Un nouveau swing pour flamenco : Caractérisation du locus flamenco, un gène non codant régulateur des éléments transposables par ARN interférence dans les tissus reproducteurs de Drosophila melanogaster

Coline Goriaux

UNIVERSITE BLAISE PASCAL UNIVERSITE D'AUVERGNE

Année 2014 N° d'ordre :

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

Thèse

Présentée à l'Université d'Auvergne

pour l'obtention du grade de DOCTEUR
(Décret du 5 juillet 1984)

Spécialité : Génétique

soutenue le

21 octobre 2014

Goriaux Coline

Un nouveau swing pour *flamenco* : Caractérisation du locus *flamenco*, un gène non codant régulateur des éléments transposables par ARN interférence dans les tissus reproducteurs de *Drosophila melanogaster*.

Président : Dr Vincent Sapin

Rapporteur : Dr Stéphane Ronsseray
Dr Constance Ciaudo

Examinateur : Dr Thierry Forné

Directeur de thèse : Dr Chantal Vaury
Co-directeur : Dr Emilie Brasset

« Instabilités génétiques et contrôle par le génome de l'hôte »
UMR -CNRS 6293 Clermont-Université /INSERM U1103, Clermont Ferrand

Résumé

Ces dernières années, de nombreuses études transcriptomiques à grande échelle ont clairement mis en évidence que la grande majorité du génome des eucaryotes est transcrite. Ce réseau complexe de transcrits inclus des petits ARN non codants qui interviennent généralement en tant que régulateurs transcriptionnels, post-transcriptionnels et/ou traductionnels de l'expression de certains ARNm cibles spécifiques. Ils sont classés selon leur origine biologique et leur mode d'action. Une catégorie de petits ARN non codants, les Piwi-interacting RNAs (piRNA), maintient l'intégrité du génome dans les tissus reproducteurs des métazoaires en réprimant les éléments transposables endogènes, des séquences ADN capables de se déplacer et de se dupliquer à l'intérieur du génome. Les piRNA sont produits par deux mécanismes : i) La biogenèse primaire à partir de longs ARN simple brin produits par certains loci spécifiques du génome, les clusters de piRNA, des loci énigmatiques, localisés dans les régions hétérochromatiques et composés de fragments d'éléments transposables actifs, ii) La boucle d'amplification appelée ping-pong.

Durant ma thèse, j'ai étudié un cluster de piRNA majeurs dans les cellules somatiques des gonades femelles de *Drosophila melanogaster*, le locus *flamenco*. Tout d'abord, j'ai mis en évidence que la transcription de *flamenco* est initiée à partir d'un promoteur contenant une séquence INR et un élément DPE, reconnu par l'ARN polymerase II, et qu'elle nécessite la présence du facteur de transcription *Cubitus Interruptus*. Ensuite, j'ai montré que le transcrit de *flamenco* subit de l'épissage alternatif pour générer divers précurseurs ARN qui seront ensuite maturés en piRNA. De plus, j'ai montré que le promoteur de *flamenco* serait suffisant pour déclencher l'adressage du transcrit vers la voie de maturation des piRNA. Dans un autre axe, je me suis intéressée à l'organisation tridimensionnelle du locus *flamenco* au sein du noyau en recherchant ses partenaires d'interaction en utilisant la technique de 4C (capture de la conformation des chromosomes). J'ai pu voir que *flamenco* semble interagir physiquement avec des régions génomiques fortement transcrites en cis. En trans, le locus *flamenco* interagit majoritairement avec des régions génomiques péricentromériques et avec d'autres clusters de piRNA. Cette disposition tridimensionnelle particulière pourrait être le reflet d'une organisation fonctionnelle.

Dans l'ensemble, ces travaux permettent de mieux comprendre l'expression et le fonctionnement du locus *flamenco* et ouvrent la voie vers de nouvelles recherches prometteuses.

Abstract

The past few years it has become clear from many transcriptomic studies that most of the eukaryotic genome is pervasively transcribed. This complex network of transcripts include several types of small RNAs classified as non-coding RNAs. The vast majority of small RNA act as transcriptional, posttranscriptional and/or translational regulators, controlling specific target mRNAs involved in various cellular functions. They are classified based on their biogenesis and mode of action.

A subclass of small non-coding RNAs, the Piwi-interacting RNAs (piRNA), ensures genomic stability by silencing endogenous transposable elements, endogenous sequence that are able to move and duplicate into the genome, in both germline and somatic gonadal tissues of metazoan. piRNA are produced through two mechanisms, i) The primary processing pathway from long single-stranded precursors produced by some specific loci in the genome, the piRNA clusters, ii) The secondary pathway by the amplification loop called the ping-pong. piRNA clusters are enigmatic loci localized in heterochromatic region and composed of transposable element fragments.

During my PhD, I studied a major piRNA cluster in the somatic cells of *Drosophila melanogaster* female gonads, the *flamenco* locus. First, I demonstrated that *flamenco* transcription is initiated from an RNA Polymerase II promoter containing Inr and DPE elements, and requires the transcription factor, Cubitus interruptus. Then, I showed that the *flamenco* precursor transcript undergoes differential alternative splicing to generate diverse RNA precursors that are processed into piRNA. Moreover, I showed that the *flamenco* promoter could be sufficient to target transcripts into the piRNA processing pathway. In an other hand, I was interested to the tridimensional nuclear organization of the *flamenco* locus using the 4c technology. I saw that the *flamenco* locus interacts physically with strongly transcribed genomic region in cis. In trans, the *flamenco* interacts with other peri-centromeric genomic regions and with two other piRNA cluster. This particularly three-dimensional positioning could be the reflect of a functional organization.

In the main, this work allows to better understand the expression and the mode of action of the *flamenco* locus and pave the way for new promising research.

Remerciements

En premier lieu, je souhaite remercier ceux qui ont accepté d'évaluer ce travail de thèse : Dr Constance Ciaudo, Dr Stéphane Ronsseray, Dr Thierry Forné et Dr Vincent Sapin. Je vous remercie également de vous être déplacés jusqu'à Clermont-Ferrand pour cette occasion.

Je remercie chaleureusement Chantal Vaury de m'avoir accepté au sein de l'équipe « Instabilités génétiques et contrôle par le génome de l'hôte » dans laquelle j'ai passée trois superbes années. En plus d'être une chef d'équipe remarquable, tu diriges le GReD avec une énergie communicative incroyable. Je te souhaite que tes efforts soient récompensés par une longue vie du GReD pleine de découvertes et de publications !

Merci beaucoup Emilie pour m'avoir guidé tout au long de cette thèse. Tu as réussi à m'accorder suffisamment de confiance pour me laisser la liberté dont j'avais besoin tout en m'apportant toute l'aide qu'un doctorant peut espérer de son chef ! J'ai vraiment pu m'épanouir à fond et apprendre un maximum de chose. Je te remercie pour la souplesse dont tu as fait part pendant ces trois ans, notamment pour les corrections à des heures indues pour des présentations le lendemain ou autres… Avec ton énergie et tes idées bouillonnantes, nul doute que l'équipe sera très productive à l'avenir ! En tout cas, j'espère avoir été à la hauteur de tes attentes en tant que ta première doctorante !

Un grand merci à Nathalie ! Pour ton aide, tes conseils, ton soutien, ton enthousiasme quotidien, …! Ta présence dans l'équipe est une chance incroyable et je suis confiante pour tous les doctorants à venir car je sais que tu seras là pour les aider.

Je remercie également Rana, Manue, Cynthia et Silke pour le formidable esprit d'équipe et d'entraide qui règne au laboratoire.

En plus de ces remerciements très professionnels je tiens vraiment à remercier une nouvelle fois toute mon équipe pour leur qualités humaines incroyables. C'était vraiment un plaisir de travailler avec vous !

Un grand, grand merci à mes trois sauveurs… nos bio-informaticiens Pierre, Yoan et Romain que j'ai tour à tour sollicité pour des demandes plus ou moins intelligentes, en tout cas toujours désespérées… ! J'ai appris beaucoup de chose grâce à vous (et même à utiliser

des lignes de commande, la classe !) sur ces trucs bizarres que sont les ordinateurs et qui ne sont peut-être pas si nuls que ça finalement.

Je remercie tous les drosophilistes de la fac de médecine avec qui j'ai partagé ma paillasse durant ces trois ans. Merci pour tous vos conseils qui m'ont bien aidée et pour votre bonne humeur de tous les jours. Ca y est je m'en vais, plus besoin de planquer vos stylos !

Je remercie toute l'équipe des manges-tôt ! Je crois qu'on aura marqué la cafét avec nos débats enflammés qui, il faut bien l'avouer, ne valent pas toujours l'ardeur que l'on y met. En tout cas, ces pauses de midi étaient très appréciables pour se défouler entre deux manips !

Je remercie tout le personnel technique et administratif du GRe*D.* Vous faites vraiment du super boulot pour nous permettre de bosser dans de bonnes conditions, merci !

Je remercie toutes les personnes que j'ai pu rencontrer dans le milieu professionnel qui m'ont aidée et apportée leur soutien durant ces trois ans.

Merci à la compagnie pouet pouet pour tous ces supers moments de joie, de folie et de bonheur indispensable à mon équilibre psychique dans ce monde bien trop rationnel ! Et bien sur, je salue tout particulièrement Marion, et Ju. Franchement, rien que pour vous rencontrer, ça valait le coup de venir à Clermont-Ferrand.

Un grand merci à django django, zoufri maracas, systema solar, chinese man,… ainsi qu'aux djs K.*D.*S et N-u et aux radio nova et meuh pour leur soutien inconditionnel dans les bons jours et surtout dans les mauvais jours. Je remercie également Chafifou pour m'avoir accompagné, bon gré mal gré, pendant ses (trop) longues heures d'écriture.

Plus sérieusement, je remercie grandement ma famille pour leur soutien sans faille et leurs encouragements. Mes grands parents pour m'avoir proposée une halte pendant tous ces allers retours ! Vivre à Clermont nous a été l'opportunité de se voir plus souvent, c'était super ! Merci à Thomas et Jietka, je vous souhaite plein de bonheur pour votre nouvelle vie à Prague ! Et surtout un très grand merci à mes parents pour leur soutien de tous les jours. Thomas et moi sommes si chanceux de vous avoir comme parents. Vous nous avez appris tellement et surtout à être heureux et libres, ce qui n'a pas de prix.

Enfin, un très grand merci à Arthur qui m'a soutenue et aidée au quotidien. Merci de m'avoir accompagnée dans cette ville si éloignée de la mer ! Maintenant, je suis libre et avec

toi à mes cotés, le monde ressemble à immense un terrain de jeu que j’ai très envie d’explorer !

De façon générale, cette thèse est dédicacée à tous les chercheurs, scientifiques et savants fous avec qui mes rapports furent aussi divers qu’enrichissants!

Sommaire

Index des figures

pétunia

Introduction

I. Organisation du génome et régulation de l'expression génique

A. Caractérisation du génome

1. Emergence de l'ère de la génétique

Bien que témoin quotidien de nombreux phénomènes héréditaires, la société humaine possédait, au début du XIXème siècle, une notion encore très floue des mécanismes contrôlant la transmission des caractères d'un organisme à un autre, au cours des générations. Les découvertes successives de plusieurs scientifiques ont permis à la communauté scientifique de décrypter les bases fonctionnelles et structurelles de l'hérédité.

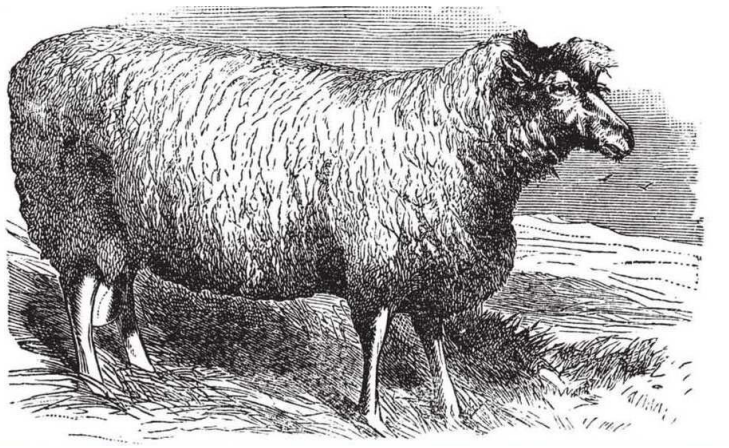
a) Découverte des lois fondamentales de l'hérédité

Afin de comprendre les principes de fonctionnement de l'hérédité, de nombreuses réflexions collectives étaient menées, notamment au nord-est de l'Europe où une société savante d'éleveurs de moutons a longtemps débattu sur l'hérédité des caractéristiques de la laine de mouton. Imre Festetics, un éleveur ovin membre de cette société, publia plusieurs papiers en 1819 dans lesquels il récapitule les conclusions obtenues sur la transmission des caractéristiques de la laine de mouton à partir de 15 ans d'observations et de croisements, les nommant « lois génétiques de la nature » (Poczai et al. 2014) **(Figure 1 A et B)**. Festetics avait observé plusieurs principes fondamentaux de l'hérédité sur l'importance de la sélection des parents et la variabilité intergénérationnelle des caractères transmis. Malheureusement pour lui, si la qualité de sa production laineuse fût reconnue, la complexité de son raisonnement scientifique, sur un caractère que l'on sait maintenant multigénique, empêcha la propagation de ses idées.

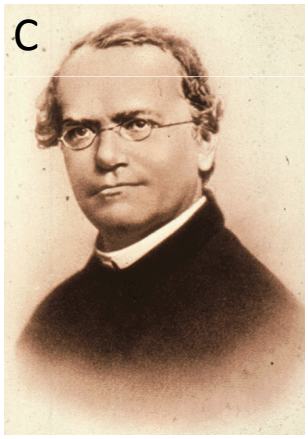
En 1866, Gregor Mendel, un moine tchèque du monastère de Brno, publia les conclusions de ses travaux réalisés sur le petit pois où il observa la transmission de divers caractères phénotypiques à travers plusieurs générations (MENDEL 1950) **(Figure 1 C et D**). Ses conclusions supposent qu'un caractère est déterminé par 2 composantes d'un même facteur au sein d'un individu : ce sont les deux allèles d'un gène. Mendel, aujourd'hui reconnu comme le père de la génétique, propose trois lois qui définissent la transmission de

A

B



C



D

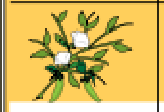
Graine		Fleur	Cosse		Tige	
Forme	Cotylédons	Couleur	Forme	Couleur	Emplacement	Taille
						
Gris & lisse	Jaune	Blanc	Plein	Jaune	Cosse axiale Fleur tout du long	Long (~3m)
						
Blanc & Ridé	Vert	Violet	Étroit	Vert	Cosse terminales Fleurs en haut	Court (~30 cm)
1	2	3	4	5	6	7

Figure 1 : Les pères fondateurs de la génétique et leurs modèles d'étude

A) Imre Festetics (1764-1847)

B) Gravure d'un bélier Dishley, espèce obtenue par Robert Bakewell dont les travaux influencèrent grandement Festetics.

C) Gregor Mendel (1822-1884)

D) Tableau récapitulatif des différents caractères phénotypiques du petit pois étudiés par Mendel

ces allèles : i) L'uniformité des hybrides de première génération ii) La disjonction des allèles lors de la méiose iii) La ségrégation indépendante des caractères héréditaires multiples.

Si certains se demandaient comment les caractères se transmettent de façon stable à travers les générations, d'autres s'intéressaient plutôt à l'acquisition de nouveaux caractères et à la naissance de nouvelles espèces. Les prémices de la paléontologie avaient fourni plusieurs exemples de fossiles appartenant à des espèces disparues, prouvant que les espèces n'étaient pas stables dans le temps. Jean-Baptiste de Lamarck, un naturaliste français, proposa une première théorie sur la transformation du vivant dans son livre *Philosophie zoologique,* publié en 1809. Dans cet ouvrage, il suppose que la vie tend à être toujours plus complexe de façon intrinsèque et que l'émergence des différents organes est due à une utilisation intensive en réponse à un stimulus environnemental **(Figure 2 A et B)**. De plus, il stipule que lorsque le stimulus a été maintenu suffisamment longtemps, la modification de l'organe sera transmise à la descendance. Sans qu'il n'en ait clairement établi les clauses, Lamarck sera considéré comme le précurseur de l'hérédité des caractères acquis.

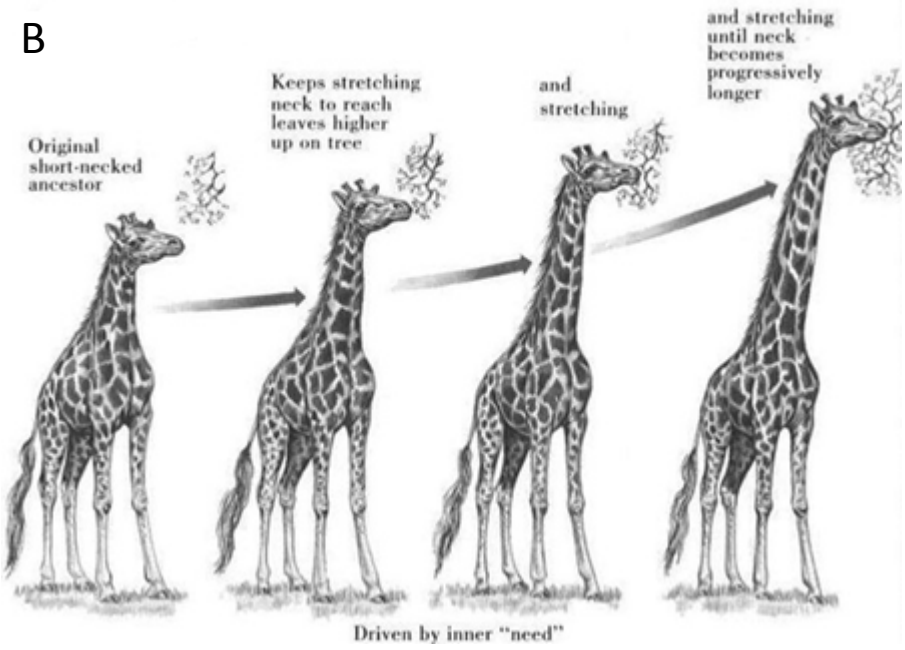
Charles Darwin, naturaliste anglais également convaincu de la transformation des espèces, proposa une nouvelle théorie dans son traité *L'origine des espèces* paru en 1859. Ses observations réalisées au cours de deux voyages dans les îles Galápagos l'amenèrent à penser que, à l'intérieur de la masse d'individus qui constitue une espèce, ceux qui dévient légèrement peuvent acquérir des capacités différentes d'adaptation à leur environnement **(Figure 2 C et D)**. Si ces différences peuvent être transmises à la descendance et si elles leurs confèrent un avantage reproductif, ces nouvelles caractéristiques seront plus fréquemment transmises à la descendance, ce qui crée une sorte de sélection des individus les mieux adaptés pouvant conduire à l'émergence de nouvelles espèces. Si Darwin ne proposa aucune théorie concernant l'origine de ces variations, ses idées provoqueront une grande polémique et sont maintenant reconnues comme la base de la théorie de l'évolution des espèces.

b) Découverte du support moléculaire de l'hérédité

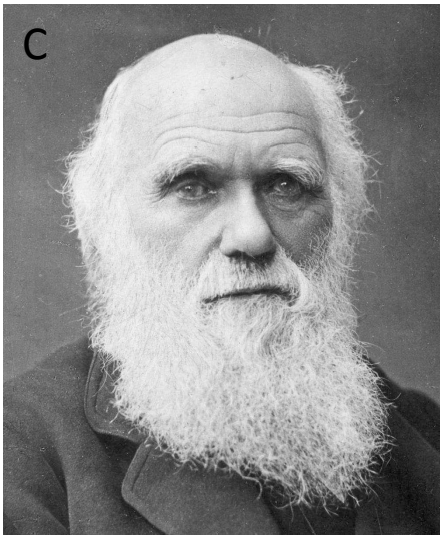
Si les principes de la transmission des caractères à la descendance avaient été établis, le support moléculaire de ces informations génétiques restait à définir. Thomas Hunt Morgan, zoologiste à l'université de Columbia, travaillait au début du XXème siècle sur la mouche du vinaigre, *Drosophila melanogaster*, ou drosophile. Ses résultats sur des mutations spontanées démontrèrent que les caractères phénotypiques ségrégent avec les chromosomes, des entités

A

B



C



D

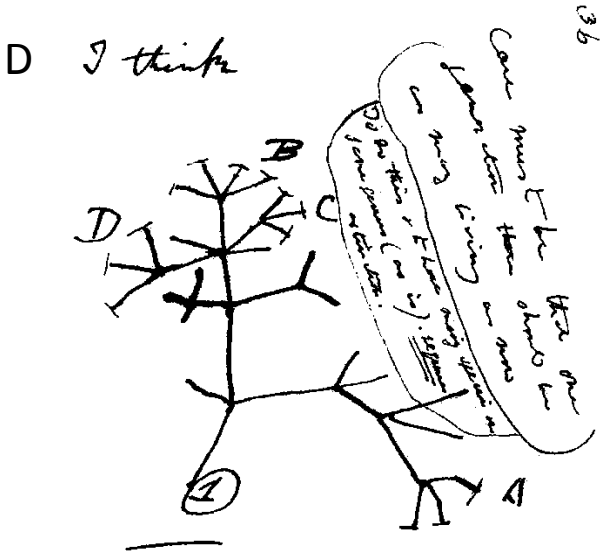


Figure 2 : Jean-Baptiste de Lamarck et Charles Darwin, deux théoriciens de l'évolution

A) Jean-Baptiste de Lamarck (1744-1829).

B) Illustration de la théorie évolutive de Lamarck exemplifiée par le cou de la girafe.

[http://www.bio.miami.edu/ecosummer/lectures/lec_evolution.html]

C) Charles Darwin (1809-1882).

D) Arbre phylogénétique extrait des carnets de Darwin.

[http://www.hominides.com/html/theories/theories-evolutionnisme.php]

observables au microscopique dans le noyau des cellules, après coloration (Morgan, 1911) **(Figure 3A)**. Il établit la théorie chromosomique de l'hérédité dans laquelle il proposa que les gènes soient des entités concrètes présentes sur les chromosomes. Il réalisa les premières cartes génétiques de la drosophile et observa que la ségrégation commune de certains caractères dépend de leur espacement le long des chromosomes, espacement mesuré en centimorgan (cM). Ses travaux furent reconnus par un prix Nobel de Physiologie ou Médecine en 1933 et ils ouvrirent la voie à de très nombreuses études sur la drosophile, faisant de cet organisme un des modèles biologiques les plus utilisés.

Les chromosomes étant constitués de plusieurs types de molécules (protéines et acides nucléiques), la question du porteur de l'hérédité n'était que partiellement résolue. Il fallut attendre 1944 pour que Oswald Avery apporte la réponse finale. Avery était un médecin américain qui poursuivait les travaux de Frederick Griffith sur la transformation bactérienne chez les pneumocoques. Griffith avait observé une transmission de caractère entre deux souches différentes de pneumocoques, sans connaître le facteur nécessaire à cette transformation. Avery reproduisit cette expérience et montra que l'acide désoxyribonucléique (ADN), une molécule identifiée en 1869 par Friedrich Miescher, est nécessaire et suffisant à cette transmission (Avery 1944).

L'ADN était alors connu pour être composé de sucres, de phosphates et de quatre bases azotées différentes (Adénine, Cytosine, Guanine et Thymine) mais la structure de cet ensemble était inconnue. La compréhension de l'enchevêtrement de ces différents constituants au sein de la molécule d'ADN fut possible grâce aux travaux de cristallographie aux rayons X réalisés par Rosalind Franklin, une biologiste travaillant au King's College de Londres. Les clichés ainsi obtenus furent interprétés par James Watson et Francis Crick, biologistes à l'université de Cambridge, comme une structure en double hélice en 1953 (WATSON & CRICK 1953). En faisant le lien avec les règles de Chargaff, qui stipule que dans l'ADN de n'importe quelle cellule la quantité de bases adénines est égale à la quantité de bases thymines ainsi que la quantité de bases cytosines est égale à la quantité de bases guanines, Watson et Crick en déduisirent que cette hélice est composée de deux brins d'ADN complémentaires, où chaque base d'un brin est associé à sa base complémentaire sur l'autre brin (une adénine en face d'une thymine et une cytosine en face d'une guanine) **(Figure 3B)**.

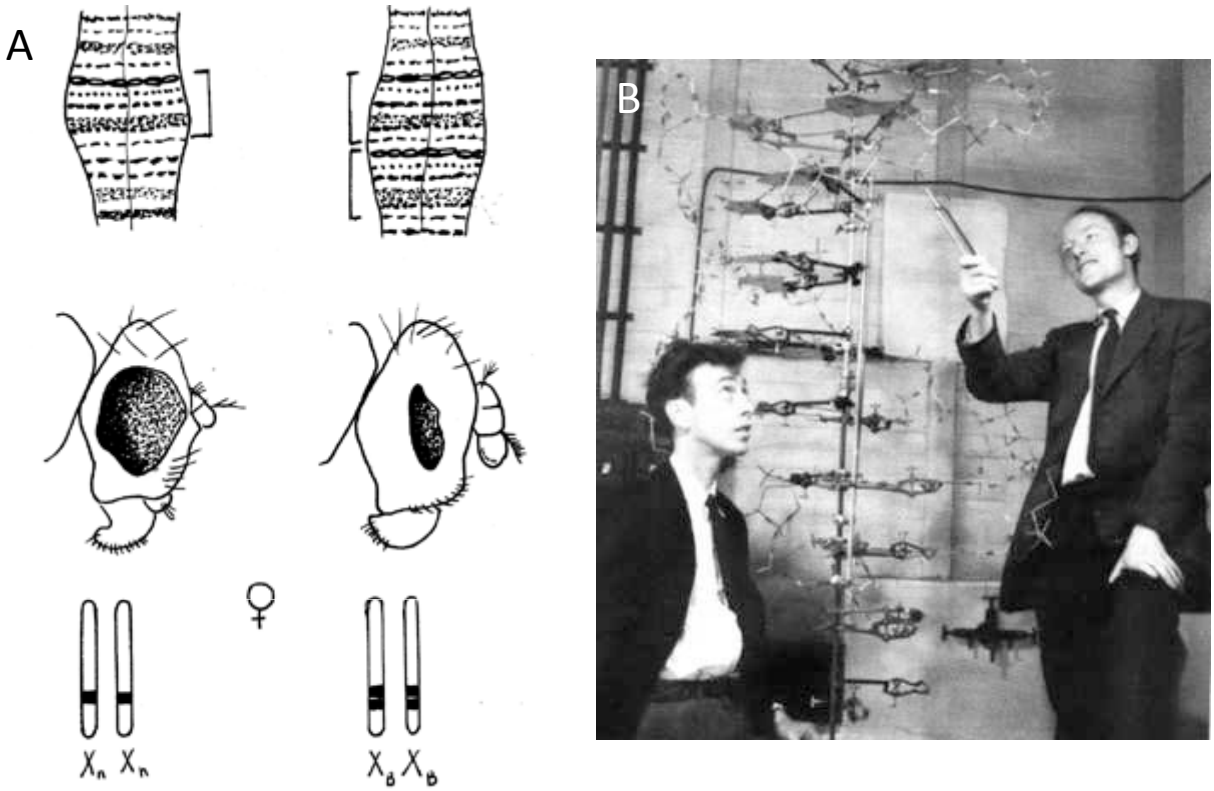


Figure 3 : L'ADN, support moléculaire de l'hérédité

A) Schéma de Thomas Hunt Morgan qui montre qu'une anomalie localisée sur un chromosome, en l'occurrence la duplication d'une zone, est responsable d'une malformation de l'œil.
[http://www.svtenligne.peda-go.fr/spip.php?article255]

B) Photo de James Watson et Francis Crick devant une maquette de la structure en double hélice de l'ADN.

c) L’ère de l’hérédité inclusive

Toutes ces études ont permis de construire un modèle où chaque organisme est défini par un ensemble de gènes, le génome, responsable de l’ensemble de ses caractéristiques phénotypiques. Ces gènes sont des entités indépendantes définies par une séquence ADN propre et portées par les chromosomes de chaque cellule de l’organisme. Ils seront transmis à la descendance par l’intermédiaire des cellules germinales, les gamètes, qui formeront la base d’un nouvel organisme. Le modèle de la sélection naturelle appliqué aux gènes semblait répondre aux questions de l’hérédité et de l’évolution des espèces. Pourtant, certaines observations vinrent remettre en question ce paradigme, notamment lors de la réalisation des études d’associations à échelles génomiques chez l’humain (GWAS) pour lesquelles de nombreuses études statistiques avaient été réalisées afin de retrouver les gènes associés à un phénotype particulier ou à une maladie. Même si de nombreuses variations génétiques ont pu être associées à un caractère phénotypique, l’influence de ces variations semblait être très limitée et loin d’être déterminante (Maher 2008). Les scientifiques ont donc supposé qu’il existe une hérédité non génétique, indépendante des gènes, que l’on appelle hérédité inclusive.

L’hérédité inclusive englobe toute information transmise à la descendance, que ce soit par l’intermédiaire des gènes (hérédité génétique) ou part d’autres facteurs (hérédité non génétique). Elle a tout d’abord trouvé un support dans l’épigénétique, qui correspond à la modification stable et héritable de l’expression des gènes sans changement de la séquence nucléotidique. L’épigénétique représente l’ensemble des marques physico-chimiques que portent l’ADN et ses protéines associées et qui influence la transcription des séquences géniques. Ces marques sont retrouvées chez tous les organismes et sont nécessaires au développement harmonieux de l’individu en permettant, entre autres, la différenciation cellulaire. De nombreuses études ont montré que ces marques peuvent être influencées par l’environnement, ce qui a relancé l’idée de la transmission des caractères acquis de Lamarck. Mais l’hérédité inclusive va encore au-delà de l’épigénétique et prend également en compte les mécanismes héréditaires dus aux effets parentaux et à l’hérédité écologique et culturelle (Danchin et al. 2011) **(Figure 4)**.

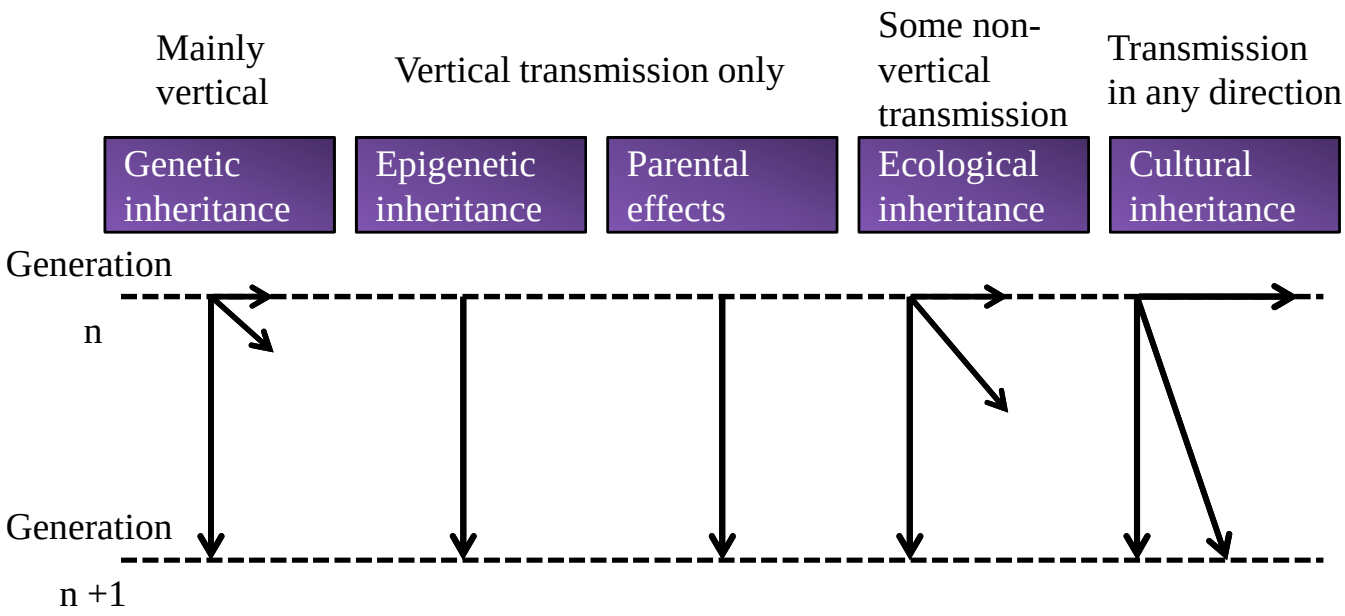


Figure 4 : Les différents types d'hérédité et leurs modes de transmission

Schéma représentant les différentes types d'hérédité (rectangle violet) et les modes de transmission associés (flèches noires). [Danchin *et al*. 2011]

2. Organisation du génome

a) Composition et structure moléculaire du génome

Le génome désigne la totalité de l'information génétique d'une espèce. Les premiers constituants du génome à avoir été découverts sont les gènes qui s'expriment en produisant des protéines, en passant par un intermédiaire sous forme d'Acide RiboNucléique ou ARN. Mis à part chez les procaryotes où la densité de gène est très élevée, les séquences codantes sont le plus souvent largement minoritaires, notamment chez les eucaryotes. Les séquences codantes au sens strict, les exons, ne représentent même que quelques pourcentages du génome total (<1,5% chez l'humain, (Venter et al. 2001)). Les séquences non codantes sont de plusieurs types : les introns, les pseudogènes, les séquences répétées et les séquences uniques non codantes **(Figure 5)**. Chez l'humain, les introns, des séquences de tailles très variables qui fragmentent les gènes et ne seront pas traduites en protéine, représentent autour de 25 % du génome. Les pseudogènes sont des séquences apparentées à des gènes connus mais qui ne seront pas à l'origine d'une protéine. On en recense plus de 8000 chez l'humain (Zhang et al. 2003). Les séquences répétées peuvent être très répétées ou moyennement répétées. Dans la première catégorie, on trouve les ADN « satellites » qui représentent 7 % du génome humain. Ce sont de petites séquences de quelques paires de bases (ou pb) mais tellement répétées qu'elles peuvent s'étendre sur plusieurs millions de pb, notamment au niveau des centromères. Les séquences moyennement répétées sont largement constituées d'éléments transposables (les ETs), des séquences ADN capable de se déplacer et de se dupliquer de façon autonome à l'intérieur des génomes. On estime que le génome humain contient plus de 45 % de séquences apparentées aux ETs (éléments encore actifs ou ancestraux). Les séquences uniques non codantes sont extrêmement diverses par leur composition et leur localisation. Si certaines ne semblent pas fonctionnelles, d'autres peuvent avoir des fonctions régulatrices ou produire des ARN dits non codants qui pourront jouer un rôle dans divers mécanismes biologiques.

La double hélice ADN est organisée à l'intérieur de la cellule. Chez les eucaryotes, l'ADN est constitué de plusieurs chromosomes réunis à l'intérieur d'un compartiment cellulaire, le noyau. Les cellules eucaryotes possèdent un nombre variable de chromosomes, allant de 1 à plusieurs centaines. Chaque chromosome est composé d'un centromère et de deux bras chromosomiques distincts protégés à leurs extrémités par des télomères. La quantité totale d'ADN dans une cellule peut s'étendre sur de très longues distances. On estime que,

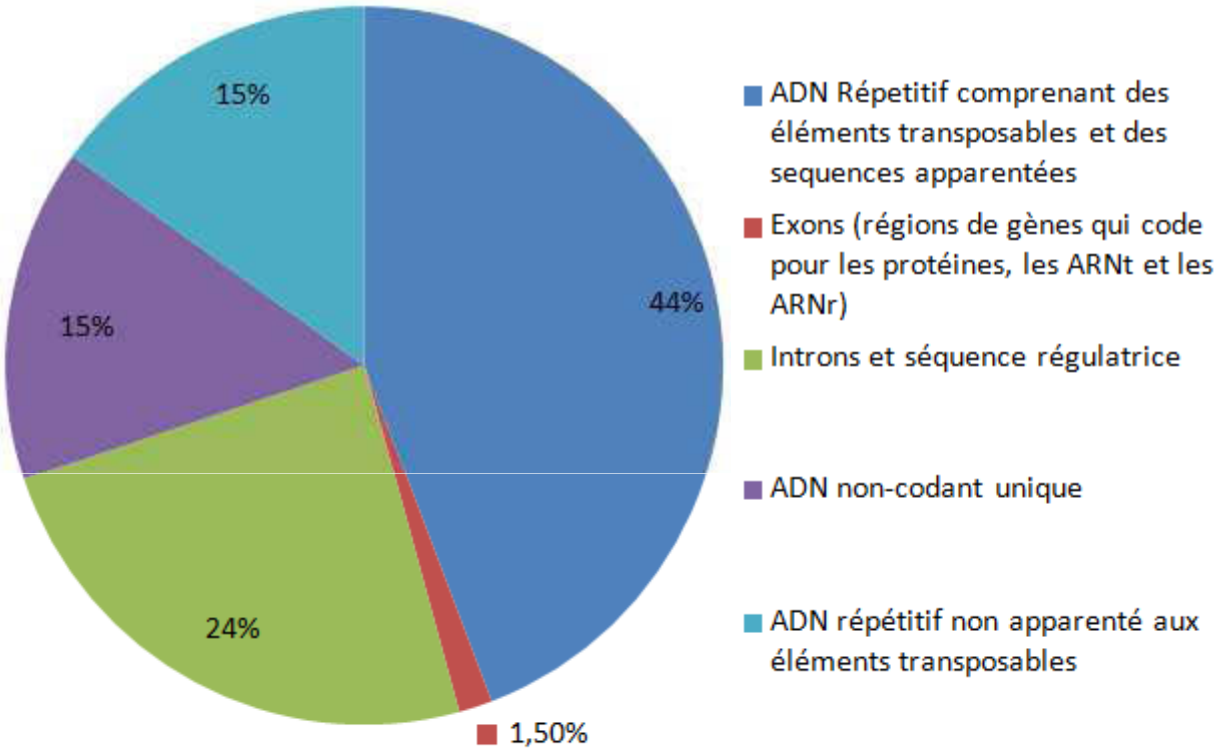


Figure 5 : Les différents types de séquences ADN dans le génome humain

Représentation graphique de la proportion des différents types de séquences ADN retrouvés dans le génome humain.

[http://fr.wikipedia.org/wiki/ADN_non_codant]

mis bout à bout, l'ADN contenu dans une cellule humaine s'étendrait sur 2 mètres de long. Or cette structure est confinée dans un espace très restreint (le noyau cellulaire possède généralement un diamètre de quelques μm). L'ADN va donc être compacté au maximum grâce à différentes protéines. L'ensemble formé par l'ADN et les protéines est appelé chromatine.

La chromatine a été nommée ainsi en 1880 par Walther Flemming pour son affinité forte à la coloration. C'est sous cette forme que l'ADN est présent dans les noyaux des cellules. On estime que la chromatine est constituée d'environ 30% d'ADN et 60% de protéines, dont la moitié sont des Histones, des protéines extrêmement bien conservées dans l'évolution et qui possèdent une forte charge électrique positive leur conférant ainsi une forte affinité pour l'ADN qui est globalement chargé négativement. Lorsque la chromatine est isolée à partir des cellules, on peut observer en microscopie électronique une fibre de 11 nanomètres, dite « en collier de perles » (Olins AL 1974). Des analyses supplémentaires ont montré que cette structure particulière correspond à des segments d'ADN de 146pb enroulés autour du nucléosome, une entité protéique composée de huit protéines qui s'organisent en cylindres (deux exemplaires de chacune des histones H2A, H2B, H3 et H4) (Luger et al. 1997). Le collier de perle est le degré de compaction minimal de l'ADN dans les cellules. L'espace internucléosomal va fixer des protéines dites internucléosomales, notamment l'Histone H1, qui vont replier les nucléosomes entre eux pour former une nouvelle fibre appelée solénoïde de 30nm de diamètre (Adkins et al. 2004). La chromatine va ensuite subir d'autres étapes de compaction qui varient d'une séquence ADN à une autre ou d'une cellule à une autre et peuvent aboutir à une fibre de 1400 nm de diamètre visible en microscopie optique, lors de la mitose **(Figure 6)**.

b) Compaction et décompaction de la chromatine

Bien que fortement compactée, la séquence ADN doit rester accessible à toute la machinerie protéique nécessaire à de nombreux mécanismes biologiques tels que la transcription des gènes ou la réplication de l'ADN, d'où la nécessité de pouvoir naviguer entre les différents niveaux de compaction. Par exemple, l'activité transcriptionnelle est fortement influencée par le degré de compaction de la chromatine. Trois mécanismes principaux vont être capable de réguler ce degré de compaction : la méthylation de l'ADN, la modification des histones et l'intervention des facteurs de remodelage de la chromatine.

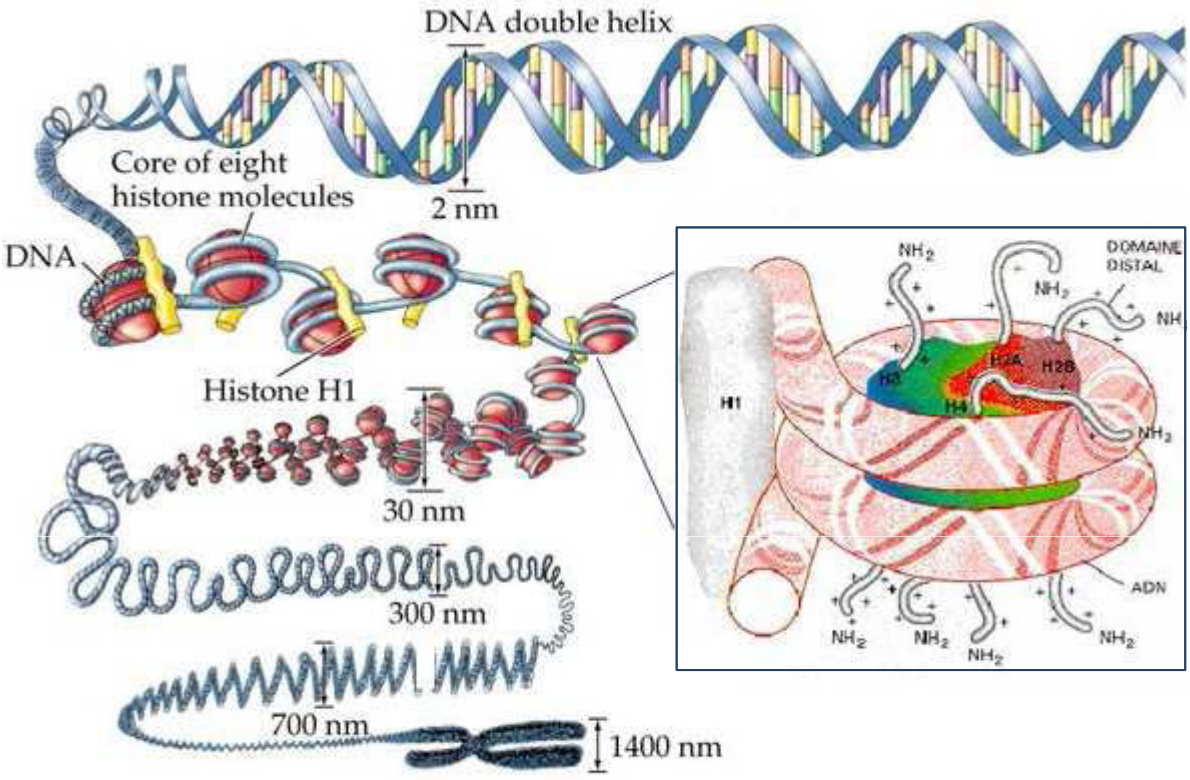


Figure 6 : Les différents niveaux de compaction de la chromatine

Illustration de la conformation spatiale de la double hélice ADN à l'intérieur des noyaux. L'encadré représente un zoom effectué sur un nucléosome.

[Grunstein *et al*, 1992]

La méthylation de l'ADN est un processus réversible qui ajoute un groupement méthyle sur le carbone 5 d'une base cytosine ou, plus rarement, adénosine. Cette marque épigénétique, dite stable car transmise à la descendance lors de la mitose et de la méiose, est associée à une fermeture de la chromatine et donc à une répression de l'expression des gènes. Elle est retrouvée dans de très nombreux organismes, de la bactérie à l'homme, attestant de son origine évolutive précoce. Pourtant, cette marque n'est retrouvée que très faiblement chez certaines espèces, comme *D. melanogaster* et semble totalement absente chez les levures (Capuano et al. 2014). Chez les vertébrés, la méthylation est réalisée presque uniquement en contexte symétrique, sur des dinucléotides CG (70 à 80 % des CG du génome humain sont méthylés), tandis que les plantes présentent des cytosines méthylées également dans des contextes non symétriques (CHH par exemple). La méthylation est assurée par différentes enzymes ADN méthyl-transférase. Dnmt3a et Dnmt3b sont responsables de l'établissement de-novo de la méthylation, notamment dans l'embryon (Okano et al. 1999). Dnmt1 assure la maintenance des marques de méthylation lors de la réplication de l'ADN (Bestor 1992; Leonhardt et al. 1992). La méthylation de l'ADN peut inhiber l'activité transcriptionnelle en perturbant la fixation des facteurs de transcription sur leur site de fixation (Watt & Molloy 1988). De plus, cette marque est reconnue par des protéines possédant un domaine Methyl-CpG binding (protéines appartenant aux familles MBD, Kaiso ou SRA) qui peuvent agir directement sur l'architecture de la chromatine ou qui recrutent des enzymes de modification des histones, telles que des Histones methyl-transferases ou Histones Déacétylases (Nan et al. 1998) **(Figure 7A)**. Des défauts de méthylation d'ADN sont associés à une létalité embryonnaire chez les mammifères ce qui témoigne de l'importance de cette marque (Li et al. 1992). Elle a en effet été impliquée dans des mécanismes variés tels que la répression des gènes et des ETs, l'empreinte parentale ou l'inactivation du chromosome X.

Les histones sont des protéines possédant un domaine central globulaire extrêmement conservé capable de fixer l'ADN et les autres histones, ainsi qu'une extrémité amino-terminale libre riche en résidus lysine et arginine et dépourvue de structure secondaire. Les histones peuvent subir de nombreux types de modifications post-traductionnelles (PTMs) qui permettront de moduler la compaction de la chromatine. L'acétylation et la méthylation des résidus lysines furent les premières PTMs à avoir été découvertes (ALLFREY et al. 1964). Depuis, il a été recensé des cas d'ubiquitinylation et sumolyation des résidus lysines, méthylation et citrullination des arginines, ADP-ribosylation et phophorylation des serines, threonines et tyrosines,... (Bannister & Kouzarides 2011) **(Figure 7B)**. La majorité de ces

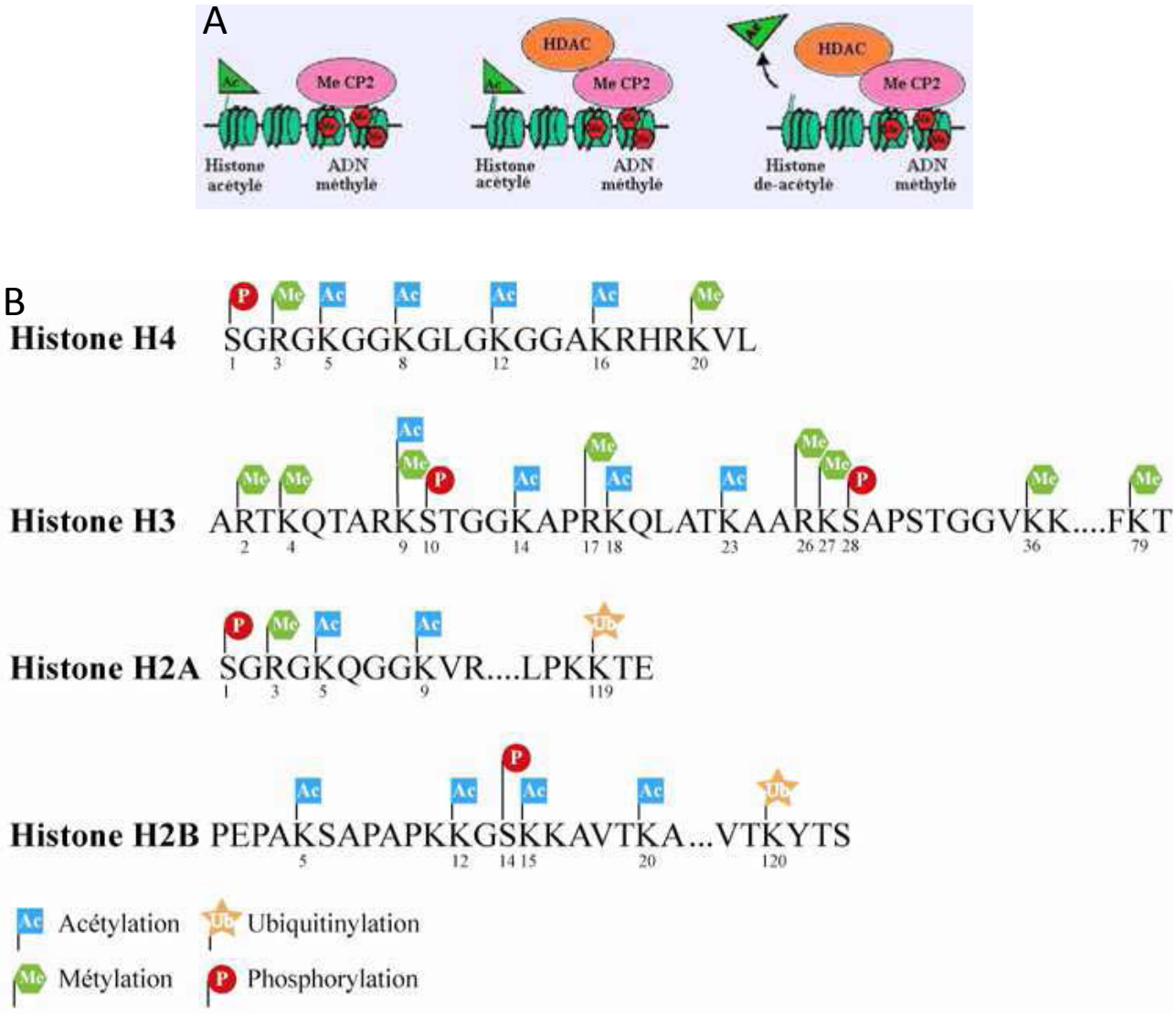


Figure 7 : Méthylation de l'ADN et modifications post-traductionnelles (PTMs) des histones

A) La protéine Methyl-CpG Binding Protein 2 (MeCP2) fixe les cytosines méthylées de l'ADN et recrute des Histones Déacétylases afin d'ôter les groupements Acétyles des histones.

[http://atlasgeneticsoncology.org/Educ/HeterochromID30058FS.html]

B) Récapitulatifs des différentes PTMs des Histones.

[http://theses.ulaval.ca/archimede/fichiers/22435/ch02.html]

modifications a lieu sur l'extrémité amino-terminale libre des histones. Chaque modification, que l'on appelle aussi marque d'histone, est mise en place par une certaine catégorie d'enzymes. L'ensemble de ces marques sur une région génomique constitue un code qui dictera le degré d'ouverture de la chromatine. Certaines marques ont un effet direct sur l'interaction des histones avec l'ADN, comme par exemple l'acétylation des lysines, mise en place par des Histones Acétyle-Transférase (HAT), qui entraine une diminution de la charge négative des histones et donc une diminution de son affinité pour l'ADN. Le nucléosome va être plus relâché ce qui facilitera la transcription de la séquence ADN (Shogren-Knaak et al. 2006). Cette marque est donc généralement associée à une augmentation de l'activité transcriptionnelle. D'autres marques vont plutôt recruter des protéines qui elles seront capables d'agir directement sur la compaction de la chromatine. C'est le cas de la tri-méthylation de la lysine 9 de l'histone H3, ou H3K9me^3, qui recrute la protéine HP1 (Lachner et al. 2001; Bannister et al. 2001); ou de la méthylation de la lysine 27 de l'histone H3, H3K27me^3, qui recrutera des protéines du type Polycomb (Cao & Zhang 2004). Ces deux marques seront associées à une fermeture de la chromatine, et donc à une répression transcriptionnelle même s'il a été mis en évidence que certains gènes transcriptionnellement actifs présentent ces marques chromatiniennes (les gènes *light* et *rolled* par exemple chez *D. melanogaster*)(Lu et al. 2000). A l'inverse, la méthylation des lysines 4 et 36 de l'histone H3 est associée à une ouverture de la chromatine. En plus des marques d'histones, le nucléosome peut également présenter des variants d'histones aux rôles variés.

De nombreuses protéines peuvent également influencer directement ou indirectement le niveau de compaction de la chromatine. Ce sont les facteurs généraux de remodelage, composés de nombreuses familles différentes sur leur mode d'action, leur domaine catalytique et leur domaine d'interaction (Muchardt & Yaniv 1999). Par exemple, les protéines des sous-familles Swi/Snf2 et CHD utilisent de l'ATP pour casser les liaisons ADN-Histones et ainsi faire glisser la séquence ADN le long du nucléosome ou remodeler celui-ci afin de rendre accessibles différentes séquences ADN (Corona et al. 1999). Les enzymes de la sous-famille Swi2 possèdent un bromo-domaine qui leur permettra de cibler les zones où les histones sont majoritairement acétylées tandis que les protéines CHD possèdent un chromo-domaine qui reconnaîtra les histones méthylées (Grüne et al. 2003). Les mécanismes d'action sont également variés, ces enzymes pouvant agir par exemple sur le surenroulement des nucléosomes, l'accessibilité de la séquence ADN et la structure du nucléosome.

Tous ces facteurs permettent la mise en place de différents niveaux de compaction de la chromatine. Historiquement, Emil Heitz mis en évidence en 1928 deux niveaux de compaction du génome : l'euchromatine et l'hétérochromatine dont la coloration diffère à l'intérieur du noyau **(Figure 8A)**. L'euchromatine est associé à des marques de chromatine ouverte (acétylation des lysines, méthylation de la lysine 4 de l'histone H3) et donc à un état transcriptionnel actif. Elle contient la majeure partie des gènes et est localisée plutôt à l'intérieur des bras chromosomiques. L'hétérochromatine est associée à des marques de chromatine fermée (méthylation de la lysine 9 de l'histone H3) et généralement à un état transcriptionnel inactif. Elle ne contient d'ailleurs que peu de gènes. On distingue l'hétérochromatine constitutive, qui comprend les zones péricentromériques et télomériques, de l'hétérochromatine facultative, localisée plutôt dans les régions euchromatiques et dont le degré de compaction varie en fonction du stade de développement ou du tissu cellulaire. Depuis, l'évolution des technologies et notamment des techniques de séquençage associées à l'Immuno-Précipitation de la Chromatine (ChIP-seq) a permis de mettre en évidence plusieurs types de chromatines, désignés par des noms de couleurs (Filion et al. 2010). Cinq types de chromatine, définis selon les protéines non histones qu'elle contient, ont ainsi été mis en évidence chez la drosophile : trois types d'hétérochromatine (bleue, verte et noire) et deux d'euchromatine (rouge et jaune). Ces cinq types de chromatine sont également caractérisés par des marques d'histones différentes **(Figure 8B)**.

c) L'organisation tridimensionnelle du génome

Le noyau est une structure cellulaire séparée du cytoplasme par une double membrane phospholipidique et qui contient l'ensemble du génome de la cellule et de très nombreuses protéines nécessaires à son fonctionnement. Il en résulte un espace de quelques micromètres cubes extrêmement dense qui s'est révélé être hautement organisé. Ainsi, le noyau s'organise autour de plusieurs domaines nucléaires possédant diverses fonctions spécialisées et regroupant des acteurs protéiques et/ou ribonucléiques différents. De plus, les séquences ADN adoptent une conformation spatiale particulière qui semble dépendante de la régulation transcriptionnelle qu'elles subissent. Avec l'évolution progressive des techniques, de nombreuses expériences ont permis de montrer que la chromatine ne forme pas « un plat de spaghettis » désorganisé à l'intérieur du noyau mais possède une structure tridimensionnelle complexe et fonctionnelle.

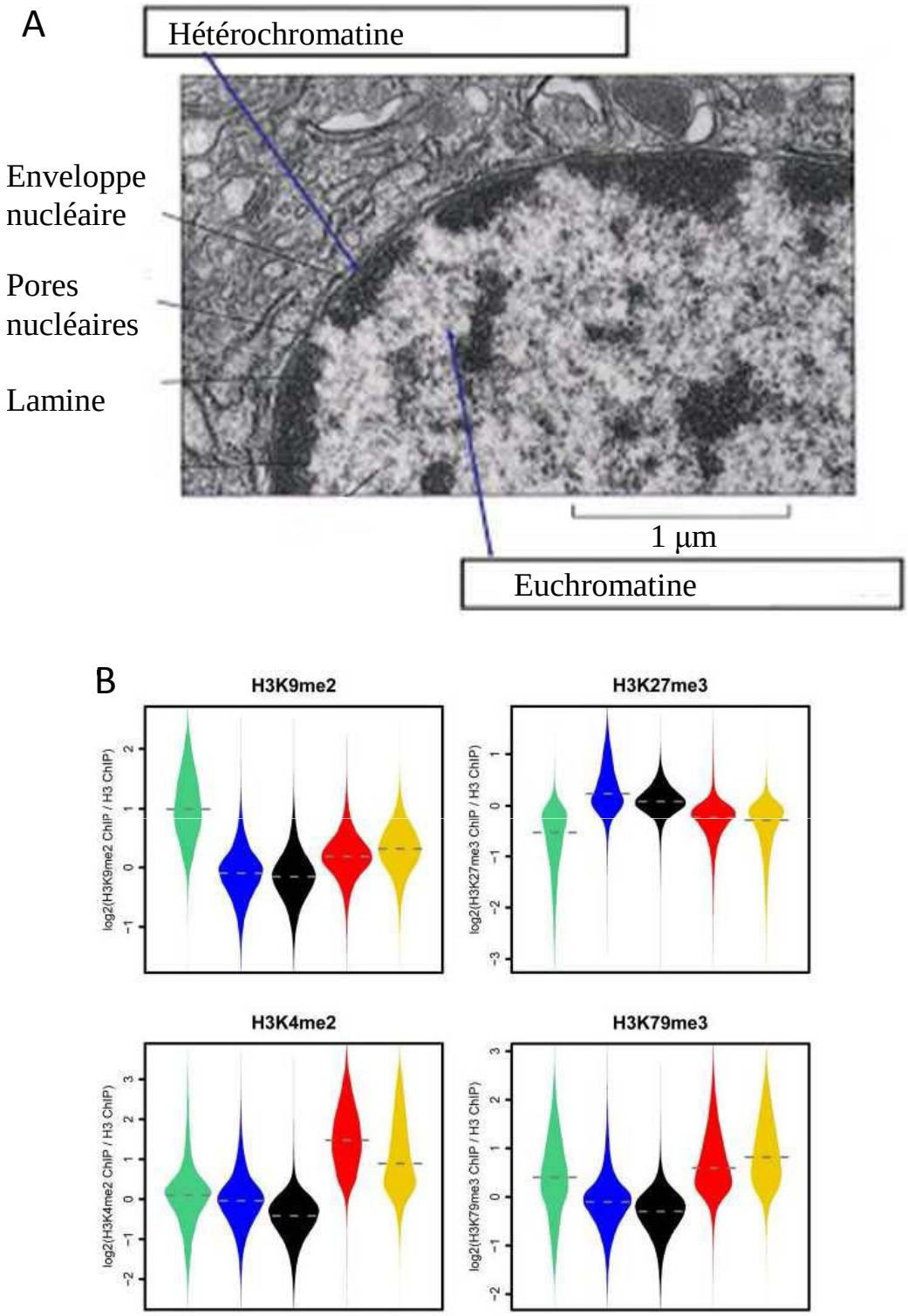


Figure 8 : Les différents contextes chromatiniens

A) Image d'un noyau obtenue par microscopie électronique. L'hétérochromatine correspond aux zones sombres du noyau, les zones claires étant de l'euchromatine.

[http://genomedarkmatter.blogspot.com]

B) Enrichissement en diverses modifications d'histones des différents types de chromatine obtenu par ChIP-seq sur des noyaux de drosophile [Filion *et al*. 2010]

A la fin du XVIIIème siècle, F. Fontana, un abbé italien nota la présence de corps ronds à l'intérieur du noyau des cellules : il venait de découvrir le nucléole, un des domaines nucléaires les plus spacieux et les plus étudiés. Des analyses supplémentaires ont révélé, des années plus tard, que ce domaine est spécialisé dans la transcription des ARN ribosomiques (ou ARNr) et la formation des ribosomes (Birnstiel et al. 1966). De nombreux autres domaines nucléaires ont ensuite été mis à jour par des techniques d'immunofluorescence permettant d'observer les répartitions hétérogènes des protéines **(Figure 9)**. Ainsi, on peut citer les corps de Cajal, où a lieu la maturation des snRNP (small nucleolar RiboNucléoProtéine), les granules interchromatiniens ou « speckles », impliqués dans les phénomènes d'épissage, les centres de transcription, riches en facteurs protéiques indispensables à la transcription, … (Misteli 2005). S'il n'existe qu'un seul nucléole par noyau, dont la taille peut atteindre plusieurs micromètres de diamètre, la plupart des autres domaines nucléaires sont présents en plusieurs exemplaires.

Si les chromosomes apparaissent bien distincts lors de la mitose, leur conformation pendant l'interphase a été source de débats pendant de nombreuses années. En 1909, Theodor Boveri, un biologiste allemand qui s'intéressait à l'organisation de la chromatine dans le noyau, déduisit de ses observations une répartition spécifique des différents chromosomes pendant l'interphase qu'il appela « territoires chromosomiques », ou TC. Il établit plusieurs règles de ces territoires : i) L'ordre des TC ne change pas pendant l'interphase et la prophase, ii) Le voisinage des TC évolue entre la prophase et la métaphase, iii) L'ordre des TC établit sur la plaque métaphasique sera conservé jusqu'à la fin de la mitose et donc sera symétrique entre les deux cellules filles. Toutefois, les expérimentations de microscopie électronique ne permirent pas d'observer ces différents TC et il fallut attendre 1977 pour qu'une équipe américaine réussisse, par un traitement chimique approprié, à observer les TC durant l'interphase (Stack et al. 1977). Depuis, les techniques d'Hybridation In Situ à Fluorescence, ou FISH, appliquées aux chromosomes entiers (Cremer et al. 1988) ont permis d'observer les TC dans tous les organismes étudiés à ce jour **(Figure 10A)**. Les TC peuvent varier selon l'espèce et les tissus concernés même s'il semble que les chromosomes possédant une faible concentration de gène soient situés plutôt vers la périphérie et les autres au centre du noyau. Deux modèles différents tentent de définir les limites de ces territoires. Le premier, imaginé par l'équipe Cremer, suppose l'existence d'un milieu inter-chromatinien sans chromatine séparant les différents territoires et contenant divers facteurs protéiques comme ceux nécessaires à la transcription (Albiez et al. 2006). A l'inverse, selon l'équipe Pombo, les

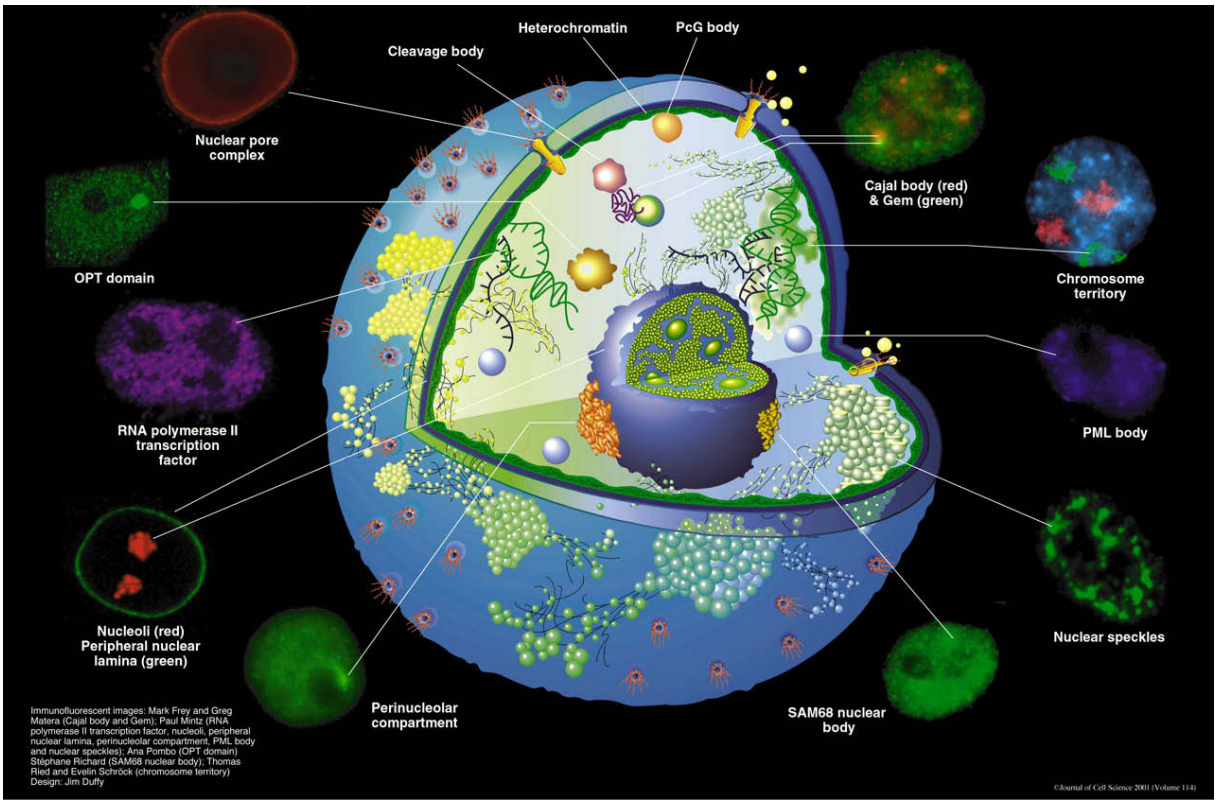


Figure 9 : Le noyau, un organite intracellulaire hautement organisé

Représentation schématique des différents domaines et corps nucléaires d'un noyau de cellule eucaryote ainsi que les images obtenues par microscopie à fluorescence de certains de ces domaines. [Spector DL, 2001]

différents territoires sont séparés par une zone contenant un entremêlât de chromatine provenant des différents chromosomes voisins (Branco & Pombo 2006) **(Figure 10B)**.

Bien que répartis en territoires chromosomiques distincts, des interactions entre chromosomes demeurent cependant nécessaires notamment lors des événements de réparation ou de réplication de l'ADN, mais encore lors de la régulation de l'expression génique. De manière générale, la position d'un gène à l'intérieur du noyau ne semble pas être une caractéristique aléatoire. Dès 1928, E. Heitz observa que l'euchromatine, riche en gène, est localisée à l'intérieur du noyau tandis que l'hétérochromatine, pauvre en gène est repoussée vers les parois membranaires. A l'intérieur des TC, les régions actives transcriptionnellement sont plutôt présentes à la périphérie tandis que les régions inactives sont regroupées vers le centre. Il a également été possible de positionner artificiellement un locus donné à la périphérie nucléaire, en forçant son interaction avec les protéines lamines qui tapissent la paroi nucléaire, ce qui a parfois entrainé une diminution de l'expression de ce dernier (Reddy et al. 2008). De même, il a été montré que le locus de la β-globine est progressivement repositionné vers le centre du noyau au cours de la maturation des érythrocytes dans le foie de la souris, ce qui coïncide avec l'initiation de la transcription des gènes de globine (Ragoczy et al. 2006).

De nombreuses boucles d'interaction physique ont également été mises en évidence, notamment par des techniques de double FISH. C'est le phénomène de « gene kissing » (Spilianakis et al. 2005), qui implique majoritairement des interactions physiques entre des gènes et leurs séquences ADN régulatrices **(Figure 10C)**. De même, il a été observé une colocalisation des gènes conjointement régulés. Un des exemples les plus connus concerne les gènes soumis à la régulation par des protéines du type Polycomb (ou PcG). Les gènes ayant une fonction dans le développement de l'organisme, appelés gènes homéotiques, doivent être réprimés lorsqu'ils ne sont plus nécessaires et cette répression doit être maintenue durant toute la vie de l'individu. Le maintien de cette répression est assuré par les PcG qui sont capables de condenser directement et indirectement la chromatine. Des centaines de gènes sont concernés par cette régulation. Or, ces gènes se regroupent par des interactions longues distances au sein de foyers nucléaires contenant les PcG. Si ce regroupement est empêché, des mutations touchant le développement harmonieux de l'individu sont observées (Bantignies et al. 2011) ;

Ces différentes interactions longues-distances à l'intérieur du noyau font entrevoir un réseau complexe de chromatine difficilement modélisable dans sa globalité. Jusqu'au début

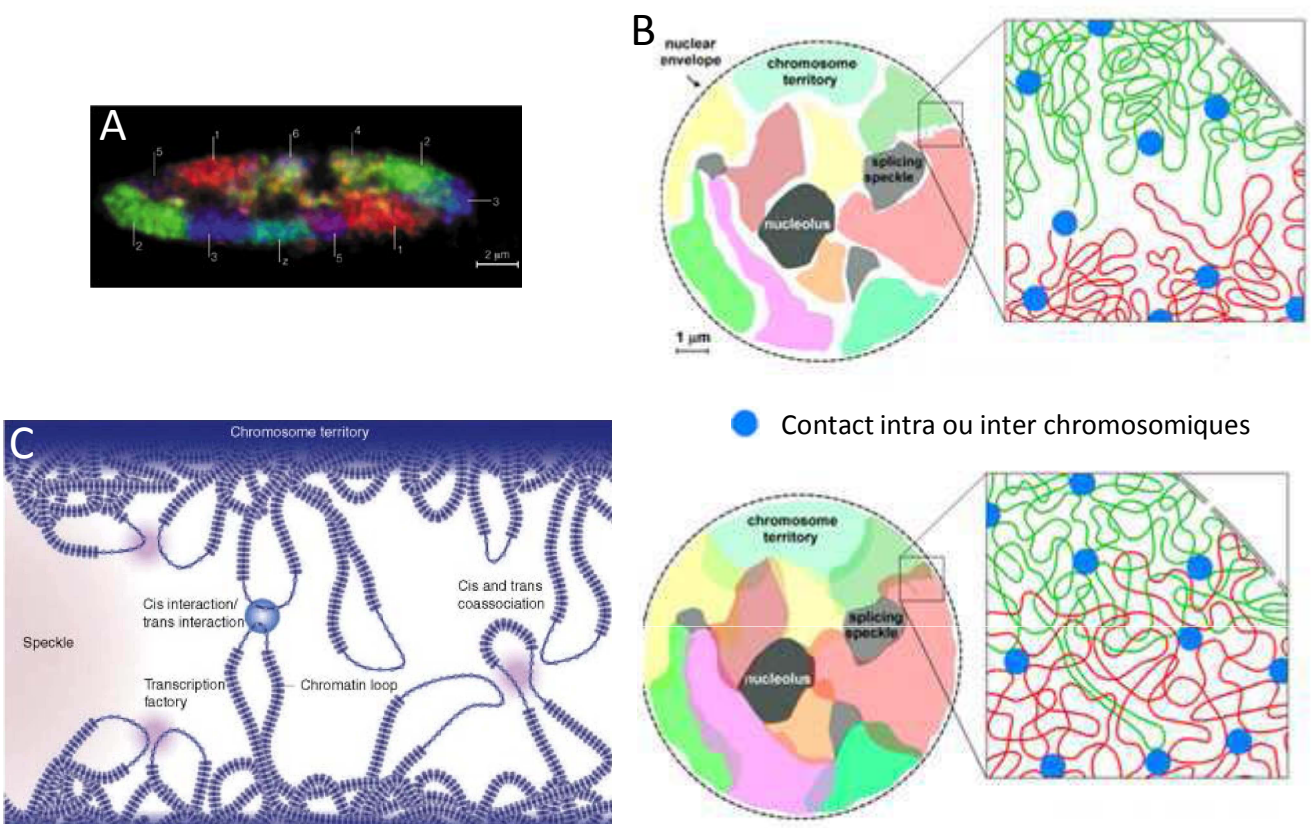


Figure 10 : Organisation des chromosomes à l'intérieur du noyau.

A) Hybridation fluorescente in situ à ADN de chaque chromosome sur un noyau de poulet.

B) Un premier modèle d'organisation (schéma du haut) suppose que les chromosomes sont séparés par un milieu interchromatinien sans chromatine. Un deuxième modèle (schéma du bas) prédit une frontière floue entre deux territoires composée d'un entremêlât de fibres chromatiniennes.

C) Représentation schématique des interactions longues distances entre différentes régions du génome. [Cremer et Cremer, 2010]

des années 2000, seules les techniques de FISH permettaient de localiser un ou plusieurs loci à l'intérieur d'un noyau. Depuis une dizaine d'années, de nouvelles techniques en plein essor permettent de localiser les différents partenaires d'interactions au sein d'une vaste population de noyaux cellulaires. La technique de 3C (Chromosome Conformation Capture), mise au point en 2002 (Dekker et al. 2002), et ses dérivées permettent de connaître les partenaires d'interactions d'un ou plusieurs loci grâce à un ingénieux couplage de digestion et ligation enzymatique de l'ADN **(Figure 11)**. Si le 3C permet seulement de quantifier l'interaction entre deux loci connus, la technique de 4C (Circularized Chromosome Conformation Capture)associée au séquençage haut-débit va révéler l'ensemble des loci interagissant avec un locus donné. Au-delà, les premières techniques de 5C (Chromosome Conformation Capture Carbon Copy) ou de HI-C fournissent une carte détaillée de l'ensemble des boucles d'interactions à l'intérieur des noyaux.

Ces techniques ont permis de mettre en évidence de nombreuses interactions, notamment entre les gènes régulés et leurs séquences régulatrices. Ces regroupements ont été appelés ACH pour Active Chromatin Hub. Là encore, le locus de la β-globine fournit un exemple tout à fait intéressant. Ce locus de 200 kilobases (kb) contient 5 gènes qui s'expriment successivement au cours de la différenciation des érythrocytes. Cet échange est contrôlé par le locus régulateur LCR situé quelques dizaines de kb en amont des gènes. Or, des expériences de 3C ont permis de montrer que le locus LCR forme un ACH spécifiquement avec le gène actif (Tolhuis et al. 2002). De nombreux autres ACH ont ainsi été mis en évidence dans de nombreux organismes. Les expériences de 3C ont également permis de montrer une boucle d'interaction entre le départ et la fin d'un gène, surtout chez les gènes fortement exprimés. Cette boucle serait associée à une augmentation de l'activité transcriptionnelle grâce à un rechargement facilité de l'ARN polymérase (Németh et al. 2008). Les expériences plus globales de 4C ont confirmé les changements de partenaires d'interaction en fonction de l'activité transcriptionnelle des gènes. Les gènes homéotiques ont fourni des exemples probants de ces variations. Organisés en clusters de plusieurs dizaines de gènes, il a été montré dans les tissus embryonnaires de souris que les gènes homéotiques actifs d'un même cluster se regroupent en un foyer activateur de la transcription tandis que ceux qui sont inactifs forment un autre foyer ayant peu de contact avec les gènes actifs (Noordermeer et al. 2011). Le passage des gènes d'un foyer à un autre a pu être observé par l'équipe en analysant des tissus appartenant à différentes lignées cellulaires.

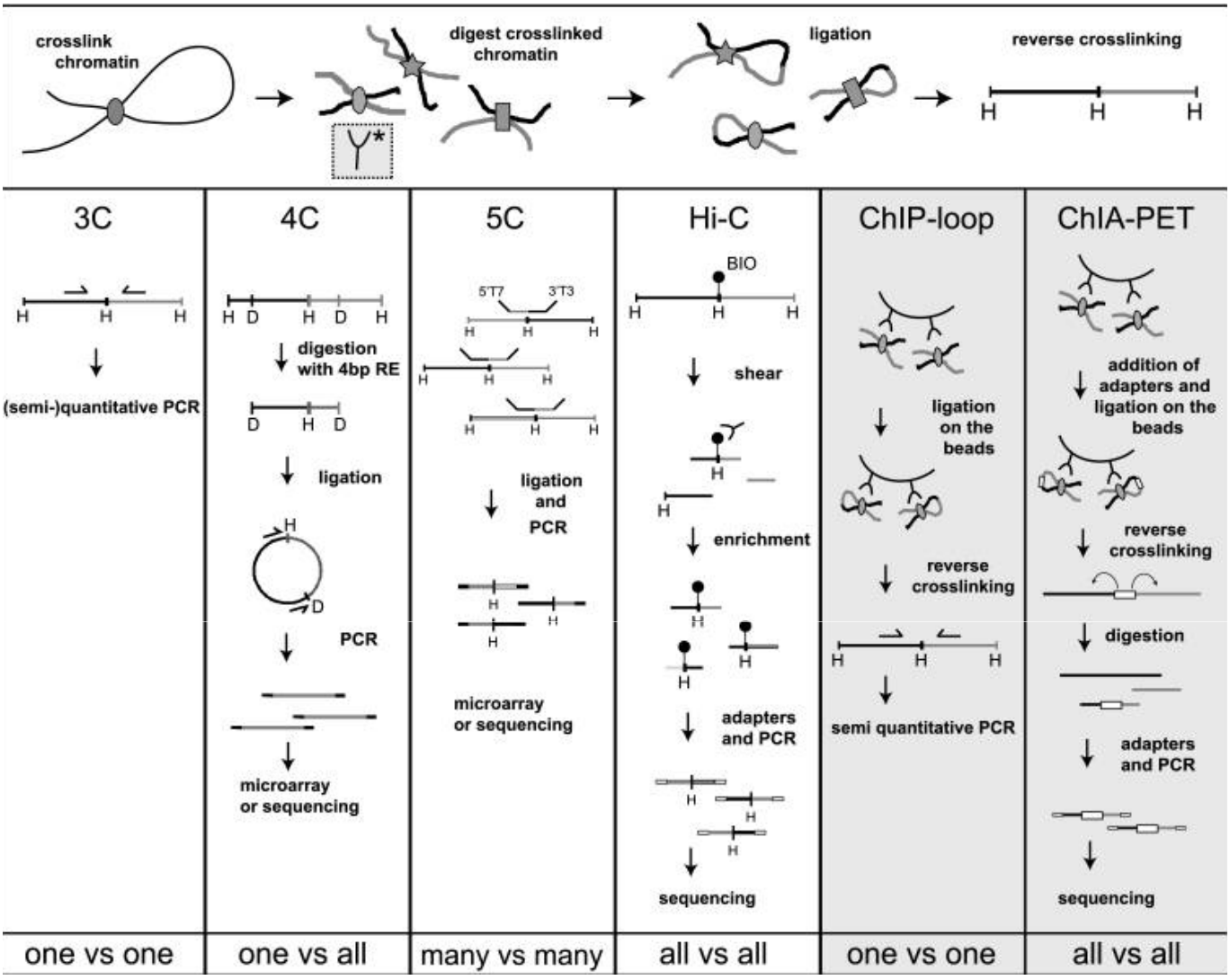


Figure 11 : Vue d'ensemble des méthodes dérivées du 3C (Chromosome Conformation Capture)

Le schéma horizontal représente les étapes de fixation, digestion et ligation commune à toutes ces méthodes. Les schémas verticaux représentent les étapes spécifiques à chacune des méthodes

[de Wit E et de Laat W, 2012]

B. Expression du génome

1. Les différents types d'ARN

Jusqu'au début des années 2000, la communauté scientifique envisageait le génome comme une suite d'unités transcriptionnelles produisant un ARNmessager (ARNm) unique qui sera traduit en une protéine donnée. Les techniques de séquençage haut débit ont depuis permis de réaliser des études transcriptomiques à grande échelle qui ont révélé que, loin d'être restreint aux gènes codants, la quasi-totalité du génome est transcrite (environ 90% chez l'homme) produisant de nombreux ARNm mais également un grand nombre d'ARN ne correspondant à aucune protéine, que l'on les appelle les ARNnc (ARN non codants) (Birney et al. 2007).

a) Les ARN messagers

Les gènes codants produisent un ARNm composé d'une séquence codante (CDS, Coding DNA Sequence) entourée de deux régions non traduites, les 5'UnTranslated Region et 3'Untranslated Region (5'UTR et 3'UTR). Le CDS est lui-même morcelé en plusieurs parties : les exons, qui seront traduits en acides aminés et les introns qui seront éliminés de l'ARNm mature par le phénomène d'épissage. Cet ARNm sera ensuite traduit en protéine par les ribosomes grâce au code génétique. En effet, la séquence nucléotidique des gènes représente un code dans lequel chaque triplet de nucléotides (64 possibilités) correspond à un des 20 acides aminés qui composent les protéines. Ce code est universel : on le retrouve dans toutes les espèces vivantes avec quelques variantes observées chez l'ADN mitochondrial de certaines espèces. De plus, le code génétique est non ambigu, un triplet équivaut à un seul acide aminé, et redondant, un acide aminé est codé par plusieurs triplets.

Les gènes codants sont extrêmement divers de par leur taille, leur expression et leur produit final. Par exemple, certains gènes, comme celui de la dystrophine chez les vertébrés, s'étendent sur plusieurs milliers de kb tandis que d'autres ne font que quelques centaines de pb, comme par exemple l'histone H1a qui s'étend sur 781 nucléotides. La taille moyenne d'un gène humain est de 27 kb pour une séquence codante de 1,3kb (Birney et al. 2007). Les écarts d'expression sont tout aussi impressionnants puisque certains gènes ne vont être exprimés que ponctuellement dans quelques cellules tandis que d'autres, appelés gènes de ménage, vont être

exprimés fortement dans tous les types cellulaires. Ces derniers sont couramment utilisés en biologie moléculaire à des fins de normalisation. Enfin, les protéines produites par ces gènes vont fournir des activités aussi diverses que des fonctions de structure, comme les composants du cytosquelette, ou des fonctions enzymatiques.

De nombreuses études ont montré que l'image simpliste du gène produisant un ARN devait être remise en question. En effet, la plupart des gènes se sont transformés en une pléiade de transcrits possibles, dans lesquels le départ et la terminaison de la transcription ainsi que l'épissage peut varier. Des cas de trans-épissage, fusion de deux ARN différents, ont également été reportés. Tout cela génère un nombre d'ARNm bien supérieurs au nombre de gène, en accord avec les dizaines de milliers de protéines différentes synthétisées.

b) Les ARN non codants

Les séquençages haut débit ont donc révélé qu'il n'existait pas que des ARN codants. Tout d'abord, de nombreux ARN se sont superposés à la transcription des gènes codant les protéines chez les eucaryotes **(Figure 12)**. Chaque unité de transcription est enchevêtrée dans un réseau complexe de transcrit. Par exemple, de nombreux ARN en orientation inverse (dits Antisens, ou AS) de taille variable sont retrouvés tout le long des gènes, en position tête à tête, totalement chevauchante ou queue à queue. De plus, la région autour du promoteur des gènes semble être soumise à de nombreux événements de transcription opportunistes. Ces événements produisent de nombreux transcrits non canoniques qui seront détruits par les cellules, notamment par le complexe protéique de l'exosome, qui possède une activité exoribonucléolytique 3'-5'. Ainsi, la mutation d'une des protéines de l'exosome chez les levures comme chez les humains provoque une forte augmentation des transcrits associés aux promoteurs (Wyers et al. 2005). On retrouve aussi deux types de petits ARN associés généralement aux gènes les plus fortement exprimés et localisés au niveau du promoteur ou du site de terminaison de la transcription, les PASR (Promoter Associated Short RNAs) et les TASR (Telomeric Associated Short RNAs) (Kapranov et al. 2007). Les PASR font souvent 26, 38 ou 50 nucléotides et correspondraient à un arrêt précoce de la transcription tandis que les TASRs ont une taille variable de 20 à 200 nucléotides et seraient dus au clivage de transcrits plus longs (Valen et al. 2011).

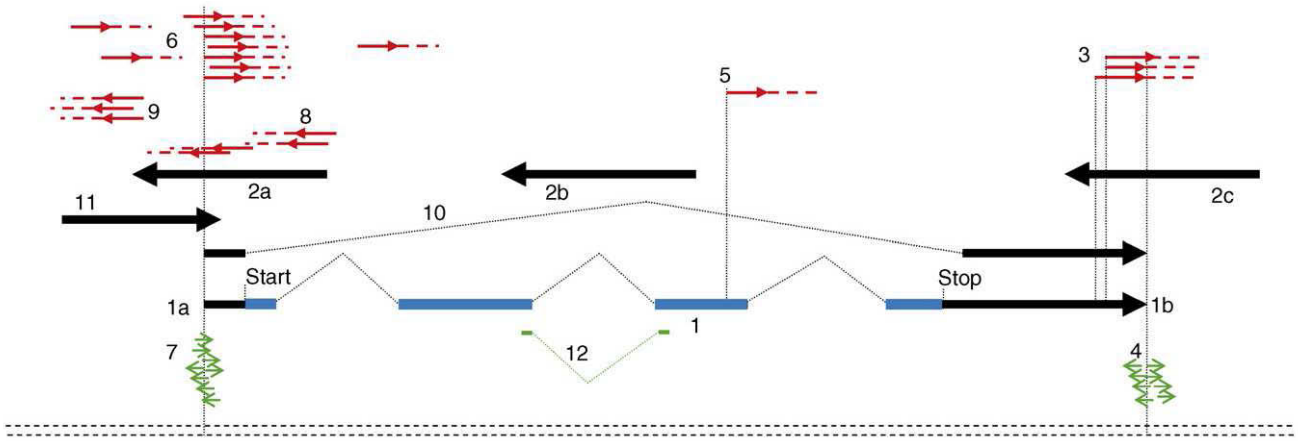


Figure 12 : Un seul locus peut produire de nombreux ARN

Représentation schématique des différents ARN dont la séquence peut être semblable ou complémentaire inverse à celle d'un ARN messager. (1) ARNm principal avec son départ (1a), son site de terminaison (1b) et son CDS figuré en bleu. (2) ARN Antisens en configuration tête à tête (2a), totalement chevauchante (2b) ou queue à queue (2c). (3) Transcrits identifiés à l'intérieur du 3'UTR (3'-Untranslated Region). (4) TASRs (Telomere Associated small RNAs). (5) Transcrits initiés par un promoteur alternatif. (6) Transcrits associés au promoteur. (7) PASRs (Promoter Associated Small RNAs). (8) Transcription AS à partir du premier exon ou du premier intron. (9) ARN issus d'un promoteur bi-directionnel. (10) ARN non codant chevauchant les sites d'initiation et de terminaison de la transcription de l'ARNm. (11) PALRs (Promoter Associated Long RNAs)

[Carninci P *et al*, 2008]

Au-delà des unités de transcription des gènes codant, il existe une grande variété d'ARN non codants qui diffèrent par leur origine, leur structure et leur fonction. Ainsi, les transcriptomes contiennent en très grand nombre des ARN dits de structure. Ces ARN possèdent une forte structure secondaire, un repliement de la molécule d'ARN, qui sera importante pour leur fonction. Il existe deux catégories d'ARN de structure : les ARN de transfert (ARNt) et les ARN ribosomiques (ARNr). Les ARNt sont composés de 70 à 100 nucléotides et seront chargés de transmettre les acides aminés lors de la traduction. Chaque ARNt, une vingtaine, va être responsable du transport d'un acide aminé vers la chaîne protéique en construction. Les ARNr constitueront la majeure partie et la fonction catalytique des ribosomes, également impliqués dans la traduction. Il en existe 4 chez les eucaryotes et 3 chez les procaryotes. Les ARNr proviennent de gènes organisés en tandem et très fortement répétés afin de pourvoir à la très forte demande de la cellule pour ces ARN (jusqu'à 80% de la masse totale d'ARN d'une cellule). On trouve également d'autres ARN non codants qui participent aux mécanismes cellulaires, appelés les snARN (small nuclear RNAs), dont les snoARN (small nucleolar RNAs) qui participent à l'épissage.

De plus, il existe de nombreux ARNnc de grande taille (lncARN, long non coding RNA, >200 nucléotides, dont 9640 espèces ont été dernièrement recensées par GENCODE chez l'humain) qui interviennent dans divers processus de régulation de l'expression génique. Un exemple bien documenté est celui de l'ARNnc *Xist* (X Inhibitory Specific Transcript), responsable de l'inactivation du chromosome X dans les cellules de mammifères femelles. Cet ARNnc de 19 kb, produit en excès par un des deux chromosomes X, recouvre l'ensemble de ce chromosome et recrute la protéine PRC2 (Polycomb repressive Complex 2) qui va méthyler la lysine 27 de l'histone H3 entrainant l'hétérochromatisation de ce chromosome et son inhibition transcriptionnelle (Wutz 2011) **(Figure 13)**. Des exemples de lncARN activateurs de la transcription ont également été mis en évidence comme l'ARN *Mistral* qui active l'expression de deux gènes homéotiques chez la souris (Bertani et al. 2011).

Les génomes produisent aussi une multitude de petits ARN régulateurs (<30 nucléotides) qui interviennent dans les mécanismes d'ARN interférence, qui se divisent en trois voies : les siRNA, les miRNA et les piRNA. Ces voies ont en commun d'associer un petit ARN avec une protéine de la famille Argonaute pour former un complexe RISC (RNA Induced Silencing Complex). Ce complexe va cibler un ARN précis grâce à la complémentarité de séquence entre le petit ARN et l'ARN cible, entrainant l'inhibition de

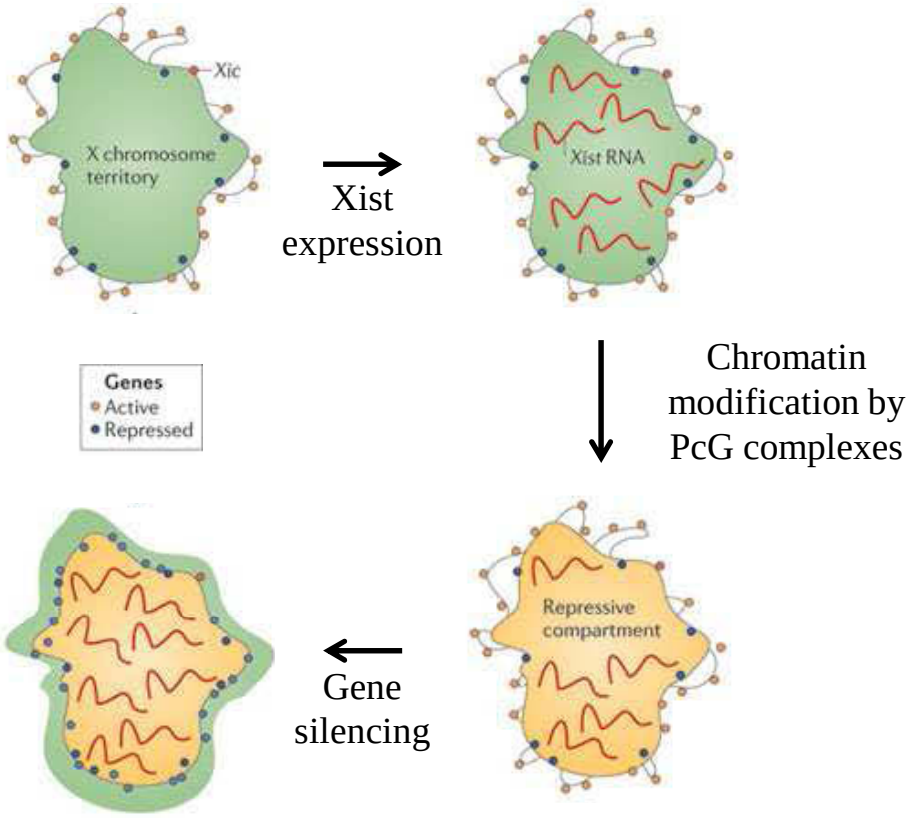


Figure 13 : Inactivation du chromosome X par le long ARN non codant Xist

Représentation schématique du déroulement de l'inactivation du chromosome X. Le gène Xic produit l'ARN non codant Xist qui se localise au centre du territoire nucléaire du chromosome X et déclenche la formation d'un compartiment répressif qui va inhiber la transcription des gènes environnants, notamment par l'intervention des protéines Polycomb.

[Adapté de Wutz A, 2011]

l'expression de cet ARN, soit par TGS (Transcriptionnal Gene Silencing) soit par PTGS (Post-Transcriptionnal Gene Silencing).

Devant une telle diversité d'ARN qui semblent jouer des rôles de plus en plus significatifs pour le fonctionnement cellulaire, la communauté scientifique s'est une nouvelle fois interrogée. Les protéines, et notamment les enzymes, sont codées à partir d'ARN mais les ARN nécessitent des protéines (ARN polymérase, etc) pour être synthétisés. La biologie s'est donc retrouvée encore une fois devant le paradoxe de l'œuf et la poule : qui de l'ARN ou de la protéine est apparu en premier dans l'évolution ? Ce débat porte ici sur les premiers stades de la vie sur terre, avant même de pouvoir parler de cellules vivantes. Dans les années 1980s, deux scientifiques indépendants, Tom Cech et Sydney Altman, ont démontré l'existence chez le cilié de ribozymes, des ARN intervenant dans les ribosomes et doués de propriétés catalytiques semblables aux enzymes (Kruger et al. 1982; Guerrier-Takada et al. 1983). Ces expériences proposent un nouveau modèle de l'apparition de la vie sur Terre qui aurait commencé par des ARN auto-réplicatifs et dotés de fonctions enzymatiques puis aurait vu l'émergence d'une molécule plus solide pour conserver l'information, l'ADN. Cette théorie replace l'ARN au centre du dogme de la biologie et avec la découverte de plus en plus d'ARN non codants ayant une fonction propre, certains biologistes n'hésitent plus à parler d'un monde ARN, où l'ARN devient un acteur majeur de la biologie cellulaire.

2. Machinerie transcriptionnelle et post-transcriptionnelle

Le génome est transcrit en ARN par des ARN polymérases ADN-dépendantes ainsi que de nombreux cofacteurs formant une structure complexe. Il existe 5 ARN polymérases chez les végétaux et 3 chez les animaux, chacune étant responsable de la transcription de certaines portions génomiques. L'ARN polymérase I est responsable de la transcription des gènes produisant les ARNt et l'ARN U6 du complexe d'épissage ainsi que les ARNr sauf celui de la sous-unité 5S qui est transcrit par l'ARN polymérase III. L'ARN polymérase II transcrit les gènes codants pour des protéines ainsi que de nombreux ARN non codants. Chez les plantes, les ARN polymérase IV et V sont requises pour la transcription et le fonctionnement de nombreux siRNA.

a) La transcription des ARN

La transcription d'un gène est initiée au niveau de son promoteur proximal, qui comprend une centaine de pb autour du départ de transcription, le TSS (transcriptional start site). Cette région contient des séquences nécessaires à la fixation du complexe protéique minimal requis pour la transcription. Il existe plusieurs séquences fonctionnelles dont deux principales : la TATA-box et l'Inr. La TATA- box est une séquence riche en nucléotides Adénine et Thymine située entre 25 à 40 pb en amont du TSS des gènes des eucaryotes supérieurs et jusqu'à 120 pb chez les levures. C'est la séquence retrouvée le plus fréquemment (30 à 40% des gènes de drosophile) et dont la séquence stricte peut largement dériver, la rendant difficile à identifier uniquement par analyse de séquence (Ohler et al. 2002). L'Inr, ou élément Initiateur, est une séquence consensus qui contient le TSS. Ces séquences se répartissent différemment au sein des promoteurs qui peuvent être composites (présence de la TATA-box et de l'Inr), dirigés par la TATA-box ou par l'Inr, ou encore nuls (aucune des deux séquences). Les promoteurs nuls ont souvent des départs de transcription multiples, ce qui montre l'importance de ces séquences dans la spécificité de l'initiation de la transcription. Des analyses *in silico* plus approfondies ont montré que les promoteurs des gènes contiennent d'autres séquences plus ou moins consensus, comme par exemple la séquence BRE (The B Recognition Element) en amont du TSS et la DPE (Dowstream Promoter Element) en aval.

L'ARN polymérase responsable de la transcription des gènes codants pour une protéine est l'ARN polymérase II, un complexe protéique de plusieurs centaines de kDa et composé de 10 à 12 sous-unités protéiques spécifiquement impliquées dans la reconnaissance du TSS, la vitesse de transcription ou l'interaction avec des modulateurs de la transcription. Initialement mise en évidence par sa capacité à produire un ARNm *in vitro*, la spécificité de l'initiation de la transcription n'est obtenue que par l'ajout de cofacteurs nécessaires pour ancrer l'ARN polymérase au début de la portion d'ADN à transcrire. Ce sont les 6 facteurs de transcription généraux (GTFs) spécifiques à l'ARN polymérase II : TFIIA, TFIIB, TFIID, TFIIE, TFIIF et TFIIH. Ces protéines, qui font partie des mieux conservées de l'évolution, vont se fixer, selon un ordre strict également conservé, sur les séquences consensus contenues dans le promoteur des gènes. Grâce à de nombreuses analyses *in vitro*, l'ordre d'assemblage ainsi que la fonction des différents GTFs sont bien connus (Gralla 1996). La première protéine à se fixer sur l'ADN est le facteur de transcription TFIID, notamment grâce à sa sous-unité TBP (TAT A Binding Protein) qui reconnaît la TATA-Box. TBP est une sous-unité

présente dans chaque complexe d'initiation de la transcription, quelle que soit l'ARN polymérase, et qui va induire une courbure particulière de l'ADN permettant aux autres facteurs de se fixer. Cette sous-unité va également recruter de nombreux facteurs protéiques intervenant dans l'initiation de la transcription et appelés TAFs pour TBP-associated factors. TFIIA et TFIIB viennent se fixer sur TFIID et recrutent TFIIH qui possède la capacité à recruter l'ARN polymérase **(Figure 14).** Les facteurs de transcription TFIIE et TFIIH viennent consolider le complexe et ouvrir la double hélice ADN grâce aux activités hélicase de TFIIH. Cette mise en place étape par étape du complexe protéique de la transcription a été depuis remise en question par des études montrant que, *in vivo,* l'holoenzyme composée de l'ARN polymérase et des GTFs préexiste au sein des noyaux (Greenblatt 1997).

Suite à cette phase d'initiation, la transcription doit passer en phase d'élongation où l'ARN sera synthétisé. L'élongation est enclenchée par un échange des cofacteurs interagissant avec l'ARN polymérase. Cet échange est provoqué par la phosphorylation, grâce à l'activité kinase de TFIIH, du domaine CTD contenu dans la sous unité RPB1 de l'ARN polymérase. Le domaine CTD consiste en la répétition d'un motif heptapeptide (Tyr-Ser-Pro-Thr-Ser-Pro-Ser) dont le nombre de répétition varie de 26 chez la levure à 52 chez l'homme. Le domaine CTD va d'abord être phosphorylé sur la sérine en position 5 du domaine CTD ce qui entrainera le détachement des GTFs du promoteur afin de libérer l'ARN polymérase. Par la suite de l'élongation, le domaine CTD sera phosphorylé sur la sérine 2. L'élongation est, pour de nombreux gènes, l'étape limitante de la synthèse des ARN et de nombreux facteurs peuvent influer sur la processivité de l'ARN polymérase, ce qui en fait une nouvelle étape du contrôle de l'expression génique. Les nombreux facteurs d'élongation forment un complexe protéique qu'on appelle le SEC (Super Complex of Elongation).

La terminaison de l'élongation est déclenchée par la présence d'une séquence particulière sur l'ADN, le signal de poly-adénylation. Lorsque l'ARN polymérase transcrit ce signal présent à l'extrémité 3' des gènes, les enzymes CPSF (Cleavage Polyadenylation Specificity Factor) et CstF (Cleavage Stimulation Factor) vont cliver l'ARN et le détacher de l'ARN polymérase. Cette étape est particulièrement importante pour éviter que la transcription ne « déborde » sur les gènes voisins et pour assurer un stock suffisant d'ARN polymérase disponible.

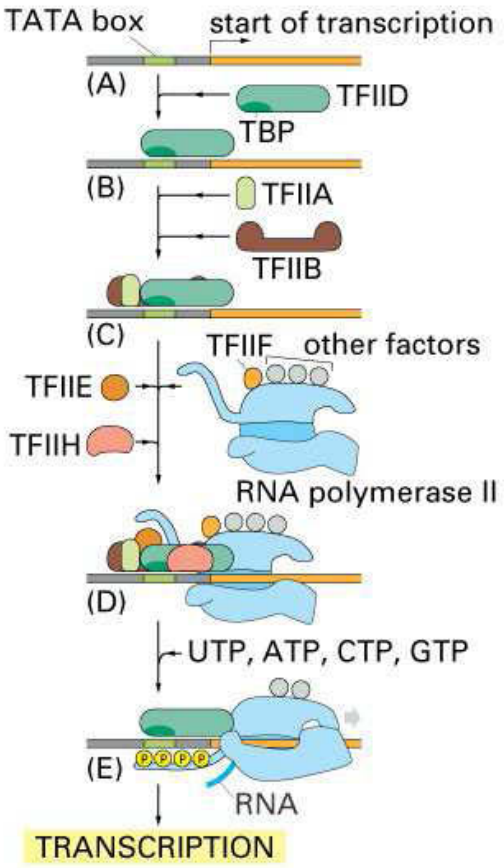


Figure 14 : Les étapes de l'initiation de la transcription

Le promoteur contient des séquences consensus, ici une TATA box (A), qui vont permettre la fixation séquentielle de plusieurs facteurs de transcription généraux (B et C). Ce recrutement permet la mise en place d'une conformation particulière de l'ADN propice à la fixation de l'ADN polymérase (D). TFIIH phosphoryle l'extrémité CTD de l'ARN polymérase nécessaire à l'initiation de la transcription. [Essential cell Biology, 2004 Garland Science]

b) La maturation des ARN

L'ARN issu de la transcription subit plusieurs étapes de maturation : l'ajout de la coiffe et de la queue poly-Adénosine, ou queue poly-A, l'édition et l'épissage des introns. L'ajout de la coiffe est une modification affectant l'extrémité 5' des ARN qui a lieu peu de temps après le début de l'élongation de l'ARN polymérase II, grâce au recrutement d'enzymes responsables de l'ajout d'une méthyl-guanosine tri-phosphate (Ho 1998). La queue poly-A consiste en une suite de nucléotides adénosines (jusqu'à 200) mis en place par une poly-A-polymérase suite à la terminaison de la transcription et au clivage de l'ARN (Hirose & Manley 1998). L'édition est un processus relativement rare permettant d'ajouter ou d'ôter quelques nucléotides, de remplacer ou de modifier une base (Brennicke et al. 1999). L'épissage est une étape complexe qui permet d'enlever les introns du futur ARNm.

En 1977, plusieurs équipes découvrent simultanément que les gènes peuvent être fragmentés et que les ARNm subissent des transformations qui les délestent de certaines séquences, les introns (Chow et al. 1977; Berget et al. 1977). Si l'origine des introns est encore soumise à controverse (il est possible que les ETs ait joué un rôle dans leur apparition (Roy 2004)), ils ont fortement colonisé les génomes, notamment eucaryotes, où la plupart des gènes possèdent plusieurs introns. D'ailleurs, ils constituent, en nombre de pb, la majeure partie des gènes, les séquences codantes étant très limitées.

Les introns possèdent des tailles extrêmement variables, allant de moins de 100 pb à plusieurs dizaines de kb, tandis que les exons semblent eux relativement fixes autour de 200 pb. L'épissage des introns est réalisé par un complexe ribonucléoprotéique (RNP), le spliceosome, qui fait intervenir des protéines couplées à des ARN (Will & Lührmann 2011). Le spliceosome est composé de 5 sous-unités RNP différentes, les snRNP (small nuclear RNP) U1, U2, U4, U5 et U6, associées à de très nombreux cofacteurs. L'ensemble atteint plusieurs milliers de kDa. Chaque snRNP contient un snARN particulier, 7 protéines dites Sm et de 7 à 21 protéines autres. Les snRNA sont de petits ARN de moins de 250 pb, cappés mais non polyadénylés et transcrits par l'ARN polymérase II à partir d'unités de transcription similaires aux gènes codant les protéines. Ces ARN vont être exportés dans le cytoplasme par un complexe d'exportation, où ils seront maturés et assemblés avec les protéines Sm pour former les sous-unités RNP du spliceosome. Le Spliceosome va ensuite retourner dans le noyau pour réaliser l'épissage. Il reconnaît les introns grâce à plusieurs séquences consensus situées à l'intérieur des introns : le 5'SS (splicing site) : GU, le 3'SS : AG et le BPS (Branch

Point Sequence). Les introns sont clivés par deux réactions successives de transestérification **(Figure 15)**. En plus du spliceosome, l'épissage requiert plusieurs complexes protéiques : NTC, le complexe Prp19/CDC5L ; EJC, le complexe d'assemblage des exons ; CBP, le complexe de fixation de la coiffe ; RES, le complexe de rétention et d'épissage ; TREX, le complexe d'export.

c) Les Facteurs de transcription et le complexe mediator

Selon le modèle précédent où la transcription est déclenchée par la fixation des GTFs sur des séquences consensus, la présence de ces séquences serait nécessaire et suffisante pour déclencher l'événement de transcription. De nombreuses études ont donc été menées pour rechercher ces promoteurs basaux dans le génome et en déduire ainsi la liste des gènes potentiels. Cependant, ces prédictions se sont révélées en grande partie fausses car le promoteur basal ne suffit généralement pas à déclencher la transcription. De nombreux autres facteurs sont nécessaires, appelés les facteurs de transcriptions (FTs), qui sont capables de moduler la transcription des gènes. Pendant longtemps, les biologistes pensaient que chaque gène était transcrit de façon immuable dans le temps et l'espace. D'un autre côté, ils se demandaient comment un organisme découlant d'une seule cellule initiale (l'embryon) pouvait posséder autant de types cellulaires différents. On sait maintenant que l'expression de chaque gène est finement contrôlée au niveau spatio-temporel notamment grâce à une combinaison de différents FTs. Chaque type cellulaire exprime une combinaison spécifique de FTs qui sera responsable de son identité.

Les FTs sont des protéines qui possèdent au moins un domaine de liaison à l'ADN capable de fixer une séquence consensus (élément de réponse ou BS pour Binding Site) située à une distance plus ou moins grande du TSS. L'ensemble des éléments de réponse d'un gène forme le promoteur dit distal. Les FTs contiennent également un Trans-Activating Domaine (TAD) qui est capable de moduler l'activité transcriptionnelle du gène ciblé par divers mécanismes. Tout d'abord, certains FTs ont la capacité de stabiliser le complexe d'initiation de la transcription. Ensuite, de nombreux FTs sont capables de modifier l'acétylation des Histones, soit parce qu'ils possèdent une activité HAT ou HDAC (Histone DéACétylase), soit parce qu'ils vont recruter des protéines qui possèdent ces activités. Ou encore, les FTs peuvent également recruter divers coactivateurs ou corépresseurs. Enfin, un FT peut tout simplement bloquer l'accès du BS à un autre FT et donc l'empêcher de fonctionner (Gill 2001)

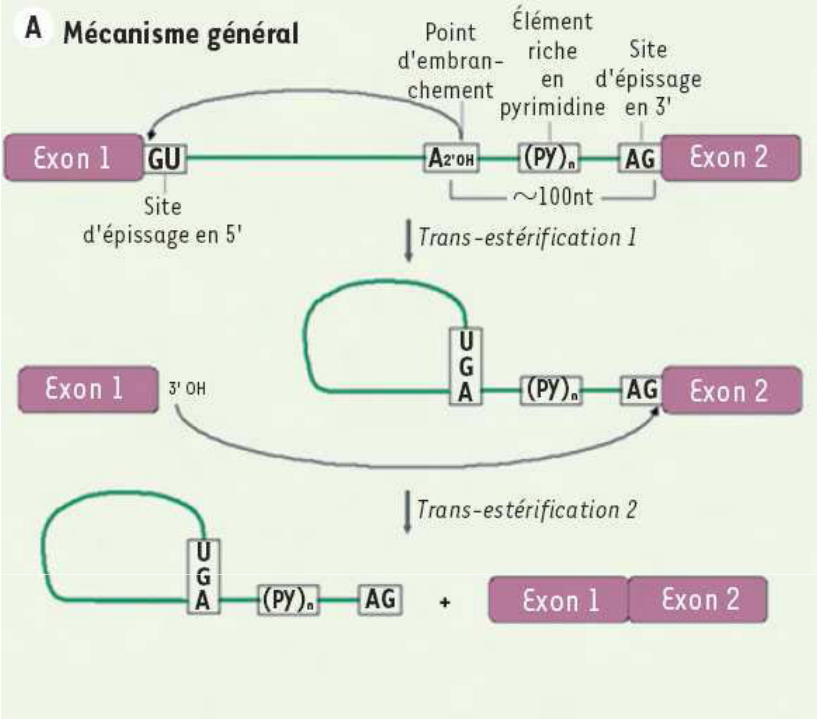


Figure 15 : Principe de l'épissage

L'épissage est dirigé par plusieurs séquences consensus : un site d'épissage en 5' et en 3' de l'intron ainsi qu'un point d'embranchement et une séquence riche en base pyrimidique à l'intérieur de l'intron. Deux réactions de trans-estérification sont responsables du détachement de l'intron sous forme d'un lasso.

[http://id.erudit.org/iderudit/013504ar]

Les FTs vont ainsi réguler l'expression des gènes qui ne doivent être exprimés qu'à certains moments, comme par exemple les gènes homéotiques ou, de façon plus large, ceux qui participent au développement de l'organisme, les gènes du cycle cellulaire, de l'homéostasie, etc. Les FTs vont également permettre à la cellule de répondre à divers stimuli environnementaux (choc thermique par exemple) ou à des infections cellulaires. Afin de pouvoir réguler les gènes au moment opportun, les FTs sont eux-mêmes soumis à une régulation très stricte de leur activité. En effet, la majorité des FTs sont activables par des modifications post-traductionnelles telles que la phosphorylation ou la fixation d'une sous-unité activatrice. De même, la transcription des FTs est elle-même souvent contrôlée par d'autres FTs. L'activation des FTs est contrôlée par diverses voies de signalisation cellulaire qui permettent de transduire un signal extérieur à la cellule jusqu'au noyau.

Plusieurs FTs sont souvent nécessaires pour permettre une expression correcte d'un gène. L'ensemble de ces FTs va interagir physiquement avec un complexe appelé Mediator (MED), une structure protéique composée d'une trentaine de sous-unités qui peut atteindre plusieurs milliers de kDa. MED forme une sorte de plateforme fixé au complexe d'initiation de la transcription et sur laquelle peuvent s'arrimer l'ensemble des FTs régulant un gène grâce à ses nombreux domaines d'interaction **(Figure 16)**. Cet arrimage est rendu possible par le rapprochement physique des séquences génomiques contenant les promoteurs et les éléments de réponse où sont fixés les FTs. MED est alors capable d'intégrer l'ensemble des signaux provenant des FTs régulant un même gène. Cet ensemble sera communiqué au complexe d'initiation de la transcription sous forme d'un signal global activateur ou inhibiteur de la transcription (Malik & Roeder 2010).

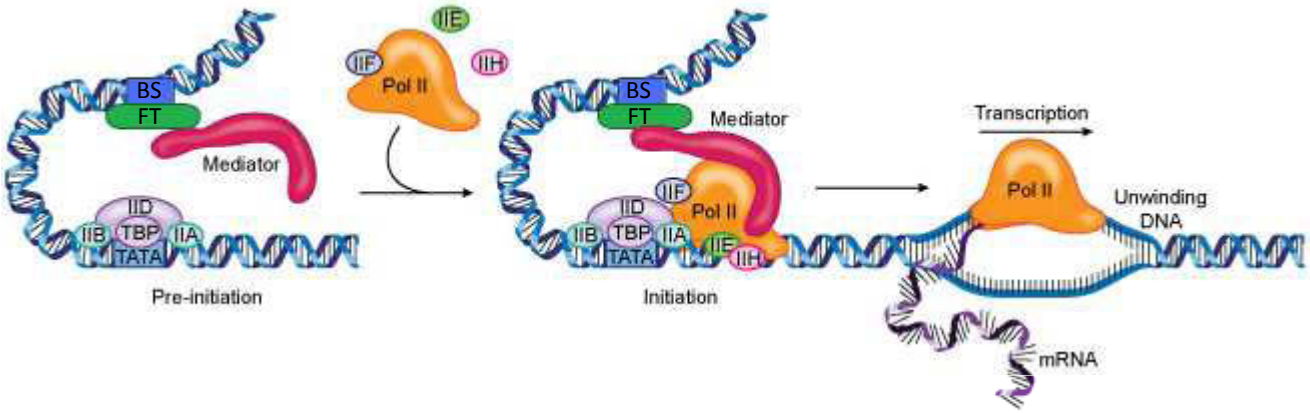


Figure 16 : Régulation de la transcription par un facteur de transcription

Les facteurs de transcription (FT) sont fixés à leur site de fixation (BS, Binding Site) et recrutent le complexe Mediator qui interagit avec l'ARN polymérase II et retransmet le signal activateur ou inhibiteur.

[https://mutagenetix.utsouthwestern.edu/phenotypic/phenotypic_rec.cfm?pk=399]

II. Les éléments transposables, des séquences ADN régulatrices et régulées.

A. Les éléments transposables

1. Découverte des éléments transposables.

a) Les travaux de McClintock sur le maïs

Dans les années 1940, Barbara McClintock, cytogénéticienne au laboratoire Cold Spring Harbor, travaillait sur le génome du maïs. Grâce à des techniques de coloration très précises qu'elle avait mises au point, elle s'intéressait aux mécanismes de cassure des chromosomes. En 1948, elle émit une hypothèse originale pouvant expliquer la mosaïque de couleur observée sur l'aleurone du grain de mais, caractéristique phénotypique contrôlée par le gène aleurone-color gene situé sur le bras court du chromosome 9 du maïs **(Figure 17)**. Ces résultats montrèrent que ce gène contient un élément génétique indépendant qui empêche son expression, appelé Dissociator (Ds), et que la présence ailleurs dans le génome d'un autre gène, appelé Activator (Ac), entraine le déplacement de Ds, permettant ainsi de recouvrer l'expression de l'aleurone-color gene (McCLINTOCK 1950). Ces résultats nécessitaient d'admettre deux faits révolutionnaires pour la biologie contemporaine : i) L'expression des gènes peut être modulée, ii) Il existe des gènes capables de se déplacer à l'intérieur du génome, appelés éléments transposables (ETs).

Cette hypothèse reçut un accueil froid de la communauté scientifique internationale, qualifié de « perplexe, voire hostile » par McClintock elle-même (Ravindran 2012). Il fallut attendre des années pour que d'autres scientifiques vinrent corroborer ses résultats. Finalement, la communauté scientifique comprit enfin la portée de la découverte de McClintock et lui attribua le prix Nobel de Physiologie ou de Médecine en 1983.

b) La conquête des autres organismes

En 1969, James Alan Shapiro, qui réalisait un post-doctorat à l'université Pasteur, décrivit l'apparition spontanée d'une mutation dans l'opéron lactose d'*E. coli*. Les analyses qu'il mena le firent admettre que « l'explication la plus simple de l'apparition de cette mutation est l'insertion d'une séquence étrangère au sein de l'opéron » (Shapiro 1969).

Figure 17 : Mosaïcisme de la coloration du grain de maïs.

Epi de maïs présentant un mosaïcisme au niveau de l'expression du gène de coloration de l'aleurone du grain, dû à la mobilisation de l'élément transposable dissociator.

[http://fr.wikipedia.org/wiki/Barbara_McClintock]

Shapiro venait de découvrir les séquences IS, la forme la plus simple d'ETs que l'on trouve généralement chez les procaryotes. Ce fut la deuxième preuve d'existence des ETs. Par la suite de très nombreux exemples d'ETs furent décrits chez les procaryotes.

En 1976, Georges Picard, un généticien de l'université de Clermont Ferrand décrivit l'apparition de stérilité dans la descendance femelle de croisements effectués entre différentes souches de drosophile, du type Inducer (I) ou Reactive (R) (Picard 1976). Cette incompatibilité entre les souches fut appelée dysgénésie des hybrides **(Figure 18)**. De façon curieuse, la stérilité apparaît uniquement lorsqu'on croise une femelle R avec un mâle I et elle est due à la présence d'un élément chromosomique particulier appelé le Facteur I. Des études complémentaires ont montré que ce facteur était un ET qui avait envahi la souche I mais n'était pas présent dans la souche R. Ce fut le début d'une longue histoire de l'étude des ETs à Clermont-Ferrand, qui se poursuit encore aujourd'hui. En 1977, une autre équipe montra un exemple similaire de dysgénésie des hybrides chez la drosophile associée à un autre élément transposable, l'élément P (Kidwell et al. 1977).

Suite à la découverte des ETs chez les procaryotes et chez certains métazoaires, la communauté scientifique commença à reconsidérer certains résultats et découvrit que les ETs étaient également présent chez les mammifères (Whitney & Lamoreux 1982). A ce jour, les ETs ont été trouvés dans tous ou presque tous les organismes séquencés à ce jour et représentent un champ de recherche extrêmement prolifique. Des centaines d'équipes à travers le monde étudient les ETs, sur des organismes aussi variés que le melon et la grenouille *Rana esculenta,* en passant par la paramécie.

2. Structures et mécanismes de transposition

Depuis leur découverte, de nombreux types d'ETs ont été découverts. S'ils partagent la caractéristique de se déplacer à l'intérieur des génomes, les structures et les mécanismes de transposition varient fortement, ce qui rend difficile l'identification automatisée des ETs lors de l'annotation des génomes. Un système de classification a progressivement été mis en place jusqu'à la proposition de Wicker, en 2007, basée sur la structure, le fonctionnement et l'évolution des ETs (Wicker et al. 2007). Cette classification est partagée en deux grandes classes : les rétrotransposons et les transposons à ADN.

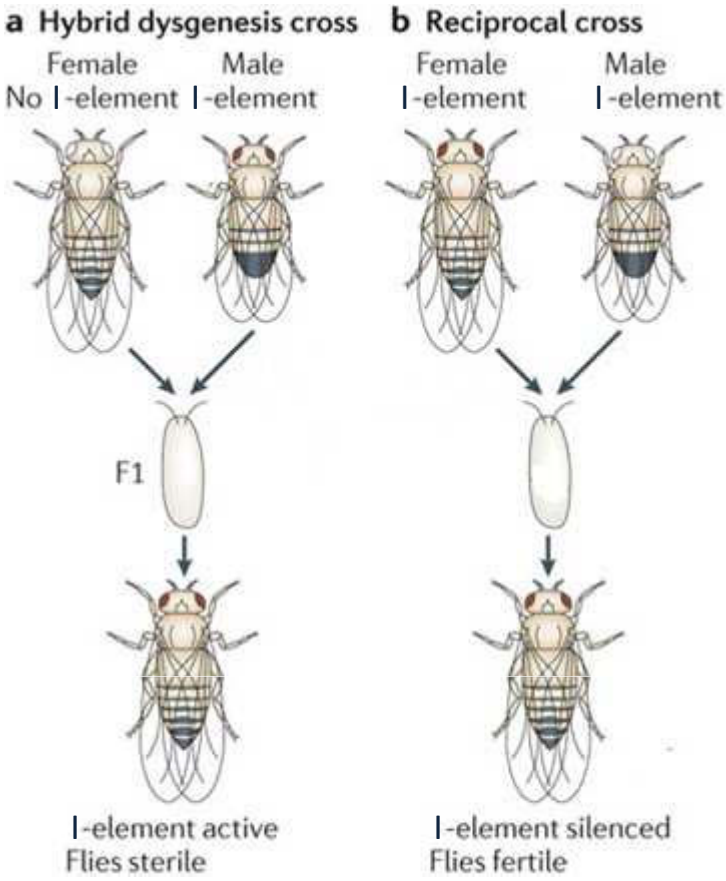


Figure 18 : La dysgénésie des hybrides

Lorsque deux lignées différentes de drosophile sont croisées, l'une contenant l'élément transposable I et l'autre non, la descendance peut être de deux phénotypes différents. Dans le cas (a), c'est le père qui contient l'élément I. La descendance sera caractérisée par une dérégulation de l'élément I et une stérilité. Dans le cas (b), la mère contient l'élément I. La descendance sera alors fertile et l'élément I régulé.

[Adapté de Siomi *et al*, 2011]

a) Les ETs de type I : les rétrotransposons

Les rétrotransposons ont la caractéristique d'utiliser un intermédiaire ARN qui sera rétrotranscrit en ADN puis intégré dans le génome. Ceci leur permet de transposer sur un mode « copier/coller », également appelé transposition réplicative, et donc de se dupliquer à chaque événement de transposition, expliquant la forte représentation de cette catégorie dans les génomes, en particulier chez les plantes. On distingue deux catégories de rétrotransposons : les rétrotransposons à LTR et ceux qui n'en possèdent pas **(Figure 19)**.

Les rétrotransposons à LTR (Long-Terminal Repeat) contiennent deux séquences répétées à leurs extrémités, d'une taille variant de plusieurs centaines de pb à plus de 5 kb. Leur structure est extrêmement similaire à celle des rétrovirus ce qui suggère une origine commune. La séquence interne contient au minimum deux cadres de lectures : *gag*, qui code pour une protéine de structure permettant de former des particules similaires aux particules virales et *pol*, qui code pour toutes les protéines nécessaires au mécanisme de transposition qui sont : une protéase aspartique, une transcriptase inverse (qui permet de rétrotranscrire l'ARN), une RNase H (qui dégrade la copie ARN), et une intégrase (qui intègre la néo-copie ADN dans le génome). Certains rétrotransposons possèdent également un troisième cadre de lecture, *env*, qui code pour des glycoprotéines transmembranaires impliquées dans l'infection de cellule à cellule (Leblanc et al. 1997; Pélisson et al. 1994). L'intégration de la copie ADN dans le génome entraine une duplication du site d'intégration de 4 à 6 pb, caractéristique utilisée pour repérer ces éléments dans le génome.

Les rétrotransposons sans LTR ne possèdent pas d'extrémités répétées mais possèdent une séquence riche en adénosine à leur extrémité 3'. Il en existe 4 catégories : les DIRS-like, les Penelope-like, les LINEs et les SINEs. Les LINEs (Long Interpersed Nuclear Element) sont retrouvés dans la quasi-totalité du monde vivant. Le facteur I découvert par G. Picard fait partie de cette catégorie. Ces éléments contiennent au moins un cadre de lecture codant pour les différentes protéines nécessaires à leur rétrotransposition, et parfois un deuxième situé en 5' des éléments et dont le rôle reste encore à définir. Les LINEs sont également responsables de la mobilisation des SINEs (Short Interpersed Nuclear Element), des rétrotransposons ne possédant pas de transcriptase inverse.

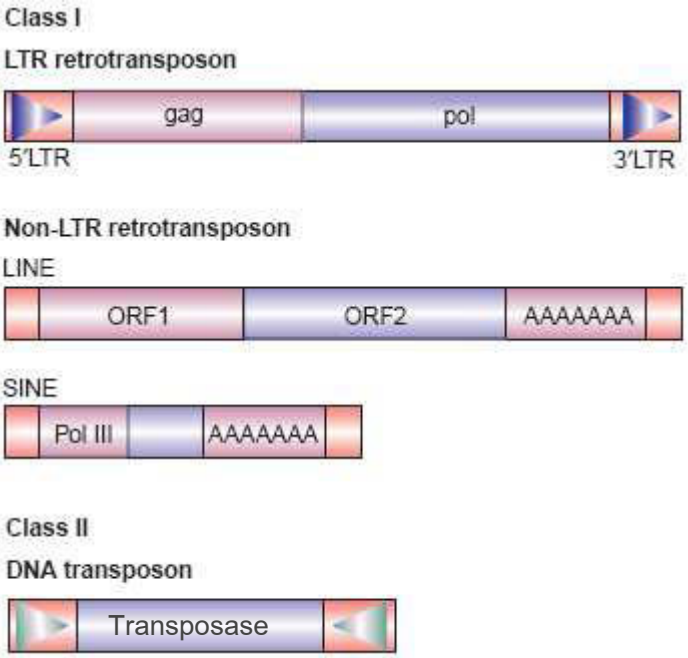


Figure 19 : Structure des deux classes d'ETs

La classe I correspond aux rétrotransposons et se compose d'éléments à LTR ou sans LTR. Certains ETs à LTR contiennent également un cadre de lecture similaire à *env* des rétrovirus. La classe II correspond aux ETs à ADN.

[Adapté de Wong *et al*, 2004]

b) Les ETs de type II : les transposons à ADN

Les ETs à ADN, ou transposons, transposent directement sous une forme ADN. Une majorité de ces ETs ne copie pas leur séquence lors de la transposition mais utilise plutôt le principe du « couper-coller », ou transposition conservative **(Figure 20)**. C'est-à-dire que ces ETs vont s'exciser de leur locus d'insertion et se réintégrer ailleurs dans le génome. L'augmentation du nombre de copies de ces éléments n'est donc généralement pas due à leur mobilisation mais à l'addition de mécanismes cellulaires tels que la réplication du génome ou la recombinaison homologue lors des réparations des cassures doubles brins. Le système Ac/Ds découvert par McClintock ainsi que l'élément P de la drosophile font partie de cette catégorie d'ETs.

La structure la plus simple de ces éléments est constituée d'un cadre de lecture codant pour une protéine transposase entouré de séquences ITR (Internal Terminal Repeats). C'est le cas des séquences IS découvertes par Shapiro. La transposase a la capacité de s'associer aux ITR en reconnaissant des séquences cibles et de couper le brin d'ADN afin de permettre à la fois l'excision de l'ET mais également sa réinsertion. Une autre catégorie d'ETs à ADN transpose sous la forme d'ADN tout en ayant la capacité de se multiplier. Cette catégorie, qui contient notamment les ETs du type Hélitron, transpose par un système de cercle circulant où seulement un brin d'ADN est coupé au locus d'excision et d'insertion. Le brin manquant sera reconstitué par la cellule à partir du brin restant, dupliquant ainsi la séquence de base.

c) Fréquence de transposition

Si les génomes sont tous plus ou moins envahis par les ETs, les événements de transposition restent rares car la très grande majorité des ETs est devenue défective à cause des nombreuses mutations les affectant. Toutefois on estime la fréquence de transposition des ETs à 10^{-6} événements par génération et par élément. De façon intéressante, de nombreuses études montrent une augmentation de l'activité des ETs lorsque l'organisme hôte subit un stress. Ainsi, certaines études ont montré que la répression exercée sur les ETs est relâchée lors de stress environnementaux, tels qu'un choc thermique ou hydrique chez les plantes (Vasilyeva et al. 1999). Ce relâchement entraine une remobilisation des ETs ce qui génère des mutations permettant une évolution des organismes qui aboutirait à l'émergence d'une nouvelle espèce plus résistante aux nouvelles conditions environnementales.

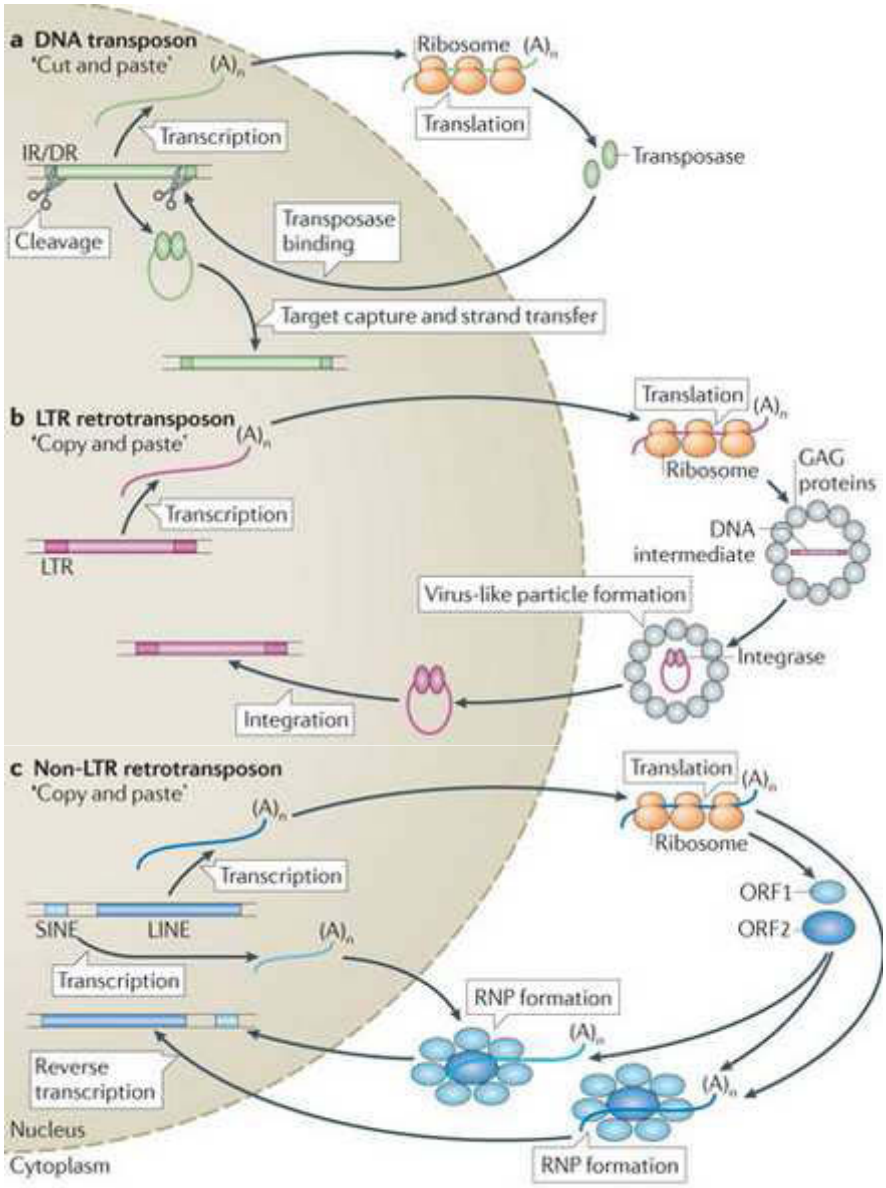


Figure 20 : Mode de transposition des ETs

Les transposons à ADN transposent selon un mode « couper-coller » grâce à une transposase. Les rétrotransposons transposent par l'intermédiaire d'un ARN selon un mode « copier-coller ».

[Siomi *et al*, 2011]

D’autres études ont également montré que les ETs s’inséraient parfois préférentiellement dans certains sites génomiques. Par exemple, le rétrotransposon *gypsy* que l’on retrouve chez la drosophile s’insère fréquemment dans deux gènes, *ovo* et *cut*, situés dans des parties euchromatiques de chromosome (Dej et al. 1998). Une autre stratégie adoptée par les ETs consiste à contrôler le tissu où aura lieu la transposition. L’élément P est en cela un exemple intéressant. L’élément P code pour une transposase de 87 kDa. Toutefois, dans les cellules somatiques, un épissage particulier résulte en la production d’une protéine de 66 kDa répressive de la transposition qui empêche la transposition dans ces tissus (Nouaud & Anxolabéhère 1997). La transposase dans son intégrité est exprimée spécifiquement dans les cellules germinales de la drosophile. Ceci permet à l’élément P de transposer uniquement dans des séquences ADN qui seront transmises à la descendance.

3. Impact sur les génomes

a) Proportion dans les génomes

L’importance des ETs dans la structure des génomes n’a été réellement comprise que lors des séquençages génomiques haut-débit réalisés dans les années 2000. En effet, les ETs ont alors été retrouvés dans quasi toutes les espèces séquencées à ce jour, dans des proportions allant de très faible chez les procaryotes (0,1%) à très forte (>80%) **(Figure 21)**. Les végétaux possèdent souvent les génomes les plus riches en ETs (jusqu’à 85% chez certaines espèces de maïs ou de blé) même si, chez les métazoaires, les génomes des amphibiens sont également riches en ETs (77% pour la grenouille *Rana esculenta*). Chez l’humain, on estime à 45% la part des ETs dans le génome et la proportion atteint 18% chez *Drosophila melanogaster* (Biémont & Vieira 2006).

Certaines études ont montré qu’une trop forte proportion en ETs dans un génome pouvait être néfaste pour l’organisme, notamment chez la drosophile où l’augmentation du nombre d’insertion d’ETs serait corrélée à une baisse du taux d’éclosion des œufs (Pasyukova et al. 2004). Toutefois, plusieurs études ont récemment montré qu’une forte présence d’ETs générerait des organismes ayant une plus forte capacité d’adaptation (Oliver & Greene 2011).

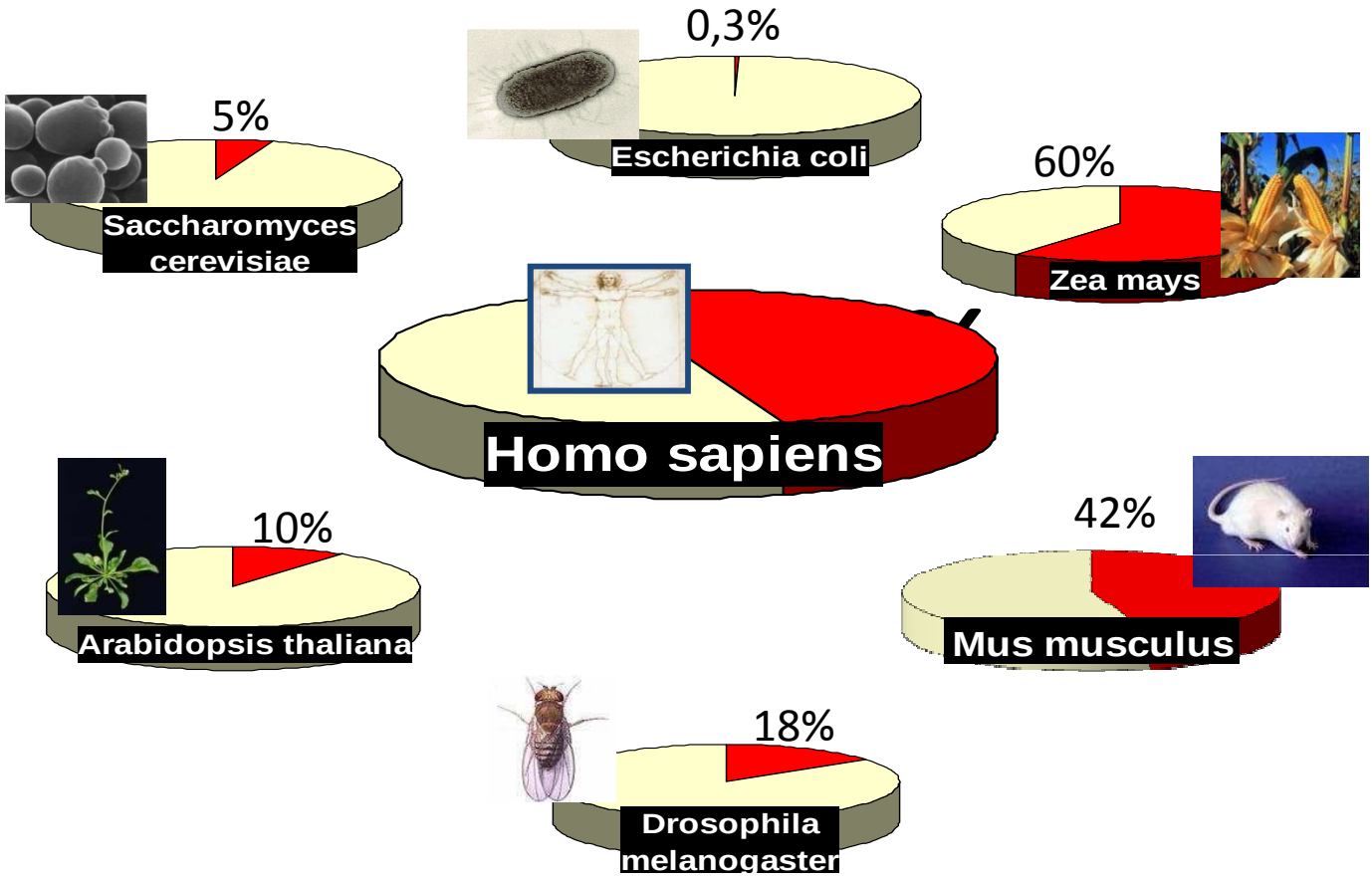


Figure 21 : Proportion des ETs dans différents génomes

Les ETs sont retrouvés dans tous les organismes mais dans des proportions différentes

[Adapté de C. Biémont]

b) Instabilités génétiques

De par leur capacité à se déplacer, les ETs sont intrinsèquement vecteurs d'instabilités génétiques. C'est d'ailleurs grâce à cette propriété que McClintock les a découvert. Un ET qui s'insère dans ou à proximité d'un gène peut i) Perturber le cadre de lecture entrainant la formation d'un allèle nul, ii) Apporter des signes d'épissages ou un nouveau promoteur qui permettront l'apparition de nouveaux exons, entrainant l'émergence d'un nouvel allèle du gène, iii) Isoler un gène de ses séquences régulatrices par la présence de séquences insulatrices dans les ETs, iv) Entrainer le ciblage des voies de répression des ETs sur ce gène et donc réprimer son expression (Bourque 2009) **(Figure 22)**. Ces diverses conséquences ont de très nombreuses répercutions sur l'expression génique des organismes hôtes. Ainsi, on connaît maintenant de nombreuses maladies humaines qui sont dues à la perturbation d'un gène par une néo-insertion d'ET. Chez l'humain, 65 maladies héréditaires causées par les ETs ont été recensées, comme certains cas d'hémophilie ou de myopathie de Duchesne. De même, l'apparition de cancer a pu être reliée à la mobilisation des ETs (Morse et al. 1988). Toutefois, de nombreuses voies s'élèvent pour défendre l'idée que les ETs sont également des collaborateurs très précieux pour l'établissement des réseaux complexes de transcription des gènes, en apportant de très nombreux promoteurs et séquences régulatrices. Il a été montré chez la souris et l'humain que 18,1% et 31,4% des TSS, respectivement, sont issus d'un ET (Faulkner et al. 2009). De plus, la plupart de ces TSS sont caractérisés par une expression tissu-spécifique.

En plus de ces effets directs sur l'expression génique, les ETs influent également beaucoup sur la structure des génomes. En effet, la présence de séquences répétées à l'intérieur des génomes, que ce soient des ETs ou non, est source de recombinaisons ectopiques, notamment lors du fonctionnement de la voie de réparation homologue (Bennetzen 2005). Ce mécanisme, qui nécessite la juxtaposition des deux séquences alléliques du génome, peut engendrer des recombinaisons entre deux ETs situés à diverses locations dans le génome. Ce genre de recombinaison va alors provoquer de gros réarrangements chromosomiques tels que des délétions ou duplications de grandes tailles, mais également des inversions chromosomiques.

Enfin, la présence dans le noyau de nombreuses protéines possédant une activité endonucléase, et donc capable de générer des cassures doubles brins, est un danger pour la stabilité du génome. Il a d'ailleurs été montré que la transposase du rétrotransposon L1

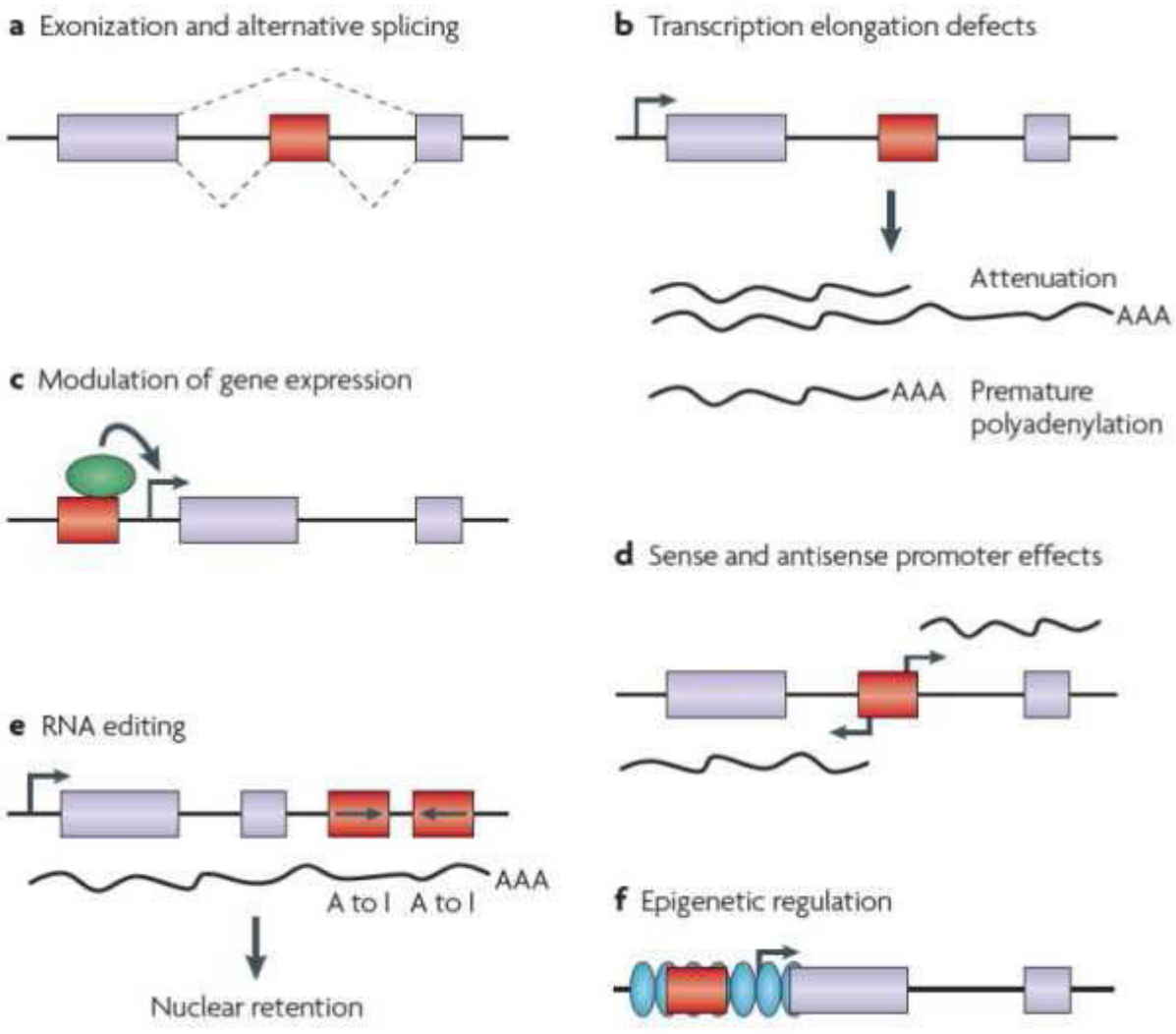


Figure 22 : Impact des ETs sur l'expression des gènes

Les ETs (en rouge) sont capables de perturber l'expression des gènes par plusieurs mécanismes. En effet ils sont responsables de : (a) L'apport de nouveaux sites d'épissages. (b) La perturbation de la terminaison de la transcription (c) L'apport de nouveaux sites de régulation de l'expression génique. (d) L'apport de nouveaux promoteurs. (e) L'apport de nouveaux signaux d'édition. (f) Le ciblage par des mécanismes de régulation épigénétique.

[Cordaux R and Batzer MA, 2009]

générait plus de cassures doubles brins que d'événements d'insertions de L1 dans une culture de cellules Hela (Gasior et al. 2006). Ces cassures double brins peuvent être sources de mutations, notamment lors de la réparation par le mécanisme NHEJ (Non Homologous End Joining) qui génère des mutations de petite taille.

c) Domestication

La cohabitation entre les ETs et les hôtes donne lieu à des exemples de domestication de la part des organismes hôtes dont certains cas sont très bien documentés. Par exemple, on peut citer la drosophile qui utilise deux ETs pour assurer les fonctions télomériques. Chez les eucaryotes, les télomères sont généralement constitués d'une petite séquence de 2 à 8 nucléotides qui est répétée un très grand nombre de fois à chaque extrémité chromosomique. Chez la drosophile, il a été montré que les télomères sont constitués de séquences répétées en tandem étrangement longues, 14,5 kb, qui sont en fait deux ETs, *Het-A* et *Tart* (Levis et al. 1993). Cette situation originale a été retrouvée au moins une fois chez un autre eucaryote, *Giardia lamblia*, un protozoaire flagellé qui utilise *GilM* et *GilT*, deux rétrotransposons, comme structure télomérique (Gladyshev & Arkhipova 2007).

Chez les vertébrés, le système immunitaire adaptatif repose sur la production de lymphocytes possédant des récepteurs membranaires spécifiques à un antigène donné. L'incroyable variété de récepteurs nécessaires est rendue possible grâce à un système de recombinaison des gènes nécessaires à la fabrication de ces récepteurs, le système de recombinaison V(D)J. Or il a été prouvé que cette recombinaison utilise deux protéines, RAG1 et RAG2, qui forment en s'hétérodimérisant, une structure proche des transposases et qui va reconnaître des séquences spécifiques pour couper l'ADN. Les protéines RAG1 et RAG2 proviendraient d'un ET devenu immobile (Kapitonov & Jurka 2005).

d) Spéciation et moteurs évolutifs

Malgré tous les risques d'instabilités engendrés par les ETs, ils semblent avoir colonisé tous les génomes et ce dans des proportions importantes. De plus en plus, les études montrent que ce sont des constituants fonctionnels du génome. C'est pourquoi de plus en plus de scientifiques s'accordent pour dire que les ETs ne seraient pas que des « parasites » de l'ADN, ou de « l'ADN poubelle », mais qu'ils contribueraient fortement au fonctionnement

des organismes et notamment au niveau évolutif (Piskurek & Jackson 2012). On les appelle les moteurs de l'évolution. Les ETs pourraient contribuer aux phénomènes de spéciation par plusieurs mécanismes. Tout d'abord, les ETs sont à l'origine de mutations qui peuvent, à terme, aboutir à l'apparition d'une nouvelle espèce. Mais les ETs peuvent également provoquer des incompatibilités génétiques de façon directe, notamment dans les cas de dysgénésie des hybrides où la présence d'un ET dans une lignée d'une espèce empêche son croisement avec une autre lignée qui en est dépourvue.

De façon générale, il semble que si les individus peuvent subir les conséquences parfois fatales de la mobilisation des ETs, l'ensemble de l'espèce est favorisé par la présence d'ETs actifs dans son génome qui sont un outil majeur de l'évolution des espèces. Toutefois, la mobilisation des ETs doit être finement régulée afin que la relation entre les ETs et leurs hôtes soit optimisée. Pour cela, les organismes ont mis en place des mécanismes de répression des ETs qui utilisent notamment les voies d'ARN interférence.

B. Les ETs sont régulés par ARN interférence.

1. Les différentes voies d'ARN interférence

En 1990, l'équipe de Jorgensen en Californie a voulu augmenter la couleur pourpre des fleurs de pétunia en introduisant dans ces plantes un transgène contenant plusieurs copies du gène responsable de la coloration de la fleur (Napoli et al. 1990). De façon surprenante, non seulement les fleurs ne sont pas devenues plus colorées mais certaines sont même devenues blanches **(Figure 23)**. Le transgène avait donc la capacité de réprimer en trans le gène endogène. Cet effet ne fut vraiment compris qu'en 1998, grâce aux travaux de Fire et Mello (prix Nobel de Physiologie ou Médecine en 2006), où ils montrèrent que l'introduction d'un ARN double brin (db) dans une cellule entrainait la répression de l'ARN simple brin (sb) complémentaire chez *C. elegans* (Fire et al. 1998). Ce phénomène fut appelé l'interférence par ARN ou ARNi et fut rapidement retrouvé chez tous les organismes. Le principe de l'ARNi est de combiner un ARN de petite taille (de 20 à 30 nucléotides) avec une protéine de la famille Argonaute. Ce complexe est appelé RISC pour RNA Induced Silencing Complex. Depuis, de nombreuses études ont montré que l'ARNi est un mécanisme universel divisé en 3 voies, les siRNA, les miRNA et les piRNA, qui diffèrent par l'origine et la structure des petits ARN utilisés, les protéines Argonaute impliquées et leurs fonctions.

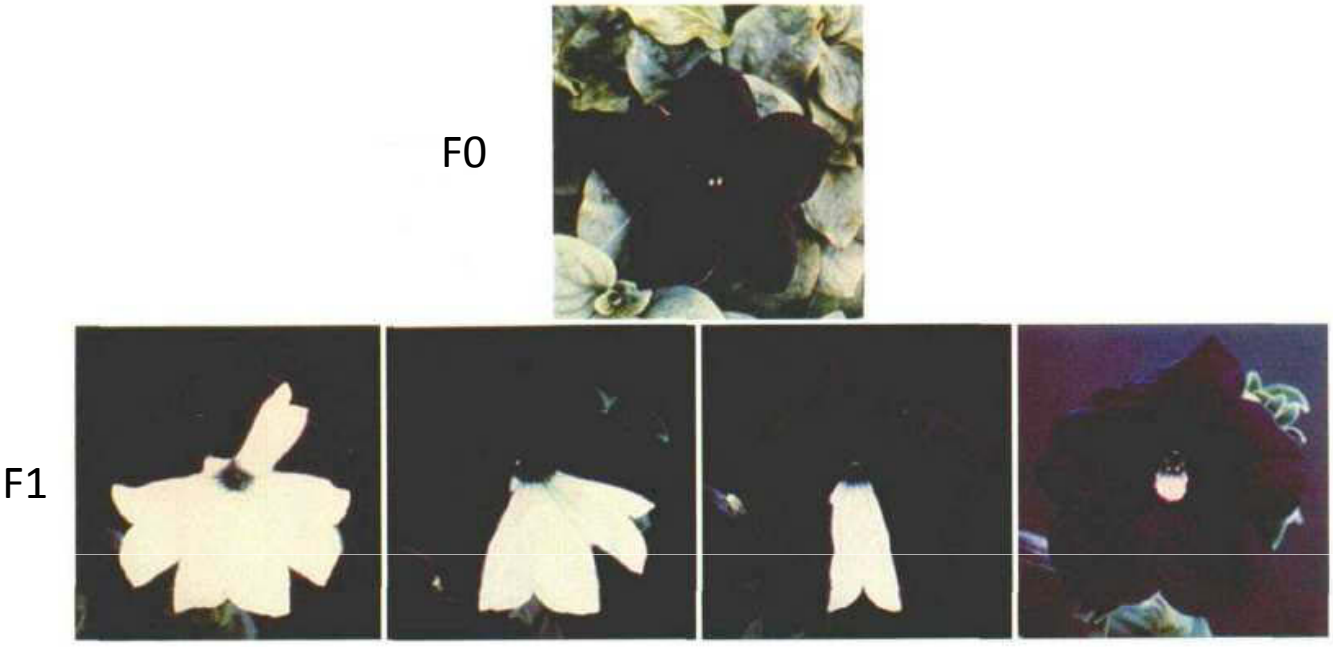


Figure 23 : Inhibition d'un gène endogène par introduction d'un transgène chez le pétunia

La photo du dessus représente le phénotype parental d'une fleur de pétunia (F0). En introduisant un transgène qui contient plusieurs copies du gène responsable de la couleur pourpre, les expérimentateurs ont obtenu une descendance dans laquelle ce gène était réprimé dans certaines cellules (F1). Ce phénomène fut d'abord appelé co-suppression avant de parler d'ARN interférence. [Napoli *et al*, 1990]

a) Les siRNA

Les siRNA sont maturés à partir d'une structure en ARN db qui peut provenir de plusieurs mécanismes. L'une des principales sources de siRNA correspond à l'expression de séquences géniques extérieures aux génomes, tels les virus et les transgènes. Leurs ARN vont s'hybrider avec des ARN de séquences complémentaires préalablement produits par la cellule et former ainsi une structure db qui sera reconnue et maturée en siRNA. Ce mécanisme de défense adaptatif est un système immunitaire très important chez les plantes et les insectes et il a été récemment observé chez les mammifères (Maillard et al. 2013). Ainsi, les siRNA sont considérés comme des défenseurs du génome car ils sont capables de réprimer les infections virales mais également de défendre le génome contre les séquences parasites venues de l'intérieur comme les ETs. En plus de cette voie, les structures db peuvent également provenir d'ARN issus de la transcription de loci génomiques particuliers formant des transcrits bidirectionnels ou de pseudogène en orientation inverse (Ghildiyal & Zamore 2009).

Le long ARN db est reconnu par une protéine Dicer, qui possède un domaine de liaison à l'ARN db. Après avoir reconnu et fixé cet ARN, la protéine Dicer va le couper afin de produire des ARN db de 21 nucléotides, avec des extrémités sortantes de 2 nucléotides **(Figure 24)**. Le brin contenant l'extrémité 5' sortante est monophosporylé et servira de brin guide, tandis que l'autre est appelé brin complémentaire. Le dimère est introduit dans une protéine Argonaute (Ago 2 chez les drosophiles et Ago 1 à 4 chez les mammifères) avec l'intervention d'un complexe de protéines chaperonnes telles que Hsp90. L'activité slicer de la protéine Ago 2 va alors cliver la liaison phosphodiester entre les nucléotides 9 et 10 du brin complémentaire, ce qui va entrainer son éviction du complexe RISC (Matranga et al. 2005). Dans le cas d'un chargement dans une protéine Ago 1, 3 ou 4, qui ne possèdent pas d'activité slicer, la séparation des deux brins sera effectuée par un mécanisme indépendant de cette fonction, plus lent.

Une fois le complexe RISC formé, celui-ci va reconnaître des ARN cibles parfaitement complémentaires sur toute la longueur du siRNA guide et le dégrader progressivement grâce à l'activité slicer de la protéine Ago (Liu et al. 2004). Si la complémentarité n'est pas parfaite, ou si la protéine Ago ne possède pas d'activité slicer, la traduction de l'ARN ciblé va être inhibée de façon similaire à l'action des miRNA (voir ci-dessous). Chez les plantes, les vers et les champignons, un mécanisme d'amplification des siRNA a été mis en évidence. Dans ces organismes, les ARN clivés par un complexe RISC

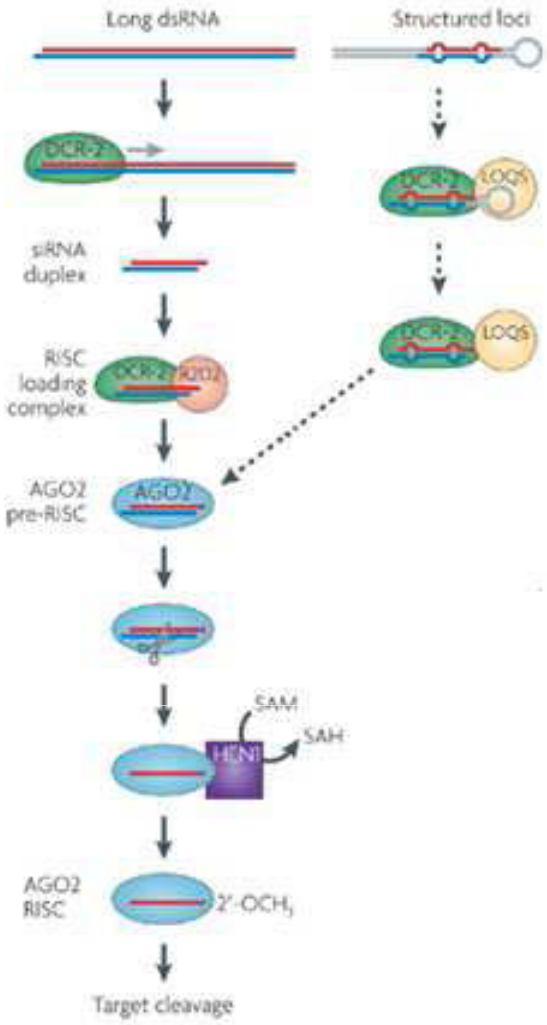


Figure 24 : La voie des siRNA chez la drosophile

La protéine Dicer-2 fixe et transforme les ARN précurseurs (double brin) db en duplex de siRNA, en partenariat avec la protéine R2D2 ou Loquacious (Loqs). Ces duplex sont chargés par la protéine Ago2 et l'un des deux brins sera détruit pour former le complexe RISC. Ce complexe ciblera les ARN parfaitement complémentaires pour les dégrader.

[Ghildiyal M and Zamore PD, 2009]

vont être recrutés par une ARN-dépendante-ARN polymérase (RdRP) qui va utiliser ces ARN comme matrice pour synthétiser un nouveau brin d'ARN (Pak & Fire 2007). Cette synthèse va former un nouveau long ARN db qui sera pris en charge par Dicer pour être maturé en nouveaux siRNA. Ce mécanisme permet de renforcer la répression exercée sur les ARN ciblés.

b) Les miRNA

En 1993, deux équipes découvrent chez C. elegans le gène lin-4 qui régule l'expression d'un autre gène, lin-14 (Wightman et al. 1993; Lee et al. 1993). L'originalité de cette découverte est due au fait que la régulation s'effectue par la production d'un petit ARN de 22 nucléotides, complémentaire en séquence au gène lin-14. Ce mécanisme fait bien évidemment penser à un système d'ARN interférence, et ce n'est que plus tard que l'on se rendit compte qu'une nouvelle voie ubiquitaire d'ARN interférence avait été mise en évidence, la voie des miRNA. Cette voie est plus généralement impliquée dans la régulation homéostatique des mécanismes cellulaires. De nombreux miRNA sont impliqués dans des voies de signalisation et permettent de réguler finement certains gènes au cours du développement.

Les miRNA sont longs de 22 nucléotides et possèdent une extrémité 5' monophosphorylée. Si les siRNA sont retrouvés en grand nombre dans les cellules, le nombre de miRNA différents est limité et n'excède pas plusieurs centaines. Par contre, certains miRNA sont très fortement exprimés. Ils peuvent posséder leur propre unité de transcription ou provenir de la maturation particulière d'introns, voir de quelques exons (Cai et al. 2004). Dans tous les cas, les miRNA proviennent en grande majorité de structures db appelées pri-miRNA. Si quelques mécanismes alternatifs existent, la biogenèse des miRNA correspond à une succession d'étapes bien définies **(Figure 25)**. La présence de séquences complémentaires de 33 nucléotides séparées l'une de l'autre par une dizaine de nucléotides permet la formation de structures dites « en épingle à cheveux » par hybridation des séquences complémentaires. Cette structure particulière formant un ARN db va être reconnue par une protéine du type Drosha, possédant une activité RNAse III capable de séparer la structure en hairpin du reste de l'ARN (Lee et al. 2003). Cette nouvelle structure est appelée pré-miRNA et sera exportée dans le cytoplasme, par l'exportine 5, où elle sera reconnue et clivée par la protéine Dicer pour former un ARN db de 22 nucléotides avec des extrémités

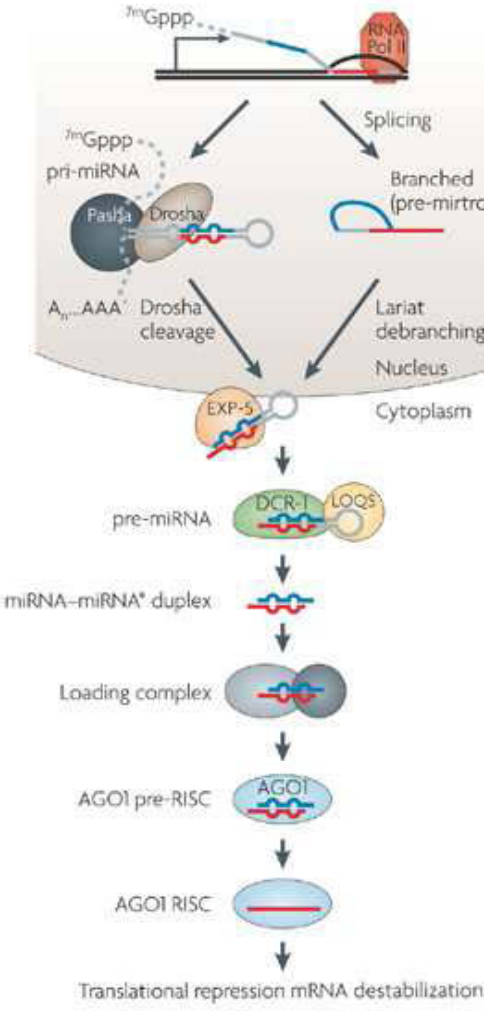


Figure 25 : La voie des miRNA chez la drosophile

Les miRNA proviennent d'unités de transcription du génome qui produisent une structure ARN particulière appelée pri-miRNA. Ce pri-miRNA est reconnu par la protéine Drosha qui le coupe pour libérer le pré-miRNA qui sera exporté dans le cytoplasme pour être pris en charge par Dicer-1. Le pré-miRNA sera alors découpé en duplex de miRNA et pris en charge par la protéine Ago1. L'un des deux brins sera détruit pour former le complexe RISC qui ciblera les ARN partiellement complémentaires afin d'inhiber leur traduction ou de diminuer leur demi-vie.

[Ghildiyal M and Zamore PD, 2009]

sortantes de 2 nucléotides, appelé miRNA/miRNA* duplex. Ce dimère est prise en charge selon un mécanisme ATP dépendant par une protéine Argonaute qui va expulser le brin le moins stable chimiquement afin de former un complexe RISC mature : une protéine Argonaute chargée d'un petit ARN sb (Kawamata et al. 2009).

Le complexe RISC des miRNA est localisé dans une structure cytoplasmique particulière, les P-bodies, grâce à une interaction directe entre la protéine Ago du complexe RISC et la protéine TNRC6 des P-bodies. TNRC6 va alors recruter deux complexes ayant des activités enzymatiques déadénylases, CCR4-NOT et PAN2-PAN3. La fixation du complexe RISC sur un ARNm va donc entrainer la suppression de la queue poly-A de cet ARNm entrainant ainsi sa dégradation (Behm-Ansmant et al. 2006). Le ciblage par un complexe RISC-miRNA permet aussi d'inhiber l'initiation de la traduction sans que les mécanismes moléculaires de cette inhibition soient connus.

Si les drosophiles possèdent deux protéines Dicer, Dcr1 et Dcr2, qui vont chacune agir respectivement sur les précurseurs des miRNA ou des siRNA, les autres métazoaires n'en possèdent qu'une. Cette unique protéine Dicer a une affinité plus forte pour les pré-miRNA que pour les longs ARN db. Cependant il a été rapporté que les cellules de l'ovocyte de la souris expriment une isoforme particulière qui ne possède plus de domaine hélicase et qui a une affinité très forte pour les longs ARN db (Flemr et al. 2013).

Si le nombre de miRNA est limité, le nombre de ses cibles semble lui disproportionné. En effet, la conformation particulière du miRNA à l'intérieur du complexe RISC fait que seuls les 6 à 7 premiers nucléotides de l'extrémité 5' sont strictement nécessaires à la reconnaissance de la cible (Lewis et al. 2003). Un appariement aspécifique peut se produire avec le reste de l'extrémité 3', ce qui génère un nombre impressionnant de cibles potentielles pour un miRNA donné. En fait, il semble que seuls quelques ARNm réagissent à la présence d'un miRNA par une variation significative et fonctionnelle de leur concentration. Une théorie originale d'Hervé Seitz propose que seuls ceux-là soient les cibles biologiques réelles des miRNA. Le reste des ARN visés aurait un rôle d'inhibiteurs compétitifs des miRNA (Seitz 2009). Dans cette théorie, les ARN régulés deviennent ainsi régulateurs, et vice-versa.

c) Les piRNA

Récemment, une nouvelle voie d'ARN interférence a été découverte spécifiquement dans les tissus reproducteurs des métazoaires (Brennecke et al. 2007; Aravin et al. 2006). Cette voie implique des protéines Argonaute de la sous-famille PIWI et des petits ARN appelés piRNA, pour PIWI-Interacting RNAs. Elle est largement impliquée dans la régulation des ETs, régulation particulièrement importante dans ces tissus (Aravin et al. 2007). En effet, la lignée germinale est le seul type cellulaire à transmettre son génome à la descendance. Il est donc primordial pour les espèces de préserver au mieux l'intégrité génomique de ces cellules et d'y contrôler finement la mobilisation des ETs.

Découverte depuis peu, la voie des piRNA est moins bien caractérisée que la voie des siRNA et des miRNA et beaucoup de zones d'ombres subsistent, d'autant plus que cette voie présente de nombreuses particularités : i) la maturation des piRNA est indépendante des protéines Dicer, ce qui suggère qu'ils ne proviennent pas d'un précurseur db mais sb, ii) les piRNA sont des petits ARN de 24 à 30 nucléotides, soit légèrement plus long que les si- et les miRNA, iii) alors que le nombre de miRNA à l'intérieur d'une espèce dépasse rarement les centaines, il est apparu dès les premiers séquençages, que les piRNA sont extrêmement complexes et divers (plusieurs centaines de milliers de séquences différentes), iv) contrairement à certains siRNA, il n'a pas été observé de clivage en phase des précurseurs sb, et v) les piRNA interagissent avec une sous-famille de protéines Argonautes, la sous-famille PIWI, constituée de trois éléments : Piwi, Aubergine (Aub) et Argonaute 3 (Ago3). Malgré tout, de nombreuses avancées ont été faites dans la compréhension de la voie des piRNA et notamment dans leur biogenèse (Ishizu et al. 2012). Ainsi, deux voies de biogenèse des piRNA ont été découvertes chez la drosophile : la voie de biogenèse des piRNA primaires et la boucle d'amplification du « ping-pong ». Ces voies sont conservées dans de nombreux organismes tels que la souris, le Xénope et le poisson zèbre.

La voie de biogenèse des piRNA primaires correspond à la maturation de longs ARN précurseurs sb provenant de quelques loci ponctuels dans le génome, appelés clusters de piRNA. Ces clusters, qui seront mieux détaillés dans la suite de l'introduction, peuvent être de taille très variée et sont généralement constitués de séquences homologues aux ETs. Dans les cellules somatiques qui entourent et protègent la lignée germinale de la drosophile femelle, les cellules folliculaires, la voie de biogenèse des piRNA primaires est la seule voie active de production de piRNA. Ce tissu s'est donc révélé particulièrement utile pour identifier les

différents facteurs importants pour leur biogenèse. Grâce à de nombreux cribles réalisés *in vivo* et *ex-vivo*, de nombreuses protéines ont été impliquées dans la voie de biogenèse des piRNA primaires, bien que le rôle précis de chacune soit encore flou.

Dans les cellules folliculaires de drosophile, la maturation des longs ARN précurseurs sb a lieu dans les Yb-bodies, une structure cytoplasmique péri-nucléaire qui contient notamment les protéines Yb, Armitage (Armi), Shutdown (Shu), Sister of Yb (SoYb), and Vreteno (Vret) (Saito et al. 2010; Handler et al. 2013) **(Figure 26)**. Chacune de ces protéines possède un domaine protéique qui la relie au métabolisme des ARN. Par exemple, Yb et Armi sont des hélicases présumées tandis que Vret contient un domaine de reconnaissance des ARN. La mutation d'une seule de ces protéines entraine une perte de la biogenèse des piRNA primaires. Yb serait nécessaire pour le recrutement des précurseurs ARN ainsi que pour la formation du complexe piRISC en collaboration avec la protéine Armi. Les piRNA issus de la voie de biogenèse primaire forment un complexe piRISC avec les protéines Piwi et Aub uniquement. Dans les cellules folliculaires, seule Piwi est présente. Les précurseurs ARN doivent subir deux clivages afin de générer les extrémités 5' et 3' des piRNA matures. L'activité endonucléase spécifique sb de la protéine Zucchini (Zuc) a été impliquée dans la formation des extrémités 5' monophosphorylé des piRNA (Ipsaro et al. 2012; Nishimasu et al. 2012). Zuc est une protéine qui appartient à la famille des phospholipases D et qui est présente à la surface de la membrane des mitochondries. Ces dernières sont d'ailleurs souvent retrouvées à proximité des Yb-bodies. Bien que Zuc soit capable de cliver n'importe quelle séquence ARN sb, la majorité des piRNA possède un nucléotide Uridine en position 5'. Cette caractéristique est probablement due à une affinité plus forte de Piwi et Aub pour ces piRNA. L'extrémité 3' est clivée par une endonucléase encore inconnue. De façon intéressante, la taille des piRNA varie légèrement en fonction de la protéine PIWI à laquelle ils sont associés. L'extrémité 3' serait donc générée après la formation du complexe piRISC et la protéine PIWI servirait de mesure. Suite au clivage de l'extrémité 3', la protéine Hen1 fixe un groupement méthyle sur le carbone 2' de cette extrémité (Horwich et al. 2007).

Une fois le complexe piRISC formé, celui-ci retourne dans le noyau grâce au domaine NLS de la protéine Piwi. Même si Piwi possède une activité endonucléase, celle-ci n'est pas nécessaire à son rôle dans la régulation des ETs, contrairement à son signal de localisation nucléaire (Darricarrère et al. 2013). Récemment plusieurs équipes ont observé que, une fois dans le noyau, le complexe piRISC interagit physiquement avec les protéines Mael et Gstf1 afin de réguler au niveau transcriptionnel les ETs (Dönertas et al. 2013; Rozhkov et al. 2013;

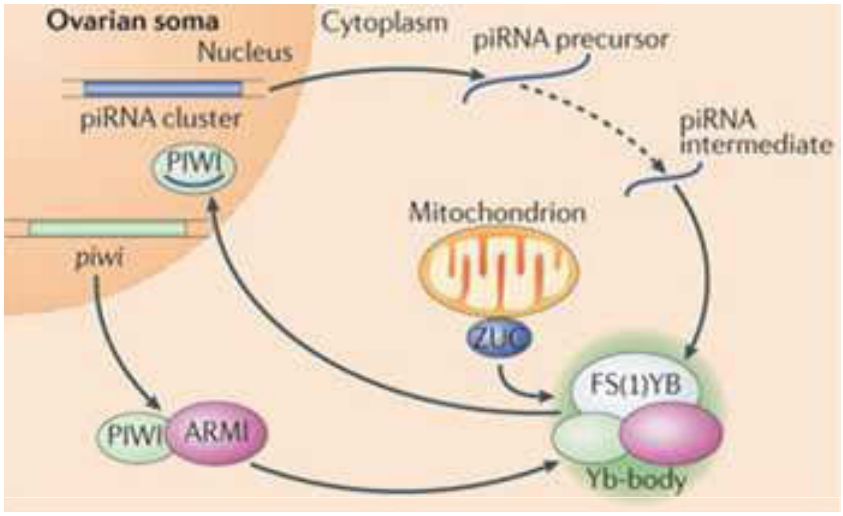


Figure 26 : La voie de biogenèse des piRNA primaires dans les cellules folliculaires de la drosophile

Les ARN précurseurs issus des clusters de piRNA sont transportés dans le cytoplasme vers les Yb-bodies où a lieu la maturation en piRNA et qui contiennent les protéines Piwi, Armi, et FS(1)Yb. La protéine mitochondriale Zucchini (Zuc) est responsable du clivage de l'extrémité 5' des piRNA matures qui sont ensuite chargés dans la protéine Piwi. Le complexe Piwi-piRNA retourne dans le noyau pour réguler au niveau transcriptionnel les ETs.

[Siomi *et al*, 2011]

Sienski et al. 2012). La mutation d'une de ces 3 protéines entraine une diminution de la marque H3K9me^3 au niveau des ETs ainsi qu'une augmentation de la présence de l'ARN polymérase II sur leur promoteur. Toutefois, même si les marques chromatiniennes sont considérées comme stable, il a été montré que la régulation des ETs nécessite la présence constante de la protéine Piwi (Dufourt et al. 2011).

Dans les cellules de la lignée germinale, les choses sont un peu plus compliquées, notamment par la présence de deux protéines PIWI supplémentaires : Aub et Ago3. Si la voie de biogenèse des piRNA primaires est toujours présente dans ces cellules, une voie d'amplification secondaire existe, appelée le mécanisme du ping-pong (Brennecke et al. 2007) **(Figure 27)**. Dans cette voie, les piRNA primaires chargés par Aub sont localisés dans une structure cytoplasmique appelée le « nuage » qui entoure le noyau. Ce complexe va alors réguler au niveau post-transcriptionnel les ETs en ciblant leur ARN. Ces ARN vont être clivés par la fonction endonucléase de la protéine Aub. Ce clivage, qui a lieu entre le dixième et onzième nucléotides de la zone ciblée par le piRNA (Gunawardane et al. 2007), va générer l'extrémité 5' d'un nouveau piRNA, dit secondaire, qui sera complémentaire sur 10 pb avec le piRNA primaire. L'ensemble des piRNA secondaires est d'ailleurs caractérisé par une majorité de nucléotide Adénosine en position 10. Ces piRNA sont alors pris en charge par la protéine Ago3, formant un nouveau complexe piRISC. Ce nouveau complexe Ago3-piRNA est capable de cibler l'ARN issu des clusters de piRNA et de les cliver en piRNA primaires formant ainsi une boucle d'amplification de la répression s'exerçant sur les ETs actifs.

2. Les mystères des clusters de piRNA

a) La découverte de *flamenco* chez *D. melanogaster*

En 1983, une équipe du centre CNRS de Gif-sur-Yvette caractérise 3 nouvelles mutations dominantes générées par mutagenèse EMS du gène *ovo,* qui code pour un facteur de transcription primordiale pour le développement précoce des gamètes femelles. De façon surprenante, le croisement d'une des 3 mutations, *ovoD1*, avec une souche contenant de nombreuses mutations sur le chromosome X, la souche MG, entraine la réversion de la mutation en un allèle *ovo* nul et récessif ainsi que l'apparition de mutations ailleurs dans le génome. L'hypothèse la plus simple était donc que ce croisement entraine la mobilisation d'un ET qui allait s'insérer préférentiellement dans *ovo*. Cette hypothèse fut confirmée par

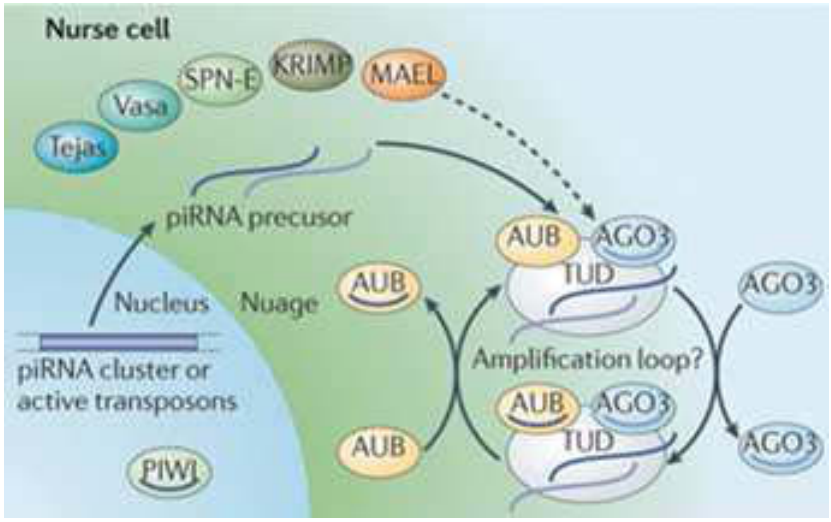


Figure 27 : La voie d'amplification du ping-pong dans les cellules germinales de la drosophile

Le mécanisme du ping-pong a lieu dans le nuage où la plupart des protéines impliquées dans cette voie d'amplification des piRNA sont localisées. La protéine Aubergine (Aub), chargée avec des piRNA primaires, va cibler les transcrits des ETs et les maturer en piRNA secondaires qui seront pris en charge par la protéine Argonaute 3 (Ago3). Ce nouveau complexe va cibler les transcrits issus des clusters de piRNA et les transformer en nouveaux piRNA primaires. Ce mécanisme nécessite de nombreuses protéines partenaires telles Vasa, Spindle-E, … et l'ensemble colocaliserait grâce à la présence de protéines TUDOR possédant de nombreux domaines d'interactions.

[Siomi *et al*, 2011]

l'étude de l'équipe de Madeleine Gans en 1989, qui prouva que les cas de reversions étaient dus à l'insertion de l'ET *gypsy* (et parfois *copia*), un rétrotransposon à LTR de structure très proche des rétrovirus, dans un site préférentiel localisé à l'intérieur du gène *ovo* (Mével-Ninio et al. 1989). Les auteurs supposèrent que le gène *ovo* contienne une séquence ADN spécifique ciblée par *gypsy*, de même que d'autres gènes où de nombreux cas d'insertions furent recensés (*cut* par exemple). Concernant la souche MG, les auteurs ne purent que constater un grand nombre de copie *gypsy* par rapport à une souche sauvage mais ne firent aucune hypothèse sur le mécanisme à l'origine de ces dérégulations. La réversion de la mutation *OvoD1* étant facilement identifiable par l'apparition de femelles fertiles dans la descendance, l'équipe d'Alain Bucheton en 1995 en profita pour caractériser la dérégulation de *gypsy* induite par la souche MG (Prud et al. 1995). Tout d'abord, ils montrèrent que le taux de mobilisation de *gypsy* est proportionnel à la température et inversement proportionnel à l'âge de la mère. De plus, la dérégulation de *gyspy* semble soumise à un strict effet maternel puisqu'elle affecte uniquement la descendance mâle ou femelle dont la mère provient de la souche MG. Enfin, les auteurs purent associer la propriété de la souche MG avec la mutation récessive d'un gène localisé au locus 20A1, dans l'hétérochromatine péricentromérique du chromosome X que les auteurs appelèrent *flamenco*, en référence à la musique jouée par les populations gitanes (*gypsy* en anglais). L'équipe du Dr Bucheton montra également que *flamenco* contrôle l'expression de *gypsy* uniquement dans les cellules folliculaires des ovaires de drosophile (Pélisson et al. 1994). L'allèle contenu dans la souche MG est dit permissif pour la mobilisation de *gypsy*, tandis que les souches sauvages contiennent un allèle restrictif.

En parallèle à ces travaux, d'autres équipes étaient à la recherche d'événements drastiques de dérégulation d'ETs. En 1999, notre équipe découvrit une lignée dans laquelle deux nouveaux rétrotransposons, *ZAM* et *Idefix*, sont dérégulés et s'insèrent préférentiellement dans le gène *white* (Desset et al. 1999). L'insertion de ces éléments entraine une modification de l'expression de ce gène, visible au niveau de la coloration de l'œil. En 2003, elle montra que cette dérégulation est due à une mutation située au locus 20A2 du chromosome X et donc à proximité du gène *flamenco* (Desset et al. 2003). Cette mutation entraine une dérégulation de *ZAM* et *Idefix* spécifiquement dans les cellules folliculaires des ovaires de drosophile **(Figure 28)**. Par contre, elle est incapable de déréguler *gyspsy* et est donc indépendante du gène *flamenco.* Les locus 20A1 et 20A2 sont inclus dans l'hétérochromatine péricentromérique du chromosome X, connue pour contenir de nombreuses séquences défectives d'ETs. Les auteurs proposèrent donc que cette région joue

Figure 28 : les fish de zam et idefix en stable et unstable

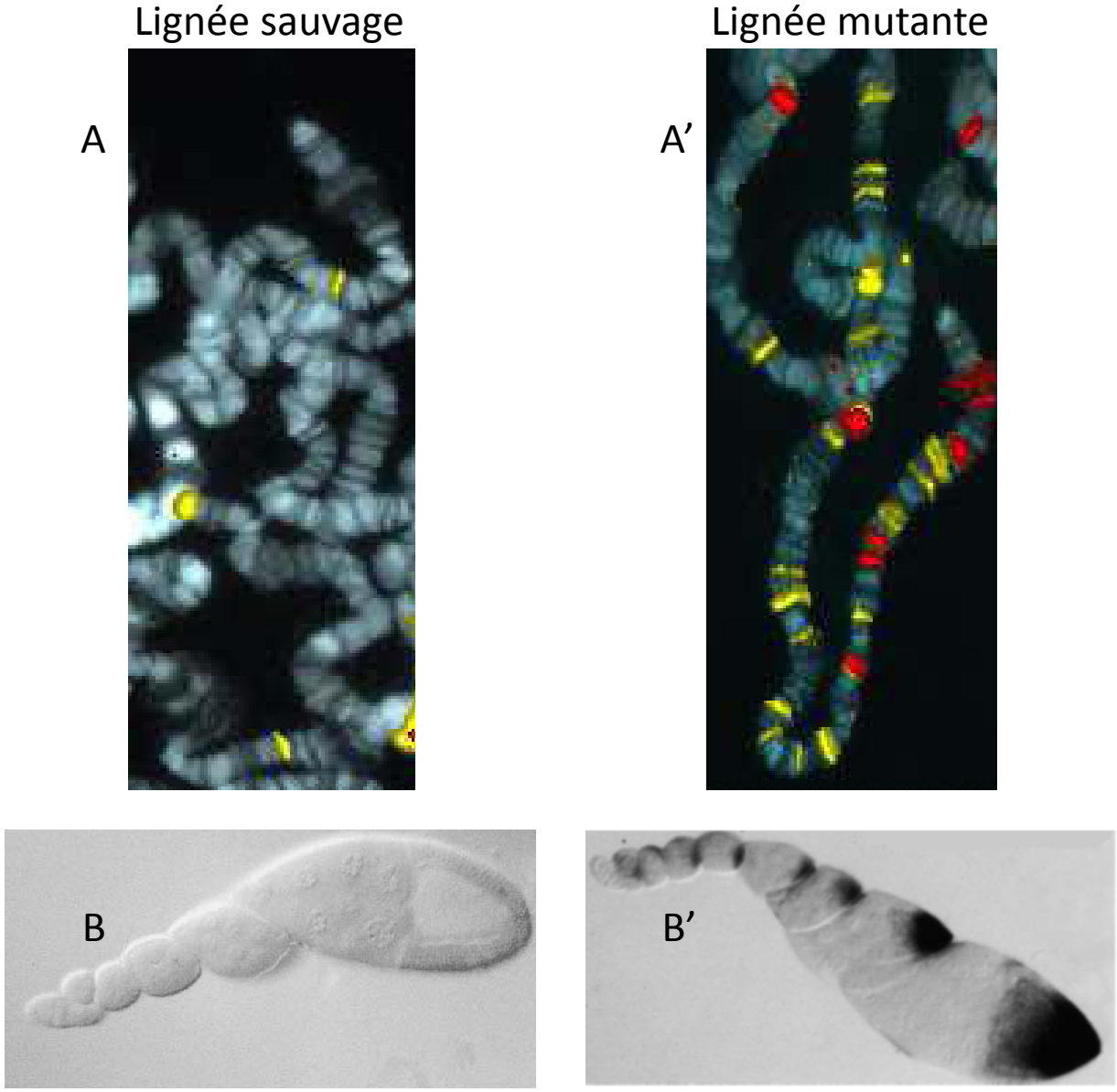


Figure 28 : Les lignées mutantes pour *flamenco* sont caractérisées par des dérégulations d'ETs

A et A'. Hybridation in situ en Fluorescence sur ADN de deux ETs, *ZAM* et *Idefix,* sur des chromosomes polytènes d'une lignée sauvage pour le locus *flamenco*/COM (à gauche) et une lignée mutante (à droite).

B et B'. Profil d'expression de la protéine LacZ dans les ovaires de drosophile dont le gène LacZ est contenu dans un transgène et fusionné à l'ET ZAM ce qui entraîne sa répression dans la lignée sauvage (à gauche) mais pas dans la lignée mutante pour le locus *flamenco*/COM (à droite).

[Adapté de Desset *et al*, 1999 et 2003]

un rôle central dans la régulation des ETs par un mécanisme d'ARN interférence et l'appelèrent COM pour Centre Organisateur de la mobilisation.

Ces arguments furent étayés par l'équipe du Dr Pelisson en 2004 qui montra que la régulation induite par *flamenco* est dépendante d'une protéine de la famille Argonaute, Piwi (Sarot et al. 2004). De plus, des petits ARN de 25 à 27 nucléotides complémentaires à la séquence nucléotidique de *gypsy* furent observés. En 2007, l'équipe du Dr Hannon séquença les petits ARN immuno-précipités avec Piwi, Ago3 et Aub, les trois protéines Argonautes de la sous-famille PIWI présentes dans les tissus reproducteurs de la drosophile femelle (Brennecke et al. 2007). De nombreux petits ARN provenant des loci *flamenco* et COM furent retrouvés co-immunoprécipités avec la protéine Piwi, prouvant que ces deux gènes font en fait partie d'un seul et unique locus, appelé *flamenco*, produisant de nombreux petits ARN non codants, les piRNA. Ce locus s'étend sur plus de 180 kb et il est constitué de plus de 80% d'ETs, majoritairement insérés dans un seul sens **(Figure 29A)**. L'hypothèse est donc que ce locus soit transcrit dans le sens opposé à l'insertion des ETs afin de produire des piRNA à la séquence complémentaire de celles des ARN des ETs, des piRNA dits antisens. Cette étude montra également que le génome de la drosophile contient de nombreux loci similaires à *flamenco*, que les auteurs appelèrent des clusters de piRNA. Par la suite, une étude réalisée en 2009 montra que *flamenco* produit des piRNA uniquement dans les cellules folliculaires des ovaires de drosophile, où il est d'ailleurs le cluster de piRNA majoritaire (Malone et al. 2009) **(Figure 29B)**.

La régulation des ETs dans les cellules folliculaires est particulièrement importante puisque des liens étroits existent entre cette lignée et la lignée germinale. Ainsi, des techniques de mosaïcisme ont permis de montrer que la descendance d'une lignée germinale restrictive pour l'ET *gypsy* mais entourée de cellules folliculaires permissives contient des néo-insertions de *gypsy,* ce qui suppose un passage de l'ET entre les cellules folliculaires et les cellules germinales (Chalvet et al. 1999). Or, *gypsy* est un rétrotransposon à LTR qui contient le gène *env* permettant de générer des particules infectieuses. L'hypothèse est donc que *gypsy* migre des cellules folliculaires à l'ovocyte en formant des particules pro-virales. Ces particules ont d'ailleurs été observées par deux équipes différentes (Song et al. 1997; Lécher et al. 1997). De la même façon, des particules pro-virales de *ZAM* ont été observées entre les cellules folliculaires et l'ovocyte (Brasset et al. 2006). Ces particules empruntent des voies d'exocytose et d'endocytose pour circuler entre ces deux types cellulaires.

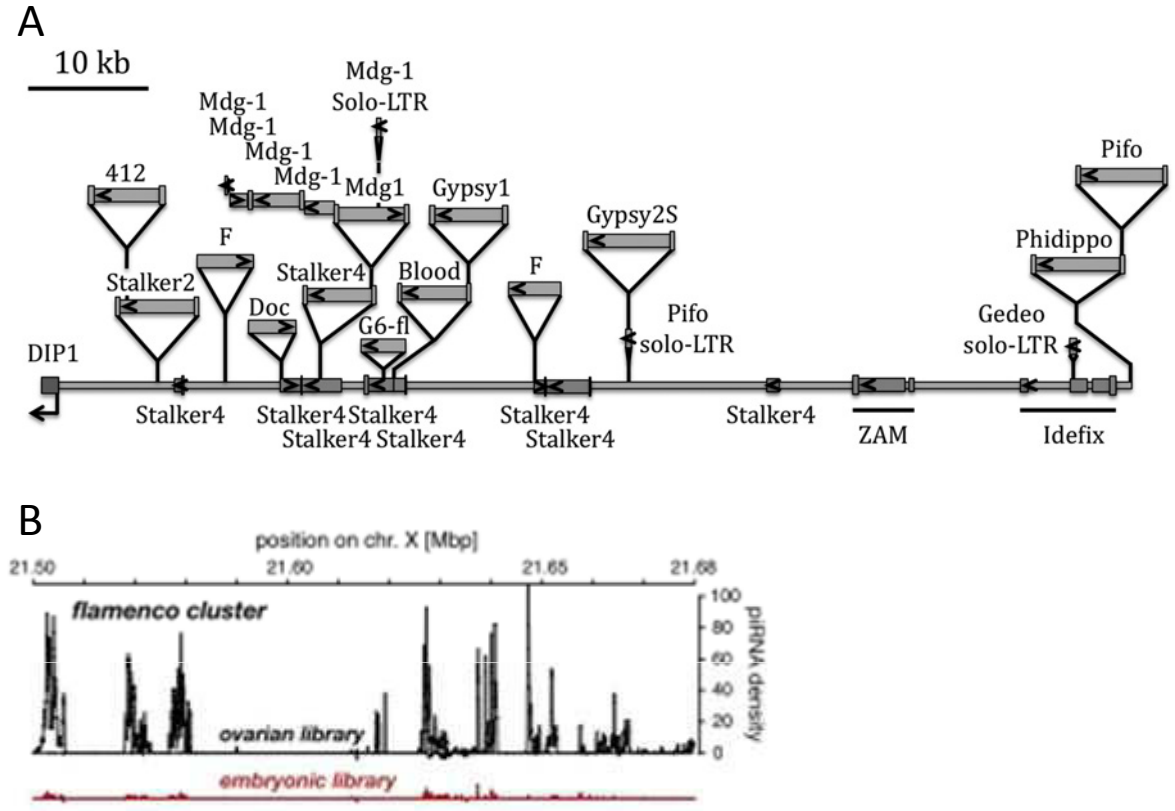


Figure 29 : Structure du locus *flamenco* et production de piRNA

A. Représentation schématique des différents ETs insérés dans *flamenco* et de leur sens d'insertion, indiqué par les flèches noires. Les triangles représentent les ETs récemment intégrés au locus.

[Zanni *et al*, 2013]

B. Densité de piRNA unique le long du locus *flamenco* dans les ARN totaux d'ovaires (au-dessus) ou d'embryon (en-dessous).

[Malone *et al*, 2009]

b) Structure et transcription des clusters de piRNA

Chez la drosophile, 142 clusters de piRNA produisent 90% des piRNA contenus dans les ovaires. De nombreux gènes produisent également des piRNAs, notamment au niveau de leur 3'UTR (Robine et al. 2009). Les clusters de piRNAs, longs de quelques kb à plusieurs centaines de kb, sont le plus souvent localisés dans l'hétérochromatine des chromosomes et généralement constitués de nombreux ETs défectifs et/ou tronqués enchevêtrés les uns dans les autres. Au total, ils couvrent environ 5% du génome de la drosophile. Ils sont classés en deux catégories : les clusters unidirectionnels qui produisent des piRNA à partir d'un seul brin, et les clusters double-directionnels qui en produisent à partir des deux brins d'ADN. Généralement, les clusters unidirectionnels sont actifs uniquement dans les cellules folliculaires, tandis que les clusters double-directionnels sont plutôt transcrits dans les cellules germinales. Le cluster *flamenco* est un exemple typique de cluster unidirectionnel tandis que les clusters double-directionnels sont souvent représentés par le cluster majoritaire de piRNA en cellules germinales, le cluster 42AB. De récentes études ont montré que ces clusters étaient différents par leur mode de fonctionnement.

Les clusters unidirectionnels tels que *flamenco* sont caractérisés par un pic de l'ARN polymérase II et de la marque d'histone activatrice H3K4me au niveau de leur TSS ainsi que par la production d'ARN coiffés (Mohn et al. 2014a). De façon intéressante, des expériences de mutagenèse ont montré que l'insertion d'un transgène au début de *flamenco* est suffisante pour empêcher la production de piRNA sur tout le locus (Prud et al. 2001). Il existe ainsi deux lignées de drosophile, KGP et BGP, dans laquelle l'insertion d'un transgène dérivé de l'ET P au début de *flamenco* entraine une perte de la quasi-totalité des piRNA provenant de ce cluster. Il en résulte une dérégulation des ETs *gypsy*, *ZAM* et *Idefix* ainsi que des défauts de développement des ovaires. Ces données semblent indiquer que les clusters unidirectionnels sont transcrits à partir d'un promoteur unique. Leur transcription produit de long ARN sb qui couvrent la totalité des clusters. Il a récemment été montré dans les cellules folliculaires de la drosophile que les ARN précurseurs issus des différents clusters unidirectionnels se réunissent en un seul foyer nucléaire localisé à la périphérie, appelé Dot COM (Dennis et al. 2013) **(Figure 30)**. Ce foyer est localisé à proximité des Yb-bodies, des structures cytoplasmiques où a lieu la maturation en piRNA. A l'inverse, une autre équipe a déterminé, en utilisant la microscopie électronique sur des cellules folliculaires en culture (les cellules OSS), que ce foyer, renommé *flam* Bodies, est situé en dehors du noyau et nécessite la présence de la protéine Yb pour se former (Murota et al. 2014a). Des analyses complémentaires sont donc

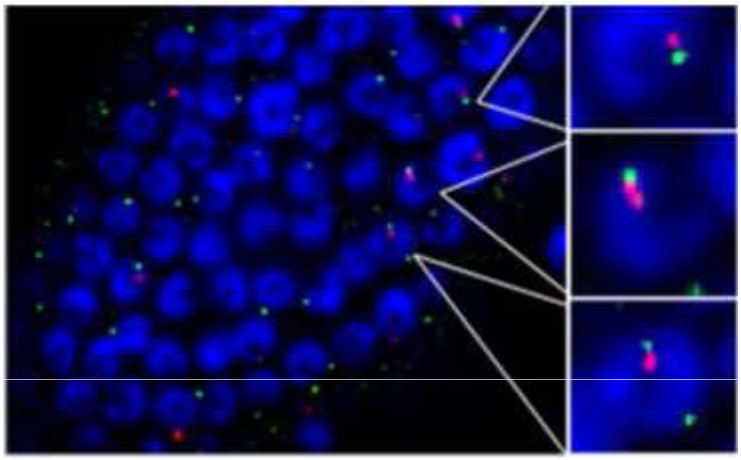


Figure 30 : Les ARN de *flamenco* forment un foci nucléaire appelé DOT-COM

Hybridation in situ en Fluorescence de l'ARN de *flamenco* (en rouge) et immunomarquage fluorescent de la protéine Armi dans les cellules folliculaires des ovaires de drosophile.

[Dennis *et al*, 2013]

encore nécessaires pour statuer sur la localisation de ce foyer, même si on ne peut exclure des différences fonctionnelles et mécanistiques entre les cellules folliculaires des ovaires de drosophile et les cellules folliculaires en culture.

Les mécanismes intervenant dans la transcription des clusters germinaux semblent très différents. Ces loci apparaissent comme très particuliers dans le génome de la drosophile puisque caractérisés par la présence d'ARN non coiffés sans signatures claires de promoteurs à leur extrémité. Des expériences de ChIP ont permis de mettre en évidence deux protéines spécifiquement fixées sur les clusters germinaux, Rhino (Rhi), une protéine hétérochromatique paralogue à HP1, et Cutoff (Cuff), une protéine similaire à un facteur de terminaison de la transcription chez la levure (Rai1) (Pane et al. 2011; Klattenhoff et al. 2009). De plus, les clusters germinaux sont fortement caractérisés par la marque d'histone de fermeture de la chromatine H3K9me. La perte de cette marque entraine une diminution de la transcription des clusters germinaux ainsi que de la production des piRNA (Rangan et al. 2011). Plus récemment, l'équipe du Dr Brennecke a développé un modèle dans lequel les clusters germinaux double-directionnels sont ciblés par un complexe RDC comprenant les protéines Rhi, Cuff et Deadlock (Del) (Mohn et al. 2014a) **(Figure 31)**. Cette association est permise grâce au chromo-domaine de la protéine Rhino qui interagit avec la marque de méthylation H3K9 présente sur ces clusters. Ce complexe permettrait la transcription non canonique des clusters germinaux. En effet, ces clusters sont caractérisés par une absence de pic d'ARN polymérase II à leur extrémité. Leurs transcriptions ne seraient donc pas dues à l'existence d'un promoteur spécifiquement défini de part et d'autre de ces clusters mais plutôt à la fuite de l'ARN polymérase II en cours de transcription provenant des unités transcriptionnelles voisines. Les transcrits néo-synthétisés par ce mécanisme ne pouvant subir le processus de coiffage, la présence de la protéine Cuff permettrait de les protéger de la dégradation. De plus, le complexe RDC, en collaboration avec la protéine UAP56 impliquée dans les processus d'épissage, empêcherait l'épissage de ces ARN (Zhang et al. 2014). Il en résulte une transcription non canonique des clusters germinaux capable de générer des transcrits non-coiffés en ignorant les divers signes d'épissage et de poly-adénylation. Ces transcrits seraient alors pris en charge par la protéine UAP56 pour être transportés vers le nuage, *via* les pores nucléaires et grâce à l'interaction avec la protéine Vasa, un constituant essentiel du nuage (Zhang et al. 2012).

Chez la souris on dénombre également une centaine de clusters dont la taille peut atteindre 100 kb et qui produisent 90% des piRNA présents dans les gonades mâles. Dans cet

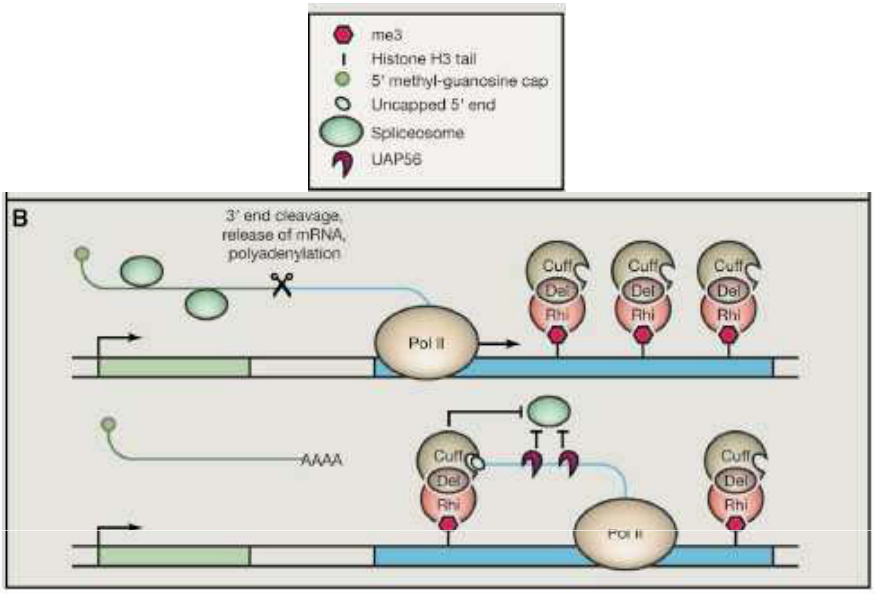


Figure 31 : Transcription des clusters de piRNA dans les cellules germinales de drosophile

A. Les clusters de piRNA germinaux (en bleu) sont transcrits par l'échappement de l'ARN polymérase II à partir de gènes convergents (en vert). La présence de marque H3K9me^3 permet le recrutement du complexe RDC (Rhino-DeadLock-Cutoff) qui protège de la dégradation les ARN non coiffés issus de cette transcription non canonique. La protéine Cutoff, en partenariat avec le protéine UAP56, empêche également l'épissage de ces ARN.

[Sapetschnig A and Miska EA, 2014]

organisme, il semble que deux populations de piRNA différentes se succèdent en fonction du stade développemental. Avant que les spermatocytes n'atteignent le stade pachytène de la méiose, les piRNA produits sont majoritairement homologues aux ETs de la souris. A ce stade, les clusters de piRNA ressemblent aux clusters double-directionnels de la drosophile et produisent des piRNA à partir des deux brins d'ADN d'un même locus. Une fois le stade pachytène atteint, une nouvelle population de piRNA est produite qui correspond majoritairement à des séquences non codantes uniques et dont le rôle précis est encore inconnu. Ces piRNA proviennent majoritairement de clusters unidirectionnels. Quelques clusters produisent des piRNA à partir des deux brins d'ADN mais ces piRNA ne sont pas chevauchant et proviennent de deux événements de transcription indépendants et divergents à partir d'une même région promotrice. On parle alors de clusters bidirectionnels. Grâce à l'analyse *in silico* de l'ensemble des promoteurs des clusters du stade pachytène, il a été mis en évidence l'implication du facteur de transcription A-myb dans l'activation de la transcription de ces clusters (Li et al. 2013).

Enfin, chez *C. elegans*, chaque piRNA provient d'une unité de transcription spécifique possédant son propre promoteur reconnu par l'ARN polymérase II (Ruby et al. 2006). Cette transcription est activée par la fixation d'un facteur de transcription de la famille Forkhead sur une séquence consensus située 40 pb en amont.

c) Les clusters de piRNA, des pièges à ETs ?

Plus d'une centaine d'ETs différents ont été identifiés chez la drosophile. Ces différents ETs proviennent de famille évolutives différentes et sont donc extrêmement divergents en séquence et mode de fonctionnement. Afin d'assurer une régulation fine de tous ces ETs, la première difficulté à laquelle se heurte les organismes hôtes est d'être capable de reconnaître ces différents ETs et de les distinguer des gènes endogènes. L'hypothèse a été émise que les clusters de piRNA constituent une sorte de répertoire des ETs en agissant comme des pièges à ETs, se servant ainsi de la seule caractéristique commune des ETs : la capacité de transposition (Bergman et al. 2006). Lors de l'introduction d'un nouvel ET dans un organisme ou d'une dérégulation soudaine pour des raisons diverses, cet ET n'est d'abord soumis à aucune régulation de la part de l'hôte et peut donc se déplacer librement. L'un de ses événements de transposition le fera, tôt ou tard, s'insérer dans un cluster de piRNA. Une fois qu'il est inséré dans un de ces cluster, la cellule hôte serait alors capable de produire des

piRNA dirigés contre lui et donc de le réguler. Cette régulation pourrait s'exercer sur toutes les copies actives de l'ET situées en trans dans le génome mais également sur les copies d'ETs apparentés.

Afin de vérifier cette hypothèse, une étude a été menée par notre équipe chez la drosophile (Zanni et al. 2013). Elle a permis de mettre à jour les différences de structure génomique du cluster de piRNA *flamenco* entre une lignée capable de réprimer deux rétrotransposons, *ZAM* et *Idefix*, lignée dite restrictive, et une lignée permissive qui est incapable de les réprimer. Il a été observé une absence des séquences de *ZAM* et *Idefix* dans la séquence génomique de *flamenco* de la lignée permissive, contrairement à la lignée restrictive. De plus, une réannotation totale du locus a été réalisée avec de nouveaux outils d'annotation des ETs. Cette réannotation a permis d'identifier que : i) Sur les 52 ETs recensés sur *flamenco*, 49 sont présents en copie unique, ce qui va en faveur de l'hypothèse qu'une seule insertion d'un ET dans un cluster de piRNA est suffisante à sa répression, ii) Le locus *flamenco* semble extrêmement dynamique d'un point de vue structural avec l'occurrence de nombreux événements de duplication ou délétion à l'intérieur du locus mais également de très fréquentes insertions d'ETs, ce qui suggère un ciblage préférentiel des ETs dans ce cluster iii) Un certain nombre d'ETs qui étaient considérés comme appartenant à une certaine famille d'ETs de *D. melanogaster* partagent finalement plus d'identité avec des ETs d'autres espèces de drosophile du sous-groupe *melanogaster*, ce qui va en faveur d'une transmission horizontale d'ETs entre ces espèces.

Ainsi, au fur et à mesure des événements de transposition, les clusters de piRNA deviendraient des répertoires contenant l'ensemble des ETs présents dans le génome. Ceci constituerait une sorte d'immunité adaptative contre la mobilisation des ETs.

d) L'adressage des transcrits à la voie des piRNA

Les clusters de piRNA seraient donc des répertoires contenant l'ensemble des séquences qui doivent être mises sous silence. C'est pourquoi il est capital pour la cellule d'identifier correctement les ARN provenant de ces clusters afin de les envoyer vers la voie de maturation en piRNA. Toutefois, les mécanismes permettant cette reconnaissance sont encore peu compris.

En 2012, une étude réalisée chez la drosophile et la souris a montré que l'insertion d'une séquence hétérologue à l'intérieur d'un cluster permet de produire des piRNA à partir de cette séquence (Muerdter et al. 2012). De plus, des clusters insérés de façon ectopique ailleurs dans le génome, et notamment dans des régions euchromatiques, produisent de nombreux piRNA. Dans l'ensemble, ces données montrent que la reconnaissance des ARN issus des clusters de piRNA ne dépendrait pas de facteurs localisés dans les séquences environnantes aux clusters et que n'importe quelle séquence peut produire des piRNA à partir du moment où elle est introduite dans un de ces clusters. Toutefois, aucune étude bioinformatique n'a permis de retrouver une séquence consensus à proximité des clusters de piRNA chez la drosophile ou chez la souris. De même, aucune structure secondaire particulière n'a été retrouvée sur leurs transcrits. L'équipe de Stéphane Ronsseray a montré que la présence dans le génome d'un cluster de piRNA artificiel contenant des séquences homologues à un autre locus non producteur pouvait transformer ce locus « naïf » en un fort cluster de piRNA (de Vanssay et al. 2012). Cette transformation nécessiterait la présence d'un facteur hérité par la mère. Il a également été montré que des transgènes contenant des séquences d'un ET réprimé par les piRNA, l'élément I, peuvent également se mettre à produire des piRNA (Olovnikov et al. 2013). Toutes ces données soulignent l'importance de la préexistence de piRNA homologues dans la reconnaissance des clusters de piRNA. Ainsi, les piRNAs transmis à la descendance par la mère seraient nécessaires à la reconnaissance des clusters de piRNAs dans la drosophile fille.

De récents papiers ont permis de confirmer cette hypothèse (Mohn et al. 2014a; Le Thomas et al. 2014) **(Figure 32)**. Tout d'abord, il a été montré que les complexes piRISC hérités de la mère, et plus particulièrement les complexes Aub-piRNAs, permettent de spécifier les ARN issus des clusters de piRNA en initiant le mécanisme du ping-pong. Les complexes Piwi-piRNA peuvent, eux, cibler les séquences génomiques des clusters de piRNA afin d'entrainer leur hétérochromatisation par la marque H3K9me^3. Cette marque sera reconnue par le chromo-domaine de Rhino ce qui recrutera l'ensemble du complexe RDC et permettra la transcription non canonique de ces clusters. De façon intrigante, le ciblage artificiel d'une région génomique par Rhino ne serait pas suffisant pour déclencher la maturation de ses transcrits en piRNA. La production d'ARN antisens, à partir du même locus ou d'une région homologue ailleurs dans le génome, est nécessaire (Zhang et al. 2014). La fonction de ces ARN sens et antisens reste à découvrir, puisque la formation d'ARN db a, jusque là, été écartée dans la voie des piRNA.

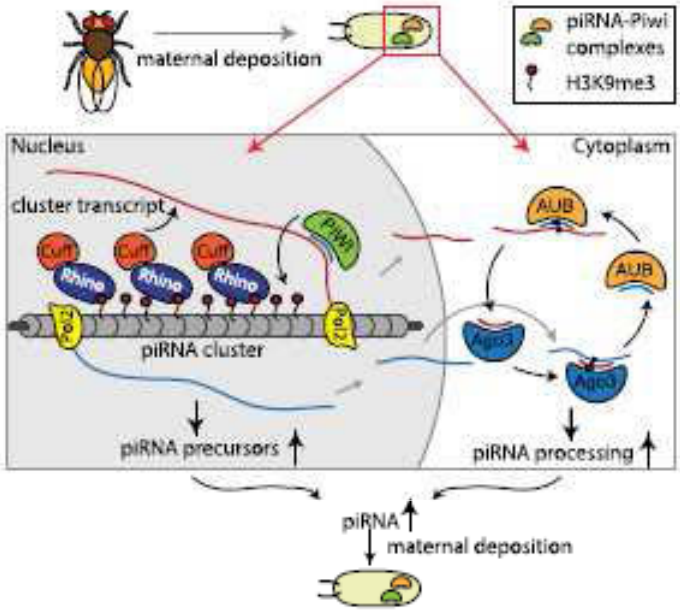


Figure 32 : Rôle des piRNA hérités de la mère dans la définition des clusters de piRNA

Les complexes piRISC hérités de la mère agissent de deux façons. Tout d'abord, ils initient le mécanisme ping-pong dans le noyau. De plus, ils ciblent les clusters de piRNA dans le noyau et déclenchent l'installation de la marque $H3K9me^3$. Cette marque, avec l'aide de facteurs encore inconnus, sera reconnue par le complexe RDC.

[Le Thomas *et al*, 2014]

III. Le modèle d'étude

A. *Drosophila melanogaster* et l'appareil reproducteur femelle

1. La drosophile

La drosophile fut introduite dans les laboratoires en 1901 à l'université de Harvard où le président du club entomologiste de Cambridge, Charles W. Woodworth, ayant repéré ce petit organisme invasif, proposa à William E. Castle de l'utiliser pour ses recherches en génétique. Quelques scientifiques se mirent alors à l'utiliser, dont Thomas H. Morgan qui la rendit célèbre en publiant la théorie chromosomique de l'hérédité grâce à ses recherches sur la drosophile. La drosophile fut donc le premier organisme à posséder une carte génétique, ce qui permit à de nombreux « drosophilistes » de créer de très nombreuses lignées au génotype et phénotype connus. La génétique de la drosophile est facilitée par le fait qu'elle ne possède que 4 paires de chromosomes, dont une paire de chromosome sexuel (XY) et une paire de chromosomes de taille très réduite, largement hétérochromatique et contenant peu de gène, les chromosomes 4. De plus, il n'y a pas de recombinaison génétique lors de la méiose chez le mâle. Enfin, la drosophile possède un cycle de reproduction très rapide, peut produire de nombreux descendants et nécessite des conditions d'élevage simples et peu coûteuses. Grâce à tous ces avantages, la drosophile devint très vite un des organismes les plus utilisés en recherche biologique, notamment dans le domaine de la génétique.

La drosophile est un insecte diptère de la famille des drosophilidae. Elle est également appelée la mouche du vinaigre pour son aptitude à coloniser les fruits en décomposition. Le cycle de vie de cet animal est très rapide et son espérance de vie est de 30 jours **(Figure 33)**. La durée de développement de la drosophile est inversement proportionnelle à la température d'incubation et peut varier de 7 jours à 28°C à plus de 45 jours à 12 °C. La femelle peut pondre jusqu'à 400 œufs qui écloront au bout d'une dizaine d'heures pour donner naissance à des larves. Les larves vont se développer pendant 4 à 6 jours puis s'encapsulent dans une pupe pour subir une métamorphose de 5 jours à l'issue de laquelle les drosophiles adultes émergeront. Une dizaine d'heures après son émergence, une drosophile femelle peut s'accoupler avec un mâle et commencer à pondre. La drosophile femelle possède une spermathèque qui lui permet de conserver les spermatozoïdes du premier mâle avec lequel elle s'accouple. L'expérimentateur doit donc s'assurer de la virginité des femelles (en les récupérant suffisamment jeunes) avec lesquelles il veut entreprendre un croisement.

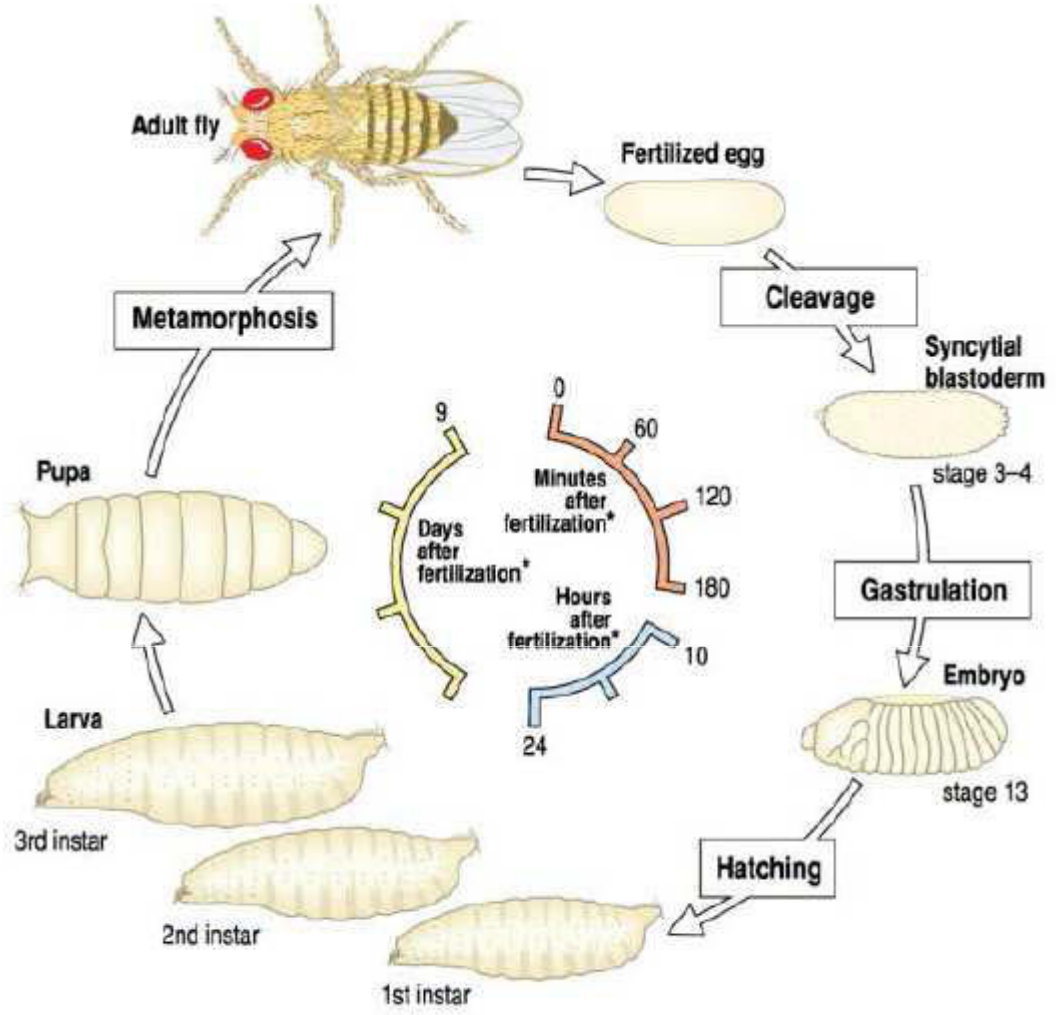


Figure 33 : Cycle de vie de la drosophile

Représentation schématique des différentes étapes de la vie de la drosophile. Les durées sont données pour une incubation à 25°C

[Biologie du développement, Lewis Wolpert (1999)]

Le génome de la drosophile s'étend sur 180 millions de pb et a été séquencé en 2000. Plus de 15 000 gènes y ont été recensés s'étendant sur 20% du génome. Basé sur une recherche de conservation inter-espèces des séquences, il semble que plus de la moitié des 80% de séquences restantes soient fonctionnellement importantes (Halligan & Keightley 2006). 50% des protéines de la drosophile possèdent une protéine homologue chez l'homme, ce qui permet d'utiliser ce modèle pour étudier de nombreuses pathologies humaines aussi diverses et variées que la maladie de Parkinson, l'obésité ou l'abus de drogue. A ces fins, de nombreux outils génétiques y ont été développés tels que la transgénèse, la mutagenèse, les systèmes de recombinaison ou même la réalisation de clones de cellules mutantes au sein d'un organisme de génotype sauvage.

L'étude des ETs chez la drosophile remonte à la découverte du phénomène de la dysgénésie des hybrides à la fin des années 1970. Depuis de nombreuses analyses y ont été menées sur leur activation, leur évolution mais surtout sur leur répression, notamment depuis la découverte de la voie des piRNA dans les tissus reproducteurs de la drosophile femelle.

2. Les ovaires de la drosophile femelle

Les tissus reproducteurs de la drosophile femelle occupent une large place dans l'abdomen de l'individu. Ils sont constitués d'une paire d'ovaires qui contiennent chacun 15 à 20 ovarioles où aura lieu l'ovogenèse qui dure une dizaine de jours **(Figure 34)**. Dans la partie antérieure des ovarioles se situe une niche qui contient 2 cellules souches de la lignée germinale. Ces cellules souches se divisent de façon asymétrique : une cellule fille reste dans la niche où elle continue à exprimer les caractéristiques d'une cellule souche ; l'autre cellule fille, localisée en peu plus en postérieur dans ce qu'on appelle le germanium, commence à se différencier en ovocyte. Cette cellule va d'abord subir 4 divisions successives pour donner naissance à 16 cellules dont une seule deviendra l'ovocyte. Les autres cellules deviendront des cellules nourricières qui participeront à la lignée germinale en déversant leurs ARN et protéines dans l'ovocyte à l'issue de l'ovogenèse.

L'ensemble composé d'un ovocyte et de ses 15 cellules nourricières est encapsulé par un épithélium folliculaire ovarien composé de cellules folliculaires. Ces cellules proviennent de cellules souches SSC (Somatic Stem Cells), situées dans le germarium des ovarioles. L'ensemble de la lignée germinale et des cellules folliculaires forme ce que l'on appelle un follicule qui doit subir une longue phase de maturation de plusieurs jours avant que l'œuf

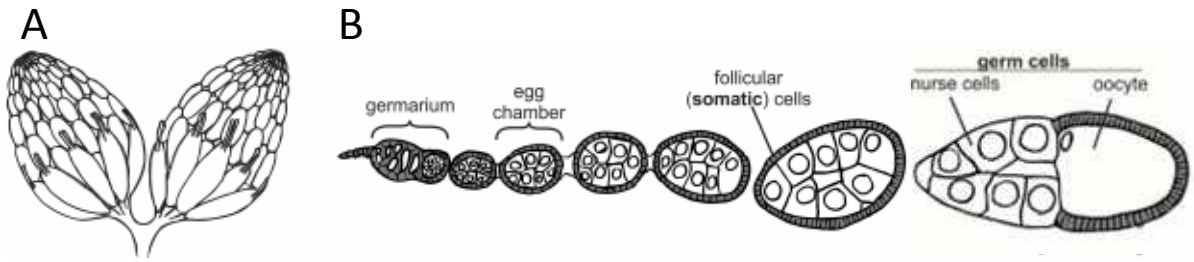


Figure 34 : Organisation de l'organe reproducteur femelle de la drosophile

A. Représentation schématique des deux ovaires qui contiennent la lignée germinale femelle. Chacun de ces ovaires est constitué d'une quinzaine d'ovarioles
B. Représentation schématique d'un ovariole, constitué d'une suite de follicules à différents stades de maturation. Chaque follicule contient un ovocyte, 15 cellules nourricières et des cellules folliculaires qui entourent et protègent l'ovocyte. Au pôle apical de l'ovariole se situe un germarium qui contient les cellules souches de la lignée germinale et des cellules folliculaires, ainsi que les cystes en division.

[Olovnikov IA and Kalmykova AI, 2013]

mature ne soit pondu. Un ovariole contient généralement 6 ou 7 follicules à des stades de maturation différents et qui se succèdent selon l'axe antéro-postérieur. Les cellules folliculaires entourent et protègent l'ovocyte et sont reliées entre elles par des jonctions serrées. Elles sont indissociables de la lignée germinale ce qui rend difficiles les analyses séparées *in vivo*. Pour cela, une lignée cellulaire dérivée des cellules folliculaires a été isolée.

3. Les cellules OSS/OSC

En 2006, une équipe de Chicago s'intéressait aux propriétés des cellules souches à l'origine de la lignée germinale et des cellules folliculaires de la drosophile femelle (Niki et al. 2006). Afin de pouvoir les étudier plus facilement, ils réussirent à mettre en culture ces différentes cellules à partir d'une souche mutante *bam*$^{-}$, un facteur de différenciation. Ils ont ainsi généré deux lignées stables : les cellules *bam*$^{-}$ GSC, qui contiennent un mélange de cellules germinales et folliculaires, et les cellules OSS ou OSC (Ovarian Somatic Sheet Cells), qui contiennent uniquement des cellules souches des cellules folliculaires, les SSC. Ces cellules expriment notamment la protéine FasIII, qui est un marqueur des cellules folliculaires.

Ces cellules semi-adhérentes se cultivent à 25°C avec du milieu de culture complémenté avec du broyat de drosophile. Elles ont été largement utilisées pour étudier la voie des piRNA et notamment la voie de biogenèse des piRNA primaires, qui est la seule ayant lieu dans ces cellules. Les séquençages des petits ARN ont prouvé que ces cellules contenaient bien des piRNA et elles se sont révélées particulièrement utiles pour réaliser des cribles de mutants par ARNi.

IV. Objectifs de l'étude

Bien qu'il ait été découvert chez la drosophile il y a une trentaine d'années, le cluster de piRNA *flamenco* reste encore bien mystérieux. Producteur majoritaire de piRNA dans les cellules folliculaires, il produit 79%, 30% et 33% de la totalité des piRNA homologues aux ETs *ZAM*, *Idefix* et *Gypsy* respectivement. Il est l'unique cluster pour lequel il existe des allèles nuls, ce qui suggère un rôle particulièrement important et peut-être un fonctionnement spécifique de ce cluster. Pourtant, si de nombreuses avancées ont été faites récemment sur le fonctionnement des clusters de piRNA de la souris, ou des clusters germinaux de la drosophile, les mécanismes nécessaires à la production de piRNA par *flamenco* sont encore largement méconnus.

L'objectif de ma thèse est d'aller plus loin dans la compréhension et l'analyse du locus *flamenco*, et notamment sur trois caractéristiques de ce locus. Tout d'abord, je me suis intéressée aux facteurs importants pour sa transcription ainsi qu'à la structure du ou des transcrit(s) issu(s)de *flamenco*. Ensuite, j'ai cherché à comprendre de quelle façon l'ARN de *flamenco* est adressé à la voie de maturation des piRNA. Et enfin, je me suis intéressée à l'environnement nucléaire de ce locus atypique et à ses partenaires d'interactions génomiques.

Résultats

I. Caractérisation de la transcription et de l'épissage du locus *flamenco*.

A. Introduction

En 2013, les facteurs impliqués dans la transcription du locus *flamenco* ainsi que la structure du ou des transcrit(s) issu(s) de ce locus sont encore méconnus. De plus, les quelques données obtenues sont contradictoires. En effet, certaines données supposent que *flamenco* produit un seul et unique long transcrit le long de ses 180 kb, battant ainsi largement le record du gène de la dystrophine long de 130 kb. Et pourtant, ce long transcrit n'a jamais pu être observé et seuls des ARN de petites tailles ont pu être identifiés par Northern Blot (de quelques dizaines de pb à plusieurs milliers) (Saito et al. 2010; Murota et al. 2014b). D'un autre côté, l'expression de ce locus semble être contrôlée puisqu'il n'est exprimé que dans les cellules folliculaires des ovaires de drosophile mais aucun facteur impliqué dans cette régulation n'a, jusqu'à présent, été mis en évidence. Récemment, de nouvelles données ont permis de mieux comprendre la transcription des clusters de piRNA actifs dans les cellules germinales de la drosophile. Mais les mécanismes mis en place ne semblent pas s'appliquer aux clusters de piRNA actifs dans les cellules folliculaires tel le locus *flamenco*. Nous avons donc souhaité étudier les caractéristiques de la transcription du locus *flamenco* ainsi que la structure de son ou de ses transcrit(s).

Afin de répondre à la première problématique, j'ai développé un système de double gènes rapporteurs dans les cellules OSS. Le premier gène rapporteur, codant pour la luciférase firefly, m'a permis de quantifier l'activité transcriptionnelle d'une séquence génomique d'intérêt. L'autre gène rapporteur, qui code pour la luciférase renilla, m'a permis de normaliser les résultats de mes expérimentations. Un gène rapporteur code pour une protéine exogène à l'organisme qui est facilement quantifiable. Ainsi, en clonant un gène rapporteur en aval d'une séquence génomique, l'activité transcriptionnelle de cette séquence pourra être évaluée en dosant la quantité de protéines produites. Les gènes des luciférases firefly et rénilla produisent deux enzymes, chacune capable de catalyser une réaction chimique produisant de la lumière suite à l'ajout d'un substrat spécifique dans l'extrait cellulaire à analyser. La lumière dégagée est mesurée par un luminomètre ce qui nous permet d'en

déduire la quantité d'enzyme initialement présente dans l'extrait. L'avantage d'utiliser les gènes de luciférase firefly et rénilla réside dans le fait que l'activité de chacune de ces enzymes peut être mesurée séparément de façon séquentielle dans un unique extrait cellulaire. L'activité de la luciférase firefly est d'abord mesurée en ajoutant son substrat spécifique. Puis la réaction est stoppée et le substrat de la luciférase rénilla est ajoutée.

Dans le but de caractériser le promoteur de *flamenco*, j'ai cloné des séquences génomiques d'intérêts en amont du gène de la luciférase firefly. Les plasmides ainsi obtenus ont été transfectés transitoirement dans les cellules OSS, où la transcription de *flamenco* est maximale. De plus, de nombreux piRNA provenant de *flamenco* ont été recensés dans ces cellules, ce qui prouve que la voie des piRNA primaires y est active. L'utilisation des cellules, en plus d'être plus simple à manipuler, permet de travailler sur un seul type cellulaire et donc de ne pas diluer le signal, comme dans les ovaires où les ARN issus des cellules folliculaires ne représentent qu'une part faible des ARN totaux.

Suite à la transfection des plasmides, l'activité de la luciférase firefly est dosée. Cette activité est directement proportionnelle à la quantité d'enzyme luciférase firefly produite et donc à l'activité transcriptionnelle de la séquence génomique en amont de ce gène. Toutefois, cette valeur peut être impactée par certains paramètres intrinsèques à l'expérimentation tels que le nombre de cellules transfectées ou l'efficacité de la transfection et du broyage des cellules. Dans le but de comparer plusieurs expériences indépendantes entre elles, il est nécessaire de co-transfecter un plasmide de référence dont l'activité transcriptionnelle n'est pas censée varier à l'intérieur des cellules OSS. Au cours de mes expériences de quantification de la luciférase firefly, j'ai donc utilisé un plasmide de référence qui contient le gène de la luciférase rénilla sous contrôle du promoteur du gène actine, un gène fortement exprimé et de façon ubiquitaire chez la drosophile. Ce promoteur a produit une quantité forte et homogène de luciférase rénilla dans les cellules OSS ce qui nous a permis de normaliser les activités obtenues à partir de nos différentes constructions. Grâce à ce système, j'ai pu obtenir des résultats fortement reproductibles qui m'ont permis d'estimer au mieux l'activité transcriptionnelle de chaque séquence génomique clonée en amont du gène de la luciférase firefly.

Afin d'étudier la structure du ou des transcrit(s) isssu(s) de *flamenco*, des techniques classiques de biologie moléculaire telles que des PCR sur ARN rétro-transcrits, ou RT-PCR, ont été utilisées. Afin d'obtenir une vision plus exhaustive, j'ai tiré profit du séquençage des

longs transcrits réalisé par l'équipe du Dr J. Brennecke dans un extrait d'ARN totaux issus des cellules OSS. Ces transcrits ont été obtenus à partir de cellules transfectées avec des siRNA dirigés contre la GFP, une protéine qui n'existe pas chez la drosophile . Ce séquençage a été utilisé comme contrôle expérimental par l'équipe de J. Brennecke.

L'ensemble de ces résultats a permis la rédaction d'un article de recherche publié en janvier 2014 dans le journal Embo Reports, retranscrit ci-après.

B. Résultats

Scientific Report

TRANSPARENT PROCESS

EMBO reports

Transcriptional properties and splicing of the *flamenco* piRNA cluster

Coline Goriaux[1,2,3,†], Sophie Desset[1,2,3,†], Yoan Renaud[1,2,3], Chantal Vaury[1,2,3,**] & Emilie Brasset[1,2,3,*]

Abstract

In *Drosophila*, the piRNA cluster, *flamenco*, produces most of the piRNAs (PIWI-interacting RNAs) that silence transposable elements in the somatic follicle cells during oogenesis. These piRNAs are thought to be processed from a long single-stranded precursor transcript. Here, we demonstrate that *flamenco* transcription is initiated from an RNA polymerase II promoter containing an initiator motif (Inr) and downstream promoter element (DPE) and requires the transcription factor, Cubitus interruptus. We show that the *flamenco* precursor transcript undergoes differential alternative splicing to generate diverse RNA precursors that are processed to piRNAs. Our data reveal dynamic processing steps giving rise to piRNA cluster precursors.

Keywords *Drosophila*; *flamenco*; piRNA clusters; transcription; transposable elements
Subject Categories RNA Biology; Transcription
DOI 10.1002/embr.201337898 | Received 19 August 2013 | Revised 29 January 2014 | Accepted 31 January 2014

Introduction

Small non-coding RNAs can induce gene silencing through specific base pairing with target molecules. A subclass of small non-coding RNAs (23–29 nt) that interact specifically with the PIWI clade of Argonaute proteins, the PIWI-interacting RNAs (piRNAs), ensures genomic stability by repressing the expression and transposition of transposable elements (TE) in reproductive tissues including *Drosophila* germline and surrounding somatic follicle cells. Most piRNAs are derived from presumed long, single-stranded precursor transcripts encoded by genomic loci known as piRNA clusters [1].

In somatic cells, a major piRNA cluster, *flamenco* (*flam*), controls several TEs such as *gypsy*, *Idefix* and *ZAM* [1–5]. This 180-kb locus is located at the boundary between euchromatin and pericentromeric heterochromatin on the *Drosophila* X-chromosome, proximal to the *DIP1* gene. It harbours many defective transposons similarly oriented to produce antisense transcripts capable of silencing active transposon mRNAs. We recently reported that the *flam* RNA precursor is transported from the genomic site where it is produced to a perinuclear structure called Dot COM juxtaposed with cytoplasmic Yb bodies where primary piRNA biogenesis occurs [6,7]. Promoters and transcription factors involved in piRNA cluster transcription are starting to be identified. In *Drosophila melanogaster*, Rhino and Cutoff are required for transcription/processing of germinal bidirectional piRNA clusters. In mice, the transcription factor MYB-related protein A has been reported to drive transcription of specific piRNA clusters [8–10]. To provide further understanding of piRNA cluster transcription, we undertook a comprehensive characterization of *flam* expression. We identified its transcription start site (TSS) and a transcription factor critical for its transcription in follicle cells. Our results also demonstrated that the *flam* transcript is alternatively spliced to generate multiple and distinct precursors.

Results and Discussion

An Inr-DPE Pol II promoter promotes *flam* piRNA cluster transcription

To identify the *flam* TSS, we performed 5′RACE experiments on four independent RNA extracts from *Drosophila* ovaries and ovarian somatic stem (OSS) cells (Supplementary Table S1). From the capped RNA fraction, a TSS located at position 21,502,918 (flybase version FB2011_08) was identified in all the independent amplifications from both ovary and OSS cell RNA extracts (Fig 1A). Several other TSSs (a total of 10) were occasionally amplified but were found in only one of the experiments performed. These data suggest that the *flam* transcripts are initiated from a major promoter located 1733 bp upstream of *DIP1*.

To gain a better understanding of the core promoter of *flam*, we examined the motifs located upstream and downstream of the TSS. Based on the consensus initiator element (Inr) sequence TCAGTY obtained by computational analysis of thousands of *Drosophila* core

1 Clermont Université, Université d'Auvergne, Clermont-Ferrand, France
2 Inserm, Unité 1103, Clermont-Ferrand, France
3 CNRS, UMR 6293, Clermont-Ferrand, France
*Corresponding author. Tel: +33 4 7317 8184; Fax: +33 4 7327 6132; E-mail: emilie.brasset@udamail.fr
**Corresponding author. Tel: +33 4 7317 8184; Fax: +33 4 7327 6132; E-mail: chantal.vaury@udamail.fr
†These authors contributed equally to this work.

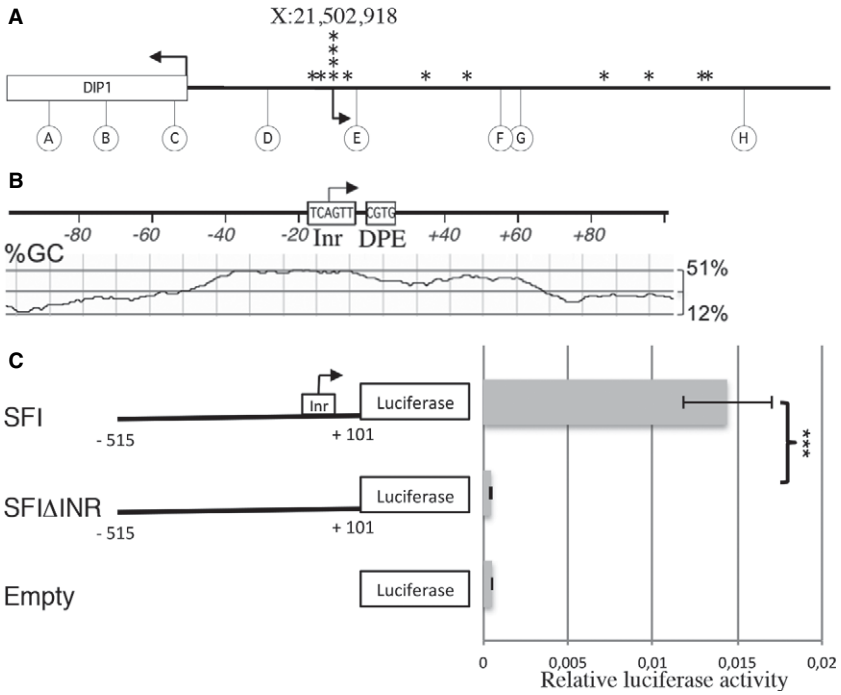


Figure 1. The *flam* locus is transcribed from an Inr-downstream promoter element (DPE) promoter.

A 5′RACE experiment on total RNA from ovary or ovarian somatic stem (OSS) cells. The arrowheads indicate the transcription start site (TSS) of the *DIP1* gene (X:21,501,185), the major TSS that initiates *flam* transcription (X:21,502,918), and the stars indicate the additional TSSs. Circles indicate primers used for 5′ RACE.

B *In silico* sequence analysis of the major *flam* TSS. The Inr and DPE motif are depicted, and the GC content is represented below (window size: 30 nucleotides).

C Schematic representation of the SF reporter constructs (left panel). Numbers indicate the fragment length from the TSS. SFΔInr carries a deletion of the predicted Inr sequence from −2 to +4. OSS cells were co-transfected with the firefly luciferase reporters carrying different *flam* promoter fragments and with an actin Renilla luciferase reporter construct (Supplementary Methods). Firefly luciferase (Fluc) activity was normalized to that of the Renilla luciferase (Rluc). Data are presented as means ($n = 4$). Error bars represent ± s.d.; Students *t*-test: ***$P < 0.001$.

promoters [11,12], we found that only the major TSS contains a consensus Inr sequence TCAGTT. In this Inr element, the A nucleotide corresponds to the +1 position of the core promoter (Fig 1B). Further analysis did not reveal a consensus TATA box, where the upstream T is usually located at −31 or −30 nt relative to the A +1 (or G +1) position in the Inr. However, a CGTG tetramer was characterized at +23 to +26 bp of the major TSS as a downstream promoter element (DPE), which is typically over-represented in many *Drosophila* TATA-less promoters. Like many *Drosophila* and mammalian promoters [13,14], a wide area in the vicinity of the major *flam* TSS (from −50 to +70 bp) displays a significant increase in GC content, which is known as a "GC hill." Aside from this major TSS, no other TSSs identified in this experiment displayed such promoter characteristics. Overall, these data designate the TSS located at 21,502,918 as the main promoter of the *flam* piRNA cluster.

To assess the potential of the *flam* Inr core promoter to drive transcription, the promoter region (SFI) including 515 bp upstream of the TSS and 101 bp of the transcribed sequence was cloned upstream of the luciferase reporter gene at the ATG start codon of the coding region. Transcriptional activities were measured in transient transfection experiments in OSS cells. Our results indicate that this *flam* fragment is sufficient to promote high-level expression of the *luciferase* reporter gene since an almost 30-fold enhancement of transcription of the firefly *luciferase* gene was observed compared to the empty plasmid (Fig 1C). Then, we generated a new reporter, SFIΔInr, that lacks the Inr sequence. This deleted reporter resulted in a significant decrease in luciferase expression compared to the transcriptional enhancement exhibited by the wild-type SFI. These results confirm the importance of the Inr sequence for promoting transcription of the *flam* locus.

The presence of an Inr core promoter and a cap structure indicates that RNA polymerase II (Pol II) could be responsible for *flam* transcription. In order to test this hypothesis, we treated OSS cells with alpha-amanitin, an inhibitor of initiation and elongation of Pol II. Transcription efficiency of the *flam* locus was determined by RT-qPCR using primer pairs spanning three different regions of *flam*. *18S* ribosomal RNA known to be transcribed by RNA polymerase I (Pol I) was used as a reference gene for normalization.

We found up to tenfold decreases in *flam*-derived long RNAs in cells cultured in the presence of alpha-amanitin, indicating that *flam* transcription is indeed Pol II dependent (Fig 2A). The amount of *rp49* transcripts (known to be transcribed by Pol II) is shown as a positive control. Moreover, using Pol I or Pol III inhibitors [15,16], we confirmed that *flam* transcripts are indeed products of Pol II (Supplementary Fig S1).

Then, we performed ChIP-qPCR experiments using an antibody against the initiating form of Pol II. We found that Pol II was more strongly recruited immediately downstream of the *flam* TSS than elsewhere within the gene body (Fig 2B). Thus, Pol II is the polymerase involved in *flam* piRNA cluster transcription. These results extend findings obtained in mouse testes, in which piRNA precursor

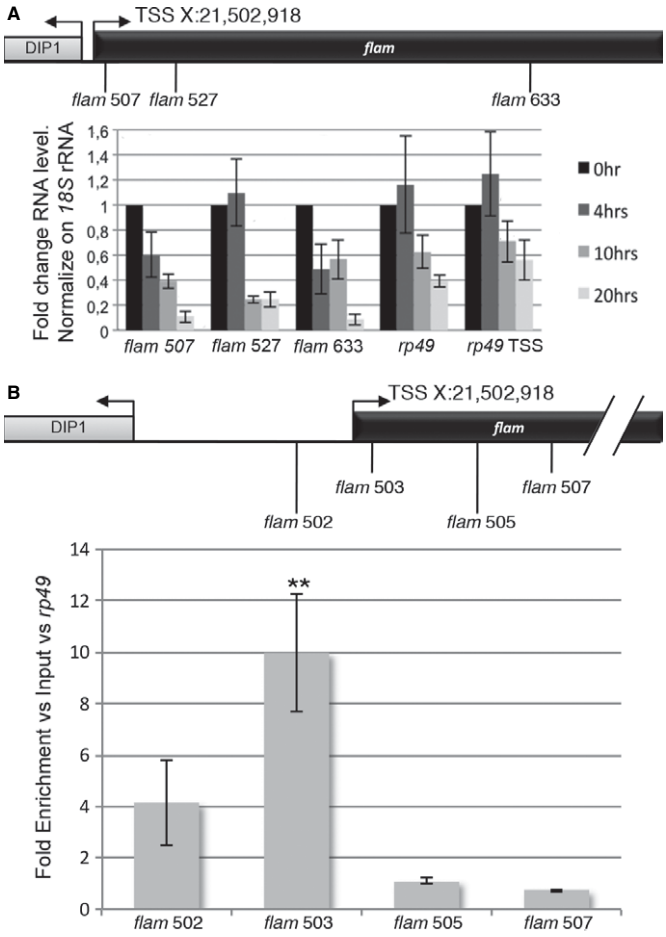


Figure 2. The *flam* locus is transcribed by RNA polymerase II.

A RT-qPCR on total RNAs from ovarian somatic stem (OSS) cells treated with α-amanitin from 0 to 20 h. The positions of primers along the *flam* sequence are indicated above. The amount of *rp49* transcript is shown as a positive control. The first primer set is located in exon 3, and the second primer set in exon 1 of the *rp49 gene*. qPCR data were normalized to *18S* rRNA. Data are presented as means (n = 4). Error bars represent ± s.d. Expression in non-treated OSS cells was set to one.

B ChIP-qPCR analysis was carried out on the *flam* promoter with a specific antibody against phospho-S5 RNA polymerase II. The positions of primers used for PCR amplification are indicated above. Enrichments were calculated versus Input and *rp49*. Data are presented as means (n = 5). Error bars represent ± s.e.m.; Students t-test: $^{**}P < 0.01$.

transcripts have been described to be canonical Pol II transcripts bearing 5′caps and 3′ poly(A) [10].

The transcription factor, Cubitus interruptus, is required to activate transcription of the *flam* locus

To identify cis-regulatory sequences, we constructed serially deleted promoter-*luciferase* reporter plasmids containing various lengths of the *flam* promoter region from either −1,624 bp (SF), −515 bp (SFI) or −356 bp (SFII) upstream to +101 bp downstream of the TSS. When the SF construct was used for transfection, efficient reporter activity was detected (Fig 3A). Deletion of the region from −1,624 to −515 (SFI) did not result in any significant change in promoter activity. On the contrary, further deletion to −356 (SFII) caused an eightfold decrease in promoter activity compared to the SFI construct. Finally, a NC construct corresponding to SFI in which the *flam* fragment comprised between −515 and −356 has been replaced by a 159-bp fragment of a non-promoting sequence, confirmed that the region located downstream of position X: 21,502,403 (−515 bp) and upstream of position X: 21,502,562 (−356 bp) contains critical *cis*-elements required for the transcriptional activation of the locus.

Within the −515; −356 region, nine potential transcription factor-binding sites were identified using genomatix MatInspector (Fig 3B). Based on the modENCODE dataset, four of them are expressed in OSS cells: Broad (Br), Big-brother (BgB), Doublesex (Dsx) and Cubitus interruptus (Ci). To specifically analyse the involvement of these factors in *flam* transcription, we performed successive deletions of each of their predicted binding sites (Fig 3C). The expression of each construct significantly decreased when compared with the SFI control but the most severe reduction (tenfold) was observed with SFI deleted for the Ci binding site, which was similar to the levels seen with the SFII construct. This suggests that the Ci binding site is necessary for the activation of *flam* transcription.

Several lines of evidence further implicated Ci in regulating *flam* transcription. First, Ci is expressed in follicle cells from the germarium to stage 6 egg chambers (Fig 4A) (Supplementary Fig S2) [17]. Second, based on ChIP assays, we found that Ci is 10- to 12-fold more recruited around the TSS and its predicted binding site than elsewhere in the locus (Fig 4B). Third, mutant clones generated by mitotic recombination using flies [*y-hs-flp*; *FRT42D P*[*Ci*+] /*FRT42D hs-MYC 45*; *Ci94*/*Ci94*] indicated that the *flam* transcript level decreases in *Ci* mutants in a manner similar to the decrease observed for *ptc* transcripts, a gene known to be activated by Ci, but not producer of piRNAs [18] (Supplementary Fig S2). Fourth, siRNA-mediated knockdown of *Ci* in OSS cells led to a decrease in *flam* transcripts two days post-transfection (Fig 4C). In contrast, the production of piRNAs and the TE mRNA levels were not significantly affected (Supplementary Fig S3). However, an upregulation of TE expression was observed 4 days post-infection (Fig 4D). A delay is observed between disruption of *flam* transcription and TE deregulation possibly due to stability and abundance of *flam* piRNAs.

Finally, evidence that Ci is involved in *flam* transcription was also provided by an analysis of the *flam* mutation present in the BG lines [19]. In this line, a P-element insertion at the 5′ end of *flam* results in an absence of the precursor transcripts encoded by *flam* [1]. When examined in detail, we found that the P-insertion occurred at position X:21,502,538 (−380 bp from the TSS), a position that disrupts the Ci binding site. Considered together, these data strongly suggest a role for Ci in the activation of *flam* transcription.

In *Drosophila* somatic follicle cells, the major sources of piRNAs are the *flam* locus and the cluster 2. Thus, we examined the cluster 2 promoter and found an Inr consensus sequence (21,390,615) 108 bp upstream of the first piRNA, and a Ci binding site 2,846 bp upstream of the Inr (Supplementary Fig S4). Furthermore, Ci mutants led to a decrease in *cluster 2* expression (Fig 4C and Supplementary Fig S2). These data suggest that Ci might also contribute to the transcription of other piRNA clusters in these cells.

A comparative analysis of the *flam* promoter region performed across several *Drosophila* species, *D. sechellia, D. simulans, D. yakuba, D. erecta,* was then performed. These species diverged from a common ancestor approximately 10 million years ago

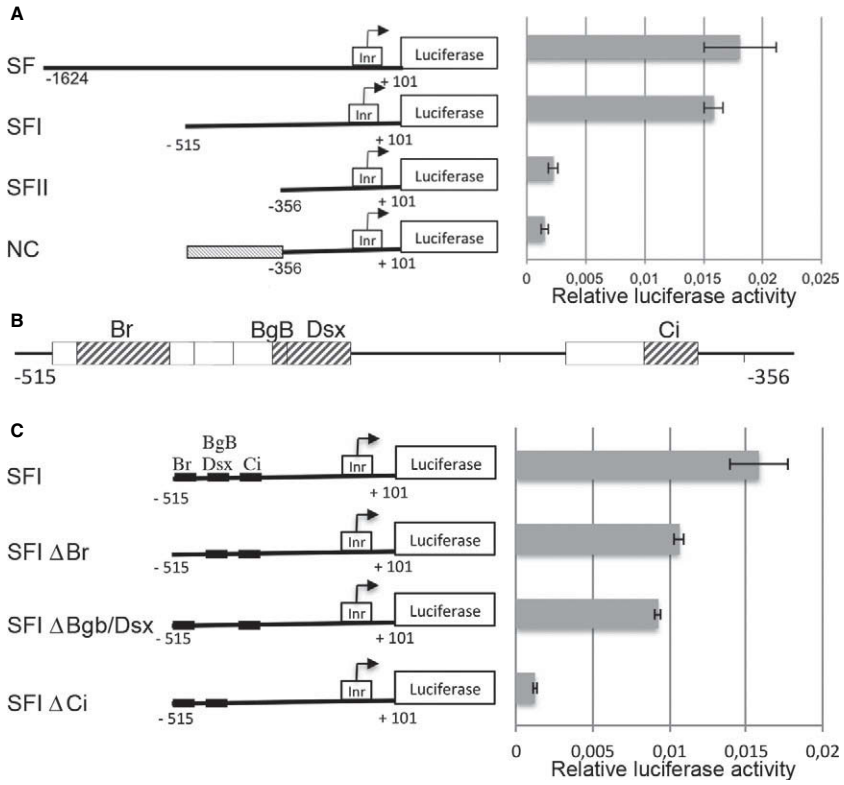


Figure 3. A functional Ci binding site is required for *flam* transcription.

A Schematic representations of the SF reporter constructs (left panel). The relative luciferase activity (Fluc/Rluc) was measured as described previously (right panel). Data are presented as means (n = 4). Error bars represent ± s.d. In the negative control construct (NC), a 159-bp fragment of a non-promoting sequence taken within the *gfp* gene was cloned upstream the −356 to fill the space between −515 and −356.

B Genomatix *in silico* analysis of the region from −515 to −356 upstream of the transcription start site (TSS). Boxes indicate transcription factor-binding sites. Grey boxes indicate transcription factors known to be expressed in ovarian somatic stem (OSS) cells (modEncode data).

C Schematic representation of firefly reporters carrying deletions of each predicted transcription factor-binding site. The Fluc/Rluc activity was measured as described previously (right panel). Data are presented as means (n = 4). Error bars represent ± s.d.

[20,21]. We found that *flam* orthologs are located on the pericentromeric X-chromosome close to the *DIP1* gene in *D. simulans* and *D. erecta*, similar to *D. melanogaster*, whereas they are still assigned in a scaffold in *D. yakuba* and *D. sechellia* (Supplementary Table S2). A multiple alignment revealed two highly conserved regions located at positions (−14;+37) and (−398; −372) according to the *D. melanogaster flam* TSS. The first (−14;+37) corresponds to the Inr-DPE core promoter suggesting a high conservation of its function. The second (−398; −372) includes the Ci binding site (Supplementary Fig S4). Then, we plotted uniquely mapping piRNAs that could be assigned to the putative *D. erecta flam* locus [5]. We found that, like in *D. melanogaster*, the density of piRNAs is very low close to the *flam* presumptive promoter and it highly increases 1 kb downstream (Supplementary Fig S4). This analysis of the *flam* promoter sequence across several *Drosophila* species confirms that the Inr-DPE and the Ci binding site are necessary motifs for *flam* transcription.

The *flam* transcript is alternatively spliced and gives rise to multiple *flam* precursors

The *flam* piRNA cluster has been proposed to produce a long single-stranded precursor RNA that is processed into primary piRNAs in the cytoplasmic Yb bodies [6,22]. We sought to better characterize this proposed long precursor. Fragments amplified from the 5′RACE experiments described above to localize the TSS were systematically sequenced. This allowed the identification of an intron located between bases +432 and +2067 from the *flam* promoter. Then, RT-PCR experiments were performed using a 5′ primer taken either within the first or the second exon, and 3′ primers designed along the 180 kb of this cluster. Figure 5A shows structures of *flam* transcripts deduced from sequencing of RT-PCR products. Different patterns of intron splicing were detected. The intron sizes are extremely diverse and range from 0.7 kb to 158 kb. Interestingly, the first exon (exon 1: 21,502,918…21,503,349) was found to be constitutively spliced since it is always present within the processed RNAs. By contrast, downstream of this first common exon, the other exons differ indicating that they result from alternative splicing. Analysis of *flam* spliced transcripts revealed that the majority of the intron boundaries obey the GT-AG rule (Supplementary Tables S3 and S4).

To verify our findings, we interrogated publicly available RNA-seq libraries [23] and found that indeed very few reads corresponding to intron 1 have been reported compared to the number of reads mapping exon 1 or exon 2 (Fig 5B). We found that 84% and

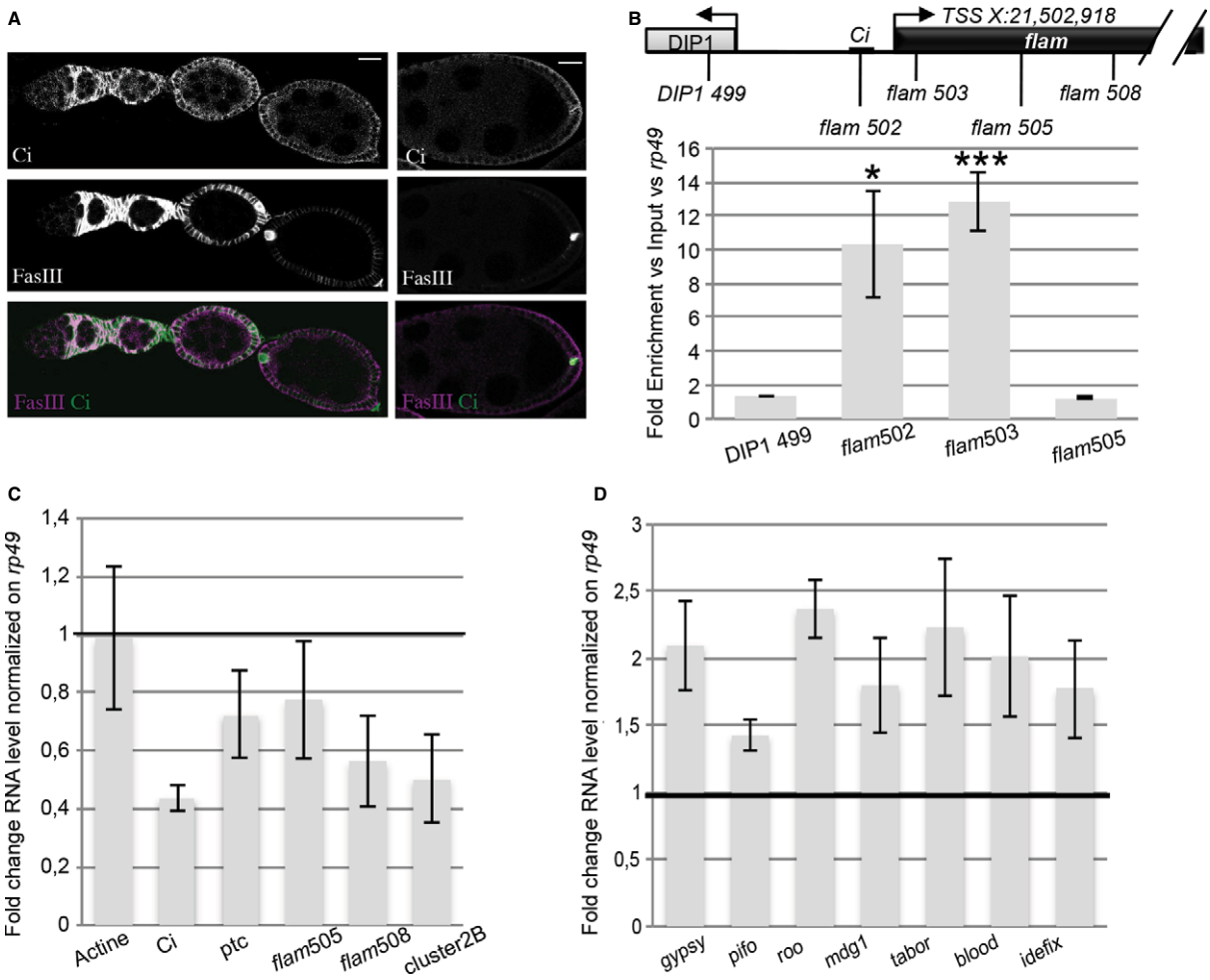


Figure 4. Ci activates *flam* transcription *in vivo*.

A Ci and FasIII (a follicle cell marker) immunostaining of a *Drosophila melanogaster* ovariole. Left panel shows early stages and right panel a stage 8. Scale bar, 20 μm.

B ChIP-qPCR analysis was carried out on the *flam* promoter with a specific Ci antibody. The positions of primers used for PCR amplification are indicated above. Enrichments were calculated versus Input and *rp49*. Data are presented as means (n = 5). Error bars represent ± s.e.m., *P < 0.05, ***P < 0.001.

C RNA level in ovarian somatic stem (OSS) cells transfected with siRNAs against Ci as compared to OSS cells transfected with siRNAs against GFP. The positions of primers used for PCR amplification are indicated above. *ptc* is a gene known to be activated by Ci in follicle cells. Data are presented as means (n = 4). Error bars represent ± s.e.m.

D Fold change in RNA level of diverse somatic transposable elements in OSS cells transfected with siRNAs against Ci and compared to OSS cells transfected with siRNAs against GFP, 4 days post-transfection. Data are normalized to *rp49* expression and are presented as means (n = 4). Error bars represent ± s.e.m.

16% of reads mapped the first exon–exon and intron–exon junction, respectively (Fig 5C). Then, we extended this analysis to 21 major piRNA clusters expressed in ovaries and found that seven of them contain introns (Supplementary Fig S5). These data suggest that several piRNA clusters including *flam* are transcribed as a long primary multi-kilobase RNA transcript before being spliced.

To determine whether these spliced RNAs are processed into piRNAs, we sequenced small RNAs from OSS cells and searched for reads that align uniquely to the identified *flam* spliced junctions. Reads spanning exon junctions were identified. Furthermore, we found that piRNAs encompassing the exon 1/intron 1 junction are under-represented compared to piRNAs matching the splice junction (Fig 5C). These results further indicate that *flam* transcripts are processed into piRNAs after the precursor is spliced. Although the diversity of alternatively spliced transcripts of *flam* is likely underestimated, it can be predicted that the multiple splicing events contribute to create a high diversity of *flam* precursors.

In $flam^{KG}$ mutant, the *KG* transgene is localized at position 21,505,285 downstream of the TSS, at the beginning of intron 2. Nevertheless, homozygote $flam^{KG}$ mutant females exhibit atrophic ovaries like $flam^{BG}$ females [24]. This ovarian phenotype has been attributed to an absence of *flam* transcription. If the reason why $flam^{BG}$ transcription is affected can be explained by disruption of the Ci binding site, the reason why *flam* transcription is also affected in the $flam^{KG}$ mutant remains obscure. It can be proposed that either the correct transcription of *flam* or the stability of its transcripts is affected. We have shown that the *KG* transgene is

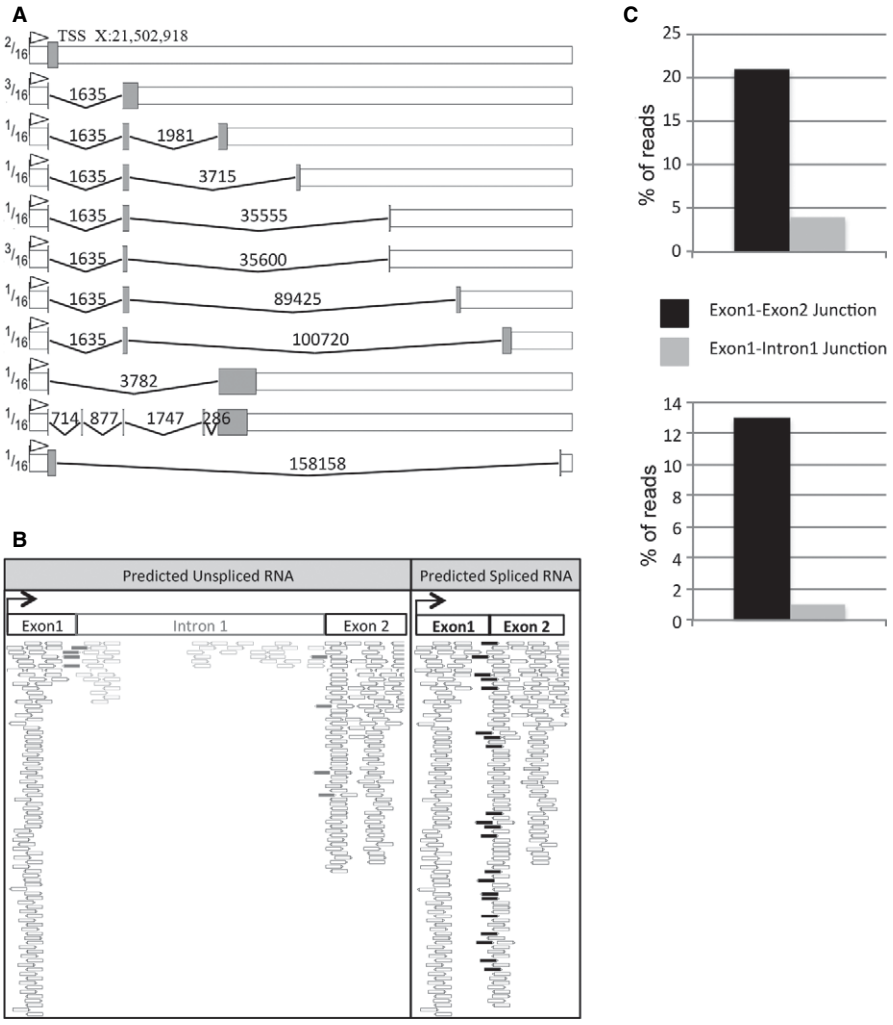


Figure 5. *flam* transcripts are alternatively spliced before piRNA processing.

A Representation of alternatively spliced RNAs identified by RT-PCR experiments. Grey boxes represent exons and peaked lines introns, numbers above introns indicate the length. The ratio of individual alternative transcripts to the total transcripts is indicated on the left part of the figure.

B Mapping of reads from Sienski *et al* (2012) sequencing data on exon–1–intron–1–exon–2 or on exon–1–exon–2 predicted *flam* transcripts. White reads mapped in exons, grey reads between exon 1 or 2 and intron 1, and black reads between the 2 exons.

C Percentage of reads (RNAs or piRNAs) corresponding either to the predicted non-spliced exon 1/intron 1 junction or to the predicted spliced exon 1/exon 2 junction in ovarian somatic stem (OSS) cells.

located at the border of the second intron. Disruption of this site might prevent its recognition as a donor site. Since almost all the spliced transcripts detected in WT *flam* alleles contain this spliced border, it might then be anticipated that this donor site plays a crucial role in generating the pool of alternative spliced RNAs. *flam* mutation due to *KG* insertion would then lead to unstable *flam* transcripts and thus, as for the *BG* insertion, to a phenotype of atrophic ovaries.

Overall, *flam* precursors display two characteristics: first, they display distinct structures resulting from alternative splicing, and second, they all share the first exon at their 5′ end. Future work is needed to elucidate the function of this common 5′ end. A likely hypothesis is that it helps to transfer RNA precursors from their site of transcription to Dot COM at the nuclear membrane facing the cytoplasmic Yb bodies, where they are processed to piRNAs. Recently, UAP56, a helicase of the exon junction complex (EJC), has been shown to play a role in the transport of germline precursor piRNA transcripts to the nuclear pore [25]. It remains to be clarified whether the recruitment of the EJC necessary for *flam* splicing also plays a role in the stabilization, surveillance and transport of the *flam* precursors.

Many TE families are known to originate from recent horizontal transfer between *Drosophila* species [26]. Recently, we have reported that many of these new TEs preferentially insert within heterochromatic regions such as the *flam* locus [27]. Thus, the dynamic nature of this piRNA cluster suggests that novel motifs for splicing are constantly gained or lost resulting in distinct pools of

flam precursors. Such stochastic splicing depending on structural modifications affecting piRNA loci might help genomes to rapidly react against new TE invasions.

Materials and Methods

Drosophila strains

ChIP experiments were performed on the W^{1118} line. Clonal analyses were performed on flies with the following genotype: *y-hs-flp; FRT42D P[Ci+]/FRT42D hs-Myc;; ci94/ci94*. Flies were heat-shocked three times in 12 h and then dissected 7 days later.

RNA extraction and RT-qPCR analysis

Total RNAs from 15 ovaries or OSS cells were extracted with Trizol. After DNase treatment, cDNA was synthesized from 1 μg RNA using random primers and SuperScript III Reverse Transcriptase. qPCR was performed to assay levels of *flam*. *18S* or *rp49* RNA was used for the normalization. Fold changes were calculated using the delta Ct method [28]. Primers are listed in Supplementary Table S5.

RNA analyses

Small RNAs from OSS cells were extracted by Trizol. Deep sequencing was performed by Fasteris S.A. (Geneva/CH) on an Illumina Hi-Seq 2000 (Fasteris). RNA-seq libraries were analysed with *bowtie* mappers and were visualized using http://genomeview.org/. Small RNA sequencing data were analysed with NucBase [29]. For researches of mRNA or piRNA across exon junctions, reads were mapped on the reconstituted junction using *bowtie* [30] for mRNA and NucBase for small RNA.

Supplementary information for this article is available online: http://embor.embopress.org

Acknowledgements

The OSS cell line was a gift of Yuzo Niki (Ibaraki University). Flies with the following genotypes *y-hs-FLP; FRT42D ci[+];;ci[94]* and *y; FRT42D hs-MYC 45/CyO;;ci[94]/y[+]* were kindly provided by Robert Holmgren. We are grateful to Françoise Pellissier and Agostinha De Sousa for technical assistance. CG received a graduate grant from the Ministere de lEnseignement Superieur et de la Recherche (MESR). This work was supported by grants from the Region Auvergne, European Union (FEDER), the Ligue régionale contre le cancer and the Association Nationale de la Recherche (ANR) (project plasTiSiPi). We thank all members of our group for helpful discussion.

Author contributions

EB and CV conceived and designed the experiments. CG and SD performed most of the experiments. YR and CG analysed bioinformatic data. EB and CV wrote the manuscript.

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, Sachidanandam R, Hannon GJ (2007) Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. *Cell* 128: 1089 – 1103
2. Prudhomme N, Gans M, Masson M, Terzian C, Bucheton A (1995) Flamenco, a gene controlling the gypsy retrovirus of *Drosophila melanogaster*. *Genetics* 139: 697 – 711
3. Desset S, Meignin C, Dastugue B, Vaury C (2003) COM, a heterochromatic locus governing the control of independent endogenous retroviruses from *Drosophila melanogaster*. *Genetics* 164: 501 – 509
4. Sarot E, Payen-Groschêne G, Bucheton A, Pelisson A (2004) Evidence for a piwi-dependent RNA silencing of the gypsy endogenous retrovirus by the *Drosophila melanogaster flamenco* gene. *Genetics* 166: 1313 – 1321
5. Malone CD, Brennecke J, Dus M, Stark A, Mccombie WR, Sachidanandam R, Hannon GJ (2009) Specialized piRNA pathways act in germline and somatic tissues of the Drosophila ovary. *Cell* 137: 522 – 535
6. Saito K, Ishizu H, Komai M, Kotani H, Kawamura Y, Nishida KM, Siomi H, Siomi MC (2010) Roles for the Yb body components Armitage and Yb in primary piRNA biogenesis in Drosophila. *Genes Dev* 24: 2493 – 2498
7. Dennis C, Zanni V, Brasset E, Zhang L, Eymery A, Jensen S, Rong Y, Vaury C (2013) Dot COM, a nuclear transit center for the primary piRNA pathway in Drosophila. *PLoS ONE* 8: e72752
8. Klattenhoff C, Xi H, Li C, Lee S, Xu J, Khurana JS, Zhang F, Schultz N, Koppetsch BS, Nowosielska A *et al* (2009) The Drosophila HP1 homolog rhino is required for transposon silencing and piRNA production by dual-strand clusters. *Cell* 138: 1137 – 1149
9. Pane A, Jiang P, Zhao DY, Singh M, Schüpbach T (2011) The Cutoff protein regulates piRNA cluster expression and piRNA production in the Drosophila germline. *EMBO J* 30: 4601 – 4615
10. Li XZ, Roy CK, Dong X, Bolcun-Filas E, Wang J, Han BW, Xu J, Moore MJ, Schimenti JC, Weng Z *et al* (2013) An ancient transcription factor initiates the burst of piRNA production during early meiosis in mouse testes. *Mol Cell* 50: 67 – 81
11. Lo K, Smale ST (1996) Generality of a functional initiator consensus sequence. *Gene* 182: 13 – 22
12. Ohler U, Liao G-C, Niemann H, Rubin GM (2002) Computational analysis of core promoters in the *Drosophila* genome. *Genome Biol* 3: research0087-0087.12
13. Arkhipova IR (1995) Promoter elements in *Drosophila melanogaster* revealed by sequence analysis. *Genetics* 139: 1359 – 1369.
14. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CAM, Taylor MS, Engström PG, Frith MC *et al* (2006) Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet* 38: 626 – 635
15. Bensaude O (2011) Inhibiting eukaryotic transcription. Which compound to choose? How to evaluate its activity?. *Transcription* 2: 103 – 108.
16. Yee NS, Zhou W, Chun SG, Liang IC, Yee RK (2012) Targeting developmental regulators of zebrafish exocrine pancreas as a therapeutic approach in human pancreatic cancer. *Biol Open* 1: 295 – 307
17. Forbes AJ, Spradling AC, Ingham PW, Lin H (1996) The role of segment polarity genes during early oogenesis in Drosophila. *Development* 122: 3283 – 3294
18. Méthot N, Basler K (1999) Hedgehog controls limb development by regulating the activities of distinct transcriptional activator and repressor forms of Cubitus interruptus. *Cell* 96: 819 – 831

19. Bellen HJ (2004) The BDGP gene disruption project: single transposon insertions associated With 40% of Drosophila genes. *Genetics* 167: 761 – 781
20. Russo CA, Takezaki N, Nei M (1995) Molecular phylogeny and divergence times of drosophilid species. *Mol Biol Evol* 12: 391 – 404
21. Tamura K (2004) Temporal patterns of fruit fly (Drosophila) evolution revealed by mutation clocks. *Mol Biol Evol* 21: 36 – 44
22. Qi H, Watanabe T, Ku H-Y, Liu N, Zhong M, Lin H (2011) The Yb body, a major site for PIWI-associated RNA biogenesis and a gateway for PIWI expression and transport to the nucleus in somatic cells. *J Biol Chem* 286: 3789 – 3797
23. Sienski G, Dönertas D, Brennecke J (2012) Transcriptional silencing of transposons by PIWI and maelstrom and its impact on chromatin state and gene expression. *Cell* 151: 964 – 980
24. Mével-Ninio M, Pelisson A, Kinder J, Campos AR, Bucheton A (2007) The flamenco locus controls the gypsy and ZAM retroviruses and is required for Drosophila oogenesis. *Genetics* 175: 1615 – 1624
25. Zhang F, Wang J, Xu J, Zhang Z, Koppetsch BS, Schultz N, Vreven T, Meignin C, Davis I, Zamore PD *et al* (2012) UAP56 Couples piRNA clusters to the perinuclear transposon silencing machinery. *Cell* 151: 871 – 884
26. Bartolomé C, Bello X, Maside X (2009) Widespread evidence for horizontal transfer of transposable elements across Drosophila genomes. *Genome Biol* 10: R22
27. Zanni V, Eymery A, Coiffet M, Zytnicki M, Luyten I, Quesneville H, Vaury C, Jensen S (2013) Distribution, evolution, and diversity of retrotransposons at the flamenco locus reflect the regulatory properties of piRNA clusters. *Proc Natl Acad Sci* 110: 19842 – 19847.
28. Livak KJ, Schmittgen TD (2001) Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta CT}$ method. *Methods* 25: 402 – 408
29. Dufourt J, Pouchin P, Peyret P, Brasset E, Vaury C (2013) NucBase, an easy to use read mapper for small RNAs. *Mob DNA* 4: 1
30. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Meth* 9: 357 – 359

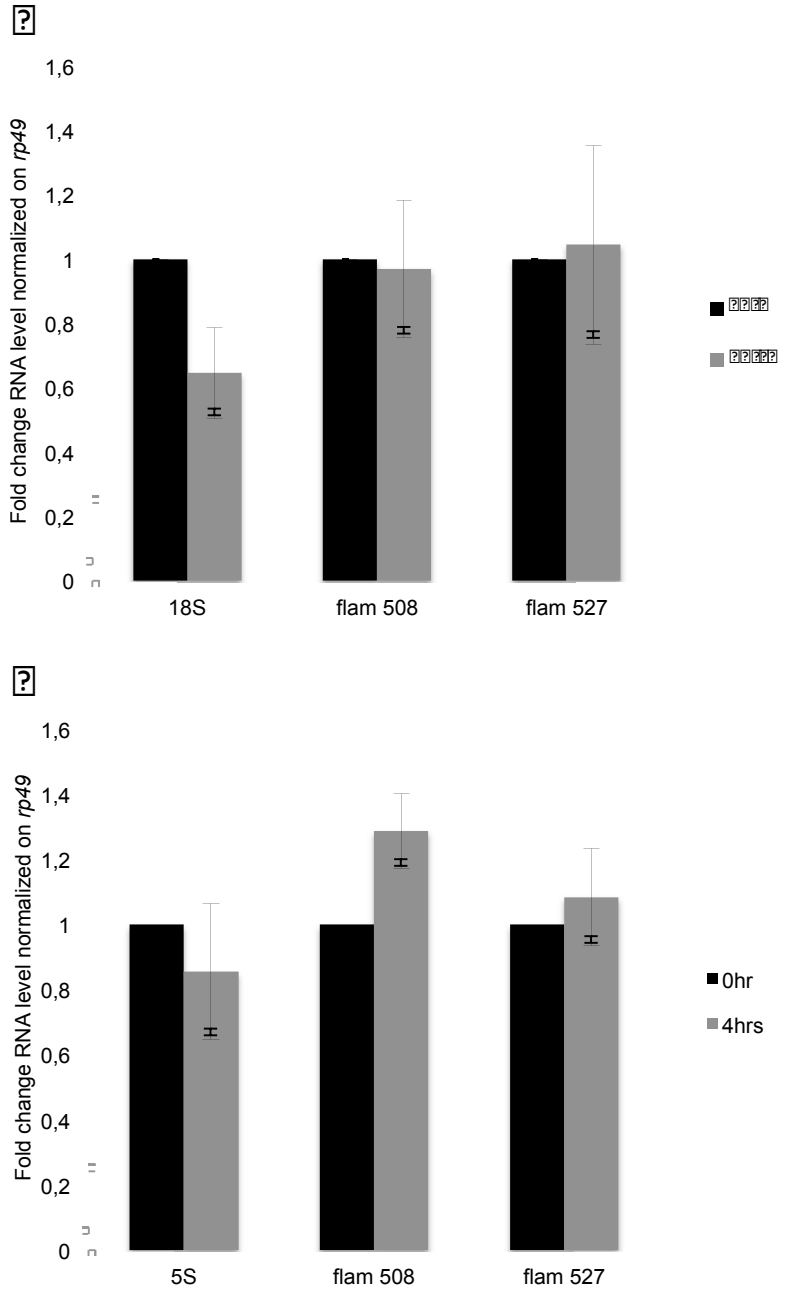
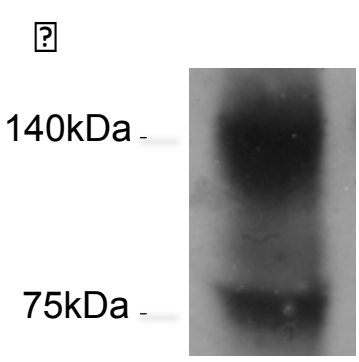
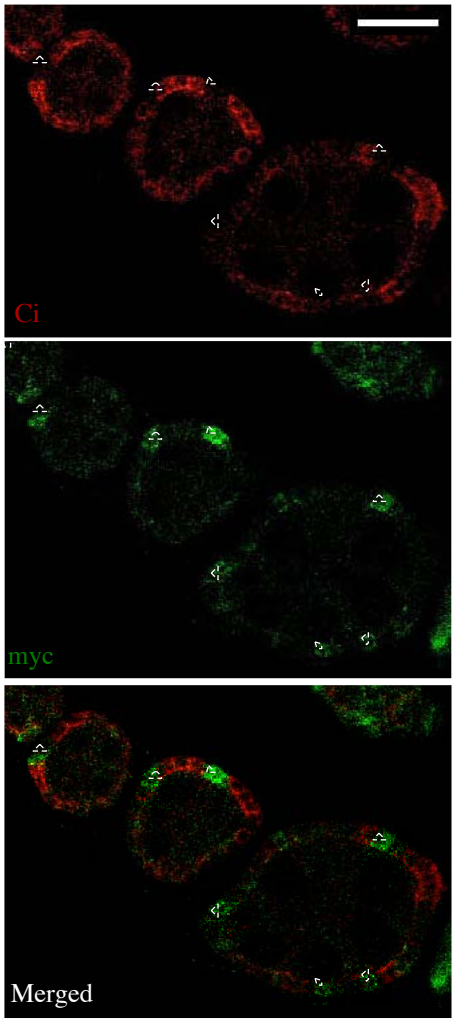
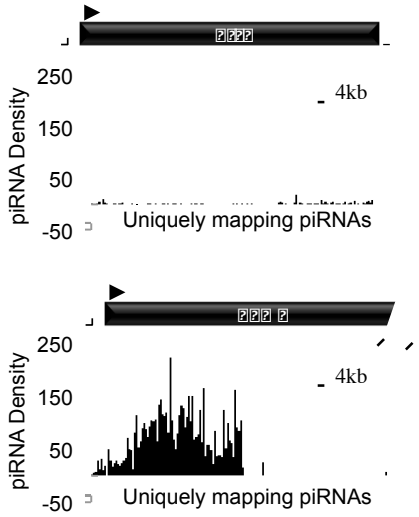


Figure S1: (A) RT-qPCR on total RNAs from OSS cells treated with Actinomycin D during 4 hours. 18S ribosomal RNA known to be transcribed by Pol I was used as positive control. Data are normalized to *rp49* expression and are presented as means (n=4). Error bars represent ± s.e.m. Expression in untreated OSS cells (0hr) was set to one. (B) RT-qPCR on total RNAs from OSS cells treated with a Pol III inhibitor (ML-60218) during 4 hours. 5S ribosomal RNA known to be transcribed by Pol III was used as positive control. Data are normalized to *rp49* expression and are presented as means (n=3). Error bars represent ± s.e.m. Expression in untreated OSS cells (0hr) was set to one.


140kDa
75kDa


Ci
myc
Merged


piRNA Density
250
150
50
-50
4kb
Uniquely mapping piRNAs
piRNA Density
250
150
50
-50
4kb
Uniquely mapping piRNAs

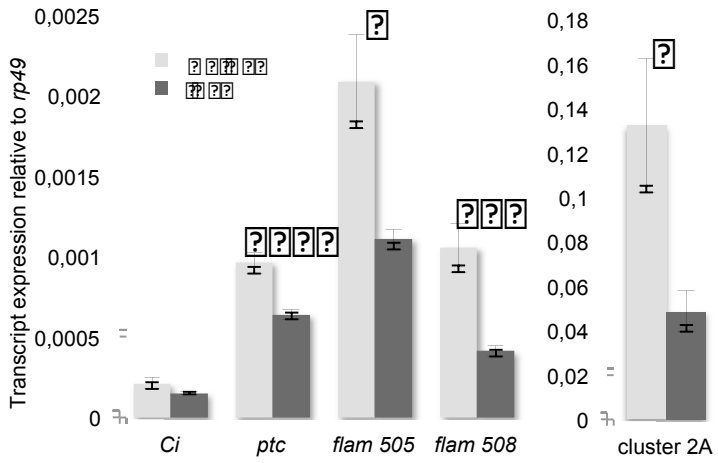

Transcript expression relative to rp49
0,0025
0,002
0,0015
0,001
0,0005
0
Ci
ptc
flam 505
flam 508
0,18
0,16
0,14
0,12
0,1
0,08
0,06
0,04
0,02
0
cluster 2A

Figure S2: (A) Western Blot against Ci on OSS cells protein extraction. The two isoforms of Ci protein are detected in OSS cells. (B) Mutant clones for Ci. Ci (red) and Myc (green) immunostaining of ovarioles from *hs-flp*; *FRT42D P[Ci+] /FRT42D hs-MYC 45*; *Ci94*/*Ci94*. No Ci expression is detected in homozygote Myc cells visualized by a white dotted line. Scale bar, 20µm (C) piRNA profiles (genome unique; sense up, antisense down) from OSS cells mapping to the *ptc* gene (a gene known to be activated by Ci) and the piRNA cluster *flam.* (D) RT-qPCR on total RNAs from control ovaries (no HS) or from ovaries with a clonal population mutant for Ci (HS). The positions of primers used for PCR amplification are indicated in the Fig 4. *ptc* is a gene known to be activated by Ci in follicle cells. Data are presented as means (n>5). Error bars represent ± s.e.m.; Student's *t*-test: *=P<0,05; **=P<0,01; ***=P<0,001.

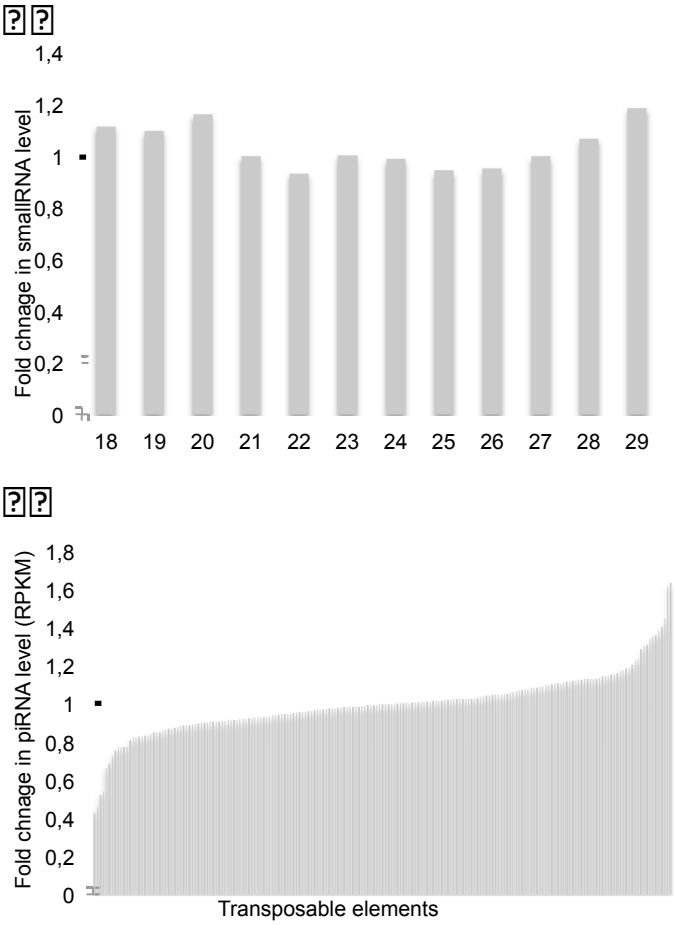
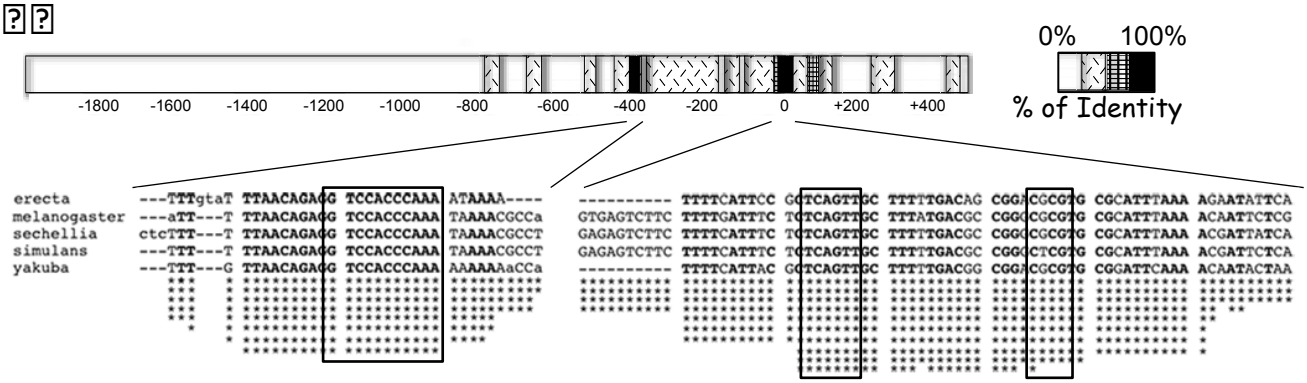
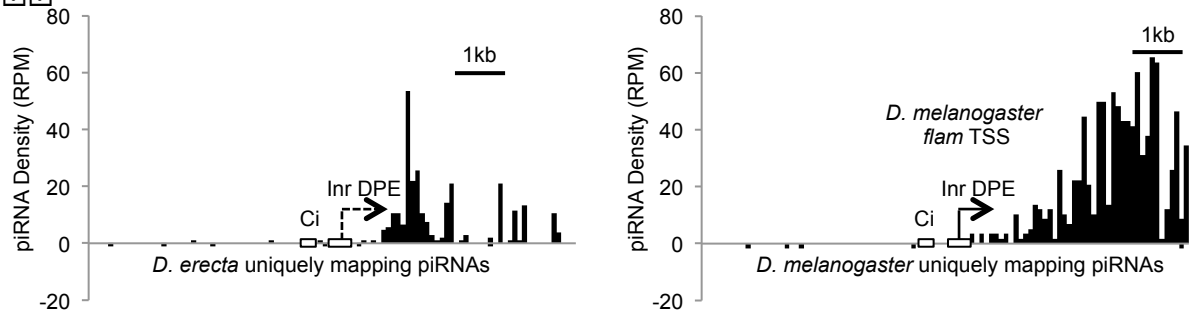


Figure S3: (A) Graph represents the fold change in length repartition of total small RNA sequenced from OSS cells transfected with siRNAs against Ci compared to OSS cells transfected with siRNAs against GFP, two days post-transfection. (B) Graph represents the fold change of transposable element piRNAs in OSS cells transfected with siRNAs against Ci and compared to OSS cells transfected with siRNAs against GFP. piRNAs are counted and normalised per RPKM on the 238 transposable elements indexed in repbase.

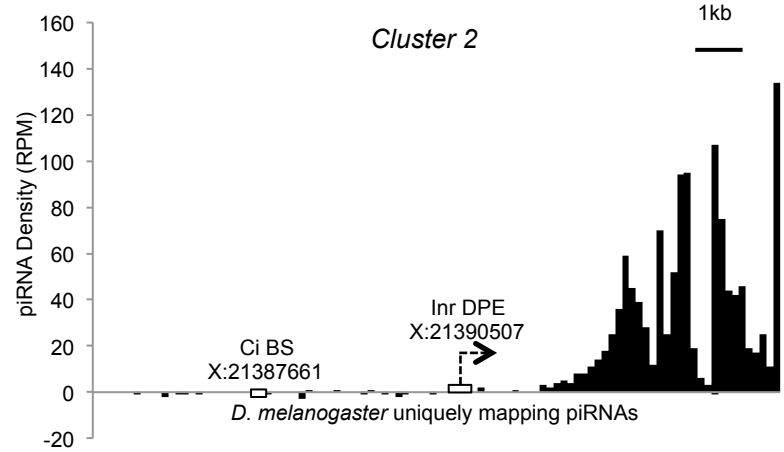
??


0%
100%
% of Identity
-1800
-1600
-1400
-1200
-1000
-800
-600
-400
-200
0
+200
+400
erecta
melanogaster
sechellia
simulans
yakuba

??


piRNA Density (RPM)
1kb
Inr DPE
Ci
D. erecta uniquely mapping piRNAs
D. melanogaster
flam TSS
D. melanogaster uniquely mapping piRNAs

??


Cluster 2
1kb
piRNA Density (RPM)
Inr DPE
X:21390507
Ci BS
X:21387661
D. melanogaster uniquely mapping piRNAs

??????? ???? ????????????

Figure S4: (A) piRNA profiles (genome unique; sense up, antisense down) from OSS cells mapping to the beginning of piRNA cluster 2. Positions of the putative Inr-DPE sequences and the Ci binding site are indicated by rectangles. Dotted arrow indicated the putative cluster 2 promoter. (B) DNA sequence alignments of the *flam* promoter from different *Drosophila melanogaster* subgroup species from 2kb upstream to 500bp downstream of the TSS. The two more conserved regions are detailed below. The Ci binding site, Inr and DPE are indicated by rectangles. (C) Uniquely mapping *D. erecta and D. melanogaster* piRNAs are plotted over the putative *flam* cluster (normalized per million of reads) [29]. The Ci binding site and the Inr-DPE sequence are indicated. *D. melanogaster flam* TSS and the putative *D. erecta flam* promoter are represented by an arrow and dotted arrow respectively.

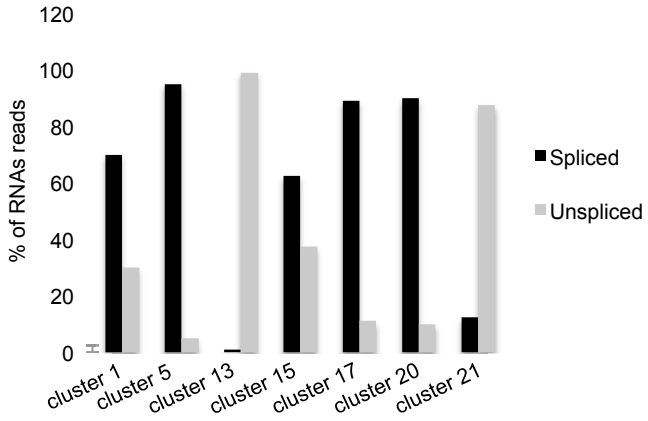


Figure S5: Reads encoded by several piRNA clusters reveal splicing events. Graphs represent the percentage of reads from [21] corresponding either to the predicted non-spliced Exon/Intron junction (grey) or to the predicted spliced Exon/Exon junction (black) from 7/21 piRNA clusters analyzed. Spliced reads have been visualised thanks to tophat.

		Ovaries		OSS cells	
TSS	position	Extraction1	Extraction2	Extraction1	Extraction2
1	21502904			X	
2	21502915		X		
3	21502918	X	X	X	XX
4	21502983			X	
5	21503199		X		
6	21504137		X		
7	21504639	X			
8	21506316	X			
9	21506855	X			
10	21507493	X			
11	21507571	X			

Table S1: RNA extractions in which the different TSSs were found.

species	flam's TSS position
melanogaster	X:21502918
simulans	X:16575589
sechellia	Scaffold423:1584
yakuba	X:20625060
erecta	Scaffold4690:17965703

Table S2: Best putative *flam* TSS position in *Drosophila melanogaster* subgroup species.

	length	Donor site		Acceptor site	
21506829..21507115	286	tgaacttacg	GTaatatcc	tgtttttcAG	atttcagatt
21503350..21504064	714	ttttcaagcg	GTaagtgttt	cacctttAG	atattagctt
21504108..21504985	877	tacaattgtg	GTacgtgatt	tttttttcAG	agcattggac
21503350..21504985	1635	ttttcaagcg	GTaagtgtta	tttttttcAG	agcattggac
21505027..21506774	1747	ccggtttgct	Gcaagtattt	tttattttAG	ctccttttt
21505134..21507115	1981	accagccaag	GTatttaata	tgtttttcAG	atttcagatt
21505134..21508851	3715	accagccaag	GTatttaata	tatctgatAG	tcggggaact
21503350..21507132	3782	ttttcaagcg	GTaagtgttt	attaccattt	ggctatgagg
21505134..21540689	35555	accagccaag	GTatttaata	tgatttaaat	tatcatgctc
21505089..21540689	35600	cgtgtttttt	ttttgcgct	gatttaaatt	atcatgctc
21505134..21594560	89423	accagccaag	GTatttaata	tttccatcAG	tgtcccgtct
21503500..21661658	158158	tcctcctact	GTtttggaat	cgccggcgct	gcttccggtg

Table S3: Sequence upstream and downstream of the donor and acceptor sites of *flam* introns. Grey represents intron. Consensus sites are written in capital letters and underlined.

Exon	position	length
1	21504986..X	
2	21507116..X	
3	21508851..X	
4	21540689..X	
5	21594560..X	
6	21507132..X	
7	21507116..X	
8	21504986..21505026	40
9	21504065..21504107	42
10	21506775..21506828	53
11	21504986..21505088	102
12	21504986..21505133	147
13	21502918..21503349	431
14	21502918..21503500	582

Table S4: Coordinates and length of different exons find in *flam*. X represented undefined end.

Experiment	Primer Name	Primer sequence
q-PCR and RT-qPCR	Dip1 499 F	TCGTCCGACCAGCAGAAT
	Dip1 499 R	TCCTCTATATTGTCGACGGA
	flam 502 F	CGAAAACAAATCAGGATCAAA
	flam 502 R	AAAACTTTTGCGACAGTATTAGGT
	flam 503 F	CGCATTTAAAACAATTCTCG
	flam 503 R	TTTCGTTGTTGTTTCGCTTAG
	flam 505 F	TTATTTCTATGCCGGTTTGC
	flam 505 R	GTTCGCTTGAAAGCTAGGAA
	flam 507 F	TTGGCTATGAGGATCAGACA
	flam 507 R	CTTCAAAGCGATTCATTCCT
	flam 527 F	GTGGCTTCACAAAACACGAC
	flam 527 R	GCCGGTCCTAAATATCTTCTC
	flam 508 F	ATTCTCCTTTCTCAGGATGC
	flam 508 R	GCATTGCTACCTTACGTTTC
	flam 633 F	ATGTAACAGGTATAGATGTAGTA T
	flam 633 R	CACATAGTCTTAAGCACGCCT
	Ci R	TCTCATTACTGTGTGTTCGATTT
	Ci F	GCAGTATATGCTTGTTGTGCAT
	Ptc R	GAGGGAAGCCAGCTGTTG
	Ptc F	CATCACAGAGGCGGGATT
	Cluster2A F	GCCTACGCAGAGGCCTAAGT
	Cluster2A R	CAGATGTGGTCCAGTTGTGC
	Cluster2B F	CTGCTTTGTGCTTGGAGATG
	Cluster2B R	TCTGCACAGATTCTGAAATTGAA
	Actine F	AAGTTGCTGCTCTGGTTGTCG
	Actine R	GCCACACGCAGCTCATTGTAG
	Blood F	GACTTACATGGCATGGATTGA
	Bloood R	GAATTCTTAAATCAAATCGGCAG
	Gypsy F	AGTTGTGTATCTGGCCACGT
	Gypsy R	CTTTGCCGAAAATATGCAATGT
	Idefix F	TCCAGACCAACCAAAGAAGC
	Idefix R	TCCATTGTTCCTGTTTGGAA
	Mdg1 F	ATATGAATTCCATATCGTACCTG
	Mdg1 R	CCTCAAAGTGAACCAATCTTC
	Pifo F	GCATTCAACGCCAATAACCT
	Pifo R	TTGTCTCAACTCCGTTGTCG
	Roo F	CGTCTGCAATGTACTGGCTCT
	Roo R	TTCTTCACTTTCGGACTGAATG
	Tabor R	ACGTTGTTCACGACATTAGCCG
	Tabor F	GGGTTGGTTCGGATCTGACG

Experiment	Primer Name	Primer sequence
5’- and 3’-race	A	ATCAAGGAGGAGCCGATCTC
	B	ATTTGTTAACCGGACTTTGC
	C	TCGTCGTGCATTTGCTGGAA
	D	ATGCAGCGCAGTTTGACACT
	E	TACACTGCTGCTGAGCTGAT
	F	AAACACGACCGTTCGCTTGA
	G	CAAACAGCACTGCGAATCACCTTTAA
	H	CAGAAAATTAAGCGGAAGCCCCAC
	I	CACCCAAGATTTTGTACATTTTCAAGC
	J	AAGATTTTGTACATTTTCAAGCGAGCA

Tables S5: Sequences of primers used in this study.

C. Conclusion et perspectives

Les expériences de 5'race-PCR réalisées sur des ARN extraits d'ovaires ou de cellules OSS et associées à des analyses *in silico* ont permis de mettre en évidence un promoteur majoritaire de *flamenco* à la position X:21.502.918 (release 5.57). Ce promoteur possède de nombreuses caractéristiques des promoteurs canoniques des gènes reconnus par l'ARN polymérase II, telle la présence de séquences consensus caractéristiques. Des expériences de ChIP et d'inhibition de l'ARN polymérase II ont d'ailleurs démontré l'implication de cette polymérase dans la transcription de *flamenco*. L'analyse des activités transcriptionnelles de séquences génomiques spécifiques grâce à l'utilisation du système de double gènes rapporteurs dans les cellules OSS a confirmé le rôle du promoteur mis en évidence et a également permis de montrer que cette transcription est activée par le facteur de transcription (FT) Cubitus interruptus (Ci).

Dans l'ensemble, ces données montrent que le locus *flamenco* possède sa propre unité de transcription de façon comparable aux gènes codants pour des protéines. Cette caractéristique est commune aux clusters unidirectionnels du stade pachytène des cellules germinales mâles de la souris. De plus, même si beaucoup d'ARN non codants sont issus d'événements de transcription dus à des phénomènes d'échappement et d'aspécificité de l'ARN polymérase II, un grand nombre d'ARN non codants produits par les génomes sont également transcrits à partir d'un promoteur canonique de l'ARN polymérase II dont l'activité est contrôlée par un ou plusieurs FT(s) (Guttman et al. 2009). Cette régulation permet leur expression spécifique dans certains types cellulaires où leur fonction est requise.

La voie de signalisation Hedgehog contrôle de nombreux processus développementaux, notamment lors de la mise en place des disques imaginaux de l'aile chez la larve de drosophile. La mutation d'une des protéines de cette voie est d'ailleurs létale pour l'organisme dès les stades larvaires. Cette voie est conservée chez les mammifères où elle est particulièrement importante pour le développement correct de l'embryon. Une fois adulte, la voie Hedgehog est nécessaire au maintien de certaines cellules souches, notamment dans les ovaires de drosophile où elle joue un rôle pour le maintien et la division des cellules somatique souches (SSC) qui sont à l'origine des cellules folliculaires. La protéine Hedgehog (hh) est exprimée au pôle apical de l'ovariole, dans les cellules de la niche, et diffuse à travers les cellules de la gaine interne (ISC) jusqu'à atteindre les SSC **(Figure 35)**. A la surface de ces cellules, hh inhibe ptc, un récepteur membranaire à l'origine d'une cascade d'activation

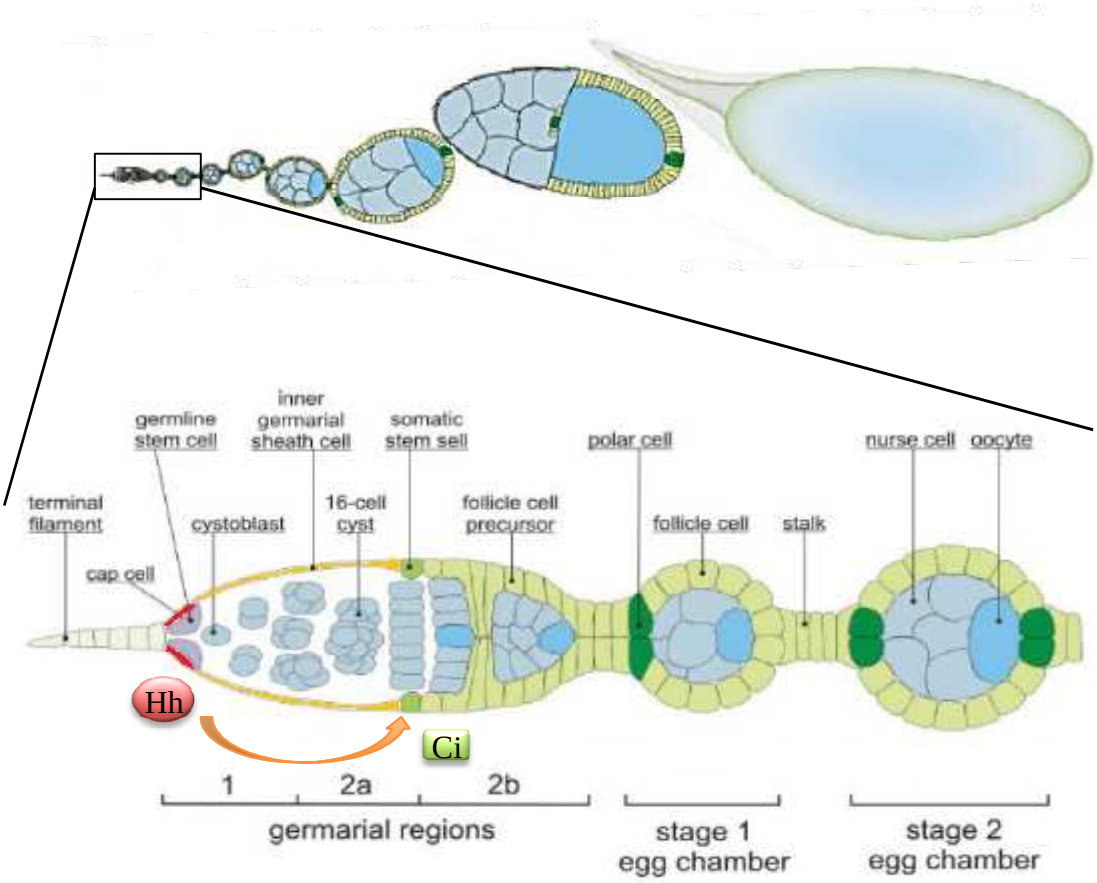


Figure 35 : La voie Hedgehog dans les ovaires de drosophile

La protéine Hedgehog est exprimée dans les cellules de la cap (en rouge), au pôle apical de l'ovariole, puis se déplace à l'intérieur des cellules de la gaine interne (en orange) jusqu'au cellules souches des cellules folliculaires (en vert pomme) où elle activera le facteur de transcription Ci. Cette voie est requise pour la division des cellules souches et le maintien des cellules folliculaires.

[Adapté de Sally Horne-Badovinac and David Bilder, Developmental Dynamic]

qui conduit à l'inhibition de Ci. La présence de hh empêche donc cette inhibition et permet la relocalisation de Ci dans le noyau où il pourra activer de nombreux gènes cibles, dont *flamenco*. Toutefois, l'expression spécifique de *flamenco* dans les cellules folliculaires nécessite certainement l'implication d'autres facteurs de transcription, puisque Ci est actif dans d'autres tissus, notamment aux stades larvaires, dans lesquels l'expression de *flamenco* n'est pas attendue. Pour étudier cette hypothèse, le système de double gènes rapporteurs luciférases peut à nouveau être utilisé afin d'analyser l'impact sur l'activité transcriptionnelle du promoteur de *flamenco* de séquences génomiques situées plus en amont ou en aval de celui-ci. Il est également possible de générer des délétions de taille réduite à l'intérieur du promoteur afin de cibler l'analyse sur une zone en particulier.

Dans une autre partie de l'étude nous avons vu, grâce à des expériences de RT-PCR ainsi que l'analyse d'un séquençage des longs ARN, que le long ARN issu de la transcription de *flamenco* est épissé de façon alternative avant d'être maturé en piRNA. En lignée germinale, il a récemment été démontré que l'épissage et le coiffage des transcrits issus des piRNA clusters germinaux étaient bloqués par le complexe RDC formé des protéines Rhino-Cuff-Del en partenariat avec la protéine UAP56 (Mohn et al. 2014b; Zhang et al. 2014). Toutefois, ces protéines ne sont pas présentes sur les clusters de piRNA folliculaires et j'ai pu montré que le transcrit issu de *flamenco* est épissé. Ainsi, mes données suggèrent que, à l'instar du mode de transcription et de la structure chromatinienne, l'épissage des ARN précurseurs de piRNA diffère entre les transcrits issus de clusters germinaux et ceux issus de clusters folliculaires.

En cellules somatiques, le premier exon de *flamenco* est conservé dans une grande majorité de transcrits quel que soit l'épissage en aval. Cet exon pourrait donc être nécessaire à la maturation de l'ARN de *flamenco* en piRNA en contenant, par exemple, des séquences qui seraient reconnues par des facteurs qui permettraient l'adressage du transcrit de *flamenco* après épissage vers les Yb-bodies cytoplasmiques où l'ARN sera maturé en piRNA. De même, un intron situé au début de *flamenco* (+432 ;+2067) semble être majoritairement épissé en comparaison des autres introns qui ne sont retrouvés épissés que dans quelques séquences. Il est possible que l'épissage de cet intron permette le recrutement de facteurs d'épissages qui pourraient jouer un rôle dans la maturation en piRNA. Par exemple, dans les cellules germinales, la protéine UAP56 qui fait partie du complexe d'épissage EJC (Exon Junction Complex), fixe les transcrits des clusters germinaux pour les transférer vers le nuage où ils seront maturés en piRNA (Zhang et al. 2012).

La fonction de l'épissage alternatif observé sur le transcrit de *flamenco* après le premier intron pourrait avoir plusieurs fonctions. La première serait de répondre à une nécessité mécanique de réduction de la taille du transcrit issu de *flamenco*. En effet, il n'existe pas d'autres ARN aussi longs et même l'ARN de la dystrophine est considérablement réduit après épissage (les différents ARNm issus du transcrit de la dystrophine n'excèdent pas 14kb). Il est fort probable qu'une taille trop importante soit incompatible avec le bon déroulement de certains processus biologiques. La réduction de la taille du transcrit issu de *flamenco* pourrait être une nécessité afin d'assurer sa maturation en piRNA. Face à cette nécessité, un épissage alternatif permet de réduire la taille du transcrit de *flamenco* tout en garantissant la présence de toutes les séquences génomiques de *flamenco* dans au moins une structure ARN épissée. Ceci permettrait d'obtenir tout un pool d'ARN de structures différentes dont l'ensemble permettra de produire des piRNA homologues à la totalité du locus *flamenco*. Comme l'a vu notre équipe, le locus *flamenco* est un locus génomique particulièrement dynamique qui subit de nombreux événements de recombinaison, délétion ou insertion d'ETs. Afin d'assurer la régulation d'un ET qui s'insérerait à n'importe quel endroit du locus, il est donc primordial que chaque séquence de *flamenco* puisse produire des piRNA.

II. Analyse de l'adressage des transcrits *flamenco* à la voie des piRNA

A. Introduction

L'une des grandes inconnues de la voie des piRNA est de comprendre comment la cellule reconnaît les transcrits issus des clusters de piRNA et les adresse à la voie de maturation des piRNA. Des données récentes montrent que les piRNA hérités de la mère seraient impliqués dans la reconnaissance de la séquence génomique des piRNA clusters germinaux chez la drosophile en favorisant l'émergence d'un contexte chromatinien spécifique au niveau de ces clusters. Toutefois, ce mécanisme ne semble pas pouvoir s'appliquer au locus *flamenco*. En effet, celui-ci est caractérisé par une transcription similaire aux gènes classiques et une structure chromatinienne exempte de marques hétérochromatiques ou des protéines Rhino et Cutoff. De plus, on ne retrouve aucun piRNA provenant de *flamenco* dans les séquençages des petits ARN issus d'embryons précoces, contrairement aux piRNA des clusters germinaux, ce qui montre que les complexes piRISC formés dans les cellules folliculaires ne sont pas transmis à la descendance. Il semble donc que ni ces piRNA ni la structure chromatinienne ne soient en mesure d'expliquer l'adressage du transcrit *flamenco* vers la voie des piRNA. On s'est donc demandé si le promoteur de *flamenco* est suffisant pour déclencher la maturation du transcrit de *flamenco* en piRNA.

Si c'est le cas, on peut s'attendre à trouver des piRNA provenant de l'ARN luciférase dans les cellules transfectées avec un plasmide contenant le gène de la luciférase firefly sous contrôle du promoteur de *flamenco*. Afin de répondre à cette question, nous avons donc séquencé les petits ARN présents dans les cellules OSS transfectées avec trois plasmides différents : le plasmide SFI, qui semble être le promoteur minimal nécessaire à une expression correcte de *flamenco*, le plasmide SFET, qui contient la totalité du promoteur proximal de *flamenco* à partir du gène DIP1 et jusqu'aux premières séquences d'ETs en aval du TSS, ou le plasmide pGL3control, qui contient le promoteur viral SV40 **(Figure 36)**.

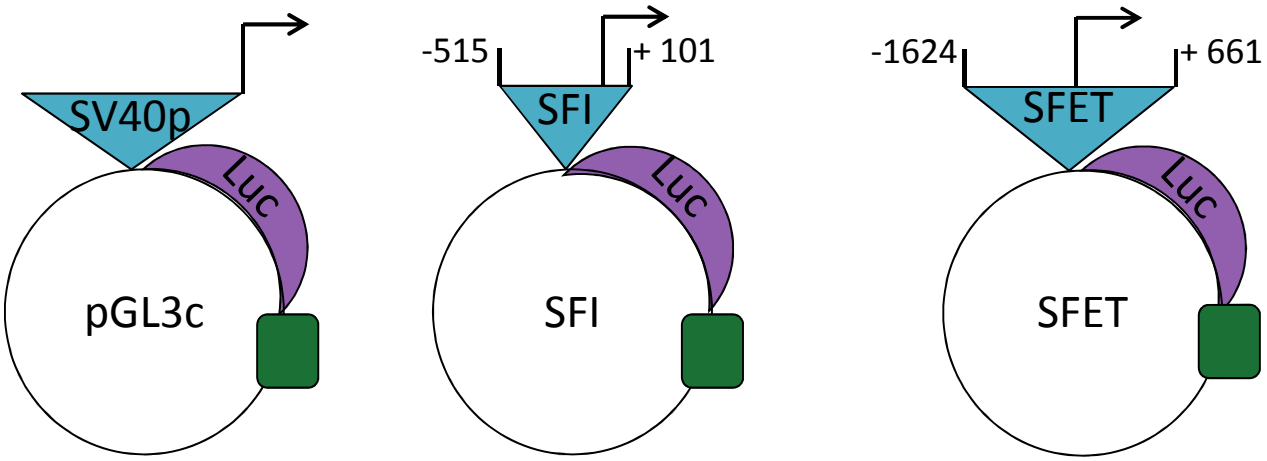


Figure 36 : Les différents plasmides transfectés en cellules OSS

Présentation de la carte schématique des trois plasmides utilisés pour étudier l'adressage du transcrit *flamenco* à la voie de maturation des piRNA. Le gène codant pour la luciférase firefly est représenté en violet, le signal de poly-adénylation SV40 en vert. Les différents promoteurs sont figurés par les triangles bleu. Les promoteurs SFI et SFET correspondent à des portions plus ou moins grandes du promoteur de *flamenco*. Les chiffres indiquent le premier et le dernier nucléotides de la séquence génomique de *flamenco* incluse dans le plasmide. La flèche indique le départ de la transcription

B. Résultats

Les séquençages ont été analysés par le logiciel Nucbase afin de rechercher des petits ARN homologues à la séquence des plasmides transfectés (Dufourt et al. 2013). Afin de ne pas perturber les résultats avec les petits ARN endogènes, seules les séquences non retrouvées sur le génome de la drosophile ont été analysées, c'est-à-dire la totalité des plasmides pGL3 sans les séquences promotrices de *flamenco*. Le nombre de reads homologues à la séquence de pGL3 en sens et en antisens a été normalisé en million de reads **(Figure 37 A)**. Les résultats ont ensuite été exprimés en nombre de reads sur 50pb afin de pouvoir comparer la distribution de ces reads le long de la séquence du plasmide **(Figure 37 B)**. De plus, la distribution des tailles des petits ARN retrouvés a été calculée ainsi que la proportion de petits ARN possédant un nucléotide Uridine en première position **(Figure 37 C)**.

Lors du séquençage des petits ARN provenant des cellules OSS transfectées avec le plasmide pGL3 control, un nombre non significatif de petits ARN homologues à la luciférase a été retrouvé. En revanche, plusieurs dizaines de petits ARN par million de reads sont retrouvés spécifiquement homologues au site de poly-adénylation SV40 situé à la fin de la séquence codante de la luciférase. Ces petits ARN sont de la taille des piRNA (entre 24 et 29 nucléotides), mais ils ne possèdent pas de biais de nucléotide Uridine en position 5’ et sont orientés en antisens par rapport au sens de transcription déterminé par le promoteur SV40.

Les plasmides contenant des séquences promotrices de *flamenco* présentent de nombreux petits ARN homologues non seulement à la luciférase mais aussi à l’ensemble du plasmide. Ces petits ARN peuvent être divisés en deux catégories. La première partie correspond aux petits ARN antisens homologues à la séquence de poly-adénylation SV40 et qui ont également été retrouvés dans les cellules OSS transfectées avec le plasmide pGL3control. La deuxième partie contient de nombreux petits ARN orientés en sens et majoritairement long de 24 à 29 nucléotides. De plus, ils présentent un fort biais Uridine pour leur premier nucléotide (environ 75%). Ils présentent donc de nombreuses caractéristiques des piRNA. Le plasmide SFI produit seulement quelques petits ARN homologues à la luciférase par million de séquence, tandis que le plasmide SFET en produit plusieurs dizaines.

Les différences de quantité de petits ARN peuvent refléter une différence de quantité d’ARN luciférase. Pour cela, j’ai analysé l’activité transcriptionnelle de ces trois promoteurs en quantifiant par RT-qPCR l’ARN de la luciférase firefly ainsi que l’activité enzymatique de

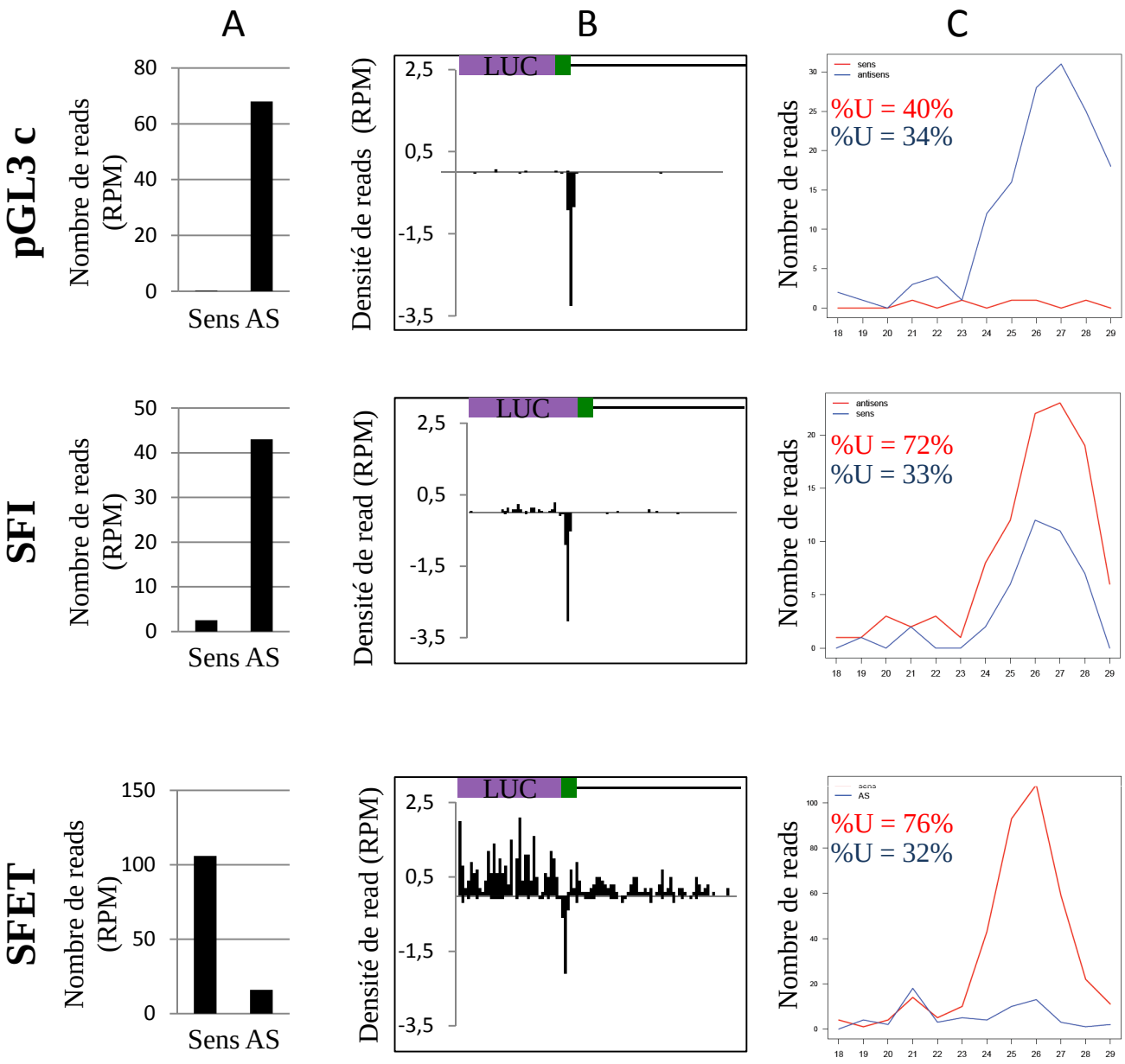


Figure 37 : Analyse des petits ARN présents dans les cellules OSS transfectées

Les lignes 1, 2 et 3 correspondent aux cellules transfectées avec le plasmide pGL3control, SFI et SFET respectivement.

A. Nombre de reads homologues à la séquence du plasmide pGL3 dans les cellules OSS transfectées. Les reads sont répertoriés en fonction de leur orientation sens ou antisens
B. Répartition des reads le long de la séquence du plasmide pGL3. Le nombre de reads différents sur 50pb normalisé par million de séquences est représenté.
C. Répartition de la taille des reads. Les reads sens sont représentés en rouge, les antisens en bleu. Le pourcentage de reads commençant par un U dans les reads sens ou antisens est indiqué

cette protéine **(Figure 38)**. On observe moins d'ARN luciférase dans les cellules transfectées avec le plasmide SFI en comparaison de celles transfectées avec le plasmide SFET. La différence de quantité d'ARN est similaire à la différence du nombre de petits ARN présents dans les cellules transfectées par ces deux plasmides. De plus, le plasmide pGL3control possède une très faible activité transcriptionnelle, voire quasi-nulle. Le dosage de l'activité enzymatique confirme les résultats de RT-qPCR.

C. Conclusion et perspectives

De nombreux petits ARN homologues à la luciférase ont été observés dans les cellules OSS transfectées avec des plasmides contenant le promoteur de *flamenco*. Ces petits ARN présentent de nombreuses caractéristiques des piRNA primaires : la taille, le biais de nucléotide U en première position et l'orientation en sens par rapport au transcrit. De plus, ils ne sont pas retrouvés dans les cellules OSS transfectées avec le plasmide pGL3control. Il semble donc que la présence du promoteur de *flamenco* soit suffisante pour déclencher l'adressage des transcrits à la voie de maturation des piRNA. Les petits ARN sont plus nombreux à partir de transcrits issus du promoteur complet de *flamenco* (plasmide SFET), en comparaison de ceux issus du promoteur minimal (SFI). Toutefois cette différence peut être due à la différence d'activité transcriptionnelle observée entre ces deux séquences promotrices. Les signaux nécessaires à l'adressage des transcrits à la voie de maturation des piRNA seraient donc contenus dans le plasmide SFI et donc entre les nucléotide -515 ; +101 du locus *flamenco*. Ces signaux pourraient servir à recruter des protéines permettant de transférer les transcrits vers les Yb-bodies où ils seront maturés en piRNA. Une analyse approfondie de cette portion génomique est maintenant nécessaire.

Le plasmide pGL3control ne produit pas de petits ARN homologues à la luciférase et orientés en sens. Mais il possède une activité transcriptionnelle quasi nulle et ne produit que très peu d'ARN luciférase. Après vérification, il semble en effet que le promoteur SV40 contenu dans ce plasmide soit peu actif dans les cellules de drosophile (Qin et al. 2010). C'est pourquoi ce contrôle ne nous permet pas encore de vérifier que l'adressage des transcrits à la voie des piRNA est une caractéristique spécifique des plasmides contenant des séquences promotrices de *flamenco*. Il est donc nécessaire de construire un nouveau plasmide contenant

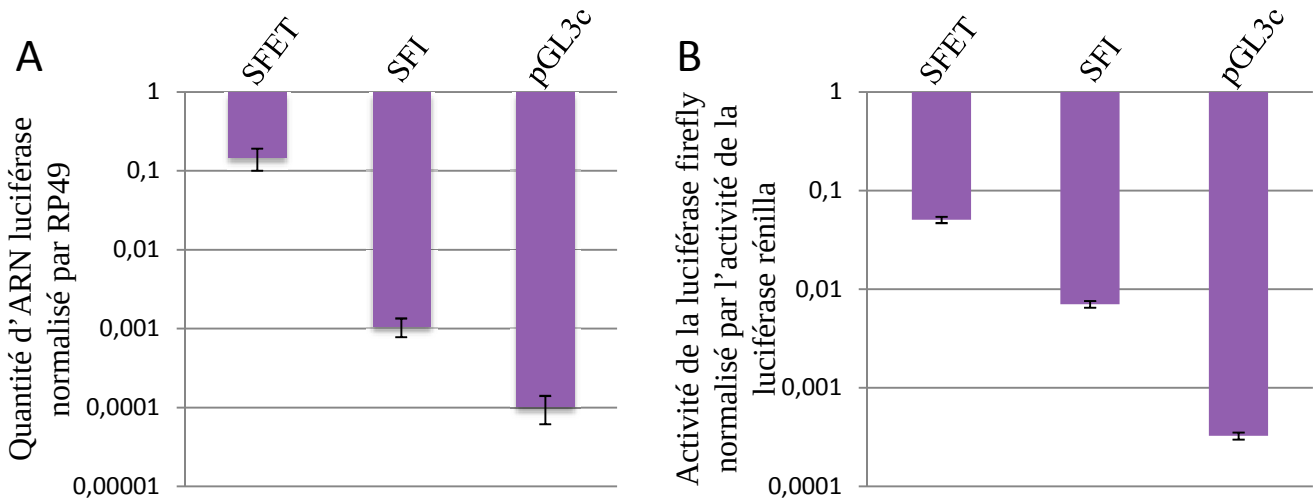


Figure 38 : Quantification de l'activité transcriptionnel du promoteur des plasmides utilisés

A. Quantification de l'ARN luciférase par RT-qPCR. Les résultats sont normalisés par rapport au gène de ménage rp49. L'échelle des ordonnées est logarithmique.
B. Dosage de l'activité de la luciférase firefly. L'échelle des ordonnées est logarithmique.

un gène rapporteur sous contrôle d'un promoteur actif en cellules OSS (par exemple, le promoteur de l'actine).

Les petits ARN sens sont retrouvés en majorité sur la luciférase mais également après le signal de poly adénylation SV40, ce qui suggère que le signal d'arrêt de la transcription situé à la fin du gène de la luciférase firefly est mal interprété par la cellule. Pourtant, ce signal est reconnu pour être efficace (Carswell & Alwine 1989). Même si on ne peut exclure que, comme le promoteur, ce signal soit mal reconnu par les cellules OSS, c'est peut être une indication que le promoteur de *flamenco* est capable de déclencher un mécanisme d'inhibition des signaux de poly-adénylations de façon similaire aux mécanismes observés sur les clusters germinaux. Ces signaux sont d'ailleurs largement présents dans la séquence génomique de *flamenco*, notamment à l'intérieur des séquences d'ETs. Si on sait que le locus *flamenco* n'est pas fixé par le complexe RDC (Rhino-Cuff-Del), on peut imaginer qu'un autre complexe protéique au rôle similaire fixe le promoteur de *flamenco* en somatique.

En plus des petits ARN homologues à la luciférase, des petits ARN particuliers ont été retrouvés. Ces petits ARN sont longs de 24 à 29 nucléotides mais orientés en antisens par rapport au sens de la transcription de la luciférase et ne présentent pas de biais de U au niveau de leur premier nucléotide. On ne peut donc pas les considérer comme des piRNA primaires. Ces petits ARN sont homologues à la séquence du signal de poly-adénylation SV40. Ces petits ARN sont retrouvés également dans les ARN de cellules OSS non transfectées (voir Annexe 1). Pourtant, cette séquence n'est pas retrouvée dans le génome de la drosophile. Les cellules OSS sont dérivées d'une lignée de drosophile contenant une mutation du gène *bam* par insertion d'un transgène dérivé de l'élément P. Il est possible que ce transgène contienne une séquence homologue au signal de poly-adénylation SV40 et que cette séquence soit transcrite en orientation inverse avant d'être dégradé en petit ARN. De façon surprenante, le nombre de ces petits ARN diminue dans les cellules transfectées avec les plasmides contenant le promoteur de *flamenco*, en particulier dans les cellules transfectées par le plasmide SFET qui produit beaucoup de piRNA. Ainsi, la transcription de l'ARN précurseur de ces petits ARN ou sa dégradation en petits ARN semblent être perturbées par l'introduction de ces plasmides.

Afin d'aller plus loin sur cette problématique, une collaboration a été débutée avec une équipe de Grenoble dirigée par le Dr. Ramesh Pillai. Cette équipe, qui travaille également sur la voie des piRNA depuis plusieurs années, possède des connaissances solides

en biochimie et dans les techniques d’analyses des complexes protéiques. Ils sont actuellement en train d’analyser l’interaction des petits ARN homologues au gène de la luciférase avec la protéine Piwi. Par la suite, notre projet de collaboration visera à précipiter l’ARN luciférase et les protéines associées. Cette expérimentation devrait permettre d'identifier les facteurs protéiques impliqués dans le transport de l'ARN précurseur produit par les piRNA clusters jusqu’au site cytoplasmique où la maturation des piRNA est effectuée.

III. Analyse de l'environnement nucléaire de *flamenco*

A. Introduction

Le cluster de piRNA *flamenco* est un locus particulièrement atypique. En effet, il est localisé dans une région hétérochromatique du génome de la drosophile mais semble fortement exprimé et exempt de marques de fermetures de la chromatine. Son transcrit est coiffé et épissé mais il est ensuite maturé en piRNA. Ainsi, *flamenco* semble être à mi-chemin entre un gène classique et un cluster de piRNA de type germinale. Il a été montré que les séquences génomiques à l'intérieur du noyau peuvent se regrouper au sein de foyers nucléaires avec d'autres séquences génomiques subissant le même type de régulation. On a donc souhaité analyser l'environnement nucléaire et les partenaires d'interactions génomiques du locus *flamenco*.

Pour répondre à cette question, une expérience de 4C (Circularized Chromosome Conformation Capture) suivie d'un séquençage haut-débit (4C-seq) a été réalisée sur les cellules OSS en prenant pour cible une séquence localisée dans la séquence promotrice de *flamenco* **(Figure 39)**. Le 4C-seq permet de connaître l'ensemble des partenaires d'interaction à l'échelle génomique de la séquence ciblée. Pour cela, les interactions physiques entre séquences génomiques sont tout d'abord fixées à l'intérieur des noyaux d'environ 2 million de cellules. Puis, en combinant deux étapes de digestions enzymatiques suivies de ligations, on obtient un ensemble de molécules ADN circularisées dont chacune contient les deux séquences génomiques impliquées dans une interaction physique. Une PCR inverse est ensuite réalisée en prenant des amorces situées de part et d'autre de la séquence que l'on souhaite cibler (dans notre cas le promoteur de *flamenco*). Cette PCR permet d'amplifier l'ensemble des séquences génomiques qui interagissaient physiquement dans les noyaux avec le promoteur de *flamenco* au moment de la fixation. Un séquençage de cette PCR permet d'identifier ces séquences. Plus une séquence interagit fréquemment avec le promoteur de *flamenco* et plus celle-ci sera retrouvée fréquemment dans le séquençage.

L'échantillon biologique a été préparé par Nathalie Gueguen, technicienne dans l'équipe, puis séquencée et traitée par l'équipe du Dr W. De Laat, au Pays-Bas. Cette équipe est spécialisée dans l'étude des interactions longues distances depuis de nombreuses années. Elle a notamment développé un pipeline informatique qui permet le traitement automatisé des données de séquençage des expériences de 4C. Ce pipeline a été utilisé pour le traitement des

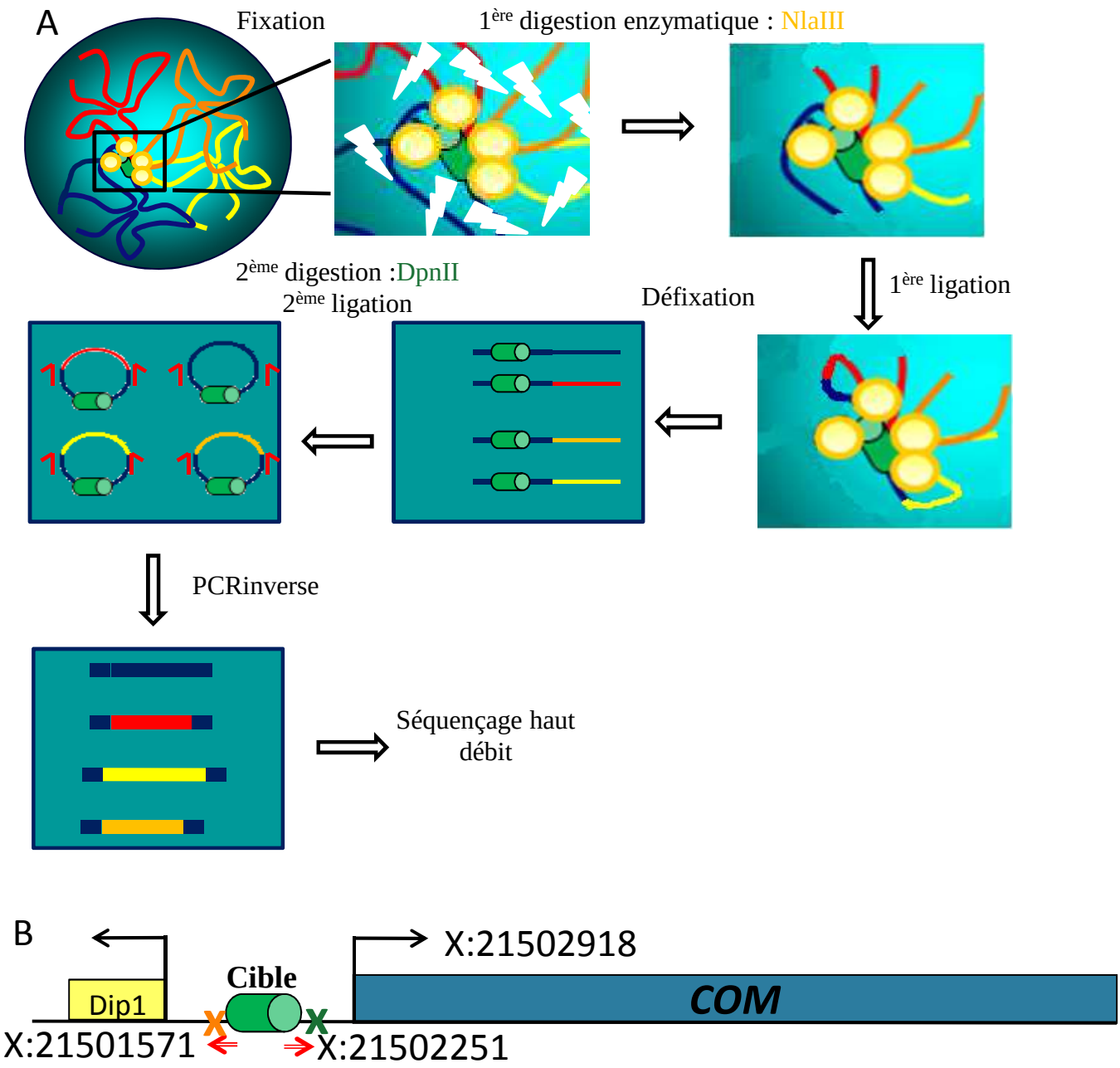


Figure 39 : Présentation de l'expérience de 4C réalisée

A. Représentation schématique des différentes étapes de l'expérience de 4C. Le cylindre vert représente la cible.

B. Localisation de la cible sur le promoteur de *flamenco*, représentée par un cylindre vert. Les flèches rouges représentent les amorces utilisées pour la PCR inverse. La croix orange représente le site de restriction NlaIII et la croix verte celui de DpnII

données de notre expérience de 4C. En utilisant un modèle probabiliste pour estimer le nombre d'interactions que l'on peut attendre avec chaque zone du génome, et donc le nombre de reads correspondant à cette zone, le pipeline calcule l'enrichissement en reads de chaque zone du génome par rapport à l'estimation. Les régions génomiques où l'enrichissement est significativement supérieur à l'estimation sont identifiées comme étant des régions génomiques interagissant avec le locus *flamenco*. Mon rôle a été d'analyser la composition de ces différentes régions.

B. Matériels et méthodes

Protocole expérimental du 4C-seq

Fixation. Les cellules OSS contenues dans deux boites de culture de 75cm^3, portées à confluence, sont fixées 10 minutes à température ambiante dans du PBS agrémenté de 10% de milieu de culture et de 2% de formaldéhyde. La fixation est stoppée pendant 1 minute par l'ajout de 1,425ml de glycine 1M à 4°C.

Lyse des cellules. Le milieu est enlevé et remplacé par 4ml de tampon de lyse (Annexe 2) pendant 10 minutes à température ambiante. Les cellules sont décollées au grattoir et les noyaux cellulaires sont culottés par centrifugation (5min, 750g, 4°C). Le culot est repris dans 1 ml de tampon de lyse puis reculotté (2min, 540g, 4°C).

1ère Digestion. Le culot est incubé pendant 1 heure à 37°C sous agitation douce dans 500µl d'eau final avec 1,2X de tampon de digestion et 0,25% de SDS. Au bout d'une heure, 50µl de Triton X-100 20% est ajouté et l'ensemble est remis à incuber pendant 1 heure à 37°C sous agitation douce. 200U de l'enzyme de restriction NlaIII sont ajoutées et mis à incuber pendant 4 heures. Une fois cette première incubation achevée, 200U de l'enzyme de restriction NlaIII sont rajoutées et mis à incuber pendant 12heures. A l'issue de cette nouvelle incubation, à nouveau 200U de l'enzyme de restriction NlaIII sont rajoutées et mis à incuber pendant 4 heures. L'enzyme NlaIII est ensuite inactivée par la chaleur pendant 20 min à 65°C.

1ère Ligation. L'ADN digéré est dilué dans un volume final de 7ml d'eau contenant le tampon de ligation et 50U de ligase (T4 DNA Ligase, Roche). L'ensemble est incubé 12 heures à 18°C. Les interactions sont dé-fixées par déprotéinisation pendant 12heures à 65°C dans 85µg/ml final de protéinase K. Les ARN sont dégradés par une incubation de 45min à 37°C dans 40µg/ml final de RNAse A. L'ADN est purifié et récupéré par une extraction au phénol/chloroforme suivie d'une précipitation à l'éthanol et au sodium acétate. Le culot est resuspendu dans 150µl de tris-HCl 10mM pH=7,5.

2ème Digestion. L'ADN est digéré pendant 12 heures à 37°C dans 500 µl final d'eau contenant le tampon de digestion et 50U d'enzyme de restriction DpnII. L'enzyme DpnII est inactivée par la chaleur pendant 25 min à 65°C.

2ème Ligation. L'ADN est ligué pendant 12 heures à 18°C dans un volume final de 14ml contenant le tampon de ligation et 100U de ligase. L'ADN est précipité à l'éthanol et au sodium Acétate et repris dans 150 µl de tris-HCl 10mM pH=7,5. L'ADN est purifié sur colonne (QIAquick PCR Purification Kit, Qiagen) et dosé au Nanodrop.

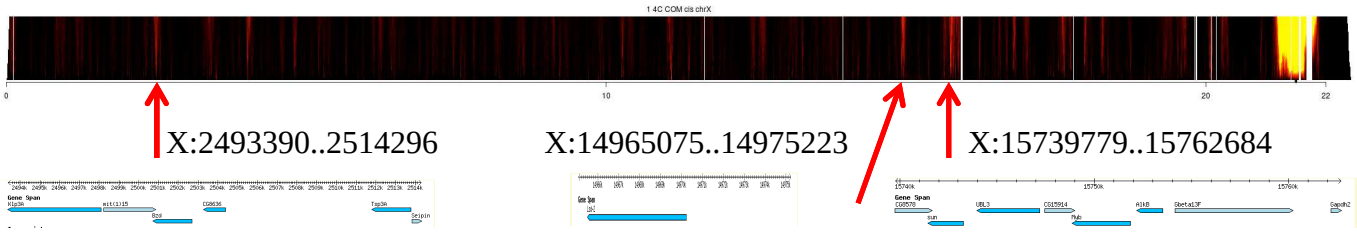
PCR inverse. Après avoir déterminé les conditions optimales, une PCR est réalisée avec une ADN polymérase du système long Expand de Roche (Expand Long Template system, Roche) sur 1,6 µg d'ADN dans 800 µl final ; 0,2mM dNTPs, ; 0,5mM chaque amorces. La PCR est purifiée sur colonne (High Pure PCR Product Purification Kit, Roche).

C. Résultats

Le séquençage a généré plus de 2,5 millions de reads dont 2,3 millions ont pu être localisés sur le génome. Le 4C permet de détecter deux types d'interactions : tout d'abord les interactions dites en cis, situées sur le même chromosome que la cible (ici le chromosome X), ainsi que les interactions en trans, situées sur les autres chromosomes. Seuls les chromosomes 2 et 3 sont inclus dans l'analyse, le chromosome 4 étant considéré comme de taille trop réduite (seulement 5000kb). 80% de ces reads se sont révélés être présents en cis, c'est-à-dire sur le chromosome X. La moyenne de la significativité de l'enrichissement de chaque région génomique présente le long d'un chromosome est représentée selon un code couleur sur un domainogramme représentant la valeur de la moyenne pour une fenêtre allant de 1 zone génomique (en bas) à 200 zones génomiques consécutives (en haut). Les valeurs p les plus hautes, proches de zéro, sont représentées en noir tandis que les plus basses, qui vont jusqu'à 10^{-8}, sont représentées en jaune. Les valeurs intermédiaires ressortent en rouge.

Concernant les interactions en cis, le pipeline a permis de détecter trois zones d'interactions significatives en dehors des régions localisées à proximité de la zone cible, qui ressortent très fortement enrichies **(Figure 40)**. Ces zones s'étendent sur 10 à 22 kb et sont situées dans des régions euchromatique du chromosome X de la drosophile. L'analyse plus détaillée de ces régions révèle qu'elles contiennent chacune plusieurs gènes. Ces gènes sont, pour la plupart, exprimés plus ou moins fortement en cellules OSS (d'après les données MODENCODE). On remarque par exemple la présence d'un gène dit « de ménage », *gapdh2*, qui présente une forte expression dans tous les types cellulaires. D'autres gènes ne sont que faiblement exprimés dans les cellules OSS. Toutefois, il s'agit généralement du type cellulaire dans lequel ils sont le plus exprimés. Dans l'ensemble, il semble donc que le locus *flamenco* interagisse majoritairement en cis avec des gènes codants des protéines et qui sont exprimés dans les cellules folliculaires.

Concernant les interactions en trans, on s'aperçoit que de nombreuses zones d'interactions (19 au total) sont également mises en évidence par le pipeline et sont réparties sur les 4 bras chromosomiques des deux autres chromosomes de la drosophile (2R, 2L, 3R et 3L) **(Figure 41)**. Ces différentes régions semblent plus hétérogènes que dans le cas des interactions en cis. Certaines contiennent des gènes et d'autres non. Globalement, peu de gènes fortement exprimés sont retrouvés. Généralement ces régions se trouvent majoritairement dans la partie péricentromérique des bras chromosomiques. Deux zones



Gène	modEncode
Klp3A	1191
mit(1)15	966
bzd	394
CG8636	10912
Tsp3A	340
Seipin	305

Gène	modEncode
Lsd-2	2051

Gène	modEncode
CG8578	541
sun	5211
UBL3	1371
CG15914	115
Myb	5963
AlkB	129
Gbeta13F	5258
Gapdh2	14450

Figure 40 : Analyse des régions d'interactions situées sur le chromosome X

La figure du haut est un domainogramme représentant par un code couleur la significativité de l'interaction sur l'ensemble du chromosome X. Les valeurs p les plus hautes, proches de zéro, sont représentées en noir tandis que les plus basses, jusqu'à 10^{-8}, sont représentées en jaune. Les valeurs intermédiaires ressortent en rouge. La zone jaune correspond aux régions autour de la zone cible qui sont surreprésentées. Les flèches rouges désignent les trois zones d'interactions mise en évidence, dont les coordonnées respectives sont indiquées à côté. En dessous est représentée une visualisation de la zone, où les gènes sont indiqués en bleu (gbrowse flybase), ainsi qu'un tableau indiquant le nom des gènes et le niveau d'expression déterminé par le projet ModEncode dans les cellules OSS.

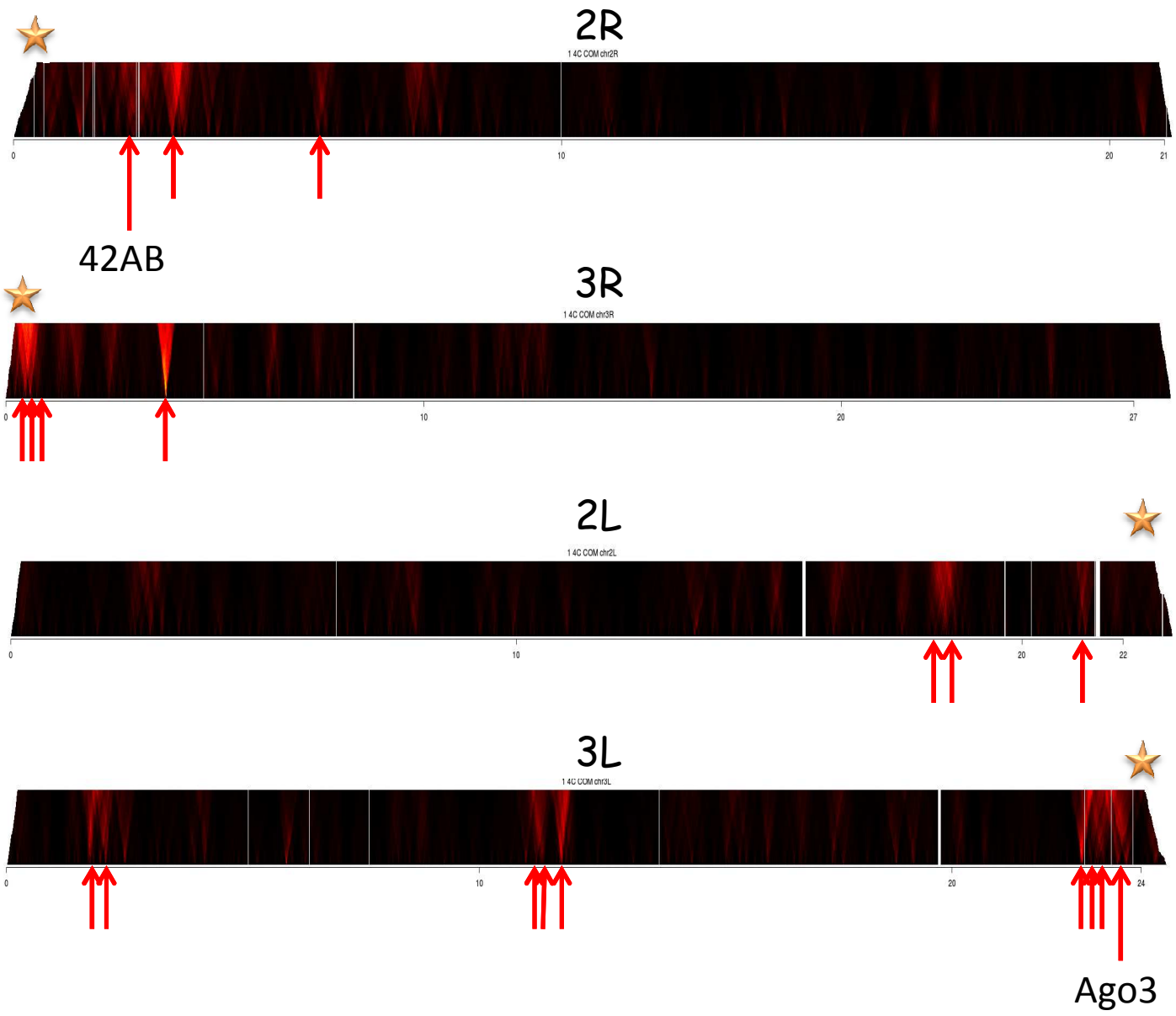


Figure 41 : Analyse des régions d'interactions en trans

Les quatre bras chromosomiques des chromosomes 2 et 3 de la drosophile sont représentés par un domainogramme représentant par un code couleur la significativité de l'interaction avec le locus *flamenco*. Les valeurs p les plus hautes, proches de zéro, sont représentées en noir tandis que les plus basses, jusqu'à 10^{-8}, sont représentées en jaune. Les valeurs intermédiaires ressortent en rouge. Les flèches rouges désignent les zones qui interagissent significativement. Les étoiles jaunes indiquent la localisation du centromère. La position des deux clusters de piRNA interagisssant avec le locus *flamenco* est indiquée.

d’interactions sont toutefois particulièrement remarquables : une zone qui correspond à une extrémité du piRNA cluster germinal situé au locus 42AB sur le chromosome 2, et une zone qui contient le gène Ago3, dont le 3’UTR produit de nombreux piRNA dans les cellules folliculaires

D. Discussion

Le traitement du séquençage réalisé sur une expérience de 4C obtenue sur des cellules OSS a montré que cette expérience est réalisable dans ce type cellulaire. De plus, les contrôles de qualité, et notamment le nombre de reads situés en cis et à proximité de la séquence ciblée, ont montré que le choix d’une cible dans une région péricentromérique et contenant beaucoup de régions répétées aux alentours n’est pas un obstacle à la réalisation et l’interprétation de l’expérience du 4C.

En cis, il apparaît clairement que *flamenco* interagit avec des gènes codants particulièrement exprimés dans les cellules OSS. Ceci confirme l’idée que le locus *flamenco* est identifié comme une région active transcriptionnellement par la cellule, ce qui est en accord avec sa structure chromatinienne. Il serait intéressant de savoir si les gènes qui interagissent avec le locus *flamenco* sont également régulés par le facteur de transcription Ci et la voie Hedgehog. Les analyses *in silico* réalisées en ce sens n’ont pas permis de répondre à cette question, il serait donc nécessaire d’envisager des expériences de quantification de l’expression de ces gènes en contexte mutant pour Ci.

Concernant les interactions en trans, de nombreuses zones d’interactions ont été identifiées dans les régions péricentromériques des chromosomes. A l’intérieur du noyau, les centromères des chromosomes sont généralement regroupés en un ou plusieurs foci appelés chromocentres. Ce regroupement entraine un rapprochement des zones péricentromériques ce qui pourrait expliquer le nombre élevé d’interactions. De plus, le locus *flamenco* semble interagir plus souvent avec la zone péricentromérique du chromosome 3 par rapport au chromosome 2. Ceci est peut-être le reflet d’une organisation spatiale particulière. Il serait intéressant d’étudier l’organisation de ces 3 centromères dans les cellules OSS par Hybridation Fluorescente In situ sur ADN.

De façon très intéressante, le locus *flamenco* semble interagir physiquement avec deux séquences génomiques produisant des piRNA : le cluster 42AB et le gène Ago3. Cela pourrait indiquer un regroupement des clusters de piRNA à l’intérieur du noyau. Ce regroupement

pourrait avoir une fonction dans la voie des piRNAs, par exemple en facilitant l'adressage des transcrits issus des clusters de piRNA vers la voie de maturation des piRNA. Très récemment une étude menée chez *Arabidopsis thaliana* a montré que des loci similaires aux piRNA clusters de la drosophile se regroupent à l'intérieur du noyau au sein d'une région nommée KNOT (Grob et al. 2014). Ainsi, le rôle joué par les interactions courtes et longues distances au sein d'un noyau entre les acteurs de la voie des piRNA pourrait se révéler important bien que jusque-là inexploré, et être élucidé très prochainement grâce à ces approches structurales menées à l'échelle des génomes.

Pour confirmer et aller plus loin dans ces résultats, 16 nouvelles expériences de 4C sont en cours de réalisation. Elles permettront d'analyser, en plus du promoteur du locus *flamenco*, trois autres cibles: l'une située sur l'extrémité du piRNA cluster en 42AB, l'une sur le promoteur d'un gène de ménage et l'une à proximité d'une insertion euchromatique d'un ET régulé par *flamenco*. De plus, ces différents loci seront analysés dans plusieurs contextes : des cellules OSS transfectées avec des siRNA dirigés contre la GFP (qui servira de condition contrôle), des cellules OSS transfectées avec des siRNA dirigés contre l'ARNm codant la protéine de clivage des piRNA Zucchini afin de désactiver la voie des piRNAs, des cellules OSS transfectées avec des siRNA dirigés contre l'ARNm codant la protéine hétérochromatique HP1 afin de déstabiliser le contexte chromatinien. Ces expériences seront également menées sur un autre type cellulaire, les cellules S2 dérivées de cellules embryonnaires dans lesquelles la voie des piRNA n'est pas actives.

Conclusion Générale

Pendant ces trois années, mon travail avait pour but de caractériser le locus *flamenco* sur plusieurs aspects de son fonctionnement : la transcription de ce locus atypique, l'adressage de son transcrit vers la voie de maturation en piRNA ainsi que l'environnement nucléaire de ce locus.

Tout d'abord mes données ont permis de mieux comprendre l'expression de ce cluster de piRNA, qui est la première étape nécessaire à la production de piRNA. Nous avons vu que sa transcription dépendait d'un promoteur canonique reconnu par l'ARN polymérase II ainsi que de la présence du facteur de transcription Ci, impliqué dans la voie de signalisation Hedgehog active dans les cellules folliculaires. Ces résultats sont en accord avec les données récentes de la littérature qui ont montré que *flamenco* possède toutes les caractéristiques principales des gènes classiques transcrits par l'ARN polymérase II. De même, nous avons vu que les transcrits issus de ce locus sont majoritairement épissés. Cet épissage est alternatif, ce qui permettrait de créer un pool d'ARN précurseurs dans lequel chacun contiendrait une portion du locus *flamenco*. Toutefois, l'ensemble de ce pool permettrait de produire des piRNA à partir de toute la séquence génomique du locus *flamenco*. Dans l'ensemble mes données, en collaboration avec une autre étude réalisée au sein de l'équipe (Dennis et al. 2013), ont contribué à l'élaboration d'un modèle de l'expression du locus *flamenco*, de sa transcription à la maturation de son transcrit en piRNA dans les Yb-bodies **(Figure 42)**.

Par la suite, j'ai montré que, dans les cellules OSS, une petite région contenant le promoteur de *flamenco* (des nucléotides -515 à +101) serait suffisante pour déclencher l'adressage des transcrits vers la voie de maturation des piRNA, et ce, quelle que soit la séquence en aval de ce promoteur. Bien que des expériences soient encore nécessaires pour confirmer ce résultat, on peut déjà affirmer que l'hypothèse la plus simple pouvant expliquer cette caractéristique est que cette région contienne une séquence particulière qui recruterait des facteurs protéiques responsables de cet adressage. Pour nous aider dans la recherche de cette séquence, l'analyse de la conservation du promoteur de flamenco qui a été réalisée entre les différentes espèces du sous-groupe de drosophile *melanogaster* pourrait nous donner des indices (Figure S4 de l'article Goriaux *et al*, 2013). En effet, ces espèces contiennent toutes un locus similaire à flamenco, composé de nombreuses séquences d'ETs et produisant des piRNAs, dont la séquence génomique du promoteur est conservée. Or cette analyse montre que seuls le site de fixation à Ci et la zone proche du promoteur sont parfaitement conservés.

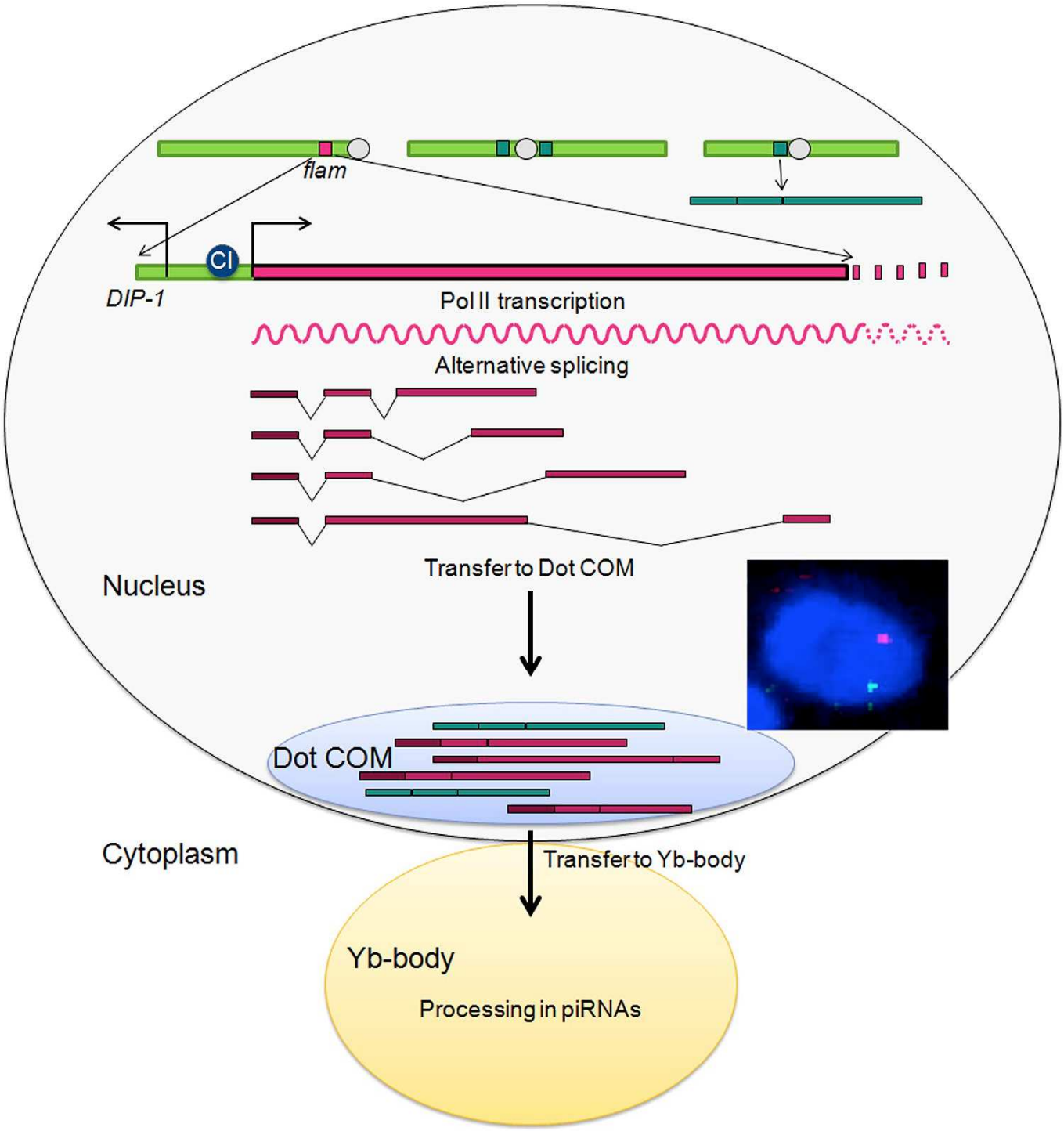


Figure 42 : Modèle de la voie de biogenèse primaire des piRNA dans les cellules folliculaires des ovaires de drosophile

Une hybridation fluorescente in situ ARN/ADN est représentée, où la séquence génomique de *flamenco* apparaît en vert, l'ARN de *flamenco* en rouge, et l'ADN est coloré au Dapi en bleu.

Le reste de la région génomique présente de nombreuses variabilités inter-espèces et aucune autre séquence ne ressort particulièrement. La capacité particulière du promoteur de *flamenco* n'est donc peut-être pas due à la présence d'une séquence spécifique. Seule la présence des marques chromatiniennes courantes (H3K4me, H3K9me) a été étudiée dans les analyses réalisées sur l'ensemble du génome. On peut donc imaginer qu'une modification d'histone plus « exotique » soit ciblée par un mécanisme inconnu sur le promoteur de *flamenco* et soit responsable de la caractéristique de ce promoteur. Dans tous les cas, il ne fait pas de doute que l'analyse poussée du promoteur de *flamenco*, et notamment grâce à la collaboration entamée avec l'équipe de Ramesh Pillai, révélera de nombreuses caractéristiques exceptionnelles de cette région génomique.

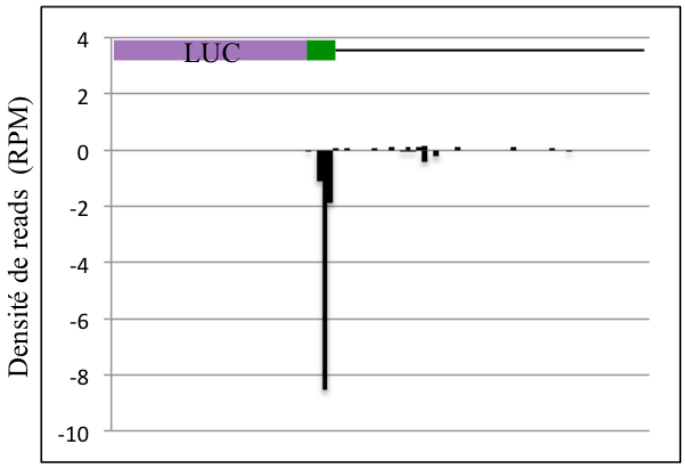
Enfin, l'analyse de l'environnement nucléaire de *flamenco* grâce à une expérience de 4C-seq s'est également révélée très intéressante. En effet, nous avons vu que le locus *flamenco* interagit, en cis, avec des gènes actifs transcriptionnellement dans les cellules folliculaires, ce qui est en accord avec l'analyse transcriptionnelle de ce locus. De plus, *flamenco* interagit avec plusieurs clusters de piRNA, ce qui pourrait révéler une colocalisation des différents clusters au sein d'un même foyer nucléaire. Cette colocalisation a été observée chez *A. thaliana* pour des loci similaires aux clusters de piRNA de la drosophile (Grob et al. 2014). Ces loci sont caractérisés par une insertion préférentielle de plusieurs ETs. Or les clusters de piRNAs, et notamment *flamenco*, semblent être préférentiellement ciblés par les ETs (Zanni et al. 2013). Ainsi, on peut imaginer que les clusters de piRNAs se regroupent au sein d'une structure nucléaire capable de recruter préférentiellement les ETs en cours de transposition afin qu'ils s'insèrent dans les piRNAs clusters. Cette structure serait donc nécessaire à l'établissement de la régulation des ETs. L'analyse des nouvelles expériences de 4C en cours de réalisation devrait nous permettre de confirmer la présence de cette structure et d'aller plus loin dans la connaissance de l'environnement nucléaire de *flamenco* et des clusters de piRNA en général.

Pour continuer l'exploration du locus *flamenco*, la technologie du CRISPR/Cas (Clustered Regularly Interspaced Short Palindromic Repeats) qui permet de créer une mutation à un endroit spécifique dans le génome devrait être extrêmement profitable. Cette technologie est dérivée d'un mécanisme bactérien et nécessite deux composants : un ARN guide et une protéine endonucléase, Cas9. L'ARN guide est composé de deux parties : une capable d'interagir avec la protéine Cas9 et une autre qui contient la séquence homologue à la région du clivage. En remplaçant la séquence homologue par n'importe quelle région

d’intérêt, et en exprimant l’ensemble de ce système dans un autre organisme, on peut générer des coupures sites spécifiques dans n’importe quelle portion d’ADN génomique. Les coupures seront réparées par le mécanisme de réparation NHEJ qui entraine fréquemment de petites mutations à l’endroit de la réparation. Cette nouvelle technique plus que prometteuse permet de générer des mutations dans n’importe quelle organisme, que ce soit dans des cultures de cellules ou dans des organismes entier en injectant les différents composants dès les stades embryonnaires. L’expérience préliminaire réalisée au laboratoire a montré que ce système pouvait être utilisé pour cibler la région génomique de *flamenco* ce qui devrait permettre, à l’avenir, de générer de très nombreux mutants de ce locus.

La poursuite de l’étude du locus *flamenco* sur les trois axes que j’ai développé permettra de mieux comprendre le fonctionnement des clusters de piRNA exprimés dans les cellules folliculaires et la voie de biogenèse des piRNA primaires de manière général. Dans l’ensemble, toutes ces données permettent de donner un nouveau swing à *flamenco* !

Annexes



Annexe 1 : Analyse des petits ARN présents dans les cellules OSS non transfectées

Répartition des reads le long de la séquence du plasmide pGL3. Le nombre de reads différents sur 50pb, normalisé par million de séquences, est représenté.

Annexe 2 : Matériels et méthodes supplémentaires de l'expérience de 4C

Composition du tampon de lyse

50mM Tris 7,5
150mM NaCl
5mM EDTA
0,5% NP-40
1% TritonX-100
Inhibiteur de protéase inhibitor (SigmaFast Protease Inhibitor Tablets)

Séquence des amorces utilisées pour la PCR inverse
flam21501571_4Cas :
AATGATACGGCGACCACCGAACACTCTTTCCCTACACGACGCTCTTCCGATCTaattg
tgtttgcaagtcatg

flam21502251_4Cs :
CAAGCAGAAGACGGCATACGAgaatatgggacagctcgact

Les nucléotides en majuscule correspondent aux adaptateurs de séquençage

Bibliographie

Adkins, N.L., Watts, M. & Georgel, P.T., 2004. To the 30-nm chromatin fiber and beyond. *Biochimica et biophysica acta*, 1677(1-3), pp.12–23. Available at: http://www.sciencedirect.com/science/article/pii/S0167478103002720 [Accessed July 15, 2014].

Albiez, H. et al., 2006. Chromatin domains and the interchromatin compartment form structurally defined and functionally interacting nuclear networks. *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, 14(7), pp.707–33. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17115328 [Accessed July 15, 2014].

ALLFREY, V.G., FAULKNER, R. & MIRSKY, A.E., 1964. ACETYLATION AND METHYLATION OF HISTONES AND THEIR POSSIBLE ROLE IN THE REGULATION OF RNA SYNTHESIS. *Proceedings of the National Academy of Sciences of the United States of America*, 51(5), pp.786–94. Available at: http://www.ncbi.nlm.nih.gov.gate2.inist.fr/pmc/articles/PMC300163/ [Accessed July 17, 2014].

Aravin, A. et al., 2006. A novel class of small RNAs bind to MILI protein in mouse testes. *Nature*, 442(7099), pp.203–7. Available at: http://www.nature.com.gate2.inist.fr/nature/journal/v442/n7099/full/nature04916.html [Accessed August 20, 2014].

Aravin, A.A., Hannon, G.J. & Brennecke, J., 2007. The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. *Science (New York, N.Y.)*, 318(5851), pp.761–4. Available at: http://www.sciencemag.org.gate2.inist.fr/content/318/5851/761.full [Accessed August 14, 2014].

Avery, O.T., 1944. STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES: INDUCTION OF TRANSFORMATION BY A DESOXYRIBONUCLEIC ACID FRACTION ISOLATED FROM PNEUMOCOCCUS TYPE III. *Journal of Experimental Medicine*, 79(2), pp.137–158. Available at: http://jem.rupress.org/cgi/content/long/79/2/137 [Accessed July 30, 2014].

Bannister, A.J. et al., 2001. Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. *Nature*, 410(6824), pp.120–4. Available at: http://www.nature.com.gate2.inist.fr/nature/journal/v410/n6824/full/410120a0.html [Accessed August 2, 2014].

Bannister, A.J. & Kouzarides, T., 2011. Regulation of chromatin by histone modifications. *Cell research*, 21(3), pp.381–95. Available at: http://www.nature.com.gate2.inist.fr/cr/journal/v21/n3/full/cr201122a.html#bib2 [Accessed August 2, 2014].

Bantignies, F. et al., 2011. Polycomb-dependent regulatory contacts between distant Hox loci in Drosophila. *Cell*, 144(2), pp.214–26. Available at:

http://www.sciencedirect.com/science/article/pii/S0092867410014856 [Accessed July 29, 2014].

Behm-Ansmant, I. et al., 2006. mRNA degradation by miRNAs and GW182 requires both CCR4:NOT deadenylase and DCP1:DCP2 decapping complexes. *Genes & development*, 20(14), pp.1885–98. Available at: http://genesdev.cshlp.org/content/20/14/1885.long [Accessed July 17, 2014].

Bennetzen, J.L., 2005. Transposable elements, gene creation and genome rearrangement in flowering plants. *Current opinion in genetics & development*, 15(6), pp.621–7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16219458 [Accessed July 27, 2014].

Berget, S.M., Moore, C. & Sharp, P.A., 1977. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proceedings of the National Academy of Sciences of the United States of America*, 74(8), pp.3171–5. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=431482&tool=pmcentrez&rendertype=abstract [Accessed August 4, 2014].

Bergman, C.M. et al., 2006. Recurrent insertion and duplication generate networks of transposable element sequences in the Drosophila melanogaster genome. *Genome biology*, 7(11), p.R112. Available at: http://genomebiology.com/2006/7/11/R112 [Accessed July 29, 2014].

Bertani, S. et al., 2011. The noncoding RNA Mistral activates Hoxa6 and Hoxa7 expression and stem cell differentiation by recruiting MLL1 to chromatin. *Molecular cell*, 43(6), pp.1040–6. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3176448&tool=pmcentrez&rendertype=abstract [Accessed July 23, 2014].

Bestor, T.H., 1992. Activation of mammalian DNA methyltransferase by cleavage of a Zn binding regulatory domain. *The EMBO journal*, 11(7), pp.2611–7. Available at: /pmc/articles/PMC556736/?report=abstract [Accessed August 2, 2014].

Biémont, C. & Vieira, C., 2006. Genetics: junk DNA as an evolutionary force. *Nature*, 443(7111), pp.521–4. Available at: http://www.nature.com.gate2.inist.fr/nature/journal/v443/n7111/full/443521a.html [Accessed August 4, 2014].

Birney, E. et al., 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146), pp.799–816. Available at: http://www.nature.com.gate2.inist.fr/nature/journal/v447/n7146/full/nature05874.html [Accessed July 9, 2014].

Birnstiel, M.L. et al., 1966. Localization of the ribosomal DNA complements in the nucleolar organizer region of Xenopus laevis. *National Cancer Institute monograph*, 23, pp.431–47. Available at: http://www.ncbi.nlm.nih.gov/pubmed/5963987 [Accessed August 2, 2014].

Bourque, G., 2009. Transposable elements in gene regulation and in the evolution of vertebrate genomes. *Current opinion in genetics & development*, 19(6), pp.607–12.

Available at: http://www.sciencedirect.com/science/article/pii/S0959437X09001725 [Accessed July 27, 2014].

Branco, M.R. & Pombo, A., 2006. Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. P. Becker, ed. *PLoS biology*, 4(5), p.e138. Available at: http://dx.plos.org/10.1371/journal.pbio.0040138 [Accessed July 20, 2014].

Brasset, E. et al., 2006. Viral particles of the endogenous retrovirus ZAM from Drosophila melanogaster use a pre-existing endosome/exosome pathway for transfer to the oocyte. *Retrovirology*, 3(1), p.25. Available at: http://www.retrovirology.com/content/3/1/25 [Accessed August 4, 2014].

Brennecke, J. et al., 2007. Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. *Cell*, 128(6), pp.1089–103. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17346786 [Accessed July 9, 2014].

Brennicke, A., Marchfelder, A. & Binder, S., 1999. RNA editing. *FEMS Microbiology Reviews*, 23(3), pp.297–316. Available at: http://onlinelibrary.wiley.com.gate2.inist.fr/doi/10.1111/j.1574-6976.1999.tb00401.x/full [Accessed July 15, 2014].

Cai, X., Hagedorn, C.H. & Cullen, B.R., 2004. Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *RNA (New York, N.Y.)*, 10(12), pp.1957–66. Available at: http://rnajournal.cshlp.org.gate2.inist.fr/content/10/12/1957.long [Accessed July 25, 2014].

Cao, R. & Zhang, Y., 2004. The functions of E(Z)/EZH2-mediated methylation of lysine 27 in histone H3. *Current opinion in genetics & development*, 14(2), pp.155–64. Available at: http://www.sciencedirect.com/science/article/pii/S0959437X0400022X [Accessed July 21, 2014].

Capuano, F. et al., 2014. Cytosine DNA methylation is found in Drosophila melanogaster but absent in Saccharomyces cerevisiae, Schizosaccharomyces pombe, and other yeast species. *Analytical chemistry*, 86(8), pp.3697–702. Available at: http://pubs.acs.org.gate2.inist.fr/doi/abs/10.1021/ac500447w [Accessed July 21, 2014].

Carswell, S. & Alwine, J.C., 1989. Efficiency of utilization of the simian virus 40 late polyadenylation site: effects of upstream sequences. *Molecular and cellular biology*, 9(10), pp.4248–58. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=362504&tool=pmcentrez&rendertype=abstract [Accessed August 7, 2014].

Chalvet, F. et al., 1999. Proviral amplification of the Gypsy endogenous retrovirus of Drosophila melanogaster involves env-independent invasion of the female germline. *EMBO Journal*, 18(9), pp.2659–2669.

Chow, L.T. et al., 1977. An amazing sequence arrangement at the 5′ ends of adenovirus 2 messenger RNA. *Cell*, 12(1), pp.1–8. Available at:

http://www.sciencedirect.com/science/article/pii/0092867477901805 [Accessed August 4, 2014].

Corona, D.F.. et al., 1999. ISWI Is an ATP-Dependent Nucleosome Remodeling Factor. *Molecular Cell*, 3(2), pp.239–245. Available at: http://www.sciencedirect.com/science/article/pii/S1097276500803147 [Accessed August 2, 2014].

Cremer, T. et al., 1988. Detection of chromosome aberrations in metaphase and interphase tumor cells by in situ hybridization using chromosome-specific library probes. *Human Genetics*, 80(3), pp.235–246. Available at: http://link.springer.com/10.1007/BF01790091 [Accessed August 2, 2014].

Danchin, É. et al., 2011. Beyond DNA: integrating inclusive inheritance into an extended theory of evolution. *Nature reviews. Genetics*, 12(7), pp.475–86. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21681209 [Accessed July 15, 2014].

Darricarrère, N. et al., 2013. Function of Piwi, a nuclear Piwi/Argonaute protein, is independent of its slicer activity. *Proceedings of the National Academy of Sciences of the United States of America*, 110(4), pp.1297–302. Available at: http://www.pnas.org/content/110/4/1297.full [Accessed August 20, 2014].

Dej, K.J. et al., 1998. A hotspot for the Drosophila gypsy retroelement in the ovo locus. *Nucleic acids research*, 26(17), pp.4019–25. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=147786&tool=pmcentrez&rendertype=abstract [Accessed August 4, 2014].

Dekker, J. et al., 2002. Capturing chromosome conformation. *Science (New York, N.Y.)*, 295(5558), pp.1306–11. Available at: http://www.sciencemag.org.gate2.inist.fr/content/295/5558/1306.full [Accessed July 13, 2014].

Dennis, C. et al., 2013. "Dot COM", a nuclear transit center for the primary piRNA pathway in Drosophila. *PloS one*, 8(9), p.e72752. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3767702&tool=pmcentrez&rendertype=abstract [Accessed July 28, 2014].

Desset, S. et al., 2003. COM, a heterochromatic locus governing the control of independent endogenous retroviruses from Drosophila melanogaster. *Genetics*, 164(2), pp.501–9. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1462594&tool=pmcentrez&rendertype=abstract.

Desset, S. et al., 1999. Mobilization of two retroelements, ZAM and Idefix, in a novel unstable line of Drosophila melanogaster. *Molecular biology and evolution*, 16(1), pp.54–66.

Dönertas, D., Sienski, G. & Brennecke, J., 2013. Drosophila Gtsf1 is an essential component of the Piwi-mediated transcriptional silencing complex. *Genes & development*, 27(15),

pp.1693–705. Available at: http://genesdev.cshlp.org/content/27/15/1693.long [Accessed July 23, 2014].

Dufourt, J. et al., 2013. NucBase, an easy to use read mapper for small RNAs. *Mobile DNA*, 4(1), p.1. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3554603&tool=pmcentrez&rendertype=abstract [Accessed September 3, 2014].

Dufourt, J. et al., 2011. Polycomb group-dependent, heterochromatin protein 1-independent, chromatin structures silence retrotransposons in somatic tissues outside ovaries. *DNA research : an international journal for rapid publication of reports on genes and genomes*, 18(6), pp.451–61. Available at: http://dnaresearch.oxfordjournals.org/content/18/6/451.full [Accessed July 31, 2014].

Faulkner, G.J. et al., 2009. The regulated retrotransposon transcriptome of mammalian cells. *Nature genetics*, 41(5), pp.563–71. Available at: http://www.nature.com.gate2.inist.fr/ng/journal/v41/n5/full/ng.368.html [Accessed August 18, 2014].

Filion, G.J. et al., 2010. Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. *Cell*, 143(2), pp.212–24. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3119929&tool=pmcentrez&rendertype=abstract [Accessed July 10, 2014].

Fire, A. et al., 1998. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature*, 391(6669), pp.806–11. Available at: http://www.nature.com.gate2.inist.fr/nature/journal/v391/n6669/full/391806a0.html [Accessed August 2, 2014].

Flemr, M. et al., 2013. A retrotransposon-driven dicer isoform directs endogenous small interfering RNA production in mouse oocytes. *Cell*, 155(4), pp.807–16. Available at: http://www.sciencedirect.com/science/article/pii/S0092867413012282 [Accessed August 19, 2014].

Gasior, S.L. et al., 2006. The human LINE-1 retrotransposon creates DNA double-strand breaks. *Journal of molecular biology*, 357(5), pp.1383–93. Available at: http://www.sciencedirect.com/science/article/pii/S0022283606001422 [Accessed August 1, 2014].

Ghildiyal, M. & Zamore, P.D., 2009. Small silencing RNAs: an expanding universe. *Nature reviews. Genetics*, 10(2), pp.94–108. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2724769&tool=pmcentrez&rendertype=abstract [Accessed July 13, 2014].

Gill, G., 2001. Regulation of the initiation of eukaryotic transcription. *Essays in biochemistry*, 37, pp.33–43. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11758455 [Accessed August 4, 2014].

Gladyshev, E.A. & Arkhipova, I.R., 2007. Telomere-associated endonuclease-deficient Penelope-like retroelements in diverse eukaryotes. *Proceedings of the National Academy*

of Sciences of the United States of America, 104(22), pp.9352–7. Available at: http://www.pnas.org/content/104/22/9352.full [Accessed August 4, 2014].

Gralla, J.D., 1996. *RNA Polymerase and Associated Factors Part A*, Elsevier. Available at: http://www.sciencedirect.com/science/article/pii/S0076687996730094 [Accessed August 4, 2014].

Greenblatt, J., 1997. RNA polymerase II holoenzyme and transcriptional regulation. *Current Opinion in Cell Biology*, 9(3), pp.310–319. Available at: http://www.sciencedirect.com/science/article/pii/S0955067497800026 [Accessed August 4, 2014].

Grob, S., Schmid, M.W. & Grossniklaus, U., 2014. Hi-C Analysis in Arabidopsis Identifies the KNOT, a Structure with Similarities to the flamenco Locus of Drosophila. *Molecular Cell*. Available at: http://www.sciencedirect.com/science/article/pii/S1097276514006029 [Accessed August 15, 2014].

Grüne, T. et al., 2003. Crystal Structure and Functional Analysis of a Nucleosome Recognition Module of the Remodeling Factor ISWI. *Molecular Cell*, 12(2), pp.449–460. Available at: http://www.sciencedirect.com/science/article/pii/S1097276503002739 [Accessed August 2, 2014].

Guerrier-Takada, C. et al., 1983. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell*, 35(3), pp.849–857. Available at: http://www.sciencedirect.com/science/article/pii/0092867483901174 [Accessed July 16, 2014].

Gunawardane, L.S. et al., 2007. A slicer-mediated mechanism for repeat-associated siRNA 5' end formation in Drosophila. *Science (New York, N.Y.)*, 315(5818), pp.1587–90. Available at: http://www.sciencemag.org.gate2.inist.fr/content/315/5818/1587.full [Accessed July 18, 2014].

Guttman, M. et al., 2009. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, 458(7235), pp.223–7. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2754849&tool=pmcentrez&rendertype=abstract [Accessed July 9, 2014].

Halligan, D.L. & Keightley, P.D., 2006. Ubiquitous selective constraints in the Drosophila genome revealed by a genome-wide interspecies comparison. *Genome research*, 16(7), pp.875–84. Available at: http://genome.cshlp.org.gate2.inist.fr/content/16/7/875.full [Accessed July 31, 2014].

Handler, D. et al., 2013. The genetic makeup of the Drosophila piRNA pathway. *Molecular cell*, 50(5), pp.762–77. Available at: http://www.sciencedirect.com/science/article/pii/S1097276513003365 [Accessed July 14, 2014].

Hirose, Y. & Manley, J.L., 1998. RNA polymerase II is an essential mRNA polyadenylation factor. *Nature*, 395(6697), pp.93–6. Available at:

http://www.nature.com.gate2.inist.fr/nature/journal/v395/n6697/full/395093a0.html [Accessed August 4, 2014].

Ho, C.K., 1998. The Guanylyltransferase Domain of Mammalian mRNA Capping Enzyme Binds to the Phosphorylated Carboxyl-terminal Domain of RNA Polymerase II. *Journal of Biological Chemistry*, 273(16), pp.9577–9585. Available at: http://www.jbc.org.gate2.inist.fr/content/273/16/9577.full [Accessed August 4, 2014].

Horwich, M.D. et al., 2007. The Drosophila RNA methyltransferase, DmHen1, modifies germline piRNAs and single-stranded siRNAs in RISC. *Current biology : CB*, 17(14), pp.1265–72. Available at: http://www.sciencedirect.com/science/article/pii/S0960982207015679 [Accessed July 13, 2014].

Ipsaro, J.J. et al., 2012. The structural biochemistry of Zucchini implicates it as a nuclease in piRNA biogenesis. *Nature*, 491(7423), pp.279–83. Available at: http://www.nature.com.gate2.inist.fr/nature/journal/v491/n7423/full/nature11502.html [Accessed August 13, 2014].

Ishizu, H., Siomi, H. & Siomi, M.C., 2012. Biology of PIWI-interacting RNAs: new insights into biogenesis and function inside and outside of germlines. *Genes & development*, 26(21), pp.2361–73. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3489994&tool=pmcentrez&rendertype=abstract [Accessed July 15, 2014].

Kapitonov, V. V & Jurka, J., 2005. RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. D. Nemazee, ed. *PLoS biology*, 3(6), p.e181. Available at: http://dx.plos.org/10.1371/journal.pbio.0030181 [Accessed July 28, 2014].

Kapranov, P. et al., 2007. RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science (New York, N.Y.)*, 316(5830), pp.1484–8. Available at: http://www.sciencemag.org.gate2.inist.fr/content/316/5830/1484.full [Accessed July 10, 2014].

Kawamata, T., Seitz, H. & Tomari, Y., 2009. Structural determinants of miRNAs for RISC loading and slicer-independent unwinding. *Nature structural & molecular biology*, 16(9), pp.953–60. Available at: http://www.nature.com.gate2.inist.fr/nsmb/journal/v16/n9/full/nsmb.1630.html [Accessed August 20, 2014].

Kidwell, M.G., Kidwell, J.F. & Sved, J.A., 1977. Hybrid Dysgenesis in DROSOPHILA MELANOGASTER: A Syndrome of Aberrant Traits Including Mutation, Sterility and Male Recombination. *Genetics*, 86(4), pp.813–33. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1213713&tool=pmcentrez&rendertype=abstract [Accessed August 27, 2014].

Klattenhoff, C. et al., 2009. The Drosophila HP1 homolog Rhino is required for transposon silencing and piRNA production by dual-strand clusters. *Cell*, 138(6), pp.1137–49. Available at:

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2770713&tool=pmcentrez&rendertype=abstract [Accessed July 18, 2014].

Kruger, K. et al., 1982. Self-splicing RNA: Autoexcision and autocyclization of the ribosomal RNA intervening sequence of tetrahymena. *Cell*, 31(1), pp.147–157. Available at: http://www.sciencedirect.com/science/article/pii/0092867482904147 [Accessed July 16, 2014].

Lachner, M. et al., 2001. Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. *Nature*, 410(6824), pp.116–20. Available at: http://www.nature.com.gate2.inist.fr/nature/journal/v410/n6824/full/410116a0.html [Accessed August 2, 2014].

Leblanc, P. et al., 1997. Invertebrate retroviruses: ZAM a new candidate in D.melanogaster. *The EMBO journal*, 16(24), pp.7521–31. Available at: http://emboj.embopress.org/content/16/24/7521.abstract [Accessed August 4, 2014].

Lécher, P., Bucheton, a & Pélisson, a, 1997. Expression of the Drosophila retrovirus gypsy as ultrastructurally detectable particles in the ovaries of flies carrying a permissive flamenco allele. *The Journal of general virology*, 78 (Pt 9), pp.2379–2388.

Lee, R.C., Feinbaum, R.L. & Ambros, V., 1993. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75(5), pp.843–854. Available at: http://www.sciencedirect.com/science/article/pii/009286749390529Y [Accessed July 31, 2014].

Lee, Y. et al., 2003. The nuclear RNase III Drosha initiates microRNA processing. *Nature*, 425(6956), pp.415–9. Available at: http://www.nature.com.gate2.inist.fr/nature/journal/v425/n6956/full/nature01957.html [Accessed August 20, 2014].

Leonhardt, H. et al., 1992. A targeting sequence directs DNA methyltransferase to sites of DNA replication in mammalian nuclei. *Cell*, 71(5), pp.865–873. Available at: http://www.sciencedirect.com/science/article/pii/009286749290561P [Accessed August 2, 2014].

Levis, R.W. et al., 1993. Transposons in place of telomeric repeats at a Drosophila telomere. *Cell*, 75(6), pp.1083–1093. Available at: http://www.sciencedirect.com/science/article/pii/009286749390318K [Accessed August 20, 2014].

Lewis, B.P. et al., 2003. Prediction of Mammalian MicroRNA Targets. *Cell*, 115(7), pp.787–798. Available at: http://www.sciencedirect.com/science/article/pii/S0092867403010183 [Accessed July 18, 2014].

Li, E., Bestor, T.H. & Jaenisch, R., 1992. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell*, 69(6), pp.915–926. Available at: http://www.sciencedirect.com/science/article/pii/009286749290611F [Accessed August 2, 2014].

Li, X.Z. et al., 2013. Article An Ancient Transcription Factor Initiates the Burst of piRNA Production during Early Meiosis in Mouse Testes. *Molecular Cell*, 50(1), pp.67–81. Available at: http://dx.doi.org/10.1016/j.molcel.2013.02.016.

Liu, J. et al., 2004. Argonaute2 is the catalytic engine of mammalian RNAi. *Science (New York, N.Y.)*, 305(5689), pp.1437–41. Available at: http://www.sciencemag.org.gate2.inist.fr/content/305/5689/1437.full [Accessed July 15, 2014].

Lu, B.Y. et al., 2000. Heterochromatin protein 1 is required for the normal expression of two heterochromatin genes in Drosophila. *Genetics*, 155(2), pp.699–708. Available at: http://www.genetics.org/content/155/2/699.full [Accessed August 2, 2014].

Luger, K. et al., 1997. Characterization of nucleosome core particles containing histone proteins made in bacteria. *Journal of molecular biology*, 272(3), pp.301–11. Available at: http://www.sciencedirect.com/science/article/pii/S0022283697912353 [Accessed July 23, 2014].

Maher, B., 2008. The case of the missing heritability. , 456(November).

Maillard, P. V et al., 2013. Antiviral RNA interference in mammalian cells. *Science (New York, N.Y.)*, 342(6155), pp.235–8. Available at: http://www.sciencemag.org.gate2.inist.fr/content/342/6155/235.full [Accessed July 11, 2014].

Malik, S. & Roeder, R.G., 2010. The metazoan Mediator co-activator complex as an integrative hub for transcriptional regulation. *Nature reviews. Genetics*, 11(11), pp.761–72. Available at: http://www.nature.com.gate2.inist.fr/nrg/journal/v11/n11/full/nrg2901.html [Accessed August 1, 2014].

Malone, C.D. et al., 2009. Specialized piRNA pathways act in germline and somatic tissues of the Drosophila ovary. *Cell*, 137(3), pp.522–35. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2882632&tool=pmcentrez&rendertype=abstract [Accessed July 9, 2014].

Matranga, C. et al., 2005. Passenger-strand cleavage facilitates assembly of siRNA into Ago2-containing RNAi enzyme complexes. *Cell*, 123(4), pp.607–20. Available at: http://www.sciencedirect.com/science/article/pii/S0092867405009220 [Accessed July 16, 2014].

McCLINTOCK, B., 1950. The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences of the United States of America*, 36(6), pp.344–55. Available at: http://www.ncbi.nlm.nih.gov.gate2.inist.fr/pmc/articles/PMC1063197/ [Accessed July 22, 2014].

MENDEL, G., 1950. Gregor Mendel's letters to Carl Nägeli, 1866-1873. *Genetics*, 35(5:2), pp.1–29. Available at: http://www.ncbi.nlm.nih.gov/pubmed/14773778 [Accessed August 2, 2014].

Mével-Ninio, M., Mariol, M.C. & Gans, M., 1989. Mobilization of the gypsy and copia retrotransposons in Drosophila melanogaster induces reversion of the ovo dominant female-sterile mutations: molecular analysis of revertant alleles. *The EMBO journal*, 8(5), pp.1549–1558.

Misteli, T., 2005. Concepts in nuclear architecture. *BioEssays : news and reviews in molecular, cellular and developmental biology*, 27(5), pp.477–87. Available at: http://onlinelibrary.wiley.com.gate2.inist.fr/doi/10.1002/bies.20226/abstract [Accessed July 24, 2014].

Mohn, F. et al., 2014a. The rhino-deadlock-cutoff complex licenses noncanonical transcription of dual-strand piRNA clusters in Drosophila. *Cell*, 157(6), pp.1364–79. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24906153 [Accessed July 18, 2014].

Mohn, F. et al., 2014b. The rhino-deadlock-cutoff complex licenses noncanonical transcription of dual-strand piRNA clusters in Drosophila. *Cell*, 157(6), pp.1364–79. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24906153 [Accessed July 18, 2014].

Morgan, T., chromosome and associative inheritance. Available at: http://www.sciencemag.org.gate2.inist.fr/content/34/880/636.long [Accessed July 30, 2014].

Morse, B. et al., 1988. Insertional mutagenesis of the myc locus by a LINE-1 sequence in a human breast carcinoma. *Nature*, 333(6168), pp.87–90. Available at: http://www.nature.com.gate2.inist.fr/nature/journal/v333/n6168/abs/333087a0.html [Accessed August 4, 2014].

Muchardt, C. & Yaniv, M., 1999. ATP-dependent chromatin remodelling: SWI/SNF and Co. are on the job. *Journal of molecular biology*, 293(2), pp.187–98. Available at: http://www.sciencedirect.com/science/article/pii/S0022283699929996 [Accessed July 27, 2014].

Muerdter, F. et al., 2012. Production of artificial piRNAs in flies and mice. *RNA (New York, N.Y.)*, 18(1), pp.42–52. Available at: http://rnajournal.cshlp.org.gate2.inist.fr/content/18/1/42.full [Accessed July 14, 2014].

Murota, Y. et al., 2014a. Yb Integrates piRNA Intermediates and Processing Factors into Perinuclear Bodies to Enhance piRISC Assembly. *Cell reports*, 8(1), pp.103–13. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24953657 [Accessed July 15, 2014].

Murota, Y. et al., 2014b. Yb Integrates piRNA Intermediates and Processing Factors into Perinuclear Bodies to Enhance piRISC Assembly. *Cell reports*, 8(1), pp.103–13. Available at: http://www.sciencedirect.com/science/article/pii/S2211124714004379 [Accessed July 15, 2014].

Nan, X. et al., 1998. Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. *Nature*, 393(6683), pp.386–9. Available at: http://www.nature.com.gate2.inist.fr/nature/journal/v393/n6683/full/393386a0.html [Accessed August 2, 2014].

Napoli, C., Lemieux, C. & Jorgensen, R., 1990. Introduction of a Chimeric Chalcone Synthase Gene into Petunia Results in Reversible Co-Suppression of Homologous Genes in trans. *The Plant cell*, 2(4), pp.279–289. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=159885&tool=pmcentrez&rendertype=abstract.

Németh, A. et al., 2008. Epigenetic regulation of TTF-I-mediated promoter-terminator interactions of rRNA genes. *The EMBO journal*, 27(8), pp.1255–65. Available at: http://emboj.embopress.org/content/27/8/1255.abstract [Accessed July 31, 2014].

Niki, Y., Yamaguchi, T. & Mahowald, A.P., 2006. Establishment of stable cell lines of Drosophila germ-line stem cells. *Proceedings of the National Academy of Sciences of the United States of America*, 103(44), pp.16325–30. Available at: http://www.pnas.org/content/103/44/16325.full [Accessed August 3, 2014].

Nishimasu, H. et al., 2012. Structure and function of Zucchini endoribonuclease in piRNA biogenesis. *Nature*, 491(7423), pp.284–7. Available at: http://www.nature.com.gate2.inist.fr/nature/journal/v491/n7423/full/nature11509.html#acknowledgments [Accessed August 13, 2014].

Noordermeer, D. et al., 2011. Variegated gene expression caused by cell-specific long-range DNA interactions. *Nature cell biology*, 13(8), pp.944–51. Available at: http://www.nature.com.gate2.inist.fr/ncb/journal/v13/n8/full/ncb2278.html [Accessed July 31, 2014].

Nouaud, D. & Anxolabéhère, D., 1997. P element domestication: a stationary truncated P element may encode a 66-kDa repressor-like protein in the Drosophila montium species subgroup. *Molecular biology and evolution*, 14(11), pp.1132–44. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9364771 [Accessed August 4, 2014].

Ohler, U. et al., 2002. Computational analysis of core promoters in the Drosophila genome. *Genome biology*, 3(12), p.RESEARCH0087. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=151189&tool=pmcentrez&rendertype=abstract.

Okano, M. et al., 1999. DNA Methyltransferases Dnmt3a and Dnmt3b Are Essential for De Novo Methylation and Mammalian Development. *Cell*, 99(3), pp.247–257. Available at: http://www.sciencedirect.com/science/article/pii/S0092867400816566 [Accessed August 2, 2014].

Olins AL, O.D., 1974. spheroid chromatin units. *Science (New York, N.Y.)*, p.1974 Jan 25;183(4122):330–2. Available at: http://www.sciencemag.org.gate2.inist.fr/content/183/4122/330.long [Accessed August 2, 2014].

Oliver, K.R. & Greene, W.K., 2011. Mobile DNA and the TE-Thrust hypothesis: supporting evidence from the primates. *Mobile DNA*, 2(1), p.8. Available at: http://www.mobilednajournal.com/content/2/1/8 [Accessed August 1, 2014].

Olovnikov, I. et al., 2013. De novo piRNA cluster formation in the Drosophila germ line triggered by transgenes containing a transcribed transposon fragment. *Nucleic acids research*, 41(11), pp.5757–68. Available at: http://nar.oxfordjournals.org/content/41/11/5757.full [Accessed July 19, 2014].

Pak, J. & Fire, A., 2007. Distinct populations of primary and secondary effectors during RNAi in C. elegans. *Science (New York, N.Y.)*, 315(5809), pp.241–4. Available at: http://www.sciencemag.org.gate2.inist.fr/content/315/5809/241.full [Accessed July 29, 2014].

Pane, A. et al., 2011. The Cutoff protein regulates piRNA cluster expression and piRNA production in the Drosophila germline. *The EMBO journal*, 30(22), pp.4601–15. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3243597&tool=pmcentrez&rendertype=abstract [Accessed July 28, 2014].

Pasyukova, E.G. et al., 2004. Accumulation of transposable elements in the genome of Drosophila melanogaster is associated with a decrease in fitness. *The Journal of heredity*, 95(4), pp.284–90. Available at: http://jhered.oxfordjournals.org/content/95/4/284.full [Accessed July 22, 2014].

Pélisson, a et al., 1994. Gypsy transposition correlates with the production of a retroviral envelope-like protein under the tissue-specific control of the Drosophila flamenco gene. *The EMBO journal*, 13(18), pp.4401–4411.

Picard, G., 1976. Non Mendelian sterility in Drosophila melanogaster:hereditary transmission of I factor. , pp.107–123.

Piskurek, O. & Jackson, D.J., 2012. Transposable elements: from DNA parasites to architects of metazoan evolution. *Genes*, 3(3), pp.409–22. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3899998&tool=pmcentrez&rendertype=abstract [Accessed July 28, 2014].

Poczai, P., Bell, N. & Hyvönen, J., 2014. Imre Festetics and the Sheep Breeders' Society of Moravia: Mendel's Forgotten "Research Network". *PLoS biology*, 12(1), p.e1001772. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3897355&tool=pmcentrez&rendertype=abstract [Accessed July 28, 2014].

Prud, N. et al., 2001. Characterization of the flamenco region of droso genome.

Prud, N. et al., 1995. flamenco, a gene controlling the gypsy retrovirus of D. melanogaster. *genetics*, 711.

Qin, J.Y. et al., 2010. Systematic comparison of constitutive promoters and the doxycycline-inducible promoter. *PloS one*, 5(5), p.e10611. Available at: /pmc/articles/PMC2868906/?report=abstract [Accessed July 11, 2014].

Ragoczy, T. et al., 2006. The locus control region is required for association of the murine beta-globin locus with engaged transcription factories during erythroid maturation.

Genes & development, 20(11), pp.1447–57. Available at: http://genesdev.cshlp.org/content/20/11/1447.long [Accessed July 31, 2014].

Rangan, P. et al., 2011. piRNA production requires heterochromatin formation in Drosophila. *Current biology : CB*, 21(16), pp.1373–9. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3205116&tool=pmcentrez&rendertype=abstract [Accessed July 21, 2014].

Ravindran, S., 2012. Barbara McClintock and the discovery of jumping genes. *Proceedings of the National Academy of Sciences of the United States of America*, 109(50), pp.20198–9. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3528533&tool=pmcentrez&rendertype=abstract [Accessed July 15, 2014].

Reddy, K.L. et al., 2008. Transcriptional repression mediated by repositioning of genes to the nuclear lamina. *Nature*, 452(7184), pp.243–7. Available at: http://www.nature.com.gate2.inist.fr/nature/journal/v452/n7184/full/nature06727.html [Accessed July 31, 2014].

Robine, N. et al., 2009. A broadly conserved pathway generates 3'UTR-directed primary piRNAs. *Current biology : CB*, 19(24), pp.2066–76. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2812478&tool=pmcentrez&rendertype=abstract [Accessed July 17, 2014].

Roy, S.W., 2004. The origin of recent introns: transposons? *Genome biology*, 5(12), p.251. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=545789&tool=pmcentrez&rendertype=abstract [Accessed August 4, 2014].

Rozhkov, N. V, Hammell, M. & Hannon, G.J., 2013. Multiple roles for Piwi in silencing Drosophila transposons. *Genes & development*, 27(4), pp.400–12. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3589557&tool=pmcentrez&rendertype=abstract [Accessed July 18, 2014].

Ruby, J.G. et al., 2006. Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in C. elegans. *Cell*, 127(6), pp.1193–207. Available at: http://www.sciencedirect.com/science/article/pii/S0092867406014681 [Accessed July 23, 2014].

Saito, K. et al., 2010. Roles for the Yb body components Armitage and Yb in primary piRNA biogenesis in Drosophila. *Genes & development*, 24(22), pp.2493–8. Available at: http://genesdev.cshlp.org/content/24/22/2493.long [Accessed August 3, 2014].

Sarot, E., Bucheton, A. & Pe, A., 2004. Retrovirus by the Drosophila melanogaster flamenco Gene. , 1321(March), pp.1313–1321.

Seitz, H., 2009. Redefining microRNA targets. *Current biology : CB*, 19(10), pp.870–3. Available at: http://www.sciencedirect.com/science/article/pii/S0960982209009130 [Accessed August 5, 2014].

Shapiro, J. a, 1969. Mutations caused by the insertion of genetic material into the galactose operon of Escherichia coli. *Journal of molecular biology*, 40(1), pp.93–105. Available at: http://www.ncbi.nlm.nih.gov/pubmed/4903362.

Shogren-Knaak, M. et al., 2006. Histone H4-K16 acetylation controls chromatin structure and protein interactions. *Science (New York, N.Y.)*, 311(5762), pp.844–7. Available at: http://www.sciencemag.org.gate2.inist.fr/content/311/5762/844.full [Accessed August 2, 2014].

Sienski, G., Dönertas, D. & Brennecke, J., 2012. Transcriptional silencing of transposons by Piwi and maelstrom and its impact on chromatin state and gene expression. *Cell*, 151(5), pp.964–80. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3504300&tool=pmcentrez&rendertype=abstract [Accessed July 18, 2014].

Song, S.U. et al., 1997. Infection of the germ line by retroviral particles produced in the follicle cells: a possible mechanism for the mobilization of the gypsy retroelement of Drosophila. *Development (Cambridge, England)*, 124(14), pp.2789–2798.

Spilianakis, C.G. et al., 2005. Interchromosomal associations between alternatively expressed loci. *Nature*, 435(7042), pp.637–45. Available at: http://www.nature.com.gate2.inist.fr/nature/journal/v435/n7042/full/nature03574.html [Accessed August 2, 2014].

Stack, S.M., Brown, D.B. & Dewey, W.C., 1977. Visualization of interphase chromosomes. *Journal of cell science*, 26, pp.281–99. Available at: http://www.ncbi.nlm.nih.gov/pubmed/562895 [Accessed July 31, 2014].

Le Thomas, A. et al., 2014. Transgenerationally inherited piRNAs trigger piRNA biogenesis by changing the chromatin of piRNA clusters and inducing precursor processing. *Genes & development*, 28(15), pp.1667–80. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25085419 [Accessed August 4, 2014].

Tolhuis, B. et al., 2002. Looping and Interaction between Hypersensitive Sites in the Active β-globin Locus. *Molecular Cell*, 10(6), pp.1453–1465. Available at: http://www.sciencedirect.com/science/article/pii/S1097276502007815 [Accessed July 14, 2014].

Valen, E. et al., 2011. Biogenic mechanisms and utilization of small RNAs derived from human protein-coding genes. *Nature structural & molecular biology*, 18(9), pp.1075–82. Available at: http://www.nature.com.gate2.inist.fr/nsmb/journal/v18/n9/full/nsmb.2091.html [Accessed August 4, 2014].

De Vanssay, A. et al., 2012. Paramutation in Drosophila linked to emergence of a piRNA-producing locus. *Nature*, 490(7418), pp.112–5. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22922650 [Accessed July 13, 2014].

Vasilyeva, L.A., Bubenshchikova, E. V & Ratner, V.A., 1999. Heavy heat shock induced retrotransposon transposition in Drosophila. *Genetical research*, 74(2), pp.111–9.

Available at: http://www.ncbi.nlm.nih.gov/pubmed/10584555 [Accessed August 4, 2014].

Venter, J.C. et al., 2001. The sequence of the human genome. *Science (New York, N.Y.)*, 291(5507), pp.1304–51. Available at: http://www.sciencemag.org.gate2.inist.fr/content/291/5507/1304.full [Accessed July 10, 2014].

WATSON, J.D. & CRICK, F.H., 1953. Genetical implications of the structure of deoxyribonucleic acid. *Nature*, 171(4361), pp.964–7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/13063483 [Accessed July 30, 2014].

Watt, F. & Molloy, P.L., 1988. Cytosine methylation prevents binding to DNA of a HeLa cell transcription factor required for optimal expression of the adenovirus major late promoter. *Genes & development*, 2(9), pp.1136–43. Available at: http://www.ncbi.nlm.nih.gov/pubmed/3192075 [Accessed August 2, 2014].

Whitney, J.B. & Lamoreux, M.L., 1982. Transposable elements controlling genetic instabilities in mammals. *The Journal of heredity*, 73(1), pp.12–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/6279728 [Accessed August 4, 2014].

Wicker, T. et al., 2007. A unified classification system for eukaryotic transposable elements. *Nature reviews. Genetics*, 8(12), pp.973–82. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19238178.

Wightman, B., Ha, I. & Ruvkun, G., 1993. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell*, 75(5), pp.855–862. Available at: http://www.sciencedirect.com/science/article/pii/0092867493905304 [Accessed July 31, 2014].

Will, C.L. & Lührmann, R., 2011. Spliceosome structure and function. *Cold Spring Harbor perspectives in biology*, 3(7), p.a003707. Available at: http://cshperspectives.cshlp.org.gate2.inist.fr/content/3/7/a003707.full [Accessed July 11, 2014].

Wutz, A., 2011. Gene silencing in X-chromosome inactivation: advances in understanding facultative heterochromatin formation. *Nature reviews. Genetics*, 12(8), pp.542–53. Available at: http://www.nature.com.gate2.inist.fr/nrg/journal/v12/n8/full/nrg3035.html [Accessed August 4, 2014].

Wyers, F. et al., 2005. Cryptic pol II transcripts are degraded by a nuclear quality control pathway involving a new poly(A) polymerase. *Cell*, 121(5), pp.725–37. Available at: http://www.sciencedirect.com/science/article/pii/S0092867405004435 [Accessed July 23, 2014].

Zanni, V. et al., 2013. Distribution, evolution, and diversity of retrotransposons at the flamenco locus reflect the regulatory properties of piRNA clusters. *Proceedings of the National Academy of Sciences of the United States of America*, 110(49), pp.19842–7. Available at: http://www.pnas.org/content/110/49/19842.full [Accessed August 4, 2014].

Zhang, F. et al., 2012. UAP56 couples piRNA clusters to the perinuclear transposon silencing machinery. *Cell*, 151(4), pp.871–84. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3499805&tool=pmcentrez&rendertype=abstract [Accessed July 18, 2014].

Zhang, Z. et al., 2003. Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. *Genome research*, 13(12), pp.2541–58. Available at: http://genome.cshlp.org.gate2.inist.fr/content/13/12/2541.full [Accessed August 2, 2014].

Zhang, Z. et al., 2014. The HP1 homolog rhino anchors a nuclear complex that suppresses piRNA precursor splicing. *Cell*, 157(6), pp.1353–63. Available at: http://www.sciencedirect.com/science/article/pii/S0092867414006023 [Accessed July 18, 2014].