* gene is present in all full length copies of *tirant*. The real time PCR I ran for the *env* gene shows that it is expressed in the germinal line of only two strains of *D. simulans* (Makindu and Mayotte) and it is silenced in *D. melanogaster* strains (as described above). The ENV protein is observed only in ovaries of the *D. simulans* Makindu strain (Mayotte was not yet analysed). Interestingly, contrary to other ERVs analysed in *Drosophila*, the *tirant* ENV is located inside the nucleus of germinative cells. Further investigation is currently being held in the laboratory in order to understand such germinal localisation and is part of Fablet et al. in preparation.



## References

- Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408 (2000) 796-815.
- Finishing the euchromatic sequence of the human genome. *Nature* 431 (2004) 931-45.
- Adey, N.B., Schichman, S.A., Graham, D.K., Peterson, S.N., Edgell, M.H. and Hutchison, C.A., 3rd: Rodent L1 evolution has been driven by a single dominant lineage that has repeatedly acquired new transcriptional regulatory sequences. *Mol Biol Evol* 11 (1994) 778-89.
- Ballestar, E.: Epigenetics Lessons from Twins: Prospects for Autoimmune Disease. *Clin Rev Allergy Immunol* (2009).
- Barbot, W., Dupressoir, A., Lazar, V. and Heidmann, T.: Epigenetic regulation of an IAP retrotransposon in the aging mouse: progressive demethylation and de-silencing of the element by its repetitive induction. *Nucleic Acids Res* 30 (2002) 2365-73.
- Bartolome, C., Bello, X. and Maside, X.: Widespread evidence for horizontal transfer of transposable elements across *Drosophila* genomes. *Genome Biol* 10 (2009) R22.
- Beadle, G.W. and Tatum, E.L.: Genetic Control of Biochemical Reactions in *Neurospora*. *Proc Natl Acad Sci U S A* 27 (1941) 499-506.
- Bennetzen, J.L.: Mechanisms and rates of genome expansion and contraction in flowering plants. *Genetica* 115 (2002) 29-36.
- Bernatavichute, Y.V., Zhang, X., Cokus, S., Pellegrini, M. and Jacobsen, S.E.: Genome-wide association of histone H3 lysine nine methylation with CHG DNA methylation in *Arabidopsis thaliana*. *PLoS One* 3 (2008) e3156.
- Biemont, C.: Genome size evolution: within-species variation in genome size. *Heredity* 101 (2008) 297-8.
- Biemont, C. and Vieira, C.: Genetics: junk DNA as an evolutionary force. *Nature* 443 (2006) 521-4.
- Bohne, A., Brunet, F., Galiana-Arnoux, D., Schultheis, C. and Volff, J.N.: Transposable elements as drivers of genomic and biological diversity in vertebrates. *Chromosome Res* 16 (2008) 203-15.
- Bollati, V., Baccarelli, A., Hou, L., Bonzini, M., Fustinoni, S., Cavallo, D., Byun, H.M., Jiang, J., Marinelli, B., Pesatori, A.C., Bertazzi, P.A. and Yang, A.S.: Changes in

- DNA methylation patterns in subjects exposed to low-dose benzene. *Cancer Res* 67 (2007) 876-80.
- Borie, N., Loevenbruck, C. and Biemont, C.: Developmental expression of the 412 retrotransposon in natural populations of *D. melanogaster* and *D. simulans*. *Genet Res* 76 (2000) 217-26.
- Bossdorf, O., Richards, C.L. and Pigliucci, M.: Epigenetics for ecologists. *Ecol Lett* 11 (2008) 106-15.
- Boulesteix, M., Weiss, M. and Biemont, C.: Differences in genome size between closely related species: the *Drosophila melanogaster* species subgroup. *Mol Biol Evol* 23 (2006) 162-7.
- Brennecke, J., Aravin, A.A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R. and Hannon, G.J.: Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. *Cell* 128 (2007) 1089-103.
- Brennecke, J., Malone, C.D., Aravin, A.A., Sachidanandam, R., Stark, A. and Hannon, G.J.: An epigenetic role for maternally inherited piRNAs in transposon silencing. *Science* 322 (2008) 1387-92.
- Brown, J.D., Golden, D. and O'Neill, R.J.: Methylation perturbations in retroelements within the genome of a *Mus* interspecific hybrid correlate with double minute chromosome formation. *Genomics* 91 (2008) 267-73.
- Bucheton, A., Paro, R., Sang, H.M., Pelisson, A. and Finnegan, D.J.: The molecular basis of I-R hybrid dysgenesis in *Drosophila melanogaster*: identification, cloning, and properties of the I factor. *Cell* 38 (1984) 153-63.
- Callinan, P.A. and Batzer, M.A.: Retrotransposable elements and human disease. *Genome Dyn* 1 (2006) 104-15.
- Capy, P. and Gibert, P.: *Drosophila melanogaster*, *Drosophila simulans*: so similar yet so different. *Genetica* 120 (2004) 5-16.
- Carbone, L., Harris, R.A., Vessere, G.M., Mootnick, A.R., Humphray, S., Rogers, J., Kim, S.K., Wall, J.D., Martin, D., Jurka, J., Milosavljevic, A. and de Jong, P.J.: Evolutionary breakpoints in the gibbon suggest association between cytosine methylation and karyotype evolution. *PLoS Genet* 5 (2009) e1000538.
- Casavant, N.C., Scott, L., Cantrell, M.A., Wiggins, L.E., Baker, R.J. and Wichman, H.A.: The end of the LINE?: lack of recent L1 activity in a group of South American rodents. *Genetics* 154 (2000) 1809-17.

- Castro, J.P. and Carareto, C.M.: *Drosophila melanogaster* P transposable elements: mechanisms of transposition and regulation. *Genetica* 121 (2004) 107-18.
- Chambeyron, S., Popkova, A., Payen-Groschene, G., Brun, C., Laouini, D., Pelisson, A. and Bucheton, A.: piRNA-mediated nuclear accumulation of retrotransposon transcripts in the *Drosophila* female germline. *Proc Natl Acad Sci U S A* 105 (2008) 14964-9.
- Chinnusamy, V. and Zhu, J.K.: Epigenetic regulation of stress responses in plants. *Curr Opin Plant Biol* 12 (2009) 133-9.
- Cho, K., Lee, Y.K. and Greenhalgh, D.G.: Endogenous retroviruses in systemic response to stress signals. *Shock* 30 (2008) 105-16.
- Clark, A.G., Eisen, M.B., Smith, D.R., Bergman, C.M., Oliver, B., Markow, T.A., Kaufman, T.C., Kellis, M., Gelbart, W., Iyer, V.N., Pollard, D.A., Sackton, T.B., Larracuent, A.M., Singh, N.D., Abad, J.P., Abt, D.N., Adryan, B., Aguade, M., Akashi, H., Anderson, W.W., Aquadro, C.F., Ardell, D.H., Arguello, R., Artieri, C.G., Barbash, D.A., Barker, D., Barsanti, P., Batterham, P., Batzoglou, S., Begun, D., Bhutkar, A., Blanco, E., Bosak, S.A., Bradley, R.K., Brand, A.D., Brent, M.R., Brooks, A.N., Brown, R.H., Butlin, R.K., Caggese, C., Calvi, B.R., Bernardo de Carvalho, A., Caspi, A., Castrezana, S., Celniker, S.E., Chang, J.L., Chapple, C., Chatterji, S., Chinwalla, A., Civetta, A., Clifton, S.W., Comeron, J.M., Costello, J.C., Coyne, J.A., Daub, J., David, R.G., Delcher, A.L., Delehaunty, K., Do, C.B., Ebling, H., Edwards, K., Eickbush, T., Evans, J.D., Filipinski, A., Findeiss, S., Freyhult, E., Fulton, L., Fulton, R., Garcia, A.C., Gardiner, A., Garfield, D.A., Garvin, B.E., Gibson, G., Gilbert, D., Gnerre, S., Godfrey, J., Good, R., Gotea, V., Gravely, B., Greenberg, A.J., Griffiths-Jones, S., Gross, S., Guigo, R., Gustafson, E.A., Haerty, W., Hahn, M.W., Halligan, D.L., Halpern, A.L., Halter, G.M., Han, M.V., Heger, A., Hillier, L., Hinrichs, A.S., Holmes, I., Hoskins, R.A., Hubisz, M.J., Hultmark, D., Huntley, M.A., Jaffe, D.B., Jagadeeshan, S., et al.: Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* 450 (2007) 203-18.
- Cropley, J.E., Suter, C.M., Beckman, K.B. and Martin, D.I.: Germ-line epigenetic modification of the murine A<sub>vy</sub> allele by nutritional supplementation. *Proc Natl Acad Sci U S A* 103 (2006) 17308-12.
- Czermín, B., Melfi, R., McCabe, D., Seitz, V., Imhof, A. and Pirrotta, V.: *Drosophila* enhancer of Zeste/ESC complexes have a histone H3 methyltransferase activity that marks chromosomal Polycomb sites. *Cell* 111 (2002) 185-96.

- de Boer, J.G., Yazawa, R., Davidson, W.S. and Koop, B.F.: Bursts and horizontal evolution of DNA transposons in the speciation of pseudotetraploid salmonids. *BMC Genomics* 8 (2007) 422.
- de Wit, E., Greil, F. and van Steensel, B.: Genome-wide HP1 binding in *Drosophila*: developmental plasticity and genomic targeting signals. *Genome Res* 15 (2005) 1265-73.
- Deininger, P.L. and Batzer, M.A.: Alu repeats and human disease. *Mol Genet Metab* 67 (1999) 183-93.
- Ding, Y., Wang, X., Su, L., Zhai, J., Cao, S., Zhang, D., Liu, C., Bi, Y., Qian, Q., Cheng, Z., Chu, C. and Cao, X.: SDG714, a histone H3K9 methyltransferase, is involved in Tos17 DNA methylation and transposition in rice. *Plant Cell* 19 (2007) 9-22.
- Dobigny, G., Ozouf-Costaz, C., Waters, P.D., Bonillo, C., Coutanceau, J.P. and Volobouev, V.: LINE-1 amplification accompanies explosive genome repatterning in rodents. *Chromosome Res* 12 (2004) 787-93.
- Dowsett, A.P. and Young, M.W.: Differing levels of dispersed repetitive DNA among closely related species of *Drosophila*. *Proc Natl Acad Sci U S A* 79 (1982) 4570-4.
- Dramard, X., Heidmann, T. and Jensen, S.: Natural epigenetic protection against the I-factor, a *Drosophila* LINE retrotransposon, by remnants of ancestral invasions. *PLoS One* 2 (2007) e304.
- Ebert, A., Lein, S., Schotta, G. and Reuter, G.: Histone modification and the control of heterochromatic gene silencing in *Drosophila*. *Chromosome Res* 14 (2006) 377-92.
- Ebina, H. and Levin, H.L.: Stress management: how cells take control of their transposons. *Mol Cell* 27 (2007) 180-1.
- El-Sawy, M., Kale, S.P., Dugan, C., Nguyen, T.Q., Belancio, V., Bruch, H., Roy-Engel, A.M. and Deininger, P.L.: Nickel stimulates L1 retrotransposition by a post-transcriptional mechanism. *J Mol Biol* 354 (2005) 246-57.
- Esnault, C., Boulesteix, M., Duchemin, J.B., Koffi, A.A., Chandre, F., Dabire, R., Robert, V., Simard, F., Tripet, F., Donnelly, M.J., Fontenille, D. and Biemont, C.: High genetic differentiation between the M and S molecular forms of *Anopheles gambiae* in Africa. *PLoS One* 3 (2008) e1968.
- Fablet, M., Lerat, E., Rebollo, R., Horard, B., Burlet, N., Martinez, S., Brasset, E., Gilson, E., Vaury, C. and Vieira, C.: Genomic environment influences the dynamics of the tirant LTR retrotransposon in *Drosophila*. *Faseb J* 23 (2009) 1482-9.

- Fablet, M., McDonald, J.F., Biemont, C. and Vieira, C.: Ongoing loss of the tirant transposable element in natural populations of *Drosophila simulans*. *Gene* 375 (2006) 54-62.
- Fablet, M., Rebollo, R., Biemont, C. and Vieira, C.: The evolution of retrotransposon regulatory regions and its consequences on the *Drosophila melanogaster* and *Homo sapiens* host genomes. *Gene* 390 (2007) 84-91.
- Feschotte, C.: Transposable elements and the evolution of regulatory networks. *Nat Rev Genet* 9 (2008) 397-405.
- Gasior, S.L., Roy-Engel, A.M. and Deininger, P.L.: ERCC1/XPF limits L1 retrotransposition. *DNA Repair (Amst)* 7 (2008) 983-9.
- Gibert, J.M., Peronnet, F. and Schlotterer, C.: Phenotypic plasticity in *Drosophila* pigmentation caused by temperature sensitivity of a chromatin regulator network. *PLoS Genet* 3 (2007) e30.
- Girton, J.R. and Johansen, K.M.: Chromatin structure and the regulation of gene expression: the lessons of PEV in *Drosophila*. *Adv Genet* 61 (2008) 1-43.
- Granzotto, A., Lopes, F.R., Lerat, E., Vieira, C. and Carareto, C.M.: The evolutionary dynamics of the Helena retrotransposon revealed by sequenced *Drosophila* genomes. *BMC Evol Biol* 9 (2009) 174.
- Gregory, T.R.: Synergy between sequence and size in large-scale genomics. *Nat Rev Genet* 6 (2005) 699-708.
- Griffiths, A.J.F., Gelbart, W.M., Miller, J.H., Lewontin, R.C.: *Modern Genetic Analysis*. W. H. Freeman and Company, New York, 1999.
- Griffiths, A.J.F., Miller, J.H., Suzuki, D.T., Lewontin, R.C., Gelbart, W.M.: *An introduction to genetic analysis*. W. H. Freeman and Company, New York, 2000.
- Han, J.S. and Boeke, J.D.: LINE-1 retrotransposons: modulators of quantity and quality of mammalian gene expression? *Bioessays* 27 (2005) 775-84.
- Hashida, S.N., Uchiyama, T., Martin, C., Kishima, Y., Sano, Y. and Mikami, T.: The temperature-dependent change in methylation of the Antirrhinum transposon Tam3 is controlled by the activity of its transposase. *Plant Cell* 18 (2006) 104-18.
- Hasler, J., Samuelsson, T. and Strub, K.: Useful 'junk': Alu RNAs in the human transcriptome. *Cell Mol Life Sci* 64 (2007) 1793-800.
- Hawkins, J.S., Kim, H., Nason, J.D., Wing, R.A. and Wendel, J.F.: Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *Gossypium*. *Genome Res* 16 (2006) 1252-61.

- Hedges, D.J. and Deininger, P.L.: Inviting instability: Transposable elements, double-strand breaks, and the maintenance of genome integrity. *Mutat Res* 616 (2007) 46-59.
- Huang, X., Lu, G., Zhao, Q., Liu, X. and Han, B.: Genome-wide analysis of transposon insertion polymorphisms reveals intraspecific variation in cultivated rice. *Plant Physiol* 148 (2008) 25-40.
- Hulme, A.E., Bogerd, H.P., Cullen, B.R. and Moran, J.V.: Selective inhibition of Alu retrotransposition by APOBEC3G. *Gene* 390 (2007) 199-205.
- Ito, T.: Role of histone modification in chromatin dynamics. *J Biochem* 141 (2007) 609-14.
- Jaenisch, R. and Bird, A.: Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat Genet* 33 Suppl (2003) 245-54.
- Jakobsson, M., Hagenblad, J., Tavare, S., Sall, T., Hallden, C., Lind-Hallden, C. and Nordborg, M.: A unique recent origin of the allotetraploid species *Arabidopsis suecica*: Evidence from nuclear DNA markers. *Mol Biol Evol* 23 (2006) 1217-31.
- Johannes, F., Colot, V. and Jansen, R.C.: Epigenome dynamics: a quantitative genetics perspective. *Nat Rev Genet* 9 (2008) 883-90.
- Josefsson, C., Dilkes, B. and Comai, L.: Parent-dependent loss of gene silencing during interspecies hybridization. *Curr Biol* 16 (2006) 1322-8.
- Kashkush, K., Feldman, M. and Levy, A.A.: Gene loss, silencing and activation in a newly synthesized wheat allotetraploid. *Genetics* 160 (2002) 1651-9.
- Khan, H., Smit, A. and Boissinot, S.: Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. *Genome Res* 16 (2006) 78-87.
- Kim, J.K., Samaranyake, M. and Pradhan, S.: Epigenetic mechanisms in mammals. *Cell Mol Life Sci* 66 (2009) 596-612.
- Klenov, M.S., Lavrov, S.A., Stolyarenko, A.D., Ryazansky, S.S., Aravin, A.A., Tuschl, T. and Gvozdev, V.A.: Repeat-associated siRNAs cause chromatin silencing of retrotransposons in the *Drosophila melanogaster* germline. *Nucleic Acids Res* 35 (2007) 5430-8.
- Kondo, Y. and Issa, J.P.: Enrichment for histone H3 lysine 9 methylation at Alu repeats in human cells. *J Biol Chem* 278 (2003) 27658-62.
- Labrador, M., Farre, M., Utzet, F. and Fontdevila, A.: Interspecific hybridization increases transposition rates of *Osvaldo*. *Mol Biol Evol* 16 (1999) 931-7.

- Lachaise, D. and Silvain, J.F.: How two Afrotropical endemics made two cosmopolitan human commensals: the *Drosophila melanogaster*-*D. simulans* palaeogeographic riddle. *Genetica* 120 (2004) 17-39.
- Lamb, J.C., Yu, W., Han, F. and Birchler, J.A.: Plant chromosomes from end to end: telomeres, heterochromatin and centromeres. *Curr Opin Plant Biol* 10 (2007) 116-22.
- Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., et al.: Initial sequencing and analysis of the human genome. *Nature* 409 (2001) 860-921.
- Lau, N.C., Robine, N., Martin, R., Chung, W.J., Niki, Y., Berezikov, E. and Lai, E.C.: Abundant primary piRNAs, endo-siRNAs, and microRNAs in a *Drosophila* ovary cell line. *Genome Res* (2009).
- Lavrov, S., Dejardin, J. and Cavalli, G.: Combined immunostaining and FISH analysis of polytene chromosomes. *Methods Mol Biol* 247 (2004) 289-303.
- Lee, Y.N. and Bieniasz, P.D.: Reconstitution of an infectious human endogenous retrovirus. *PLoS Pathog* 3 (2007) e10.
- Lerat, E., Rizzon, C. and Biemont, C.: Sequence divergence within transposable element families in the *Drosophila melanogaster* genome. *Genome Res* 13 (2003) 1889-96.
- Lippman, Z. and Martienssen, R.: The role of RNA interference in heterochromatic silencing. *Nature* 431 (2004) 364-70.

- Lisch, D.: Epigenetic regulation of transposable elements in plants. *Annu Rev Plant Biol* 60 (2009) 43-66.
- Liu, B. and Wendel, J.F.: Epigenetic phenomena and the evolution of plant allopolyploids. *Mol Phylogenet Evol* 29 (2003) 365-79.
- Long, L., Ou, X., Liu, J., Lin, X., Sheng, L. and Liu, B.: The spaceflight environment can induce transpositional activation of multiple endogenous transposable elements in a genotype-dependent manner in rice. *J Plant Physiol* (2009).
- Lorenzi, H., Thiagarajan, M., Haas, B., Wortman, J., Hall, N. and Caler, E.: Genome wide survey, discovery and evolution of repetitive elements in three *Entamoeba* species. *BMC Genomics* 9 (2008) 595.
- Ma, J., Devos, K.M. and Bennetzen, J.L.: Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. *Genome Res* 14 (2004) 860-9.
- Malone, C.D., Brennecke, J., Dus, M., Stark, A., McCombie, W.R., Sachidanandam, R. and Hannon, G.J.: Specialized piRNA pathways act in germline and somatic tissues of the *Drosophila* ovary. *Cell* 137 (2009) 522-35.
- Marino-Ramirez, L., Lewis, K.C., Landsman, D. and Jordan, I.K.: Transposable elements donate lineage-specific regulatory sequences to host genomes. *Cytogenet Genome Res* 110 (2005) 333-41.
- Matyunina, L.V., Bowen, N.J. and McDonald, J.F.: LTR retrotransposons and the evolution of dosage compensation in *Drosophila*. *BMC Mol Biol* 9 (2008) 55.
- Matzke, M., Kanno, T., Huettel, B., Daxinger, L. and Matzke, A.J.: Targets of RNA-directed DNA methylation. *Curr Opin Plant Biol* 10 (2007) 512-9.
- Mc, C.B.: The origin and behavior of mutable loci in maize. *Proc Natl Acad Sci U S A* 36 (1950) 344-55.
- McClintock, B.: Induction of Instability at Selected Loci in Maize. *Genetics* 38 (1953) 579-99.
- McClintock, B.: The significance of responses of the genome to challenge. *Science* 226 (1984) 792-801.
- Messing, J., Bharti, A.K., Karlowski, W.M., Gundlach, H., Kim, H.R., Yu, Y., Wei, F., Fuks, G., Soderlund, C.A., Mayer, K.F. and Wing, R.A.: Sequence composition and genome organization of maize. *Proc Natl Acad Sci U S A* 101 (2004) 14349-54.
- Metcalfe, C.J., Bulazel, K.V., Ferreri, G.C., Schroeder-Reiter, E., Wanner, G., Rens, W., Obergfell, C., Eldridge, M.D. and O'Neill, R.J.: Genomic instability within centromeres of interspecific marsupial hybrids. *Genetics* 177 (2007) 2507-17.

- Meyerowitz, E.M. and Hogness, D.S.: Molecular organization of a *Drosophila* puff site that responds to ecdysone. *Cell* 28 (1982) 165-76.
- Michalak, P.: Epigenetic, transposon and small RNA determinants of hybrid dysfunctions. *Heredity* 102 (2009) 45-50.
- Mito, Y., Henikoff, J.G. and Henikoff, S.: Genome-scale profiling of histone H3.3 replacement patterns. *Nat Genet* 37 (2005) 1090-7.
- Morgan, T.H.: Sex Limited Inheritance in *Drosophila*. *Science* 32 (1910) 120-122.
- Mugnier, N., Biemont, C. and Vieira, C.: New regulatory regions of *Drosophila* 412 retrotransposable element generated by recombination. *Mol Biol Evol* 22 (2005) 747-57.
- Muotri, A.R., Marchetto, M.C., Coufal, N.G. and Gage, F.H.: The necessary junk: new functions for transposable elements. *Hum Mol Genet* 16 Spec No. 2 (2007) R159-67.
- Nolte, V. and Schlotterer, C.: African *Drosophila melanogaster* and *D. simulans* populations have similar levels of sequence variability, suggesting comparable effective population sizes. *Genetics* 178 (2008) 405-12.
- O'Neill, R.J., O'Neill, M.J. and Graves, J.A.: Undermethylation associated with retroelement activation and chromosome remodelling in an interspecific mammalian hybrid. *Nature* 393 (1998) 68-72.
- Obbard, D.J., Gordon, K.H., Buck, A.H. and Jiggins, F.M.: The evolution of RNAi as a defence against viruses and transposable elements. *Philos Trans R Soc Lond B Biol Sci* 364 (2009) 99-115.
- Ohno, S.: So much "junk DNA" in our genome. *Evolution of genetic systems* 23 (1972) 366-370.
- Pascale, E., Valle, E. and Furano, A.V.: Amplification of an ancestral mammalian L1 family of long interspersed repeated DNA occurred just before the murine radiation. *Proc Natl Acad Sci U S A* 87 (1990) 9481-5.
- Petrov, D.A. and Hartl, D.L.: High rate of DNA loss in the *Drosophila melanogaster* and *Drosophila virilis* species groups. *Mol Biol Evol* 15 (1998) 293-302.
- Phalke, S., Nickel, O., Walluscheck, D., Hortig, F., Onorati, M.C. and Reuter, G.: Retrotransposon silencing and telomere integrity in somatic cells of *Drosophila* depends on the cytosine-5 methyltransferase DNMT2. *Nat Genet* 41 (2009) 696-702.
- Piriyapongsa, J., Marino-Ramirez, L. and Jordan, I.K.: Origin and evolution of human microRNAs from transposable elements. *Genetics* 176 (2007) 1323-37.

- Pogribny, I.P., Tryndyak, V.P., Woods, C.G., Witt, S.E. and Rusyn, I.: Epigenetic effects of the continuous exposure to peroxisome proliferator WY-14,643 in mouse liver are dependent upon peroxisome proliferator activated receptor alpha. *Mutat Res* 625 (2007) 62-71.
- Polavarapu, N., Marino-Ramirez, L., Landsman, D., McDonald, J.F. and Jordan, I.K.: Evolutionary rates and patterns for human transcription factor binding sites derived from repetitive DNA. *BMC Genomics* 9 (2008) 226.
- Rangwala, S.H., Elumalai, R., Vanier, C., Ozkan, H., Galbraith, D.W. and Richards, E.J.: Meiotically stable natural epialleles of Sadhu, a novel Arabidopsis retroposon. *PLoS Genet* 2 (2006) e36.
- Ray, D.A., Feschotte, C., Pagan, H.J., Smith, J.D., Pritham, E.J., Arensburger, P., Atkinson, P.W. and Craig, N.L.: Multiple waves of recent DNA transposon activity in the bat, *Myotis lucifugus*. *Genome Res* 18 (2008) 717-28.
- Rebollo, R., Lerat, E., Kleine, L.L., Biemont, C. and Vieira, C.: Losing helena: the extinction of a drosophila line-like element. *BMC Genomics* 9 (2008) 149.
- Richards, E.J.: Inherited epigenetic variation--revisiting soft inheritance. *Nat Rev Genet* 7 (2006) 395-401.
- Richards, E.J.: Population epigenetics. *Curr Opin Genet Dev* 18 (2008) 221-6.
- Riddle, N.C., Shaffer, C.D. and Elgin, S.C.: A lot about a little dot - lessons learned from *Drosophila melanogaster* chromosome 4. *Biochem Cell Biol* 87 (2009) 229-41.
- Sandmann, T., Jakobsen, J.S. and Furlong, E.E.: ChIP-on-chip protocol for genome-wide analysis of transcription factor binding in *Drosophila melanogaster* embryos. *Nat Protoc* 1 (2006) 2839-55.
- Schulze, S.R. and Wallrath, L.L.: Gene regulation by chromatin structure: paradigms established in *Drosophila melanogaster*. *Annu Rev Entomol* 52 (2007) 171-92.
- Shirasu, K., Schulman, A.H., Lahaye, T. and Schulze-Lefert, P.: A contiguous 66-kb barley DNA sequence provides evidence for reversible genome expansion. *Genome Res* 10 (2000) 908-15.
- Sinzelle, L., Izsvak, Z. and Ivics, Z.: Molecular domestication of transposable elements: from detrimental parasites to useful host genes. *Cell Mol Life Sci* 66 (2009) 1073-93.
- Stribinskis, V. and Ramos, K.S.: Activation of human long interspersed nuclear element 1 retrotransposition by benzo(a)pyrene, an ubiquitous environmental carcinogen. *Cancer Res* 66 (2006) 2616-20.

- Tanurdzic, M., Vaughn, M.W., Jiang, H., Lee, T.J., Slotkin, R.K., Sosinski, B., Thompson, W.F., Doerge, R.W. and Martienssen, R.A.: Epigenomic consequences of immortalized plant cell suspension culture. *PLoS Biol* 6 (2008) 2880-95.
- Tchurikov, N.A. and Kretova, O.V.: Suffix-specific RNAi leads to silencing of F element in *Drosophila melanogaster*. *PLoS One* 2 (2007) e476.
- Trojer, P. and Reinberg, D.: Facultative heterochromatin: is there a distinctive molecular signature? *Mol Cell* 28 (2007) 1-13.
- Ungerer, M.C., Baird, S.J., Pan, J. and Rieseberg, L.H.: Rapid hybrid speciation in wild sunflowers. *Proc Natl Acad Sci U S A* 95 (1998) 11757-62.
- Ungerer, M.C., Strakosh, S.C. and Zhen, Y.: Genome expansion in three hybrid sunflower species is associated with retrotransposon proliferation. *Curr Biol* 16 (2006) R872-3.
- van de Lagemaat, L.N., Gagnier, L., Medstrand, P. and Mager, D.L.: Genomic deletions and precise removal of transposable elements mediated by short identical DNA segments in primates. *Genome Res* 15 (2005) 1243-9.
- Verneau, O., Catzeflis, F. and Furano, A.V.: Determining and dating recent rodent speciation events by using L1 (LINE-1) retrotransposons. *Proc Natl Acad Sci U S A* 95 (1998) 11284-9.
- Vieira, C., Fablet, M. and Lerat, E.: Infra- and Transspecific Clues to Understanding the Dynamics of Transposable Elements. In: Schwarz, S. (Ed.), *Transposons and the Dynamic Genome* Springer Verlag, 2009.
- Vieira, C., Lepetit, D., Dumont, S. and Biemont, C.: Wake up of transposable elements following *Drosophila simulans* worldwide colonization. *Mol Biol Evol* 16 (1999) 1251-5.
- Vitte, C. and Panaud, O.: Formation of solo-LTRs through unequal homologous recombination counterbalances amplifications of LTR retrotransposons in rice *Oryza sativa* L. *Mol Biol Evol* 20 (2003) 528-40.
- Volff, J.N., Korting, C., Meyer, A. and Schartl, M.: Evolution and discontinuous distribution of Rex3 retrotransposons in fish. *Mol Biol Evol* 18 (2001) 427-31.
- Waterland, R.A. and Jirtle, R.L.: Transposable elements: targets for early nutritional effects on epigenetic gene regulation. *Mol Cell Biol* 23 (2003) 5293-300.
- Waterland, R.A. and Jirtle, R.L.: Early nutrition, epigenetic changes at transposons and imprinted genes, and enhanced susceptibility to adult chronic diseases. *Nutrition* 20 (2004) 63-8.

- Watson, J.D. and Crick, F.H.: Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* 171 (1953) 737-8.
- Weil, C.F.: Too many ends: aberrant transposition. *Genes Dev* 23 (2009) 1032-6.
- Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P. and Schulman, A.H.: A unified classification system for eukaryotic transposable elements. *Nat Rev Genet* 8 (2007) 973-82.
- Wong, L.H. and Choo, K.H.: Evolutionary dynamics of transposable elements at the centromere. *Trends Genet* 20 (2004) 611-6.
- Yuki, S., Inouye, S., Ishimaru, S. and Saigo, K.: Nucleotide sequence characterization of a *Drosophila* retrotransposon, 412. *Eur J Biochem* 158 (1986) 403-10.
- Zhang, J., Yu, C., Pulletikurti, V., Lamb, J., Danilova, T., Weber, D.F., Birchler, J. and Peterson, T.: Alternative Ac/Ds transposition induces major chromosomal rearrangements in maize. *Genes Dev* 23 (2009) 755-65.

---

Les éléments transposables (ET) sont une source majeure de variation génétique, ce qui leur confère un rôle essentiel dans l'évolution des génomes. Certes présents dans tous les génomes analysés à ce jour, leurs proportions sont fortement variables entre espèces et aussi entre populations, suggérant une relation unique entre génome hôte et ET. Grâce à un système modèle composé de populations naturelles de deux espèces proches (*Drosophila melanogaster* et *D. simulans*) avec des quantités différentes en ET, nous avons pu comparer les relations génome hôte/ET. Nous avons pu montrer que les ET sont contrôlés par le génome hôte par des délétions internes et probablement par un système épigénétique variable. De plus, dans certaines populations analysées, des copies peuvent échapper à ce contrôle et envahir le génome. Les ET sont donc des grands créateurs de variabilité génétique mais permettent aussi une territorialisation chromatinienne du génome car ils portent des modifications épigénétiques précises et sont capables de les étendre à leurs environnements génomiques. Ceci leur confère la fonction "d'épigénétique mobile".

---

TE variation in natural populations of *Drosophila* : copy number, transcription and chromatin state

---

Transposable elements (TEs) are one major force of genome evolution thanks to their ability to create genetic variation. TEs are ubiquitous and their proportion is variable between species and also populations, suggesting that a tight relationship exists between genomes and TEs. The model system composed of the natural populations of the twin sisters *Drosophila melanogaster* and *D. simulans* is interesting to compare host/TE relationship, since both species harbour different amounts of TE copies. The *helena* element is nearly silenced in *D. simulans* natural populations despite a very high copy number. Such repression is associated to abundant internally deleted copies suggesting a regulatory mechanism of TEs based on DNA deletion. Another pathway of TE regulation is through epigenetics where the host genome is able to keep intact the DNA sequences of TEs and still silence their activities. Chromatin remodelling is well known in *drosophila* and specific histone modifications can be associated to specific chromatin domains. We observed an important variation on H3K27me3 and H3K9me2, two heterochromatic marks, on TE copies in *D. melanogaster* and *D. simulans* natural populations. Also, we show that derepressed lines of *D. simulans* exist for specific elements, have high TE transcription rates and are highly associated to non constitutive heterochromatic marks. TEs are therefore controlled by the host genome through DNA deletion and a possible chromatin remodelling mechanism. Not only genetic variability is enhanced by TEs but also epigenetic variability, allowing the host genome to be partitioned into chromatin domains. TEs are therefore mandatory to gene network regulation through their ability of "jumping epigenetics".

---

Génétique évolutive

---

MOTS-CLES

Éléments transposables, *Drosophila*, Populations naturelles, Délétion, Epigénétique

---

LBBE UMR CNRS 5558

43 Bd du 11 novembre 1918 - 69622 cedex Villeurbanne