Etude de la dynamique des repetitions dans les genomes eucaryotes: de leur formation a leur elimination

Anna-Sophie Fiston-Lavier

UNIVERSITE PIERRE & MARIE CURIE
LA SCIENCE A PARIS

Institut Formation Doctorale

THESE DE DOCTORAT DE
L'UNIVERSITE PIERRE ET MARIE CURIE

Spécialité: Bioinformatique et Génomique
Ecole Doctorale: Logique du Vivant

Présentée par:

Fiston-Lavier Anna-Sophie

Pour obtenir le grade de:

DOCTEUR de l'UNIVERSITÉ PIERRE ET MARIE CURIE

«ÉTUDE DE LA DYNAMIQUE DES RÉPÉTITIONS DANS LES GÉNOMES EUCARYOTES: DE LEUR FORMATION À LEUR ÉLIMINATION»

Soutenue le 26 Mars 2008

Devant le jury composé de:

Anxolabéhère Dominique	Président du Jury
Duret Laurent	Rapporteur
Panaud Olivier	Rapporteur
Colot Vincent	Examinateur
Deragon Jean-Marc	Examinateur
Quesneville Hadi	Directeur de thèse

UNIVERSITE PIERRE & MARIE CURIE
LA SCIENCE A PARIS

Institut Formation Doctorale

THESE DE DOCTORAT DE
L'UNIVERSITE PIERRE ET MARIE CURIE

Spécialité: Bioinformatique et Génomique
Ecole Doctorale: Logique du Vivant

Présentée par:

Fiston-Lavier Anna-Sophie

Pour obtenir le grade de:

DOCTEUR de l'UNIVERSITÉ PIERRE ET MARIE CURIE

«ÉTUDE DE LA DYNAMIQUE DES RÉPÉTITIONS DANS LES GÉNOMES EUCARYOTES: DE LEUR FORMATION À LEUR ÉLIMINATION»

Soutenue le 26 Mars 2008

Devant le jury composé de:

Anxolabéhère Dominique	Président du Jury
Duret Laurent	Rapporteur
Panaud Olivier	Rapporteur
Colot Vincent	Examinateur
Deragon Jean-Marc	Examinateur
Quesneville Hadi	Directeur de thèse

Que dire à part « Merci » !

Je tiens à exprimer ma profonde reconnaissance à **Laurent Duret** et **Olivier Panaud** qui ont accepté de juger ce mémoire avec beaucoup d'attention, ainsi qu'à **Vincent Colot** et **Jean-Marc Deragon** qui ont accepté d'en être les examinateurs.

****

Je tiens à remercier chaleureusement **Hadi Quesneville** pour avoir encadré ce travail de thèse, avec beaucoup de compétence et d'enthousiasme. Tes conseils avisés, ton optimisme et la confiance que tu m'as accordée au cours de ces années m'ont permis d'effectuer une thèse dans de très agréables conditions de travail. Malgré son poste de directeur de recherche, tu as également su rester disponible jusqu'à la fin. Merci !

****

Ces trois années de thèses n'auraient jamais été ce qu'elles ont été sans les membres du couloir 32-42 4ème étage de l'Institut Jacques Monod. Je me souviens encore de mes premiers pas dans ce long couloir, des murs en briques rouges et ces odeurs de milieux « Droso »… y retourner réveille en moi, une pointe de nostalgie, et cela je le dois à ces occupants.
C'est pourquoi, je tiens à remercier **Dominique Anxolabéhère** pour m'avoir chaleureusement accueilli au sein du laboratoire depuis toutes ces années. Merci pour votre confiance, vos encouragements et vos conseils de tous les instants. Je suis également très sensible à votre présence dans mon jury de thèse. …*milesker* Dominique.

****

Je tiens à remercier toutes les personnes qui ont contribué de près ou de loin, chacune à sa manière, à la réalisation de cette thèse.
Je commencerais par exprimer toute mon amitié à **Danielle Nouaud**. Je garderais précieusement en mémoire tous les bons souvenirs des congrès. Merci pour votre soutien et vos précieux conseils jusqu'aux derniers instants.
Je tiens à dire un grand merci à **Stéphane Ronserray** pour sa disponibilité, son écoute et ses encouragements depuis toutes ces années.
J'adresse également mes remerciements chaleureux à, **Dominique Higuet**, mon voisin de bureau toujours disponible pour des discussions scientifiques ou existentialistes.
Merci aussi à tous mes anciens collègues du laboratoire : **Olivier Andrieu, Paula Graca, Valérie Delmarre, Laure Teysset, Anne-laure Todeschini, Thibaut Josse, Eric Bonnivard**, **Clémentine Vitte**, **Micheline Jacques, Hervé Merçot, Yves Terrat, Charles Vejnar**, **Nicolas Buisine, Sébastien Tempel**…

Je tiens particulièrement à remercier mes anciens collègues et maintenant amis : **Elodie, Delphine, Hamid et Fabricio** pour leur présence qui a permis de rendre ces trois années de

thèse très enrichissantes et encore plus agréables. Merci à mes deux compères et amis **Terry** et **Mathieu** qui m'ont apporté un véritable soutien durant ses derniers mois.
Je n'oublie pas le personnel administratif sans qui rien n'aurait été possible : **Stéphane Hoyez** (mon sauveur)**, Danielle Merkiled** (le soleil de la martinique), **Elisabeth Roux-mendras** (toujours prête). Merci !

****

Je tiens à remercier les membres de l'URGI: **Michael, Isabelle Le., Delphine, Isabelle Ly., Timothée, Victoria, Sophie, Cyril, Sébastien, Nacer, Erik, Daphné, Joëlle, Claire, Véronique, Valérie, Christian** et **Emmanuelle**. Je suis heureuse d'avoir fait ce « petit » détour sur Evry pour passer ces quelques mois en votre compagnie. Mon seul regret sera de n'avoir pas pu passer plus de temps et dans de meilleures conditions (c'est-à-dire sans le stress de la rédaction). Bonne continuation à tous.
Une spéciale dédicace à la crew de l'URGI : Daphné, Joëlle et Nacer qui ont su rendre ces derniers mois plus agréables de part leur amitié à toute épreuve. Un grand merci !
Timothée, bon courage pour ta thèse, je ne m'inquiète pas du tout pour toi…tu es de surcroit, entre de « bonnes mains » !

****

Je tiens à remercier vivement toutes les personnes qui peut-être sans le savoir ont participé à mes choix professionels : **Monsieur Michel Fournier, Madame Cécile Butor, Madame Monik Amour**…
Je ne pouvais oublier le « promoteur » de la première licence-Maitrise de Bioinformatique de France, le professeur **Serge Hazout**. Je lui dédicace ce manuscrit car sans lui rien n'aurait été possible.
Merci à mes tuteurs de thèse **Jean-Michel Camadro** et **Olivier Vallon** pour m'avoir suivi durant ces trois années.
Je remercie également **Bernard Dujon** et **Casey Bergman** pour l'intérêt qu'ils ont portés à mon travail de thèse.
Je tiens également à faire-part de ma reconnaissance envers toutes les rencontres que j'ai pu faire durant des ces trois années, *via* l'enseignement. Je pense plus particulièrement à **Christiane Durieux** et **Denis Mestivier**.

****

J'avoue que le monde de la recherche n'est pas très évident à comprendre. Il est également vrai que certaines périodes de la thèse peuvent s'avérer très stressantes. C'est pourquoi, je tiens également à remercier ma famille, ma belle-famille et mes amis pour leur patience.
Je suis très fière de remercier tous mes amis pour leur soutien, leurs encouragements, leurs disponibilités et leur écoute : **Ludivine, Célia, Christophe, Emérence, Lélia, Guillaume, Eloïse, Raphaël, Elodie, Cyril, Laetitia, Aissam, Fabricio, Delphine, Manu, Hamid, Mya, Michel, Sylvia, Julien, Alexa, François, Annie, Lolo, Olivier, Lili, La « oulalaz team », Martine (et ses deux**

adorables chipies : Victoria et Adeline), Brunette, Boniface, Benoît, Baudouin…et tous les autres.

Merci à ma sœur bien-aimée, **Fabienne** avec qui j'ai écrit mes premières lignes de programmation, pour ta patiente, tes encouragements et toute l'attention que tu me portes.

Cyril, ma moitié, mon amour, mon mari…. je te remercie pour tout ce que tu m'as apporté et ce que tu continues à m'apporter chaque jour. Ta présence est pour moi un véritable bonheur. Ce travail est aussi le tient. Merci !

Maman, papa, je vous dédie ce travail, sans vous je ne serais pas où j'en suis aujourd'hui. Je vous remercie pour ce que vous êtes : des parents extraordinaires. Vous m'avez toujours fait confiance, vous avez cru en moi et vous m'avez soutenu même dans mes choix les plus périlleux. Merci.

"La façon dont on trouve n'est pas celle dont on prouve."
Albert Einstein

«ÉTUDE DE LA DYNAMIQUE DES RÉPÉTITIONS DANS LES GÉNOMES EUCARYOTES: DE LEUR FORMATION À LEUR ÉLIMINATION»

Résumé

De la bactérie à l'homme, dispersées ou en tandem, les répétitions peuvent représenter jusqu'à 90 % de la séquence génomique. Malgré leur impact sur la plasticité et l'évolution des génomes eucaryotes, leurs mécanismes de formation sont encore très spéculatifs. L'insertion continue de nouvelles répétitions devrait conduire à une augmentation constante de la taille des génomes. Or, il ne semble pas que ce soit le cas. Y a t-il régulation de la taille des génomes? Le processus de régulation est-il le même dans l'euchromatine et l'hétérochromatine?

Afin d'étudier la dynamique des répétitions, j'ai développé un ensemble de programmes informatiques pour la détection des duplications segmentaires (DS) et des répétitions en tandem (RT). A partir des caractéristiques des DS détectées chez *Drosophila melanogaster*, j'ai proposé un modèle de formation des DS, basé sur un modèle de recombinaison homologue non-allélique. J'ai également identifié les traces de l'implication des éléments transposables (ET) dans ce processus.

Afin de caractériser la relation existante entre les répétitions et la structure de la chromatine, j'ai ensuite réalisé une analyse comparative de la dynamique des répétitions euchromatiques et hétérochromatiques. Pour ce travail, nous avons choisi comme modèle d'étude *Arabidopsis thaliana*.

La construction d'arbres phylogénétiques des séquences répétées m'a permis de dater les répétitions. Nous suggérons ainsi une propagation par « vague » des ET. J'ai ensuite estimé les forces d'élimination des copies d'ET. Nos résultats suggèrent que dans l'euchromatine, la pression de sélection due aux gènes induit l'élimination des répétitions. Dans l'hétérochromatine, la faible densité en gènes permet de maintenir une forte densité en ET. Pourtant, les estimations du taux de perte en ADN, prédisent un turnover aussi rapide dans l'euchromatine que dans l'hétérochromatine. Afin de contre-balancer l'insertion des ET dans l'hétérochromatine, nous pouvons invoquer la recombinaison homologue non-allélique.

Mots-clés

Eléments transposables, duplications segmentaires, satellites, cassures double-brin d'ADN, réarrangements, recombinaison ectopique, structure de la chromatine.

«REPEAT DYNAMICS IN EUKARYOTIC GENOMES: FROM THEIR FORMATION TO THEIR ELIMINATION»

Abstract

From bacteria to human, interspersed or in tandem, repeated sequences can cover more than 90 % of a genomic sequence. Despite their impact on the evolution and the plasticity of eukaryotic genomes, their mechanism of propagation remains still unclear. The continuous insertion of new copies should induce the increase of genome size. What are the selection pressures involved in the genome size regulation? Do these selection strength are the same in euchromatic and heterochromatic regions? In order to highlight the repeat dynamics, I first developed computational pipelines to detect segmental duplications (SDs) and tandem repeat arrays (TRs).
The SD features of the detected in the *Drosophila melanogaster* genome, allowed us to propose a non-allelic homologous recombination mechanism as a SD formation model. This process can be induced by repeats such as TEs. Indeed, I showed the traces of transposable elements (TEs) at their breakpoints.
To understand the relationship between the repeats and the chromatin structure, we investigated the repeat evolutionary dynamics by comparing their features in heterochromatin and euchromatin domains in *Arabidopsis thaliana*. We constructed phylogenetic trees of repeats to estimate their divergence in euchromatin and heterochromatin. The tree topology of TE families reflects transpositions by “burst”. In order to explain, the size and divergence variations of the TE copies between these two chromatic domains, we estimated the strength of repeat elimination into these regions. Our analysis suggests that the gene selection pressure effect induces in euchromatin the repeat elimination, although, in heterochromatin, the gene paucity allows to maintain the high TE density. However, the DNA loss rate estimations suggest the same fast turnover in the both chromatin domains. To counteract the TE insertion in heterochromatin, we proposed that non-allelic homologous recombination may play a significant role. This process allows to eliminate rapidly lots of copies.

Keywords

Transposable elements, segmental duplications, satellites, double-strand break DNA repair, rearrangement, non-allelic homologous recombination, chromatine structure.

Table des matières

1. INTRODUCTION GENERALE 1

1.1. Les différents types de répétitions des génomes 3
1.1.1. Les Éléments Transposables (ET) 3
1.1.2. Les Duplications Segmentaires (DS) 5
1.1.3. Les Répétitions en Tandem (RT) 6

1.2. Leur dynamique 7
1.2.1. Propagation des ET 7
1.2.2. Modèles de formation des DS 14
1.2.3. Expansion des RT 17

1.3. Evolution et impact fonctionnel 19
1.3.1. Evolution des familles multigéniques 19
1.3.2. Rôle de duplications géniques 20
1.3.3. L'impact fonctionnel des ET 21

1.4. Des acteurs de la plasticité des génomes 23
1.4.1. Les transpositions alternatives 23
1.4.2. La macrotransposition 23
1.4.3. La recombinaison entre copies de répétitions 25
1.4.4. La formation des solo-LTR 25

1.5. Problématiques 26
1.5.1. L'évolution de la taille des génomes 26
1.5.2. Le plan de travail 28

2. LES METHODES DE DETECTION DES REPETITIONS 31

2.1. Deux approches de détection des répétitions 31
2.1.1. Les méthodes de novo 31
2.1.2. Les méthodes avec connaissance a priori 31
2.1.3. REPET : un pipeline d'annotation des ET 32

2.2. SegDupPipeline: un pipeline de détection des DS 33
2.2.1. Etape n°1: Détecter l'ensemble des répétitions d'un génome 34
2.2.2. Etape n°2: Eliminer du jeu de données les autres répétitions 35
2.2.3. Etape n°3: Détection de duplications « segmentaires » 36
2.2.4. Etape n°4: Eliminer les évènements de transposition 37
2.2.5. Etape n°5: Construction des familles de duplication 38

2.3. Comment détecter les RT ? 39
2.3.1. Pals et Piler-TA, des programmes spécifiques 39
2.3.2. Une méthode de détection 41
2.3.3. TandemPipeline: un pipeline de détection des RT 41

2.4. Les applications 42

3. LA DYNAMIQUE DES DS CHEZ *DROSOPHILA MELANOGASTER* 45

***3.1. Pourquoi avoir choisi Drosophila melanogaster ?* 46**
3.1.1. Son séquençage 46
3.1.2. Et les autres Drosophilidés 47
3.1.3. Un génome de choix 48

***3.2. Obtention des données* 49**
3.2.1. Les annotations des répétitions 49
3.2.2. La création d'un jeu de séquences contrôles 49

***3.3. Les méthodes utilisées* 50**
3.3.1. Calcul de la fraction recouverte en séquence 50
3.3.2. Détection des répétitions aux extrémités des DS 50
3.3.3. Analyse des points de cassure des DS 50
3.3.4. Création des blocs de synténie 51
3.3.5. Identification de la séquence matrice 51
3.3.6. Etude de la divergence entre les copies de duplication 52

***3.4. Les Caractéristiques des DS chez D. melanogaster* 53**
3.4.1. Leurs caractéristiques générales 53
3.4.2. Les familles multigéniques 58
3.4.3. Les minisatellites et duplications en tandem 59

***La Réparation Homologue (RH), un mécanisme de formation de duplications* 61**
3.4.4. Les modèles de réparation des Cassures Double-brin (CDB) d'ADN 61
3.4.5. La RH des CDB d'ADN 63
3.4.6. Le modèle SSA (« Single-Strand Annealing ») 65
3.4.7. Le modèle DSBR (« Double-Strand Break Repair ») 65
3.4.8. Le modèle BIR (« Break-Induced Replication ») 67
3.4.9. modèle SDSA (« Synthesis-Dependent Strand Annealing ») 69
3.4.10. Un modèle de formation des DS, un variant du SDSA 71

***3.5. Identification des traces du mécanisme* 73**
3.5.1. Les traces de la ré-invasion du brin 73
3.5.2. Un modèle affiné 78

***3.6. Les ET, inducteurs du processus de duplication* 80**
3.6.1. Présence de répétitions aux extrémités des duplications 80
3.6.2. Recombinaison entre éléments aux points de jonction des DS 81

***3.7. Conclusion/Discussion* 83**
3.7.1. Un mécanisme dépendant de la taille du génome 83
3.7.2. Le DDSA, un modèle de formation des DS 83

4. LE TURNOVER DES SEQUENCES 87

***4.1. Pourquoi avoir choisi Arabidopsis thaliana ?* 88**
4.1.1. Un bon modèle d'étude 88
4.1.2. hk4S, le *knob* du chromosome 4 89
4.1.3. Une grande part des régions hétérochromatiques séquencées 92

4.1.4. Les annotations disponibles 93

4.2. Méthodes de datation des répétitions 94
4.2.1. Les méthodes classiques de datation des séquences 94
4.2.2. La longueur des branches terminales des arbres 96
4.2.3. DTscore, pour les arbres de duplication en tandem 98

4.3. Méthodes d'estimation des vitesses évolutives 99
4.3.1. Taux de petites délétions par maximum de vraisemblance 99
4.3.2. Pression de sélection due aux gènes 101
4.3.3. Taux de recombinaison homologue non-allélique 101

4.4. Les répétitions chez A. thaliana 102
4.4.1. Leur distribution chromosomique 103
4.4.2. Caractéristiques des ET 104
4.4.3. Caractéristiques des DS 107
4.4.4. Caractéristiques des RT 110

4.5. Dynamique comparée des répétitions 113
4.5.1. Insertion des ET par vague 113
4.5.2. Dynamique d'insertion des DS 117
4.5.3. Dynamiques d'expansion des RT 119

4.6. Les forces d'élimination des répétitions 120
4.6.1. Effet de la sélection 121
4.6.2. Impact de la recombinaison ectopique 124
4.6.3. Les petites délétions 126

4.7. Impact des répétitions sur la plasticité du génome 130
4.7.1. Dynamique de formation des DS 130
4.7.2. Perte et gain d'ADN 132

4.8. Conclusion/Discussion 133
4.8.1. Deux profils de dynamique 133
4.8.2. Dynamique d'élimination des ET 134
4.8.3. Les modèles de dynamique 135
4.8.4. Le turn-over des séquences 137
4.8.5. La dynamique d'insertion des répétitions dans *hk4S* 139

5. DISCUSSION GENERALE 143

5.1.1. Des caractéristiques des DS 143
5.1.2. La compartimentation des répétitions 145
5.1.3. Contraction de taille des génomes médiée par les répétitions 145
5.1.4. Les ET, éléments moteurs de l'augmentation de la taille des génomes 147

5.2. Vers une approche de Génomique populationnelle 148
5.2.1. Les données génomiques disponibles 148
5.2.2. Génomique comparative et Génétique des populations 148

6. ANNEXES 150

6.1. Article 1: «Detection of transposable elements by their compositional bias» 151

6.2. Article 2: « A model of segmental duplication formation in Drosophila melanogaster » 152

6.2.1. Annexe1: Coordonnées des DS détectées 153

6.2.2. Annexe2: Analyse des 12 évènements de duplication sélectionnées 154

6.2.3. Annexe3: Analyse des «*clusters*» de RT 155

6.2.4. Annexe4: Analyse de points de cassure des duplications 156

6.3. Article 3 (en cours d'écriture) 157

Figures

FIGURE 1. CLASSIFICATION DES ET. 2
FIGURE 2. MECANISMES DE LA RETROTRANSPOSITION. 8
FIGURE 3. MECANISME DE TRANSPOSITION. 10
FIGURE 4. MODELE DE FORMATION DES ELEMENTS MITE. 11
FIGURE 5. MODELES DE FORMATION DES ELEMENTS DE CLASSE III. 12
FIGURE 6. MODELES DE FORMATION DES DS CHEZ LES MAMMIFERES. 14
FIGURE 7. PROCESSUS D'EXPANSION ET DE CONTRACTION DES RT. 17
FIGURE 8. MECANISME DE FORMATION DES RT *VIA* LA CONVERSION GENIQUE. 18
FIGURE 9. LES TROIS MODELES D'EVOLUTION DES FAMILLES MULTIGENIQUES. 19
FIGURE 10. LA TRANSPOSITION ALTERNATIVE. 22
FIGURE 11. LA MACROTRANSPOSITION 22
FIGURE 12. LES MECANISMES EVOLUTIFS DES REPETITIONS PAR RH. 24
FIGURE 13. LE PIPELINE DE DETECTION DES DS. 33
FIGURE 14. ETAPE D'ELIMINATION DES ET ET DES SATELLITES. 35
FIGURE 15. DEUXIEME ETAPE D'ELIMINATION DES ET. 37
FIGURE 16. DETECTION DES RT PAR PALS/PILER-TA. 40
FIGURE 17. DETECTION DU «*CLUSTER*» DES GENES *HISTONE*. 40
FIGURE 18. LE PIPELINE DE DETECTION DES RT. 42
FIGURE 19. LES 12 ESPECES DE DROSOPHILES. 47
FIGURE 20. IDENTIFICATION DE LA SEQUENCE MATRICE. 51
FIGURE 21. POURCENTAGE D'IDENTITE PAR NOMBRE DE COPIES DE DUPLICATION. 53
FIGURE 22. TAILLE DES COPIES EN FONCTION DU NOMBRE DE COPIES PAR GROUPE. 54
FIGURE 23. NOMBRE DE COPIES PAR GROUPE DE DUPLICATION. 54
FIGURE 24. DISTRIBUTION CHROMOSOMIQUE DES DES ET ET DES DS DETECTEES. 55
FIGURE 25. LE MODELE SSA. 64
FIGURE 26. LE MODELE DSBR DE SZOSTAK. 64
FIGURE 27. LES TROIS MODELES BIR. 66
FIGURE 28. LE MODELE SDSA AVEC FORMATION D'UNE BULLE DE MIGRATION. 68
FIGURE 29. LE MODELE DDSA. 70
FIGURE 30. TRACES ATTENDUES D'APRES UN PROCESSUS DE RE-INVASION. 74
FIGURE 31. ANALYSE DES POINTS DE CASSURE DES DS. 82
FIGURE 32. HISTOIRE EVOLUTIVE DU GENOME DE *A. THALIANA*. 88
FIGURE 33. *HK4S*, LE KNOB DU CHROMOSOME 4 DE *A. THALIANA*. 89
FIGURE 34. SCHEMA DE L'HISTOIRE EVOLUTIVE DE *HK4S*. 90
FIGURE 35. LE *KNOB* DU CHROMOSOME 4 (*HK4S*). 90
FIGURE 36. QUELQUES TERMES PHYLOGENETIQUES. 95
FIGURE 37. RECOUVREMENT EN SEQUENCES DES CHROMOSOMES DE *A. THALIANA*. 102
FIGURE 38. PROPORTION DES ET PAR SUPERFAMILLES. 104
FIGURE 39. DISTRIBUTION DE LA TAILLE DES COPIES. 105
FIGURE 40. NOMBRE DE COPIES PAR GROUPE DE SEQUENCES DUPLIQUEES. 107
FIGURE 41. POURCENTAGE D'IDENTITE DES DUPLICATIONS DETECTEES. 108
FIGURE 42. TAILLE DES DUPLICATIONS DETECTEES SELON LE NOMBRE DE COPIES. 109
FIGURE 43. DISTRIBUTION DE LA TAILLE DES RT. 111
FIGURE 44. DISTRIBUTION DE LA DIVERGENCE DES ET PAR DOMAINE. 113
FIGURE 45. DIVERGENCE DES ET PAR SUPERFAMILLE. 114

FIGURE 46. EXEMPLES D'ARBRES DE TRANSPOSITION. 116
FIGURE 47. DISTRIBUTION DE LA DIVERGENCE DES DS PAR DOMAINE. 117
FIGURE 48. EXEMPLES D'ARBRES DE RTET DE DS. 118
FIGURE 49. DISTRIBUTION DE LA DIVERGENCE DES RT PAR DOMAINE. 119
FIGURE 50. DEUX EXEMPLES D'ARBRES DE « *CLUSTERS* » DE RT. 120
FIGURE 51. RELATION: DENSITE EN ET/DENSITE EN GENES/TAUX DE RECOMBINAISON. 122
FIGURE 52. GRAPHIQUE DES REGRESSIONS TAILLE/DISTANCE ET TAILLE/DIVERGENCE. 123
FIGURE 53. RELATION TAILLE DES COPIES/DELETIONS. 128
FIGURE 54. EFFET « HILL-ROBERTSON ». 135
FIGURE 55. DYNAMIQUE DES ET CHEZ *A. THALIANA*. 136

Tableaux

TABLEAU 1. DUPLICATIONS INTERCHROMOSOMIQUES ET INTRACHROMOSOMIQUES. 56
TABLEAU 2. PROPORTIONS DE DS CHEZ CERTAINS GENOMES EUCARYOTES. 58
TABLEAU 3. LES GROUPES A PLUS DE 5 COPIES. 59
TABLEAU 4. TRACES DU PROCESSUS DE RE-INVASION DU BRIN SELON LE DDSA. 76
TABLEAU 5. PRESENCE D'ET ET DE MICROSATELLITES AUX EXTREMITES DES DS. 81
TABLEAU 6. RECOUVREMENT DES CHROMOSOMES EN SEQUENCES. 103
TABLEAU 7. REPARTITION ET DENSITE CHROMOSOMIQUE DES DS PAR DOMAINE. 107
TABLEAU 8. LA COMPOSITION DES DS DETECTEES. 110
TABLEAU 9. TAUX DE RETROELEMENTS A LTR. 124
TABLEAU 10. TAUX DE PETITES DELETIONS PAR DOMAINE. 126
TABLEAU 11. PRESENCE DE REPETITIONS AUX EXTREMITES DES DS CHEZ *A. THALIANA*. 130

Abréviations

ET: Elément Transposable
DS: Duplication Segmentaire
RT: Répétition en Tandem
RI: Région Interne
CDB: Cassure Double-Brin
RH: Réparation Homologue

pb: paire de base
kb: kilobase
Mb: Megabase

LTR: «*Long Terminal Repeats*»
TSD: «*Target Site Duplication*»
LINE: «*Long INterspersed Elements*»
SINE: «*Short INterspersed Elements*»
TIR: «*Terminal Inverted Repeats*»
MITE: «*Miniature Inverted-repeats Transposable Elements*»
NAHR: «*Non-Allelic Homologous Recombination*»
SDSA: «*Single Dependant-Strand-Annealing*»
DDSA: «*Duplication Dependant-Strand-Annealing*»
RU: «*Repbase Update*»

1. Introduction générale

Les séquences d'ADN répétées ont longtemps été considérées comme des éléments parasites des génomes. Elles se révèlent aujourd'hui comme des éléments moteurs de l'évolution. De la bactérie à l'homme, dispersées ou en tandem, elles peuvent représenter jusqu'à 90 % de la séquence d'un génome tel que pour le génome du blé. Chez l'homme, plus de 50 % du génome se compose de répétitions. Même si leur impact reste encore flou, il est clair qu'elles participent activement à l'organisation et à l'évolution des génomes. En effet, l'insertion de ces séquences peut être délétère ou peut perturber le fonctionnement des génomes. Pourtant, leurs dynamiques et mécanismes de formation restent encore obscurs.

On distingue trois types de répétitions: (1) les éléments transposables (ET), (2) les satellites ou répétitions en tandem (RT), (3) et les duplications segmentaires (DS). Elles présentent toutes trois des caractéristiques qui leur sont propre: organisation, nombre de copies, taille, composition, etc. Ces caractéristiques peuvent quelque peu différer d'un organisme à un autre. Dans les génomes eucaryotes, les régions hétérochromatiques en sont enrichies, ce qui suggère une relation entre la structure de la chromatine et la présence de répétitions dans ces régions.

Dans ce premier chapitre, après un rappel des principales caractéristiques des répétitions, je présenterai les modèles actuels permettant d'expliquer la formation et l'évolution des répétitions. Puis, j'introduirai la problématique de ma thèse en abordant leur impact fonctionel et structural sur les génomes.

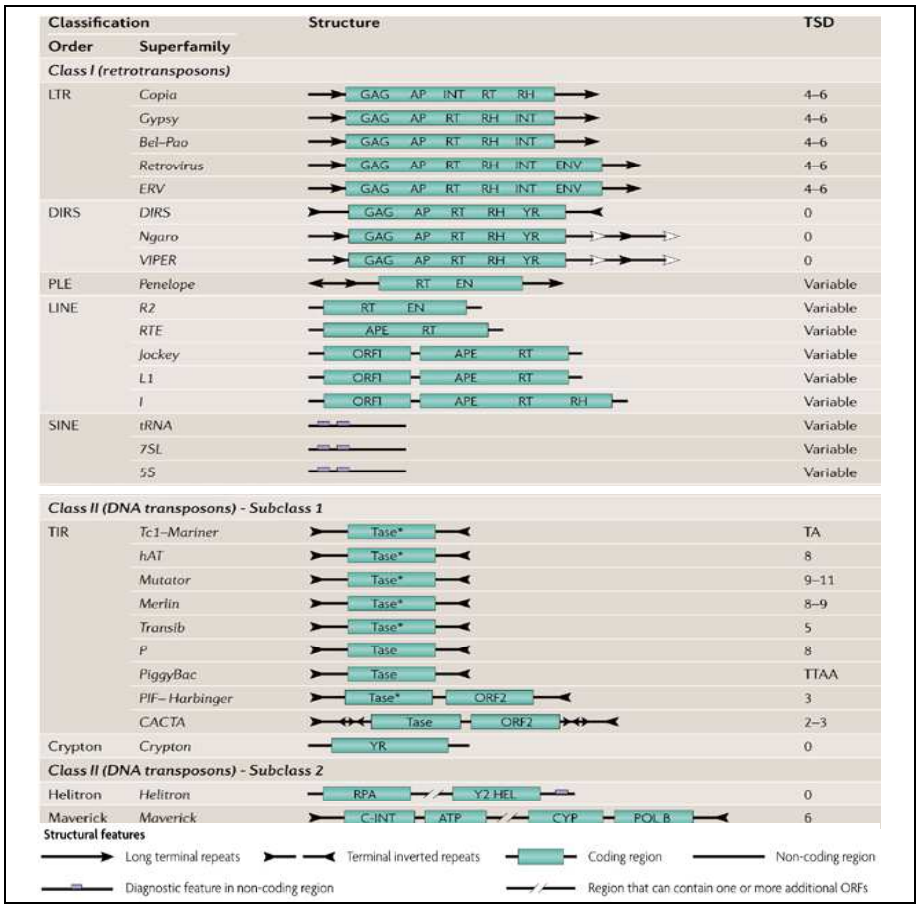


Figure 1. Classification des ET.

Les ET sont subdivisés en deux classes. Les éléments se distinguent d'abord en fonction de leur intermédiaire de transposition (ARN ou ADN), puis de leur structure. Les éléments de classe I ou éléments à intermédiaire ARN ne nécessitent qu'une seule cassure pour la transposition. On distingue parmi les rétroéléments, les éléments avec ou sans LTR, des répétitions directes présentes aux extrémités des séquences. Pour les transposons ou éléments de classe II, les éléments transposent *via* un intermédiaire ADN. Leur mécanisme est initié par 2 cassures d'ADN. On distingue deux sous-classes de transposons d'après leur mécanisme de transposition: Les éléments de la sous-classe 1 se propageant *via* des transposases alors que ceux de la sous-classe 2 ont des mécanismes de propagation différents (Wicker *et al.* 2007).

1.1. Les différents types de répétitions des génomes

1.1.1. Les Éléments Transposables (ET)

Les ET ont été découverts par Barbara Mc Clintock (1950). Elle interpréta de façon inédite des mutations instables du maïs, en les associant à des déplacements spontanés de fragments d'ADN dans le génome. Ces éléments sont capables de sauter d'un endroit à un autre d'un génome: ils transposent. Les ET peuvent représenter jusqu'à 90 % de la séquence d'un génome comme celui du blé (SanMiguel et al. 1998). Chez l'homme, près de la moitié du génome correspond à des ET (Lander et al. 2001). Considérés comme hautement mutagènes, ils s'insèrent fréquemment dans les régions codantes ou régulatrices.

Plusieurs systèmes de classification des ET ont été proposés. Certains auteurs les classent d'après le mécanisme de transposition, et d'autres d'après leur structure. Finnegan a proposé, en 1989, la première classification selon leur intermédiaire de transposition (Finnegan 1989). On distingue ainsi les éléments de classe I, aussi appelé rétrotransposons, qui transposent *via* un intermédiaire ARN selon un mécanisme de « copier-coller », et les éléments de classe II ou transposons qui transposent *via* un intermédiaire ADN selon un mécanisme de « couper-coller ». Pendant de très nombreuses années, cette classification a été considérée comme la classification de référence. L'observation d'éléments au mécanisme différent des deux autres, a conduit à la création d'une troisième classe (ou classe III). L'accumulation de nouvelles données et l'émergence de nouvelles techniques d'étude ont fait place à une nouvelle classification basée sur le mécanisme (Curcio et Derbyshire 2003). Elle tient compte du nombre de cassures nécessaire à la transposition.

Wicker *et al.* ont proposé en 2007 une nouvelle classification réconciliant intermédiaire et mécanisme de transposition. Cette classification permet de garder la notion d'élément de classe I et de classe II (Figure 1). Une nomenclature a également été proposée. N'étant pas encore admise par tous, celle-ci ne sera pas utilisée dans ce manuscrit (Wicker et al. 2007).

Les rétrotransposons (Figure 1**)**, appelés aussi éléments de classe I, sont d'abord transcrits sous forme d'ARN. Ils produisent un autre enzyme, la transcriptase reverse, qui synthétise une molécule d'ADN à partir de cet ARN. Puis, le nouveau fragment d'ADN synthétisé s'insère ailleurs dans le génome. On distingue deux grands groupes de rétrotransposons: les rétroéléments sans LTR pour « *Long Terminal Repeats* » et les rétroéléments à LTR. Ces derniers possèdent les domaines de la glycoprotéine de capside (ou « gag ») et de la polymérase (ou « pol »). Ils sont de plus bornés par deux répétitions directes: les LTR. Ces LTR se compose de 3 éléments: U3, R et U5. Les éléments R sont des courtes répétitions alors que les éléments U3 et U5 correspondent à des régions ternimales uniques. Seule l'absence du domaine « env » de l'enveloppe différencie les rétroéléments à LTR des rétrovirus.
Parmi les rétroéléments dépourvus de LTR, on distingue les LINE pour « *Long INterspersed Elements* » et les SINE pour « *Short INterspersed Elements* ». Les éléments de ces deux superfamilles ont en 3'terminale une queue poly-A ou un microsatellite de taille variable.

Les transposons (Figure 1) ou éléments de classe II, transposent *via* un intermédiaire ADN. Ces éléments, s'ils sont autonomes, codent une transposase. Grâce à celle-ci, une copie est excisée de son site donneur pour s'insérer ailleurs dans le génome. D'après le mécanisme de transposition, on distingue deux sous-classes de transposons.
Le mécanisme des éléments de la sous-classe 1, est orchestré par la transposase. Celle-ci induit une Cassure Double-brin (CDB) au site donneur qui est ensuite réparée. Les éléments dits non-autonomes car dépourvus de transposase, peuvent se propager grâce à la transposase d'un autre ET. Les transposons sont caractérisés par des séquences répétées terminales et inversées appelées des TIR pour « *Terminal Inverted Repeats* ». On retrouve dans cette classe les éléments dit MITE pour « *Miniature Inverted-repeats Transposable Elements* ». La taille de ces éléments n'atteint que quelques centaines de paires de bases (pb). Même si leur mécanisme de propagation demeure encore mal connu, la présence de TIR et de TSD suggère que ces éléments dérivent d'éléments de classe II, et sont mobilisés par des éléments autonomes de leur famille d'origine. Poutant, la région interne ne présente aucune similitude avec celle des transposons d'origine.
Dans la deuxième sous-classe, on retrouve les *Hélitrons* et les *Polintons* (ou *Maverick*). Les mécanismes proposés diffèrent de ceux des éléments de la première sous-classe.

Les *Hélitrons* ont un domaine protéique HEL. Ce domaine se compose d'une hélicase et un domaine initiateur de la réplication. Le mécanisme réplicatif est connu sous le nom de « *rolling-circle* » (Kapitonov et Jurka 2001, 2007). La séquence de ces éléments contient également une région palindromique à proximité de la région 3' terminale (de 10 à 12 pb avant la fin de l'élément). Cette région palindromique forme une structure en épingle à cheveux de moins de 20 pb. Toutes les copies de cette superfamille sont bornées en 5' par TC et en 3' par CTRR (la lettre R correspond à une base purine).

Les *Polintons* se composent d'une intégrase et d'une polymérase. Des TIR ont également été décrits aux bornes des *Polintons*. Bien que plusieurs modèles aient été proposés, le mécanisme de transposition de ces éléments reste encore indéterminé.

1.1.2. Les Duplications Segmentaires (DS)

A l'opposé des ET, les DS sont généralement des répétitions en faible nombre de copies (Nous fixons ce seuil à ≤ 5 copies). Ces répétitions peuvent couvrir plusieurs kilobases (kb). Lorsqu'elles concernent des chromosomes entiers, on parle d'hyperploïdie. Des duplications peuvent également se produire à l'échelle du génome entier et conduire ainsi à de la polyploïdie. Chez les primates, la taille des DS détectées varie entre 1 et 400 kb. Les copies partagent une forte identité de séquences entre elles (> 90 %). Elles auraient vu le jour lors des 35 derniers millions d'années. Des analyses de FISH (« *Fluorescent In Situ Hybridization* ») et *in silico* ont révélé que plus de 5 % de la séquence du génome humain correspond à des DS (Bailey *et al.* 2001; 2002; Cheung *et al.* 2003; She *et al.* 2004). Les DS humain ayant de 98 % à 100 % d'identité de séquence sont réparties tout le long des chromosomes. Pourtant, des zones d'accumulation de DS ont été identifiées. Leur localisation semble liée à la distribution des points chauds de recombinaison (Ji *et al.* 2000; Armengol *et al.* 2003; Locke *et al.* 2003). Pour les DS moins conservées (de 90 % à 98 %), on observe un enrichissement au niveau des régions péricentromériques et subtélomériques. On distingue donc 3 types de DS, en fonction de leur localisation et leur structure: les DS péricentromériques, les DS subtélomériques et les DS dispersées euchromatiques (Bailey et Eichler 2006). Les DS peuvent aussi bien être composées d'ET que de gènes. Chez les primates, ces segments dupliqués apparaissent riches en ET. Les segments ne recouvrant qu'un seul gène, et cela entièrement, sont appelés « duplications géniques ». D'autres segments peuvent recouvrir plusieurs copies d'ET et plusieurs gènes. Une famille multigénique

correspond à un groupe de gènes partageant des fonctions similaires ou proches. Les gènes de ces familles, issus de telles duplications, sont généralement organisés en « *clusters* » (Nei et Rooney 2005).

1.1.3. Les Répétitions en Tandem (RT)

Les séquences d'ADN satellites sont des séquences répétées en tandem. Elles sont composées de deux ou plusieurs répétitions adjacentes d'un même segment d'ADN. Elles se forment lorsqu'un fragment d'ADN est copié en deux fragments adjacents. On les trouve chez tous les eucaryotes. Ces répétitions représentent environ 3 % du génome humain. Elles sont majoritairement localisées au niveau des télomères et des centromères. L'impact de ces séquences dans de nombreuses maladies humaines a attiré l'attention de la communauté scientifique.

On distingue trois types de satellites: les satellites (Csink et Henikoff 1998), les minisatellites (Jarman et Wells 1989) et les microsatellites (Karlin et Burge 1995). On les différencie principalement par la taille des unités de répétitions. Les microsatellites sont des séquences dont l'unité varie de 1 à 6 pb. La taille des unités des minisatellites va de 6 pb à une centaine de paires de bases et les satellites jusqu'à plusieurs centaines de paires de bases. On associe aux microsatellites et minisatellites différents noms tel que SSR pour « *Simple Sequence Repeats* » ou STR pour « *Short Tandem Repeats* » ou encore VNTR pour « *Variable Number of Tandem Repeats* ».

1.2. Leur dynamique

1.2.1. Propagation des ET

On estime le taux de transposition par élément à 10^{-3} ou 10^{-4} évènements par génération chez *Drosophila melanogaster*. Le taux d'excision dite « précise » est estimé entre 10^{-6} et 10^{-10} par élément et par génération, soit environ 10 000 fois plus faible que le taux de transposition (Harada *et al.* 1990; Nuzhdin et Mackay 1995; 1997; Maside *et al.* 2000; 2001). Le taux d'insertion des copies étant plus important que le taux d'élimination par excision, on devrait s'attendre à une accumulation d'ET, c'est-à-dire une constante augmentation de la taille des génomes.

Or, ce n'est pas ce que l'on observe. Il y a donc régulation du nombre de copies d'ET. Les insertions délétères altérant l'expression des gènes étant contre-sélectionnées, la régulation du nombre d'ET peut donc s'exercer *via* la sélection (Biemont et al. 1997; Nuzhdin 1999).

Les premières études ont suggéré une insertion aléatoire des ET dans un génome. Mais certains sites dits « chauds » sont plus propices à l'insertion d'ET. Ces régions sont généralement riches en satellites. Les ET se retrouvent aussi en abondance dans les régions hétérochromatiques.

En fonction du type d'élément, le mécanisme de transposition étant différent, on ne s'attend pas à observer une dynamique d'insertion similaire. En effet, la propagation d'un transposon conduit à l'excision de la copie au site donneur. Une CDB est alors créée. Différentes voies de réparation rentrent ensuite en compétition: des mécanismes de RH pour « Réparation Homologue » et des mécanismes non-homologues.

Si la cassure est réparée *via* un mécanisme de RH, une autre copie de cet élément peut servir de séquence matrice. Une nouvelle copie est ainsi insérée au site de cassure. Par contre, si cette cassure est réparée *via* un mécanisme de réparation non-homologue, le nombre de copies peut rester inchangé. Seule la localisation d'une copie aura été modifiée. Le mécanisme de transposition des éléments de classe I, quant à lui, conduit toujours à l'augmentation du nombre de copies.

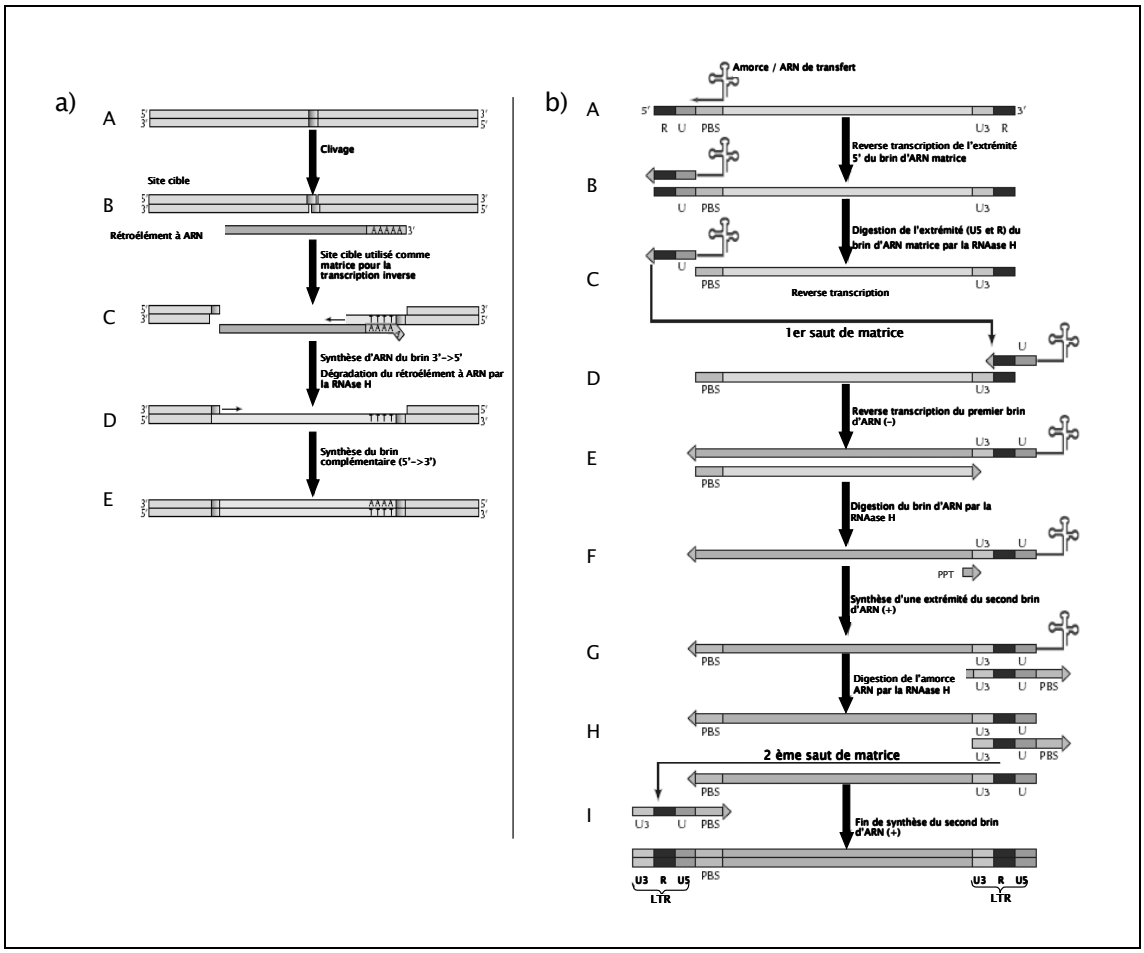


Figure 2. Mécanismes de la rétrotransposition.

a) Mecanisme de la réverse transcription des rétroéléments sans LTR. (**A**) l'endonucléase crée une cassure simple brin au site d'insertion. (**B**) Le brin d'ARN du rétroélément s'apparie au niveau de la queue poly-A. (**C**) Transcription inverse et clivage du brin complémentaire. (**D**) Le brin d'ARN est dégradé par le RNAase H. (**E**) Synthèse du deuxième brin d'ADN. **b) Mécanisme de rétrotransposition des éléments à LTR.** (**A**) Le mécanisme est initié par la fixation de l'ARN de transfert sur le site de liaison: le PBS. (**B**) Synthèse du brin d'ARN (-). (**C**) Digestion par la RNAaseH de R et U5. (**D**) Premier saut de matrice de la région d'ARN nouvellement séquencée de l'extrémité 5' vers l'extrémité 3'. (**E**) Synthèse d'ARN du brin (-) puis décrochage du PBS et le clivage au site U3. (**F**) Digestion par la RNAaseH du brin (+) de l'ARN matrice. (**G**) Synthèse du nouveau brin d'ARN (+), initiée au niveau du PPT puis dégradation par la RNAase H de l'amorce. (**H**) Second saut de matrice: le brin d'ARN nouvellement synthétisé s'apparie au niveau l'extrémité 3' *via* les PBS. (**I**) Synthèse du deuxième brin d'ARN.

La transposition par « copier-coller » ou rétrotransposition

Les mécanismes de rétrotransposition des éléments à LTR et des éléments sans LTR, sont totalement différents (Figure 2). Les mécanismes de rétrotransposition font principalement appel à deux protéines: une transcriptase inverse et une endonucléase.

La propagation des transposons à ARN sans LTR se déroule en 4 grandes étapes (Figure 2 a) (Christensen et Eickbush 2005): (A) l'endonucléase du rétroélément crée une cassure simple brin au site d'insertion. (B/C) la transcription inverse de l'extrémité 3' OH libre peut alors être initiée. Pour cela, le brin d'ARN du rétroélément s'apparie au niveau de la queue poly-A (ou microsatelitte). La synthèse d'ARN conduit au clivage du brin d'ADN complémentaire (D) En fin de synthèse d'ADN, le brin d'ARN est dégradé par le RNAase H. (E) La synthèse du deuxième brin d'ADN utilise comme matrice, le brin nouvellement synthétisé.

Le cycle des éléments à LTR ressemblent énormément à celui des rétrovirus (Figure 2). Le mécanisme est initié par la fixation de l'ARN de transfert sur le site de liaison: le PBS pour « «*Primer Binding Site* » (Figure 2bA). La synthèse du brin d'ARN (-) est initiée à partir de l'amorce d'ARNt (Figure 2bB). La RNAaseH dégrade ensuite l'extrémité 5' de l'élément (R et U5) (Figure 2bC), ce qui indui la recherche d'une nouvelle région d'appariement. L'action de la RNAes H induit par conséquent un saut de matrice de la région d'ARN nouvellement séquencée de l'extrémité 5' vers l'extrémité 3' (Figure 2 bD). La synthèse d'ARN du brin (-) se finit par le décrochage du PBS puis le clivage au site U3 (Figure 2bE). La RNAse H dégrade alors le brin (+) de l'ARN matrice (Figure 2bF). La synthèse du nouveau brin d'ARN (+) peut être alors initiée au niveau du PPT pour « *PolyPurine Tract* » (Figure 2 bG). Après dégradation par la RNAase H de l'amorce de la synthèse du deuxième brin d'ARN, il y a un second saut de matrice (Figure 2 bH): le brin d'ARN nouvellement synthétisé s'apparie au niveau l'extrémité 3' *via* les PBS, ce qui permet de continuer la synthèse du deuxième brin d'ARN (Figure 2 bI).

La transposition par« couper-coller »

La formation d'un dimère à partir de deux molécules de transposase est la première étape de la transposition des éléments de classe II à TIR (Figure 3) (Reznikoff *et al.* 1999). La transposase se fixe au niveau des deux TIR afin d'induire leur appariement. Le transposon forme ainsi une boucle. La transposase coupe les deux brins d'ADN grâce à son activité endonucléase. L'élément s'excise ainsi du site donneur et s'insère ailleurs dans le génome. La CDB produite au site donneur peut être réparée par un mécanisme de RH*. Dans ce cas, la séquence d'une autre copie de cet élément (généralement celle sur la chromatide sœur du site correspondant) servira de séquence matrice. Au site receveur, l'insertion de la copie excisée crée une nouvelle CDB. Cette cassure conduit à la formation d'extrémités cohésives complémentaires.

Les mécanismes de transposition des éléments de classe I et de classe II finissent par la réparation des bases manquantes aux bornes de la copie insérée. Cette étape induit la formation de deux petites duplications directes (< 10 pb) aux bornes de l'élément: les TSD pour « *Target Site Duplications* » (Figure 3). La présence de TSD est donc une marque de l'activité de la copie.

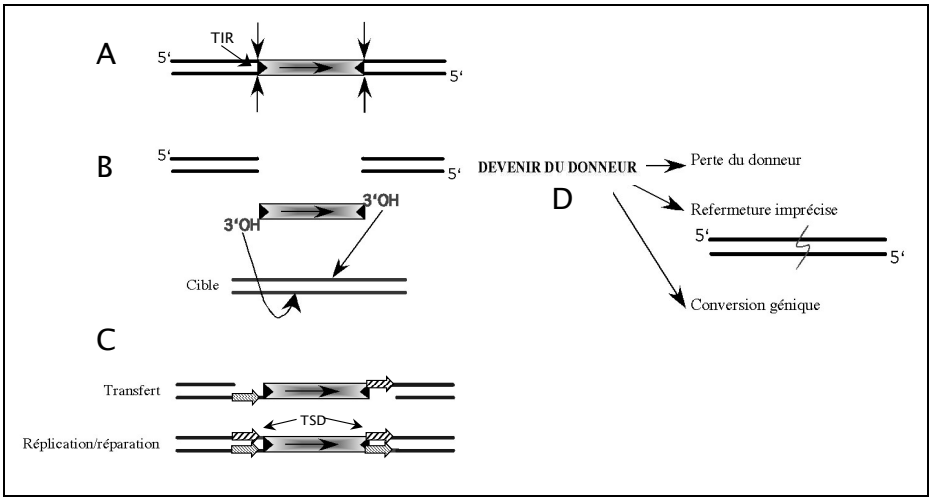


Figure 3. Mécanisme de transposition.

A. L'excisé de l'élément au site donneur se fait au niveau des TIR. Elle crée ainsi une CDB. **B.** La copie excisée s'insère au niveau d'un site accepteur. L'insertion nécessite une autre CDB. **C.** Cette cassure conduit à la formation d'extrémités cohésives complémentaires. La réparation des bases manquantes induit la formation des TSD. **D.** La CDB générée au site donneur pourra être réparée par conversion génique, ce qui permet de restaurer l'élément. Elle peut également être réparée par ligation des extrémités libres, ou fermeture imprécise (Reznikoff *et al.*, 1999).

* Réparation Homologue

Modèle de formation des éléments MITE

Un modèle de formation des MITE a été proposé par Feschotte *et al.* en 2002 (Feschotte et al. 2002). Ces auteurs considèrent les éléments MITE comme des éléments non-autonomes dérivant de transposons autonomes. D'après leur modèle, la formation des éléments MITE se déroule en deux étapes (Figure 4). La première correspond à l'amplification du nombre de copies de plusieurs familles de transposons autonomes proches en séquence. Les copies vont ensuite diverger. Seules les régions des TIR restent conservées entre les copies. Pour expliquer la petite taille de ces éléments, ils proposent une réparation incomplète de la CDB (Engels *et al.* 1990; Rubin et Levy 1997). Les éléments non-autonomes transposent ensuite grâce à une transposase amenée en *trans* par des copies autonomes de la même famille. On parle de « *trans-mobilisation* ». Lorsque la transposase appartient à une autre famille d'ET, on parle de « *cross-mobilisation* » (Figure 4).

Plus récemment, Yang *et al.*, ont mis en évidence l'activité de transposition d'un élément MITE chez *Arabidopsis thaliana* : *mPing*. Trois sources de transposase ont été caractérisées : celles des éléments autonomes *Ping* et *Pong*, et celle dérivée de l'ADNc d'un transcript de *Ping* (Yang *et al.* 2007). D'après leurs résultats, la mobilisation de la transposase d'éléments autonomes permet bien la propagation de ces éléments de classe II non-autonomes.

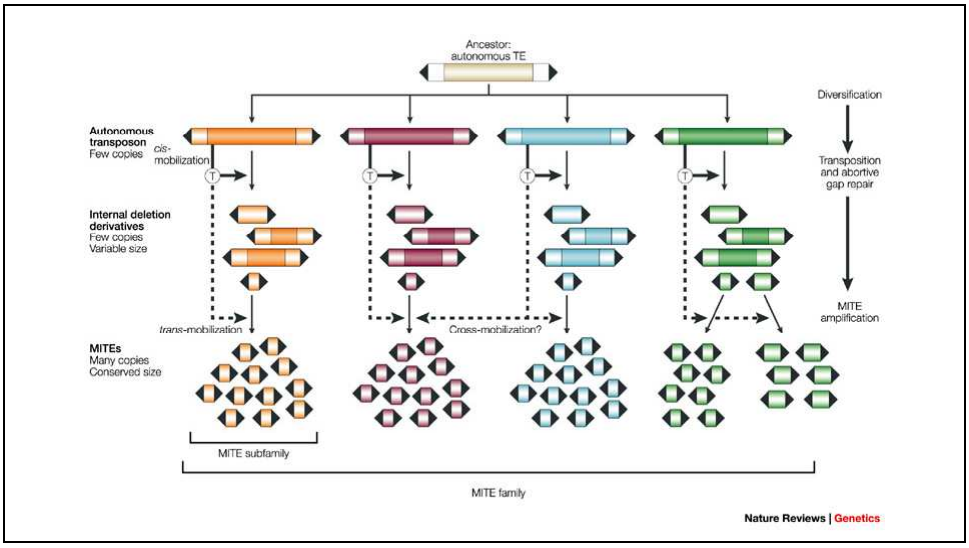


Figure 4. Modèle de formation des éléments MITE.

Ce modèle prédit dans un premier temps, l'amplification du nombre de copies de plusieurs familles de transposons autonomes provenant d'un même ancêtre commun. Les régions des TIR restant conservées entre les copies (malgré une forte divergence entre elles), des processus de transposition impliquant les copies non-autonomes d'une famille avec les transposases de la même famille (« *trans-mobilization* ») ou des autres familles (« *cross-mobilization* ») pourront avoir lieu (Feschotte *et al.*, 2002).

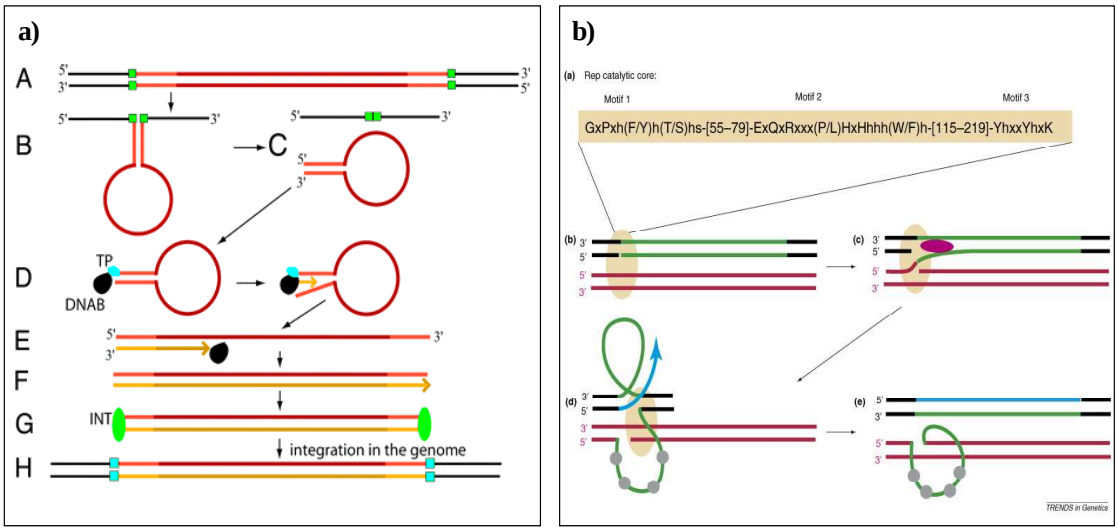


Figure 5. Modèles de formation des éléments de classe III.

a)-Modèle de formation des *Polintons* (Kapitonov et Jurka 2006). (**A**) l'intégrase induit l'excision simple brin de l'élément. (**B/C**) Le brin libre forme alors une structure en forme de boucle. (**D/E/F**) réplication du brin libre *via* la polymérase POLB du *Polinton*. (**G/H**) La séquence double-brin s'intègre à un site cible par un mécanisme identique aux transposons **b)**-Modèle de formation des *Helitrons via* un mécanisme de « *rolling-circle* » (Kapitonov et Jurka 2007). (**a/b**) Fixation de la protéine sur un motif sur la séquence donneuse et la receveuse. (**c**) Clivage puis synthèse d'ADN. (**d**) Le brin donneur sert ensuite de matrice pour la synthèse d'ADN du deuxième brin. (**e**) Le brin de la séquence donneuse est inséré au site receveur.

Transposition des Polintons.

De récentes études ont également permis de proposer un mécanisme de « couper-coller » comme mécanisme de propagation des *Polintons*. En effet, la présence de TIR et de TSD suggère un mécanisme proche de celui des transposons de classe II et de sous-classe 1. Cependant, les caractéristiques de ces séquences suggèrent certaines particularités. La polymérase de cet élément appartient à la famille des polymérases des bactériophages, des adénovirus et des plasmides linéaires.

Au vu de ces caractéristiques, Kapitonov et Jurka ont proposé en 2006, un modèle de formation des Polintons. Premièrement, durant la réplication, l'intégrase induit l'excision simple brin de l'élément. Le brin libre forme alors une structure en forme de boucle. Il y a deuxièmement réplication du brin libre *via* la polymérase POLB du *Polinton*. Puis en fin de synthèse, la séquence double-brin s'intègre à un site cible par un mécanisme identique aux transposons (Kapitonov et Jurka 2006) (Figure 5).

Transposition par « *rolling-circle* »

Les *Hélitrons* sont supposés transposer *via* un mécanisme de « *rolling circle* » (Kapitonov et Jurka 2007). Ce modèle de transposition est inspiré du monde bactérien. La protéine Rep (pour « *Replication initiator* ») reconnaît spécifiquement un motif conservé de la séquence d'ADN. La protéine se fixe sur ce motif, à la fois sur la séquence donneuse et la receveuse. Le clivage à ce site de la séquence donneuse, permet d'initier la synthèse d'ADN. En effet, la CDB de la séquence receveuse induit la formation d'un hétéroduplex par ligation d'une extrémité de la séquence donneuse avec une des extrémités de la séquence receveuse. Puis, le brin donneur sert de matrice pour la synthèse d'ADN du deuxième brin. En fin de synthèse, la séquence donneuse se compose du brin nouvellement synthétisé et du brin matrice. Au site receveur, le brin de la séquence donneuse est inséré mais ne génère pas de TSD (Figure 5).

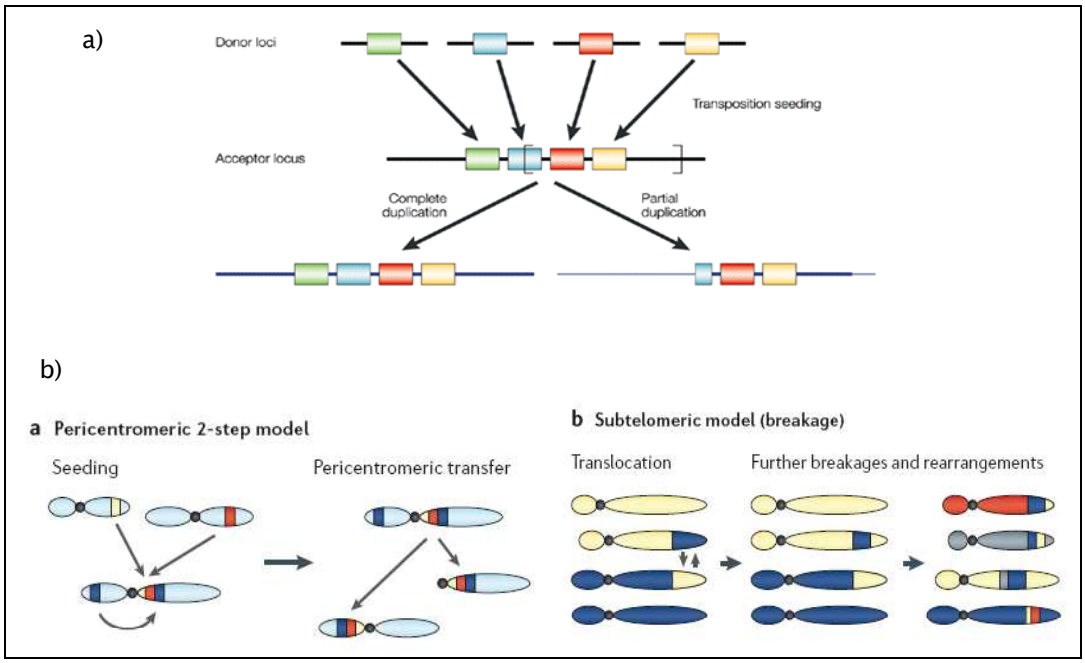


Figure 6. Modèles de formation des DS chez les mammifères.

a) Le modèle de Samonte et Eichler se déroule en deux étapes: (1) Transpositions de différents sites donneur vers un site accepteur; (2) Duplication partielle ou totale de la nouvelle région créée. **b)** ***a***-Bailey et Eichler ont également proposé des modèles pour expliquer la formation des duplications chez l'homme. Le modèle de formation des duplications péricentromériques est proche de celui de Samonte et Eichler. Lors de la première étape, les sites accepteurs des transpositions sont localisés dans les régions péricentromériques. ***b***-Pour expliquer la formation des duplications subtélomériques, une série de translocations peut conduire à la création de régions mosaïque dans les régions subtélomériques. Selon ce modèle, les DS créées dans ces régions sont contiguës et de même orientation (Samonte et Eichler 2002; Bailey et Eichler 2006).

1.2.2. Modèles de formation des DS

Après des analyses phylogénétiques et comparatives de DS chez les primates, Samonte et Eichler ont proposé un modèle de formation (Samonte et Eichler 2002). Ce modèle prédit un processus en deux étapes (Figure 6 a). La première correspond à une série de petites duplications ou transpositions. La duplication des régions de plusieurs sites donneurs vers un site accepteur conduit à la formation d'une séquence dite « mosaïque ». Celle-ci créée au site accepteur est ensuite dupliquée totalement ou partiellement. Ce modèle permet d'expliquer l'enrichissement en répétitions des DS détectées chez les primates. Mais il ne décrit pas le mécanisme moléculaire mis à l'œuvre.

Afin d'expliquer la formation des duplications péricentromériques chez l'homme, Bailey et Eichler proposent un modèle de formation, en deux étapes, inspiré du modèle de Samonte et Eichler (Figure 6 b.a) (Bailey et Eichler 2006). D'après ce modèle, la région mosaïque est crée dans une région péricentromérique. Cette séquence est ensuite dupliquée complètement ou partiellement au niveau d'une autre région péricentromérique.

Des copies de duplications de même orientation, également majoritairement interchromosomiques se retrouvent souvent localisées dans les régions subtélomériques. Bailey *et al.* ont proposé qu'une série de translocations suivie par des réarrangements pourrait conduire à la formation des duplications subtélomériques (Figure 6 b.b).

Les autres duplications détectées dans le génome humain sont dispersées le long des chromosomes avec une distribution non-aléatoire. En effet, les DS observées se retrouvent préférentiellement à proximité de leur séquence donneuse. Ce phénomène est appelé le « *duplication shadowing* ». D'après ce dernier, une région flanquée par une duplication a 10 fois plus de chance d'être impliquée dans une autre duplication qu'une autre région aléatoirement (Lander *et al.* 2001; Newman *et al.* 2005). Le mécanisme de NAHR pour « *Non-Allelic Homologous recombination* » a été suggéré suite à l'observation d'une forte densité en répétitions aux points de jonction des DS chez les eucaryotes, comme par exemple les éléments *Alu* qui représentent 24 % des extrémités des récentes DS, chez l'homme (Bailey et Eichler 2006). Chez la levure *Saccharomyces cerevisiae*, des rétroéléments à LTR ont été identifiés aux points de cassure de duplications (Koszul et al. 2004).

Les modèles de NAHR apparaissent comme de bons modèles pour expliquer la formation des DS chez les eucaryotes. La forte densité en DS dans les régions hétérochromatiques suggère un mécanisme en faveur d'insertions dans l'hétérochromatine.

1.2.3. Expansion des RT

Une courte séquence peut être dupliquée en tandem lors de la réplication. Le glissement de la polymérase permet d'expliquer la formation de ces répétitions. En effet, parfois l'ADN polymérase est bloquée par une structure secondaire de l'ADN et « bégaye » sur cette portion d'ADN: la polymérase revient en arrière de un à plusieurs nucléotides, relit la séquence et la recopie en créant une nouvelle unité du satellite sur le brin nouvellement créé (Ellegren 2004).

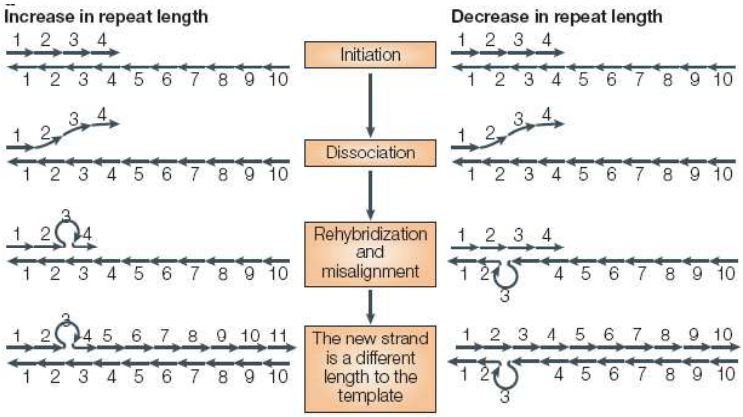


Figure 7. Processus d'expansion et de contraction des RT.

Le glissement de la polymérase conduit à l'expansion ou la contraction d'un «*cluster*» de RT (Ellegren 2004). Si le glissement se fait sur la séquence donneuse, le nombre de copies du nouveau «*cluster*» augmente (à gauche). Si le glissement se produit sur la séquence en cours de synthèse, le nombre de copies du nouveau «*cluster*» est moins important que celui de la séquence donneuse (à droite).

Pendant la réplication, le déplacement de la polymérase peut également induire la formation de boucles de contraction. Si cette boucle s'opère au niveau du brin en cours de synthèse, il y a alors expansion de la région par l'augmentation du nombre de copies. Par contre, si la contraction s'opère au niveau du brin matrice, le nombre de copies de répétitions diminue (Ellegren 2004). Au-delà d'un certain nombre de copies, la région satellite acquiert une dynamique plus importante. Plus il y a d'unités de répétitions et plus la variation du nombre d'unités devient importante.

Par conversion génique ou crossing-over inégal, une région satellite entière peut être dupliquée ailleurs dans le génome (Figure 8). Cette propagation peut se faire lors de la RH* d'une CDB d'ADN (Richard et Paques 2000). Le mécanisme de réparation peut également faire varier le nombre de copies de répétitions soit par glissement de la polymérase ou arrêt précoce de la synthèse (Figure 8).

Les mécanismes de formation des RT sont également responsables de l'évolution de ces répétitions. A l'opposé, les ET et les DS évoluent par des mécanismes différents de ceux de leurs formations.

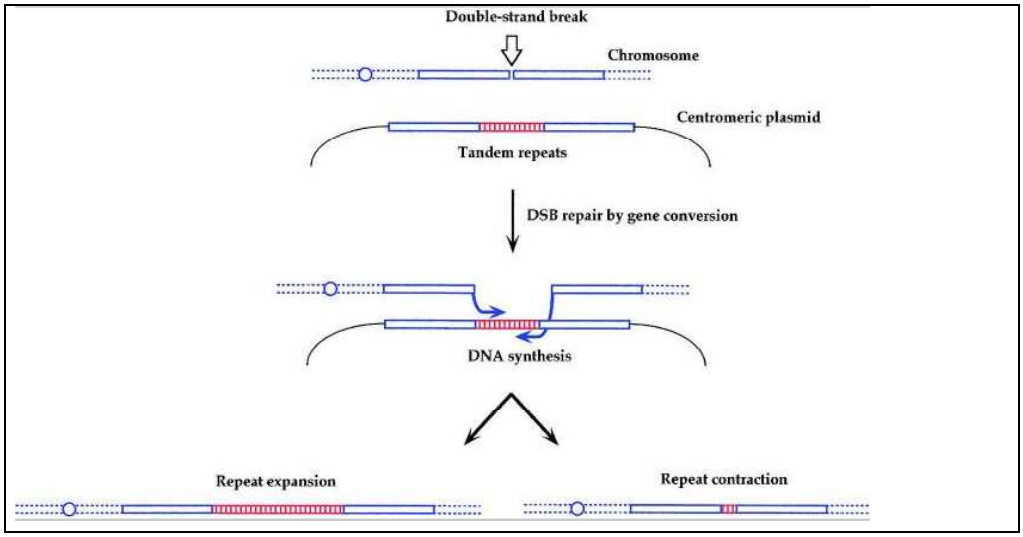


Figure 8. Mécanisme de formation des RT *via* la conversion génique.

Après une CDB d'ADN, un «*cluster*» de répétitions en tandem peut à l'aide d'une région homologue, conduire à la duplication du «*cluster*» de répétitions au site de cassure. Un arrêt précoce de la réparation peut conduire à une diminution importante du nombre d'unités (Richard et Paques 2000).

* Réparation homologue

1.3. Evolution et impact fonctionnel

1.3.1. Evolution des familles multigéniques

La dynamique des DS dépend de la composition en séquences des segments dupliqués. Les duplications géniques suivent trois grandes voies d'évolution: (1) la subfonctionnalisation, c'est-à-dire la spécialisation de fonction entre la copie donneuse et la receveuse; (2) la néofonctionnalisation, soit la création d'une nouvelle fonction pour la nouvelle copie; et la nonfonctionnalisation, c'est-à-dire l'élimination de la nouvelle copie. Même si cette dernière intervient le plus fréquemment, certaines copies de gènes persistent et donnent naissance aux familles multigéniques.

Depuis les années 1970, la détection et l'analyse détaillée de familles multigéniques a conduit à l'élaboration de différents modèles d'évolution des gènes. Le premier modèle d'évolution de gènes dupliqués a été appelé modèle de « *divergent evolution* » (Figure 9). Ce modèle a été proposé après l'analyse d'une famille bien connue: les *Globines*. La relation phylogénétique entre les protéines de *Globines* a permis de suggérer une divergence croissante entre les gènes. Or, les régions intergéniques des gènes dupliqués sont plus similaires dans une même espèce qu'entre deux espèces proches. Afin d'expliquer cette observation, un autre modèle a été proposé. Il a été nommé modèle d'évolution concertée (Figure 9) (Brown *et al.* 1984; Nei et Rooney 2005).

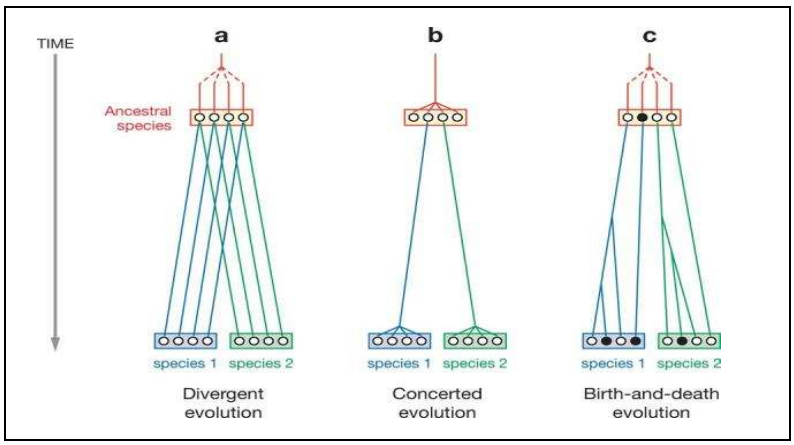


Figure 9. Les trois modèles d'évolution des familles multigéniques.

L'analyse des familles multigéniques a conduit à proposer trois grands modèles d'évolution: le modèle de « *divergent evolution* », le modèle d'évolution concertée et le modèle de « *birth-and-death* ». Les cercles blanc et noir correspondent respectivement aux gènes fonctionnels et aux pseudogènes (Nei et Rooney 2005).

D'après ce modèle, un ensemble de gènes évolue de manière concertée et non pas indépendamment les uns des autres dans un organisme donné. Les mutations d'un gène sont copiées chez les autres gènes par conversion génique. Ces mécanismes permettent ainsi d'homogénéiser les mutations entre les copies. Après spéciation, les deux groupes de gènes évoluent indépendamment. Ce modèle permet d'expliquer les observations faites sur la plupart des familles multigéniques tel que le « *cluster* » des gènes des ARN ribosomiques qui est l'exemple le plus représentatif de l'évolution concertée. Pourtant, ce modèle ne permet pas d'expliquer la dynamique évolutive d'autres familles multigéniques.

En effet, l'analyse des gènes du système immunitaires et du MHC (ou « *Major Histocompatibility Complex* ») a révélé des profils d'évolution assez différents. Pour expliquer les observations, un nouveau modèle appelé « *birth-and-death* » a donc été proposé par Nei et Rooney (Nei et Rooney 2005) (Figure 9). D'après ce modèle, certains des nouveaux gènes créés par duplication pourront se fixer dans le génome alors que la plupart seront inactivées et/ou délétés.

1.3.2. Rôle de duplications géniques

En modifiant une région codante ou régulatrice, les duplications de séquences peuvent être à l'origine de nouveautés génétiques. Ils jouent ainsi, dans certains cas, un rôle adaptatif. Le nombre et l'organisation des duplications géniques peuvent avoir des conséquences importantes pour l'organisme. Prenons par exemple le «*cluster*» des gènes des *Histones* chez la Drosophile. On distingue 4 classes de genes *Histone* en fonction de leur profil d'expression. A chaque stade de développement, seuls certains gènes sont exprimés. Il existe des gènes dits précoces et des gènes dits tardifs. L'expression des gènes est de plus également tissu-spécifique (Maxson *et al.* 1983). Dans de nombreux cas, les gènes dupliquées ont un rôle de sauvegarde: si l'expression d'un gène essentiel est éteinte, une autre copie de ce gène peut reprendre le relais.

1.3.3. L'impact fonctionnel des ET

De nombreuses études suggèrent que des introns seraient dérivés d'ET après la domestication de ces derniers (Giroux et al. 1994). Certains ET peuvent également être essentiels, voire vitaux pour la survie de l'hôte. Chez *A. thaliana*, le gène « *Daysleeper* » est essentiel à la croissance de la plante. La région codante de ce gène partage plusieurs caractéristiques avec la transposase des éléments de la superfamille des *hAT* (Bundock et Hooykaas 2005). Bundock et Hooykaas suggèrent qu'il y ait eu domestication de la transposase par l'hôte afin d'intégrer différentes fonctions cellulaires (Bundock et Hooykaas 2005). L'absence de cette protéine entraîne un développement anormal de la plante.

On soupçonne également les ET d'être à l'origine de la modification de la structure de la chromatine *via* la formation de petits fragments d'ARN abbérent.

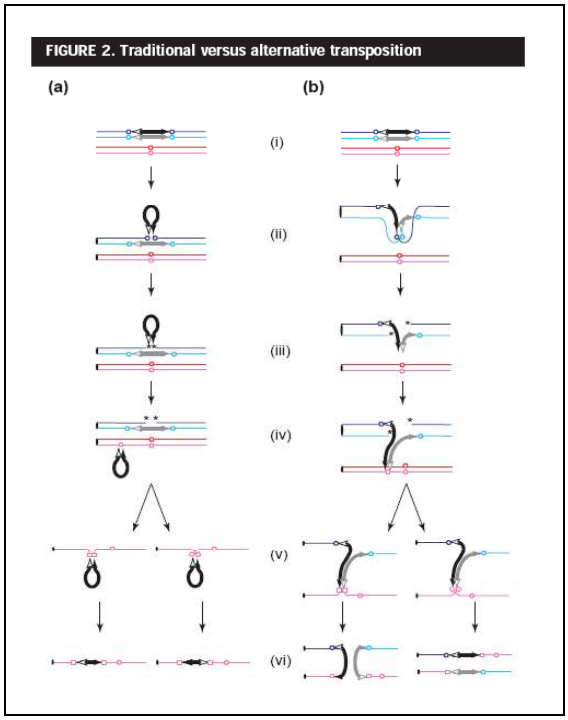


Figure 10. La transposition alternative.

(a) Mécanisme de transposition classique. L'appariement précédent l'étape d'excision se fait entre les deux TIR d'une seule copie. **(b)** Exemple de mécanisme de transposition alternatif. L'appariement se déroule entre deux TIR de deux copies distinctes (Gray 2000).

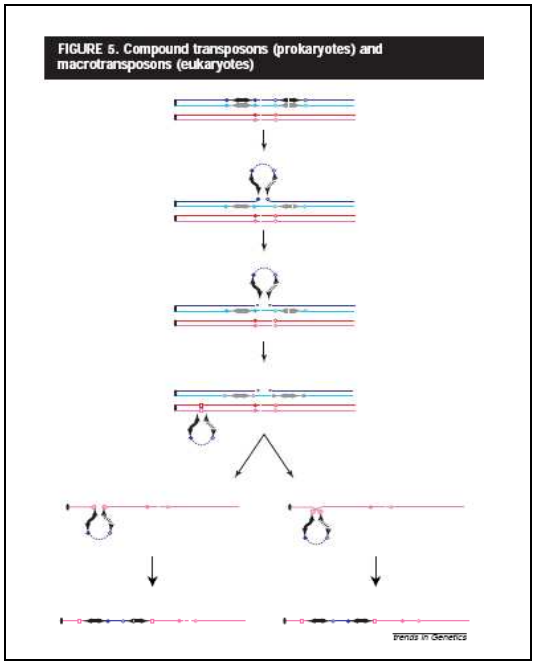


Figure 11. La macrotransposition.

Le mécanisme de transposition est similaire à celui de la transposition classique. Ce processus est induit par l'appariement de deux TIR de deux copies distinctes et contiguës. D'après ce processus, la région interne aux deux copies peut être dupliquée ou simplement déplacée (Gray 2000).

1.4. Des acteurs de la plasticité des génomes

1.4.1. Les transpositions alternatives

L'impact des répétitions sur la plasticité des génomes peut être direct ou indirect. On appelle « effet direct », l'impact dû au mécanisme de propagation des répétitions. L'insertion de répétitions est à lui seul un processus important d'augmentation de taille des génomes. Mais, durant le processus de transposition d'éléments de classe II, des erreurs de réparation des CDB peuvent conduire aussi à des transferts de brins et donc le déplacement ou la duplication de matériel génétique. Les ET sont aussi des acteurs important de la plasticité des génomes grâce aux mécanismes de transposition dits « alternatifs » qui peuvent conduire à des réarrangements de grande envergure (Figure 10). Parmi les processus de transposition alternatifs, on distingue, en particulier, le cas où deux TIR de deux copies différentes s'apparient. En fonction des molécules impliquées, l'excision conduit alors à différents types de réarrangements. Dans l'exemple présenté sur la Figure 10, l'excision peut conduire à la formation d'un chromosome acentrique et d'un chromosome dicentrique, ou à la formation de délétions et duplications. Ce phénomène a déjà été observé chez de nombreux organismes tel que la bactérie (avec les ET IS*10*/Tn*10*), le maïs (avec les ET *Ac/Ds*), le tabac (avec les ET *Ac/Ds*), et la Drosophile (avec l'*élément P*) (Preston *et al.* 1996). La plupart des réarrangements identifiés comme induits par les ET correspondent à des délétions, des duplications ou des inversions. Ce biais dans l'observation des réarrangements peut être dû à la faible viabilité des organismes ayant subit les autres types de réarrangements.

1.4.2. La macrotransposition

La grande différence entre la transposition classique et la transposition alternative est le choix des TIR impliqués dans le mécanisme de transposition. Ce dernier peut également choisir deux TIR de deux copies contiguës séparées de plusieurs kb. La transposition conduit alors à la duplication d'un grand fragment génomique. Ce mécanisme est appelé: « macrotransposition » (Figure 11). Un tel mécanisme a été identifié chez les procaryotes et les eucaryotes. Par exemple, l'élément Tn*10* s'est formé grâce à deux copies d'un élément IS*10*. Des évènements similaires ont été montrés chez les plantes pour l'élément *Ac/Ds* (Kunze 1996) et chez la Drosophile entre deux copies de l'élément *mariner* (Gray 2000).

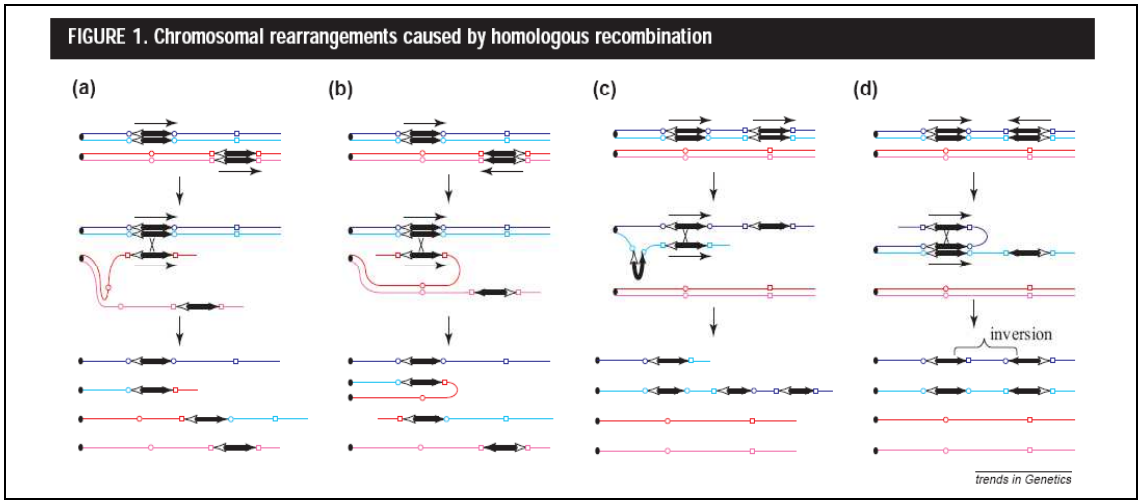


Figure 12. Les mécanismes évolutifs des répétitions par recombinaison homologue.

En fonction de l'orientation et la localisation des répétitions, la recombinaison homologue ne conduit pas au même résultat. **(a)** Si les copies en sens direct sont localisées sur deux chromosomes différents, la recombinaison conduit à une translocation. **(b)** Si les deux copies sont en orientation inverse, la recombinaison conduit à la formation d'un chromosome dicentrique et d'un chromosome acentrique. **(c)** La recombinaison entre deux copies présentes sur le même chromosome en sens direct conduit à une duplication et/ou une délétion. **(d)** Quand les deux copies sont sur le même chromosome mais inversées, la recombinaison mène à l'inversion de la région comprise en ces deux répétitions (Gray 2000).

1.4.3. La recombinaison entre copies de répétitions

On appelle « effet indirect », les réarrangements par RH* ectopique ou NAHR dus à la présence des répétitions dans les génomes. En effet, l'abondance de copies de séquences similaires dans un génome peut conduire à des réarrangements chromosomiques importants, et ainsi à la réorganisation de celui-ci. Les évènements induits peuvent conduire à des inversions, des délétions et également des duplications (Figure 12) (Gray 2000; Bailey et Eichler 2006). Le produit de ces réarrangements dépend de la localisation et l'orientation des copies. Si les copies sont en sens direct et localisées sur deux chromosomes différents, la recombinaison conduit à une translocation. Si les deux copies sont en orientation inverse, la recombinaison mène à la formation d'un chromosome dicentrique et d'un chromosome acentrique. Si les deux copies sont localisées sur le même chromosome, la recombinaison entre deux répétitions directes conduit à une duplication et/ou une délétion. Quand les deux copies sont inversées, la recombinaison conduit à l'inversion de la région comprise en ces deux répétitions (Figure 12).

1.4.4. La formation des solo-LTR

La recombinaison homologue non-allélique semble plus fréquente quand une répétition est à proximité de la cassure. Or, durant la transposition de rétroéléments à LTR, les séquences des deux LTR sont rendues identiques. On observe alors souvent la recombinaison entre le LTR 5' et le LTR 3' d'un même élément. Cette recombinaison conduit à ce qu'on appelle un solo-LTR. La relation entre la formation des solo-LTR et la recombinaison homologue ectopique a été montrée chez la levure (Roeder et Fink 1980). On peut s'attendre à observer dans les régions hautement recombinogènes une plus forte densité en solo-LTR en comparaison avec les LTR complets. Les rétroéléments à LTR, pouvant atteindre plusieurs kb et hautement recombinogènes, apparaissent comme des acteurs important de la variation de taille des génomes. Cette relation entre taille des génomes et rétroéléments à LTR a été étudiée chez les plantes (Shirasu *et al.* 2000; Devos *et al.* 2002; Vitte *et al.* 2007).

* Réparation homologue

1.5. Problématique

1.5.1. L'évolution de la taille des génomes

Durant ces dix dernières années, deux grandes idées de l'évolution de la taille des génomes s'affrontent. La première penche pour une dynamique constante, ce qui conduit à une augmentation continue de la taille des génomes. La deuxième propose un maintien de la taille des génomes, impliquant la régulation de celle-ci.

Même si de nombreuses copies sont éliminées juste après leur insertion, certaines se fixent dans la population, ce qui conduit à une augmentation constante de la taille des génomes. En absence de sélection, l'augmentation de la taille des génomes *via* l'insertion d'ET, de grandes duplications ou par la polyploidisation, affecte la croissance et l'encombrement spatial des chromosomes dans le noyau. Bennetzen et Kellogg, vont jusqu'à parler de la tendance à l'obésité des génomes (Bennetzen et Kellogg 1997). Mais cet état induit une limite d'accès des protéines aux séquences génomiques et donc à la mort cellulaire. On peut penser que la sélection retarde ce processus sur plusieurs millions d'années. Durant ce laps de temps, certaines insertions peuvent alors avoir un impact évolutif impliquant une adaptation structurale de l'organisme à l'augmentation de la taille de son génome. Alexander Vinogradov a proposé l'absence de contraintes adaptative due à la polyploïdisation. Son modèle prédit une augmentation constante de la taille des génomes (Vinogradov 2003; 2004).

Nous pouvons également supposer une forte régulation des répétitions pour un maintien de la taille du génome vers une taille minimale nécessaire pour la cohésion de l'information génétique dans d'un génome. Dans ce cas, la taille actuelle d'un génome serait donc proche de la taille minimale. Cette hypothèse repose sur le fait que plus la taille augmente et plus l'activité cellulaire est réduite. La sélection va donc tendre à équilibrer la variation de taille.

Deux modèles ont été proposés pour expliquer ce contrôle par la sélection du nombre de copies des répétitions. Le premier modèle est appelé : le « modèle de rupture de gènes ». L'insertion de répétitions est ici contre-sélectionnée au niveau des gènes ou des régions régulatrices (Nuzhdin 1999). Ce modèle prédit une faible densité en répétitions dans les régions riches en gènes. Une étude réalisée chez *A. thaliana* a montré une corrélation négative entre la densité en gènes et celle en ET (Wright et al. 2003).

L'autre modèle, dit « modèle de recombinaison ectopique », propose une sélection contre les effets délétères de la recombinaison ectopique à cause des remaniements qu'elle provoque. Ce modèle prédit une accumulation de répétitions dans les régions à faible de taux de recombinaison, car peu éliminées. Ce modèle prédit également une sélection plus forte contre les copies de grandes tailles et contre les copies de familles hautement répétées puisque plus sujettes à recombiner (Charlesworth *et al.* 1986; Montgomery *et al.* 1987; Dray et Gloor 1997; Petrov *et al.* 2003).

L'implication directe les protéines PIWI dans la biogénèse de petits ARN: les « piRNA » a récemment été montrée chez la Drosophile et le Zebrafish. Ces petits ARN sont reponsables du « *silencinsg* » d'ET dans les lignées germinales. Par conséquent, ces séquences sont également impliquées dans le maintien de copies d'ET (Brennecke *et al.* 2007; O'Donnell et Boeke 2007). L'implication de ces piRNA a été montrée chez plusieurs espèces eucaryotes telles que la souris. Ces strucutres agissent également directement sur l'activité de transposition des ET (Carmell *et al.* 2007).

De récentes études réalisées chez les plantes et la Drosophile, ont cherché à confronter ces modèles. Ces études ont montré que la recombinaison ectopique semble être avec les petites délétions, des forces majeures permettant de limiter la taille des répétitions et donc des génomes (Wicker *et al.* 2003; Bennetzen *et al.* 2005; Vitte *et al.* 2007; Wicker *et al.* 2007). En effet, un grand nombre de rétroéléments à LTR sont fortement délétés ou sous forme de solo-LTR (Shirasu *et al.* 2000; Devos *et al.* 2002; Vitte *et al.* 2007). Malheureusement, ces études n'ont pas permis d'éclaircir le modèle d'évolution de la taille des génomes. On ne sait toujours pas si la variation de taille est principalement due à des variations du rapport du taux d'insertion/efficacité d'élimination. D'autres auteurs ont discuté de la variation de taille des génomes à la lumière de leurs caractéristiques génomiques et environnementales, mais aucune corrélation n'a pu être mise en évidence (Vinogradov 2003; Bennetzen et al. 2005).

1.5.2. Le plan de travail

Mon sujet de thèse s'inscrit dans l'étude de la plasticité des génomes eucaryotes. Le but de mon travail a été de déterminer et de discuter de l'organisation et de l'impact des répétitions sur la plasticité des génomes.

Dans un premier temps, pour annoter les répétitions, j'ai développé un pipeline informatique de détection des DS et un autre pour la détection des RT. Ces deux pipelines, utilisés pour les génomes de *Drosophila melanogaster* et de *Arabidopsis thaliana* seront présentés dans le chapitre 2.

Deuxièmement, afin de mieux comprendre la dynamique d'insertion des répétitions, j'ai analysé les DS détectées chez *D. melanogaster* et proposé un modèle pour expliquer leur formation. J'ai ensuite cherché à montrer l'impact des ET dans ce processus. L'ensemble de ce travail sera exposé dans le chapitre 3.

Dans la plupart des génomes eucaryotes séquencés et annotés, les répétitions montrent une densité plus élevée dans l'hétérochromatine. Un déséquilibre entre insertion et élimination est-il suffisant pour expliquer la forte densité en répétitions dans les régions hétérochromatiques? L'insertion des répétitions est-elle plus tolérée dans ces régions?

Troisièmement, pour répondre à ces questions, j'ai comparé la dynamique des répétitions de l'euchromatine avec celles de l'hétérochromatine chez *A. thaliana*. Pour cela, j'ai retracé l'histoire évolutive des répétitions et estimer les forces qui tendent à les éliminer. J'ai pu ainsi discuter de la relation entre la présence de répétitions et la structure de la chromatine. Ce travail vous sera présenté dans le chapitre 4.

Enfin, le cinquième chapitre est une discussion générale sur la plasticité des génomes qui reprend les différents résultats obtenus durant ma thèse.

.

2. Les méthodes de détection des répétitions

Détecter et annoter les répétitions sont les étapes initiales pour l'étude de leur dynamique. Dans cette optique, j'ai développé, durant ma thèse, des outils de détection des DS et des RT. Je vous présenterai dans ce chapitre les méthodes utilisées pour cette tâche.

2.1. Deux approches de détection des répétitions

2.1.1. Les méthodes de novo

Les méthodes *de novo* de détection cherchent à identifier les répétitions sans aucune connaissance *a priori* sur celles-ci. La stratégie la plus courante consiste à détecter dans un premier temps l'ensemble des répétitions d'un génome. Puis, les copies sont regroupées en familles de répétitions. Une séquence consensus pourra être ensuite créée pour chacune des familles. Le jeu de données peut contenir différents types de répétitions tel que les ET, les RT et les DS. On peut les trier selon leur nombre de copies. En effet, les ET sont généralement plus répétés que les DS. Mais, la détection de copies très divergentes s'avère plus problématique *via* ces méthodes. En effet, la plupart des copies d'ET sont très dégénérées. Différents degrés de similarité de séquence parmi les copies peuvent conduire à la formation de « familles » composées de plusieurs sous-familles. L'emboîtement des copies représente également un obstacle majeur pour leur identification. En effet, l'imbrication des copies les unes dans les autres induit la formation de « meta-familles », c'est-à-dire des grandes familles composées d'un grand nombre de copies de répétitions de différentes familles emboîtées les unes dans les autres.

2.1.2. Les méthodes avec connaissance a priori

A l'opposé de la méthode *de novo*, la méthode de détection des répétitions avec connaissance *a priori*, consiste à utiliser une banque de séquences de référence de répétitions pour détecter par alignement des copies et/ou de nouvelles familles. La banque de données la plus utilisée se nomme « *Repbase Update* » (RU). Cette banque de données contient des séquences répétées présentes dans différentes espèces eucaryotes. Elle a été développée en 1990 sous la direction de Jerzy Jurka, au

GIRI « *Genetic Information Research Institute* » (Jurka 2000). La plupart des séquences sont des séquences consensus de familles de répétitions. L'avantage de cette méthode est sans nul doute sa spécificité et sa sensibilité. Cependant, les résultats de la détection dépendent de la qualité des séquences de référence utilisées. La construction des séquences consensus de référence représente donc une étape cruciale pour cette approche.

2.1.3. REPET : un pipeline d'annotation des ET

La détection peut également se baser sur le biais de composition nucléotidique des ET. En effet, de nombreuses études ont suggéré un enrichissement des ET en bases A/T (Shields et Sharp 1989; Lerat *et al.* 2000). Cette approche consiste à utiliser des programmes basés sur les modèles de Markov cachés (ou HMM pour « *Hidden Markov Models* »). A cet effet, une approche basée sur cette méthode, à laquelle j'ai contribué, a été développée au laboratoire: TE-HMM (Andrieu et al. 2004) (6.1.Article 1: «Detection of transposable elements by their compositional bias»).

Un pipeline d'annotation des ET, combinant différents programmes dont le TE-HMM, a ensuite été développé (Quesneville et al. 2005). Celui-ci a été utilisé pour l'annotation des ET euchromatiques (Release 4) de *D. melanogaster*. Les annotations obtenues ont été incorporées dans la base de données d'annotations des Drosophilidés: « *FlyBase* » (http://flybase.bio.indiana.edu). Ce pipeline est aujourd'hui utilisé pour l'annotation officielle des ET de nombreux génomes eucaryotes dont celui d'*A. thaliana*. J'ai donc utilisé les résultats de ce pipeline comme annotation des ET.

2.2. SegDupPipeline: un pipeline de détection des DS

La détection des DS soulève certains problèmes: (i) les duplications peuvent se composer d'autres répétitions, en conséquence de quoi, il s'avère difficile de les différencier des autres répétitions; (ii) Des évènements de larges insertions et délétions qui ont eu lieu après la duplication, peuvent perturber la détection des grandes duplications. J'ai donc développé un pipeline de détection des DS en gardant à l'esprit ces différents obstacles. Les données sont stockées dans une base de données MySQL, afin d'optimiser la recherche et le transfère des informations (Fiston-Lavier *et al.* 2007) (6.2.Article 2: « A model of segmental duplication formation in Drosophila melanogaster »).

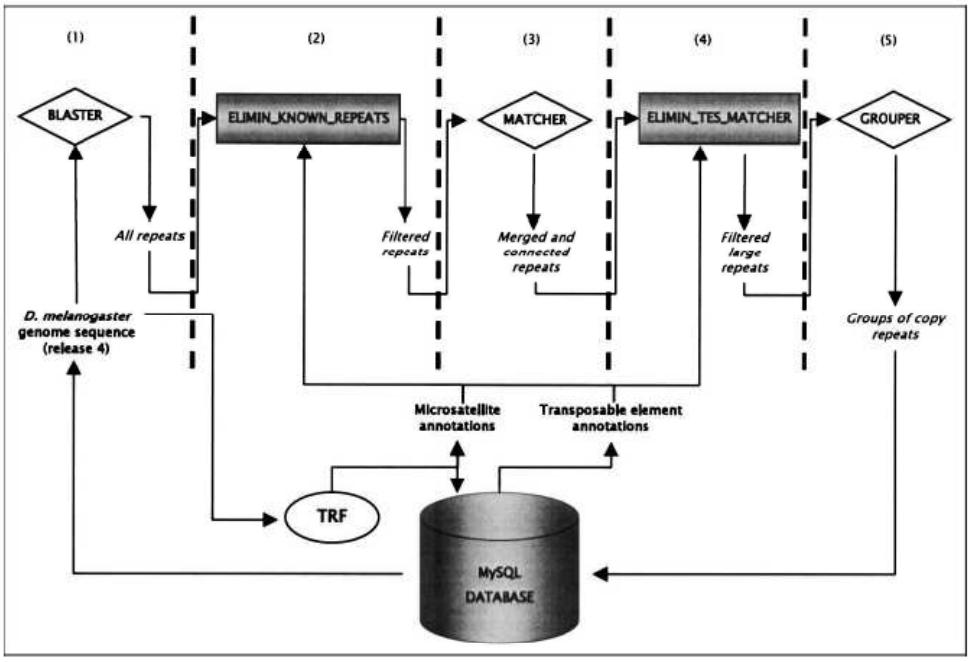


Figure 13. Le pipeline de détection des DS.

(1) Détection par BLAST de l'ensemble des répétitions d'un génome *via* BLASTER. (2) Première étape d'élimination des ET et microsatellites. (3) Etape de connexion des fragments contigus et fusion des fragments chevauchants *via* MATCHER. (4) Deuxième étape d'élimination des ET à la structure plus complexe. (5) Regroupement par famille de duplication *via* GROUPER (Fiston-Lavier *et al.* 2007).

2.2.1. Etape n°1: Détecter l'ensemble des répétitions d'un génome

La première étape du pipeline a pour but de détecter toutes les répétitions d'un génome (Figure 13). Le programme BLASTER implémenté au laboratoire (Quesneville *et al.* 2003; 2005), utilise le programme BLASTN du NCBI afin de comparer un génome contre lui-même (Altschul *et al.* 1990; 1997; Quesneville *et al.* 2003; 2005). Il réalise une comparaison entre une banque de séquences « requêtes » et une banque de séquences « sujettes ». Les séquences « requêtes » et « sujettes » sont tout d'abord découpées en fragments chevauchants. Puis, chaque fragment est aligné avec la banque « sujette » par BLAST (ici BLASTN). Les résultats du BLAST sont des alignements locaux (ou HSP pour « *High Scoring Pair* ») qui seront ensuite traités. Certains sont filtrés en fonction de critères paramétrables tels que le score, la e-value, la longueur de l'alignement, le pourcentage d'identité. Enfin, les HSP chevauchants sont fusionnés. N'étant pas limité dans la taille des séquences des banques, BLASTER peut traiter des génomes entiers. Lorsque les deux banques sont identiques, BLAST permet de détecter les répétitions de cette banque. En conséquence de quoi, il permet de détecter l'ensemble des répétitions d'un génome. Les données de sortie de BLASTER peuvent ensuite être traitées par les programmes MATCHER pour l'étape d'annotation, ou GROUPER pour regrouper les copies de répétitions en famille (Quesneville *et al.* 2003; 2005).

2.2.2. Etape n°2: Eliminer du jeu de données les autres répétitions

Il faut ensuite distinguer et éliminer de notre jeu de répétitions toutes les répétitions annotées comme ET et microsatellites. J'ai implémenté le programme ELIMIN_KNOWN_REPEATS qui traite l'ensemble des répétitions d'un génome détectées par BLASTER (Figure 13). Il interroge une base de données d'annotations des répétitions afin d'éliminer, de l'ensemble des répétitions détectées, les couples de répétitions -HSP- pour lesquels au moins l'une des deux répétitions est complètement incluse dans une répétition connue: ET, satellites ou microsatellites (Figure 14). L'une des difficultés du processus d'annotation des répétitions est de définir précisement les bornes de la séquence. Afin de réduire le risque de garder certains faux-positifs, l'utilisateur est libre de choisir une valeur d'extension des extrémités des répétitions annotées. La valeur d'extension doit tenir compte de la confiance attribuée dans les annotations utilisées, au risque d'éliminer certains vrais-positifs. Cette valeur a été fixée à 5 pb (valeur par défaut).

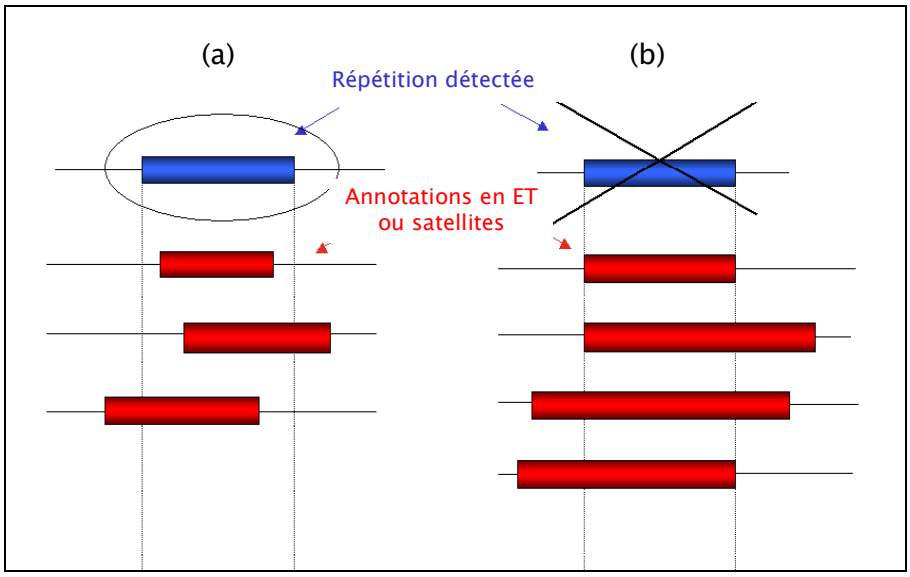


Figure 14. Etape d'élimination des ET et des satellites.

(a) La répétition détectée sera conservée si elle ne « *matche* » pas entièrement avec une annotation d'ET ou satellites. **(b)** La répétition détectée est éliminée car elle est complètement incluse dans une annotation d'ET ou satellites.

2.2.3. Etape n°3: Détection de duplications « segmentaires »

A cette étape, le jeu de données se compose de répétitions ne correspondant ni à des ET, ni à satellites ou microsatellites. Or, après duplication, une séquence peut subir des évènements d'insertion et/ou délétion. Une étape de connexion ou défragmentation des HSP est donc nécessaire. Cette étape s'avère cruciale car, elle permet de rassembler deux fragments de répétition d'une même DS.

Cette étape est réalisée *via* le programme MATCHER (Quesneville *et al.* 2003; 2005). MATCHER cherche ici à connecter des HSP contigus. Le programme évalue les connexions entre deux HSP par un score faisant intervenir le score d'alignement des deux HSP et une pénalité de « *gap* » et de « *mismatch* » dépendant de la distance entre les deux HSP. MATCHER recherche une chaîne optimale de connexion (de score maximal positif) des HSP, par un algorithme de programmation dynamique. Les HSP trouvés dans la chaîne sont ensuite retirés et une nouvelle recherche est effectuée, ce qui permet itérativement d'obtenir toutes les chaînes d'HSP possibles. A l'issue de cette étape, la duplication est détectée entièrement.

2.2.4. Etape n°4: Eliminer les évènements de transposition

Malgré la première étape de filtre, après la connexion des fragments contigus, des faux-positifs peuvent persister: le jeu de données peut encore contenir des copies d'ET. Entre autres, les copies bornées par des microsatellites ne sont pas éliminées par le premier filtre (Figure 15). En effet, pour les éléments type SINE et LINE, la variation de tailles de la queue poly-A entre copies peut permettre à certaines copies d'échapper au filtre. Afin d'éliminer ces faux-positifs, j'ai implémenté le programme ELIMIN_TES_MATCHER qui détecte et élimine les larges répétitions d'un ET borné par des microsatellites (Figure 15). Même si certaines copies d'ET ont pu être générées par duplication et pas par transposition, obtenir un jeu de données propre, permettra de ne pas biaiser les analyses. Le programme ELIMIN_TES_MATCHER élimine les répétitions qui ont plus de 99 % de recouvrement en ET ne correspondant qu'à un seul élément. Pour les répétitions ayant entre 95 % et 99 % de recouvrement en ET avec un seul élément, le programme élimine les répétitions bornées par des microsatellites ou bornées par moins de 20 pb de séquence non-ET et non-satellites (valeur par défaut).

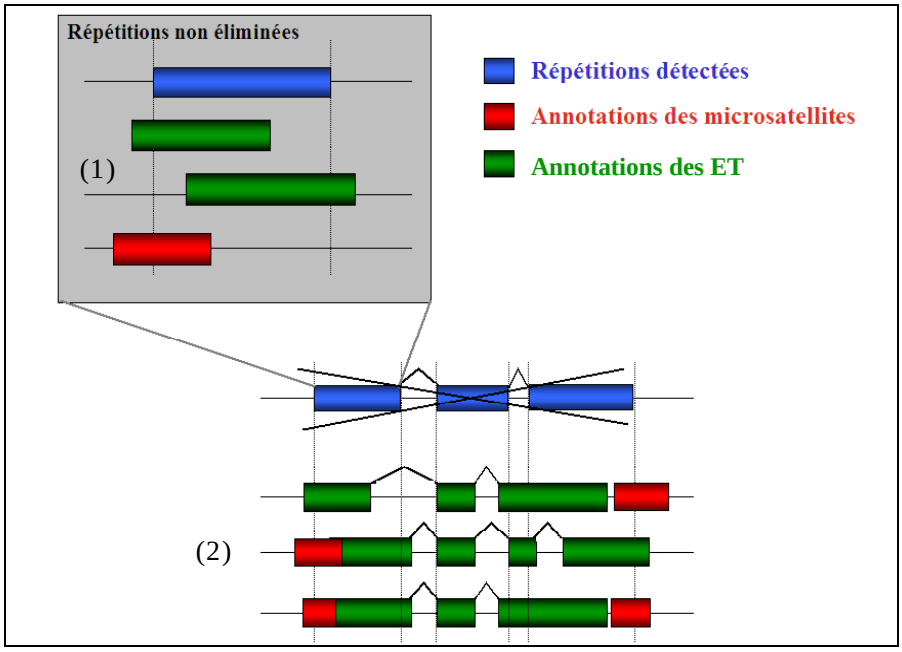


Figure 15. Deuxième étape d'élimination des ET.

Une répétition peut ne pas avoir été éliminée lors du premier filtre. Pourtant, cette répétition peut appartenir à une annotation d'ET. Si l'on considère après la connexion la partie recouverte par l'ensemble, on pourra observer deux cas de figure: **(1)** la répétition correspond entièrement à un ET, ce qui induit son l'élimination dès le premier filtre; **(2)** la répétition correspond à une région d'un ET recouvert à plus de 99% ou bornée par des microsatellites.

2.2.5. Etape n°5: Construction des familles de duplication

La dernière étape du pipeline consiste à regrouper toutes les copies d'une même duplication. Pour cela, le pipeline utilise le programme GROUPER, également développé au laboratoire (Quesneville *et al.* 2003; 2005). Il permet de reconstruire les familles d'ET. Ce programme traite les sorties de BLASTER en réalisant des groupes de séquences similaires. Pour ce faire, il commence par connecter les répétitions contiguës par programmation dynamique. De cette manière, comme le programme MATCHER, il tente de reconstruire les éléments morcelés. Puis, *via* un algorithme de « *clustering* » simple lien, il regroupe les séquences partageant des régions similaires. A l'issu de cette étape, on obtient des groupes correspondant à des copies similaires.

2.3. Comment détecter les RT ?

La difficulté de la détection des RT est de délimiter précisément les unités de répétitions ainsi que définir les limites des «*clusters*» de répétitions, malgré la présence d'inserts ou de grandes délétions chevauchant plusieurs unités. Pour leur détection, nous avions à notre disposition plusieurs programmes, dont certains développés au laboratoire (BLASTER et MATCHER, déjà présentés plus haut). D'autres programmes tel que Piler, ont été développés spécifiquement pour la détection des RT (Edgar et Myers 2005). Après la présentation de ces programmes, je discuterai des limites des différentes méthodes. Une approche combinant ces programmes s'est avérée plus sensible.

2.3.1. Pals et Piler-TA, des programmes spécifiques

Pals réalise des alignements locaux à partir de séquences pouvant faire quelques centaines de bases. Les hits détectés pourront être traités par le programme Piler (Edgar et Myers 2005). Le programme Pals recherche tous les hits ayant une taille minimale et une identité de séquence minimum donnée. Ces valeurs sont par défaut de 400 pb et 94 % d'identité. Pals fournit ainsi au programme Piler des répétitions bien conservées et pauvres en gaps. Ces séquences peuvent ensuite, à l'aide du programme Piler-TA, être regroupées par «*cluster*» de RT. Chaque «*cluster*» peut être visualisé comme une pyramide en « *DotPlot* » (Figure 16).

Cette méthode ne tient pas compte des insertions qui peuvent se produire au sein du «*cluster*». Par exemple, au niveau du «*cluster*» des gènes des *Histones* chez *D. melanogaster*, un élément *roo* s'est inséré. L'insertion de cette séquence perturbe la détection du «*cluster*» (Figure 17).

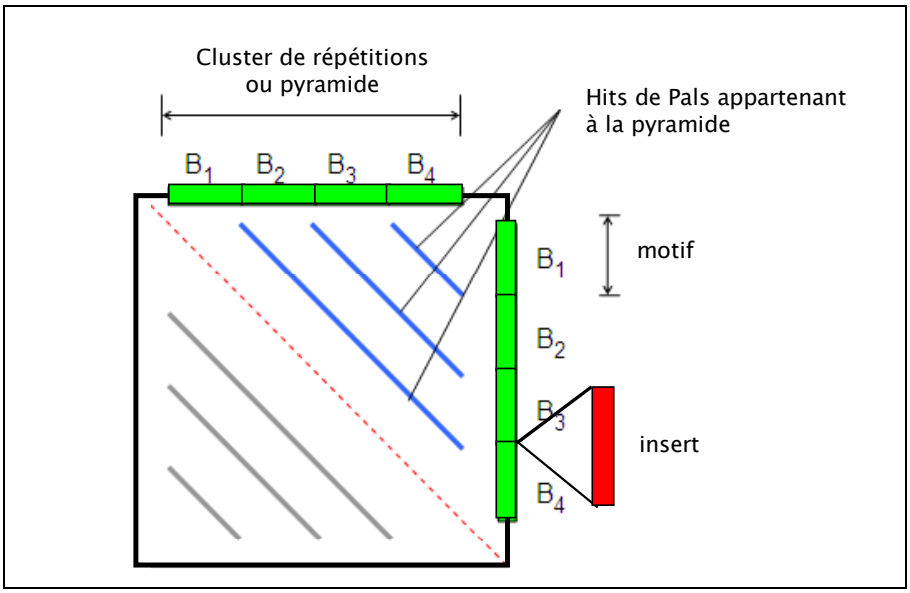


Figure 16. Détection des RT par Pals/Piler-TA.

Schéma du « *DotPlot* » d'une région de RT détectée par Pals/Piler-TA. Chaque unité de répétition (en vert) « *matche* » avec les autres unités du «*cluster*», ainsi qu'avec des régions composées de plusieurs motifs strictement en tandem. Une pyramide se compose de hits imbriqués les uns dans les autres. La base de la pyramide correspond au «*cluster*» de RT. L'insertion d'une région (en rouge) peut perturber la détection d'une pyramide.

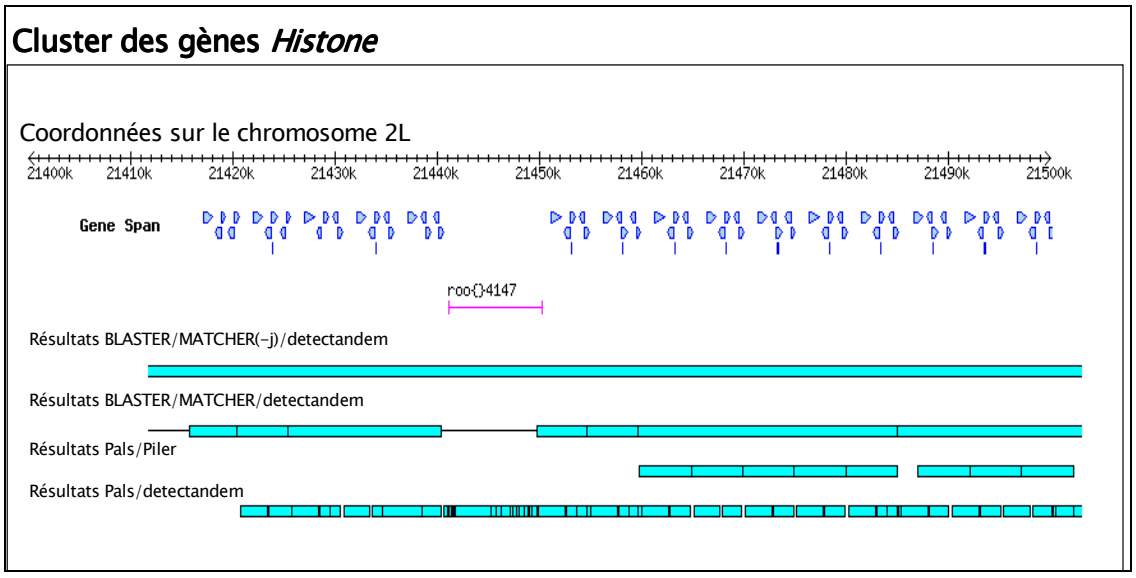


Figure 17. Détection du «*cluster*» des gènes *Histone*.

Visualisation *via* Gbrowse du «*cluster*» des gènes *Histone* chez *D. melanogaster*. Ce « cluster » est localisé sur le bras chromosomique 2L de *D. melanogaster*. Les annotations des gènes correspondent aux triangles bleus. Le trait rose indique une copie de l'élément *roo*. Celle-ci semble s'être insérée dans le «*cluster*». Les annotations en bleu turquoise correspondent aux différents résultats de détection obtenus par les différentes combinaisons des approches présentées: BLASTER/MATCHER/Detectandem.py et Pals/Piler. Alors que l'insertion est identifiée par la première approche, l'approche Pals/Piler limitée aux répétitions strictement en tandem, n'identifie pas tout le «*cluster*». Afin de montrer la puissance de l'approche Pals/Piler pour la détection des unités de répétitions, j'ai également représenté le résultat de Pals/Detectandem.py.

2.3.2. Une méthode de détection

L'exécution des programmes BLASTER et MATCHER pour une séquence génomique contre elle-même, permet d'identifier l'ensemble des répétitions d'un génome. Le programme MATCHER doit être utilisé sans l'option « -j », qui connecte les fragments contigus. Afin de détecter les RT parmi l'ensemble des répétitions, j'ai implémenté un programme: « Detectandem.py ». Pour chaque hit, ce programme calcule la distance minimale entre les répétitions contiguës. Si cette distance est inférieure à 10 pb, le programme connecte les régions. Malheureusement, cette approche ne permet ni d'obtenir toutes les RT, ni d'identifier précisément les unités répétées. Mais, cette approche a l'avantage de pouvoir gérer les insertions et les délétions dans ces séquences répétées (Figure 17). Afin de pouvoir détecter précisément l'unité de répétition et de détecter entièrement le «*cluster*» de répétitions sans être perturbé par d'éventuelles insertions ou délétions, nous avons combiné les deux approches présentées.

2.3.3. TandemPipeline: un pipeline de détection des RT

La détection de l'ensemble des RT est effectuée *via* l'approche Pals/Piler-TA. Pour chaque «*cluster*», les séquences des unités de répétitions sont extraites. Une séquence consensus peut être alors créée après alignement multiple des unités. L'exécution du programme BLASTN *via* BLASTER avec les séquences consensus contre les séquences génomiques permet d'identifier les unités de RT, même très divergentes. Les programmes MATCHER et « Detectandem.py » permettent enfin de délimiter les bornes du «*cluster*» (Figure 16).

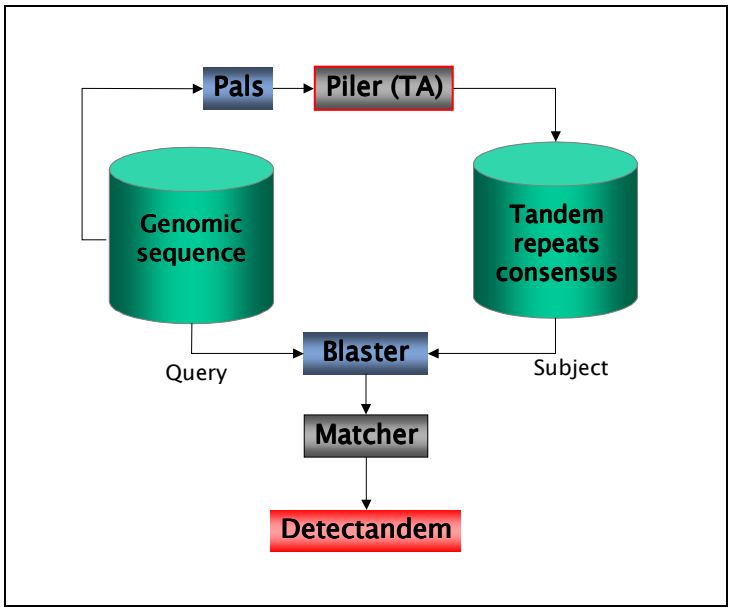


Figure 18. Le pipeline de détection des RT.

En exécutant les programmes Pals/Piler-TA sur un génome, on détecte les unités bien conservées et non-interrompues des RT. Pour chaque « *cluster* », la création d'une séquence consensus, va permettre de rechercher les autres unités de répétition plus divergentes et plus éloignées à l'aide des programmes BLASTER/MATCHER/Detectandem.py.

2.4. Les applications

La qualité des résultats de ces pipelines dépend de la qualité des annotations disponibles, ainsi que de la qualité des alignements. Le pipeline de détection des DS a été utilisé pour les génomes de *D. melanogaster* et *A. thaliana*. Pour ces deux génomes, les annotations en ET sont disponibles. Elles ont été réalisées au laboratoire. Ces annotations correspondent aux annotations officielles (http://flybase.net/; http://www.arabidopsis.org/). Pour *D. melanogaster*, la version 4 de la séquence génomique avec ses annotations a été utilisée.

Les deux pipelines de détection ont été utilisés sur la version 6 de la séquence et de l'annotation de *A. thaliana*. Le pipeline de détection des RT a été testé à l'aide des séquences en tandem de *D. melanogaster*.

3. La dynamique des DS chez *Drosophila melanogaster*

Les DS apparaissent aujourd'hui comme l'un des éléments moteurs de l'évolution des génomes de part leur taille et leur abondance. Pourtant, à l'heure actuelle, leur mécanisme de formation reste encore obscur. Plusieurs modèles de formation des DS ont déjà été proposés chez les eucaryotes. Certains suggèrent des modèles de réparation ectopique de cassures d'ADN tels que les modèles de NAHR (Richardson et al. 1998; Koszul et al. 2004; Linardopoulou et al. 2005; Bailey et Eichler 2006). Ces modèles ne donnent pas de détails sur le mécanisme de formation.

Ayant à notre disposition les annotations des ET et satellites du génome de D. melanogaster, nous avons pu détecter les DS et rechercher d'éventuelles traces du mécanisme de NAHR. Cette travail nous a conduit à proposer un mécanisme de formation des DS, basé sur un modèle de NAHR: le DDSA pour « Duplication-Dependant-Strand-Annealing ». Ce modèle est basé sur le SDSA pour « Synthesis-Dependant-Strand-Annealing », un modèle de recombinaison homologue.

3.1. Pourquoi avoir choisi *Drosophila melanogaster* ?

3.1.1. Son séquençage

Le génome diploïde de *D. melanogaster* est formé de 4 paires de chromosomes, avec la première paire correspondant aux chromosomes sexuels. La taille du génome est de l'ordre de 180 Mb, dont 60 Mb d'hétérochromatine et 120 Mb d'euchromatine. En 1998, le séquençage du génome de la Drosophile a été initié par le groupe du BDGP en ce qui concerne les chromosomes 2, 3 et 4, et par le groupe de l'EDGP pour le chromosome sexuel X. En mai 1998, une collaboration pour le séquençage voit le jour entre le HHMI (« *Howard Hughes Medical Institue* ») en la personne de Gerald M. Rubin et Craig Venter de la compagnie « *Celera Genomics Corporation* ». Ils utilisent alors la méthode de séquençage du WGS (« *Whole genome shotgun* »). Cette méthode de séquençage n'avait jamais été utilisée et testée sur des génomes plus grands que les génomes bactériens (quelques Mb). L'inquiétude était de savoir si elle allait permettre le séquençage de génomes complexes, c'est-à-dire riches en répétitions (Weber et Myers 1997; Green 2007). Courant septembre 1999, une partie euchromatique du génome est achevée. Ce fut la première démonstration de l'efficacité de l'approche par WGS dans le séquençage d'organismes pluricellulaires.

La première séquence complète du génome de *D. melanogaster* a été obtenue en mars 2000 (Adams *et al.* 2000; Myers *et al.* 2000; Rubin et Lewis 2000). Depuis la Release 1, il y a eu quatre « *Releases* » (Celniker et al. 2002). Au moment de notre étude, la Release 4 sortie en Avril 2004, était disponible. La Release 5, prévue comme Release finale, est actuellement disponible. Chaque Release augmente en qualité grâce à la correction d'erreurs, aussi bien au niveau de l'assemblage que de la séquence. Il ne restait que 23 gaps dans la séquence euchromatique de la Release 4. Parallèlement au séquençage de l'euchromatine, un autre groupe, le DHGP, a été chargé de séquencer l'hétérochromatine. Les séquences euchromatiques et hétérochromatiques ont été réunies dans la Release 5.

3.1.2. Et les autres Drosophilidés

Actuellement, 12 espèces de Drosophilidés sont séquencées et disponibles (Clark et al. 2007). *D. simulans*, l'espèce jumelle de *D. melanogaster*, partage un grand nombre de caractéristiques communes avec *D. melanogaster* aussi bien morphologique que génomique. Son génome est également composé de 4 paires de chromosomes. Le projet de séquençage du génome de *D. simulans* soutenu par le NHGRI (« *National Human Genome Research Institute* ») et le NIH (« *National Institutes of Health* »), a voulu mettre en évidence le polymorphisme inter-souches. Six souches de *D. simulans* ont donc été séquencées à partir de plasmides avec une couverture de 1X pour chaque souche. Une autre souche, la souche w[501] a également été séquencée à environ 3X à partir de banques de plamides et de fosmides. L'ensemble des résultats de séquençage a ensuite été utilisé pour créer une séquence mosaïque de *D. simulans* (Figure 19).

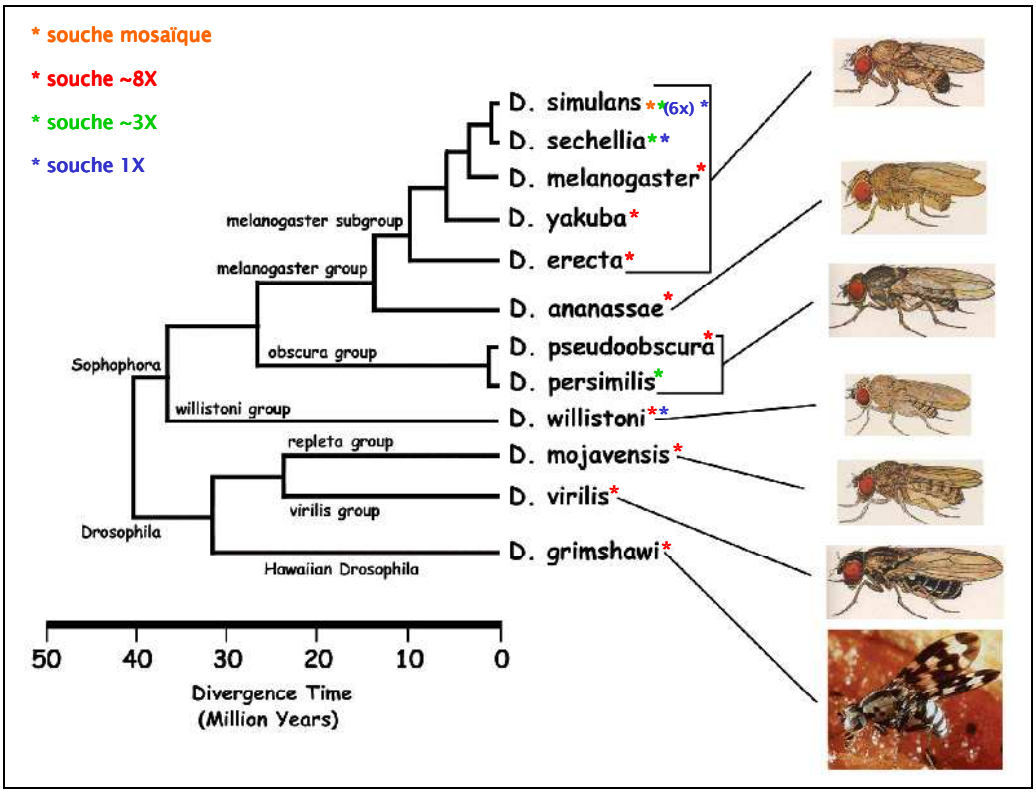


Figure 19. Les 12 espèces de Drosophiles.

Relation entre des 12 espèces de Drosophiles, dont le génome actuellement est séquencé. Chaque sous-groupe est associé à une image montrant les caractéristiques phénotypiques des Drosophiles du sous-groupe. (Source: http://flybase.bio.indiana.edu/). La couverture de séquençage de ces génomes varie entre 1X et 8X. La majorité d'entres ont été séquencé en 8X. Huit souches de *D. similans* ont été séquencées, dont 6 en 1X, une en 3X et une souche mosaïque.

Au moment de notre étude, la séquence génomique (Release 1) de *Drosophila. yakuba* était disponible et de très bonne qualité (couverture de 8X). Comme *D. simulans* et *D. melanogaster*, cette espèce diploïde présente 4 paires de chromosomes dont une paire sexuelle. Les deux espèces *D. melanogaster* et *D. yakuba* ont divergé, il y a environ 10 millions d'années. La séquence des autres génomes de Drosophile, plus éloignés phylogénétiquement de *D. melanogaster* par rapport à *D. yakuba*, n'était pas encore disponible ou de bonne qualité au moment de notre étude. A l'heure actuelle, au moins une souche par espèce a été séquencée avec une couverture de 8X pour la majorité des Drosophilidés (Figure 19).

3.1.3. Un génome de choix

Le génome de *D. melanogaster* nous a permis de travailler avec une séquence génomique de qualité. Au début de ma thèse, les annotations en gènes et ET pour ce génome ainsi que les séquences génomiques de *D. yakuba* et *D. simulans* étaient également disponibles. L'idée étant d'orienter et de dater les évènements impliquant des répétitions, nous étions donc bien placés pour initier cette étude. On s'attendait à identifier des évènements spécifiques de *D. melanogaster* et très récents, ce qui peut permettre d'identifier d'éventuelles traces des mécanismes.

3.2. Obtention des données

3.2.1. Les annotations des répétitions

Le pipeline SegDupPipeline a été utilisé sur la séquence génomique de la Release 4 de *D. melanogaster* (http://www.fruitfly.org/). Pour les étapes d'élimination des ET, j'ai utilisé les annotations officielles de Flybase, produites au laboratoire. J'ai également utilisé les annotations des potentiels nouveaux ET obtenues par BLASTER avec le programme TBLASTX contre l'ensemble des ET eucaryotes connus de « *Repbase Update* ». Pour les satellites et microsatellites, j'ai utilisé les annotations obtenues par le programme TRF pour « *Tandem Repeat Finder* » (Benson 1999).

Pour la première étape du pipeline de détection des DS, nous avons choisi d'utiliser une e-value à $1x10^{-300}$ pour la détection initiale des répétitions par BLASTER. Cette valeur a pour but d'orienter la détection vers des duplications récentes. Nous espérons ainsi identifier d'éventuelles traces laissées par le mécanisme de formation.

3.2.2. La création d'un jeu de séquences contrôles

Afin de pouvoir discuter de la pertinence des caractéristiques observées pour les DS détectées, nous voulu comparer les caractéristiques observées pour les DS avec celles obtenues pour des séquences contrôles. Pour cela, nous avons choisi d'effectuer un tirage aléatoire de 100 fois le nombre de copies de DS détectées.

Pour chaque copie:

1- on définit une fenêtre centrée sur celle-ci. Les bornes de cette fenêtre correspondent à plus ou moins 50 kb des extrémités de la copie.

2- on tire au hasard un couple de coordonnées dans la fenêtre, de manière à ce que la région sélectionnée soit de la même taille que la copie de DS.

Les séquences contrôles correspondent donc à des séquences tirées aléatoirement dans le même environnement génomique et de même taille que les DS détectées. Cette approche permet ainsi de s'affranchir d'éventuels biais dû à la localisation ou à la taille des séquences. Ce jeu de données a été créé à l'aide du programme CREATED_TEST_SET.

3.3. Les méthodes utilisées

3.3.1. Calcul de la fraction recouverte en séquence

L'analyse de la composition en séquence des duplications détectées a été réalisée à l'aide du programme FIND_ANNOT. Ce interroge une base de données afin d'y extraire les annotations en ET, satellites, microsatellites et gènes. A partir de ces annotations, le programme calcule le pourcentage de recouvrement des duplications en ET, microsatellites et gènes.

3.3.2. Détection des répétitions aux extrémités des DS

Le programme FIND_ANNOT peut également être utilisé pour détecter la présence de répétitions aux extrémités des séquences. Après avoir comptabilisé le nombre de répétitions chevauchant les extrémités, le programme range les séquences en trois classes: séquence bornée par des répétitions de part et d'autre; séquence avec une répétition chevauchant une extrémité; séquence dépourvue de répétitions à ses extrémités.

3.3.3. Analyse des points de cassure des DS

Pour cette analyse, seuls les évènements uniques de duplication (duplications à 2 copies) montrant la présence de répétitions à leurs extrémités sont sélectionnés. Chaque séquence est préalablement étendue de 500 pb en 5' et en 3'. J'ai ensuite réalisé, pour chaque événement de duplication, un alignement 2 à 2 avec le programme ClustalW (Thompson et al. 1994). Afin de positionner précisément les répétitions (ET ou satellites) aux extrémités des duplications, les séquences des répétitions ont été ajoutées à l'alignement réalisé *via* du programme ClustalW. Ce dernier permet l'alignement d'une séquence sur un profil, c'est-à-dire sur le résultat d'un alignement. L'identification des points de cassure se fait par la recherche d'un changement brusque dans la qualité d'alignement. En effet, on s'attend à ce que les régions flanquantes des deux copies de DS ne s'alignent pas correctement.

3.3.4. Création des blocs de synténie

Les régions synténiques entre les génomes de Drosophiles ont été obtenues par la combinaison des programmes BLASTER -utilisant BLASTN- et MATCHER. La qualité des séquences des génomes, pris en compte, permet d'être sûr de la qualité des régions synténiques identifiées.

3.3.5. Identification de la séquence matrice

Afin de détecter des évènements de duplication récents et spécifiques de *D. melanogaster*, nous avions fait le choix d'utiliser les génomes de *D. simulans* et *D. yakuba*. Or, l'identité de séquence moyenne entre les séquences génomiques de *D melanogaster* et *D. simulans* est de 97 %. On s'attend donc difficilement à identifier les évènements de duplication spécifique *D. melanogaster*. Par conséquent, seul le génome de *D. yakuba* a été utilisé. Ce génome est un bon compromis puisque l'on s'attend à travailler avec un jeu de données plus conséquent, tout en espérant identifier d'éventuelles traces laissées des mécanismes récents (moins de 10 millions d'années). Nous avons cherché la séquence matrice de duplications, en recherchant de manière automatique la copie commune aux deux génomes dans des régions synténiques (Figure 20).

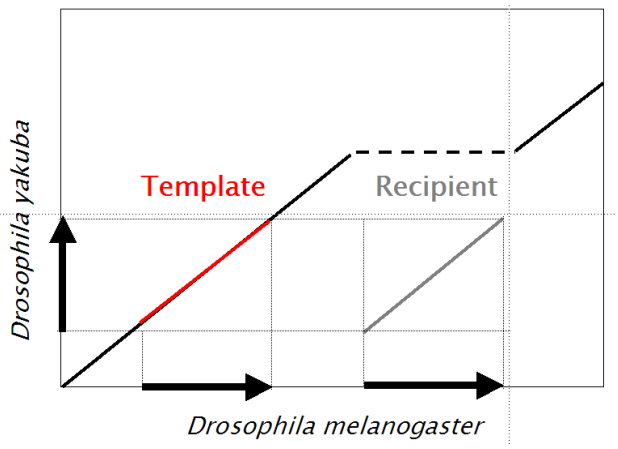


Figure 20. Identification de la séquence matrice.

Le schéma d'un « *DotPlot* » représentant une région synténique entre les deux génomes: *D. melanogaster* et *D. yakuba*. D'après l'exemple présenté ici, la région synténique recouvre une duplication intrachromosomique directe spécifique de *D. melanogaster*. La région synténique est représentée par un trait noir. Une seule des deux copies étant présente chez *D. yakuba*, on suppose que la copie commune au deux génomes, correspond à la séquence matrice (« *template* »; trait rouge) de l'évènement de duplication qui a eu lieu chez *D. melanogaster*. Que les évènements soient interchromosomiques ou intrachromomiques, seuls les cas où les régions flanquantes des deux copies sont conservées, ont été sélectionnés.

3.3.6. Etude de la divergence entre les copies de duplication

Pour les évènements uniques de duplication (duplications à 2 copies) pour lesquels nous avons pu identifier la séquence matrice, j'ai estimé l'âge des deux copies inférant un arbre de duplication à partir de l'alignement multiple des copies de la DS avec la séquence unique présente dans l'autre espèce *via* ClustalW (Thompson *et al.* 1994). La reconstruction des arbres a été réalisée à l'aide du programme PhyML (Guindon et Gascuel 2003). Pour chacun des arbres, j'ai ensuite extrait les longueurs des branches terminales, de bons estimateurs de l'âge des séquences. Ceci permet ainsi de dater l'évènement de duplication chez *D. melanogaster*.

3.4. Les Caractéristiques des DS chez *D. melanogaster*

3.4.1. Leurs caractéristiques générales

A l'issue du pipeline, j'ai obtenu un jeu de données composé de 444 séquences réparties en 138 groupes (Annexe1: Coordonnées des DS détectées). L'ensemble de ces duplications représente 1.4 % du génome de *D. melanogaster* (1.66 Mb / 118.35 Mb). Le nombre de copies de duplication varie de 2 à 32 (Figure 23). Plus de la moitié des copies de duplication partage une identité de séquence supérieure à 97 % (Figure 21). La taille des séquences varie de 346 pb à 81.1 kb (Figure 22).

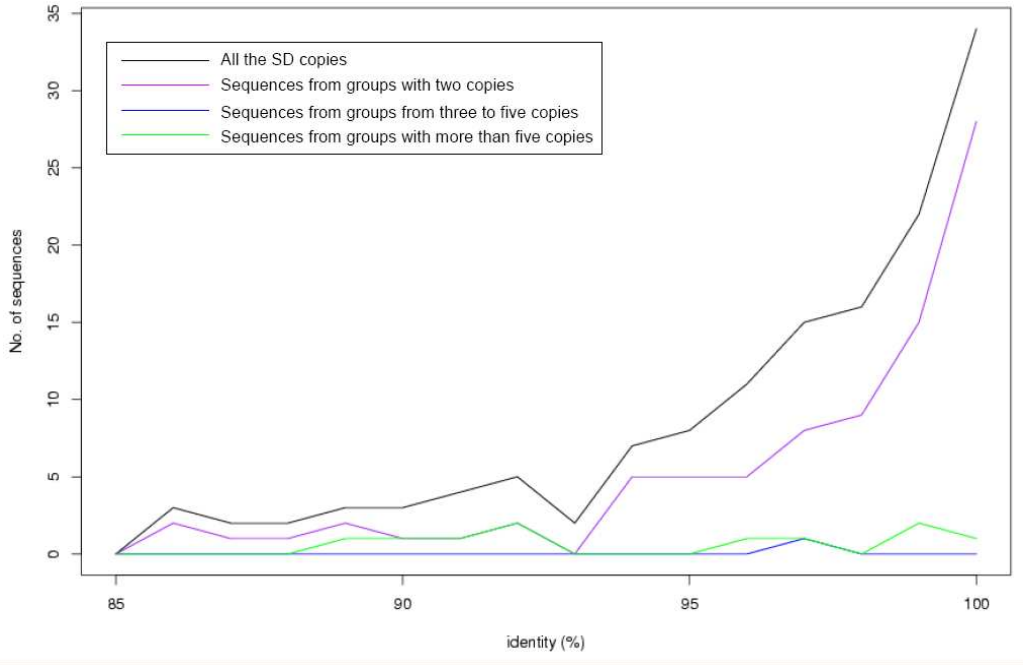


Figure 21. Pourcentage d'identité par nombre de copies de duplication.

Les différentes courbes représentent les distributions par classe de nombre de copies par groupe. La distribuion de l'identité des séquences des groupes à plus de trois copies est quasi-uniforme entre 85 et 100 %. Les séquences des groupes à deux copies ont dans plus de 75 % des cas plus de 95 % d'identité (Fiston-Lavier et al. 2007).

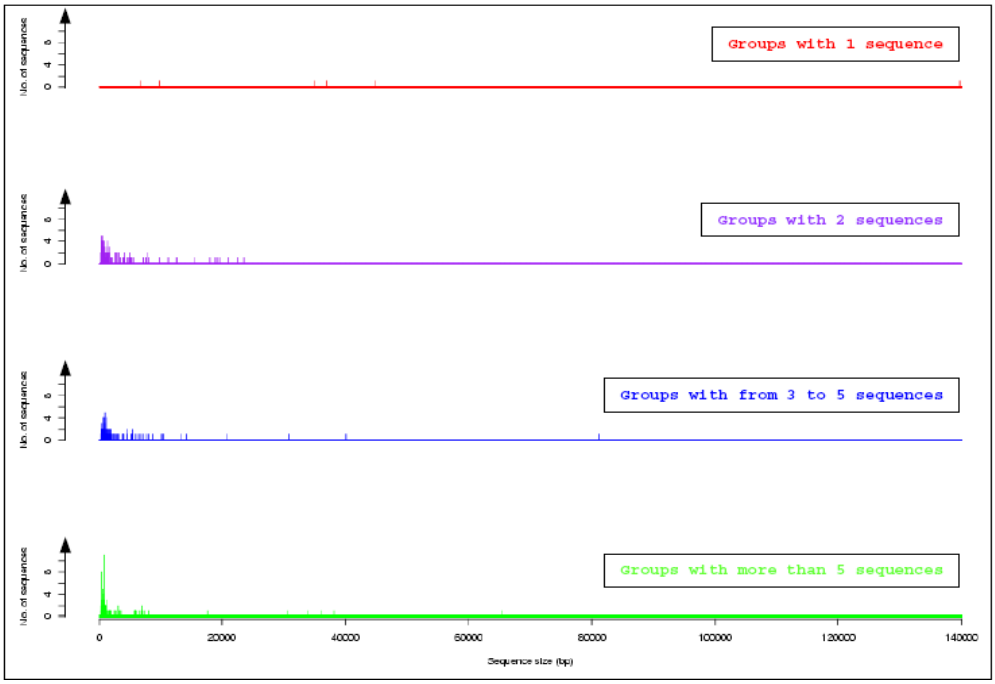


Figure 22. Taille des copies en fonction du nombre de copies par groupe.

Chaque ligne correspond à une classe de duplication en fonction du nombre de copies. La première ligne correspond à la distribution des tailles des copies des groupes à une copie. La seconde ligne correspond à la distribution des tailles des séquences des groupes à deux copies. La troisième ligne présente la distribution de taille des séquences des groupes ayant entre 3 et 5 copies. La dernière ligne correspond aux séquences des groupes à plus de 5 copies (Fiston-Lavier et al. 2007).

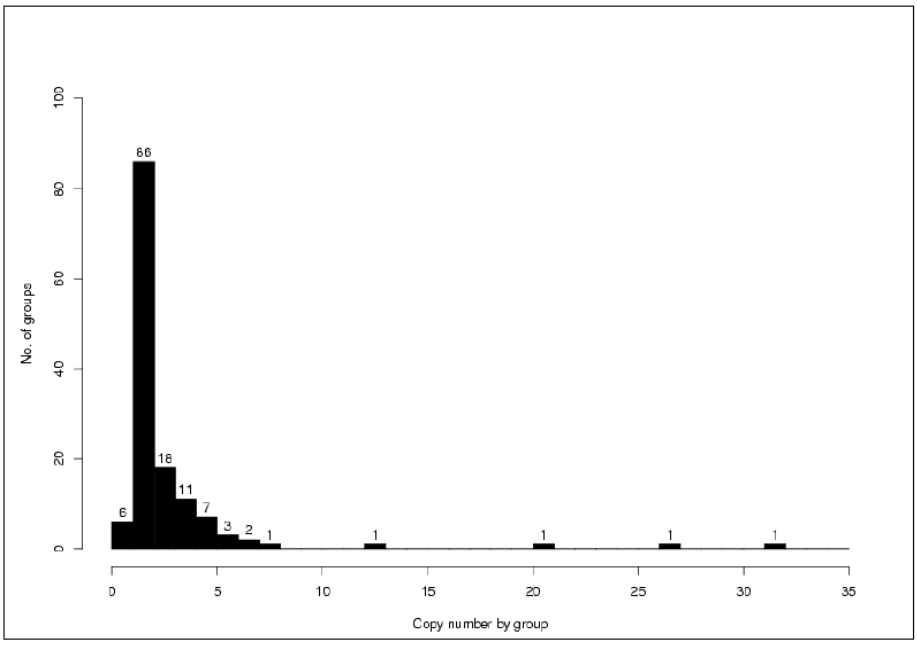


Figure 23. Nombre de copies par groupe de duplication.

Le nombre de groupes est indiqué au-dessus de chaque barre. On observe une majorité de groupes à deux copies. Les 6 groupes à 1 séquence correspondent à des régions composées de répétitions proches ou en tandem (exemple: les gènes des *Histones*). Ces répétitions chevauchantes ou contiguës sont connectées les unes aux autres pour ne former qu'une seule séquence par notre algorithme de clustering (Fiston-Lavier et al. 2007).

Les régions hétérochromatiques enrichies en duplications

Les chromosomes X et 4 ont une densité en duplications de 8.59 et 5.22 DS/Mb, respectivement. Leurs densités sont deux fois plus grandes que celles des autres chromosomes, c'est-à-dire les chromosomes 2 et 3 (avec des densités de 2.62 et 2.18 DS/Mb). Le chromosome 4 est connu pour être enrichi en domaines hétérochromatiques (Sun et al. 2000). Les régions péricentromériques des chromosomes X, 2 et 3, comme les régions subtélomériques des chromosomes X et 3, apparaissent enrichies en DS (Figure 24). La région centrale du chromosome X présente une très forte densité en DS (Figure 24). Les coordonnées cytogénétiques de cette région sont 10F à 13A. Les analyses de cette région, *via* l'analyse des chromosomes polytènes, ont montré que cette région est sous-répliquée, ce qui révèle un état hétérochromatinien (Ashburner 1989). Kaufmann avait déjà suggéré en 1939, l'état hétérochromatinien de cette région pour y expliquer la forte fréquence des cassures d'ADN (Kaufmann 1939). Ces observations nous permettent de suggérer une distribution chromosomique des DS non-uniformes. En effet, nous les retrouvons préférentiellement dans les régions hétérochromatiques.

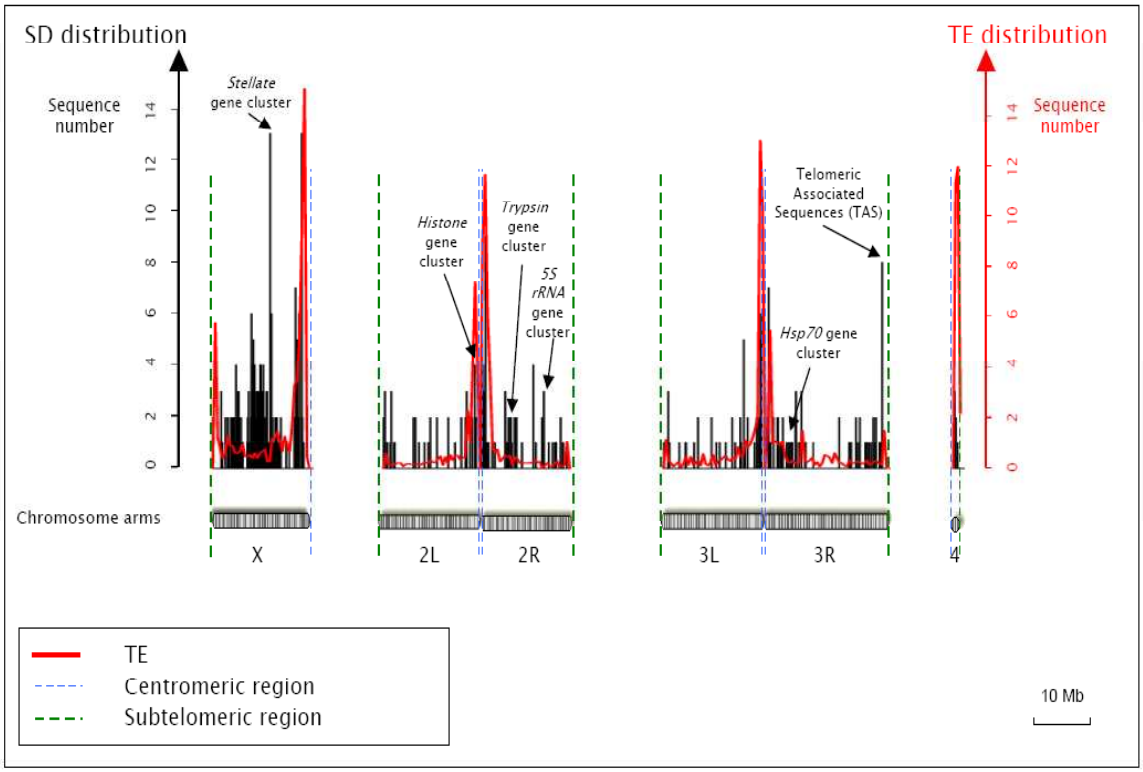


Figure 24. Distribution chromosomique des des ET et des DS détectées.

La distribution des DS est illustrée par des histogrammes noirs. La courbe rouge superposée correspond à la distribution en ET. Les régions centromériques et subtélomériques sont signalées par des lignes verticales en pointillées vertes et bleues (Fiston-Lavier et al. 2007).

Une majorité d'évènements de duplication intrachromosomiques

Afin de ne prendre en compte que les évènements simples, nous avons sélectionné tous les évènements uniques de duplication (duplications à 2 copies). La forte identité de séquence observée entre les deux copies nous permet de supposer que l'une des deux copies dérive de l'autre. Sur les 172 copies, soit 86 évènements uniques de duplication, 70.35 % (121/172) copies sont localisées sur les chromosomes 2 et 3 (Tableau 1). Cette observation est en fait attendue puisqu'on s'attend à ce que, plus un chromosome soit grand et plus on s'attend à y observer un grand nombre de duplications. Lorsque l'on rapporte cet effectif à la taille des chromosomes, le chromosome 4 montre une plus forte densité en DS que les autres chromosomes. On observe également une forte densité au niveau du chromosome X.

On observe une majorité de duplications intrachromosomiques (86 % soit 74 sur 86; Tableau 1). Les duplications interchromosomiques et intrachromosomiques diffèrent significativement en taille. La taille moyenne des copies est respectivement de l'ordre de 3.1 kb et 2.1 kb (Test du $\chi 2$ (Pearson)): $\chi 2=552.22$, ddl=40, P-value$<2.2\text{x}10^{-16}$). Les copies ne semblent pas montrer de préférence chromosomique (Test du $\chi 2$ (Pearson): $\chi 2=8$, ddl=6, P-value=0.24).

Chromosomal arm of Sequence 2 / Chromosomal arm of Sequence 1	X	2L	2R	3L	3R	4	Total number of sequence on each chromosome arm
X	40	1	3	1	2	1	48
2L		22	0	1	0	0	24
2R			32	1	2	0	38
3L				26	0	0	29
3R					26	0	30
4						2	3

Tableau 1. Duplications interchromosomiques et intrachromosomiques.

Ce tableau correspond à la table de contingence du nombre de duplications par bras chromosomique. Seuls les évènements uniques de duplication ont été pris en compte (groupes à 2 copies). Les valeurs sur la diagonale correspondent aux duplications intrachromosomiques. Les autres indiquent le nombre de duplications interchromosomiques. Par exemple, il y a 22 duplications intrachromosomiques sur le bras 2L et une copie de duplication interchromosomique. L'autre copie de cette duplication est localisée sur le chromosome X (Fiston-Lavier et al. 2007).

Les ET, les principaux composants des DS

Afin d'avoir une idée plus précise de la composition en type de séquence des DS, j'ai calculé la fraction recouverte en gènes, en ET et en microsatellites des duplications. Les taux de recouvrement ont ensuite été comparés avec ceux obtenus pour les séquences contrôles (voir 3.2.2. La création d'un jeu de séquences contrôles). Cette étape de comparaison s'est avérée cruciale pour s'affranchir des biais dus à la taille et à la localisation des séquences détectées.

Les résultats indiquent un enrichissement significatif des DS détectées en gènes (26.32 % contre 19.26 % pour les séquences aléatoire; Test de Student: t=4.73, ddl=44842, P-value=2.23×10^{-6}). Ce résultat peut s'expliquer par la fixation de duplications géniques. Une observation similaire est faite pour les taux de recouvrement en ET (recouvrement moyen en ET de 21.44 %) en comparaison avec le jeu de séquence contrôle (recouvrement moyen en ET de 13.75 %; Test de Student: t=5.40, ddl=44842, P-value=6.65×10^{-8}). Les copies d'une même duplication semblent majoritairement regroupées au sein de certaines régions denses en ET. La probabilité pour une séquence tirées aléatoirement dans ces régions de recouvrir la séquence d'un ET est donc très élevée. Or, les séquences tirées aléatoirement dans ces régions (jeu de séquences contrôles), à l'opposé des DS, n'apparaissent pas enrichies en ET. Ce résultat suggère fortement l'implication des ET dans le processus de duplication.

On observe un appauvrissement des DS en microsatellites par comparaison avec les séquences aléatoires (9.79 % contre 12.98 % pour les séquences tirées aléatoirement; Test de Student: t=-7.63, ddl=41995, P-value=2.46×10^{-14}).

Des DS de petite taille

Quarante-neuf pourcents des duplications font plus de 1 kb mais seulement 7.21 % des duplications sont supérieures à 10 kb (Tableau 2). Le génome de *D. melanogaster* apparaît plus pauvre en grandes duplications par comparaison avec les génomes de mammifères (Tableau 2). Chez *D. melanogaster*, on observe une majorité de DS non-géniques (61.71 %). Or les duplications géniques s'avèrent être plus petites que les autres duplications. Ces observations renforcent l'observation de la petite taille des DS chez cet organisme en comparaison avec les génomes de plus grande taille et suggèrent des mécanismes de formation des duplications dépendant de la taille des génomes.

SD size	SDs size distribution: percentage (number of sequences)	Percentage of genome (%)			
		Fly (our results)	Fly	Worm	Human
> 1 kb	49.32 (219)	1.28	**1.20**	**4.25**	**3.25**
> 5 kb	16.44 (73)	1.02	**0.37**	**1.50**	**2.86**
> 10 kb	7.21 (32)	0.79	**0.08**	**0.66**	**2.52**

Tableau 2. Proportions de DS chez certains génomes eucaryotes.

Ce tableau reprend les estimations de proportions de DS chez *D. melanogaster*, *Caenorhabditis elegans* et l'homme. Ces estimation ont été obtenues par Samonte et Eichler (2002). L'approche de détection utilisée permet d'identifier un plus grand nombre de fragments de plus de 5 kb (1.02 % contre 0.37 %). Les résultats suggèrent un plus grand nombre de petites duplications chez *D. melanogaster* en comparaison avec l'homme (Fiston-Lavier et al. 2007).

3.4.2. Les familles multigéniques

Plusieurs classes de DS existent: les DS composées de séquences répétées (ET et microsatellites); les DS sans séquences répétées mais avec des gènes; les DS aux composantes diverses (séquences répétées, uniques et gènes) et les DS correspondant à des satellites. Parmi les 38.29 % duplications géniques, on identifie des familles multigéniques connues. Leurs copies peuvent être en tandem, comme pour les gènes des *Histones*, des *récepteurs des Lipoprotéines*, du *5S ARNr*, des *Hsp70B* et la famille des gènes *Stellate* ou dispersées comme les gènes de la *Beta-tubuline* et de l'*Actine*. On les retrouve majoritairement dans les régions euchromatiques, et en plus grand nombre sur les chromosomes X et 2R. Leur nombre de copies varie de 2 à 100. La famille ayant 100 copies correspond en fait aux gènes des *ARNr 5S* de 143 pb chacun. La taille moyenne des gènes dupliqués est de 618 pb. La forte identité de séquence observée pour les duplications géniques peut être expliquée soit par une forte pression de sélection s'exerçant sur des copies anciennes soit parce qu'elles sont issues d'évènements récents. Sachant que 61.71 % des duplications détectées ne contiennent aucune région génique et que 26.20 % des gènes identifiés ne correspondent pas à des familles multigéniques connues, on peut suggérer que la majorité des duplications détectées soit issue d'évènements récents, au vu de la forte identité de séquence globalement observée. Dans cette situation, on s'attend à identifier d'éventuelles traces laissées par le mécanisme de formation des duplications.

3.4.3. Les minisatellites et duplications en tandem

Une même région génomique peut être impliquée complètement ou partiellement dans plusieurs évènements de duplication. Trente-six duplications ont entre 3 et 5 copies. Seulement 10 duplications ont plus de 5 copies avec un maximum de 32 copies (Figure 23). Ces duplications ont une identité de séquence (identité moyenne=94.07 %) significativement plus faible que celles des groupes à faible nombre de copies (de 2 à 5 copies; identité moyenne=96.73 %; Test de Student: t=2.61, ddl=28, P-value=0.014). Ces répétitions sont significativement plus petites (taille médiane=829 pb contre 1 kb; Test du $\chi 2$ (Pearson): $\chi 2$=7477.27, ddl=90, P-value<$2.2x10^{-16}$). De plus, près des ¾ des répétitions avec plus de 5 copies sont localisées sur le chromosome X. On y trouve plus particulièrement une forte densité en minisatellites. Les minisatellites identifiés sont annotés comme SARDM et XDMR (Tableau 3) (Waring et Pollack 1987; Ashburner 1989). Ces minisatellites ont déjà été décrits dans la banque de données RU (Jurka 2000).

Group n°		49	44	103	54	99	19	117	20	110	106
No. of sequences		6	6	6	7	7	8	13	21	27	32
mean size (bp)		950	1061	6266	441	6207	785	793	4436	6641	936
min size / max size (bp)		607/ 1121	919/1141	649/33754	441/441	513/30617	367/1199	787/829	434/8124	459/65477	576/3253
identity (%)		95.98	88.20	90.20	96.61	91.53	98.71	98,98	99.80	91.19	89,45
Sequence coverage (%)	Gene	0.33	83.26	3.31	0.00	1.74	0.00	66.37	0.00	9.90	0.00
	TE	85.08	0.00	0.02	79.12	0.91	96.96	0.00	98.22	0.34	0.00
	Satellite	15.17	4.43	9.73	9.44	9.76	13.80	6.23	10.52	8.23	9.92
Chromosomal location of all sequences		2/3	X/2R/3	X	2R/3R	X	X/2L/3R/4	X	X/2/3	X (only one on 3R)	X (only one on 2L)
Particularity		1360 *	Actin	XDMR & SAR_DM	Invader4 *	gt,CG32797 ,CG32733, SAR_DM	Jockey*	*Stellate* cluster	297,1731,rover	XDMR & SAR_DM	XDMR & SAR_DM

* TE flanked by repeated regions greater than 20 bp which do not correspond to TE or satellite regions.

Tableau 3. Les groupes à plus de 5 copies.

Ce tableau réunit les informations sur les groupes au plus grand nombre de copies (plus de 5 copies). Quatre groupes correspondent à des faux-positifs, 4 à des minisatellites et 2 à des familles multigéniques. Les groupes 49,54 et 19 correspondent à des DS composées majoritairement par des ET et avec plus de 20 pb répété ne correspondant ni à des ET, ni à des RT. (Fiston-Lavier et al. 2007).

Dans les groupes à plus de 5 copies, on y retrouve également des familles multigéniques tel que les gènes *Stellate* ou les gènes de l'*Actine*. Les gènes *Stellate* euchromatiques sont organisés en tandem de 13 copies. Les copies ont une taille de 787 pb. Les copies partagent entre elles une identité moyenne de 98.98 %. Pour expliquer la conservation entre les gènes, Tulin *et al.* suggère une évolution concertée du « *cluster* » (Tulin *et al.* 1997). Il est localisé au centre du chromosome X au niveau des bandes cytogénétiques 12E1-2. La présence d'un « *cluster* » *Stellate* hétérochromatique a également été notée sur le chromosome X (Abramov et al. 2005). Des variants localisés sur la bande h26 du chromosome X, sont interrompus par des éléments *Copia-like*, des LINE et des ADNr. Cette structure est désignée comme SCLR pour « *Stellate-Copia-LINE Ribosomal DNA* » (Shevelyov 1992; Nurminsky *et al.* 1994). Un autre « *cluster* » est également localisé sur la bande h26, à proximté du SCLR. Abramov *et al.* ont testé l'impact des réarrangements entre les « *clusters* » des deux domaines chromatiniens. Pour expliquer la réplication des gènes hétérochromatiques normalement sous-répliqués, ils proposent un phénomène d'euchromatinisation (Abramov *et al.* 2005).

3.5. La Réparation Homologue (RH), un mécanisme de formation de duplications

3.5.1. Les modèles de réparation des Cassures Double-brin (CDB) d'ADN

Les CDB, connues pour être une cause majeure d'instabilités génétiques, ont des origines très diverses. Elles peuvent survenir suite à l'action de divers agents mutagènes exogènes ou endogènes. Ils peuvent également apparaître spontanément à n'importe quel stade du cycle cellulaire. Un des processus fréquents à l'origine des CDB est la réplication de l'ADN. Ce mécanisme génère environ 10 000 lésions spontanées de CDB par cellule humaine et par jour (Burkart et al. 1999). Les topoisomérases I et II interviennent pour réguler l'enroulement de l'ADN en induisant des coupures qu'elles réparent ensuite. Elles forment une liaison covalente avec l'ADN pour changer son statut hélicoïdal. Ce processus est nécessaire pour la réplication, la recombinaison, la ségrégation chromosomique et la transcription. La topoisomérase I génère des cassures simple brin tandis que la topoisomérase II génère des CDB. Les CDB sont indispensables à la séparation des chromatides sœurs pendant la mitose et la méiose. Elles permettent également d'initier la recombinaison homologue. Ces deux enzymes sont donc impliqués dans la stabilisation du génome mais peuvent également promouvoir des recombinaisons illégitimes qui peuvent engendrer la formation d'aberrations chromosomiques.

Une autre source importante de dommages spontanés de l'ADN est le stress oxydatif provoqué par différentes réactions redox dans les métabolismes aérobies. Ces dommages sont pris en charge par le BER pour « *Base Excision Repair* ». La première étape de ce mécanisme est la création d'une cassure simple brin. Les séquences de microsatellites en subissant des processus dynamiques d'expansion et de contraction peuvent aussi être la source de telles cassures (Pfeiffer et al. 2000).

Afin de réparer une cassure d'ADN, plusieurs mécanismes de réparation vont rentrer en compétition. Les cassures simple-brin d'ADN sont prises en charge par le système SSBR pour « *Single-Stranded Break Repair* ». Pour la réparation des CDB, il existe deux grandes voies de réparation: la Réparation Homologue (ou RH) et non-homologue. La RH se caractérise par une étape de recherche de région homologue. La région sélectionnée sert ainsi de séquence matrice pour la synthèse d'ADN au site de cassure. Dans ce cas, la réparation est conservative. En effet, elle restaure la séquence à

l'identique. Après la CDB, une digestion exonucléasique peut exposer les extrémités 3'OH lésées. Lors de la réparation non-homologue, les extrémités libres sont raboutées grâce à une région de micro-homologie. La réparation est alors non-conservative car il y a perte d'ADN.

3.5.2. La RH des CDB d'ADN

Les mécanismes de « Réparation Homologue », permettent le maintien de l'intégrité génomique par la réparation précise des CDB d'ADN. Le mécanisme de RH prédomine en fin de phase S et en phase G2 du cycle cellulaire, lorsque les chromatides sœurs sont disponibles. En effet, ces mécanismes utilisent comme matrice une séquence homologue présente au niveau de la chromatide sœur (en méiose) ou du chromosome homologue (en mitose) (Rong et Golic 2003). Afin d'expliquer comment les erreurs de RH peuvent conduire à des réarrangements chromosomiques, je ferai d'abord une description détaillée des modèles de RH des CDB d'ADN. Au moins quatre mécanismes de recombinaison homologue peuvent expliquer la réparation des CDB d'ADN dans les cellules mitotiques: (1) le « *Single-Strand Annealing* » ou SSA (Paques et Haber 1999); (2) le « *Break-Induced Replication* » ou BIR (Mosig 1987; Kogoma 1997); (3) le « *Double-Strand Break Repair* » ou DSBR (Szostak et al. 1983) et (4) le « *Synthesis-Dependent Strand Annealing* » ou SDSA (Nassif et al. 1994). Ces modèles sont conservés chez tous les eucaryotes. Les trois derniers modèles sont des processus dits conservatifs, aboutissant à deux séquences intactes. Le modèle SSA est un processus non-conservatif où le produit de la réparation est plus ou moins délété. L'ensemble de ces mécanismes initiés par une CDB d'ADN commence par la détection des deux extrémités d'ADN générées par la cassure. Les extrémités sont ensuite dégradées par une activité exonucléase en 5' → 3', ce qui conduit à la formation de brèches simples brins sortants des extrémités 3'OH. De nombreuses études ont permis de montrer que la digestion peut se produire sur une très longue distance. Chez *Saccharomyces cerevisiae,* la digestion exonucléasique des brins lésés est souvent supérieure à 1 kb (Sun *et al.* 1991; Lee *et al.* 1998).

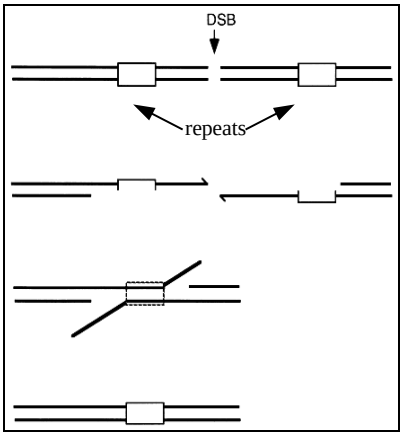


Figure 25. Le modèle SSA.

Ce mécanisme se produit lorsque la cassure a lieu entre deux copies d'une même répétition. Après la digestion exonucléasique des extrémités 5'Phosphate, les deux extrémités 3' OH libres sortantes s'apparient au niveau de la répétition. Les extrémités libres simple brin non-appariées sont ensuite dégradées et la synthèse d'ADN comble les régions lésées. Ce mécanisme conduit à la délétion de la région incluse entre les deux copies de répétition (Paques et Haber 1999).

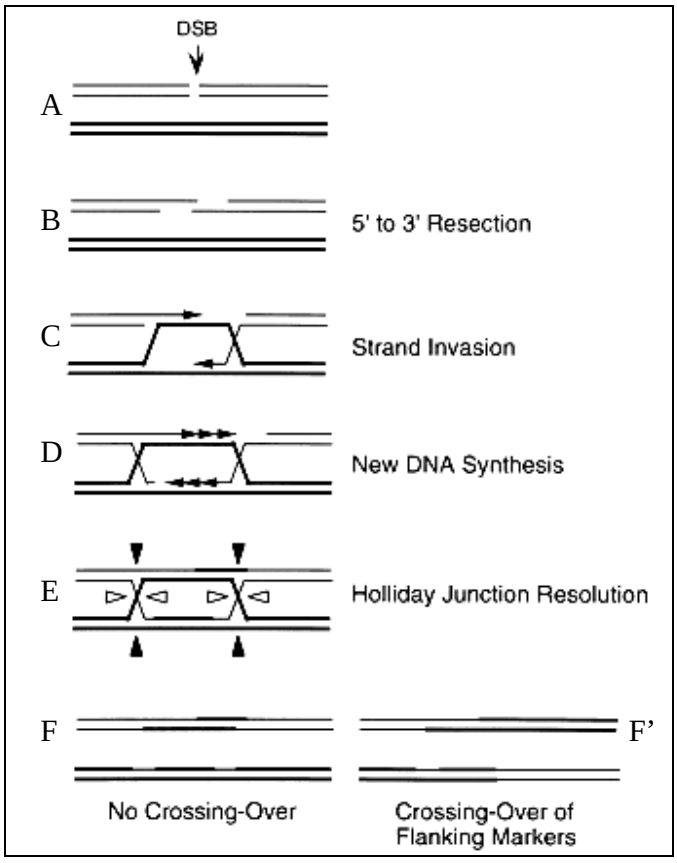


Figure 26. Le modèle DSBR de Szostak.

(**A**) CDB d'ADN. (**B**) Puis digestion des extrémités libre 5'P. (**C**) L'une des deux extrémités 3' OH libre sortante envahit un duplex matrice. (**D**) Un hétéroduplex se forme entre les deux brins du duplex matrice et ceux de la molécule lésée. (**E**) La formation d'un hétéroduplex conduit à la formation de jonctions de Holliday. Ces jonctions pourront être résolues de deux manières possibles: *via* la conversion génique (**F**) ou *via* le crossing-over (**F'**) (Szostak *et al.* 1983; Paques et Haber 1999).

3.5.3. Le modèle SSA (« Single-Strand Annealing »)

Le mécanisme du SSA n'a lieu que lorsque la dégradation s'étend jusqu'à exposer une région d'homologie d'au moins 400 pb entre les deux extrémités 3' OH sortantes (données établies chez *Saccharomyces cerevisae*). Ces deux régions homologues vont alors pouvoir s'apparier (Figure 25). Cette homologie conduit à la disparition d'une des deux répétitions avec la région interne -entre les deux répétitions-, c'est-à-dire à une délétion lorsqu'elle survient au niveau d'un même chromosome (ou intrachromosomique) ou à une translocation si elle à lieu entre deux chromosomes différents (ou interchromosomique). Ce mécanisme est donc non-conservatif (Ozenberger et Roeder 1991).

3.5.4. Le modèle DSBR (« Double-Strand Break Repair »)

Les deux extrémités 3'OH libre crées par la cassure servent d'amorces pour une synthèse d'ADN. Une des extrémités 3'OH libre sortante envahit le chromosome homologue ou la chromatide sœur au site d'homologie et s'y apparie (Figure 26; B, C). Chez *Saccharomyces cerevisiae*, la taille de la région homologue varie entre 25 pb et > 2 kb. Les séquences donneuse et matrice se doivent de présenter une similarité quasi-parfaite (Ira et Haber 2002).

Le second brin de la molécule matrice est déplacé et s'apparie avec le brin qui lui est complémentaire sur la molécule endommagée (Figure 26; C). La structure ainsi formée est appelée « *D-Loop* ». Il y a ensuite formation de doubles jonctions de Holliday (Figure 26; D; (Holliday 1964). Ces jonctions correspondent à des liaisons covalentes induites par la formation de l'hétéroduplex entre les quatre brins d'ADN. Elles sont ensuite clivées pour permettre la séparation des deux molécules d'ADN (Figure 26; E). Ce mécanisme conduit à une réparation semi-conservative, dans la mesure où un nouveau brin synthétisé est présent à la fois sur les brins « donneur » et « receveur ». Le clivage conduit à deux résolutions différentes: (1) la séparation peut s'effectuer avec un échange complet de séquence (crossing-over; Figure 26, F') ou (2) simple transfert d'une séquence sur l'autre (conversion génique; Figure 26, F). Ce modèle est aujourd'hui le modèle le plus souvent admis pour expliquer la recombinaison homologue, et en particulier la recombinaison durant la méiose (Szostak et al. 1983). Il permet de rendre compte du lien fort qui peut exister entre la conversion (échange non réciproque) et le crossing-over (échange réciproque).

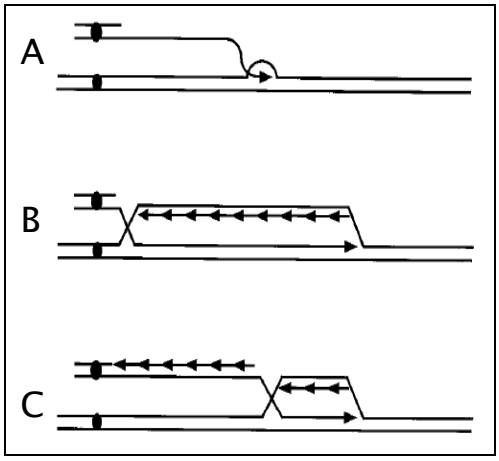


Figure 27. Les trois modèles BIR.

(A) L'une des extrémités envahit un duplex matrice pour former une bulle. Cette dernière migre ensuite le long de la séquence matrice afin de permettre la synthèse d'ADN. Le second brin lésé est réparé après la réparation du premier brin. **(B)** Formation d'une vraie fourche de réplication. Il y a synthèse des deux brins précoces dans le sens 5'→ 3' et tardif dans le sens inverse, par l'intermédiaire de fragments d'Okazaki. **(C)** Le dernier modèle est un modèle hybride des deux autres. D'après ce modèle, après la formation d'une bulle, il y a synthèse simultanée des deux brins lésés. La réparation se fait par conversion génique comme pour le DSBR (Paques et Haber 1999).

3.5.5. Le modèle BIR (« Break-Induced Replication »)

Le modèle BIR propose l'invasion d'une des deux extrémités 3' au niveau d'une courte région homologue. La synthèse d'ADN peut ensuite se faire sur plusieurs kb, voire des chromosomes entiers (Mosig 1987; Kogoma 1997; Morrow *et al.* 1997; Haber 1999). On distingue trois modèles pour expliquer le BIR:

1. Le complexe formé après l'invasion d'une extrémité, appelée « bulle » (Formosa et Alberts 1986), est converti en une petite fourche de réplication unidirectionnelle. La migration de la bulle permet ainsi la synthèse d'une région de taille importante. La synthèse du brin complémentaire est initiée après la dissociation du brin nouvellement synthétisé et son appariement au niveau de son duplex d'origine (Figure 27).
2. L'extrémité 3'OH sortante peut également envahir le duplex matrice de manière à recréer une vraie fourche de réplication (Figure 27). D'après ce modèle, la synthèse des deux brins se réalise comme durant la réplication mais de manière unidirectionnelle: il existe ainsi un brin direct ou brin précoce et un brin retardé ou tardif. Le brin direct correspond au brin qui a initié la formation de cette fourche. Sa synthèse se fait de manière continue dans le sens 5' → 3'. A l'opposé, pour l'autre brin, la synthèse se réalise de manière discontinue dans le sens 3' → 5' *via* les fragments d'Okazaki. Dans ce cas la synthèse est semi-conservative. La structure formée sera ensuite résolue par des endonucléases.
3. Si les deux extrémités de la cassure sont toutes deux homologues au niveau d'une même région chromosomique, cette région sert de matrice. Les deux brins sont réparés par conversion génique (voir le modèle DSBR; Figure 27).

Le BIR induit la synthèse d'une région de très grande taille. Si en fin de synthèse du premier brin, celui-ci ne retourne pas s'apparier à son duplex d'origine, la duplication va s'étendre du site de cassure jusqu'à l'extrémité du chromosome. Ce modèle est reconnu comme un mécanisme de maintien des télomères chez la levure et les mammifères (Lydeard et al. 2007).

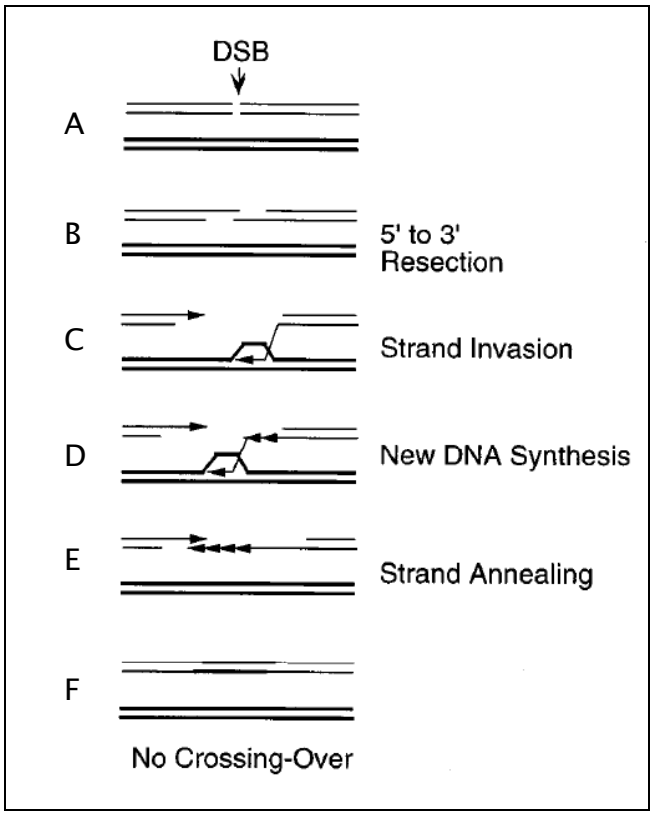


Figure 28. Le modèle SDSA avec formation d'une bulle de migration.

(**A**) Après la CDB, (**B**) les extrémités libre 5' P sont dégradées. L'une des deux extrémités 3' OH libre sortante envahit un duplex matrice (**C**). Ceci va conduire à la formation d'une bulle, c'est-à-dire un désappariement local. L'hétéroduplex formé migre ensuite le long de la séquence matrice pour permettre la synthèse d'ADN (**D**). Le brin nouvellement synthétisé revient s'apparier au niveau de son duplex d'origine ce qui induit la réparation du deuxième brin lésé (**E**). Ce modèle décrit un mécanisme de conversion génique donc sans crossing-over (**F**) (Paques et Haber 1999).

3.5.6. modèle SDSA (« Synthesis-Dependent Strand Annealing »)

Le DSBR ne permet pas d'expliquer les réparations homologues des CDB par conversion génique chez tous les organismes et en particulier chez *D. melanogaster*. En effet, l'étude de la réparation de cassures provoquées par la transposition de l'*élément P* chez cet organisme a montré un faible taux de crossing-over. Le modèle de Szostak *et al.* (DSBR) ne permettant pas d'expliquer l'absence de traces de réparation *via* un crossing-over, le modèle SDSA a donc été proposé comme modèle de RH chez *D. melanogaster*. Ce modèle est considéré comme la voie principale de RH chez *D. melanogaster*. Sa particularité concerne le déplacement du brin d'ADN nouvellement synthétisé le long de la séquence matrice puis son retour en fin de réparation en s'appariant à la molécule lésée (Figure 28; C et D). Ceci permet d'initier la synthèse du deuxième brin lésé (Figure 28; E). Après la digestion exonucléasique et l'invasion d'un des deux brins lésés, les topoisomérases ou les hélicases induisent la formation d'une bulle de réplication (Figure 28). Cette bulle correspond au désappariement local des deux brins du duplex matrice. L'hétéroduplex formé restant de petite taille, le brin complémentaire ne s'intègre pas dans l'hétéroduplex. Il ne peut donc pas y avoir la formation de jonctions de Holliday (Figure 28) (Formosa et Alberts 1986; Nassif *et al.* 1994). La synthèse d'ADN est ici conservative. La synthèse d'ADN peut également être bidirectionnelle: Les deux extrémités 3' OH libres sont alors réparées indépendamment. Les séquences matrices peuvent correspondre soit à la même région homologue, soit à deux régions homologues différentes. Les deux brins nouvellement synthétisés peuvent ensuite se lier soit par réparation non-homologue (NHEJ) ou SSA.

Un modèle de type SDSA avec formation de crossing-over a également été proposé par Ferguson et Holloman (1996). Ils proposent que l'invasion soit initiée par une des extrémités 3' OH libre. Le déplacement de la « *D-Loop* » créée par le premier brin va induire l'appariement de l'extrémité du second brin. Les jonctions de Holliday sont ensuite résolues.

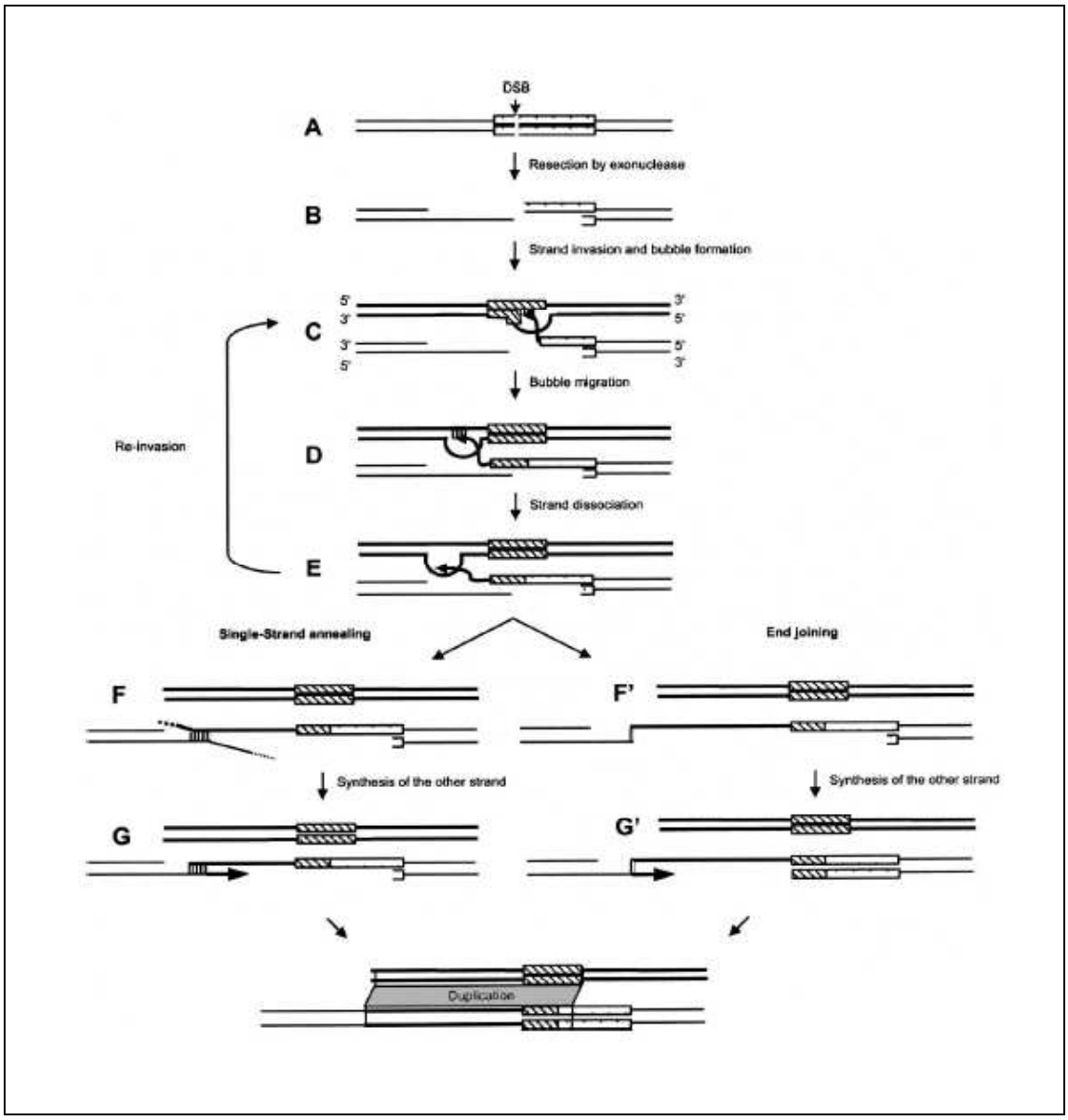


Figure 29. Le modèle DDSA.

(**A**) CDB. (**B**) Digestion des extrémités libre 5' P. (**C**) L'une des deux extrémités 3' OH libre sortante envahit un duplex matrice. Ce qui va conduire à la formation d'une bulle, c'est-à-dire un désappariement local. (**D**) Cet hétéroduplex migre ensuite le long de la séquence matrice pour permettre la synthèse d'ADN. (**E**) Le brin nouvellement synthétisé se dissocie du duplex matrice. Il pourra alors soit (**F et F'**) retourner s'apparier au niveau de son duplex d'origine soit (**C**) sur une séquence matrice. Pour revenir s'apparier au niveau de son duplex d'origine, deux voies sont possibles: Le SSA (**F**) ou le NHEJ (**F'**). Dans les deux cas, la synthèse du deuxième brin lésé est initiée (**G et G'**). Ce modèle prédit ainsi la duplication de la séquence matrice au site de cassure. D'après ce modèle, la nouvelle copie de duplication est bornée par au moins une copie chimère d'une répétition (Fiston-Lavier et al. 2007).

3.5.7. Un modèle de formation des DS, un variant du SDSA

Quand la réparation est fidèle, l'appariement se fait au niveau de la région allélique de la chromatide sœur ou du chromosome homologue. Mais une région ectopique peut être utilisée comme région homologue. Ce mécanisme appelé NAHR utilise comme matrice une région ectopique et conduit à la duplication de cette région, au site de cassure. Des répétitions du génome telles que les ET peuvent induire le choix d'une région ectopique comme matrice.

Nous proposons un modèle basé sur le modèle SDSA pour expliquer la formation des DS chez *D. melanogaster*. Il est considéré comme le modèle préférentiel de RH chez cet organisme (Engels *et al.* 1990; Nassif *et al.* 1994; Rong et Golic 2003). Le modèle BIR peut également conduire à la duplication d'une région génomique. Mais, il conduit dans la plupart des cas à des translocations terminales non-réciproques. Pour souligner son apparentement avec le SDSA, nous l'avons appelé DDSA pour « *Duplication-Dependent Strand Annealing* ». Ce modèle prédit qu'après la cassure et la digestion des extrémités 5' P (Phosphate) lésés (Figure 29; A, B), une extrémité exposant une répétition envahit une région homologue correspondant à une autre copie de cette répétition (Figure 29; C). L'extrémité 3' OH libre peut alors s'apparier au niveau du brin complémentaire de la région homologue. Cet appariement conduit à la séparation des deux brins et à la formation d'un l'hétéroduplex qui encastre le brin en cours de synthèse. Cet hétéroduplex, également appelé « bulle de synthèse», migre le long de la séquence matrice permettant ainsi la synthèse d'ADN (Figure 29; D). A l'opposé de la D-loop formée d'après le modèle DSBR de Szostak, cet hétéroduplex est instable. En effet, la D-loop apparaît comme un hétéroduplex stable grâce à l'appariement des deux brins lésés et la formation des jonctions de Holliday. Pour le SDSA, un seul brin forme l'hétéroduplex. McVey *et al.* (McVey et al. 2004) a montré qu'on pouvait observer des décrochages suivis de ré-invasions pendant la synthèse du brin (Figure 29; E). La ré-invasion peut se faire aussi bien au niveau de la même séquence que sur une autre séquence. Dans le premier cas (ré-invasion sur la même séquence matrice), la bulle de synthèse étant déjà formée, une micro-homologie suffit pour reprendre la synthèse. Dans le deuxième cas, le mécanisme de réparation est réinitialisé.

A l'issu de la synthèse d'ADN, les deux brins sont liés soit selon un mécanisme de SSA sur une très courte homologie (Sugawara et al. 1997) (Figure 29; F, G), soit par un mécanisme de NHEJ par micro-homologie (Paques et Haber 1999) (Figure 29; F', G'). Ce mécanisme conduit ainsi à la duplication d'une région ectopique au site de cassure.

On peut également considérer une réparation bidirectionelle simultanée (pas nécessairement synchrone) des deux extrémités lésées utilisant deux séquences matrices distinctes. Dans ce cas, la résultante correspondra à la duplication de deux régions ectopiques au même site.

Pour valider ce modèle, j'ai d'abord fait l'inventaire des différentes traces attendues sur les séquences de DS suivant ce modèle. Puis, j'ai recherché au niveau des séquences des DS identifiées, les traces attendues et testées la pertinence de celles-ci.

3.6. Identification des traces du mécanisme

3.6.1. Les traces de la ré-invasion du brin

D'après le modèle DDSA, des décrochages du brin peuvent avoir lieu en cours de synthèse (McVey et al. 2004). Le brin ainsi libre peut se raccrocher sur la même séquence matrice, vers l'arrière ou vers l'avant par rapport au site de décrochage.

Les traces attendues

La bulle de synthèse n'étant formée que par un seul brin de la molécule donneuse, cet hétéroduplex instable peut permettre une succession de dissociations suivies de ré-invasions du brin pendant la synthèse. A la fin du processus, la séquence nouvellement synthétisée retourne s'apparier au niveau de son duplex d'origine. Dans le cas où la ré-invasion s'effectuerait au niveau de la même séquence matrice, on doit pouvoir observer des traces de ce processus. Ces traces ne seront observables qu'au niveau de l'alignement entre la séquence matrice et la nouvelle copie. Les traces dépendent de l'orientation de la ré-invasion par rapport au site de décrochage du brin. Elles correspondent à des insertions ou délétions. Ces réarrangements ont lieu sur la séquence nouvellement synthétisée, alors que la séquence matrice reste intacte. Elles sont donc identifiées au niveau de l'alignement des 2 séquences comme des « *gaps* » localisés sur une des deux séquences selon le processus de la ré-invasion. En effet, en fonction de l'orientation du processus de ré-invasion, on doit observer des gaps spécifiquement sur la séquence donneuse ou la receveuse et par conséquent les répétitions sur l'autre séquence. Lors d'une ré-invasion vers l'arrière par rapport au site de dissociation, le processus ré-utilise une même région comme matrice et cela de manière successive. On s'attend donc à observer une nouvelle répétition sur la séquence nouvellement synthétisée. Une seule des deux copies de cette répétition est présente dans la séquence matrice alors que l'autre y apparaît deux fois en tandem, dont une en face du « gap ».

Dans le cas inverse, c'est-à-dire une ré-invasion vers l'avant par rapport au site de décrochage, il y a un saut de matrice. Une région de la séquence matrice n'est donc pas recopiée. La recherche d'une homologie est nécessaire pour reprendre la synthèse d'ADN. Le complexe de synthèse d'ADN étant déjà en place sur la séquence matrice, une courte homologie -ou micro-homologie de quelques paires de bases- sera suffisante. Dans ce cas, on s'attend à observer un gap sur la séquence nouvellement synthétisée,

contigu à des traces de micro-homologie sur la séquence matrice. L'une de ces petites répétitions doit être retrouvée au niveau du gap et l'autre à la borne opposée du gap avec une région unique entre les deux répétitions (Figure 30).

Dans le cas de la ré-invasion en arrière, l'étape de recherche de micro-homologies sera également nécessaire mais ces traces seront confondues avec la répétition en tandem. La distinction entre les deux types de traces est donc dépendante de l'orientation du gap (séquence matrice ou nouvelle copie) et de la présence d'une région unique entre les deux répétitions (Figure 30).

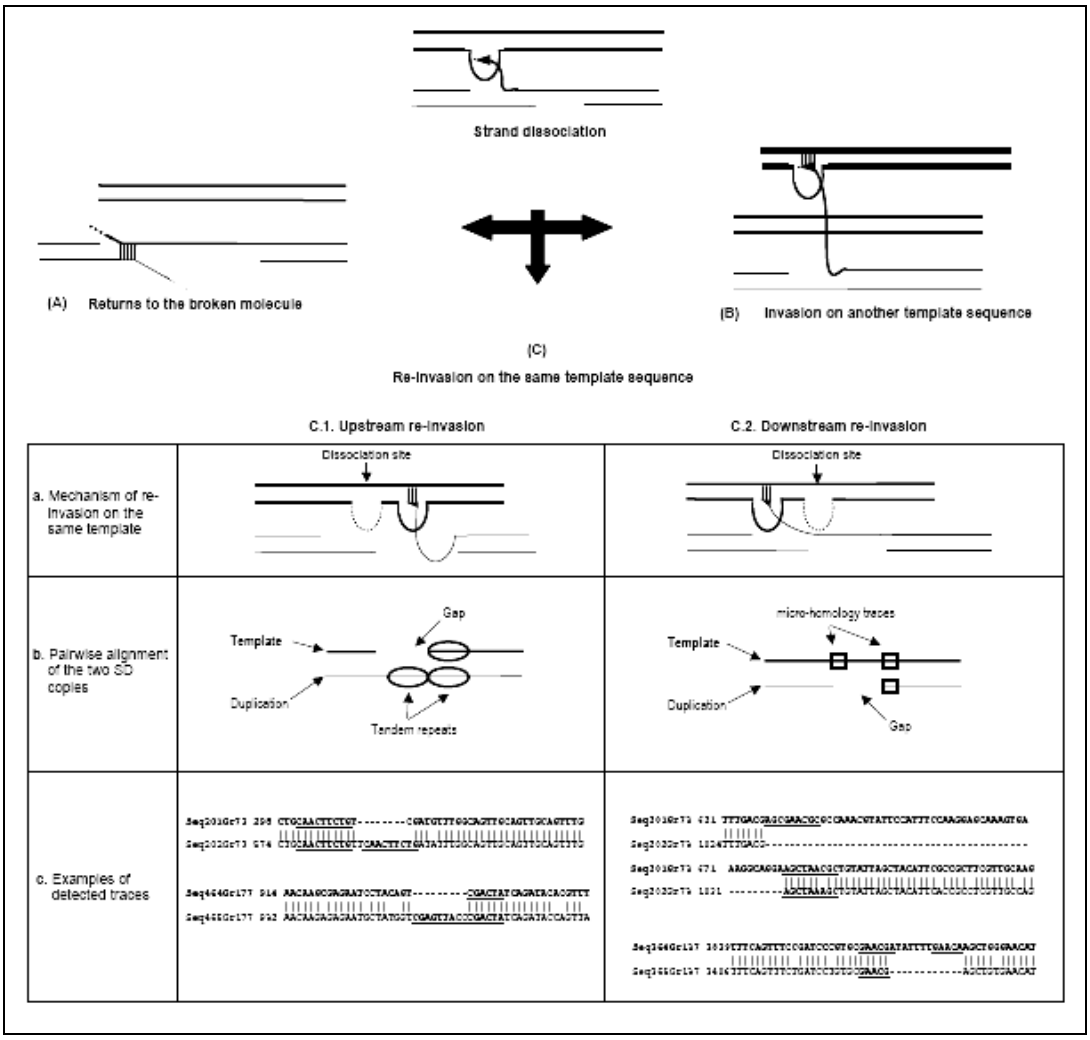


Figure 30. Traces attendues d'après un processus de ré-invasion.

(A), (B) et (C) représentent les différentes possibilités de ré-invasion du brin après dissociation. (C) Etant le cas où le brin ré-envahit la même séquence matrice. Dans ce cas, deux types de traces sont attendues: (C1) Schéma des traces attendues, si ré-invasion vers l'arrière. (C2) si ré-invasion vers l'avant par rapport au site de dissociation (Fiston-Lavier et al. 2007).

Méthodes de détection

Afin de détecter les traces laissées par ces deux processus, une approche de génomique comparative utilisant le génome *D. yakuba* a été mise en place. Par simplicité, nous avons choisi de travailler avec les séquences impliquées dans des évènements uniques de duplication (duplications à 2 copies) du génome de *D. melanogaster*, soit 86 évènements. Pour chacun de ces évènements, il y a une séquence donneuse et une séquence receveuse et seules les situations où une seule des deux copies est présente dans le génome de *D. yakuba* sont informatives pour identifier la copie donneuse.

Malgré la haute qualité des séquences génomiques, un certain nombre de cas se sont avérés douteux. En effet, lorsque les deux copies sont strictement en tandem chez *D. melanogaster* alors qu'une seule copie est retrouvée chez *D. yakuba*, il est difficile de savoir si les deux copies n'ont pas donné lieu à un artefact d'assemblage. Pour éviter à ce problème, seules les copies contiguës séparées par une région interne synténique ont été étudiées.

Nous avons ainsi pu identifier les séquences matrices, avec sécurité, pour 12 évènements de duplication qui ont eu lieu chez *D. melanogaster* après la spéciation entre les deux génomes (c'est-à-dire il y a moins de 10 millions d'années. Ensuite, nous avons recherché les traces attendues d'après le modèle DDSA pour ces 12 évènements de duplication. Pour chacune des duplications, les alignements des copies 2 à 2 ont été réalisés. Ces alignements ont été obtenus *via* un programme d'alignement local ne pénalisant pas les grands gaps, développé au laboratoire. Afin d'avoir la sécurité de recouvrir complètement les séquences de duplication, les copies de duplication ont été étendues de 500 pb de part et d'autre. En effet, l'extension de HSP peut être limitée par l'algorithme de BLAST utilisé ici. Les paramètres d'alignement utilisés sont 10 pour le poids du « *match* », 12 comme pénalité de « *mismatch* », 16 comme la pénalité d'ouverture de gap, 4 pour la pénalité d'extension de gap, et 100 pb comme pénalité de longueur du gap au-delà duquel il n'y a plus de pénalité pour son extension. Seuls les gaps entre 5 et 100 pb ont été pris en compte. Les gaps de moins de 5 pb risquent d'être trop fréquents. Ils peuvent aussi être liés à un mauvais alignement de séquences. Or ces gaps ainsi que ceux de plus de 100 pb peuvent être dus à d'autres mécanismes. L'analyse manuelle des gaps sélectionnés permet d'identifier des traces malgré une forte divergence entre les répétitions. Afin de ne pas prendre en compte les RT générées par un autre mécanisme que par la ré-invasion de brin pendant la réparation de CDB, les RT avec plus de deux copies n'ont pas été prises en compte. En effet, les mécanismes

d'expansion pourraient aussi être à l'origine de leur formation. De plus, afin d'être sûr que les traces de micro-homologie identifiées ne sont pas dues au hasard, seul les répétitions de plus de 3 pb et contiguës à une séquence interne unique ont été considérées.

Types of traces / Gaps location		Tandem repeats	Micro-homology traces	No traces	Total number of gaps
template sequence	All gaps	**8** [(a)]	2	7	17
	Intergenic and intronic gaps	**7** [(a)]	1	6	14
duplicated sequence	All gaps	4	**13** [(b)]	19	36
	Intergenic and intronic gaps	2	**6** [(b)]	12	21

Tableau 4. Traces du processus de ré-invasion du brin selon le DDSA.

Ce tableau comptabilise les différentes traces identifiées au niveau des alignements de séquence 2 à 2 des duplications sélectionnées. On détecte ainsi 8 gaps (sur 17) sur la séquence matrice avec des RT situées sur la séquence donneuse. Sept gaps sur 8 sont localisés dans des régions intergéniques ou introniques. [(a)] Traces de ré-invasion vers l'arrière. [(b)] Traces de ré-invasion vers l'avant (Fiston-Lavier et al. 2007).

Nous avons observé deux fois plus de gaps sur la séquence matrice (36 gaps) que sur la séquence nouvellement synthétisée (17 gaps). Parmi les gaps localisés sur la séquence matrice, 8 sont associés à des RT localisées sur la séquence nouvellement synthétisée. Seuls 4 gaps sur 36 (gaps localisés sur la nouvelle séquence) sont associés à des RT (Tableau 4; Test du $\chi 2$: $\chi 2 = 14.0046$, ddl = 1, P-value = 1.824×10^{-4}). Les traces observées –gaps dans la matrice associés aux RT– correspondent à des traces de ré-invasion en arrière. Nous avons également détecté significativement plus de traces attendues selon le processus de ré-invasion en avant, c'est-à-dire des gaps sur la séquence nouvellement synthétisée associé à des traces de micro-homologie (Tableau 4; Test du $\chi 2$: $\chi 2 = 4.29$, ddl = 1, P-value = 3.834×10^{-3}).

A cause de la pression de sélection exercée sur les gènes, on s'attend à ce qu'une duplication contenant un gène évolue sous contrainte. Ainsi, la pression de sélection peut biaiser nos observations. En effet, les biais d'occurrence des traces observées dans

le cas de duplications géniques pourraient être explicables par des mécanismes de sélection. Afin de s'affranchir de ce biais, nous avons refait cette analyse en ne prenant en compte que les traces localisées au niveau des régions introniques ou intergéniques. Le biais d'occurrence des traces apparaît toujours: 7 contre 2 gaps associés à des RT et 6 contre 1 gap associé à des traces de micro-homologies (Tableau 4). Ces observations soutiennent le modèle de migration de la bulle comme mécanisme de formation des DS.
L'observation d'une seule copie chez *D. yakuba*, pourrait également suggérer une autre histoire évolutive: un évènement de duplication chez l'ancêtre commun suivit de l'élimination d'une des deux copies chez *D. yakuba*. La présence de traces pourrait dans ce cas s'expliquer par d'autres évènements tels que du « *gap repair* » pour les traces de micro-homologie. Cependant, si la présence des copies de duplication chez *D. melanogaster* est due à un évènement de duplication qui a eu lieu après la spéciation entre les deux organismes, on s'attend à observer une plus grande divergence entre la copie de *D. yakuba* et celles de *D. melanogaster*. Des arbres phylogénétiques des trois séquences nous ont permis de vérifier la divergence entre les copies (Annexe2: Analyse des 12 évènements de duplication sélectionnées). Dans la majorité des cas (10/12), les séquences sont plus proches entre les deux copies de *D. melanogaster* qu'entre ces copies et celle de *D. yakuba*. Ces observations confirment l'hypothèse d'évènements de duplications récents spécifique de *D. melanogaster*, sans biais de sélection sur une des copies (longueurs de branches équivalentes sur l'arbre phylogénétique).

3.6.2. Un modèle affiné

Séquestration du brin en cours de synthèse après sa dissociation

On observe deux fois plus de gaps au niveau de la séquence nouvellement synthétisée. Cette observation suggère une plus grande facilité du mécanisme à réaliser des sauts de matrice vers l'avant. Nous proposons une séquestration du brin dans le complexe de formation après sa dissociation. En restant dans la bulle, alors que celle-ci continue encore sa migration le long de la séquence matrice, le brin dissocié peut facilement se raccrocher au complexe quelques bases plus loin. A l'opposé, une ré-invasion vers l'arrière nécessite la ré-initialisation du processus avec la re-formation du complexe vers l'arrière. La taille des gaps correspondant à la distance entre les sites de dissociation et de ré-invasion, nous suggérons une ré-invasion à proximité du site de dissociation (16 pb de distance en moyenne). Ces observations permettent de proposer que le complexe de synthèse formé continue encore sa migration après dissociation du brin, ce qui permet à la séquence de se ré-apparier facilement au complexe et cela à proximité du site de dissociation.

Une recherche d'homologie préférentiellement à proximité du site de cassure

Les évènements uniques de duplication (duplications à 2 copies) sont en majorité des duplications intrachromosomiques (86 % intra contre 14 % inter), c'est-à-dire sur le même chromosome. Afin d'estimer la distance entre les copies de duplications intrachromosomiques, nous avons virtuellement assemblé les deux bras des chromosomes sans tenir compte des régions pericentromériques non-assemblées. Pour les chromosomes 2 et 3, les bras L pour « *Left* » et R pour « *Right* » ont ainsi été réunis. Seuls les coordonnées des annotations sur les bras R ont ensuite été convertis.

Les deux copies de duplications intrachromosomiques sont distantes de 192 pb à 13.7Mb. La majorité des copies sont localisées à proximité les unes des autres: dans 50 % des cas, elles se retrouvent à moins de 14kb l'une de l'autre.

Une étude réalisée par Rong et Gloic en 2003, a suggéré que dans les cellules mitotiques, la séquence matrice est préférentiellement localisée au niveau de la chromatide sœur ou sur le chromosome homologue. Nos résultats permettent de suggérer que la séquence matrice est choisie préférentiellement en *cis* de la cassure,

c'est-à-dire préférentiellement à proximité. En effet, la médiane de la distance observée entre les copies est de 14 kb et les ¾ des duplications sont distantes à moins de 50 kb.

La détection de duplications interchromosomiques est un argument en faveur d'un mécanisme de NAHR. Après des analyses de la recombinaison de chromosomes hétérologues dans des cellules souches, Richardson *et al.* (Richardson et al. 1998) ont proposé un modèle de recombinaison homologue pour expliquer la formation des duplications interchromosomiques chez les mammifères. En effet, la conformation nucléaire peut induire le rapprochement de deux régions homologues de deux chromosomes différents. D'après notre modèle du DDSA, les ET apparaissent comme de très bons candidats pour induire la formation de ce type de duplication. Ces répétitions dispersées partagent une forte identité de séquence, ce qui permet une longue et forte similitude entre les copies permettant d'initier la recombinaison. La suite de mon travail a donc consisté mettre en évidence l'implication de ces séquences dans le processus de duplication.

3.7. Les ET, inducteurs du processus de duplication

Les ET, par leur activité de transposition, peuvent créer des CDB d'ADN. Comme montré chez les mammifères (Bailey et al. 2001), les DS de *D. melanogaster* apparaissent plus denses dans les régions hétérochromatiques, régions connues pour leur forte densité en ET. On peut supposer que ces régions sont sujettes à un grand nombre de CDB générées par l'excision des ET durant la transposition. Lorsque la cassure se situe dans une région dense en ET, la recherche d'homologie pourrait utiliser les ET comme initiateurs du processus de réparation en raison du grand nombre de copies proches et de la forte identité de séquence entre les copies. De surcroît, ces régions hétérochromatiques sont connues pour être plus tolérantes à l'insertion de répétitions. Cette tolérance expliquerait aussi la forte densité en ET.

Lorsque l'on compare la distribution en DS et en ET, il semble y avoir une corrélation positive entre ces deux types de répétitions. Seule la région centrale du chromosome X n'affiche pas cette relation (Figure 24). Dans la suite de ce manuscrit, je mentionnerai à nouveau les particularités de cette région. Les ET, composants majoritaires des DS apparaissent donc comme de bons inducteurs du processus de duplication d'après le modèle DDSA.

Si des ET sont à l'origine de duplications, on s'attend à identifier deux types de traces. Premièrement, d'après les mécanismes de HR, la séquence répétée peut recouvrir en plus de la séquence dupliquée, la région contiguë. On s'attend donc à détecter des traces d'ET aux extrémités des duplications. Deuxièmement, ces copies devraient correspondre à des copies chimères, résultats du changement de matrice. J'ai donc recherché ces traces parmi les DS détectées.

3.7.1. Présence de répétitions aux extrémités des duplications

Afin de tester cette prédiction, j'ai comptabilisé les cas où la DS présente aucune, une ou deux répétitions(s) à ses extrémités. Pour cette analyse, j'ai recherché pour chaque borne de ces DS les annotations en ET et en microsatellites. Afin de tester les observations, les résultats obtenus ont ensuite été comparés aux résultats obtenus pour les séquences génomiques choisies de manière aléatoire (La création d'un jeu de séquences contrôles).

Type of repeats	Type of sequence analyzed	Percentage of sequences (No. of sequences)		
		No repeat	One repeat	One repeat on each end
Transposable elements	**SD sequences**	**80.85 % (359)**	**13.96 % (62)**	**5.18 % (23)**
	Control sequences	*85.84 % (38115)*	*9.47 % (4207)*	*4.68 % (2078)*
Microsatellites	**SD sequences**	**69.37 % (308)**	**26.58 % (118)**	**4.05 % (18)**
	Control sequences	*75.80 % (33657)*	*22.28 % (9891)*	*1.92 % (852)*

Tableau 5. Présence d'ET et de microsatellites aux extrémités des DS.

Pour chaque DS, les répétitions (ET et microsatellites) aux extrémités ont été comptabilisées. On distingue trois cas de figure: la copie ne présente aucune répétition aux extrémités, une répétition est localisée à une des extrémités de la copie, la copie est bornée par deux répétitions. Les valeurs entre parenthèses correspondent aux effectifs (Fiston-Lavier et al. 2007).

Les DS dans la majorité des cas, ne présentent pas de traces de répétitions à leurs extrémités. Ces traces ont pu être éliminées après la duplication par des délétions ou des mutations. Par ailleurs, il y a moins de DS avec deux répétitions de part et d'autre que de DS avec la présence d'une répétition à une de ses extrémités. Le même résultat est obtenu avec les séquences tirées au hasard. Cependant, la comparaison des résultats pour les duplications et pour les séquences tirées au hasard, montre une différence significative, avec plus de répétitions aux extrémités des DS (Tableau 5; Test du $\chi 2$ de conformité: $\chi 2 = 10.99$, ddl = 2, P-value = 4.09×10^{-3}). On retrouve principalement des copies de la famille des *INE1*. Ce résultat était attendu car cette famille d'éléments avec plus de 2000 copies, correspond à la famille la plus représentée dans le génome de *D. melanogaster* (Quesneville et al. 2005; Bergman et al. 2006).

On observe également un enrichissement significatif des DS en microsatellites aux extrémités par comparaison avec les séquences tirées au hasard (Test du $\chi 2$: $\chi 2 = 16.63$, ddl = 2, P-value = 2.45×10^{-4}). L'enrichissement en ET des DS suggère une implication de ces répétitions dans le processus de duplication. Mais ces observations ne nous fournissent pas plus d'indications sur le mécanisme de formation des DS.

3.7.2. Recombinaison entre éléments aux points de jonction des DS

D'après le modèle de DDSA, la recherche d'homologie en position ectopique conduit à observer aux extrémités des duplications, la présence de copies d'ET chimères (Figure 31). Les méthodes d'annotation des ET ne sont pas assez sensibles pour détecter

les éléments chimères. En effet, deux copies adjacentes d'un même élément sont identifiées comme une seule et même copie. Cependant, si on aligne la copie chimère avec une des deux copies donneuses, on doit observer de part et d'autre du point de cassure, une qualité d'alignement très différente. Le point de cassure correspond à la limite entre les deux types d'alignements. La recherche des points de cassures a été réalisée au niveau des alignements entre la copie donneuse et la copie receveuse (Figure 31). Alors que la copie donneuse de la duplication présente une seule copie d'ET, on s'attend à ce que la copie receveuse soit en fait une copie chimère (Linardopoulou et al. 2005) (Figure 31). Les points de cassure de 19 (sur 86) évènements uniques de duplication (duplications à 2 copies) ont pu être ainsi identifiés (Annexe4: Analyse de points de cassure des duplications). Ces observations sont des arguments forts pour proposer que le modèle DDSA est induit par les ET pour la formation de nombreuses DS chez *D. melanogaster*.

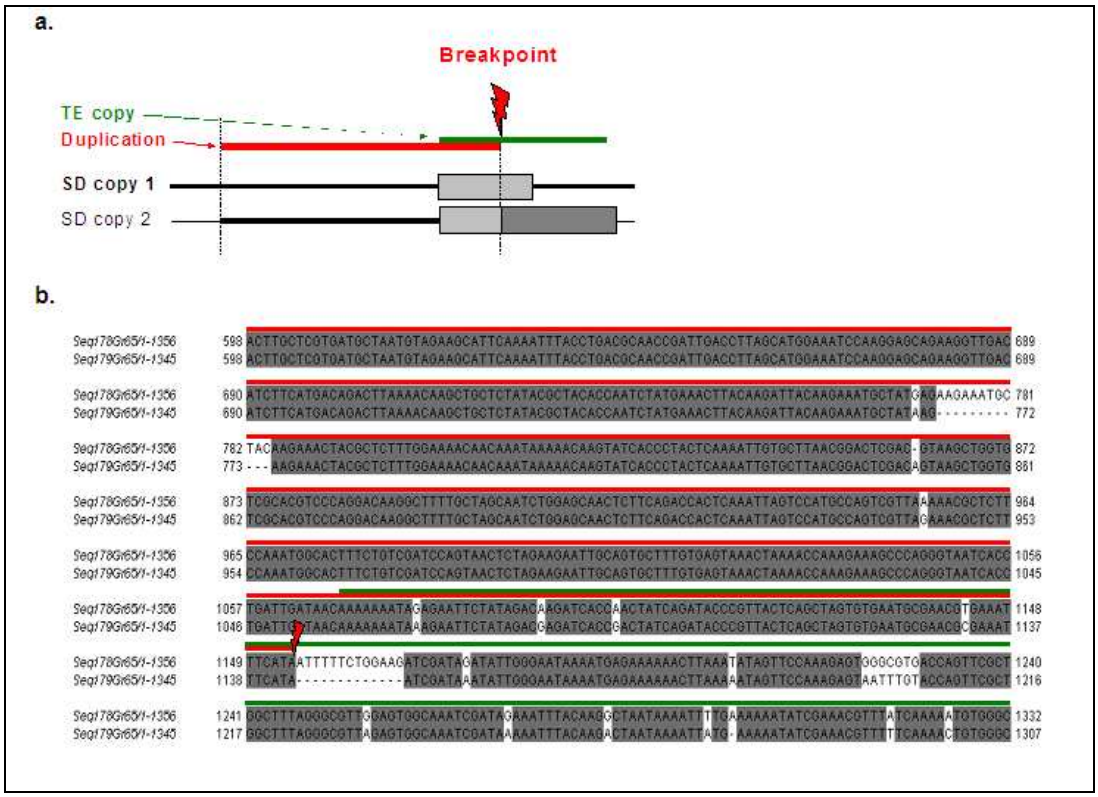


Figure 31. Analyse des points de cassure des DS.

Les traits rouge et vert repositionnent respectivement les DS et les ET sur les séquences (a) Schéma de l'observation attendue. (b) Exemple d'alignement entre la séquence donneuse et la receveuse (Fiston-Lavier *et al.* 2007).

3.8. Conclusion/Discussion

3.8.1. Un mécanisme dépendant de la taille du génome

L'ensemble de nos résultats est en faveur du modèle DDSA et permet de rejeter l'hypothèse stipulant que les traces observées seraient le résultat de processus post-duplication. Dès qu'une CDB de l'ADN apparaît, la machinerie cellulaire fait rentrer en compétition plusieurs voies de réparation. Cette machinerie de réparation se compose de plus de 20 protéines dont des endonucléases, des exonucléases, des hélicases, des kinases et ligases. Les premières protéines en scènes ont pour rôle de détecter la cassure. Par la suite en fonction des protéines disponibles, il y aura ou non recherche d'une région homologue, puis réparation et enfin ligation. Les variations des abondances relatives et des affinités de ces enzymes entre les génomes eucaryotes peuvent expliquer les différences de proportions en DS observées chez les eucaryotes (Tableau 2). Celles-ci seraient dues à une différence de processus de réparation des CDB. En effet, dans un grand génome, on s'attend à ce que les cassures soient plus fréquentes que dans un petit génome. En conséquence de quoi, les grands génomes peuvent être soumis à un plus grand nombre d'erreurs de réparation et donc plus de duplications. Si l'on compare la taille des génomes eucaryotes avec la proportion de DS, on constate en effet que plus un génome est grand et plus il est riche en DS. Chez la souris, près de 1.2 % du génomes correspondant à des DS récentes ont été identifiées dans un génome de 2.5 Gb (Waterston et al. 2002). Chez le rat (génome de 2.75 Gb), la proportion en DS s'élève à 2.92 % (*Rattus norvegicus*, v.3.1; (Rew 2004). Chez l'homme, cette proportion est d'au moins 5 % (taille du génome estimé à 3 Gb; (Bailey *et al.* 2001; Lander *et al.* 2001; Venter *et al.* 2001; 2002).

3.8.2. Le DDSA, un modèle de formation des DS

L'analyse détaillée des traces identifiées, nous permet de proposer le modèle DDSA comme un modèle permettant d'expliquer la formation des DS chez *D. melanogaster* et de suggérer un rôle des ET dans leur formation. La probabilité d'observer par chance des répétitions de 10 pb associées à un gap localisé spécifiquement sur la séquence matrice est très faible. Le même raisonnement peut être fait pour les traces de micro-homologie. Les biais de traces attendues d'après le DDSA que nous observons, sont des arguments forts en faveur du modèle. De plus, après la dissociation du brin, nos

analyses montrent une ré-invasion préférentielle vers l'avant par rapport au site de dissociation. L'ensemble des observations permet ainsi d'affiner le modèle: nous proposons une séquestration du brin après dissociation dans le complexe de synthèse de l'ADN. En effet, la distance entre le site de dissociation et le site de ré-invasion sur le même brin correspond à la taille des gaps observés. La taille moyenne des gaps observés étant de le l'ordre de 16 pb (taille médiane de 12 pb), on propose que le brin dissocié reste séquestré dans la bulle de migration alors que celle-ci continue à avancer le long de la séquence matrice. Ainsi, tant que le brin en cours de synthèse reste dans le complexe, il peut ré-envahir aisément la séquence matrice en amont. Par contre si la bulle, en continuant sa migration, se sépare du brin nouvellement synthétisé, ce dernier pourra soit envahir la même séquence matrice en retournant en arrière, soit une autre séquence matrice ou encore retourner au niveau de sa molécule d'origine. On s'attend alors à ce que la ré-invasion vers l'avant ainsi que le retour du brin synthétisé au niveau de sa molécule d'origine soit les deux cas les plus probables puisqu'ils sont moins contraints physiquement. Le modèle DDSA est le premier modèle détaillé permettant d'expliquer la formation des DS. La migration de la bulle permet des dissociations précoces au cours de la synthèse et la ré-invasion du brin dissocié au niveau de la même séquence matrice, pouvant conduire à la formation de petites duplications. Beaucoup de mécanismes de recombinaison homologue associés à des ET peuvent expliquer la formation de duplications. La présence de signatures du modèle de la bulle de migration est un argument fort en faveur d'un modèle de NAHR *via* la formation d'un hétéroduplex instable. Les mécanismes pouvant induire du crossing-over, comme le DSBR, sont donc exclus.

4. Le turnover des séquences

C'est en 1928 que Emil Heitz, à partir d'observations exclusivement histologiques, définit l'hétérochromatine comme les régions chromosomiques très condensées et très colorées dans le noyau métaphasique. Plusieurs caractéristiques biochimiques ou encore cytologiques complètent cette définition. On sait aujourd'hui que l'hétérochromatine de part son état de compaction, est peu sensible à la DNAse I. Alors que les méthylations de certaines lysines de l'Histone 3 (H3-K4, H3-K36 et H3-K79) sont corrélées avec l'activation de la transcription, la méthylation des lysines 9 (H3-K9) et 27 (H3-K27) sont des marques de l'état réprimé de la chromatine (Lachner et al. 2001; Fischle et al. 2003). La triméthylation de H3-K9 et la monométhylation de H3-K27 sont associées à la formation de l'hétérochromatine (Rice et al. 2003). La relation étroite existante entre les répétitions et la structure de la chromatine a depuis longtemps été soulignée (Lippman et al. 2004). Ces régions sont connues pour être enrichies en répétitions. Mais, on ne sait toujours pas si c'est la présence de répétitions qui induit l'hétérochromatinisation d'une région ou si les répétitions se propagent préférentiellement dans l'hétérochromatine. Dans le but d'aborder ce point, nous avons voulu estimer le turnover des répétitions (ET, DS et satellites) dans l'hétérochromatine. Cette étude implique l'analyse de la dynamique d'insertion et l'estimation des forces qui tendent à éliminer ces répétitions: délétion, recombinaison et sélection. Nous avons donc réalisé une analyse comparative de la dynamique des répétitions dans l'euchromatine et l'hétérochromatine. Au vu des caractéristiques des différentes répétitions, j'ai pu proposer un modèle de dynamique de ces séquences dans l'hétérochromatine. Estimer le turnover des séquences de l'hétérochromatine permet d'estimer le turnover des séquences génomiques qui évolue de manière neutre, c'est-à-dire sans pression de sélection due aux gènes.

4.1. Pourquoi avoir choisi *Arabidopsis thaliana* ?

4.1.1. Un bon modèle d'étude

Cette étude a été réalisée chez *A. thaliana*, plus communément appelée l'arabette des dames. *A thaliana* est la première plante au génome séquencé. Bien qu'elle ne présente aucun intérêt agronomique particulier, cette mauvaise herbe de la famille des *Brassicaceae* a suscité un certain intérêt dans le monde de la recherche. Elle apparaît comme un très bon organisme modèle pour la recherche en génétique végétale. En effet, en raison sa petite taille, cette plante, principalement autogame, se cultive et se reproduit facilement en laboratoire. Son génome estimé à 130 Mb, est réparti sur 5 chromosomes. Les chromosomes 2 et 4 sont acrocentriques alors que les chromosomes 1, 3 et 5 sont métacentriques. Le nombre de gènes est estimé à environ 30000. Les régions hétérochromatiques de ce génome sont, en grande partie, séquencées. On distingue, en plus des régions péricentromériques, des « *knobs* », c'est-à-dire des régions hétérochromatiques surnuméraires.

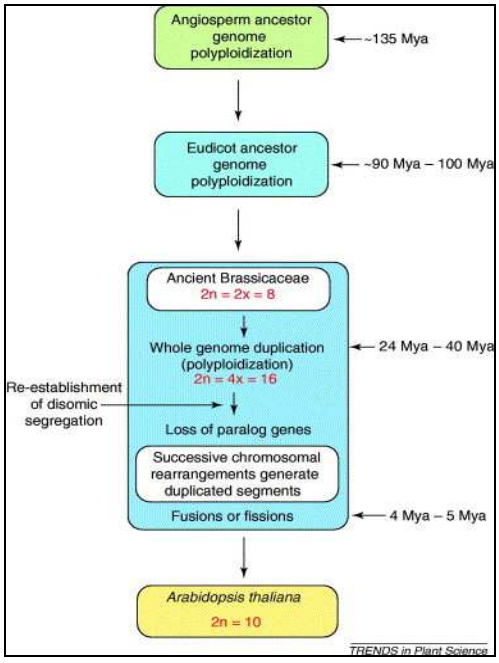


Figure 32. Histoire évolutive du génome de *A. thaliana*.

Ce schéma représente l'histoire évolutive du génome de *A. thaliana* retracée d'après des analyses de fossiles. Le génome ancêtre des Angiospermes a subi un évènement de polyploïdisation, il y a environs 135 millions d'années. Son génome fils, le génome ancêtre des Eudicots, a également subi un évènement de polyploïdisation, il y a environs 95 millions d'années. Entre 24 et 40 millions d'années, le génome ancêtre des Brassicaceae a ensuite subi une nouvelle polyploïdisation. Le génome passe alors de 2n=8 à 2n=16. Pour expliquer la taille du génome actuel *d'A. thaliana* (2n=10), d'importants réarrangements chromosomiques tels que de la fusion et de la fission ont été proposés (Henry *et al.* 2006).

Le génome d'*A. thaliana* aurait, de plus, subi au cours de son évolution plusieurs duplications globales du génome (Figure 32) (Fransz *et al.* 2000; Henry *et al.* 2006). L'histoire évolutive de ce génome montre une forte régulation du nombre de chromosomes chez le génome ancêtre. En effet, après au moins trois évènements de polyploïdisation, le génome ancêtre a subit des évènements de fusion et de fission de chromosomes. On peut donc parler de remaniement important puisque le génome est passé de 2n=16 à 2n=10, ce qui souligne le dynamisme de ce génome.

4.1.2. *hk4S*, le *knob* du chromosome 4

Les *knobs* sont des régions hétérochromatiques différentes des régions péricentromériques et subtélomériques. Ces régions hétérochromatiques sont plus couramment observées chez les plantes. Également appelés, renflements hétérochromatiques, les *knobs* ont été identifiés, pour la première fois, chez le maïs (McClintock 1929). La présence de grands « *clusters* » de RT au sein des *knobs* semble être une caractéristique de ces régions. On y retrouve plus souvent des RT de l'ordre de 180 pb. Ces régions sont également riches en rétrotransposons (Peacock *et al.* 1981; Ananiev *et al.* 1998; 2000).

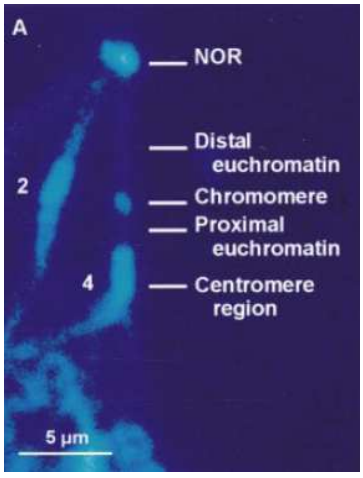


Figure 33. *hk4S*, le knob du chromosome 4 de *A. thaliana*.

Image de FISH des chromosomes 2 et 4 de *A. thaliana* au stade Pachytene. Les régions bleues claires correspondent aux régions hétérochromatiques. Seul le bras court du chromosome 4 est visible sur cette image. Le *knob* du chromosome 4 (ou *hk4S* ou chromomere) est localisé entre les régions euchromatiques distales et proximales (Fransz *et al.* 2000).

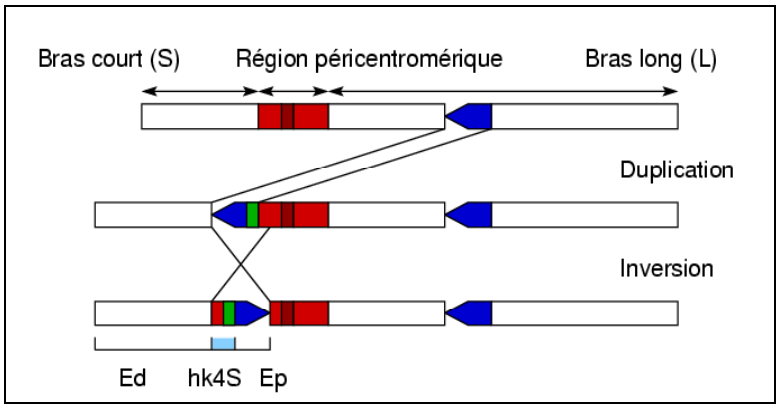


Figure 34. Schéma de l'histoire évolutive de *hk4S*.

Le modèle proposé pour expliquer la formation de *kh4S* se déroule en deux étapes. Il y aurait eu duplication d'un segment euchromatique du bras long au niveau du bras court et cela à proximité de la région péricentromérique. Ensuite, il y aurait eu hétérochromatinisation vers la région dupliquée adjacente à la région péricentromérique. La région hétérochromatinisée correspond au rectangle vert. Cette région issue de la duplication ainsi qu'une région hétérochromatique péricentromérique a ensuite subit une inversion. La nouvelle région hétérochromatique est localisée sur le bras court du chromosome 4 entre le domaine euchromatique distal (Ed) et le domaine euchromatique proximal (Ep).

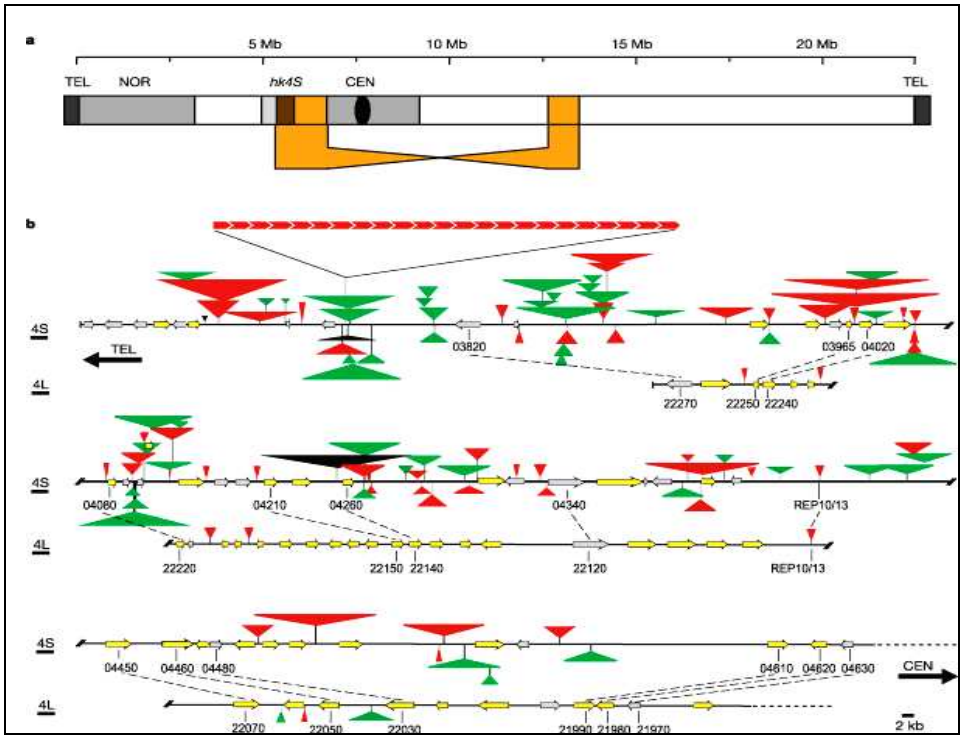


Figure 35. Le *knob* du chromosome 4 (*hk4S*).

a)-Localisation de *hk4S* et de la région donneuse qui lui est associé sur le chromosome 4. Les régions gris clair indiquent les régions hétérochromatiques. L'évènement de formation du *knob* est représenté par la bande orange. La région donneuse (en orange) est située dans la région euchromatique du bras long du chromosome 4 alors que le *knob* (en marron) est localisé à proximité de la région péricentromérique sur le bras court du chromosome 4. Cette région fait partie de l'inversion d'une plus grande région.
b)-Représentation des correspondances des annotations entre la séquence donneuse et le *knob*. Les flèches jaunes indiquent la position et l'orientation des gènes connus, et les grises, les gènes hypothétiques. Les triangles représentent l'insertion des ET (rouge pour les transposons; vert pour les rétrotransposons; noir pour les insertions plus complexes d'ET). Les flèches rouges représentent les RT (Lippman *et al.* 2004).

Deux *knobs* ont pu être identifiés chez *A. thaliana*. Appelés *hk4S* et *hk5L*, ils sont respectivement localisés sur le bras court du chromosome 4 et sur le bras long du chromosome 5 (Figure 33) (Fransz *et al.* 1998; 2000). En raison de la trop faible distance entre le *hk5L* et la région péricentromérique, les limites de cette région sont difficilement distinguables. Seules les limites de *hk4S* ont pu être identifiées. Un scénario en deux étapes a été proposé pour expliquer la formation de *hk4S* (Paul Fransz -communication personnelle; Figure 34): (1) une région euchromatique du bras long du chromosome 4 aurait été dupliquée sur le bras court du chromosome 4, à proximité de la région péricentromérique. Il y aurait eu ensuite propagation de l'hétérochromatine de la région péricentromérique en direction de la nouvelle copie de duplication. (2) Puis, une partie de la duplication ainsi qu'une région péricentromérique aurait subit une inversion. Cette nouvelle région hétérochromatique correspond donc à une région péricentromérique excentrée associée à une séquence dupliquée qui a été hétérochromatinisée (Figure 34). La taille de *hk4S* est estimée à environ 700 kb (Fransz *et al.* 2000).

L'hypothèse de l'hétérochromatinisation a été proposée après une analyse fine des deux copies de duplication. Les alignements de séquence de ces deux copies montrent bien une conservation en ordre et en orientation des gènes (Figure 35) (8 sur 33 gènes). La copie du *knob* présente un enrichissement important en ET, en comparaison avec la copie donneuse (Figure 35). En effet, alors que la séquence donneuse est pauvre en ET, la séquence receveuse est interrompue par au moins 34 rétrotransposons et 40 transposons, souvent insérés les uns dans les autres. Un « *cluster* » de RT de 23 unités d'environs 2 kb chacunes, a également été identifié. Ces RT, absentes de la copie donneuse, seraient dérivées d'ET (Lippman *et al.* 2004). En effet, elles partagent une certaine similitude de séquence avec les éléments *Mutator-like* ou *MULE*, des transposons.

La présence de répétitions associée aux caractéristiques cytologiques indiquent l'hétérochromatinisation de cette région. Le *hk4S* a été identifié chez les écotypes *Wassilevskija* et *Columbia*, mais pas chez *C24* et *Landsberg erecta*. Malgré l'absence de *knob*, l'écotype *Landsberg erecta* présente la duplication, la première étape de formation du *knob*. D'après l'histoire évolutive de ces écotypes, le *knob* du chromosome 4 résultant de l'inversion, serait donc issue d'un évènement récent.

Cette association entre répétitions et hétérochromatine a déjà été discutée par Csink et Henikoff (Csink et Henikoff 1998) puis abordée par Bennetzen (Bennetzen 2000).

Le changement de statut chromatinien étant récent, le *knob* du chromosome 4 est un bon modèle pour mieux comprendre cette relation. On s'attend à ce que les répétitions de cette région aient un statut intermédiaire entre celles de l'euchromatine et celles de l'hétérochromatine péricentromérique. Le *hk5L* laisse apparaître des caractéristiques similaires à celles des régions péricentromériques. Donc seul le *hk4S* sera pris en compte comme « *knob* » pour notre étude. Les caractéristiques des séquences de cette région hétérochromatique seront discutées dans ce chapitre.

4.1.3. Une grande part des régions hétérochromatiques séquencées

Le génome d'*A. thaliana* est l'un des rares génomes pour lequel les régions hétérochromatiques centromériques sont presque entièrement séquencées (AGI 2000). Fransz *et al.* ont positionné les limites des domaines euchromatiques et hétérochromatiques sur le chromosome 4 à partir de données de FISH (Fransz *et al.* 2000).

Afin de travailler avec un jeu de données conséquent, il nous a fallu délimiter les domaines hétérochromatiques pour tous les chromosomes. Pour cela, nous nous sommes basés sur la composition en ET et satellites de ces domaines. Tout d'abord, j'ai calculé le pourcentage de recouvrement moyen en ET et satellites des domaines hétérochromatiques pour le chromosome 4. J'ai ensuite pu déterminer empiriquement une valeur seuil permettant d'estimer approximativement les régions hétérochromatiques des autres chromosomes. Une étape d'agrégation des régions proches ayant un recouvrement supérieur au seuil permet de ne pas être biaiser par la taille des fenêtres utilisées, et ainsi de définir proprement les limites des domaines chromatiniens.

4.1.4. Les annotations disponibles

Au début de notre étude, la version 6 des annotations des gènes de « *The Arabidopsis Information Resource* » (TAIR) étaient disponibles. Ce sont donc ces annotations qui ont été utilisées pour celle-ci. Actuellement, la version 7 des annotations de gènes est en ligne: (ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR7_genome_release/).

L'annotation des DS a été réalisée à l'aide du pipeline que j'ai développé durant ma thèse (voir plus haut). En ce qui concerne les satellites, les annotations ont été effectuées au laboratoire à l'aide d'un BLAST du génome contre la banque de données de satellites de « RU » pour Repbase Update. Les annotations officielles des ET sont disponibles sur le site de TAIR (Release 5; http://www.arabidopsis.org/). Ces annotations ont été produites au laboratoire *via* une version améliorée du pipeline décrit dans (Quesneville *et al.* 2005; 2008). Différents jeux de séquences de références d'ET ont été utilisés. Le premier utilise les séquences de référence de la banque de donnée RU.

Ces séquences de référence ont été obtenues par différentes personnes et donc avec différentes méthodes. Ce jeu de données est appelé RU. Les trois autres jeux de séquences sont des mosaïques de copies génomiques construites pour optimiser différents critères (Buisine *et al.* 2008). Le jeu de données « *OptCoding* » correspond au jeu de données RU auquel des modifications au niveau des régions codantes ont été apportées. Ainsi, les régions contenant des codons stop sont remplacées par des régions de séquences de copies génomiques qui en sont dépourvues, tout en respectant la phase de lecture. Ce type de modification permet également de corriger les décalages de phase. Un troisième de jeu de données nommé « *MaxSize* » a été construit dans le but de répertorier la diversité des copies d'ET sous la forme de séquence mosaïque. A partir du jeu de données RU, un jeu de données de séquences chimères a été créé. Le quatrième jeu de données appelé « *Opt* » est une combinaison des jeux « *OptCoding* » et « *MaxSize* », visant à tirer partie des avantages des deux. Ces trois jeux de séquences permettent de détecter spécifiquement des séquences oubliées par « RU » (Buisine *et al.*, 2008).

4.2. Méthodes de datation des répétitions

4.2.1. Les méthodes classiques de datation des séquences

Plusieurs méthodes de datation d'apparition des répétitions ont été proposées. Elles ne permettent pas toutes d'estimer le même temps de divergence. La méthode la plus répandue est basée sur la datation des rétro-éléments à LTR. Elle consiste à dater la divergence entre deux LTR. En effet, lors de la transposition, les deux LTR sont rendus identiques par le processus de transposition. Après l'évènement de transposition, les deux LTR subissent des mutations de manière indépendante. La divergence entre les deux LTR permet d'estimer le temps écoulé depuis l'évènement de transposition. Cependant, de nombreuses copies de rétro-éléments sont fortement délétées et ne peuvent donc pas être datées. On détecte généralement un plus grand nombre de rétro-éléments tronqués (un ou pas de LTR associé à la région interne) ou des solo-LTR résultants de la recombinaison entre deux LTR d'un élément. Cette méthode se limite, de plus, qu'à une seule classe d'ET. Son utilisation n'étant pas représentative de toutes les familles d'ET, nous nous sommes intéressés à d'autres méthodes de datation.

Une autre approche repose sur la reconstruction phylogénétique d'une famille de répétitions. Cette approche consiste à inférer un arbre de séquences pour un ensemble de copies de répétitions. Cet arbre phylogénétique permet ainsi de retracer l'histoire évolutive d'une famille de répétitions.

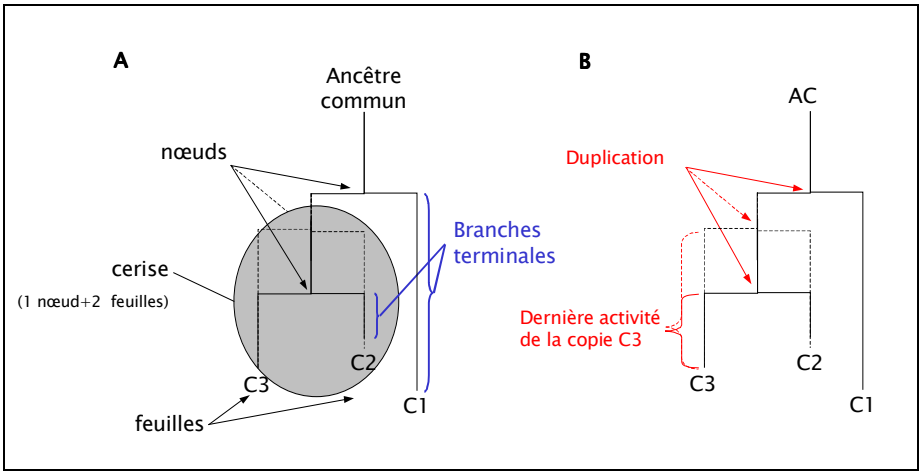


Figure 36. Quelques termes phylogénétiques.

A. Cette figure réunit le vocabulaire phylogénétique utilisé dans ce chapitre. B. Interprétation biologique de l'arbre. Les nœuds correspondent aux évènements de duplication et les longueurs des branches terminales au temps écoulé depuis la dernière activité des copies. On entend par « dernière activité », le dernier processus évolutif. Par exemple pour un ET, la dernière activité peut correspondre à la transposition de l'élément ou la duplication d'un fragment contenant cet élément.

Par exemple, si une copie est dupliquée (C1 donne C2), et que la nouvelle copie est ensuite elle-même dupliquée (C2 donne C3), on obtient une famille de 3 copies (C1, C2 et C3) (Figure 36). Ceci implique un arbre avec deux nœuds internes et 3 feuilles. Chaque nœud de l'arbre correspond à une activité de duplication. La distance entre le premier nœud (ou nœud ancêtre soit AC) et le deuxième nœud permet d'estimer le temps écoulé entre les deux évènements de duplication. La distance entre l'une des deux copies du dernier évènement et de l'ancêtre commun (AC) ne reflète pas l'âge des copies C2 et C3, mais l'âge de la famille. La longueur de la branche terminale d'une copie, ou distance entre la copie et le nœud le plus proche, permet d'estimer l'âge de sa dernière activité (Figure 36).

Les feuilles ou nœuds externes représentent les copies dans leur état actuel, alors que les nœuds internes leurs ancêtres hypothétiques. Ainsi, la distance entre une séquence consensus qui représente empiriquement un état ancestral, donc un nœud interne, et une copie, est un estimateur de l'âge de la copie. Cette distance ne reflète pas l'âge de la copie depuis sa duplication, mais la distance depuis l'apparition de la famille.

4.2.2. La longueur des branches terminales des arbres

Nous avons construit les arbres phylogénétiques de séquences pour chaque famille de répétition, puis extrait la longueur de chaque branche terminale. La longueur de la branche terminale correspond à une distance évolutive. Elle traduit l'âge de la dernière duplication de la copie. Cette distance correspond au nombre de substitutions subies au cours de l'évolution, rapporté au nombre de sites. Les évènements d'insertion et de délétion sont habituellement exclus du calcul des distances évolutives. Prendre en compte ces deux paramètres revient à complexifier le modèle évolutif, ce qui rend les calculs trop lourds.

La fiabilité des arbres dépend de la qualité des alignements et donc des séquences. Par conséquent, avant d'utiliser une méthode de reconstruction phylogénétique, une étape de pré-traitement a donc été nécessaire. Pour chaque famille de répétitions, seules les copies de plus de 100 pb ont été sélectionnées. Les alignements multiples ont ensuite été générés à partir d'alignements 2 à 2 de chaque copie avec sa séquence de référence. La méthode d'alignement est dérivée de l'algorithme de Needleman-Wunsch (Needleman et Wunsch 1970). Le programme REFAlign implémenté au laboratoire réalise des alignements globaux, tout en ne pénalisant pas les longs gaps. Les alignements 2 à 2 sont ensuite convertit en un alignement multiple (sans la séquence de référence) par simple juxtaposition. L'annotation des ET a été réalisée avec plusieurs séquences de référence par famille (une séquence par méthode d'annotation). Nous avons donc créé un alignement multiple par famille et par séquence de référence. Les différents alignements multiples (ou profils) pour une famille sont ensuite alignés les uns aux autres à l'aide du programme ClustalW (Thompson *et al.* 1994).

Trois grandes méthodes de reconstruction d'arbres phylogénétiques sur des données de séquences nucléotidiques ou protéiques existent: (i) les méthodes de parcimonie, (ii) les méthodes de distances, (iii) et les méthodes de maximum de vraisemblance. Nous avons choisi, dans notre cas, d'utiliser la méthode de maximum de vraisemblance. Cette méthode a l'avantage d'être très fiable malgré des temps de calcul plus importants.

Pour la reconstruction des arbres, nous avons choisi d'utiliser le programme PhyML (Guindon et Gascuel 2003). Des études comparatives utilisant plusieurs simulations, entre PhyML et d'autres programmes de reconstruction d'arbres, ont souligné ses performances en ce qui concerne la topologie de l'arbre et sa rapidité d'exécution.

J'ai ainsi calculé la distance évolutive entre les copies à partir de l'arbre obtenu avec PhyML. Le modèle de substitution nucléotidique utilisé est le HKY85 (Hasegawa *et al.* 1985). Ce modèle, à l'opposé du modèle de Kimura à 2 paramètres, a l'avantage de tenir compte des différences de fréquences de mutation pour les 4 bases. La valeur du rapport « transition/transversion » utilisée est de 4. Plusieurs algorithmes permettent d'inférer une phylogénie par maximum de vraisemblances. PhyML utilise une heuristique gloutonne qui permet d'inférer rapidement une phylogénie « proche » de la phylogénie optimale. L'algorithme d'agglomération considère une phylogénie initiale irrésolue. Différentes topologies de l'arbre initial sont possibles. J'ai utilisé l'approche par défaut de PhyML qui utilise un arbre construit sur les distances phylogénétiques avec la méthode BIONJ (Gascuel 1997).

Nous avons converti les valeurs de divergence obtenues en millions d'années. Pour cela, nous avons divisé la valeur de la divergence (mutations par branche et par site) par le taux de substitution. Cette valeur a été estimée pour les régions intergéniques de *A. thaliana* à 1.05×10^{-8} substitutions par site et par an (DeRose-Wilson et Gaut 2007).

L'approche présentée, comme la grande part des méthodes de reconstruction d'histoire de duplication, ne reflète pas forcement l'histoire évolutive « vraie » des répétitions. En effet, cette méthode ne peut pas s'appliquer dans le cas des RT. Elle ne tient pas compte de l'organisation des copies de répétitions. Elle ne considère pas non plus les évènements de duplication multiple, c'est-à-dire lorsqu'une séquence est à l'origine d'un ensemble de copies et cela en un seul évènement. Cette approche n'a donc pas pu être utilisée pour la reconstruction d'arbre de duplications en tandem.

4.2.3. DTscore, pour les arbres de duplication en tandem

Le problème posé par les duplications en tandem a été introduit par Fitch en 1977 (Fitch 1977). En effet, de nombreuses études soutiennent le modèle de recombinaison inégale comme le mécanisme de formation des duplications en tandem. Ce mécanisme apparaît comme prédominant pour ce type de répétitions (Fitch 1977; 2002; Elemento *et al.* 2002). Les copies se propagent de manière contiguë les unes par rapport aux autres. Un évènement de duplication peut en un seul évènement, de plus, conduire à la formation de plusieurs autres copies. Pour retracer l'histoire évolutive de ces répétitions, il faut donc tenir compte de l'ordre des répétitions et également des évènements de duplications « multiples ».

Nous avons choisi une méthode d'inférence basée sur une méthode de distance qui suit un schéma agglomératif glouton. Nous avons utilisé le programme DTscore (Elemento et Gascuel 2002). Comme son nom l'indique, l'algorithme de ce programme est basé sur une méthode de score. Le critère de score consiste à comptabiliser le nombre de fois où une paire de feuilles est considérée comme une paire externe. Le programme DTscore requiert en plus d'une matrice de distance, un fichier fournissant l'ordre des copies sur la séquence génomique.

Afin d'obtenir de bonnes reconstructions d'arbres, seules les copies de RT bien annotées ont été sélectionnées. A l'aide du programme « Detectandem.py » que j'ai implémenté, nous avons pu identifier les blocs génomiques strictement composés de RT. Pour chacun de ces blocs, j'ai extrait les séquences et sélectionné celles ayant une taille supérieure à 100 pb. Avant chaque alignement multiple, elles ont ensuite été ordonnées. Les alignements multiples ont été réalisés ici encore avec le programme REFAlign (voir plus haut). Les matrices de distance ont été construites à l'aide du programme fdnadist du package EMBOSS (Rice *et al.* 2000). Le modèle évolutif utilisé correspond au modèle de Kimura à 2 paramètres. Chacune des matrices a ensuite été associée à la liste ordonnée des copies *via* le programme DTscore. La visualisation des arbres a été réalisée à l'aide du programme DTdraw (Elemento et Gascuel 2002; 2002). Le programme DTscore ne renvoyant pas les longueurs des branches, j'ai utilisé le programme baseml du package de PAML pour calculer les longueurs moyennes des branches (Yang 1997; 2007). Ces valeurs nous ont permis ainsi d'estimer l'âge moyen de chaque «*clusters*» de RT.

4.3. Méthodes d'estimation des vitesses évolutives

4.3.1. Taux de petites délétions par maximum de vraisemblances

Les petites délétions, c'est-à-dire de moins de 100 pb, provoquent en partie, par leur accumulation, l'élimination des séquences génomiques. Lorsqu'elles ont lieu dans des régions codantes, elles peuvent être à l'origine de l'évolution des génomes *via* la modification ou la suppression de l'expression d'un gène. Après insertion, les tailles des nouvelles copies de répétitions vont tendre à diminuer *via* de petites délétions internes. Il faut noter (même si c'est trivial) que les délétions qui ont lieu aux extrémités des séquences ne seront pas décelables. Pour notre étude, nous avons fait le choix de ne travailler qu'avec les duplications simples d'ET. Nous avons donc sélectionné les « cerises » (un nœud lié à deux feuilles; Figure 36). De cette manière, nous avons pu estimer les taux de délétion avec un jeu de données ne contenant que des évènements simple de délétion. Après duplication, les deux copies subissent des évènements de délétion indépendants. Pour estimer le taux de délétion, nous nous sommes basés sur une méthode proposée par Petrov *et al.* (Petrov *et al.* 1996), puis reprise par Blumenstiel *et al.* (Blumenstiel *et al.* 2002). Petrov *et al.* ont proposé cette méthode pour analyser la vitesse d'élimination des copies d'ET chez la Drosophile. Ils ont montré que la relation existante entre le nombre de délétions et le nombre de substitutions est monotone et positive. Ce qui leur a permis de proposer une estimation du taux de délétions en fonction du taux de substitutions: soit le taux de délétion par substitution. Considérant les délétions comme des évènements rares, on peut modéliser le nombre de délétion par une loi de Poisson.

Soit pour une séquence i comportant n_i délétions:

$$P(X = n_i) = \frac{e^{-\mu_i}\,\mu_i^{n_i}}{n_i!}$$

Où μ_i correspond à l'espérance du nombre moyen de délétions pour la séquence i. La variable μ_i s'exprime de la façon suivante:

$$\mu_i = \lambda \alpha_i t_i$$

Taux de délétion — Taille de la séquence — Nombre de substitutions

L'estimateur du maximum de vraisemblance de λ est alors:

$$\lambda = \frac{\sum_i n_i}{\sum_i \alpha_i t_i}$$

Nous avons calculé l'intervalle de confiance de l'estimateur par la zone de non-rejet du test du rapport de vraisemblance entre la valeur estimée et les valeurs alentours. L'intervalle de confiance de la moyenne des tailles des délétions est estimé par « *bootstrap* ».

A partir de l'estimation du taux et de la taille des délétions, nous pouvons calculer le temps de demi-vie des copies *via* le modèle de décroissance exponentielle suivant:

$$L = L_0 \cdot e^{(-rt)}$$

Taille de la copie — Taille initiale de la copie — Temps

avec $r = ? \cdot s$

Taux de délétion — Taille moyenne des délétions

4.3.2. Pression de sélection due aux gènes

Pour chaque copie de répétitions, le programme « sqlSelectionPresGene.py » interroge une base de données d'annotations (du laboratoire), pour connaître la localisation, l'orientation et la taille des gènes avoisinants. Pour chaque copie, le programme recherche les gènes contigus. Puis, il élimine les cas où les annotations sont chevauchantes. Ensuite, il calcule et trie par ordre croissant les distances entre la copie et les gènes. La première distance correspond donc au gène le plus proche.

4.3.3. Taux de recombinaison homologue non-allélique

Pour estimer le taux de recombinaison ectopique par domaine chromatinien, nous avons utilisé comme marqueur de recombinaison les retro-éléments à LTR. Cette approche consiste à comptabiliser le produit de la recombinaison entre deux LTR: les solo-LTR. On suppose ici que le taux de recombinaison ectopique est directement corrélé à la densité en solo-LTR. Afin d'avoir une idée du taux de recombinaison dans les domaines euchromatiques et hétérochromatiques, j'ai donc estimé le taux de solo-LTR dans les différents domaines chromatinien.

Le programme « estimRateTypesLTR.py » lance le programme « TRSearch.py », implémenté au laboratoire, sur les séquences de références des retro-éléments à LTR, c'est-à-dire les éléments dits « de pleine longueur ». Le programme TRSearch.py aligne les copies d'éléments en favorisant l'alignement des régions terminales. De cette manière, ce programme permet de repositionner les LTR sur les séquences de références. Ensuite le programme estimRateTypesLTR.py identifie la présence de LTR sur chaque copie, grâce à son alignement avec la séquence de référence obtenue lors de l'annotation. J'ai pu ainsi classer les copies en fonction de leurs structures. Quatre classes de copies peuvent être mises en évidence: (1) copies composées de deux LTR contigus et séparées par une Région Interne (RI); (2) copies avec un LTR connecté à une RI; (3) copies uniquement composées par la RI; (4) copies correspondant à des solo-LTR.

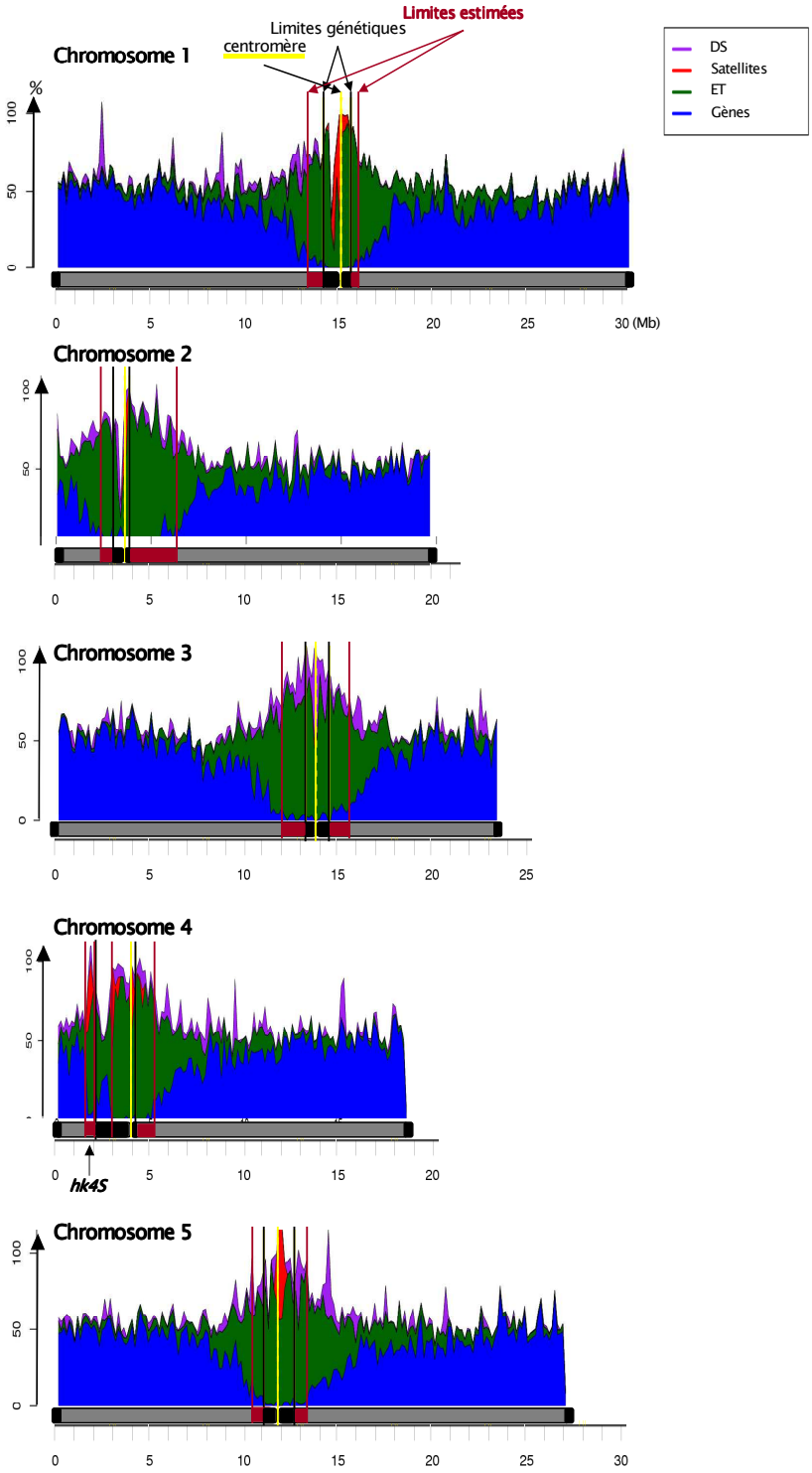


Figure 37. Recouvrement en séquences des chromosomes de *A. thaliana*.

Pour chaque chromosome, la distribution cumulée en séquences (gènes, ET, DS et satellites) est représentée (fenêtres de 150 kb chevauchantes de 5 kb). Le chromosome est schématisé en abscisse. Les limites génétiques et les limites que nous avons estimées sont respectivement indiquées par les traits noirs et rouges. Le trait jaune indique la position du centromère. Les valeurs utilisées tiennent compte du recouvrement entre les séquences: les DS géniques (recouvrement en gène > 95 %) sont comptabilisées comme gènes et non comme duplications. Les régions hétérochromatiques péricentromériques apparaissent enrichies en ET et en satellites alors que les régions euchromatiques sont principalement composées de gènes.

4.4. Les répétitions chez *A. thaliana*

4.4.1. Leur distribution chromosomique

Chez *A. thaliana*, 18.50 % du génome se compose d'ET (23.2 Mb/125.4 Mb), 18.90 % de DS (23.7 Mb/125.4 Mb) et au moins 0.68 % de satellites (0.85 Mb/125.4 Mb). Les répétitions sont majoritairement localisées dans les régions hétérochromatiques (Figure 37; Tableau 6). En effet, alors que l'euchromatine apparaît riche en gènes avec une moyenne de 40.88 % de recouvrement, les répétitions y représentent en moyenne moins de 20 %. A l'opposé dans l'hétérochromatine péricentromérique, ce sont les ET qui représentent les entités majoritaires, avec un recouvrement moyen de 70.67 %. Tandis que les satellites représentent 5.74 % en moyenne (Tableau 6). Au niveau du *knob* du chromosome 4, on observe également une forte densité en ET (64.88 %). Les satellites sont près de 4 fois plus représentés que dans l'hétérochromatine péricentromérique (19.98 %; Tableau 6). Nous n'observons pas de biais important de localisation pour les DS.

domaines chromatiniens	Eu				Hp				knob			
le type de séquence	gènes	DS	TE	satellites	gènes	DS	TE	satellites	gènes	DS	TE	satellites
chromosome 1	41,05	2,40	12,67	0,02	2,28	1,90	67,28	9,93				
chromosome 2	39,84	3,44	15,81	0,01	5,68	6,07	67,84	1,73				
chromosome 3	41,80	3,41	14,07	0,01	3,10	3,56	72,38	0,50				
chromosome 4	40,83	4,80	14,38	0,27	2,33	5,36	74,86	7,50	4,55	6,47	64,88	19,98
chromosome 5	40.71	5,72	13,63	0,02	2,37	5,24	71,00	9,05				
Tous les chromosomes	40,88	3,95	14,11	0,07	3,90	4,43	70,67	5,74	4,55	6,47	64,88	19,98

Tableau 6. Recouvrement des chromosomes en séquences.

Les valeurs indiquées correspondent au pourcentage de recouvrement moyen par chromosome et pour les différents domaines chromatiniens. Ces valeurs ne tiennent pas compte du recouvrement entre séquences. Par exemple, les valeurs de pourcentage de recouvrement en DS ne tiennent pas compte de la composition de cette séquence en ET, satellites ou gènes. Dans ce cas, une copie de duplication génique sera comptabilisée deux fois. Afin d'éviter ces doublons, les duplications géniques ainsi que celles correspondant à des satellites ont été retirées du jeu des DS. Le *knob* du chromosome 5 n'étant pas distinguable de la région péricentromérique, seul le *hk4S* a été pris en compte comme *knob*.

Alors que certaines régions en sont complètement dépourvues, d'autres régions euchromatiques et hétérochromatiques présentent de fortes densités en DS (Figure 37). Le profil de leur distribution suggère un effet de « *duplication shadowing* »: un site dense en DS sera plus sensible aux évènements de duplications (Cheng *et al.* 2005; Newman *et al.* 2005). En effet, les grandes régions dupliquées partageant une forte identité de séquence entre elles (>97 %), représentent de bons substrats pour la recombinaison homologue ectopique (Stankiewicz et Lupski 2002).

4.4.2. Caractéristiques des ET

Il a été détecté chez *A. thaliana* 31246 copies d'ET réparties en 318 familles. Ces familles se composent de 1 à 1439 copie(s). La distribution par superfamille apparaît conservée pour les 5 chromosomes: les régions euchromatiques sont plus denses en *Hélitrons*, en LINE et en *MuDR-like*; les régions hétérochromatiques péricentromériques sont plus riches en éléments *Gypsy-like* et en transposons de la superfamille *En-Spm*. La superfamille *Gypsy-like* est environs 5 fois plus représentée dans l'hétérochromatine péricentromérique que dans l'euchromatine (Figure 38). Cette observation rejoint les observations de Pereira *et al.* (Pereira 2004). En effet, leur étude a permis de suggérer une insertion préférentielle des copies des éléments de la superfamille *Gypsy-like* dans l'hétérochromatine chez *A. thaliana*. Cette superfamille a de plus, la particularité d'être très récente dans ce génome. Excepté pour cette famille, les proportions en ET peuvent être soit la conséquence d'un biais d'insertion, soit celle d'un biais d'élimination. Au vue du statut particulier de la superfamille *Gypsy-like*, certaines analyses seront réalisées sans les copies de cette superfamille.

La distribution par superfamille pour le *knob* du chromosome 4 révèle un profil intermédiaire, entre celui de l'euchromatine et celui de l'hétérochromatine péricentromérique. Les superfamilles sur-représentées dans l'euchromatine et l'hétérochromatine sont également sur-représentées dans le *hk4S* (*Helitron*, LINE, *Gypsy-like et MuDR-like*; Figure 38).

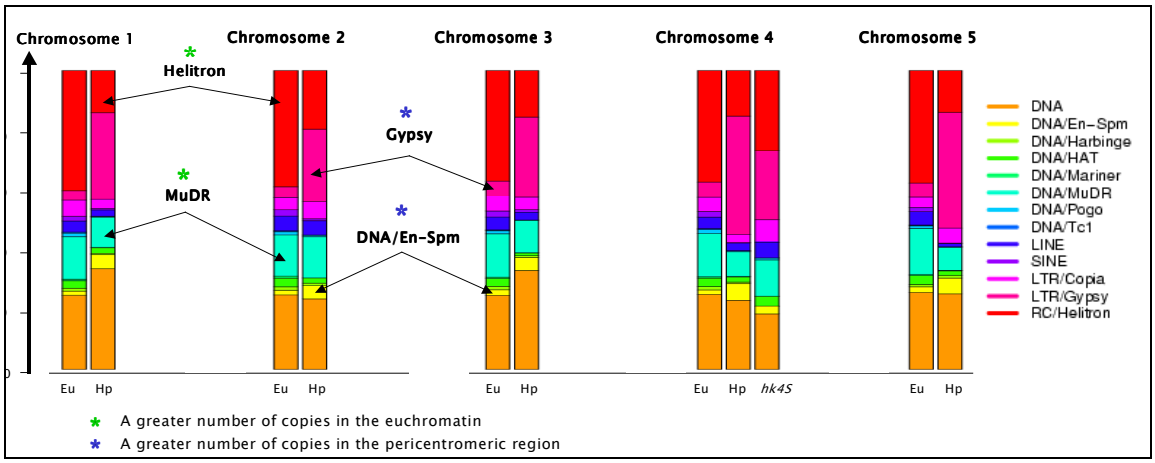


Figure 38. Proportion des ET par superfamilles.

Pour chaque chromosome, la proportion en ET est représentée par domaine chromatinien. « Eu » indique l'euchromatine; « Hp » correspond à l'hétérochromatine péricentromérique et « *hk4S* », le *knob* du chromosome 4. Pour les 5 chromosomes, les régions euchromatiques présentent une forte proportion en éléments *Helitrons* et *MuDR-like* alors que pour les régions péricentromériques, ce sont les éléments *Gypsy-like* qui apparaissent comme les éléments les plus représentés. Le profil de distribution pour le *knob* du chromosome 4 est à l'intermédiaire entre celui de l'euchromatine et celui de l'hétérochromatine.

La taille des copies varie de 9 pb à 31 kb avec une moyenne de 795 pb et une médiane à 306 pb. Les ¾ des copies annotées correspondent à des répétitions de taille inférieure à 772 pb, ce qui indique des copies fortement délétées. Les superfamilles aux copies canoniques les plus grandes (>10kb) présentent des variations de tailles plus importantes (Figure 39).

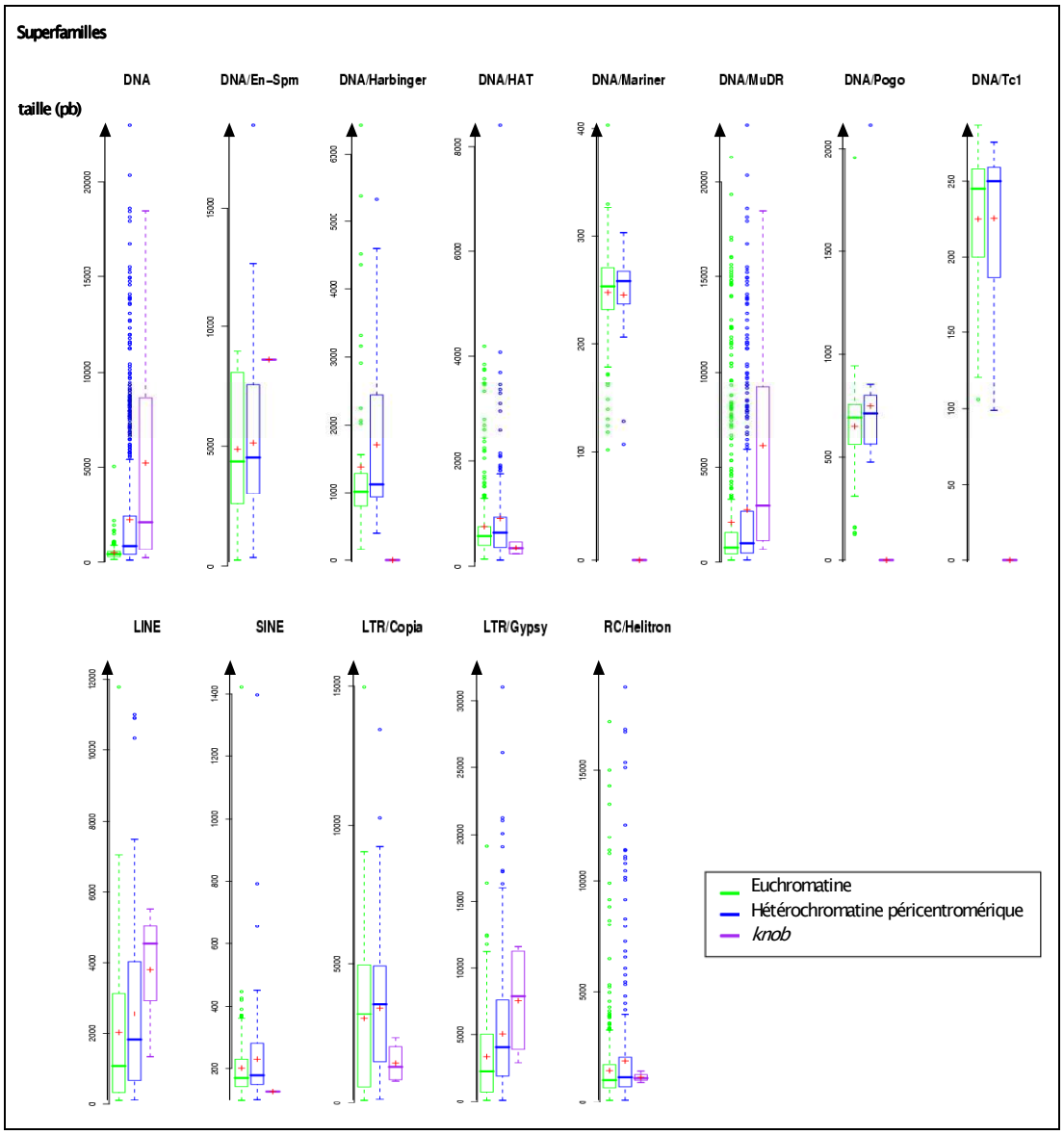


Figure 39. Distribution de la taille des copies.

Pour chaque superfamille, et pour chaque domaine chromatinien, une boîte à moustache indique les valeurs statistiques de la distribution des tailles des copies. Une boîte à moustaches fournit 5 valeurs statistiques: le bord inférieur de la boîte correspond au 1er quartile (Q25), le trait central à la médiane (Q50) et le bord supérieur au 3ème quartile (Q75). Les moustaches inférieures et supérieures délimitent les valeurs adjacentes qui sont déterminées à partir de l'écart interquartile (Q75-Q25). Des valeurs dites « *Outliers* » ou valeurs extrêmes sont représentées par des points de par et d'autre de la boîte à moustaches. Les croix rouges indiquent les moyennes. Les superfamilles les plus représentées (*DNA/En-Spm*, *MuDr-like*, *Gypsy-like*) ont les copies les plus grandes. Les copies de ces superfamilles, localisées dans le *knob*, sont en moyenne plus grandes que celles des régions péricentromériques. Les copies des régions péricentromériques sont en moyenne plus grandes que les copies euchromatiques.

Les copies des régions hétérochromatiques apparaissent en moyenne significativement plus grandes que celles de l'euchromatine (Pour toutes les familles, taille moyenne de 627 pb dans l'euchromatine contre 1283 pb dans l'hétérochromatine péricentromérique; Test de Student: t=-31.52, ddl=31873, P-value<$2.2x10^{-16}$). Les superfamilles *DNA/En-Spm, MuDr-like*, LINE et rétroéléments à LTR ont les copies les plus grandes. Sans les copies de ces superfamilles, la taille des copies apparaît effectivement plus petite. La taille moyenne des copies de l'euchromatine passe de 627 pb à 513 pb et celle de l'hétérochromatine péricentromérique passe de 1283 pb à 620 pb. La différence de taille entre les copies de l'euchromatine et de l'hétérochromatine péricentromérique, largement réduite, reste tout de même notable (Test de Student sans la superfamille *Gypsy-like:* t=-16.41, ddl=27173, P-value<$2.2x10^{-16}$; Sans la superfamille *MuDr-like*: t=-31.58, ddl=25939, P-value<$2.2x10^{-16}$; Sans les copies de toutes les superfamilles citées dans cette partie: t=-6.1958, ddl=17552, P-value=$5.93x10^{-10}$).

Dans le *knob*, seules certaines familles sont représentées. Les tailles des copies de *hk4S* ne sont pas en moyenne significativement plus grandes que celles de l'hétérochromatine péricentromérique (Test de Student: t=1.7716, ddl=7902, P-value=0.0765). Pour la superfamille *Gypsy-like*, la taille moyenne des copies passe de 1884 pb dans l'hétérochromatine péricentromérique à 2474 pb dans le *knob*. Pour la famille *DNA/MuDr-like*, leur taille passe de 1220 pb à 3246 pb. Sans les copies des superfamilles citées plus haut, la taille moyenne des copies du *knob* est divisée par trois. Les copies de ces superfamilles semblent donc biaiser l'observation. Ces copies se seraient insérées très récemment, après la formation du *knob*. Ceci permet d'expliquer la présence et la taille des copies de ces superfamilles dans ce domaine. Il reste pourtant à expliquer les différences de taille observées pour l'ensemble des copies dans l'euchromatine et l'hétérochromatine. On peut proposer soit une dynamique d'insertion plus importante dans l'hétérochromatine péricentromérique, soit une dynamique d'élimination plus importante dans l'euchromatine.

4.4.3. Caractéristiques des DS

A l'issu du pipeline, j'ai obtenu un jeu de données composé de 1930 séquences dupliquées réparties en 804 groupes. Les chromosomes 3 et 5 apparaissent enrichis en DS en comparaison avec les autres chromosomes (Tableau 7).

chromosome	domaines chromatiniens			total	densité	taille des chromosomes (Mb)
	euchromatine	heterochromatine	*knob*			
1	214	17	0	231	11,73	19,7
2	246	88	0	334	10,99	30,4
3	298	56	0	354	15,13	23,4
4	170	23	6	193	10,38	18,6
5	316	44	0	360	13,33	27

Tableau 7. Répartition et densité chromosomique des DS par domaine.

Le tableau indique, pour chaque chromosome, le nombre de copies par domaine chromatinien, la densité chromosomique (en copie/Mb) et la taille du chromosome (en Mb).

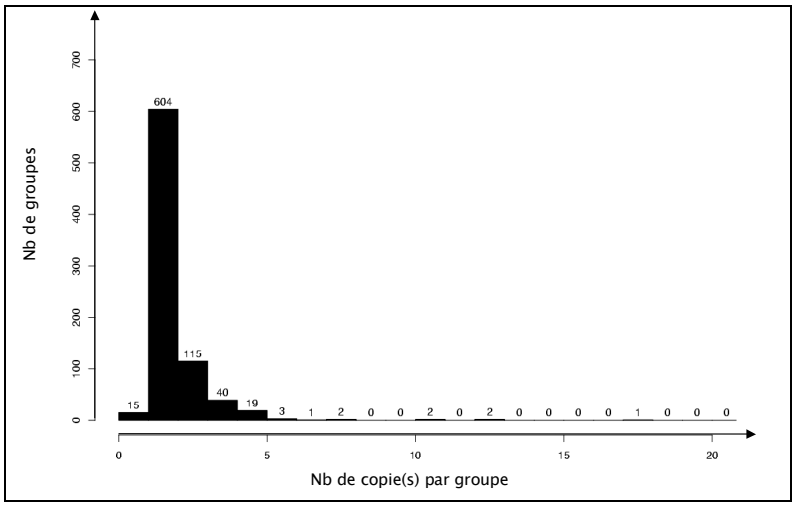


Figure 40. Nombre de copies par groupe de séquences dupliquées.

Les valeurs indiquées au-dessus de chaque barre correspondent aux nombres de groupes. Quinze groupes ne contiennent qu'une seule séquence. Les duplications correspondent dans la majorité des cas à des évènements uniques (604/804 groupes à 2 copies).

Le nombre de copies de duplication varie de 2 à 18 (Figure 40). Quinze groupes avec une seule séquence ont été identifiés (Figure 40). Ces groupes correspondent à des régions riches en duplications, contiguës voire en tandem. Ces régions sont détectées comme une seule et même séquence. En effet, la proximité ou l'emboîtement des copies induit durant l'étape de détection des DS, la connexion des différentes copies et leur regroupement en une seule séquence. Les duplications détectées correspondent dans la

majorité des cas à des évènements uniques (duplications à 2 copies). Les DS, en accord avec leur définition, sont majoritairement en faible nombre de copies (entre 2 et 5 copies; Figure 40).

Plus des ¾ des copies de duplication partagent une identité de séquence supérieure à 90 % (Figure 41). Le profil de distribution est le même pour les DS en faible nombre de copies (de 2 à 5 copies). Les DS détectées chez *A. thaliana* sont donc soit très récentes, soit bien conservées.

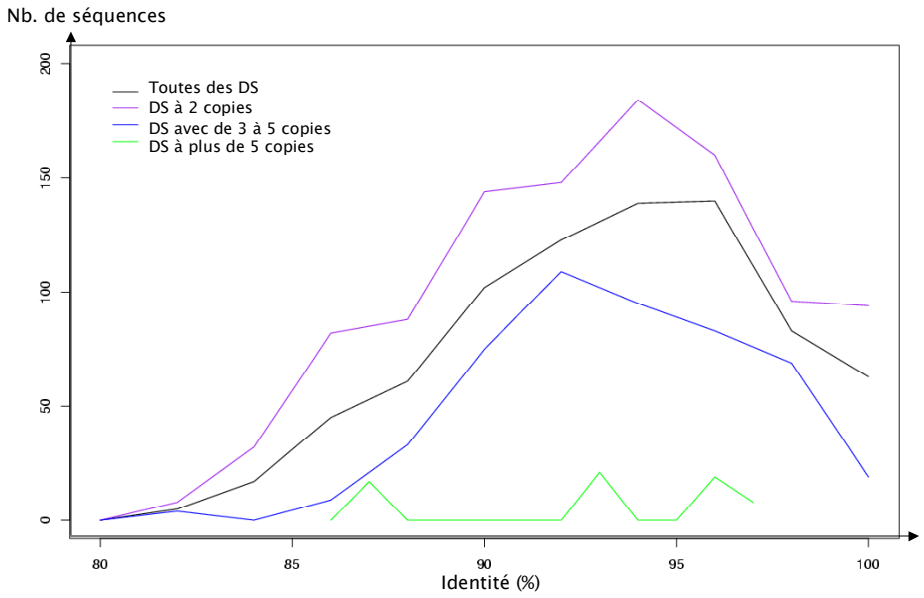


Figure 41. **Pourcentage d'identité des duplications détectées.**

Le pourcentage d'identité des DS varie entre 80 et 100 %. Chaque courbe correspond aux différentes classes de duplications déterminées d'après le nombre de copies par groupe. Le profil de distribution est le même pour toutes les types de DS. La médiane se situe aux alentours de 95 % d'identité.

La taille des séquences varie de 340 pb à 187.3 kb. La copie de 187.3 kb est un artéfact de l'étape de connexion. En la retirant du jeu de données, la taille maximale atteint presque 10 kb. Plus de 75 % des DS ont une taille inférieure à 1.8 kb (Figure 42). Les duplications au plus grand nombre de copies s'avèrent plus grandes que celles des évènements uniques (Test de Student: t=-2.15, ddl =1313, P-value=3.21×10^{-3}; Figure 42).

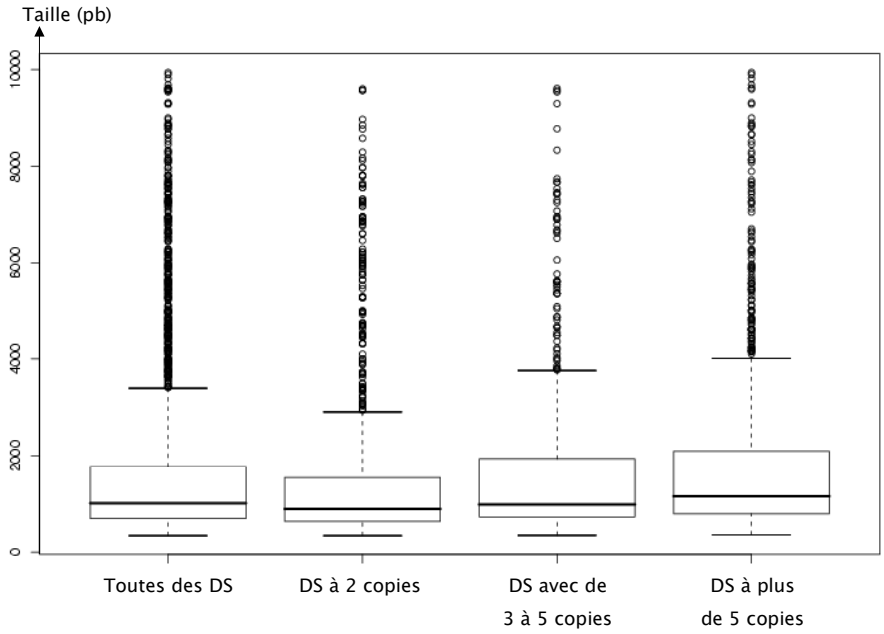


Figure 42. Taille des duplications détectées selon le nombre de copies.

Pour chaque classe, une boîte à moustache représente la distribution de taille des copies. Les valeurs statistiques sont: le minimum, le 1er quartile, la médiane, le 3ème quartile et le maximum. Les duplications aux plus grand nombre de copies (plus de 5 copies) ont une taille médiane plus grande que pour les duplications en faible nombre de copies (de 2 à 5 copies). La plus grande duplication de 187.3 kb a été retirée pour une meilleure visualisation graphique.

Pour les groupes à 2 copies, 45.70 % des duplications (276 copies) sont intrachromosomiques contre 54.30 % (328 copies) interchromosomiques. La taille moyenne des duplications intrachromosomiques est significativement plus grande que celle des duplications interchromosomiques (Test de Student: t=3.56, ddl=1206, P-value=3.85x10^{-4}). Dans les ¾ des cas, les duplications intrachromosomiques sont distantes de plus de 100 kb (avec une moyenne de 1.22 Mb).

Le génome d'*A. thaliana* aurait subi, au cours de son évolution, plusieurs duplications massives (AGI 2000; Henry *et al.* 2006). L'observation d'une forte proportion en duplications géniques en comparaison avec celle d'autres génomes eucaryotes comme *D. melanogaster*, est sans doute la conséquence de ces évènements. En effet, plus de 75 % des duplications se composent partiellement ou totalement de régions géniques chez *A. thaliana*. Chez *D. melanogaster*, cette proportion n'atteint que 38 % (Fiston-Lavier *et al.* 2007). L'histoire évolutive de ce génome peut ainsi expliquer la plus grande densité en DS dans les régions euchromatiques (Tableau 7).

Composition: *moyenne (3rd Quartile)*		Toutes des DS	DS à 2 copies	DS avec de 3 à 5 copies	DS à plus de 5 copies	Tests de Student (séquences détectées vs séquences contrôles)
Genes	séquences détectées	36.79 (78.61)	41.21 (84.72)	30.36 (68.01)	24.36 (36.68)	t = -13.6971, ddl = 194928, P-value < 2.2e-16
	séquences contrôles	27.10 (50.68)				
ET	séquences détectées	15.71 (18.72)	12.98 (5.09)	20.20 (35.73)	19.75 (25.02)	t = 10.0657, ddl = 194928, P-value < 2.2e-16
	séquences contrôles	24.22 (42.40)				
satellites	séquences détectées	10.58 (13.44)	9.15 (11.72)	12.09 (15.29)	18.14 (25.78)	t=4.844, ddl=194928, P-value=1.274e-06
	séquences contrôles	11.58 (15.36)				

Tableau 8. La composition des DS détectées.

La fraction recouverte en gènes, ET et satellites a été calculée pour les DS et les séquences contrôles. Ce calcul a été réalisé pour l'ensemble des DS, ainsi que par groupe de DS (d'après le nombre de copies par groupe). La dernière colonne indique les résultats des tests de Student comparant les fractions moyennes de l'ensemble des DS avec celles des séquences contrôles. Pour chaque calcul, la valeur moyenne est donnée. Les valeurs entre parenthèse correspondent aux 3ème Quartiles.

En moyenne près de 40 % de la séquence des duplications uniques (duplications à 2 copies) est composée de régions géniques et environs 30 % pour les séquences des duplications ayant entre 3 à 5 copies (Tableau 8). De nombreuses régions géniques de ce génome sont impliquées dans des évènements de duplication. A l'opposé de *D. melanogaster*, les DS ne sont donc pas enrichies en ET ou en satellites.

4.4.4. Caractéristiques des RT

Parmi les DS, nous avons pu identifier des duplications en tandem. Ces répétitions contiguës sont détectées comme un seul bloc, c'est-à-dire comme un groupe avec une seule séquence. Malheureusement, les répétitions détectées ne correspondent pas à l'ensemble des RT du génome. Afin d'avoir une vue plus globale, nous avons donc choisi d'utiliser comme jeu de données, les annotations des satellites obtenues avec les séquences consensus présentes dans RU (voir partie annotation).

Ce jeu de données contient 3348 séquences représentant 8 familles de satellites: la famille AR12 étant la plus représentée avec 2636 copies et la famille MI167_AT, la moins représentée avec 3 copies. Ces séquences ont comme principale caractéristique d'être préférentiellement localisées dans les régions hétérochromatiques. Seules 13.81 % des séquences sont dans l'euchromatine (Figure 37). Cette répartition peut expliquer la plus forte densité observée au niveau des chromosomes 4 et 5, qui présentent tous deux des régions hétérochromatiques surnuméraires: les *knobs*. Leurs densités en RT sont respectivement 47.42 et 42 RT/Mb contre 36.1, 16.61 et 4.95 RT/Mb pour les chromosomes 1, 2 et 3.

La taille des copies annotées varie de 19 pb à 28 kb, avec une médiane à 176 pb. Les ¾ des unités de répétitions sont inférieures à 177 pb. Les quelques grandes copies détectées (> 177 pb) correspondent à de grands « *clusters* » de RT. La taille des copies des familles AR12 et ATMSAT1 sont très conservées. Elles sont respectivement de 96 et 177 pb (Figure 43). La taille des RT est donc généralement plus petite que celle des DS (taille moyenne de 2187 pb; t-test: t=-18.2166, ddl=5276, P-value<$2.2x10^{-16}$) et des ET (taille moyenne de 795 pb; t-test: t=-19.0448, ddl=34592,P-value<$2.2x10^{-16}$).

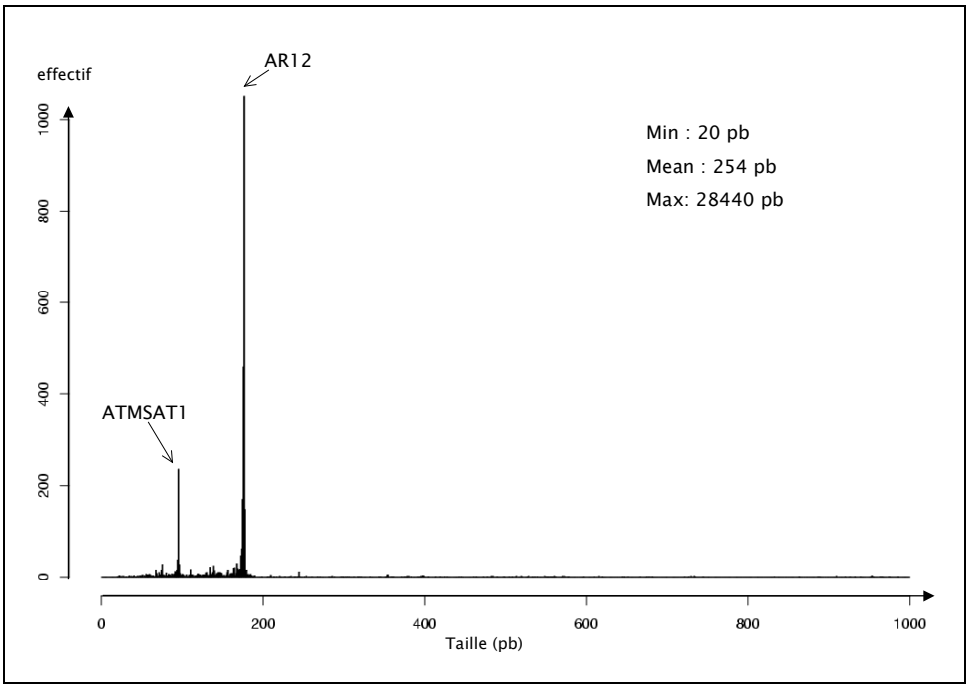


Figure 43. Distribution de la taille des RT.

La distribution des tailles des RT permet d'observer la conservation de taille entre les unités d'une même famille. Les tailles des copies des familles AR12 et ATMSAT1 sont très conservées. Elles sont respectivement de 96 et 177 pb. On les identifie par les deux plus grands pics. La fenêtre utilisée pour ce graphique est de 10 pb.

Seules des copies de la famille AR12 et ATCLUST1 sont présentent dans l'euchromatine. La taille moyenne des copies de la famille AR12 passe de 1293 pb dans l'hétérochromatine à 3279 pb dans l'euchromatine. Cette différence significative (Test de Student: t=-2.18, ddl=290, P-value=0.03) s'oppose aux observations faites pour les ET. On n'observe pas de différence significative de taille entre les copies de l'euchromatine et l'hétérochromatine pour les autres RT (Test de Student: t=-0.21, ddl=150, P-value=0.835).

4.5. Dynamique comparée des répétitions

4.5.1. Insertion des ET par vague

L'extraction des longueurs des branches terminales des arbres de transposition a permis d'estimer la divergence des copies depuis leur dernière activité de transposition. Pour 14 copies seulement, j'obtiens des valeurs de divergence supérieures à 0.60 mutations par branche et par site, dû au morcellement de ces copies et donc à un mauvais alignement. Pour la suite de notre étude, ces copies ont été exclues du jeu de données. La divergence des copies varie de 0 à 0.60 mutations par site et par branche avec une médiane à 0.07 et une moyenne de 0.11. On estime donc l'âge moyen des copies d'ET à 10.6 millions d'années, l'âge médian étant de 16 millions d'années. Les copies de l'euchromatine présentent une divergence moyenne significativement plus forte que celles de l'hétérochromatine (Test de Student: t=2.92, ddl=5027, P-value=3.52x10^{-3}; Figure 44; Figure 45).

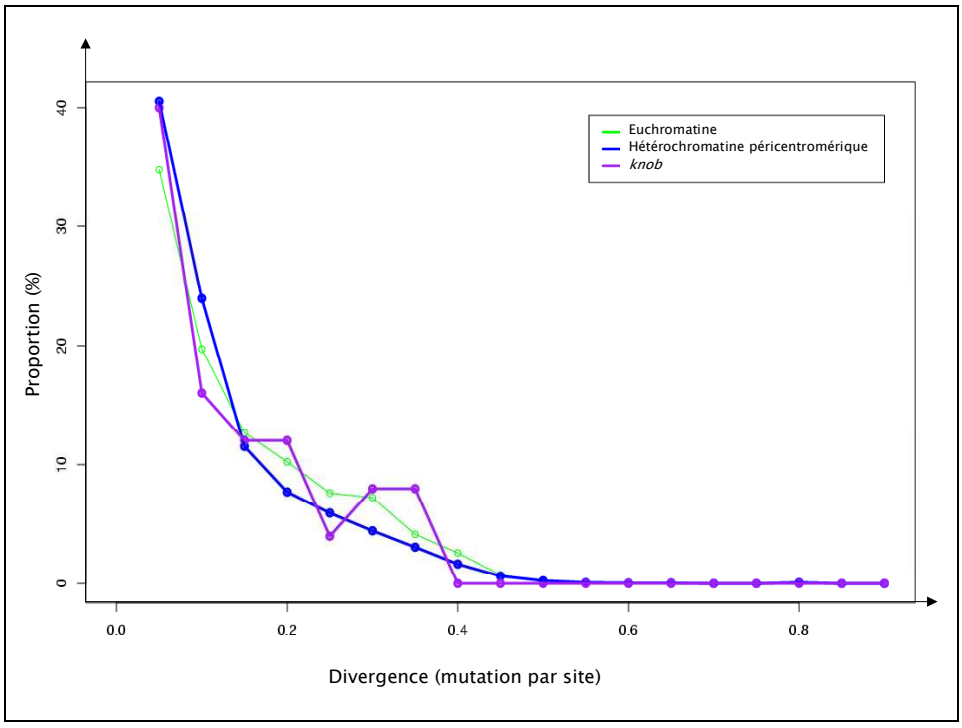


Figure 44. Distribution de la divergence des ET par domaine.

Pour chaque domaine chromatinien, nous avons représenté la proportion de copies par tranche de divergence (tous les 0.05 mutations par site). Pour les différents domaines, la majorité des copies ont une divergence inférieure à 0.2 mutations par site. Les profils pour l'euchromatine et l'hétérochromatine sont quelque peu différents: la proportion de copies ayant une divergence de moins de 0.2 mutations par site est plus importante pour l'hétérochromatine. L'inverse est observé pour les copies ayant une divergence comprise entre 0.2 et 0.5 mutations par site.

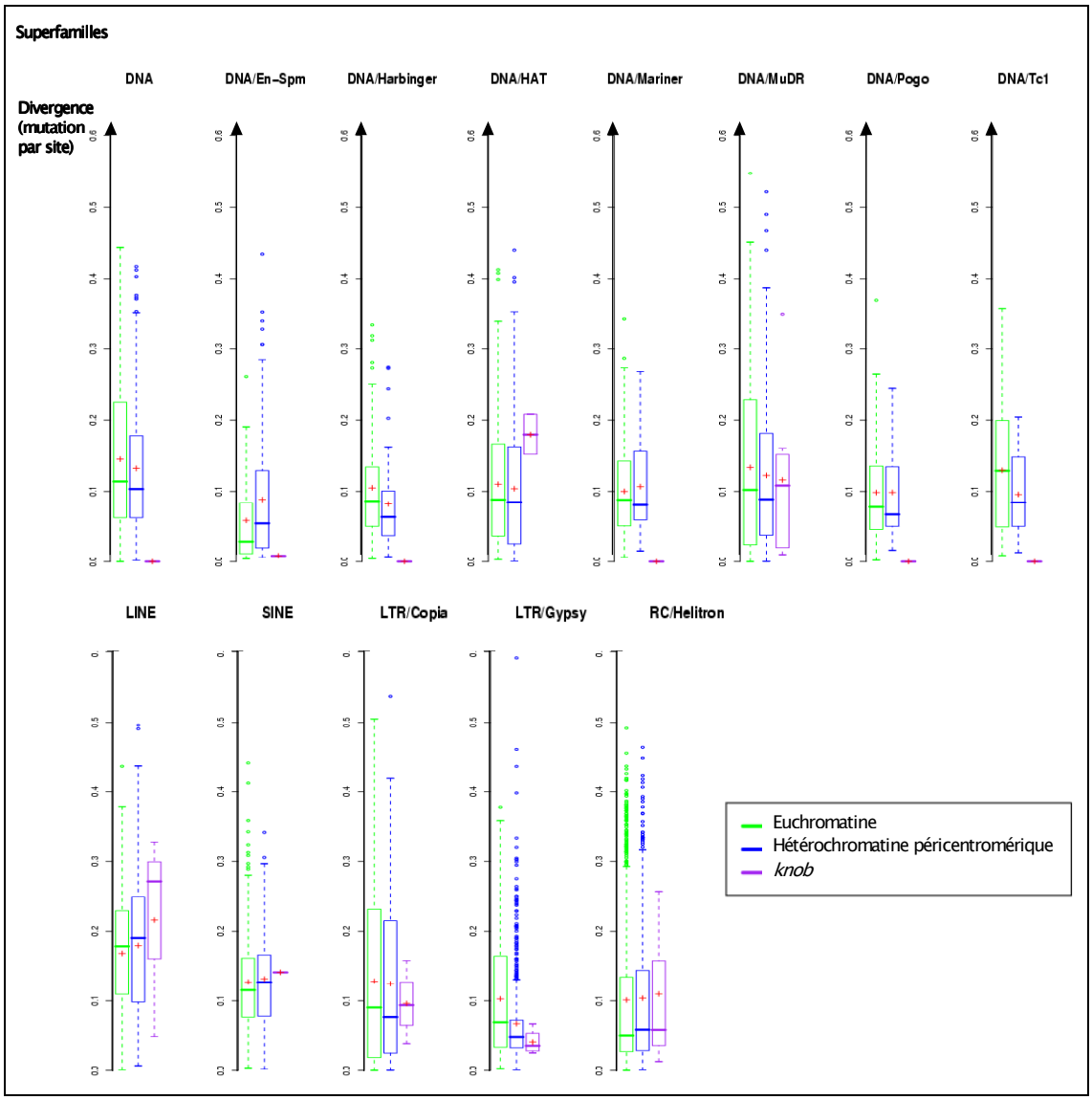


Figure 45. Divergence des ET par superfamille.

Pour chaque superfamille et pour chaque domaine chromatinien, une boîte à moustaches représente la distribution des divergences. Une boîte à moustaches fournit 5 valeurs statistiques: le bord inférieur de la boîte correspond au 1er quartile (Q25), le trait central à la médiane (Q50) et le bord supérieur au 3ème quartile (Q75). Les moustaches inférieur et supérieur délimitent les valeurs adjacentes qui sont déterminées à partir de l'écart interquartile (Q75-Q25). Des valeurs dites « *Outliers* » ou valeurs extrêmes sont représentées par des points de par et d'autre de la boîte à moustaches. Les croix rouges indiquent les moyennes. Pour la majorité des superfamilles, les copies euchromatiques semblent plus divergentes que les copies de l'hétérochromatine.

Lorsque l'on analyse la divergence par superfamille et par domaine (Figure 45), certaines superfamilles apparaissent très jeunes telles que les superfamilles *Gypsy-like* (divergence moyenne de 0.07 mutations par site) et *En-Spm* (divergence moyenne de 0.075 mutations par site). Alors que d'autres superfamilles apparaissent plus ancienne telles que les superfamilles *MuDR-like* (divergence moyenne de 0.20 mutations par site) ou *L1* (divergence moyenne de 0.18 mutations par site). Là encore, le statut hétérochromatique récent du *knob* du chromosome 4 est frappant: les copies ayant une divergence de plus 0.35 mutations par site y sont absentes. En se basant sur la copie la plus vieille, on peut dater l'âge du *knob* à environs 33 millions d'années.

Si l'on analyse la dynamique des familles d'ET, on observe quelques différences. La topologie des arbres de transposition n'est pas la même pour toutes les familles. Or, la topologie de ces arbres reflète la dynamique de transposition de la famille. Cette observation met en évidence une dynamique de propagation propre à chaque famille. Par exemple, pour les éléments de type LINE, de nombreuses études ont montré qu'après l'évènement de transposition, la nouvelle copie est souvent tronquée en 5'. L'ancienne copie reste active et continue ainsi à transposer alors que la nouvelle copie perd alors son activité de transposition. Les arbres de transposition de ces éléments montrent effectivement une topologie en « escalier » (Figure 46b).

Une tendance générale peut tout de même être mise en évidence. En effet, la topologie des arbres est en faveur d'une propagation des ET par vague. La topologie des arbres de transposition montre souvent la propagation d'une copie en un ensemble de copies en un laps de temps très court (branches internes très courtes; Figure 46).

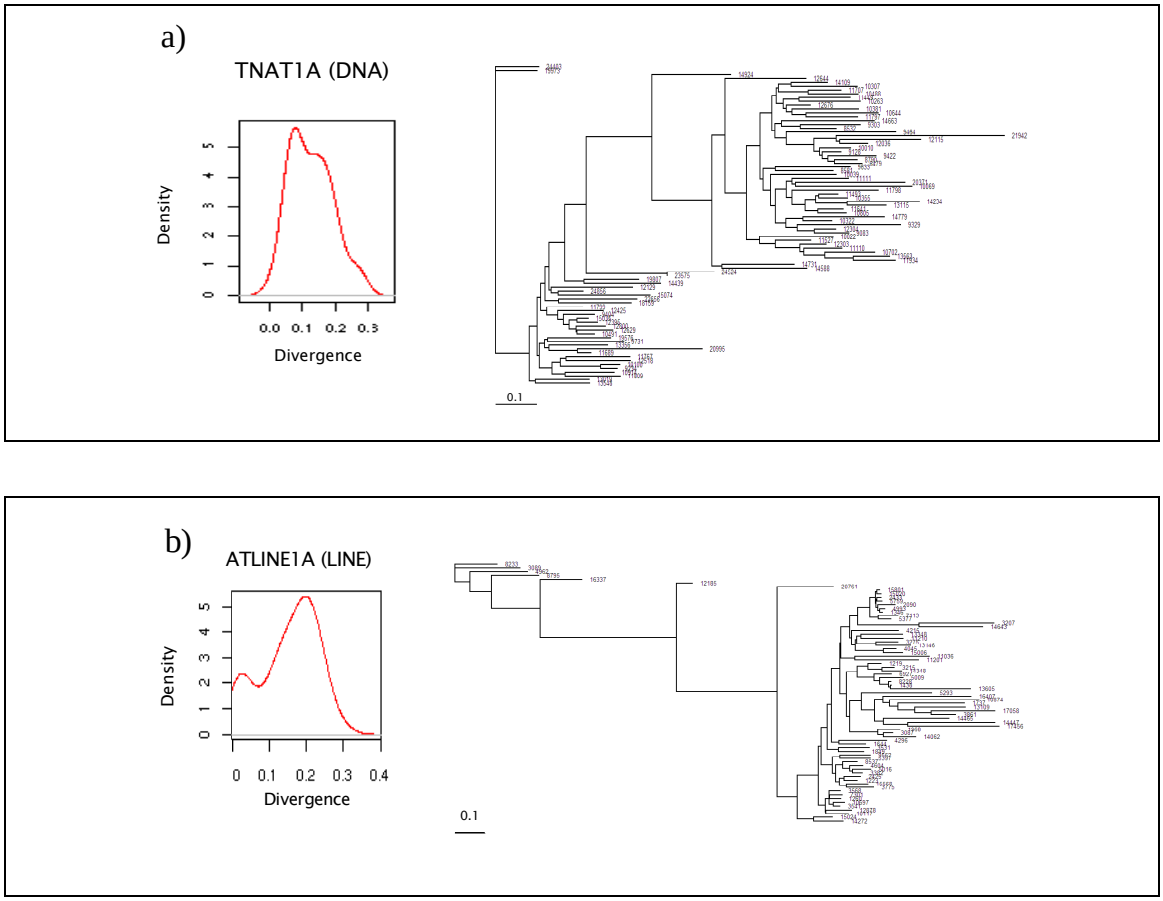


Figure 46. Exemples d'arbres de transposition.

Les graphiques présentés à gauche correspondent aux distributions des divergences de la famille. Les arbres de transposition des familles sont représentés à droite. La topologie des arbres reflète l'histoire évolutive des copies de la famille. **a)** Famille TNAT1A, transposon à ADN. La divergence moyenne de ces copies est de 0.1 mutations par site. Entre les deux vagues de transposition, on observe des évènements de transposition « simple ». **b)** Famille ATLINE1A, rétroélément de la superfamille des LINE. On oberserve clairement sur le graphique, des copies très jeunes (moins de 0.05 mutations par site) et des copies moins jeunes (0.2 mutations par site). Après transposition, la nouvelle copie LINE est souvent tronquée en 5', ce qui la rend inactive. Le profil en « escalier » de l'arbre reflète bien la dynamique de ces éléments.

4.5.2. Dynamique d'insertion des DS

Dans le but de compléter le modèle de dynamique des séquences dans les régions hétérochromatiques, nous avons également étudié la dynamique d'insertion des DS. Pour cela, j'ai utilisé une approche similaire de celle des ET pour créer des arbres de duplications. Pour cette étude, je n'ai pris en compte que les séquences appartenant aux groupes ayant entre 2 et 5 copies. Pour la suite de cette étude, seules les « vraies » DS ont été prises en compte. Ainsi les duplications géniques (DS avec plus de 95 % de recouvrement en gènes), soit 336 DS, n'ont pas été sélectionnées. Les 15 RT (les séquences des groupes à une séquence) ont également été éliminées. La divergence des 1579 copies sélectionnées oscille entre 0 et 0.90 mutations par site, avec une divergence médiane à 0.035 et un 3ème quartile à 0.056 mutations par site. Les ¾ des DS ont donc émergé, il y a moins de 6 millions d'années. Lorsque l'on regarde par domaine chromatinien, les copies de l'hétérochromatine sont plus anciennes que celle de l'euchromatine (Figure 47). La divergence des copies localisées au niveau du *knob* (6 copies) montre que certains évènements de duplications sont postérieurs à la formation du *knob* (divergence maximum de 0.036 mutations par site, soit 3.5 millions d'années). A l'opposé, les duplications géniques sont plus divergentes dans l'euchromatine (Figure 47).

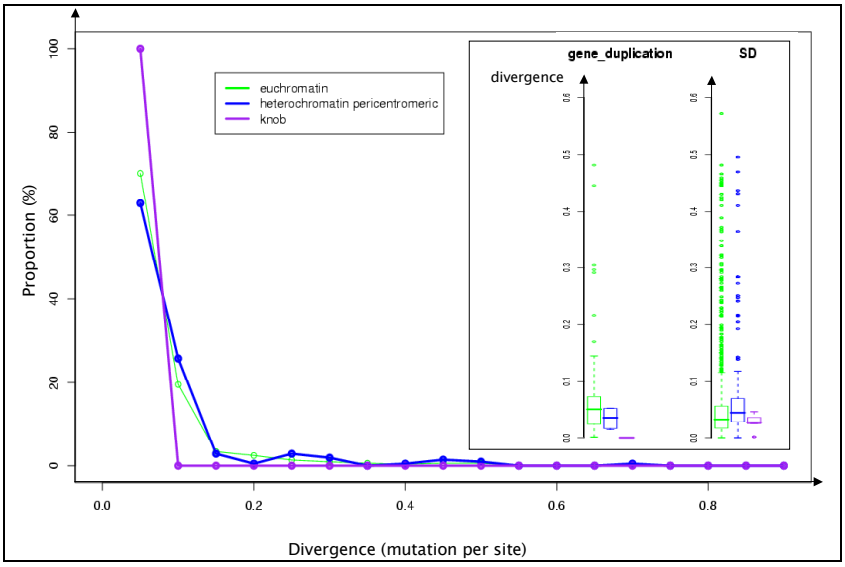


Figure 47. Distribution de la divergence des DS par domaine.

Pour chaque domaine chromatinien, la proportion de copies par tranche de divergence (tous les 0.05 mutations par site) est représentée. La majorité des copies a une divergence inférieure à 0.1 mutations par site. Les copies hétérochromatiques sont plus divergentes que celles de l'euchromatine. Pour les duplications géniques (pourcentage de recouvrement > 95 %), on observe le profil inverse. Les DS du *knob* ont une divergence très faible, inférieur à 0.036 mutations par site.

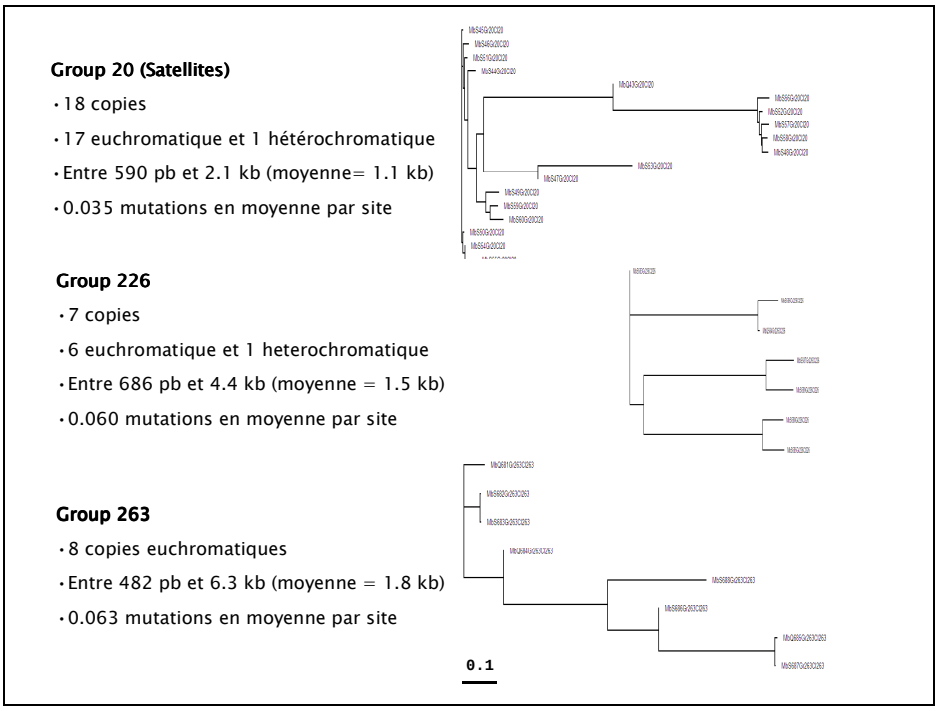


Figure 48. Exemples d'arbres de RTet de DS.

Le groupe 20 correspond à un groupe de satellites alors que les autres groupes (226 et 263) sont des « vraies » DS. Les copies de ces groupes se composent à la fois de régions d'ET, de satellites et de gènes. Les profils des arbres de duplications reflètent bien la dynamique de ces répétitions. En effet, d'après les modèles proposés, l'expansion de satellites peut conduire à une augmentation brusque du nombre de copies. Alors les évènements de duplication conduisent généralement à l'insertion d'une seule nouvelle copie.

Les copies des DS apparaissent globalement plus jeunes que celles des ET. A l'opposé des ET, on n'observe par de propagation par vague (Figure 44; Figure 47). Une copie conduit généralement à la propagation d'une seule autre copie. La nouvelle copie peut être à son tour dupliquée totalement ou partiellement. En effet, la probabilité pour qu'une région composite soit dupliquée complètement deux fois de suite est très faible. Le cas du groupe 20 (Figure 48) correspond à un groupe de minisatellites. La topologie de cette famille de répétitions est proche de celle des ET: une copie donne naissance à un ensemble de copies dans un laps de temps assez court. A l'inverse, les groupes 226 et 263 correspondent à de « vraies » DS: les séquences des copies sont composées à la fois de séquences répétées et de régions géniques. La composition en séquences des DS peut avoir une influence sur l'évolution des copies. Mais nos analyses n'ont pas permis de montrer de manière significative un quelconque impact de la composition, sur la divergence des copies (Régression multiple: longueur des branches terminales ~ composition en gènes + composition en ET + composition en satellites; $F(3,1474)=2.45$, $R^2=0.005$, R^2 ajusté=0.003, P-value=0.06).

4.5.3. Dynamiques d'expansion des RT

La détection des satellites avec les séquences de références de RU a permis d'identifier 138 « *clusters* » de RT chez *A. thaliana*. Un arbre de duplication a ensuite été inféré pour chaque bloc de répétitions *via* le programme DTscore. Le programme BASEML a ensuite permis d'estimer les longueurs des branches et ainsi l'âge moyen de chaque « *cluster* ».

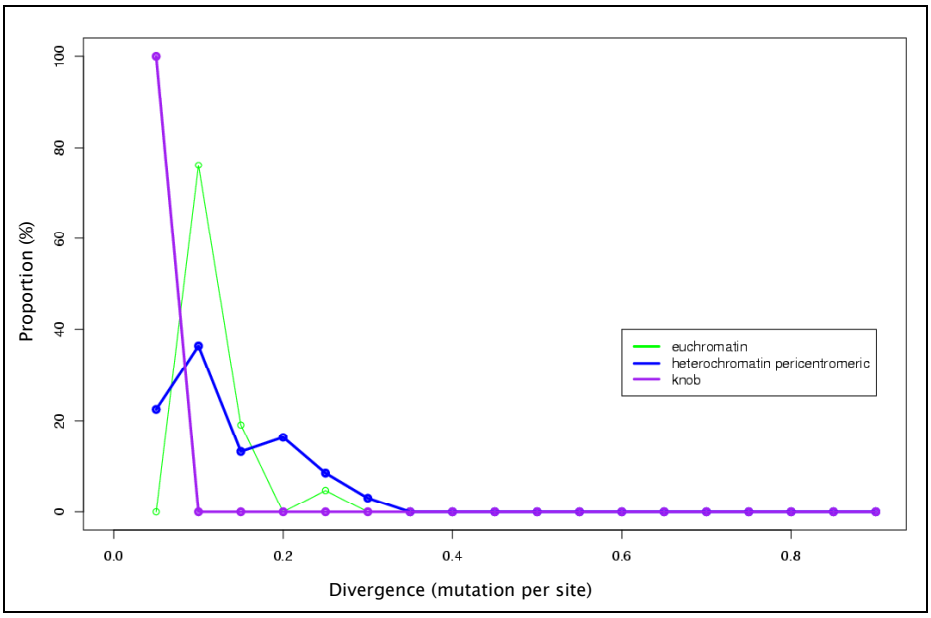


Figure 49. Distribution de la divergence des RT par domaine.

Pour chaque domaine chromatinien, nous avons représenté la proportion de copies par tranche de divergence (tous les 0.05 mutations par site). Pour les différents domaines, la majorité des copies ont une divergence inférieure à 0.4 mutations par site. Les profils pour l'euchromatine et l'hétérochromatine sont quelque peu différents: alors que la majorité des copies de l'euchromatine présente une divergence entre 0.05 et 0.2 mutation par site, les divergences des copies de l'hétérochromatine varient entre 0.05 et 0.35 mutation par site.

La divergence moyenne par «*cluster*» de RT oscille entre 0 et 0.27 mutations par site avec une moyenne à 0.12 mutations par site. Les ¾ des «*clusters*» ont une divergence moyenne inférieure à 0.19 mutations par site. La majorité des régions de RT de ce génome ont émergé, il y a environs 18 millions d'années. Lorsque l'on regarde par domaine, les divergences des copies de l'hétérochromatine et de l'euchromatine ne diffèrent pas significativement (t=-1 .16, ddl=149, P-value=0.248; Figure 49). Plusieurs profils de dynamique sont observables. Si on prend comme exemple un «*cluster*» de la famille *AR12*, le mécanisme de propagation suggère une expansion par duplication en tandem de fragments de tailles très variables (Figure 50). Les fragments étant composés eux-mêmes de RT, ces évènements apparaissent comme un « *cluster* » de RT. A

l'opposé, l'expansion des copies du « *cluster* » *ATENSAT1* s'est réalisé par insertion de nouvelles unités. De plus, la topologie de l'arbre du « *cluster* » reflète une dynamique très récente (Figure 50). Le «*cluster*» de RT localisé dans le *hk4S* serait donc issu d'un évènement très récent puisque l'on peut estimer l'âge de sa plus vieille copie à 2.8 millions d'années.

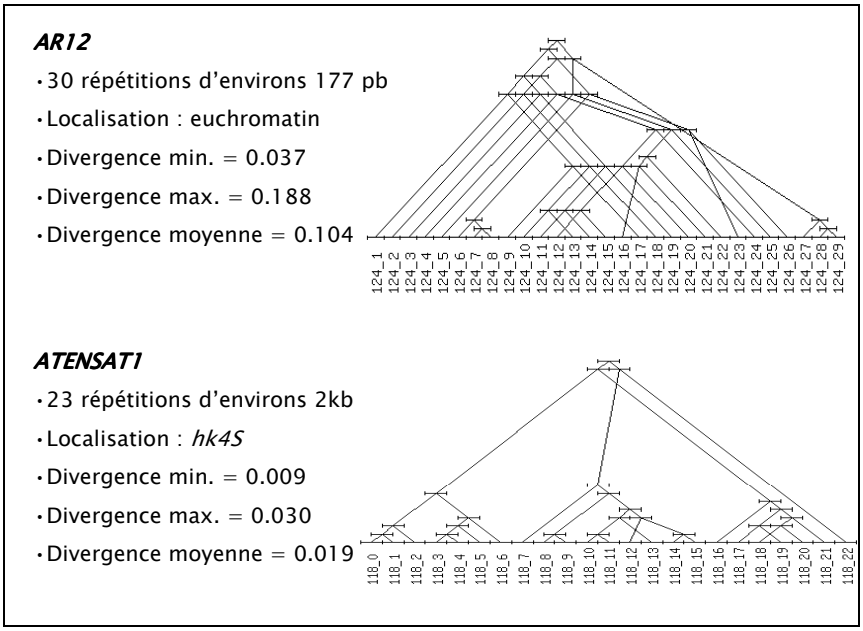


Figure 50. Deux exemples d'arbres de « *clusters* » de RT.

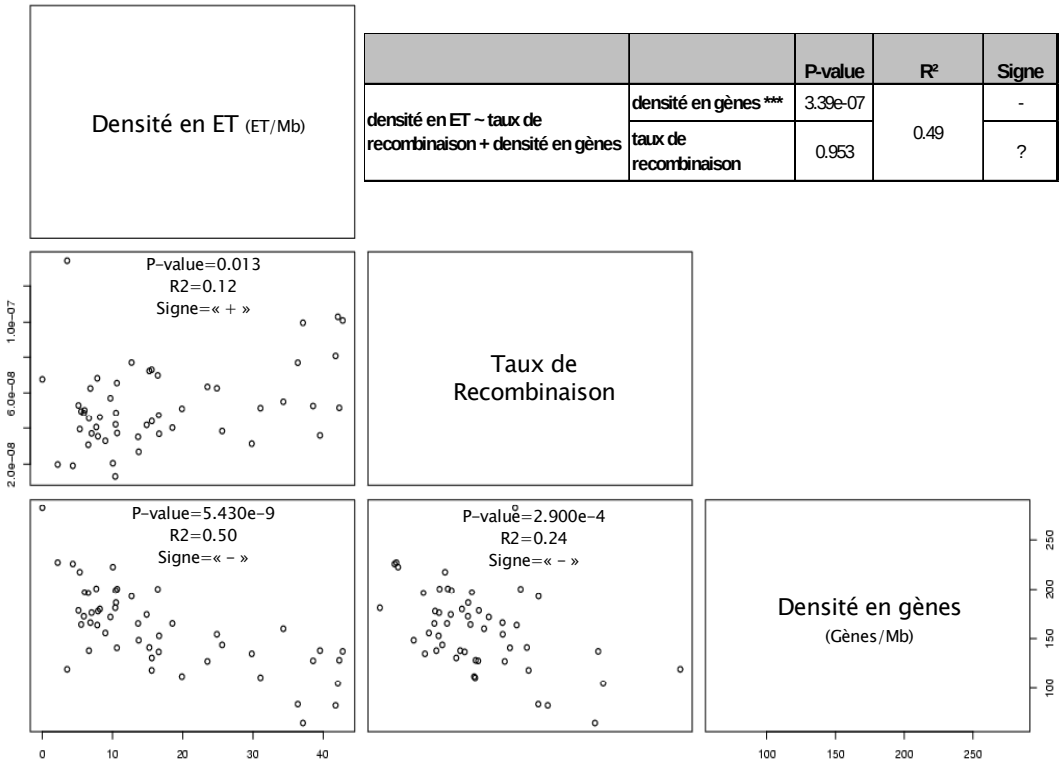
Les deux exemples correspondent à des « *clusters* » situés sur le chromosome 4. Le premier (*AR12*) est localisé dans l'euchromatine alors que le deuxième (*ATENSAT1*) est localisé dans le *knob*. Les arbres de duplication montrent deux profils de dynamique différents. Dans le premier cas, l'expansion a été réalisée par duplication de fragments composés de plusieurs RT. Dans le second cas, l'expansion du « *cluster* » a été réalisée par duplications d'unités de répétition.

4.6. Les forces d'élimination des répétitions

Afin d'expliquer les différences de taille et de dynamique observées pour les copies de répétitions dans les différents domaines chromatiniens, j'ai estimé les forces d'élimination, en utilisant comme unique jeu de données: les ET. Leur fréquence de transposition et leur impact, souvent délétère, nous autorise à dire que ces répétitions sont soumises à une dynamique évolutive plus importante que les DS ou RT. Choisir ces répétitions amène à travailler également avec une échelle de temps plus grande. En effet, la datation des copies de répétitions suggère une insertion plus ancienne des ET. Les DS et RT identifiées sont relativement récentes. Les grands effectifs en copies d'ET nous permettront, de plus, de mieux rendre compte des forces responsables de l'évolution des répétitions.

4.6.1. Effet de la sélection

Une copie insérée à proximité ou dans un gène peut avoir un effet délétère. La sélection naturelle va alors tendre à la faire disparaître. J'ai donc testé l'effet de la sélection sur les insertions de copies potentiellement délétères. J'ai analysé la relation entre la densité en gènes et le taux de recombinaison. La régression multiple n'a été réalisée qu'avec les données concernant le bras long du chromosome 4. En effet, seule la carte fine des taux de recombinaison de cette région euchromatique était disponible (Drouaud *et al.* 2006). Les résultats suggèrent un impact de la présence des gènes sur la densité en ET, et cela de manière significative (Figure 51). Plus il y a de gènes et moins il y a d'ET. Par contre, le taux de recombinaison méiotique ne permet pas d'expliquer la distribution en ET. Seul 24 % de la variation de densité en gènes peut-être expliqué par la recombinaison (Figure 51). On peut donc penser à un effet indirect du taux de recombinaison sur la densité en ET (Figure 51). Wright *et al.* ont déjà montré des résultats similaires (Wright *et al.* 2003).



		P-value	R^2	Signe
densité en ET ~ taux de recombinaison + densité en gènes	densité en gènes ***	3.39e-07	0.49	-
	taux de recombinaison	0.953		?

Figure 51. Relation: densité en ET/densité en gènes/taux de recombinaison.

Cette étude a été réalisée avec les données concernant le bras long du chromosome 4. Les valeurs de densité sont données en séquence par Mb. Elles ont été calculées à partir des annotations en ET obtenues au laboratoire et les annotations en gènes de TAIR 6. Les valeurs des taux de recombinaison méiotique sont issues des données de (Drouaud *et al.* 2006). D'après les résultats, la densité en ET peut s'expliquer de manière significative par la densité en gènes mais pas par le taux de recombinaison. Or le taux de recombinaison explique en partie la distribution des gènes. L'effet du taux de recombinaison sur la distribution en ET semble indirect comme le montre cette analyse.

Afin de mieux caractériser l'effet de la pression de sélection dû aux gènes, nous avons testé le lien entre la taille de la copie et la distance avec le gène le plus proche (Figure 52). A l'aide du programme sqlSelectionPresGene.py, j'ai obtenu pour chaque copie la distance avec le gène le plus proche. Les résultats de la régression suggèrent, de manière significative, un effet de la pression de sélection due aux gènes sur la taille des copies (Figure 52). Ces résultats indiquent que plus une copie est proche d'un gène plus elle est petite. Pour expliquer ces observations, l'hypothèse la plus parcimonieuse est que la pression de sélection du gène le plus proche va tendre à éliminer la copie insérée. On s'attend donc à ce que l'élimination soit plus forte dans les régions riches en gènes, c'est-à-dire dans l'euchromatine.

Les plus petites copies correspondent aux copies les plus anciennes (Figure 52). Ceci permet d'expliquer leur présence dans l'euchromatine. Certaines de ces copies ont pu se fixer dans la population.

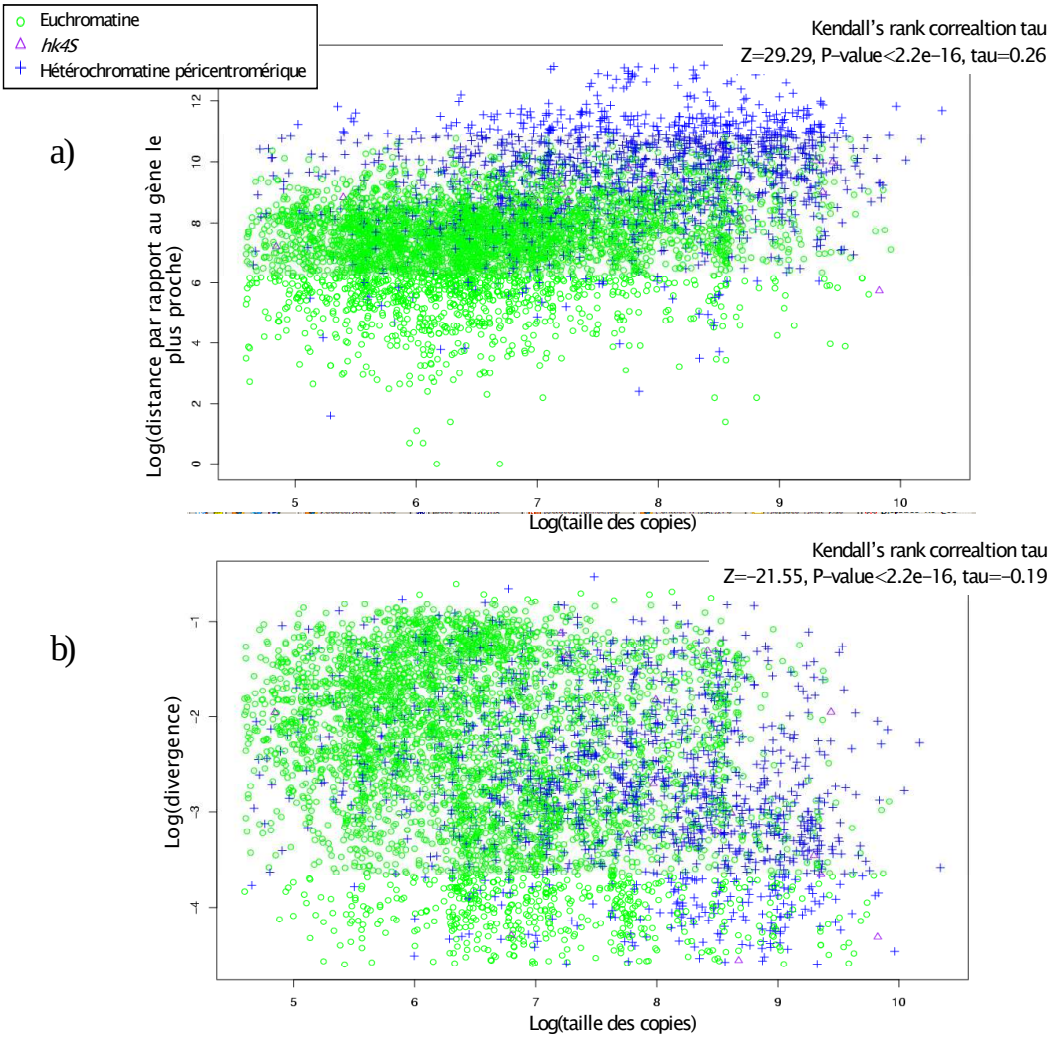


Figure 52. Graphique des régressions taille/distance et taille/divergence.

Les copies euchromatiques, péricentromériques et du *knob* sont respectivement représentées en vert, bleu et violet. **a)** Corrélation entre la taille des copies et la distance pour chaque copie avec son gène le plus proche. Plus la distance entre la copie et le gène est faible et plus le copie est petite. **b)** Corrélation entre la taille des copies et leur divergence. Les copies les plus petites correspondent aux copies les plus divergentes.

Deux mécanismes peuvent expliquer la variation de taille des copies: les délétions et la recombinaison ectopique entre copies. Afin d'estimer la part de ces mécanismes dans le processus d'élimination des copies dans les différents domaines chromatiniens, j'ai estimé la perte d'ADN due aux délétions et à la recombinaison ectopique dans l'euchromatine et l'hétérochromatine.

4.6.2. Impact de la recombinaison ectopique

La recombinaison ectopique entre les copies d'ET peut conduire à de grandes délétions ou duplications. Elle peut également aboutir à la troncature de copies ou à la formation d'éléments chimères. Afin d'estimer le taux de recombinaison ectopique, nous avons calculé le taux de solo-LTR pour chaque domaine chromatinien. Rappelons qu'un solo-LTR est le résultat de la recombinaison non-allélique entre deux LTR et que l'on suppose une corrélation directe entre le taux de recombinaison ectopique et la densité en solo-LTR.

	Euchromatine				Hétérochromatine péricentromerique			
	Solo-LTR	LTR-RI	RI	2 LTR	Solo-LTR	LTR-RI	RI	2 LTR
Gypsy-like (nb.)	103	281	342	83	131	497	364	163
Gypsy-like (% par domaine)	13,02	35,52	43,24	10,49	11,34	43,03	31,52	14,11
Copia-like (nb.)	91	318	373	142	19	71	112	30
Copia-like (% par domaine)	9,85	34,42	40,37	15,37	8,19	30,60	48,28	12,93

Tableau 9. **Taux de rétroéléments à LTR.**

Pour chaque classe de rétro-élément à LTR, le nombre de copies des superfamilles *Gypsy-like* et *Copia-like* est donné. La proportion de chaque classe pour chaque domaine a également été calculée. Donc même si il y a près de 5 fois plus de solo-LTR de type *Copia-like* dans l'euchromatine, leurs proportions dans l'hétérochromatine et dans l'euchromatine ne sont pas très différentes. Pour la superfamille *Gypsy-like*, il y a plus de solo-LTR dans l'hététochromatine mais les proportions ne différent pas non plus entre euchromatine et hétérochromatine.

De manière générale, on observe une majorité de rétro-éléments à LTR tronqués en 5' et/ou 3' (« LTR-RI » et « RI »). Pour les copies de la superfamille *Gypsy-like*, on observe deux fois plus de rétroéléments complets dans l'hétérochromatine (163 copies) que dans l'euchromatine (83 copies). Pour la superfamille *Copia-like,* on observe le profil inverse: il y a plus d'éléments complets de cette superfamille dans l'euchromatine. Sachant que les éléments de la superfamille *Gypsy-like* présentent un biais d'insertion dans l'hétérochromatine péricentromérique, nous ne pouvons pas tenir compte des résultats pour cette superfamille.

Pour la superfamille *Copia-like*, les distributions des classes de rétro-éléments (« solo-LTR », « LTR-RI », « RI » et « 2 LTR ») ne diffèrent pas significativement entre l'euchromatine et l'hétérochromatine (Test du $\chi 2$ (Pearson): $\chi 2 = 4.84$, ddl = 3, P-value=0.18). La recombinaison ectopique ne permet donc pas d'expliquer la variation de la taille des copies de l'euchromatine et de l'hétérochromatine.

Estimations du taux de délétion (en délétion /substitution /site)	Euchromatine			Hétérochromatine péricentromérique		
	Taux de délétion	Taille moyenne des délétions (pb)	Taille moyenne des copies (pb)	Taux de délétion	Taille moyenne des délétions (pb)	Taille moyenne des copies (pb)
DNA	0.094 [0.086–0.091]	19 [15–23]	**3063**	0.085 [0.079–0.088]	19 [15–24]	2714
SINE	0.087 [0.072–0.094]	6 [5–8]	420	0.11 [0.055–0.185]	3 [1–6]	**460**
Copia-like	0.042 [0.039–0.045]	17 [12–23]	3451	0.054 [0.044–0.064]	22 [9–40]	**3594**
Gypsy-like	0.050 [0.046–0.053]	22 [16–28]	4014	0.055 [0.052–0.056]	**37 [31–45]**	**5130**
LINE	**0.058 [0.054–[0.061]**	11 [10–11]	3067	0.050 [0.045–0.054]	11 [10–13]	**3126**
Helitron	**0.089 [0.083–0.090]**	14 [13–15]	2480	0.071 [0.062–0.077]	**20 [16–26]**	**2902**
All	**0.075 [0.071–0.073]**	16 [15–18]	3002	0.062 [0.059–0.062]	**25 [22–29]**	**3076**

délétions plus grandes dans l'hétérochromatine
Plus de délétions dans l'euchromatine
Plus de délétions dans l'euchromatine + délétions plus grandes dans l'hétérochromatine

Tableau 10. Taux de petites délétions par domaine.

Ce tableau récapitule les estimations du taux de délétion par subtitution et par site et de la taille des délétions (en pb). Les valeurs entre crochets correspondent aux intervalles de confiance (à 95%). Pour les éléments LINE, ainsi que les éléments *Hélitrons*, le taux de délétion estimé est significativement plus fort dans l'euchromatine. Pour les éléments des superfamilles *Gypsy-like* et *Hélitrons*, les délétions sont signicativement plus grandes dans l'hétérochromatine. Pour les superfamilles *DNA*, *SINE* et *Copia-like*, on n'observe pas de différence significative entre les taux et les tailles des délétions. Le profil général de l'ensemble des copies suggère un plus grand nombre de délétions dans l'euchromatine et des délétions plus grandes dans l'hétérochromatine.

4.6.3. Les petites délétions

Nous avons ensuite estimé la perte d'ADN *via* le mécanisme de délétion dans les différents domaines chromatiniens. Connaissant la séquence de référence, nous avons pu distinguer les délétions des insertions au niveau des alignements de séquences. Seules les délétions de moins de 100 pb ont été sélectionnées. En effet, les plus grandes délétions peuvent être dues à d'autres mécanismes tels que la recombinaison. Pour chaque famille d'ET, une estimation par maximum de vraisemblance du taux et de la taille des délétions a été réalisée. Le tableau ci-dessus résume les résultats pour différents types d'éléments.

Pour les éléments LINE, ainsi que les éléments *Hélitrons*, le taux de délétion estimé est significativement plus fort dans l'euchromatine. Pour les éléments des superfamilles *Gypsy-like* et *Hélitrons*, les délétions sont signicativement plus grandes dans l'hétérochromatine. Pour les superfamilles DNA, SINE et *Copia-like*, on n'observe pas de différence significative entre les taux et les tailles des délétions. Le profil général de l'ensemble des copies suggère un plus grand nombre de délétions dans l'euchromatine et des délétions plus grandes dans l'hétérochromatine (Tableau 10). Pour les éléments *Gypsy-like*, les délétions apparaissent plus grandes dans l'hétérochromatine que dans l'euchromatine (Tableau 10). Pour les *Hélitrons*, en plus de la différence significative de la taille des délétions, le taux de délétion est significativement plus important dans l'euchromatine que dans l'hétérochromatine (Tableau 10).

Globalement, nos estimations suggèrent un taux de délétion significativement plus élevé dans l'euchromatine et de plus grandes délétions dans l'hétérochromatine (Tableau 10).

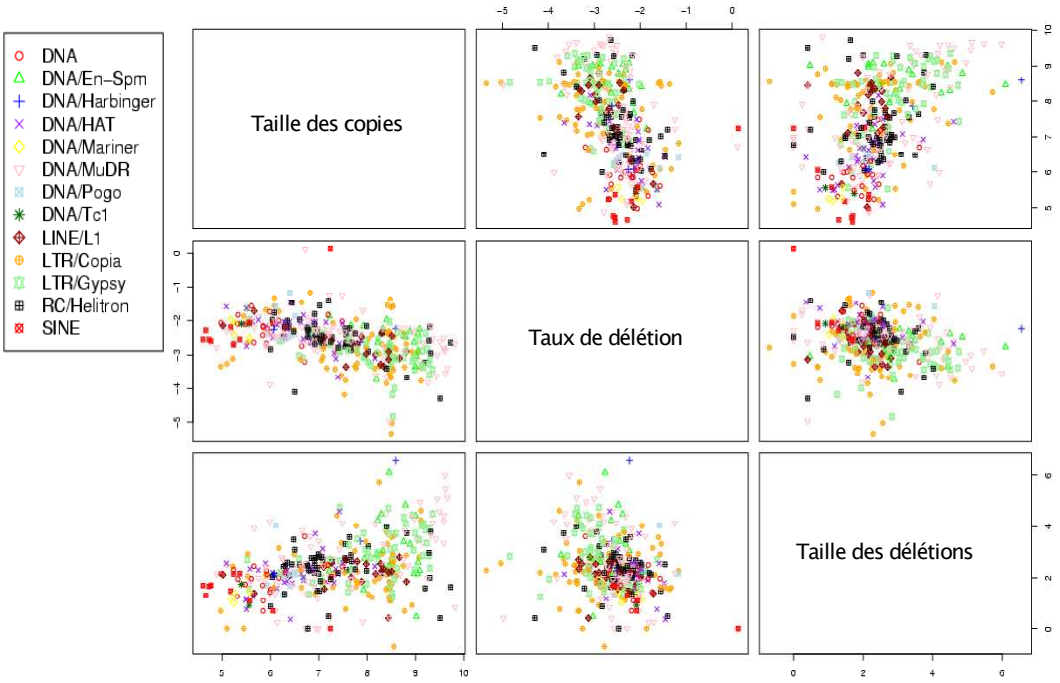


Figure 53. Relation taille des copies/délétions.

Chaque graphique met en évidence la relation entre deux des trois paramètres (taux de délétion, taille des délétions et taille des copies). Un symbole de couleur a été utilisé pour identifier chacune des superfamilles. Certaines d'entre elles se retrouvent regroupées tel que les superfamilles *Gypsy-like* et *En-spm* ou encore les SINE et les éléments ADN.

Il est important de noter que la taille des copies peut influencer le taux et la taille des délétions estimées. On observe, en effet, des relations fortes entre la taille des copies et le taux de délétion (Figure 53; Kendall's rank correlation tau: z=-10.17, P-value<$2.2x10^{-16}$) et entre la taille des copies et la taille des délétions (Figure 53; Kendall's rank correlation tau: z=8.74, P-value<$2.2x10^{-16}$). La force d'association entre la taille et le taux des délétions est moins significative. Cherchant à mettre en évidence un effet de causalité entre les délétions et la taille des copies, une analyse de régression a été réalisée. Les résultats suggèrent que plus une copie est grande, moins on observe de délétions et plus ces délétions sont grandes (F(2,386)=81.45, R2=0.30, P-value<$2.2x10^{-16}$).

Une grande délétion a plus de chance d'être interne (et donc détectable) sur une grande copie que sur une petite. Les délétions qui ont lieu aux bords des copies ne sont pas visibles. Elles ne sont donc pas comptabilisées. Donc, plus une copie est grande et plus la taille des délétions estimée sera grande. Les différences de taille des délétions n'expliquent pas l'écart de taille des copies dans l'euchromatine et l'hétérochromatine.

Par contre, les copies les plus grandes (*Gypsy-like* et *EN-spm*) présentent des taux de délétion plus faibles (en vert clair sur la Figure 53) que ceux des plus petites copies (ADN, SINE et *Pogo* ; Figure 53 voir symboles en rouge et en bleu clair). Le taux de délétions est donc plus important pour les petites copies. Or, nous avons vu que pour une même famille, les copies euchromatiques sont plus petites que les copies hétérochromatiques. Le taux de perte en ADN *via* les petites délétions est donc plus important dans l'euchromatine. Ceci explique la variation de taille des copies entre les deux domaines chromatiniens.

4.7. Impact des répétitions sur la plasticité du génome

4.7.1. Dynamique de formation des DS

Les duplications détectées correspondent majoritairement à des évènements uniques de duplications (duplications à 2 copies). Les segments se composent à la fois de régions géniques et de répétitions. Moins de 19 % des séquences sont localisées dans l'hétérochromatine. Malgré des caractéristiques différentes de celles de la Drosophile, la forte densité en ET de ce génome nous motive à suggérer un mécanisme de formation identique au modèle DDSA. Dans le but de confirmer ou d'infirmer ce modèle, il nous faut tester le rôle des répétitions dans la formation de ces répétitions.

Dans un premier temps, j'ai comptabilisé le nombre de duplications présentant des régions d'ET ou de microsatellites à leurs extrémités. La comparaison des résultats avec ceux d'un jeu de séquences contrôle (voir chapitre 3), ne montrent pas d'enrichissement significatif en ET (Tableau 11; Test du $\chi 2$ de conformité pour les ET: $\chi 2 = 2.75$, ddl = 2, P-value=0.252). Alors que chez la Drosophile, les ¾ des copies détectées ont une identité de séquence entre 95 % et 100 %, chez *A. thaliana*, cette valeur oscille entre 85 % et 95 %. Les copies de DS semblent plus divergentes. Ceci peut expliquer l'absence de traces du mécanisme induit par les ET.

Répétitions	Séquences analysées	Pourcentage de séquences (Nb de séquences)		
		Aucune répétition	une répétition à une des extrémités	une répétition à chaque extrémités
ET	DS	75.85 % (1464)	18.45 % (356)	5.70 % (110)
	séquences contrôles	76.66 % (147944)	17.16 % (33119)	6.18 % (11937)
Microsatellites ***	DS	74.46 % (1437)	23.42 % (452)	2.12 % (41)
	séquences contrôles	78.45 % (151401)	20.01 % (38622)	1.54 % (2977)

Tableau 11. Présence de répétitions aux extrémités des DS chez *A. thaliana.*

Pour chaque DS et séquences contrôles, le nombre de répétitions (ET et microsatellites) aux extrémités a été comptabilisé. On distingue trois cas: la copie ne présente aucune répétition, une répétition est localisée à une des extrémités de la copie, la copie est bornée par deux répétitions. Les valeurs entre parenthèses correspondent aux effectifs.

Par contre, les DS sont significativement enrichies en microsatellites (Tableau 11; Test du $\chi 2$ de conformité pour les microsatellites: $\chi 2= 19.35$, ddl = 2, P-value=$6.26x10^{-5}$). La détection de duplications de moins de 1 kb permet de montrer l'existence d'un mécanisme différent de ceux passant par la formation d'un hérétoduplex stable tels que les modèles DSBR ou BIR. En effet, ces deux modèles induisent la synthèse de grands fragments génomiques. Ces observations nous permettent de proposer un modèle de formation des DS, basé sur le DDSA. Les traces de microsatellites suggèrent que la formation de DS chez *A. thaliana* a été induite par ces répétitions.

4.7.2. Perte et gain d'ADN

Si nous ne considérons que les ET comme facteurs de la plasticité des génomes, nous pouvons estimer le gain et la perte totale comme la conséquence de l'insertion et de l'élimination de ces répétitions. Si l'on considère en première approximation que la taille de chaque copie au moment de son insertion correspond à la taille de la copie de référence, il suffit pour chaque famille de multiplier le nombre total de copies par la taille de la séquence de référence. Le gain dans l'euchromatine s'élève alors à 78 Mb et à 41Mb dans l'hétérochromatine péricentromérique. Pour estimer la perte après insertion, j'ai soustrait à ce gain la somme des tailles actuelles des copies du génome. La perte ainsi estimée est de 66 Mb dans l'euchromatine et de 36 Mb dans l'hétérochromatine. Le rapport perte sur gain s'élève alors à 85 % dans l'euchromatine et à 88 % dans l'hétérochromatine. Les taux relatifs de perte des séquences dans ces deux domaines chromatiniens sont identiques. En 17 millions d'années (âge de la plus ancienne transposition dans l'hétérochromatine), la région hétérochromatique a donc reçu 41 Mb et perdu 36 Mb d'ADN. Et dans l'euchromatine, en 20 millions d'années (âge de la plus ancienne transposition dans l'euchromatine), il y a eu 78 Mb de gain et 66 Mb de perte d'ADN. Ces résultats suggèrent un turnover comparable entre ces deux compartiments.

4.8. Conclusion/Discussion

4.8.1. Deux profils de dynamique

Au vue des caractéristiques mises en évidence lors de cette étude, on distingue deux profils de dynamique des répétitions: l'un pour les répétitions hautement répétées (plus de 5 copies) et l'autre pour les répétitions en faible nombre de copies (de 2 à 5 copies).

Les répétitions hautement répétées, c'est-à-dire les ET et les satellites, sont les composants majeur des régions hétérochromatiques. Ces répétitions se propagent principalement par vague, de sorte que le mécanisme conduit à la formation de plusieurs copies en un laps de temps très court. Leurs copies apparaissent significativement plus grandes dans l'hétérochromatine.

Les duplications de fragments génomiques apparaissent comme des évènements plus sporadiques. On identifie une part importante de duplications géniques chez *A. thaliana,* lorsque comparée avec les résultats obtenus chez *D. melanogaster*. La distribution chromosomique des « vraies » duplications (hors duplications géniques) est d'ailleurs uniforme. Quelques îlots de duplications ont été observés dans l'euchromatine et l'hétérochromatine. Aucune variation significative de taille des copies de DS n'a pu être mise en évidence entre les différents domaines chromatiniens.

Cette différence de dynamique entre les répétitions peut être due soit au mécanisme de propagation, soit à la composition de ces séquences. Pour estimer l'impact de la composition, nous avons testé la divergence des copies de duplication en fonction de leur composition en séquences. La dynamique des duplications géniques est plus lente que celle des duplications enrichies en ET et en satellites. Pour cette étude, nous avons donc fait le choix d'utiliser les ET comme marqueurs de la dynamique d'élimination. Nous avons ainsi travaillé avec un jeu de données plus conséquent, et cela à une échelle de temps plus fine que pour les DS ou les satellites.

4.8.2. Dynamique d'élimination des ET

Hormis en particulier les copies de la superfamille *Gypsy-like*, quel que soit le domaine, les copies d'ET ne présentent pas généralement de biais d'insertion. De plus, les copies de cette superfamille sont connues pour être récentes chez *A. thaliana*. Notre hypothèse est donc que la dynamique d'insertion des ET est la même dans les différents domaines chromatiniens. Pour l'interprétation des résultats, nous avons donc toujours tenu compte du statut atypique des copies de cette superfamille. Dans l'euchromatine, l'insertion de petites copies apparaît moins délétère. En effet, nous avons montré que la pression de sélection due aux gènes va tendre à éliminer les grandes copies (les insertions les plus délétères). Les régions hétérochromatiques tolèrent mieux l'insertion de séquences répétées. Les copies d'ET sont plus grandes. Elles semblent évoluer de façon quasi-neutre. L'évolution des séquences dans l'hétérochromatine est donc un bon reflet de l'évolution neutre du génome.

L'estimation des forces d'élimination des répétitions suggère une dynamique plus importante dans l'euchromatine. Pourtant, la perte globale due à la transposition est estimée à 85 % dans l'euchromatine et 88 % dans l'hétérochromatine. Un mécanisme d'élimination des copies dans l'hétérochromatine a dû être sous-estimé. Ce mécanisme plus important dans l'hétérochromatine permettrait de ramener le taux d'élimination dans ce domaine à une valeur proche de celle de l'euchromatine. Les copies hétérochromatiques étant significativement plus grandes, ce mécanisme ne réduit pas progressivement la taille des copies, mais, pourrait les retirer par blocs. La recombinaison ectopique entre copies identiques pourrait être ce mécanisme. Nous avons mis en évidence une proportion des solo-LTR identique dans l'hétérochromatine et dans l'euchromatine. Estimer le taux de solo-LTR revient à apprécier la force de recombinaison entre deux régions répétées (LTR) proches (quelques kilobases). La recombinaison à l'origine de grands remaniements n'a donc pas été réellement estimée ici. La pauvreté en gènes ainsi que la forte densité en répétitions devraient induire un plus grand nombre de remaniements dans l'hétérochromatine. Les réarrangements pourront alors être de plus grandes ampleurs mais aussi plus fréquents dans l'hétérochromatine. La recombinaison homologue non-allélique se révèle ainsi comme un bon mécanisme pour contre-balancer l'insertion des répétitions dans l'hétérochromatine.

4.8.3. Les modèles de dynamique

L'ensemble de nos résultats nous a permis de proposer un modèle de dynamique des séquences dans l'hétérochromatine. Pour cela, nous avons discuté des modèles de dynamiques des séquences répétées: le modèle de « recombinaison ectopique » et celui de « rupture de gènes ». Le « modèle de recombinaison ectopique » prédit une accumulation de répétitions dans les régions à faible taux de recombinaison (méiotique), car peu éliminées (Montgomery et al. 1987). En effet, les ET sont de bons substrats pour la recombinaison. Dans les régions à fort taux de recombinaison, elles seront plus rapidement éliminées (Virgin et Bailey 1998; Petrov *et al.* 2003). Nous n'avons pas pu mettre en évidence une corrélation claire entre la densité en ET et le taux de recombinaison. Mais, nous pouvons suggérer ici un effet indirect puisque la densité en gènes est corrélée au taux de recombinaison. Pour expliquer ce résultat, nous pouvons invoquer l'autofécondation. Ce processus tend à rendre les individus homozygotes. En rendant homozygote les loci, le mécanisme de recombinaison va orienter le choix de la matrice préférentiellement vers la région homologue en position allélique. Les chromosomes homologues étant identiques, les possibilités d'erreurs seront réduites, ce qui implique des réparations des CDB majoritairement conservatives.

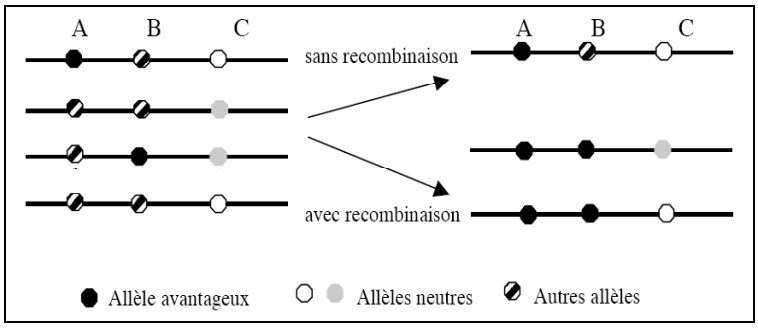


Figure 54. Effet « Hill-Robertson ».

Considérons 3 loci liés génétiquement. Parmi ces loci, le locus A correspond à un allèle avantageux. Il est soumis à faible sélection dans la population. Il augmente ainsi la *fitness* de 0,1 %. Le locus B est soumis à une forte sélection dans la population, un allèle avantageux (augmentant de 10 % la *fitness*) vient d'apparaître. Le locus C est neutre. Plusieurs allèles neutres coexistent dans la population. Ces différents allèles sont apparus par mutation *via* un processus aléatoire. Etant donné que la population est de taille finie, toutes les combinaisons d'allèles pour les différents loci ne sont pas présentes. La combinaison d'allèles dépend du mécanisme de recombinaison (Hill et Robertson 1966).

Nous pouvons également invoquer l'effet dit de « Hill-Robertson ». Celui-ci suggère un effet de la sélection plus fort lorsque le gène est situé dans une région à fort taux de recombinaison (Figure 54) (Hill et Robertson 1966). Comme le modèle de « recombinaison ectopique », ce modèle prédit également une plus forte densité en ET dans les régions à faible taux de recombinaison, si l'on considère les insertions d'ET comme délétères et non neutres. Nous avons pu apprécier la sélection exercée sur les plus grandes copies d'ET. Les copies les plus grandes sont effectivement localisées dans les régions à faible taux de recombinaison, loin des gènes et dans des régions pauvres en gènes.

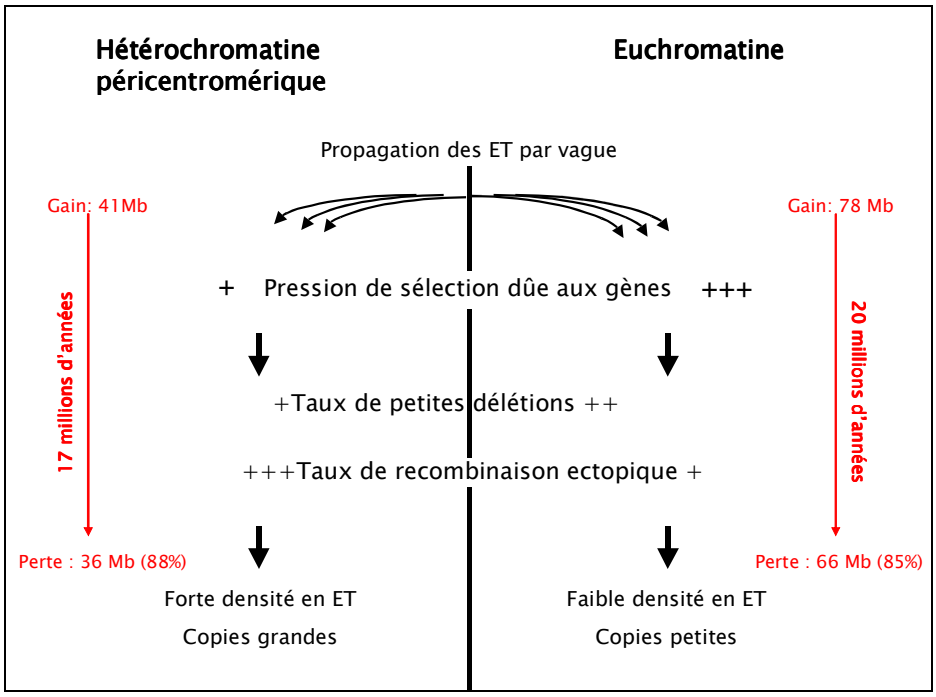


Figure 55. Dynamique des ET chez *A. thaliana*.

Ce schéma reprend l'ensemble des résultats. On suppose que la dynamique d'insertion des ET est la même dans l'euchromatine que l'hétérochromatine. Les ET se propagent par vague. Dans l'euchromatine, la pression de sélection due aux gènes va induire une élimination rapide *via* la sélection et les petites délétions. Ce qui explique la faible densité en ET ainsi que la présence de copies plus petites. A l'opposé, les copies de l'hétérochromatine sont plus grandes. Cette observation s'explique par l'absence d'une pression de sélection aussi forte que dans l'euchromatine. Dans l'hétérochromatine, pour contre-balancer l'insertion d'ET, en plus des délétions, la recombinaison ectopique (mitotique) permet d'éliminer un grand nombre de copies. Globalement, le taux de perte d'ADN induit par les ET est le même dans les deux domaines. En moins de 20 millions d'années, l'apport en ET a été de l'ordre 126 Mb. Actuellement, il ne reste que 18 Mb (126 Mb -108 Mb).

4.8.4. Le turn-over des séquences

Ces modèles n'étant pas mutuellement exclusifs, le modèle que nous proposons combine l'effet « Hill-Robertson » et le modèle de rupture de gènes. Cependant, nous avons montré que la densité en ET n'est pas expliquée par la recombinaison, ce qui laisse penser que l'effet « Hill-Robertson » joue marginalement. Une grande part des copies insérées ont probablement un effet délétère, s'il y a des gènes dans les environs. Dans les régions à faibles densité en gènes et au faible taux de recombinaison, la majorité des copies va tendre à disparaître *via* des mécanismes de délétion ou recombinaison ectopique. Dans les régions denses en gènes et à fort taux de recombinaison, ces insertions seront éliminées par la sélection avant qu'elles soient fixées dans la population. Le taux de délétion plus fort observé dans l'hétérochromatine peut également être la conséquence de la sélection. En effet, celle-ci peut sélectionner en un locus, les copies les plus délétées, ce qui augmentera artificiellement le taux de délétion mesuré. Mais, dans ces régions, des insertions d'ET peuvent aussi plus facilement se fixer dans la population en confèrant ou non un avantage évolutif à l'organisme. En effet, les associations neutres ou bénéfiques entre ET et gènes ont davantage de chance d'émerger car la combinatoire des allèles entre ces deux types de séquences est plus importante. Ceci pourrait expliquer le fait qu'on trouve des copies plus anciennes dans l'euchromatine. Seules des études de génétiques des populations pourront nous permettre d'estimer le nombre de ces copies fixées.

Le temps de demi-vie des ET dans l'hétérochromatine péricentromérique est de 0.44 mutations par site contre 0.57 dans l'euchromatine. On peut en déduire que la taille de ces éléments va donc diminuer de moitié en moins de 42 millions d'années dans l'hétérochromatine, alors que dans l'euchromatine, le processus s'effectuera en moins de 60 millions d'années. En conclusion, le turnover de ces séquences est plus lent dans les régions denses en gènes.

D'après nos analyses, le pouvoir délétère d'une séquence va dépendre de sa taille. Les séquences de grandes tailles ont un potentiel délétère fort comme le montre la décroissance de la taille des copies avec la proximité des gènes. Dans l'euchromatine, la forte densité en gènes induit une élimination rapide des séquences à effet délétère *via* la sélection et les petites délétions (Figure 55). A l'opposé, dans l'hétérochromatine, l'insertion de ces séquences est mieux tolérée. Les copies y apparaissent donc plus grandes (Figure 55). On observe des délétions plus grandes mais l'élimination par

délétions se fait plus lentement à cause de l'absence de pression de sélection. Mais, ce mécanisme seul ne suffit pas à contre-balancer l'insertion des répétitions. Même si nos résultats sous-estiment la force de ce mécanisme, la recombinaison ectopique apparaît comme un processus important d'élimination des répétitions dans l'hétérochromatine. Il peut conduire à l'élimination de nombreuses copies d'ET en un seul évènement, ce qui permettrait d'expliquer l'absence de copies anciennes dans l'hétérochromatine.

Ainsi, le turnover est très rapide aussi bien dans l'euchromatine que dans l'hétérochromatine. Ce génome a gagné 126 Mb et perdu 108 Mb (soit 86 % de perte) en moins de 20 millions d'années (Figure 55).

Nos résultats montrent que la pression de sélection due aux gènes semble être un facteur important de cette dynamique. Or dans le *hk4S,* la densité en gènes est plus importante que dans l'hétérochromatine péricentromérique. On s'attend, par conséquent, à observer un profil des caractéristiques des copies d'ET intermédiaire entre celui des copies de l'euchromatine et celui des copies de l'hétérochromatine péricentromérique.

4.8.5. La dynamique d'insertion des répétitions dans *hk4S*

L'âge de *A. thaliana* a été estimé à environs 5 millions d'années (Figure 32) (Henry *et al.* 2006)). Le *knob* est donc apparu, il y a moins de 5 millions d'années. On distingue deux classes de copies d'ET: 40 % des copies ont moins de 4.7 millions d'années et 60 % ont entre 5 et 33 millions d'années. On retrouve 0.40 % (8/1930) de DS dans cette région. On estime respectivement, l'âge maximal des satellites et des DS à moins de 3 et moins de 5 millions d'années. Les DS correspondent dans 5 cas sur 8 à des régions d'ET et des satellites.

On suppose qu'après la formation du *knob*, il y a eu, d'abord, des évènements de transposition et de duplication. Des copies d'ET plus anciennes ont ainsi pu être emmenées par les DS. Dans un second temps, une des copies de l'élément *Mutator-like* ou *MULE* a divergé et conduit à la formation du « cluster » des RT. Le profil évolutif de cette région suggère un processus d'hétérochromatinisation. En effet, en moins de 5 millions d'années, 25 des 33 gènes du *knob* ont été éliminés, un grand nombre d'ET s'y sont insérés et on y retrouve un « *cluster* » de RT.

Dans ce cas, nous proposons que les 8 gènes présents, comme la plus part des duplications géniques des génomes, vont tendre à disparaître. La présence de copies d'ET plus grandes que celles de l'hétérochromatine péricentromérique malgré des divergences égales, montre une pression de sélection encore plus faible dans cette région. Celle-ci peut s'expliquer par la pauvreté en gènes mais on peut également invoquer que la pression de sélection due à des gènes dupliqués soit moindre que celle due aux gènes uniques hétérochromatiques.

5. Discussion Générale

5.1.1. Des caractéristiques des DS

La proportion en DS varie considérablement entre les organismes eucaryotes. Elle varie de ~1 % chez la souris à près de 20 % chez *A. thaliana*. Aucune corrélation ne peut être mise en évidence entre la proportion en DS et la proportion en ET ou la taille du génome, ou même les clades. Le génome de *A. thaliana* estimé à 130Mb, présente une proportion 13 fois plus importante que celle de *D. melanogaster* (génome estimé à 180 Mb). Pourtant l'estimation de la proportion en ET est assez proche pour ces deux génomes. Entre le rat et la souris, le rapport des proportions en ET est environ égal à 2. Les proportions en DS varient également de manière importante entre ces deux génomes pourtant proches (2.92 % chez le rat et 1.2 % chez la souris; (Waterston *et al.* 2002; Rew 2004).

La taille minimale considérée pour des DS des mammifères est de 1 kb. Cette valeur n'est pas la conséquence d'une observation mais une valeur fixe définie dans la méthode de détection. N'utilisant pas les annotations des ET, Bailey *et al.* ont fait le choix, pour éliminer un grand nombre d'ET, de ne pas prendre en compte les duplications de moins de 1 kb (Bailey et al. 2001). En effet, la taille moyenne des éléments *Alu*, éléments majoritairement représentés, est de l'ordre de 300 pb. Cette méthode conduit à l'obtention d'un jeu de données pouvant contenir un certain nombre de faux-positifs. De petits fragments dupliqués peuvent aussi correspondre à des DS plus anciennes ou de réelles petites duplications produites par un mécanisme identique. L'analyse de la corrélation entre la taille des copies et leur divergence pourrait permettre d'éclaircir ce point dans ces génomes. Près de 50 % d'entre elles ont une taille inférieure à 1 kb. Les duplications de plus de 1 kb correspondent à 1.3 % de la séquence génomique. Les DS de *Caenorhabditis elegans* sont également de petite taille. D'après les données de Samonte et Eichler (Samonte et Eichler 2002), alors que 2.65 % de la séquence génomique se compose de DS de moins de 5 kb, seul 0.66 % de la séquence correspond aux DS de plus de 10 kb. Chez *A.thaliana*, on observe également près de 50 % de DS de moins de 1 kb. Seul 8 % des DS ont une taille supérieure à 5 kb et 2.64 % supérieure à 10 kb. Chez les mammifères, elles peuvent atteindre jusqu'à plusieurs centaines de kb.

Alors que chez *D. melanogaster*, nous avons mis en évidence une distribution des DS en faveur des régions riches en répétitions, c'est-à-dire l'hétérochromatine, chez *A. thaliana*, comme chez les mammifères, la répartition ne dépend pas de la structure de la chromatine. Elles sont réparties à différents sites du génome.

De récentes études comparatives entre génomes eucaryotes ont permis de montrer la présence de séquences répétées aux points de jonction des DS (Koszul *et al.* 2004; Bailey et Eichler 2006). Ces observations ont été faites aussi bien chez la levure *S. cerevisae* que l'homme. Pour comprendre le mécanisme de formation des DS chez un organisme tel que *A. thaliana*, il faut également s'assurer que les duplications ne sont pas issues de polyploïdisation. Pour cette raison, j'ai éliminé du jeu de données les duplications avec plus de 95 % de recouvrement en gènes. Ainsi, les points de jonction des DS apparaissent enrichis en microsatellites.

Le DDSA, un modèle de NAHR, pourrait être généralisé afin d'expliquer la formation des DS chez les eucaryotes. Ce modèle a été proposé pour des DS ayant une taille maximale de moins de 100 kb. Or, chez les mammifères, la taille des DS peut varier jusqu'à plusieurs centaines de kb. Comment le DDSA peut-il prédire la formation des DS de plus grande taille ? Afin de répondre à cette question, il faudrait rechercher les signatures du DDSA sur les séquences des DS chez les mammifères. D'après les processus de dissociation/ré-invasion, nous pouvons proposer qu'un mécanisme de ce type puisse relancer la synthèse et ainsi permettre de recouvrir des régions de plusieurs kb. En fonction des caractéristiques d'un génome, on peut s'attendre à ce que ces processus de ré-initialisation soient plus ou moins fréquents. Ceci permettrait d'expliquer les différences de taille des DS entre les génomes eucaryotes.

D'après le DDSA, le fait d'observer des proportions de duplications interchromosomiques et intrachromosomiques quasi-égales chez *A. thaliana*, suggère que la recherche d'homologie a autant de chance de se faire sur le même chromosome que sur un autre chromosome. La forte densité en répétitions chez cet organisme peut permettre d'expliquer ce résultat.

Pour la suite de l'étude sur les DS, il serait intéressant d'élargir cette étude aux autres génomes de Drosophile.

5.1.2. La compartimentation des répétitions

Les génomes dits compacts correspondent à des génomes tels que ceux de *A. thaliana, D. melanogaster et C. elegans*. La région euchromatique de ces génomes correspond à entre 100 et 125 Mb alors que les régions hétérochromatiques représentent entre 10 et 60 Mb (sauf *C. elegans* où il n'y en a pas).

Dans ces génomes, les ET se retrouvent préférentiellement localisés dans les régions hétérochromatiques (Kapitonov et Jurka 1999; AGI 2000). A l'opposé, chez les mammifères et plus précisément chez l'homme, les ET (majoritairement des éléments *Alu)* se retrouvent dispersés dans l'euchromatine et l'hétérochromatine (Bailey et Eichler 2006). De plus, la proportion en ET s'avère plus importante pour les grands génomes. En effet, on passe de moins de 10 % à plus de 45 % d'ET dans l'euchromatine. Dasilva *et al.* (2002) ont analysé la distribution en ET chez des petits génomes de vertébrés tels que *Tetraodon nigroviridis* et *Takifugu rubripes*. Ils ont suggéré la compartimentation des ET dans les régions hétérochromatiques comme chez les autres espèces au génome compact. La compartimentation des génomes ne dépendrait donc pas du clade mais de la taille des génomes.

Les régions chromosomiques, encore mal caractérisées à l'heure actuelle, sont les régions de transition entre l'euchromatine et l'hétérochromatine. On imagine un équilibre entre la pression de sélection due aux gènes et l'insertion de répétitions. Des vagues de transpositions seront régulées par la pression de sélection. Alors que chez les organismes aux chromosomes holocentriques, ce processus doit être très dynamique, chez *A. thaliana*, on assiste presque « en direct » à ce processus.

5.1.3. Contraction de taille des génomes médiée par les répétitions

Etant donné que la taille des copies de répétitions semble influer sur leur dynamique, on peut proposer deux modes d'élimination. On s'attend à ce que la « *fitness* » diminue avec l'augmentation du nombre de copies d'ET à effet délétère. En effet, l'augmentation du nombre de copies conduit à un taux de recombinaison ectopique plus élevé et en conséquence à plus de réarrangements chromosomiques. Pour contrecarrer cet effet, la dynamique d'élimination se fera rapidement dès l'insertion. La taille des séquences insérées, à effet délétère, va tendre à diminuer rapidement jusqu'à minimiser l'effet

délétère. Le mécanisme de recombinaison ectopique et les délétions internes de grande taille peuvent également permettre d'éliminer rapidement des fragments.

Les forces induisant une variation de la taille des génomes peuvent être soit globales telles que la polyploïdisation et la propagation des ET, ou locales comme les duplications en tandem. Dans les deux situations, les forces conduiront à d'importantes modifications structurales des génomes. Deux modèles de variation de la taille des génomes ont été mis en évidence. Chez les mammifères, les oiseaux et les poissons, les variations de taille apparaissent relativement faible. L'accumulation de petits blocs d'ADN a été proposée pour expliquer cette augmentation progressive (Gregory et Hebert 1999). A l'opposé, la taille des génomes des plantes et des invertébrés suggère des variations plus importantes. On sait aujourd'hui que la taille de certains de ces génomes a principalement augmenté par polyploïdisation (SanMiguel *et al.* 1996; 1998). Ces génomes sont également sujets à des augmentations progressives dues à l'insertion de répétitions. D'après le modèle d'évolution concertée, les duplications géniques récentes seront difficilement distinguables des anciennes. C'est ce que l'on observe pour les duplications chez *A. thaliana*. Ce génome dense en duplications est plus riche en duplications géniques que *D. melanogaster*. Or, le génome ancêtre de *A thaliana* a subit plusieurs polyploïdisations (Figure 32). Les duplications géniques détectées peuvent donc aussi bien correspondre à des duplications issues des polypoïdisations que de duplications récentes.

L'analyse comparative réalisée par Hall *et al.* (2006) a mis en évidence une importante contraction du génome ancêtre des Brassicaceae *via* un ensemble de réarrangements chromosomiques assez conséquent (Figure 32). Or ces régions sont significativement plus grandes chez *A.thaliana* que chez les espèces qui lui sont proches. Ceci suggère de récentes vagues de transpositions spécifiquement chez cet organisme. Ces évènements de transposition seraient ainsi responsables de la variation de taille des génomes entre *A. thaliana* et les espèces qui lui sont proches.

5.1.4. Les ET, éléments moteurs de l'augmentation de la taille des génomes

Les proportions en ET des génomes eucaryotes actuels varient entre ~2 et ~90 %. Douze génomes de Drosophillidés ont été récemment séquencés et annotés. Les proportions en ET de ces génomes varient entre ~2.7 % (chez *Drosophila simulans* et *Drosophila grimshawi)* et ~25 % (chez *Drosophila ananassae)*. La taille de ces génomes varie du simple au triple: entre 130 Mb et 346 Mb. Dans ces génomes, l'insertion des ET aurait un impact principalement sur la taille des régions hétérochromatiques (Clark et al. 2007).

Afin d'expliquer cette forte variation de taille de l'hétérochromatine, plusieurs mécanismes impliquant les ET peuvent être proposés tels que les duplications de grand fragments génomiques ou la macrotransposition, c'est-à-dire la transposition d'un fragment d'ADN bordé par deux éléments à ADN distant.

Les familles ne sont pas toutes représentées chez ces douze génomes. En effet, alors que certaines familles restent active chez un organisme, d'autres ont pu être amenées à disparaître chez un autre organisme.

De plus, la répartition par classe et par famille d'ET ne semble pas uniforme le long des chromosomes. La superfamille *Gypsy-like* est un bon exemple d'ET présentant un biais d'insertion préférentielle dans l'hétérochromatine. Les ET ont donc un effet majeur direct (transposition) et indirect (recombinaison ectopique et macrotransposition), sur la plasticité des génomes eucaryotes.

5.2. Vers une approche de Génomique populationnelle

5.2.1. Les données génomiques disponibles

Depuis quelques années, de nombreux génomes eucaryotes ont été séquencés. La proximité phylogénétique entre certains organismes ouvre des nombreuses perspectives d'études. En plus des séquences génomiques de qualité, les annotations en ET des 12 génomes de Drosophilidés seront prochainement disponibles. L'ensemble de ces données devrait permettre, dans les années à venir, d'étudier la dynamique des séquences à différentes échelles de temps. Il serait envisageable d'étudier les relations entre cette dynamique et les traits de vie des espèces: habitats, taille des populations, reproduction, etc. Les régions synténiques entre *A. thaliana* et *A. lyrata*, ainsi que des annotations en ET, sont également disponibles (http://gautlab.bio.uci.edu/data.html).

5.2.2. Génomique comparative et Génétique des populations

Un grand nombre de génomes complet phylogénétiquement proches étant séquencés et disponibles, la congruence entre données populationnelles et caractéristiques des génomes va pouvoir être réalisée à grande échelle, pour la première fois. La propagation des ET dans un génome est limitée à la fois par la régulation du taux de transposition et la pression de sélection exercée sur les copies après insertion. La combinaison de ces deux approches permettrait d'enrichir les modèles de dynamique des ET. Cette approche combinée pourrait être initiée par l'étude de l'impact des éléments non-LTR: éléments très abondant dans les génomes de Drosophiles. Ces éléments représentant 85 % des ET chez l'homme (Berezikov *et al.* 2000). Ces ET évoluent principalement par transfert vertical et ne s'excisent pas d'un génome. Notre analyse ne sera donc pas biaisée par d'éventuels transferts entre différentes espèces. Une étude pilote a déjà été réalisée afin de tester la faisabilité du projet. Pour cette étude, 4 familles d'élément non-LTR appartenant aux clade *Jockey* ont été utilisées (Petrov et al. 2003). De plus, seules les copies des régions à fort taux de recombinaison ont été sélectionnées. Le jeu de données utilisé ne permet pas de prendre en compte d'importants facteurs de sélection. Se limiter aux copies localisées dans les régions à fort taux de recombinaison ne permet pas d'apprécier l'effet du taux de recombinaison sur la dynamique des ET (Hill et Robertson 1966). Et se restreindre à 4 familles d'éléments non-LTR ne permet pas de tenir compte

des autres mécanismes de transposition. Une étude combinée plus approfondie en travaillant avec un jeu de données plus important pourrait être riche de nouveaux enseignements.

6. Annexes

6.1. Article 1: «Detection of transposable elements by their compositional bias»

6.2. Article 2: « A model of segmental duplication formation in Drosophila melanogaster »

6.2.1. Annexe1: Coordonnées des DS détectées

6.2.2. Annexe2: Analyse des 12 évènements de duplication sélectionnées

6.2.3. Annexe3: Analyse des «*clusters*» de RT

6.2.4. Annexe4: Analyse de points de cassure des duplications

6.3. Article 3 (en cours d’écriture)

6.3.1. Annexe5: Analyse des ET chez A. thaliana

RESUME

De la bactérie à l'homme, dispersées ou en tandem, les répétitions peuvent représenter jusqu'à 90 % de la séquence génomique. Malgré leur impact sur la plasticité et l'évolution des génomes eucaryotes, leurs mécanismes de formation sont encore très spéculatifs. L'insertion continue de nouvelles répétitions devrait conduire à une augmentation constante de la taille des génomes. Or, il ne semble pas que ce soit le cas. Y a t-il régulation de la taille des génomes ? Le processus de régulation est-il le même dans l'euchromatine et l'hétérochromatine ?

Afin d'étudier la dynamique des répétitions, j'ai développé un ensemble de programmes informatiques pour la détection des duplications segmentaires (DS) et des répétitions en tandem (RT). A partir des caractéristiques des DS détectées chez *Drosophila melanogaster*, j'ai proposé un modèle de formation des DS, basé sur un modèle de recombinaison homologue non-allélique. J'ai également identifié les traces de l'implication des éléments transposables (ET) dans ce processus.

Pour comprendre la relation entre les répétitions et la structure de la chromatine, j'ai ensuite réalisé une analyse comparative de la dynamique des répétitions euchromatiques et hétérochromatiques chez *Arabidopsis thaliana*. La construction d'arbres phylogénétiques des séquences répétées m'a permis de dater les répétitions. Nous suggérons ainsi une propagation par « vague » des ET.

J'ai ensuite estimé les forces d'élimination des copies d'ET. Nos résultats suggèrent que dans l'euchromatine, la pression de sélection due aux gènes induit l'élimination des répétitions. Dans l'hétérochromatine, la faible densité en gènes permet de maintenir une forte densité en ET. Pourtant, les estimations du taux de perte en ADN, prédisent un turnover aussi rapide dans l'euchromatine que dans l'hétérochromatine. Afin de contre-balancer l'insertion des ET dans l'hétérochromatine, nous pouvons invoquer la recombinaison homologue non-allélique.

ABSTRACT

From bacteria to human, interspersed or in tandem, repeated sequences can cover more than 90 % of a genomic sequence. Despite their impact on the evolution and the plasticity of eukaryotic genomes, their mechanism of propagation remains still unclear. The continuous insertion of new copies should induce the increase of genome size. What are the selection pressures involved in the genome size regulation? Do these selection strength are the same in euchromatic and heterochromatic regions? In order to highlight the repeat dynamics, I first developed computational pipelines to detect segmental duplications (SDs) and tandem repeat arrays (TRs).

The SD features of the detected in the *Drosophila melanogaster* genome, allowed us to propose a non-allelic homologous recombination mechanism as a SD formation model. This process can be induced by repeats such as TEs. Indeed, I showed the traces of transposable elements (TEs) at their breakpoints.

To understand the relationship between the repeats and the chromatin structure, we investigated the repeat evolutionary dynamics by comparing their features in heterochromatin and euchromatin domains in *Arabidopsis thaliana*. We constructed phylogenetic trees of repeats to estimate their divergence in euchromatin and heterochromatin. The tree topology of TE families reflects transpositions by "burst". In order to explain, the size and divergence variations of the TE copies between these two chromatic domains, we estimated the strength of repeat elimination into these regions. Our analysis suggests that the gene selection pressure effect induces in euchromatin the repeat elimination, although, in heterochromatin, the gene paucity allows to maintain the high TE density. However, the DNA loss rate estimations suggest the same fast turnover in the both chromatin domains. To counteract the TE insertion in heterochromatin, we proposed that non-allelic homologous recombination may play a significant role. This process allows to eliminate rapidly lots of copies.

Bibliographie

Abramov, Y.A., G.L. Kogan, E.V. Tolchkov, V.I. Rasheva, S.A. Lavrov, S. Bonaccorsi, I.A. Kramerova, and V.A. Gvozdev. 2005. Eu-heterochromatic rearrangements induce replication of heterochromatic sequences normally underreplicated in polytene chromosomes of Drosophila melanogaster. *Genetics* **171:** 1673-1681.

Adams, M.D. S.E. Celniker R.A. Holt C.A. Evans J.D. Gocayne P.G. Amanatides S.E. Scherer P.W. Li R.A. Hoskins R.F. Galle *et al.* 2000. The genome sequence of Drosophila melanogaster. *Science* **287:** 2185-2195.

AGI, A.G.I. 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408:** 796-815.

Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. 1990. Basic local alignment search tool. *J Mol Biol* **215:** 403-410.

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

Ananiev, E.V., R.L. Phillips, and H.W. Rines. 1998. Complex structure of knob DNA on maize chromosome 9. Retrotransposon invasion into heterochromatin. *Genetics* **149:** 2025-2037.

Ananiev, E.V., R.L. Phillips, and H.W. Rines. 2000. Complex structure of knobs and centromeric regions in maize chromosomes. *Tsitol Genet* **34:** 11-15.

Andrieu, O., A.S. Fiston, D. Anxolabehere, and H. Quesneville. 2004. Detection of transposable elements by their compositional bias. *BMC Bioinformatics* **5:** 94.

Armengol, L., M.A. Pujana, J. Cheung, S.W. Scherer, and X. Estivill. 2003. Enrichment of segmental duplications in regions of breaks of synteny between the human and mouse genomes suggest their involvement in evolutionary rearrangements. *Hum Mol Genet* **12:** 2201-2208.

Ashburner, M. 1989. Drosophila. *A Laboratory Manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.*

Bailey, J.A., A.M. Yavor, H.F. Massa, B.J. Trask, and E.E. Eichler. 2001. Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res* **11:** 1005-1017.

Bailey, J.A., Z. Gu, R.A. Clark, K. Reinert, R.V. Samonte, S. Schwartz, M.D. Adams, E.W. Myers, P.W. Li, and E.E. Eichler. 2002. Recent segmental duplications in the human genome. *Science* **297:** 1003-1007.

Bailey, J.A. and E.E. Eichler. 2006. Primate segmental duplications: crucibles of evolution, diversity and disease. *Nat Rev Genet* **7:** 552-564.

Bennetzen, J.L. and E.A. Kellogg. 1997. Do Plants Have a One-Way Ticket to Genomic Obesity? *Plant Cell* **9:** 1509-1514.

Bennetzen, J.L. 2000. Transposable element contributions to plant gene and genome evolution. *Plant Mol Biol* **42:** 251-269.

Bennetzen, J.L., J. Ma, and K.M. Devos. 2005. Mechanisms of recent genome size variation in flowering plants. *Ann Bot (Lond)* **95:** 127-132.

Benson, G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27:** 573-580.

Berezikov, E., A. Bucheton, and I. Busseau. 2000. A search for reverse transcriptase-coding sequences reveals new non-LTR retrotransposons in the genome of Drosophila melanogaster. *Genome Biol* **1:** RESEARCH0012.

Bergman, C.M., H. Quesneville, D. Anxolabéhère, and M. Ashburner. 2006. Recurrent insertion and duplication generate networks of transposable element sequences in *Drosophila melanogaster* genome. *Submitted*.

Biemont, C., C. Vieira, C. Hoogland, G. Cizeron, C. Loevenbruck, C. Arnault, and J.P. Carante. 1997. Maintenance of transposable element copy number in natural populations of Drosophila melanogaster and D. simulans. *Genetica* **100:** 161-166.

Blumenstiel, J.P., D.L. Hartl, and E.R. Lozovsky. 2002. Patterns of insertion and deletion in contrasting chromatin domains. *Mol Biol Evol* **19:** 2211-2225.

Brennecke, J., A.A. Aravin, A. Stark, M. Dus, M. Kellis, R. Sachidanandam, and G.J. Hannon. 2007. Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. *Cell* **128:** 1089-1103.

Brown, G.G., J.S. Lee, N. Brisson, and D.P. Verma. 1984. The evolution of a plant globin gene family. *J Mol Evol* **21:** 19-32.

Buisine, N., H. Quesneville, and V. Colot. 2008. Improved detection and annotation of transposable elements in genomes using multiple reference sequence sets. In *Accepted*.

Bundock, P. and P. Hooykaas. 2005. An Arabidopsis hAT-like transposase is essential for plant development. *Nature* **436:** 282-284.

Burkart, W., T. Jung, and G. Frasch. 1999. Damage pattern as a function of radiation quality and other factors. *C R Acad Sci III* **322:** 89-101.

Carmell, M.A., A. Girard, H.J. van de Kant, D. Bourc'his, T.H. Bestor, D.G. de Rooij, and G.J. Hannon. 2007. MIWI2 is essential for spermatogenesis and repression of transposons in the mouse male germline. *Dev Cell* **12:** 503-514.

Celniker, S.E., D.A. Wheeler, B. Kronmiller, J.W. Carlson, A. Halpern, S. Patel, M. Adams, M. Champe, S.P. Dugan, E. Frise *et al.* 2002. Finishing a whole-genome shotgun: release 3 of the Drosophila melanogaster euchromatic genome sequence. *Genome Biol* **3:** RESEARCH0079.

Charlesworth, B., C.H. Langley, and W. Stephan. 1986. The evolution of restricted recombination and the accumulation of repeated DNA sequences. *Genetics* **112:** 947-962.

Cheng, Z., M. Ventura, X. She, P. Khaitovich, T. Graves, K. Osoegawa, D. Church, P. DeJong, R.K. Wilson, S. Paabo *et al.* 2005. A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature* **437:** 88-93.

Cheung, J., M.D. Wilson, J. Zhang, R. Khaja, J.R. MacDonald, H.H. Heng, B.F. Koop, and S.W. Scherer. 2003. Recent segmental and gene duplications in the mouse genome. *Genome Biol* **4:** R47.

Christensen, S.M. and T.H. Eickbush. 2005. R2 target-primed reverse transcription: ordered cleavage and polymerization steps by protein subunits asymmetrically bound to the target DNA. *Mol Cell Biol* **25:** 6617-6628.

Clark, A.G. M.B. Eisen D.R. Smith C.M. Bergman B. Oliver T.A. Markow T.C. Kaufman M. Kellis W. Gelbart V.N. Iyer *et al.* 2007. Evolution of genes and genomes on the Drosophila phylogeny. *Nature* **450:** 203-218.

Csink, A.K. and S. Henikoff. 1998. Something from nothing: the evolution and utility of satellite repeats. *Trends Genet* **14:** 200-204.

Curcio, M.J. and K.M. Derbyshire. 2003. The outs and ins of transposition: from mu to kangaroo. *Nat Rev Mol Cell Biol* **4:** 865-877.

Dasilva, C., H. Hadji, C. Ozouf-Costaz, S. Nicaud, O. Jaillon, J. Weissenbach, and H. Roest Crollius. 2002. Remarkable compartmentalization of transposable elements and pseudogenes in the heterochromatin of the Tetraodon nigroviridis genome. *Proc Natl Acad Sci U S A* **99:** 13636-13641.

DeRose-Wilson, L.J. and B.S. Gaut. 2007. Transcription-related mutations and GC content drive variation in nucleotide substitution rates across the genomes of Arabidopsis thaliana and Arabidopsis lyrata. *BMC Evol Biol* **7:** 66.

Devos, K.M., J.K. Brown, and J.L. Bennetzen. 2002. Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. *Genome Res* **12:** 1075-1079.

Dray, T. and G.B. Gloor. 1997. Homology requirements for targeting heterologous sequences during P-induced gap repair in Drosophila melanogaster. *Genetics* **147:** 689-699.

Drouaud, J., C. Camilleri, P.Y. Bourguignon, A. Canaguier, A. Berard, D. Vezon, S. Giancola, D. Brunel, V. Colot, B. Prum *et al.* 2006. Variation in crossing-over rates across chromosome 4 of Arabidopsis thaliana reveals the presence of meiotic recombination "hot spots". *Genome Res* **16:** 106-114.

Edgar, R.C. and E.W. Myers. 2005. PILER: identification and classification of genomic repeats. *Bioinformatics* **21 Suppl 1:** i152-158.

Elemento, O. and O. Gascuel. 2002. An efficient and accurate distance based algorithm to reconstruct tandem duplication trees. *Bioinformatics* **18 Suppl 2:** S92-99.

Elemento, O., O. Gascuel, and M.P. Lefranc. 2002. Reconstructing the duplication history of tandemly repeated genes. *Mol Biol Evol* **19:** 278-288.

Ellegren, H. 2004. Microsatellites: simple sequences with complex evolution. *Nat Rev Genet* **5:** 435-445.

Engels, W.R., D.M. Johnson-Schlitz, W.B. Eggleston, and J. Sved. 1990. High-frequency P element loss in Drosophila is homolog dependent. *Cell* **62:** 515-525.

Ferguson, D.O. and W.K. Holloman. 1996. Recombinational repair of gaps in DNA is asymmetric in Ustilago maydis and can be explained by a migrating D-loop model. *Proc Natl Acad Sci U S A* **93:** 5419-5424.

Feschotte, C., N. Jiang, and S.R. Wessler. 2002. Plant transposable elements: where genetics meets genomics. *Nat Rev Genet* **3:** 329-341.

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

Fischle, W., Y. Wang, S.A. Jacobs, Y. Kim, C.D. Allis, and S. Khorasanizadeh. 2003. Molecular basis for the discrimination of repressive methyl-lysine marks in histone H3 by Polycomb and HP1 chromodomains. *Genes Dev* **17:** 1870-1881.

Fiston-Lavier, A.S., D. Anxolabehere, and H. Quesneville. 2007. A model of segmental duplication formation in Drosophila melanogaster. *Genome Res* **17:** 1458-1470.

Fitch, W.M. 1977. Phylogenies constrained by the crossover process as illustrated by human hemoglobins and a thirteen-cycle, eleven-amino-acid repeat in human apolipoprotein A-I. *Genetics* **86:** 623-644.

Formosa, T. and B.M. Alberts. 1986. Purification and characterization of the T4 bacteriophage uvsX protein. *J Biol Chem* **261:** 6107-6118.

Fransz, P., S. Armstrong, C. Alonso-Blanco, T.C. Fischer, R.A. Torres-Ruiz, and G. Jones. 1998. Cytogenetics for the model system Arabidopsis thaliana. *Plant J* **13:** 867-876.

Fransz, P.F., S. Armstrong, J.H. de Jong, L.D. Parnell, C. van Drunen, C. Dean, P. Zabel, T. Bisseling, and G.H. Jones. 2000. Integrated cytogenetic map of chromosome arm 4S of A. thaliana: structural organization of heterochromatic knob and centromere region. *Cell* **100:** 367-376.

Gascuel, O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol Biol Evol* **14:** 685-695.

Giroux, M.J., M. Clancy, J. Baier, L. Ingham, D. McCarty, and L.C. Hannah. 1994. De novo synthesis of an intron by the maize transposable element Dissociation. *Proc Natl Acad Sci U S A* **91:** 12150-12154.

Gray, Y.H. 2000. It takes two transposons to tango: transposable-element-mediated chromosomal rearrangements. *Trends Genet* **16:** 461-468.

Green, P. 2007. 2x genomes--does depth matter? *Genome Res* **17:** 1547-1549.
Gregory, T.R. and P.D. Hebert. 1999. The modulation of DNA content: proximate causes and ultimate consequences. *Genome Res* **9:** 317-324.
Guindon, S. and O. Gascuel. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52:** 696-704.
Haber, J.E. 1999. DNA recombination: the replication connection. *Trends Biochem Sci* **24:** 271-275.
Hall, A.E., G.C. Kettler, and D. Preuss. 2006. Dynamic evolution at pericentromeres. *Genome Res* **16:** 355-364.
Harada, K., K. Yukuhiro, and T. Mukai. 1990. Transposition rates of movable genetic elements in Drosophila melanogaster. *Proc Natl Acad Sci U S A* **87:** 3248-3252.
Hasegawa, M., H. Kishino, and T. Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol* **22:** 160-174.
Henry, Y., M. Bedhomme, and G. Blanc. 2006. History, protohistory and prehistory of the Arabidopsis thaliana chromosome complement. *Trends Plant Sci* **11:** 267-273.
Hill, W.G. and A. Robertson. 1966. The effect of linkage on limits to artificial selection. *Genet Res* **8:** 269-294.
Holliday, R. 1964. The Induction of Mitotic Recombination by Mitomycin C in Ustilago and Saccharomyces. *Genetics* **50:** 323-335.
Ira, G. and J.E. Haber. 2002. Characterization of RAD51-independent break-induced replication that acts preferentially with short homologous sequences. *Mol Cell Biol* **22:** 6384-6392.
Jarman, A.P. and R.A. Wells. 1989. Hypervariable minisatellites: recombinators or innocent bystanders? *Trends Genet* **5:** 367-371.
Ji, Y., E.E. Eichler, S. Schwartz, and R.D. Nicholls. 2000. Structure of chromosomal duplicons and their role in mediating human genomic disorders. *Genome Res* **10:** 597-610.
Jurka, J. 2000. Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet* **16:** 418-420.
Kapitonov, V.V. and J. Jurka. 1999. The long terminal repeat of an endogenous retrovirus induces alternative splicing and encodes an additional carboxy-terminal sequence in the human leptin receptor. *J Mol Evol* **48:** 248-251.
Kapitonov, V.V. and J. Jurka. 2001. Rolling-circle transposons in eukaryotes. *Proc Natl Acad Sci U S A* **98:** 8714-8719.
Kapitonov, V.V. and J. Jurka. 2006. Self-synthesizing DNA transposons in eukaryotes. *Proc Natl Acad Sci U S A* **103:** 4540-4545.
Kapitonov, V.V. and J. Jurka. 2007. Helitrons on a roll: eukaryotic rolling-circle transposons. *Trends Genet* **23:** 521-529.
Karlin, S. and C. Burge. 1995. Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet* **11:** 283-290.
Kaufmann, B.P. 1939. Distribution of Induced Breaks along the X-Chromosome of Drosophila Melanogaster. *Proc Natl Acad Sci U S A* **25:** 571-577.
Kogoma, T. 1997. Stable DNA replication: interplay between DNA replication, homologous recombination, and transcription. *Microbiol Mol Biol Rev* **61:** 212-238.
Koszul, R., S. Caburet, B. Dujon, and G. Fischer. 2004. Eucaryotic genome evolution through the spontaneous duplication of large chromosomal segments. *Embo J* **23:** 234-243.
Kunze, R. 1996. The maize transposable element activator (Ac). *Curr Top Microbiol Immunol* **204:** 161-194.

Lachner, M., D. O'Carroll, S. Rea, K. Mechtler, and T. Jenuwein. 2001. Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. *Nature* **410:** 116-120.

Lander, E.S. L.M. Linton B. Birren C. Nusbaum M.C. Zody J. Baldwin K. Devon K. Dewar M. Doyle W. FitzHugh *et al.* 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.

Lee, S.E., J.K. Moore, A. Holmes, K. Umezu, R.D. Kolodner, and J.E. Haber. 1998. Saccharomyces Ku70, mre11/rad50 and RPA proteins regulate adaptation to G2/M arrest after DNA damage. *Cell* **94:** 399-409.

Lerat, E., C. Biemont, and P. Capy. 2000. Codon usage and the origin of P elements. *Mol Biol Evol* **17:** 467-468.

Linardopoulou, E.V., E.M. Williams, Y. Fan, C. Friedman, J.M. Young, and B.J. Trask. 2005. Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. *Nature* **437:** 94-100.

Lippman, Z., A.V. Gendrel, M. Black, M.W. Vaughn, N. Dedhia, W.R. McCombie, K. Lavine, V. Mittal, B. May, K.D. Kasschau *et al.* 2004. Role of transposable elements in heterochromatin and epigenetic control. *Nature* **430:** 471-476.

Locke, D.P., N. Archidiacono, D. Misceo, M.F. Cardone, S. Deschamps, B. Roe, M. Rocchi, and E.E. Eichler. 2003. Refinement of a chimpanzee pericentric inversion breakpoint to a segmental duplication cluster. *Genome Biol* **4:** R50.

Lydeard, J.R., S. Jain, M. Yamaguchi, and J.E. Haber. 2007. Break-induced replication and telomerase-independent telomere maintenance require Pol32. *Nature* **448:** 820-823.

Maside, X., S. Assimacopoulos, and B. Charlesworth. 2000. Rates of movement of transposable elements on the second chromosome of Drosophila melanogaster. *Genet Res* **75:** 275-284.

Maside, X., C. Bartolome, S. Assimacopoulos, and B. Charlesworth. 2001. Rates of movement and distribution of transposable elements in Drosophila melanogaster: in situ hybridization vs Southern blotting data. *Genet Res* **78:** 121-136.

Maxson, R., R. Cohn, L. Kedes, and T. Mohun. 1983. Expression and organization of histone genes. *Annu Rev Genet* **17:** 239-277.

Mc, C.B. 1950. The origin and behavior of mutable loci in maize. *Proc Natl Acad Sci U S A* **36:** 344-355.

McClintock, B. 1929. Chromosome Morphology in Zea Mays. *Science* **69:** 629.

McVey, M., M. Adams, E. Staeva-Vieira, and J.J. Sekelsky. 2004. Evidence for multiple cycles of strand invasion during repair of double-strand gaps in Drosophila. *Genetics* **167:** 699-705.

Montgomery, E., B. Charlesworth, and C.H. Langley. 1987. A test for the role of natural selection in the stabilization of transposable element copy number in a population of Drosophila melanogaster. *Genet Res* **49:** 31-41.

Morrow, D.M., C. Connelly, and P. Hieter. 1997. "Break copy" duplication: a model for chromosome fragment formation in Saccharomyces cerevisiae. *Genetics* **147:** 371-382.

Mosig, G. 1987. The essential role of recombination in phage T4 growth. *Annu Rev Genet* **21:** 347-371.

Myers, E.W., G.G. Sutton, A.L. Delcher, I.M. Dew, D.P. Fasulo, M.J. Flanigan, S.A. Kravitz, C.M. Mobarry, K.H. Reinert, K.A. Remington *et al.* 2000. A whole-genome assembly of Drosophila. *Science* **287:** 2196-2204.

Nassif, N., J. Penney, S. Pal, W.R. Engels, and G.B. Gloor. 1994. Efficient copying of nonhomologous sequences from ectopic sites via P-element-induced gap repair. *Mol Cell Biol* **14:** 1613-1625.

Needleman, S.B. and C.D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48:** 443-453.

Nei, M. and A.P. Rooney. 2005. Concerted and birth-and-death evolution of multigene families. *Annu Rev Genet* **39:** 121-152.

Newman, T.L., E. Tuzun, V.A. Morrison, K.E. Hayden, M. Ventura, S.D. McGrath, M. Rocchi, and E.E. Eichler. 2005. A genome-wide survey of structural variation between human and chimpanzee. *Genome Res* **15:** 1344-1356.

Nurminsky, D.I., Y. Shevelyov, S.V. Nuzhdin, and V.A. Gvozdev. 1994. Structure, molecular evolution and maintenance of copy number of extended repeated structures in the X-heterochromatin of Drosophila melanogaster. *Chromosoma* **103:** 277-285.

Nuzhdin, S.V. and T.F. Mackay. 1995. The genomic rate of transposable element movement in Drosophila melanogaster. *Mol Biol Evol* **12:** 180-181.

Nuzhdin, S.V., E.G. Pasyukova, and T.F. Mackay. 1997. Accumulation of transposable elements in laboratory lines of Drosophila melanogaster. *Genetica* **100:** 167-175.

Nuzhdin, S.V. 1999. Sure facts, speculations, and open questions about the evolution of transposable element copy number. *Genetica* **107:** 129-137.

O'Donnell, K.A. and J.D. Boeke. 2007. Mighty Piwis defend the germline against genome intruders. *Cell* **129:** 37-44.

Ozenberger, B.A. and G.S. Roeder. 1991. A unique pathway of double-strand break repair operates in tandemly repeated genes. *Mol Cell Biol* **11:** 1222-1231.

Paques, F. and J.E. Haber. 1999. Multiple pathways of recombination induced by double-strand breaks in Saccharomyces cerevisiae. *Microbiol Mol Biol Rev* **63:** 349-404.

Peacock, W.J., E.S. Dennis, M.M. Rhoades, and A.J. Pryor. 1981. Highly repeated DNA sequence limited to knob heterochromatin in maize. *Proc Natl Acad Sci U S A* **78:** 4490-4494.

Pereira, V. 2004. Insertion bias and purifying selection of retrotransposons in the Arabidopsis thaliana genome. *Genome Biol* **5:** R79.

Petrov, D.A., E.R. Lozovskaya, and D.L. Hartl. 1996. High intrinsic rate of DNA loss in Drosophila. *Nature* **384:** 346-349.

Petrov, D.A., Y.T. Aminetzach, J.C. Davis, D. Bensasson, and A.E. Hirsh. 2003. Size matters: non-LTR retrotransposable elements and ectopic recombination in Drosophila. *Mol Biol Evol* **20:** 880-892.

Pfeiffer, P., W. Goedecke, and G. Obe. 2000. Mechanisms of DNA double-strand break repair and their potential to induce chromosomal aberrations. *Mutagenesis* **15:** 289-302.

Preston, C.R., J.A. Sved, and W.R. Engels. 1996. Flanking duplications and deletions associated with P-induced male recombination in Drosophila. *Genetics* **144:** 1623-1638.

Quesneville, H., D. Nouaud, and D. Anxolabehere. 2003. Detection of new transposable element families in Drosophila melanogaster and Anopheles gambiae genomes. *J Mol Evol* **57 Suppl 1:** S50-59.

Quesneville, H., C.M. Bergman, O. Andrieu, D. Autard, D. Nouaud, M. Ashburner, and D. Anxolabehere. 2005. Combined evidence annotation of transposable elements in genome sequences. *PLoS Comput Biol* **1:** 166-175.

Quesneville, H., N. Buisine, and V. Colot. 2008. Transposable element re-annotation of the Arabidopsis genome.
Rew, D.A. 2004. The sequencing of the rat genome. *Eur J Surg Oncol* **30:** 905-906.
Reznikoff, W.S., A. Bhasin, D.R. Davies, I.Y. Goryshin, L.A. Mahnke, T. Naumann, I. Rayment, M. Steiniger-White, and S.S. Twining. 1999. Tn5: A molecular window on transposition. *Biochem Biophys Res Commun* **266:** 729-734.
Rice, J.C., S.D. Briggs, B. Ueberheide, C.M. Barber, J. Shabanowitz, D.F. Hunt, Y. Shinkai, and C.D. Allis. 2003. Histone methyltransferases direct different degrees of methylation to define distinct chromatin domains. *Mol Cell* **12:** 1591-1598.
Rice, P., I. Longden, and A. Bleasby. 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16:** 276-277.
Richard, G.F. and F. Paques. 2000. Mini- and microsatellite expansions: the recombination connection. *EMBO Rep* **1:** 122-126.
Richardson, C., M.E. Moynahan, and M. Jasin. 1998. Double-strand break repair by interchromosomal recombination: suppression of chromosomal translocations. *Genes Dev* **12:** 3831-3842.
Roeder, G.S. and G.R. Fink. 1980. DNA rearrangements associated with a transposable element in yeast. *Cell* **21:** 239-249.
Rong, Y.S. and K.G. Golic. 2003. The homologous chromosome is an effective template for the repair of mitotic DNA double-strand breaks in Drosophila. *Genetics* **165:** 1831-1842.
Rubin, E. and A.A. Levy. 1997. Abortive gap repair: underlying mechanism for Ds element formation. *Mol Cell Biol* **17:** 6294-6302.
Rubin, G.M. and E.B. Lewis. 2000. A brief history of Drosophila's contributions to genome research. *Science* **287:** 2216-2218.
Samonte, R.V. and E.E. Eichler. 2002. Segmental duplications and the evolution of the primate genome. *Nat Rev Genet* **3:** 65-72.
SanMiguel, P., A. Tikhonov, Y.K. Jin, N. Motchoulskaia, D. Zakharov, A. Melake-Berhan, P.S. Springer, K.J. Edwards, M. Lee, Z. Avramova *et al.* 1996. Nested retrotransposons in the intergenic regions of the maize genome. *Science* **274:** 765-768.
SanMiguel, P., B.S. Gaut, A. Tikhonov, Y. Nakajima, and J.L. Bennetzen. 1998. The paleontology of intergene retrotransposons of maize. *Nat Genet* **20:** 43-45.
She, X., Z. Jiang, R.A. Clark, G. Liu, Z. Cheng, E. Tuzun, D.M. Church, G. Sutton, A.L. Halpern, and E.E. Eichler. 2004. Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature* **431:** 927-930.
Shevelyov, Y.Y. 1992. Copies of a Stellate gene variant are located in the X heterochromatin of Drosophila melanogaster and are probably expressed. *Genetics* **132:** 1033-1037.
Shields, D.C. and P.M. Sharp. 1989. Evidence that mutation patterns vary among Drosophila transposable elements. *J Mol Biol* **207:** 843-846.
Shirasu, K., A.H. Schulman, T. Lahaye, and P. Schulze-Lefert. 2000. A contiguous 66-kb barley DNA sequence provides evidence for reversible genome expansion. *Genome Res* **10:** 908-915.
Stankiewicz, P. and J.R. Lupski. 2002. Genome architecture, rearrangements and genomic disorders. *Trends Genet* **18:** 74-82.
Sugawara, N., F. Paques, M. Colaiacovo, and J.E. Haber. 1997. Role of Saccharomyces cerevisiae Msh2 and Msh3 repair proteins in double-strand break-induced recombination. *Proc Natl Acad Sci U S A* **94:** 9214-9219.

Sun, F.L., M.H. Cuaycong, C.A. Craig, L.L. Wallrath, J. Locke, and S.C. Elgin. 2000. The fourth chromosome of Drosophila melanogaster: interspersed euchromatic and heterochromatic domains. *Proc Natl Acad Sci U S A* **97:** 5340-5345.

Sun, H., D. Treco, and J.W. Szostak. 1991. Extensive 3'-overhanging, single-stranded DNA associated with the meiosis-specific double-strand breaks at the ARG4 recombination initiation site. *Cell* **64:** 1155-1161.

Szostak, J.W., T.L. Orr-Weaver, R.J. Rothstein, and F.W. Stahl. 1983. The double-strand-break repair model for recombination. *Cell* **33:** 25-35.

Thompson, J.D., D.G. Higgins, and T.J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22:** 4673-4680.

Tulin, A.V., G.L. Kogan, D. Filipp, M.D. Balakireva, and V.A. Gvozdev. 1997. Heterochromatic Stellate gene cluster in Drosophila melanogaster: structure and molecular evolution. *Genetics* **146:** 253-262.

Venter, J.C. M.D. Adams E.W. Myers P.W. Li R.J. Mural G.G. Sutton H.O. Smith M. Yandell C.A. Evans R.A. Holt *et al.* 2001. The sequence of the human genome. *Science* **291:** 1304-1351.

Vinogradov, A.E. 2003. Selfish DNA is maladaptive: evidence from the plant Red List. *Trends Genet* **19:** 609-614.

Vinogradov, A.E. 2004. Genome size and extinction risk in vertebrates. *Proc Biol Sci* **271:** 1701-1705.

Virgin, J.B. and J.P. Bailey. 1998. The M26 hotspot of Schizosaccharomyces pombe stimulates meiotic ectopic recombination and chromosomal rearrangements. *Genetics* **149:** 1191-1204.

Vitte, C., O. Panaud, and H. Quesneville. 2007. LTR retrotransposons in rice (Oryza sativa, L.): recent burst amplifications followed by rapid DNA loss. *BMC Genomics* **8:** 218.

Waring, G.L. and J.C. Pollack. 1987. Cloning and characterization of a dispersed, multicopy, X chromosome sequence in Drosophila melanogaster. *Proc Natl Acad Sci U S A* **84:** 2843-2847.

Waterston, R.H. K. Lindblad-Toh E. Birney J. Rogers J.F. Abril P. Agarwal R. Agarwala R. Ainscough M. Alexandersson P. An *et al.* 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520-562.

Weber, J.L. and E.W. Myers. 1997. Human whole-genome shotgun sequencing. *Genome Res* **7:** 401-409.

Wicker, T., N. Yahiaoui, R. Guyot, E. Schlagenhauf, Z.D. Liu, J. Dubcovsky, and B. Keller. 2003. Rapid genome divergence at orthologous low molecular weight glutenin loci of the A and Am genomes of wheat. *Plant Cell* **15:** 1186-1197.

Wicker, T., F. Sabot, A. Hua-Van, J.L. Bennetzen, P. Capy, B. Chalhoub, A. Flavell, P. Leroy, M. Morgante, O. Panaud *et al.* 2007. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet* **8:** 973-982.

Wright, S.I., N. Agrawal, and T.E. Bureau. 2003. Effects of recombination rate and gene density on transposable element distributions in Arabidopsis thaliana. *Genome Res* **13:** 1897-1903.

Yang, G., F. Zhang, C.N. Hancock, and S.R. Wessler. 2007. Transposition of the rice miniature inverted repeat transposable element mPing in Arabidopsis thaliana. *Proc Natl Acad Sci U S A* **104:** 10962-10967.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* **13:** 555-556.

Yang, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24:** 1586-1591.