Origines des séquences microsatellites dans les génomes eucaryotes

Sébastien Leclercq

UNIVERSITE MONTPELLIER II
SCIENCES ET TECHNIQUES DU LANGUEDOC

T H E S E

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITE MONTPELLIER II

Discipline : EERGP - Evolution, Ecologie, Ressources Génétiques, Paléontologie

Ecole Doctorale : SIBAGHE

présentée et soutenue publiquement

par

Sébastien Leclercq

Le 19 décembre 2007

Titre :

Origines des séquences microsatellites dans les génomes eucaryotes

- leçons tirées de la séquence du génome humain -

JURY

M.	Emmanuel Douzery	, Président
Mme.	Marie-France Sagot	, Rapporteur
M.	Guy-Franck Richard	, Rapporteur
M.	Philippe Jarne	, Directeur de thèse
M.	Eric Rivals	, Directeur de thèse
Mme	Gwenael Piganeau	, Examinateur

Avant-propos

Malgré la nature explicite du titre de cette thèse, il me semble nécessaire de commencer ce manuscrit par une brève introduction situant le contexte du travail effectué, ne serait-ce que pour prévenir l'astro-physicien un peu distrait qui se serait fait leurrer par le terme « microsatellite ». En effet, quel nom inapproprié pour ces éléments génomiques ! Ils ne sont pas du tout satellite de quoi que se soit, mais sont au contraire distribués quasiment aléatoirement dans les séquences génomiques. De plus, leur nature répétée et variable n'apparaît pas dans cette dénomination. Il aurait donc pu être plus judicieux d'utiliser les termes « STR »(pour Small Tandem Repeat) ou « VNTR »(Variable Number Tandem Repeat), mais pour des raisons poétiques et historiques [1], nous garderons le nom de « microsatellite ».

L'intérêt principal des microsatellites, et qui a stimulé les recherches sur le sujet, est leur capacité à être de bons marqueurs génétiques. Ils sont abondants dans les génomes (pour la très grande majorité des êtres vivants), neutres, co-dominants, et surtout hyper-variables en taille. Ces propriétés en ont fait un outil majeur en médecine, pour les tests de paternité, et en biologie des populations, pour évaluer le lien de parenté entre individus. Cette vision des microsatellites en tant qu'outil a donné lieu à une focalisation de la recherche sur leur hyper-variabilité et ses conséquences, c'est-à-dire essayer de comprendre comment on pouvait obtenir les distributions en taille des allèles observées dans les populations. La dynamique des éléments existants est donc aujourd'hui assez connue, comme en atteste les modèles théoriques actuels, qui s'ajustent relativement bien aux données réelles.

Il reste bien entendu un grand nombre de choses à préciser à ce niveau, surtout concernant l'impact des interruptions, mais qui demandent une connaissance fine du sujet, que je ne possédais pas au départ de ma thèse. Mon attention s'est alors porté non pas sur la dynamique des éléments existants, mais sur l'apparition de nouveaux microsatellites. La phase primordiale de création reste encore largement méconnue, et seuls quelques exemples issus de comparaisons phylogénétiques (plus quelques rares modèles théoriques) en font état. En effectuant l'état de l'art sur la question, il est apparu que deux voies majeures pouvaient conduire à l'apparition des microsatellites : l'apparition via les éléments transposables, et l'apparition *de novo*, directement à partir de la séquence génomique existante (par mutation ponctuelle). J'ai donc suivi ces deux voies de manière indépendante, ce qui a abouti aux deux travaux des chapitres 4 et 5 de cette thèse.

[1] Les microsatellites sont des satellites de très petite taille, ces derniers ayant été caractérisés pour la première fois comme des bandes en marge de la molécule ADN (donc « satellites »), lors de processus de purification de l'ADN.

Mais avant toute investigation, s'est posée la question de la méthode. De part ma formation informatique, et mes lacunes en techniques de biologie moléculaire (qui en est une conséquence), l'utilisation des génomes disponibles sous forme de séquences informatique s'est naturellement imposée. L'analyse de séquences génomiques implique l'utilisation de logiciels pour traiter ces séquences, or il en existe plusieurs dédiés à l'extraction des microsatellites. C'est pourquoi je me suis tout d'abord concentré sur la comparaison de divers algorithmes disponibles, afin d'évaluer l'influence de la méthode sur les données obtenues. Les résultats de cette étude est présentés dans le chapitre 3.

Les travaux présentés ici sont limités à l'analyse du génome humain (sauf ceux du chapitre 3, réalisés sur quatre génomes différents). Pourquoi une telle restriction, alors qu'un grand nombre de séquences d'autres organismes sont disponibles dans les banques de données ? Premièrement, parce que le génome humain contient les éléments transposables Alu, très nombreux, et dont l'association avec les microsatellites avait déjà été caractérisée. Il existait donc une base de travail à ce niveau là. Deuxièmement, la disponibilité de deux autres génomes proches séquencés (le chimpanzé et le macaque) a permis de réaliser des analyses de génomique comparative. Alors pourquoi avoir étendu le titre aux génomes eucaryotes ? Parce que je pense que les résultats présentés dans cette thèse sont valables pour l'ensemble des organismes qui possèdent les mêmes contraintes moléculaires que l'homme, du moins dans les grandes lignes. Etendre les analyses à d'autres organismes est bien sûr envisageable (et envisagé), mais demande une somme de travail qui n'était pas réalisable dans le cadre de cette thèse.

Je profite enfin de cet avant-propos pour effectuer mes remerciements, nombreux et sincères. Je remercie avant tout Philippe et Eric, qui ont su être disponibles aux moments où j'en avais besoin, et dont les discussions m'ont souvent recadré durant ces trois années. Merci à vous. Je remercie aussi Patrice, bien sûr, pour son accueil dans son équipe, même si je faisais figure d'extraterrestre au milieu de toute cette génétique des populations. Les deux autres membres de mon comité de thèse, Nicolas Galtier et Olivier Gascuel ont aussi leur place dans ces remerciements, pour leurs encouragements et leurs réflexions. Je remercie particulièrement Erick Desmarais pour les quelques discussions, trop rares mais toujours instructives, que nous avons eu à propos la relation entre microsatellites et éléments transposables. Enfin, je tiens à remercier Antoine de Daruvar et le CBiB de Bordeaux, qui m'ont donné le goût de la recherche en biologie, et m'ont extirpé de mon destin d'ingénieur informaticien dans une quelconque SSII parisienne.

Dans un cadre moins scientifique, mais tout aussi important, je me dois de remercier tous les

membres de l'équipe GenDyn, présents et passés (en particulier Guillaume, mon compagnon de ukulélé), pour le bon temps passé en leur compagnie. Viennent ensuite naturellement mes amis doctorants, du CEFE et d'ailleurs, pour les multiples soirées, repas, promenades, jeux et détentes en tout genre. Vous êtes malheureusement bien trop nombreux pour pouvoir être tous cités ici, mais mon attention se porte particulièrement sur le noyau dur que sont Anne-Violette, Claire, Anne, François et Juan. Merci d'avance à la FDMV pour le voyage à Venise. Merci à Thierry, Samuel, Manue, Emilie et Juliette, pour avoir dépassé les frontières des disciplines au nom de l'amitié. Enfin, merci à ceux qui sont loin des yeux mais au fond du coeur : mes parents, Rina (où est-tu maintenant ?), et mes amis de Touraine.

Abbréviations

CDB : Cassure Double Brin

DSBR : Double Strand Break Repair

IAM : Infinite Allele Model

indel(s) : insertion(s)-délétion(s)

LINE(s) : Long INterspersed Element(s)

Ma : Million d'années

Mb : Mégabase

NHEJ : Non Homologous End Joining

nt : nucléotide

SDSA : Synthesis-Dependant Strand Annealing

SINE(s) : Short INterspersed Element(s)

SMM : Stepwise Mutation Model

SSA : Single Strand Annealing

SSM : Slipped-Strand Mispairing

TPM : Two-Phase Model

WDP : Wraparound Dynamic Programming

Table des matières

1 Evolution moléculaire de l'ADN non-codant **1**

1.1 Le monde du non-codant . . . 1

1.2 Les mécanismes de mutation . . . 4

1.2.1 Mutations ponctuelles . . . 5

1.2.2 Déamination des cytosines méthylées . . . 5

1.2.3 Mécanismes de la recombinaison . . . 7

1.2.4 Transposition d'éléments mobiles . . . 12

1.2.5 Glissement de polymérase . . . 14

1.2.6 Fréquence des mutations . . . 15

1.3 Les séquences Alu . . . 17

1.3.1 Généralités . . . 17

1.3.2 Mécanismes de la rétrotransposition . . . 18

1.3.3 Les différentes familles . . . 20

2 Les microsatellites **23**

2.1 Description générale . . . 23

2.1.1 Les répétitions en tandem . . . 23

2.1.2 Définitions et structure moléculaire . . . 25

2.1.3 Distributions dans les génomes . . . 29

2.2 Les différentes méthodes d'analyse . . . 32

2.2.1 Les analyses de mutation directe et de variabilité . . . 32

2.2.2 Les analyses phylogénétiques . . . 33

2.2.3 Les analyses de séquences . . . 34

2.3 Les processus de mutation . . . 36

2.3.1 La théorie du glissement de polymérase (SSM) . . . 36

2.3.2 Taux de mutation corrélé à la longueur du microsatellite . . . 37

2.3.3 Pas multiples . . . 38
2.3.4 Importance du motif . . . 39
2.3.5 Interruptions stabilisantes . . . 40
2.3.6 Biais de mutation . . . 40
2.4 Le cycle de vie des microsatellites . . . 41
2.4.1 Modèles théoriques . . . 41
2.4.2 Modèle biologique . . . 43
2.4.3 Apparition des microsatellites . . . 46

3 Les limites de la détection bio-informatique 49
3.1 Approche bio-informatique . . . 49
3.1.1 Choix du type d'étude . . . 49
3.1.2 Problématique . . . 50
3.2 Méthodes et résultats . . . 53
3.2.1 Description des algorithmes . . . 53
3.2.2 Influence des paramètres . . . 59
3.2.3 Comparaison des algorithmes . . . 64
3.3 Discussion . . . 66

4 Apparition via les séquences Alu 69
4.1 Relation microsatellites - séquences Alu . . . 69
4.1.1 Etat de l'art . . . 69
4.1.2 Problématique . . . 73
4.2 Méthodes et résultats . . . 74
4.2.1 Extraction des données . . . 74
4.2.2 Méthodes de calcul de la proximité . . . 78
4.2.3 Proximité entre microsatellites et éléments Alu . . . 80
4.2.4 Influence de la famille Alu . . . 83
4.2.5 Taille des microsatellites associés aux séquences Alu . . . 86
4.3 Discussion . . . 91
4.3.1 Ré-évaluation de l'association entre microsatellites et séquences Alu . . . 91
4.3.2 Réduction de la taille des microsatellites . . . 97
4.3.3 Apparition des microsatellites à partir du polyA . . . 98
4.3.4 Apparition en interne . . . 101
4.3.5 Conclusion . . . 104

5 Apparition *de novo* **105**

5.1 Une taille minimum pour les microsatellites ? . . . 105

5.1.1 Une sur-représentation des locus courts . . . 106

5.1.2 Cas d'apparition de répétitions en tandem . . . 110

5.1.3 Estimer l'importance du glissement pour les petites tailles . . . 112

5.2 Méthodes et résultats . . . 113

5.2.1 Méthode d'alignement . . . 113

5.2.2 Calcul de la sur-représentation . . . 118

5.2.3 Etude de la micro-duplication . . . 121

5.2.4 Etude du glissement pour les répétitions de petite taille . . . 126

5.3 Discussion . . . 129

6 Dissertation et conclusion **137**

6.1 Un modèle d'apparition des microsatellites . . . 137

6.1.1 Apparition des doublons . . . 139

6.1.2 Développement des microsatellites . . . 140

6.2 Les diverses implications du modèle . . . 141

6.2.1 Sur la distribution des microsatellites . . . 141

6.2.2 Sur la construction des modèles théoriques . . . 145

6.3 Conclusion . . . 150

6.3.1 Synthèse . . . 150

6.3.2 Vers un envahissement du génome ? . . . 152

Chapitre 1

Evolution moléculaire de l'ADN non-codant

1.1 Le monde du non-codant

Nous savons, depuis les travaux de Morgan sur les drosophiles, que notre information génétique est portée par nos chromosomes, constitués d'acides désoxyribonucléiques (ADN). En 1953, Watson et Crick ([Watson and Crick, 1953]) en découvrent la structure en double hélice, à partir de laquelle s'est fondée la biologie moléculaire moderne, basée sur les théories de la réplication de l'ADN, de la transcription en acide ribonucléique (ARN), et de la traduction en protéines. La compréhension de ces mécanismes moléculaires a permis de faire un bond énorme en médecine et en pharmacologie, car les pathologies ont pu être analysées au niveau des réactions physico-chimiques des molécules (enzymes, hormones, sucres, etc.), et de leur production. Dès lors, on comprend que le gène, dont est issu chaque protéine utilisée par le métabolisme, ait été tant mis en valeur et étudié. Il reste d'ailleurs encore énormément de choses à découvrir, lorsque l'on voit le nombre de gènes de fonction inconnue dans le génome humain, pourtant l'un des plus étudiés.

Toutefois, les génomes des organismes vivants ne sont pas constitués uniquement de séquences codantes, mais possèdent aussi des fragments dits intergéniques, situés entre les différents gènes. Ces zones, dont l'ampleur n'a été découverte que dans les années 70 [Doolittle and Sapienza, 1980], ont été qualifiées d'*ADN poubelle*, c'est-à-dire n'ayant aucune fonction particulière et évoluant au gré de mutations aléatoires. Pourtant ces zones non-codantes recèlent une richesse inouïe en termes de diversité de séquences, de mécanismes moléculaires et de fonctions pour l'organisme. Il y a bien sûr la présence des introns, des promoteurs et des sites de régulation, qui permettent l'expression

correcte des gènes et qui ont donc leur importance dans le développement et la survie des organismes. D'autres zones, telles que les points chauds de recombinaison, les centromères et télomères, les isochores, n'ont pas de lien direct avec la survie de l'individu, mais jouent un rôle dans l'évolution des génomes. Enfin, il existe des éléments tels que les répétitions en tandem et les éléments transposables qui ne possèdent pas de fonction à proprement parler, mais dont la dynamique est régie par des mécanismes complexes et certainement pas strictement aléatoires.

Il est évident que l'étude du non-codant, en plus d'être passionnante, est d'un intérêt primordial pour comprendre la structure et la dynamique des génomes. En effet, la comparaison de la taille des génomes pour les divers règnes du vivant montre une corrélation très nette entre complexité des organismes et la taille de leur génome, ceux des bactériophages et virus étant très simples et très petits, alors que les eucaryotes complexes ont un génome de grande taille (Figure 1.1). Or, chez les virus et les procaryotes, les séquences codantes occupent 80 à 90% de la totalité de la séquence d'ADN, taux qui est réduit à moins de 50% en général chez les eucaryotes supérieurs, et jusqu'à 2-3% chez certains mammifères. L'expansion du non-codant a donc peut-être joué un rôle dans la diversification et la complexification du vivant. Ces résultats ne sont toutefois basés que sur une comparaison à très grande échelle, et la corrélation n'est pas forcément valable entre organismes d'un même règne. Par exemple, *A. thaliana* et le maïs, toutes deux des plantes de complexité équivalente, ont un rapport de taille de génome de plus de 20. Des exemples comme celui-ci sont souvent présentés dans la littérature, mais ils sont basés sur quelques organismes atypiques et ne reflètent pas forcément la tendance générale, comme l'explique Michael Lynch dans son ouvrage 'The origins of genome architecture' ([Lynch, 2007]).

La dynamique du non-codant répond aux mêmes contraintes que l'évolution des gènes, à savoir un équilibre entre les mutations subies par la séquence et les effets populationnels que sont la sélection et la dérive génétique. La sélection peut jouer sur l'évolution du non-codant à plusieurs niveaux. Tout d'abord, non-codant ne veut pas dire non-fonctionnel et certaines zones (généralement des sites de régulation) sont sous une forte pression sélective, parfois plus importante que certaines régions codantes ([Andolfatto, 2005]). Par ailleurs, l'évolution de certains éléments du non-codant peut parfois affecter des zones codantes et avoir une action délétère. Ce sont, par exemple, les insertions d'éléments transposables dans des gènes, ou l'expansion de microsatellites instables induisant des maladies neurodégénératives dites « à triplets » (voir chapitre 2). La sélection peut donc empêcher certaines mutations sur des zones non-codantes d'être transmises, ou même promouvoir des moyens de contrer ces mutations, telle la redondance chez les eucaryotes supérieurs ([Krakauer and Plotkin,

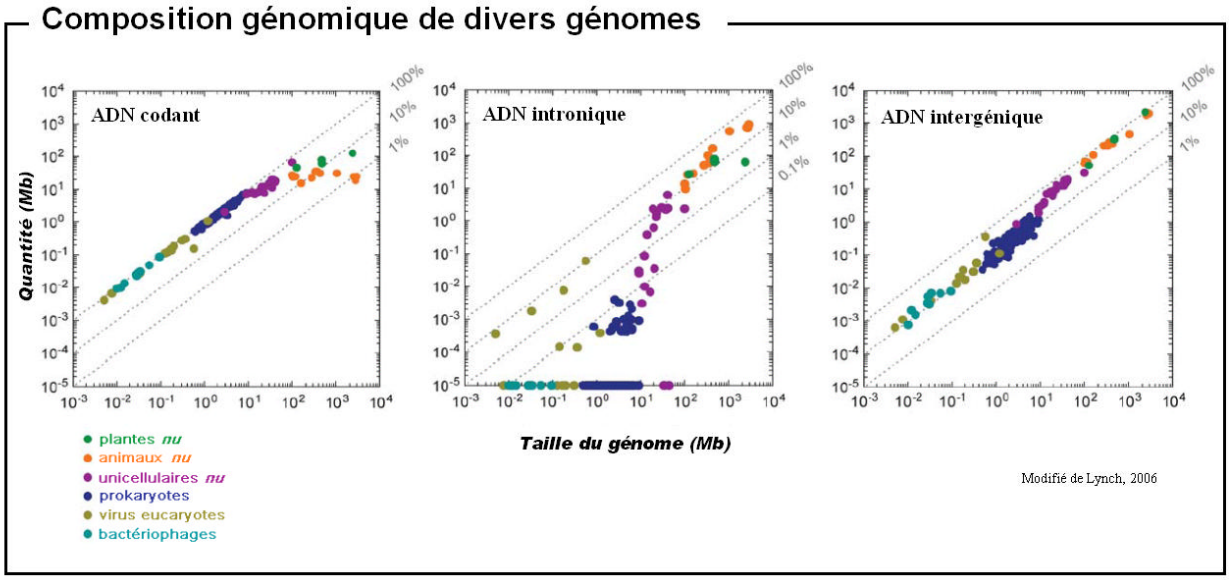


FIG. 1.1 – Composition génomique, en ADN codant, intronique, et intergénique, en fonction de la taille du génome, pour divers organismes classés par règne. les lignes pointillées indiquent le rapport entre la quantité et la taille du génome. La ligne la plus haute indique le ratio 1 :1, les lignes inférieures les ratios de 1 :10, 1 :100 et 1 :1000 respectivement.

2002]). Pour autant, les éléments génomiques concernés par cette thèse ne sont pas soumis, de manière générale, aux effets de la sélection directe. Les microsatellites sont en effet réputés neutres dans les génomes, les phénomènes de sélection et leurs implications ne seront donc que très peu abordés dans ce manuscrit.

En revanche, la dérive génétique n'est pas limitée aux régions sous sélection, et son action peut avoir des répercussions importantes sur l'organisation des génomes, particulièrement au niveau du non-codant non fonctionnel. En effet, sous le seul effet de la dérive génétique, une mutation neutre se fixe dans une population avec une probabilité égale à l'inverse de deux fois la taille efficace de cette population (valeur théorique en rapport avec le nombre d'individus lorsque la population est stable ; voir [Hartl and Clark, 1997]). Les génomes d'espèces ayant de faibles tailles efficaces de population (comme beaucoup d'eucaryotes supérieurs) peuvent donc présenter des mutations issues de mécanismes moléculaires rares, mais qui se sont fixées rapidement. Les causes et effets de la dérive génétique sont par conséquent indissociables de toute étude d'évolution moléculaire, et peuvent être (sont) souvent des sujets de thèse à part entière. Pourtant, les travaux présentés ici s'émancipent des contraintes liées à la dérive, grâce aux méthodes employées. Nous allons travailler sur des génomes complètement séquencés et disposer de millions de locus différents. Les interprétations seront alors basées sur un très grand nombre d'observations, qui ne peuvent pas tous être la conséquence

d'événements isolés, mais bien issus de mécanismes évolutifs généraux.

La suite de ce chapitre est consacrée à la présentation de la dernière contrainte qui structure les régions non-codantes des génomes : les mutations.

1.2 Les mécanismes de mutation

Les génomes sont soumis à toutes sortes de mutations, qui sont le moteur de l'évolution, selon la théorie Darwinienne. Cette thèse concerne l'évolution des microsatellites, nous allons donc nous concentrer sur les mécanismes de mutation en rapport avec ces derniers. L'influence des mutations ponctuelles et des éléments transposables sur l'apparition des microsatellites étant les sujets traités au cours de ma thèse, il est indispensable de les présenter ici. La recombinaison et le glissement de polymérase sont connus pour avoir des effets sur la dynamique des microsatellites ; ils seront donc détaillés eux aussi. De plus, les mutations pouvant être parfois délétères, les organismes vivants ont acquis des systèmes moléculaires de réparation de l'ADN. Nous ne détaillerons pas (volontairement) les systèmes de réparation, car nous supposerons que les mutations présentées sont justement celles ayant échappé à la correction. Quelques-uns de ces mécanismes seront toutefois cités dans la suite, et une brève explication de leur fonction sera alors donnée.

On appelle mutation n'importe quel changement dans une séquence d'ADN qui peut être transmise à sa descendance. Il y a deux sources majeures de mutation : les dommages physiques sur l'ADN et les erreurs lors de la réplication ([Baer et al., 2007]). Les dommages physiques peuvent être exogènes (provoqués par des rayons UV par exemple) ou endogènes (provoqués par des résidus de réactions métaboliques). Les erreurs de réplication sont caractérisées par des appariements erronés lors de la synthèse d'un nouveau brin d'ADN à partir d'un brin matrice. Les taux de mutations sont, par contre, très variables selon les mécanismes, comme nous le présenterons à la fin de cette section.

Il faut aussi souligner que la descendance n'a pas la même signification pour les organismes unicellulaires et multicellulaires. Pour un organisme multicellulaire, les seules mutations qui sont transmises à la descendance sont celles qui se produisent dans les lignées germinales. Toutes les mutations se produisant dans d'autres tissus sont appelées mutations somatiques. D'un point de vue évolutif, seules les mutations germinales ont une importance, la notion de mutation se référera donc généralement à ces dernières dans les chapitres suivants.

1.2.1 Mutations ponctuelles

Le type de mutation le plus courant que subit le génome est la mutation ponctuelle. Elle se définit comme la substitution d'une base azotée par une autre au niveau de la séquence génétique (Figure 1.2.1). La transformation d'une purine (A ou G) en une autre purine, ou d'une pyrimidine (C ou T) en une autre est appelée transition, tandis que les transformations purine-pyrimidine ou pyrimidine-purine sont appelées transversions (Figure 1.2.2). En plus des substitutions, il peut arriver que la séquence subisse des insertions ou des délétions (indels) de quelques bases.

Les mutations ponctuelles interviennent en majorité lors de la synthèse de l'ADN, à cause des erreurs de la polymérase (qui recopie le brin d'ADN). Cela signifie que la réplication de l'ADN est une source importante de mutations ponctuelles, mais il est possible qu'elles se produisent aussi lors d'autres phases de synthèse, comme par exemple lors de la réparation de cassures double-brin ou l'intégration d'éléments exogènes (voir sections 1.2.3 et 1.2.4). Parmi les mutations ponctuelles, les transitions sont nettement plus courantes que les transversions, et les indels sont plutôt rares. A titre d'exemple, on peut citer l'étude de Haag-Liautard *et al.* [Haag-Liautard et al., 2007], qui recense les mutations arrivées dans différentes lignées de drosophiles. Ils obtiennent 45% de transitions, contre 22,5% de transversions, 22,5% de indels et 10% d'événements complexes et transpositions. Pour souligner la variabilité entre les organismes, on peut comparer les résultats de Haag-Liautard *et al.* à ceux obtenus pour l'homme par Nachman *et al.* [Nachman and Crowell, 2000], qui donnent une répartition à hauteur de 66% de transitions pour 26% de transversions et 8% d'indels. Ces derniers n'ont toutefois pas relevé les événements complexes.

1.2.2 Déamination des cytosines méthylées

La déamination des cytosines est une dégradation de l'ADN assez courante, pouvant être provoquée par plusieurs types de stress (*i.e.* d'événements affectant le génome) [Wang et al., 1980]. Elle consiste en la transformation d'une base cytosine (C) en une uracile (U). Cette erreur est normalement détectée par le système de réparation de l'ADN. En revanche, si la cytosine a été méthylée (m5C), le résidu créé par la déamination est une thymine (T). Le système de réparation de l'ADN peut alors corriger le T en C aussi bien que le G complémentaire en A (Figure 1.2.3) ; seul ce dernier cas aboutit à une mutation.

La méthylation des cytosines est un phénomène fréquent chez les eucaryotes, particulièrement les vertébrés et les plantes. Sa fonction n'est pas encore clairement définie, mais il est probable qu'elle soit utilisée pour la reconnaissance de corps exogènes [Klinman et al., 1999], et qu'elle ait un

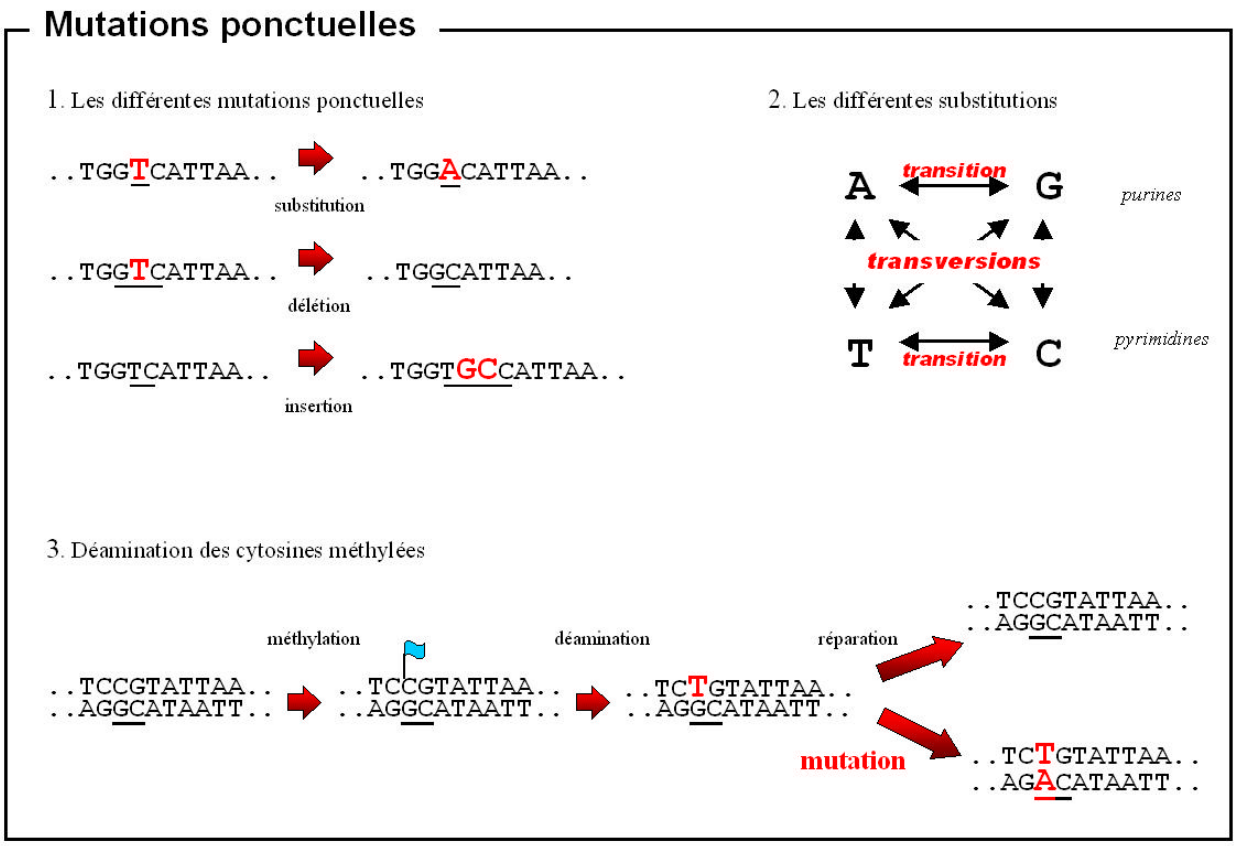


FIG. 1.2 – **1-** Les trois cas possibles de mutation ponctuelle. Les sites soulignés sont les sites de la mutation. **2-** Les différentes substitutions possibles. Les probabilités de transitions et de transversions ne sont pas identiques (voir texte). **3-** Processus de déamination des cytosines méthylées. Le site souligné est le site de mutation. Le drapeau indique la cytosine méthylée.

rôle dans les lignées germinales, pour la méiose par exemple [Oakes et al., 2007]. Cette méthylation affecte à 90% les cytosines des dinucléotides CG [Ehrlich and Wang, 1981], que l'on nomme CpG pour signifier que la guanine suit la cytosine (dans le sens 5'-3').

L'action couplée de la méthylation et de la déamination provoque donc une sur-représentation des transitions C vers T, ou G vers A, sur les sites CpG, avec un taux de 2 à 4 fois supérieur par rapport aux autres sites [Podlutsky et al., 1998]. Ces mutations ne peuvent pas être considérées comme des mutations ponctuelles, car elles sont dépendantes du processus de méthylation qui n'est pas aléatoire. Il convient de préciser ici le terme d'*îlot CpG*, qui désigne des régions chromosomiques où le taux de méthylation est inférieur aux autres régions, avoisinant parfois le zéro. Ce sont principalement des zones en 5' de gènes [Oakes et al., 2007].

1.2.3 Mécanismes de la recombinaison

La recombinaison est le mécanisme moléculaire permettant la réparation de l'ADN lorsque ce dernier subit des cassures double brin (CDB). Les CDB sont accidentelles en temps normal, causées par des stress cellulaires (radiations UV par exemple), mais elles sont provoquées lors de la phase de division méiotique des cellules. Plus de dix gènes sont par exemple responsables des CDB méiotiques chez *Saccharomyces cerevisiae*, augmentant le nombre de recombinaisons de 100 à 1000 fois par rapport aux cellules végétatives [Smith and Nicolas, 1998]. Il existe deux types principaux de recombinaison : la recombinaison homologue et la recombinaison non-homologue, qui diffèrent par leur mécanisme et par les protéines impliquées dans le processus.

Modèles DSBR et SDSA

La réparation de CDB par recombinaison homologue nécessite un appariement de la séquence à réparer avec une région d'ADN homologue. Cette région peut être le même locus sur la chromatide sœur ou le chromosome homologue, on parlera alors de recombinaison allélique. Si la zone d'homologie est une région distante, sur le même chromosome, ou sur un autre, on parlera de recombinaison ectopique. La recombinaison homologue est surtout reconnue comme le mécanisme qui permet un brassage d'information entre séquences d'ADN, via le *crossing-over*. Cet événement est d'ailleurs souvent considéré comme la phase primordiale du sexe en ce qu'elle casse les liaisons entre allèles, et permet un maintien de la diversité au sein des populations [Marais, 2002]. La recombinaison homologue peut être réalisée par plusieurs mécanismes concurrents, avec des conséquences différentes et représentés par divers modèles.

Le premier modèle est appelé modèle *DSBR* (pour Double Strand Break Repair), ou modèle de Szostak *et al.* [Szostak et al., 1983], et est celui qui autorise le plus de crossing-over (Figure 1.3.1). A la suite de la CDB, les brins d'ADN sont dégradés du 3' au 5' pour libérer de l'ADN simple-brin sur au maximum 800 nt. Ces brins vont ensuite créer des jonctions de Holliday avec la région homologue. On appelle *jonction de Holliday* la conformation de l'ADN où quatre brins sont liés simultanément (A et B sur la figure 1.3.1). Une fois l'appariement effectué, l'ADN endommagé utilise l'autre brin comme matrice pour se réparer. Les brins d'ADN sont ensuite sectionnés au niveau des jonctions, se séparant ainsi en deux brins resynthétisés. On appelle cette étape la *résolution des jonctions de Holliday*, qui peut être avec ou sans crossing-over, en fonction des brins qui ont été sectionnés. La probabililité d'avoir une résolution avec crossing-over est *a priori* la même que de celle ne pas en avoir dans les cellules en phase mitotique, mais est plus importante dans les cellules en phase méiotique [Paques and Haber, 1999].

La réparation via le SDSA (pour Synthesis-Dependant Strand Annealing) est quant à elle un mécanisme n'impliquant généralement pas de jonctions de Holliday, mais un réappariement des brins réparés. Plusieurs modèles de SDSA ont été proposés, selon que un seul brin à réparer ou les deux sont re-synthétisés à partir de la région homologue [Formosa and Alberts, 1986, Hastings, 1988, Mcgill et al., 1989]. La réparation débute là encore par la dégradation des brins d'ADN de part et d'autre de la CDB pour libérer de l'ADN simple brin. L'un des deux brins, ou les deux, envahissent ensuite la zone homologue pour entamer leur réparation (figure 1.3.2). La ou les séquences resynthétisées se séparent enfin de la séquence matrice pour se réapparier avec la séquence complémentaire. Si un seul brin avait été réparé, le deuxième brin est synthétisé en prenant la nouvelle séquence comme matrice (figure 1.3.2).

La différence fondamentale entre ces deux modèles de recombinaison est la formation ou non de jonctions de Holliday, qui induisent des événements de crossing-over. Un autre modèle de SDSA a toutefois été proposé qui autorise, dans certains cas, la formation de jonctions de Holliday [Ferguson and Holloman, 1996].

Mutations induites par le DSBR et le SDSA

Les réparations par DSBR ou SDSA sont des mécanismes extrêmement robustes, car il permettent une réparation quasiment sans modification de la séquence originale, à l'exception du réarrangement entre allèles lors des crossing-over et des éventuelles erreurs provoquées lors de la synthèse des brins à réparer. Ils sont toutefois susceptibles de produire des événements de *conversion génique*. La conversion génique peut se caractériser comme le transfert non réciproque d'information génétique d'un brin d'ADN vers un autre. Elle est la conséquence de la formation d'*hétéroduplex* lors de la réparation, c'est-à-dire l'appariement de deux brins similaires mais pas identiques. Ce peut être le cas lorsque la recombinaison s'initie sur le chromosome homologue ayant un allèle différent, ou pour les recombinaisons ectopiques (figure 1.4.1).

Les hétéroduplex présentent des sites avec des erreurs d'appariement qui sont alors normalement réparées par le système de réparation des mésappariements (*MMR*, pour MisMatch Repair). Ce dernier va corriger la majorité des sites erronés, soit en remplaçant la séquence nouvellement synthétisée par la séquence d'origine, soit l'inverse. Dans ce dernier cas, la séquence d'origine sera perdue au profit de celle provenant de la séquence homologue. La conversion génique, tout comme le crossing-over, sont des phénomènes mutagènes dans le sens où ils induisent une nouvelle séquence en combinant les fragments de deux autres.

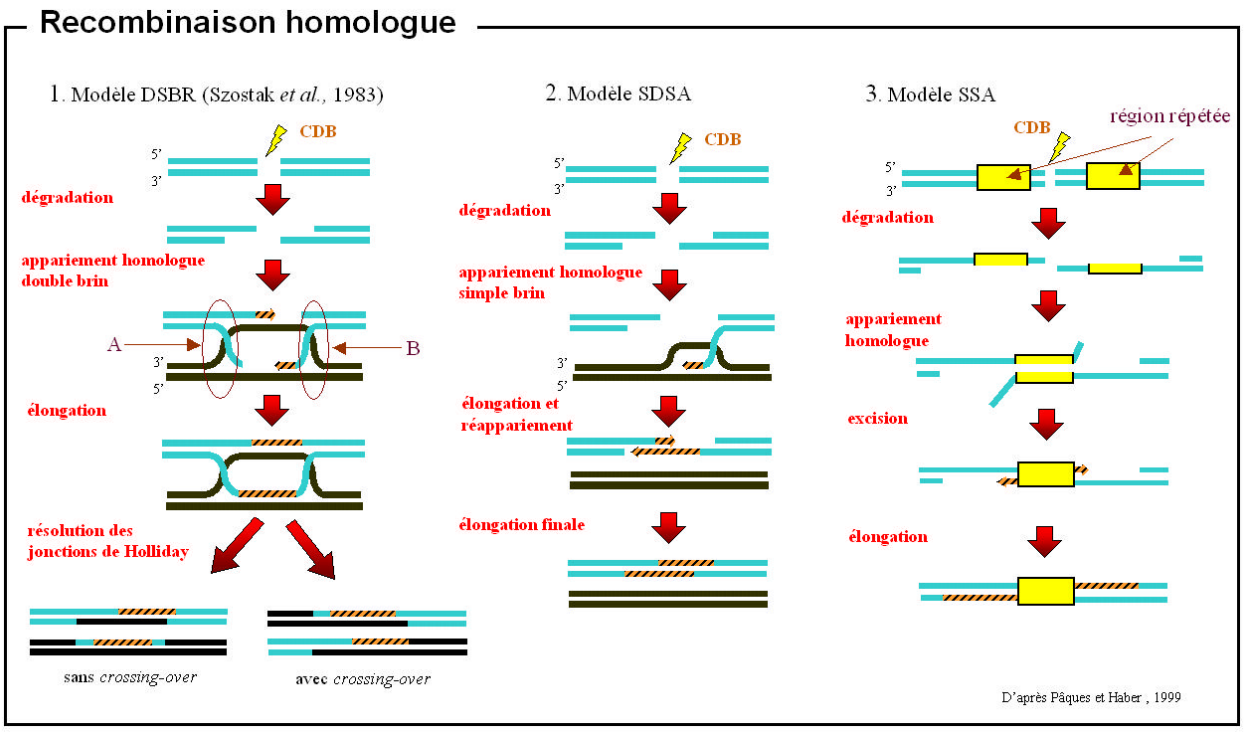


FIG. 1.3 – Les mécanismes de la recombinaison homologue après une cassure double brin (CDB). Les brins bleu clair représentent le fragment d'ADN où a lieu la CDB, les brins noirs le fragment matrice, et les brins zébrés les fragments reconstruits. **1-** Modèle DSBR (Szostak *et al.*, 1983). Les zones A et B sont les sites de jonction de Holliday. **2-** Modèle SDSA (Synthesis-Dependant Strand Annealing). Seul le modèle avec l'appariement d'un seul brin au brin matrice est représenté. **3-** Modèle SSA (Single Strand Annealing). Aucun brin matrice n'est nécessaire, l'appariement homologue se fait directement à partir de régions répétées de part et d'autre de la CDB.

Outre la conversion génique, la recombinaison homologue peut aussi impliquer des *crossing-over inégaux*. On appelle crossing-over inégal, un crossing-over lors d'une recombinaison ectopique [Lynch, 2007]. Ces événements conduisent à la duplication de la région située entre les répétitions sur l'une des molécules d'ADN, ainsi qu'à l'une de ces répétitions, et à une délétion de cette région sur l'autre molécule (figure 1.4.2). Les événements de duplication et délétion ne peuvent par contre se produire que lorsque les régions d'homologies sont situées sur deux chromatides sœurs ou chromosomes homologues.

Enfin, la recombinaison homologue, lorsque qu'elle est réalisée par SDSA, peut aboutir à un réappariement décalé si elle se produit dans une zone de répétition [Richard and Paques, 2000]. La figure 1.4.3 schématise ce processus. Une fois la synthèse terminée à partir du brin homologue, le brin réparé se réapparie à son complémentaire. Si la séquence est répétée, il est possible que le

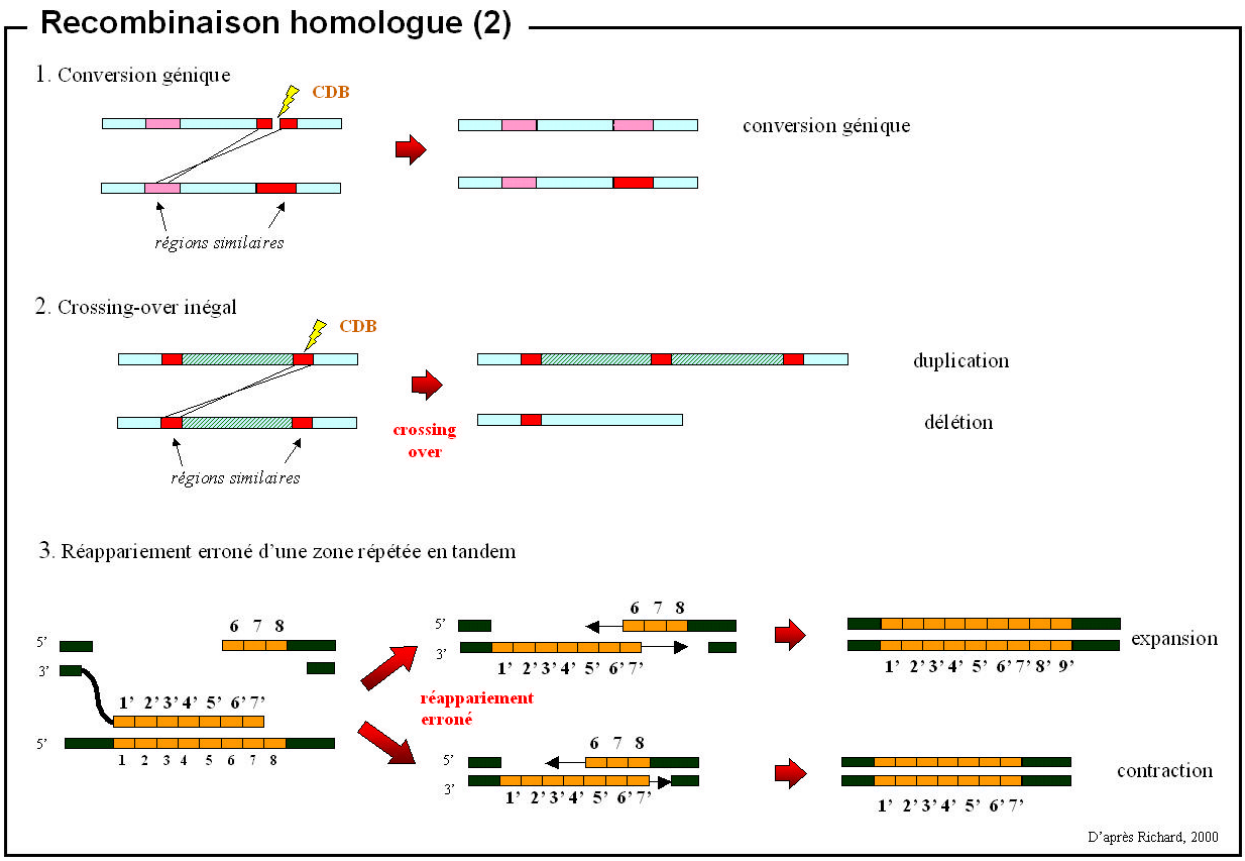


FIG. 1.4 – Les mutations provoquées lors de recombinaisons homologues **1-** Conversion génique. La séquence de la région réparée est remplacée par la séquence qui a servi à sa réparation. Le schéma est présenté pour une recombinaison ectopique mais est aussi valable pour une recombinaison allélique entre deux allèles différents sur deux chromosomes homologues. **2-** Crossing-over inégal. Le crossing-over inégal ne peut se produire que lors d'une recombinaison ectopique et provoque la duplication d'un fragment d'ADN dans une séquence, et sa délétion dans l'autre séquence. **3-** Réappariement décalé après une réparation par recombinaison SDSA, dans une zone répétée en tandem.

réappariement ne se fasse pas sur la bonne répétition. Cela peut provoquer une contraction ou une expansion du nombre de répétitions, selon que le réappariement se fait en amont ou en aval de la position correcte.

SSA

Il existe un troisième type de recombinaison homologue, mais qui ne nécessite cette fois pas de brin d'ADN matrice supplémentaire. Ce mécanisme, appelé *SSA* (pour Single Strand Annealing), apparie directement des régions homologues distantes sur les brins complémentaires (figure 1.3.3). La dégradation de l'ADN de part et d'autre de la CDB est beaucoup plus longue que pour les modèles DSBR et SDSA (pouvant aller jusqu'à 15 Kb), afin de trouver des régions répétées [Paques

and Haber, 1999]. Une fois l'appariement des régions effectuées, les fragments non-homologues sont excisés et le brin complémentaire est réparé.

Ce mécanisme, comme le crossing-over inégal, est vecteur de mutations au sens propre, car il implique la délétion de toute la région chromosomique située entre les deux zones homologues, plus une des deux zones. Il est aussi susceptible de provoquer de la conversion génique au niveau de la répétition résiduelle, car l'appariement a de grandes chances de produire un hétéroduplex. Elle ne peut par contre pas provoquer de crossing-over, puisque qu'aucune autre séquence d'ADN n'est mise à contribution.

Recombinaison non-homologue

Bien que la recombinaison homologue soit le mécanisme principal pour réparer les CDB, il existe certains cas où son recours s'avère impossible (absence de région homologue, système de réparation endommagé par des mutations, etc.). Les cellules font alors appel à un second type de mécanisme, appelé recombinaison non-homologue, ou illégitime [Paques and Haber, 1999]. Deux possibilités sont envisageables. La première est le ré-appariement direct des régions complémentaires produites par le clivage. La ligation directe permet une réparation sans erreur et sans recours à une tierce séquence (figure 1.5.1)

La seconde possibilité est un réappariement non-homologue des deux brins complémentaires, appelé *NHEJ* (pour Non Homologous End Joining), qui nécessite une micro-homologie d'environ 1 à 3 nucléotides. L'appariement peut être décalé en aval des brins (figure 1.5.2a), créant des fragments simple brin qui sont comblés par un processus de *remplissage* (« filling in » en anglais). Ce mécanisme produit une duplication de quelques bases au niveau de la zone réparée. A l'inverse, si l'appariement est décalé en amont sur les brins, quelques bases seront supprimées.

Il est aussi possible que la recherche de micro-homologie se fasse sur une distance plus grande. Ce mécanisme implique une dégradation des brins de part et d'autre de la CDB, puis un appariement au niveau de la micro-homologie, suivi d'une excision des régions non-homologues (figure 1.5.2b). Le NHEJ a pour conséquence, dans ces cas-là, la délétion de toute la région située entre les deux zones de micro-homologie (de quelques bases à plusieurs kilobases). Le processus est relativement similaire à celui du SSA mais n'implique pas les mêmes voies métaboliques, ce n'est donc pas de la recombinaison homologue.

De plus, les mécanismes de NHEJ sont facilités si la CDB a eu lieu dans une séquence répétée,

car chaque répétition peut être utilisée comme zone de micro-homologie.

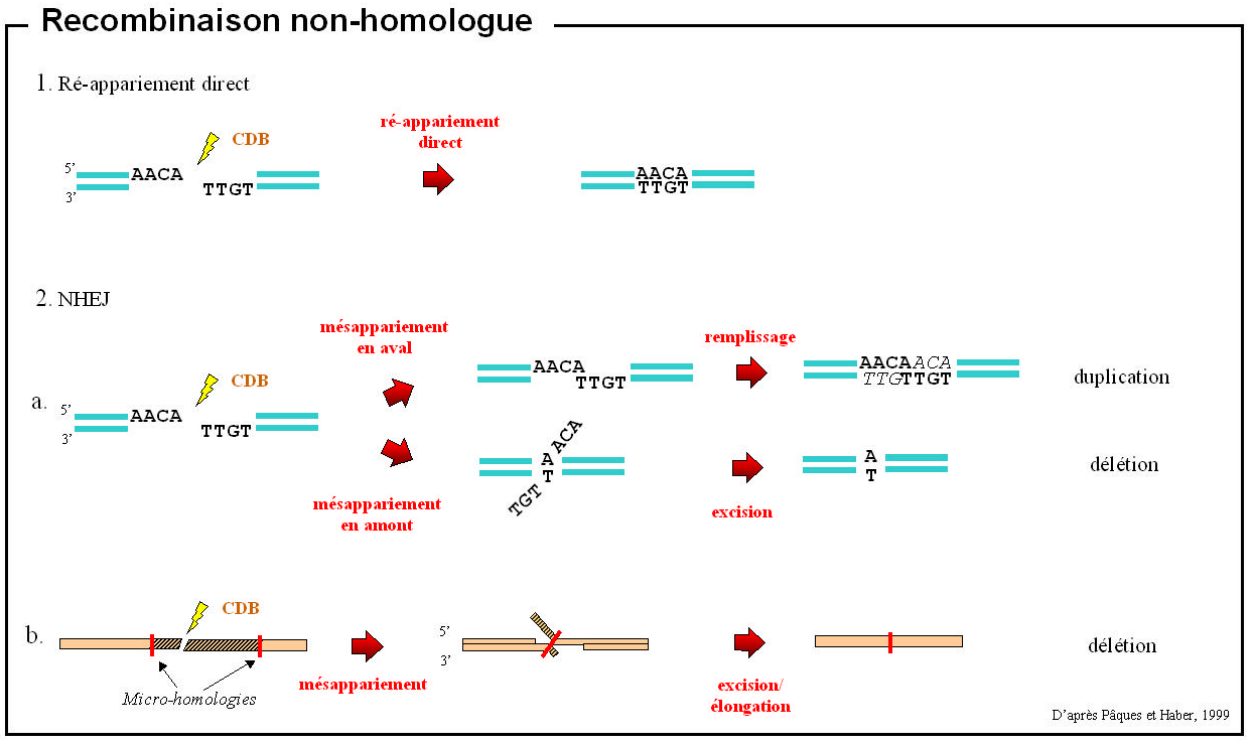


FIG. 1.5 – Les mécanismes de recombinaison non-homologue. **1-** Réappariement direct des brins clivés. La complémentarité permet une ligation des brins sans perte de séquence. **2a-** NHEJ (NonHomologous End Joining). Le réappariement se fait de manière décalée et abouti à la perte ou à la duplication de quelques bases. **2b-** Le réappariement se fait à des sites distants de micro-homologies, et aboutit à la délétion de la région située entre les deux positions.

1.2.4 Transposition d'éléments mobiles

La *transposition* est définie comme le déplacement ou la copie d'un élément chromosomique à une nouvelle position [Charlesworth et al., 1994], créant de ce fait des variations de taille plus ou moins importantes au niveau local ou global, selon l'efficacité des éléments. L'efficacité est mesurée par ce que l'on nomme l'*amplification* de ces éléments transposables, qui peut être définie comme le nombre de nouveaux éléments insérés dans le génome durant un temps donné. Les facteurs jouant sur cette amplification sont assez divers, mais le mode de transposition est l'un des principaux. On peut regrouper les éléments transposables en trois grandes classes selon leur mode de transposition : les transposons, les rétrotransposons avec LTR et les rétrotransposons sans LTR [Eickbush and Eickbush, 2005].

Les transposons

Les éléments de la première classe sont appelés simplement transposons. Ils possèdent une région fonctionnelle, qui code la transposase. Cette enzyme va reconnaître les bornes du transposon et le cliver (le détacher de la séquence d'ADN), pour le ré-insérer à une nouvelle position sur le même ou un autre chromosome (Figure 1.6.1). Ce mécanisme de « couper-coller » ne produit *a priori* pas de nouvelle séquence, mais l'amplification des transposons est tout de même assurée de deux manières. D'une part, le clivage de la séquence d'origine va produire une CDB, qui sera réparée par un événement de recombinaison. Or, nous avons vu dans le paragraphe précédent que la recombinaison provoque généralement une conversion génique, c'est-à-dire la réparation via la copie de la séquence homologue. Si cette séquence possède le transposon (ce qui est forcément le cas pour la chromatide sœur, et est possible pour le chromosome homologue), le transposon d'origine sera re-synthétisé, tandis que la copie clivée sera insérée ailleurs. D'autre part, un grand nombre de transpositions semblent survenir lors de la réplication de l'ADN. Si un transposon déjà répliqué va se ré-insérer à une position en amont de la fourche de réplication, il sera répliqué une seconde fois. Dans ces cas-là, la séquence répliquée où ne s'est pas produit le clivage contiendra deux copies de l'élément [Eickbush and Eickbush, 2005].

Les rétrotransposons avec LTR (Long Terminal Repeat)

La deuxième classe est appelée rétrotransposons avec LTR. Contrairement aux transposons, ces éléments ne sont pas excisés du génome, mais sont transcrits en ARN. Cet ARN code une enzyme particulière nommée rétrotranscriptase, qui recrée une séquence d'ADN à partir de la séquence d'ARN de l'élément transposable (Figure 1.6.2). L'ADN complémentaire ainsi reconstruit est alors réinséré à une nouvelle position dans le génome par le biais d'une seconde enzyme, l'intégrase, elle aussi codée par l'élément transposable. Ce mécanisme induit donc forcément une copie de l'élément [Eickbush and Eickbush, 2005]. Ces éléments sont nommés avec LTR, car ils possèdent une longue séquence terminale (300 à 500 nt), répétée en 3' et 5', qui semble nécessaire au processus de rétrotransposition [Eickbush and Malik, 2001].

Les rétrotransposons non LTR

Le troisième type d'éléments transposables est très proche du second, à savoir qu'il passe lui aussi par une phase de transcription en ARN, puis par une réintégration à une nouvelle position génomique. La différence tient en ce que l'ARN n'est pas rétrotranscrit directement, mais clive le site d'intégration grâce à une enzyme appelée endonucléase [Eickbush and Malik, 2001]. La séquence clivée est ensuite utilisée comme amorce pour rétrotranscrire l'ARN de l'élément transposable. Enfin,

l'ARN est éjecté, la jonction entre le brin synthétisé et le brin d'intégration est fait, et le brin d'ADN complémentaire est synthétisé (figure 1.6.3). Ces transposons sont appelés simplement rétrotransposons sans LTR car ils ne possèdent pas la séquence répétée caractéristique. Les rétrotransposons sans LTR peuvent de plus être divisés en deux sous-classes, les LINEs et SINEs [Eickbush and Malik, 2001]. Les LINEs (Long INterspersed Elements) contiennent tout le matériel génétique nécessaire à la production de leurs enzymes de rétrotransposition, alors que les SINEs (Short INterspersed Elements) ne codent aucune enzyme. Ces derniers parasitent en fait la machinerie des LINEs pour pouvoir se copier [Flavell, 1995].

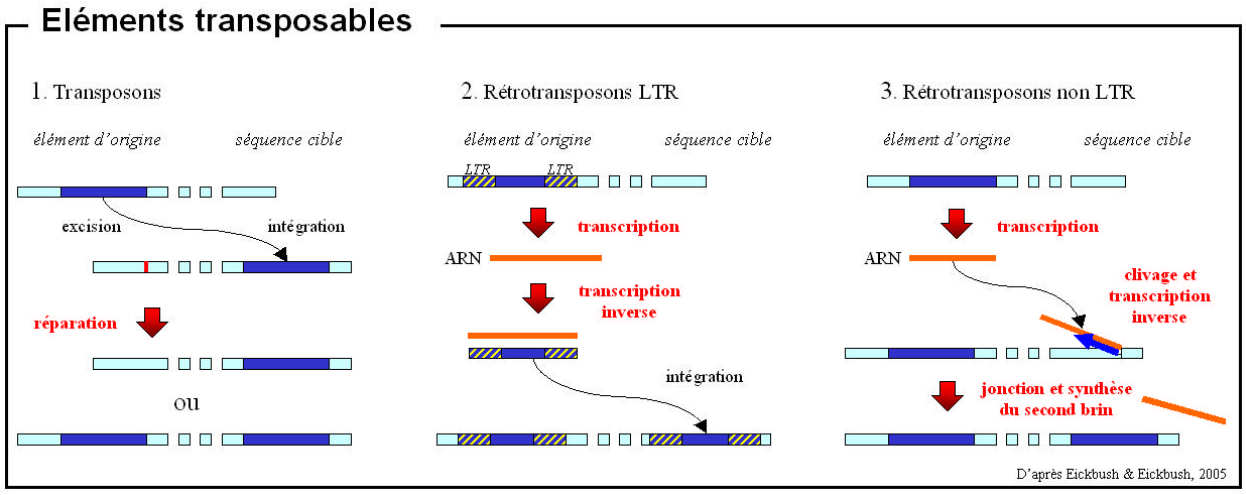


FIG. 1.6 – Les différentes classes d'éléments transposables, et leur mode de transposition. **1-** Les transposons. L'élément transposable est excisé de l'ADN, pour être ré-inséré ailleurs. La réparation resynthétise éventuellement l'élément à sa position d'origine, à partir de la séquence homologue. Ce mode de transposition n'induit pas forcément de duplication de l'élément. **2-** Les rétrotransposons avec LTR. L'élément transposable est transcrit en ARN, qui est rétrotranscrit avant d'être ré-inséré à une position distante. Ce mode de transposition induit forcément une copie de l'élément. **3-** Les rétrotransposons sans LTR. L'élément transposable est transcrit en ARN, qui utilise la séquence d'intégration comme amorce pour se rétrotranscrire. Ce mode de transposition induit aussi forcément une copie de l'élément.

1.2.5 Glissement de polymérase

La mutation par glissement (SSM, pour Slipped Strand Mispairing [Levinson and Gutman, 1987b]) intervient durant la réplication de l'ADN. Lors de la synthèse du nouveau brin par les polymérases, il peut arriver que l'extrémité du brin en synthèse se détache (désapparie) du brin matrice. A part ralentir le processus, cela ne cause pas de préjudice car le brin peut se ré-apparier grâce à la complémentarité de bases avec le brin matrice. En revanche, le désappariement dans une zone répétée en tandem peut causer un décalage lors de la réplication. En effet, le ré-appariement

peut se faire sur l'une des répétitions adjacentes car la complémentarité est respectée, on dit alors que la polymérase a « glissé » d'une ou plusieurs répétitions (Figure 1.7). Si le ré-appariement se fait en amont, la répétition sera synthétisée deux fois, il y aura donc ajout d'une répétition dans le nouveau brin d'ADN. A l'inverse, avec un ré-appariement en aval, l'une des répétitions ne sera pas synthétisée du tout, aboutissant à la perte d'une répétition dans l'ADN résultant (Figure 1.7).

Un événement de mésappariement ne produit pas systématiquement une mutation en taille, car la plupart sont détectés par le système de réparation MMR [Levinson and Gutman, 1987a, Strand et al., 1993] (qui intervient directement après la synthèse et vérifie que le nouveau brin est conforme à la matrice). En effet, le mésappariement produit une boucle, soit sur le brin matrice, soit sur le nouveau brin, qui est reconnue par le système MMR, si elle fait moins d'une dizaine de nucléotides. Selon le brin sur lequel la boucle se situe, elle sera excisée, ou le brin complémentaire sera clivé afin de re-synthétiser les bases manquantes (voir figure 1.7). Ces réparations ne sont pas systématiques, et le glissement de polymérase est considéré comme le facteur principal de l'évolution des microsatellites, lorsqu'il se produit dans les lignées germinales (voir chapitre 2).

On peut par ailleurs noter que le glissement n'est pas limité à la synthèse durant la réplication de l'ADN, mais peut aussi se produire lors des événements de SDSA [Richard and Paques, 2000]. Le principe reste le même : le brin en synthèse peut se désapparier de la séquence matrice et se réapparier de manière décalée. Si l'erreur n'est pas détectée à ce moment là, la mutation sera effective, car le brin erroné sert ensuite de matrice pour la réparation du brin complémentaire (voir section 1.2.3).

1.2.6 Fréquence des mutations

Nous avons donné jusqu'ici les principaux mécanismes de mutation qui peuvent influencer l'apparition, la disparition, et plus généralement la dynamique des microsatellites. Tous ces événements n'ont par contre pas la même probabilité de se produire dans les génomes, et beaucoup de facteurs peuvent favoriser certaines mutations plus que d'autres. Nous allons tenter dans cette section de mettre en relation les différents taux de mutations observés pour chacun des mécanismes. Ces valeurs sont extrêmement dépendantes des méthodes utilisées et organismes observés, il convient donc de s'attacher surtout aux ordres de grandeurs de ces résultats.

Les mutations ponctuelles tout d'abord. Leur taux est extrêmement variable selon les organismes, et même selon les sites, ce qui peut d'ailleurs poser un certain nombre de problèmes lors de la reconstruction des phylogénies entre espèces à partir de données moléculaires [Blanquart, 2007]. Elles ont

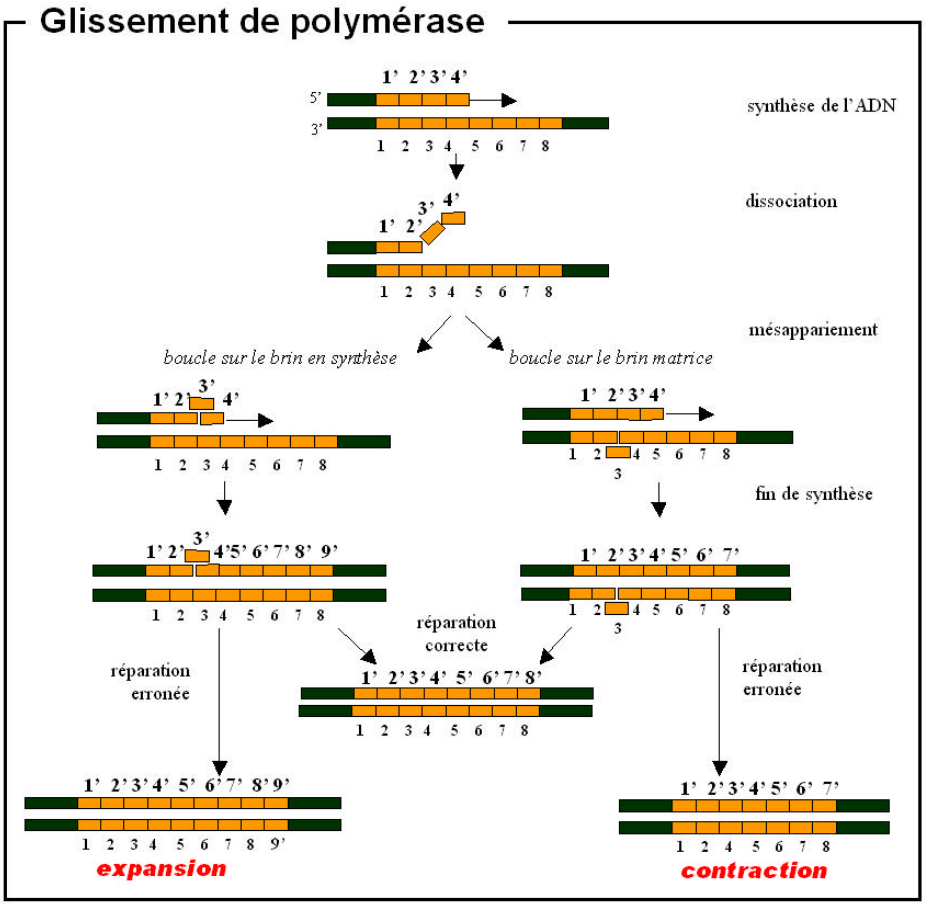


FIG. 1.7 – Le glissement de la polymérase lors de la réplication d'une séquence répétée. Les répétitions sont caractérisées par leur numéro.

été néanmoins calculées de manière globale pour divers organismes. Par exemple, chez la drosophile, le taux a été estimé à approximativement $8{,}4\text{x}10^{-9}$ mutations par site par génération [Haag-Liautard et al., 2007]. Ce taux est légèrement supérieur chez l'homme, puisqu'il a été estimé entre $1{,}3\text{x}10^{-9}$ et $2{,}7\text{x}10^{-9}$ (ces taux ayant été obtenus à partir de la divergence entre les génomes de l'homme et du chimpanzé, l'incertitude provient de l'âge supposé de la spéciation et des tailles de population efficaces ; mais voir [Kimura, 1980]). Ces taux prennent en compte les mutations aux sites CpG, qui sont environ deux à quatre fois plus fréquentes que les autres mutations ponctuelles[Podlutsky et al., 1998].

La recombinaison, quant à elle, est difficile à quantifier, car elle ne produit pas systématiquement de mutation. Elle est de plus variable selon la position génomique, avec des régions ayant de forts taux de recombinaison, ou *points chauds* de recombinaison (« recombination hot spots » en anglais) et d'autres avec de faibles taux (*points froids* ou « cold spots ») [Petes, 2001]. Li [Li, 1997] présente

par exemple des taux de conversion génique chez les levures allant de 0,5% à 18% par événement de méiose, selon les locus (avec une moyenne à 4-5%). Il faut de plus rappeler que les événements de recombinaison dans les cellules végétatives sont 100 à 1000 fois nombreuses que dans les cellules sexuelles, selon ce qui a été rapporté chez la levure [Smith and Nicolas, 1998].

Enfin, le taux de mutation par glissement est aussi relativement fréquent par rapport aux mutations ponctuelles et dépend d'un certain nombre de paramètres. Il a été estimé entre 10^{-2} et 10^{-6} événements par locus par génération pour les séquences répétées microsatellites, selon l'espèce, le nombre de répétitions, le type du motif répété [Ellegren, 2004]. Les facteurs induisant la variabilité de ces taux de glissement seront détaillés dans le chapitre 2.

1.3 Les séquences Alu

Les séquences Alu sont des éléments transposables SINEs spécifiques aux primates, de la classe des rétrotransposons sans LTR. Ces séquences ont la propriété d'être vecteurs de microsatellites et d'être très présentes dans le génome humain (plus de 1 million de copies). Le chapitre 4 de cette thèse est entièrement consacré à la relation Alu-microsatellites, et la section qui suit a pour but de donner quelques clés utiles à sa compréhension.

1.3.1 Généralités

Les éléments Alu sont constitués de deux monomères semblables d'environ 130 nucléotides (nt), reliés par une liaison $(A)_{5-6}CAT(A)_{5-6}$, nommée *linker*. L'ensemble fait à peu près 280 nt (légèrement variable selon l'élément Alu), auquel il faut ajouter une queue polyA d'environ 20 nt, elle aussi de taille variable (Figure 1.8.1). Les éléments Alu sont flanqués de deux répétitions directes de 5 à 15 nt, créées lors de l'insertion de l'élément [Batzer and Deininger, 2002]. Ils ont été identifiés comme éléments répétés dans le génome durant les années 70, et ont pris le nom de Alu car ils possèdent un site de reconnaissance pour l'enzyme de restriction *AluI*. Ils représentent à eux seuls presque 10% du génome humain, malgré leur petite taille [Lander et al., 2001].

A l'origine, les deux monomères étaient dissociés, et devaient certainement rétrotransposer indépendamment [Zietkiewicz et al., 1998]. Ils sont appelés FLA (pour Free Left Alu) et FRA (Free Right Alu), et des reliques de ces anciens éléments sont encore visibles dans le génome humain. Ces deux monomères sont issus de la séquence ARN du gène 7SL qui fait partie du complexe ribosomique [Batzer and Deininger, 2002]. A la suite d'une duplication, le gène se serait transformé en un SINE,

nommé FAM (Fossil Alu Monomer), événement qui pourrait dater d'environ 110 millions d'années (Ma) selon certaines estimations [Kapitonov and Jurka, 1996]. Le FAM aurait ensuite produit deux sous-lignées, les FLA et FRA, jusqu'à ce qu'un FRA s'insère dans la queue polyA d'un FLA, donnant naissance à l'hétérodimère Alu. L'âge des premiers éléments Alu sous leur forme complète est estimé à 57-80 Ma selon les études [Kapitonov and Jurka, 1996, Zietkiewicz et al., 1998, Price et al., 2004], ce qui est en accord avec leur présence chez les strepsirrhiniens (tels que les galagos et les lémuriens) dont la divergence avec les autres primates est estimée à 55-62 Ma [Yoder et al., 1996].

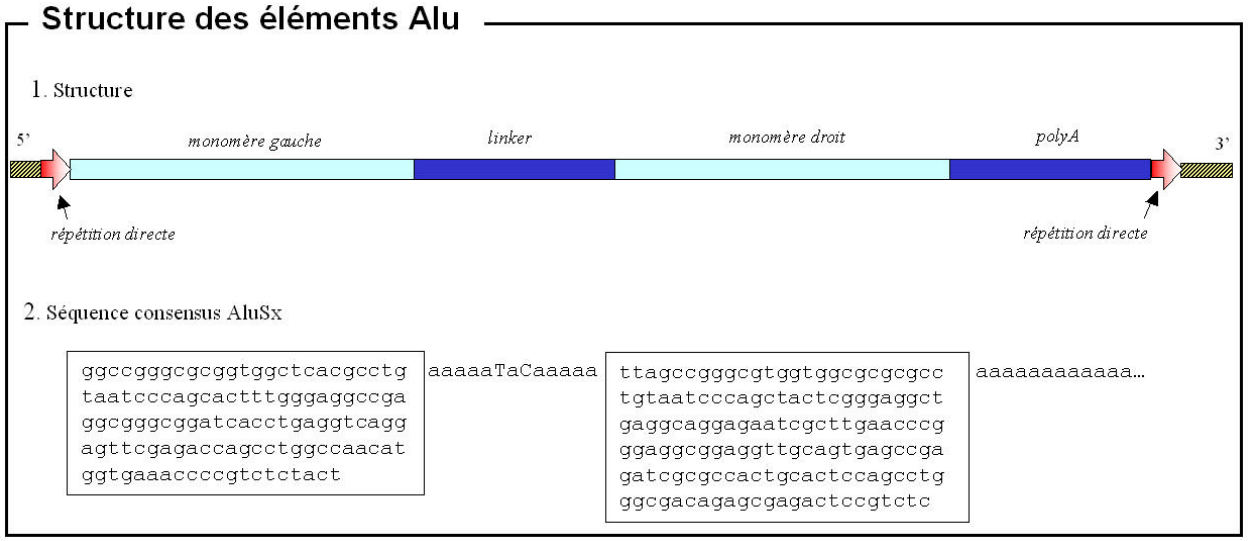


FIG. 1.8 – **1-** Structure d'une séquence Alu. L'élément Alu est composé de deux monomères semblables riches en GC d'environ 130 nt chacun. Ils sont reliés par le *l*inker, une séquence courte riche en A. La séquence se termine par une répétition polyA de taille variable, et est flanquée de deux répétitions directes de 5 à 15 nt. **2-** Séquence consensus le la famille AluSx, utilisée comme consensus général aux éléments Alu. Les deux monomères sont encadrés et leur séquence se lit de bas en haut.

1.3.2 Mécanismes de la rétrotransposition

Les éléments Alu sont répertoriés dans la classe des SINEs, qui sont des rétrotransposons de petite taille ne contenant pas de zone codante. Cela signifie qu'ils sont incapables de produire les enzymes nécessaires à leur rétrotransposition, et sont obligés de parasiter celles d'autres éléments transposables, ici les L1 [Boeke, 1997], qui sont des LINEs présents dans le génome des mammifères. La partie 5' des séquences Alu contient un site promoteur de l'ARN-polymérase III, qui permet à la séquence d'être transcrite en ARN. Ils ne possèdent, par contre, pas de site de fin de transcription, cette dernière s'arrête donc dans la région flanquante en 3', lorsqu'elle rencontre une suite de quatre thymines successives [Batzer and Deininger, 2002]. Les ARNs issus de la transcription doivent ensuite

s'apparier à la machinerie rétrotranscriptionnelle des L1. Beaucoup de questions concernant cette phase ne sont pas encore résolues. En effet, cette étape nécessite la présence d'un ARN de L1 en phase de traduction. Il faut ensuite que l'ARN du Alu soit suffisamment proche des rétrotranscriptases produites par le L1 pour pouvoir se les approprier, et les utiliser pour initier sa propre rétrotranscription.

La synchronisation des transcriptions Alu/L1 n'a pas ou très peu été étudiée, mais a peu de chances d'être simplement due au hasard. Une hypothèse a été proposée par Boeke [Boeke, 1997] concernant le rapprochement des ARNs et la récupération de la machinerie des L1 par les éléments Alu. Ils proposent que la queue polyA de Alu puisse s'apparier à la protéine chapeau ajoutée en 5' du L1 lors de sa transcription, grâce à des protéines nommées PABP (PolyA Binding Proteins). Ainsi, l'élément Alu sera automatiquement proche du L1 lorsque celui-ci sera traduit par le ribosome. Un certain nombre de PABP seraient nécessaires à la connexion avec le L1, et le nombre de PABP qui peuvent s'apparier au polyA dépend de la taille de ce dernier. Seule une queue polyA de taille importante (40 nt au minimum) semble permettre la rétrotransposition des séquences Alu [Roy-Engel et al., 2002, Dewannieux and Heidmann, 2005].

Les rétrotranscriptases produites par les L1 jouent probablement aussi le rôle d'endonucléases [Jurka, 1997], et vont permettre à l'ARN Alu de sectionner la séquence receveuse à des positions spécifiques. Un mécanisme proposé est que la protéine clive l'ADN à un site de restriction de type 5'-TTAAAA, ou 5'-TTTTAA sur le brin complémentaire. La queue polyA peut alors s'apparier à cette séquence, qui sert ensuite de promoteur pour sa rétrotranscription proprement dite (à partir du 3' ; voir figure 1.6). Le mécanisme de ré-appariement final du 5' de Alu avec la séquence génomique est encore inconnu mais l'on sait qu'il produit des répétitions directes de 5-15 nt aux extrémités de l'insertion (Figure 1.8.1).

Au-delà de ces mécanismes connus ou soupçonnés, d'autres facteurs ont certainement une importance dans la rétrotransposition des éléments Alu. Tout d'abord, ces derniers ont gardé au fil du temps une séquence génomique qui se replie selon la même structure secondaire [Zietkiewicz et al., 1998], suggérant que ce repliement joue un rôle dans l'une des étapes de la rétrotransposition. Deuxièmement, le linker semble exceptionnellement conservé au cours du temps [Roy et al., 2000]. Il est donc probable qu'il ait aussi un rôle à jouer dans la mobilité de l'élément, probablement en lien avec la structure secondaire.

Un certain nombre d'erreurs peuvent toutefois intervenir durant le processus de transposition,

les copies de Alu produites étant de fait incapables de rétrotransposition. Il se peut tout d'abord que la transcription initiale soit décalée et tronque le début de la séquence, détruisant par là les sites promoteurs. Des erreurs ponctuelles lors de la synthèse de l'ARN ou lors de la rétrotranscription pourraient avoir le même effet. Il se peut aussi que la rétrotranscription soit interrompue avant son terme, créant là encore des séquences Alu tronquées.

1.3.3 Les différentes familles

Après qu'un élément Alu a été inséré, il accumule des mutations ponctuelles comme n'importe quelle séquence génomique. Ces mutations étant aléatoires, chaque copie de Alu est censée être unique dans un génome. Toutefois, l'analyse des séquences a montré qu'il existe des mutations partagées par un certain nombre d'éléments Alu. Ces mutations, dites diagnostiques, sont en fait héritées de la séquence Alu d'origine, ce qui signifie que toutes les copies partageant ces mutations proviennent d'un Alu source commun, plus récent que l'ancêtre commun à tous les éléments Alu. Grâce à la caractérisation de ces mutations diagnostiques, il a été possible de répartir les séquences Alu en trois familles principales, les AluJ, AluS, et AluY, qui sont elles-mêmes divisées en sous-familles (tableau 1.1).

Les AluJ sont les plus anciens, et sont les Alu dimériques directement issus de la jonction entre les FLA et FRA. Ils sont décomposés en deux sous-familles, les AluJo et les AluJb, qui sont apparues à peu près en même temps, il y a environ 57 Ma [Zietkiewicz et al., 1998]. Ensuite viennent les AluS (28-48 Ma), eux-mêmes répartis en 6 sous-familles : Sz, Sg, Sq, Sp, Sc et Sx, selon la nomenclature proposée en 1996 [Batzer et al., 1996]. Ces différentes sous-familles ne sont pas également représentées dans le génome, les AluSx étant nettement plus nombreux. C'est d'ailleurs la séquence consensus de AluSx qui est utilisée comme consensus de toutes général des séquences Alu (Figure 1.8.2). Enfin, la famille des AluY regroupe tous les éléments jeunes, les plus anciens étant apparus il y a environ 20 Ma [Kapitonov and Jurka, 1996, Price et al., 2004]. Le nombre de sous-familles AluY est assez important, mais chacune comporte assez peu de copies, par rapport aux AluS et AluJ. Les sous-familles de AluY les plus étudiées sont indiquées dans le tableau 1.1, avec leurs âges approximatifs.

L'immense majorité des familles et sous-familles énumérées ici ne sont présentes qu'à l'état de fossiles dans le génome des primates actuels, c'est-à-dire que plus aucune n'a gardé la capacité de rétrotransposer. Les seuls éléments Alu encore actifs dans le génome humain sont des AluY, particulièrement jeunes, mais très peu nombreux. Certains, comme les AluYa5 et AluYb8 ont été largement

Familles Alu

Famille	Sous-famille	Age	Référence
FAM		**112**	*[1]*
J	Jb	**57 - 81**	*[1], [2], [7]*
	Jo	**57 - 81**	*[1], [2], [7]*
S(Sz)		**35 - 48**	*[1], [2], [7]*
	Sx	35 - 44	*[1], [2], [6], [7]*
	Sq	35 - 44	*[1], [2], [6]*
	Sg	31 - 41	*[1], [2], [6]*
	Sc	30 - 41	*[1], [2], [6]*
	Sp	28 - 41	*[1], [2], [6]*
Y		**19 - 24**	*[1], [7]*
	Yd1	14.96	*[6]*
	Ye5	11.53	*[6]*
	Yc3(Yd3)	12.61	*[6]*
spécifiques humains			
	Yi6	4.24	*[6]*
	Yg6	2.68 - 5.7	*[6], [9]*
	Yd6(Yd8)	2.45 - 5.3	*[6], [9]*
	Yb7	2 - 4.81	*[6], [8], [9]*
	Yb8	**2.33 - 5.4**	*[1], [3], [4], [6], [8], [9]*
	Ya5	**2.33 - 4**	*[1], [3], [4], [6], [7]*
	Ya8	1.07 - 4	*[3], [6], [7], [9]*
	Yb9	1.27	*[5]*
	Yc1	1.95 - 5.2	*[6], [9]*
	Ya5a2	0.5 - 1.1	*[3], [9]*

[1] Kapitonov & Jurka 1995, **[2]** Zietkiewicz *et al.* 1998, **[3]** Roy *et al.* 2000, **[4]** Roy-Engel *et al.* 2001, **[5]** Carroll *et al.* 2004, **[6]** Xing *et al.* 2004, **[7]** Price *et al.* 2004, **[8]** Carter *et al.* 2004, **[9]** Hedjes *et al.* 2005

TAB. 1.1 – Age des principales familles (en gras) et sous-familles Alu proposées dans la littérature. Les AluYA5 et AluYb8, particulièrement actifs, sont aussi écrits en gras. Les écarts représentent la fenêtre maximum des âges donnés dans les références citées.

étudiés [Batzer et al., 1995, Carroll et al., 2001], d'une part car ils sont relativement nombreux (2000 ou 3000 copies de chaque dans le génome humain) et d'autre part car leur insertion est parfois

la cause de maladies génétiques. Les autres éléments actifs principaux sont les AluYa5a2, AluYa8, AluYb8, AluYb9, AluYc1 et AluYc2, mais ils sont beaucoup plus rares.

Les éléments Alu présents dans le génome, qu'ils soient fossiles ou encore actifs, n'ont pas été insérés de manière régulière dans le temps, mais sont la conséquence de plusieurs vagues d'amplifications. Comme nous l'avons précisé dans le paragraphe précédent, les éléments Alu utilisent le système de rétrotransposition d'autres éléments transposables, les L1. L'activité des Alu dépend donc fortement de l'activité de ces derniers. De plus, un grand nombre de copies Alu sont tronquées ou dégradées dès leur insertion, et sont structurellement incapables de rétrotransposition. Les conditions pour qu'un nouvel élément Alu puisse se rétrotransposer sont donc très difficiles à obtenir.

Les contraintes sont même tellement fortes qu'en réalité très peu d'éléments Alu ont réussi à avoir une activité transpositionnelle importante et stable dans le temps. Ainsi, il est possible que tous les membres d'une même sous-famille soient des copies du même locus source [Deininger et al., 1992]. Cette théorie porte le nom de théorie du *gène maître* (« Master gene theory » en anglais). Une autre théorie propose que les copies nouvellement insérées soient aussi capables de rétrotransposition, et puissent donc participer à l'expansion de la famille, tant qu'elles n'accumulent pas trop de mutations [Brookfield, 1993]. Il semblerait que la réalité soit plus nuancée que le simple modèle du gène maître, puisque l'analyse détaillée des AluYa5 montre la possibilité d'avoir plusieurs copies comme source de rétrotransposition [Roy et al., 2000].

Pour résumer, le modèle le plus récent propose une expansion par « germination » [Cordaux et al., 2004], avec un locus Alu très actif (maître) capable de produire de nombreuses copies de lui-même, qui elles-mêmes peuvent parfois aider à l'expansion de la famille (germes) ou créer des sous-familles. Les AluJo et Jb originaux auraient donc été des Alu maîtres avec une forte activité, étant donné le nombre important de ces reliques dans le génome humain. L'un des deux aurait ensuite donné naissance au premier AluS qui lui-même aurait produit une série de germes (les sous-familles de AluS). Pour des raisons encore inconnues, ces germes ont eu une ou plusieurs phases de très forte activité rétrotranscriptionnelle, créant ainsi la majeure partie des séquences Alu visibles dans le génome des primates. Enfin, l'un de ces AluS se serait transformé en AluY, possédant une activité plus réduite et produisant de temps en temps de nouveaux germes [Cordaux et al., 2004, Price et al., 2004].

Chapitre 2

Les microsatellites

Ce chapitre est consacré à la description des microsatellites et de leur dynamique. Il n'a pas pour prétention d'être exhaustif sur les connaissances que nous possédons sur ces éléments génomiques, mais a pour but de décrire les propriétés utiles à la compréhension de leur apparition. La première étape est de donner une définition des microsatellites. Comme nous le verrons, il n'y a pour l'instant pas de définition arrêtée, mais plutôt une série de caractères communément admis. Quelques-unes des méthodes expérimentales utilisées pour étudier les microsatellites seront décrites, en mettant l'accent sur celles qui nous seront utiles. Nous présenterons ensuite la dynamique de mutation des microsatellites, qui leur donne leur caractère hyper-variable si intéressant. Enfin, la dernière section abordera le concept du cycle de vie des microsatellites, qui mènera à la problématique de cette thèse : comment apparaissent les microsatellites dans les génomes.

2.1 Description générale

2.1.1 Les répétitions en tandem

Les microsatellites appartiennent à un ensemble d'éléments génomiques appelés répétitions en tandem. Les répétitions en tandem sont constituées de répétitions adjacentes plus ou moins nombreuses d'un monomère (motif) donné. Certains de ces éléments ont la propriété d'être variables en nombre de répétitions, une conséquence de phénomènes de mutation particuliers (glissement de polymérase et recombinaison inégale, voir chapitre précédent). Ces répétitions en tandem peuvent être séparées en trois classes distinctes : les satellites, les minisatellites et les microsatellites.

Les satellites se définissent comme des répétitions en tandem possédant un très grand nombre de répétitions, pouvant régulièrement atteindre plusieurs mégabases [Charlesworth et al., 1994]. Ils

ont en général des périodes assez longues (une centaine de paires de base), souvent constituées de motifs répétés plus courts. Peu de travaux ont été effectués pour comprendre la dynamique des satellites, mais l'on sait toutefois que leur variabilité en taille est principalement causée par des erreurs de recombinaison (crossing-over inégaux ou ré-appariements décalés) [Stephan, 1986] (voir section 1.2.3). Leur apparition est assez mal documentée, mais pourrait être la conséquence d'un mécanisme appelé « rolling circle ». En résumé, ce mécanisme implique l'excision d'un fragment d'ADN durant la réplication, qui se referme sur lui-même et est répliqué en boucle. Les répétitions obtenues sont ensuite réinsérées dans le génome, et deviennent un satellite [Walsh, 1987, Rossi et al., 1990].

Les minisatellites sont des répétitions en tandem d'un motif compris entre dix et une centaine de paires de bases. Les répétitions sont basées sur un même motif consensus, mais peuvent contenir des variants. Ces variants sont des modifications du motif, causées par des mutations ponctuelles ou des réarrangements, mais qui sont répétées à plusieurs endroits dans l'élément, en fonction de l'histoire évolutive du minisatellite. La variabilité en taille des minisatellites repose, là encore, sur les crossing-over inégaux. Il a aussi été proposé que les minisatellites puissent évoluer par glissement, soit lors de la réplication, soit lors de recombinaisons [Richard and Paques, 2000]. Cette hypothèse semble être soutenue par d'autres travaux, qui proposent une évolution différente selon la composition du minisatellite. Ainsi les minisatellites riches en G/C évolueraient plus par recombinaison, et ceux riches en A/T plutôt par glissement [Buard and Jeffreys, 1997].

Enfin, les microsatellites sont l'équivalent des minisatellites, mais pour une période comprise entre 1 et 6 nucléotides. Plusieurs autres différences sont toutefois à relever. Tout d'abord, la définition classique des microsatellites ne propose pas la notion de variant. En effet, la période étant très courte, elle ne supporte que très peu d'imperfections. Des cas de variants sont pourtant à relever, mais sont considérés comme microsatellites complexes, comme nous le verrons dans la section suivante. En deuxième lieu, le facteur principal de variabilité des microsatellites ne semble pas être le crossing-over inégal, mais le glissement de polymérase (comme expliqué dans la section 1.2.5). La différence entre microsatellites et minisatellites est donc arbitraire, et il est probable que les modes d'évolution soient finalement assez semblables, lorsque les périodes sont similaires. C'est d'ailleurs pour cette raison que les éléments répétés de période entre 6 et 10 ne sont pour l'instant intégrés à aucune des deux classes.

2.1.2 Définitions et structure moléculaire

Il n'existe pas de définition formelle des microsatellites à l'heure actuelle. La seule caractéristique commune à tous les microsatellites est celle, déjà énoncée, d'une séquence répétée en tandem de période 1 à 6 nucléotides. Pourtant, beaucoup d'autres paramètres permettent de qualifier et classer les microsatellites. Des caractéristiques comme le nombre de répétitions, le motif en lui-même, la complexité de la séquence, sont encore sujets à débats malgré de réguliers efforts de consensus [Tautz, 1993, Chambers and MacAvoy, 2000, Ellegren, 2004, Buschiazzo and Gemmell, 2006]. Pour bien comprendre la suite du document, et particulièrement les mécanismes d'apparition des microsatellites, il est nécessaire de connaître leurs propriétés structurelles. Nous allons donc les détailler dans cette section.

Motif

Un microsatellite est défini tout d'abord par sa période (la taille du motif répété). La période des microsatellites est généralement comprise entre 1 et 6 nucléotides. Chaque période représente une classe de microsatellites, nommées respectivement mono, di, tri, tétra, penta et hexanucléotides. La période se doit aussi d'être la plus petite possible, c'est-à-dire que le motif donné ne peut être la répétition d'un motif plus court (par exemple, ATAT est un AT répété deux fois). On dit alors que le motif est indivisible.

Certaines études ne considèrent pas les mononucléotides comme des microsatellites. En règle générale, leur dynamique évolutive semble toutefois concorder avec celles des autres classes [Lai and Sun, 2003, Dieringer and Schlotterer, 2003], cette distinction n'a donc pas lieu d'être. La période maximum de 6 nucléotides est encore matière à débats, mais l'on sait que les séquences répétées de périodes plus importantes (les minisatellites) évoluent plutôt par erreurs de recombinaison (*cf.* section 2.1.1). Il y a donc une fenêtre de période entre 6 et 10 nucléotides, où l'on ne sait pas si c'est le glissement de polymérase, les erreurs de recombinaison, ou les deux qui sont majoritairement à l'origine de la variabilité des séquences. Ces classes de motifs sont toutefois intégrées à certaines analyses de microsatellites [Yeramian and Buc, 1999, Desmarais et al., 2006].

La grande majorité des études sur les microsatellites sont réalisées avec des séquences répétées de type AC/GT, pouvant laisser croire que seuls ces motifs correspondent à des microsatellites. En réalité, cette prédominance des AC est la conséquence de leur nombre important dans les génomes, du moins chez les animaux [Dokholyan et al., 2000] et de leur propension à être longs et polymorphes. Ils sont donc de fait devenus des marqueurs de choix en biologie des populations, et la matière première

des études de dynamique évolutive (souvent amorcées par des biologistes des populations). D'autres motifs sont néanmoins utilisés, tels que AT, AG/CT, CAG/CTG, et quelques tétranucléotides comme les GATA et AAAG. De plus, l'utilisation s'est élargie à tous les motifs possibles depuis que de larges fractions de génomes, voire des génomes entiers, sont disponibles dans les banques de séquences.

Taille

Un second paramètre important pour un microsatellite est sa taille, en nombre de répétitions. On décrit généralement un microsatellite sous la forme $(X)_n$, avec X le motif, et n le nombre de répétitions, même si cette notation pose plusieurs problèmes. Tout d'abord, la variabilité des microsatellites est due au glissement, qui, par définition, n'implique que des changements de taille multiples de la période (voir section 1.2.5). Il n'y a cependant aucune raison que la séquence possède un nombre entier de répétitions. Par exemple, la séquence ggATCATCATCATgg ne peut être considérée comme $(ATC)_4$, mais n'est pas non plus réellement un $(ATC)_3$. L'utilisation d'un nombre de répétitions non entier devient alors nécessaire, comme ici un $(ATC)_{3,67}$. De plus, la dénomination des microsatellites par leur nombre de répétitions peut amener une certaine confusion, lorsque l'on considère les différentes classes de motifs. Il est bien évident qu'un mononucléotide et un hexanucléotide possédant tous deux dix répétitions ne sont pas soumis aux mêmes contraintes physiques. L'hexanucléotide étant six fois plus long (en terme de nucléotides), il a par exemple beaucoup plus de chances de subir des mutations.

Un autre problème est la question de la taille minimum. Si l'on veut être formel, on peut considérer un microsatellite comme tout élément constitué d'au moins une répétition en tandem d'un motif donné. Dans les faits, une taille limite bien supérieure est généralement utilisée, soit en nombre de répétitions [Kruglyak et al., 2000], soit en paires de bases [Richard and Dujon, 1997, Toth et al., 2000], soit les deux [Jurka and Pethiyagoda, 1995]. La justification de ces limites est statistique. En effet, la définition formelle considère par exemple que tous les doublons de types AA, CC, GG ou TT sont des microsatellites, malgré la très forte probabilité de les rencontrer aléatoirement dans les génomes. Il a donc été proposé de ne considérer les microsatellites que pour des tailles où leur densité est supérieure à celle attendue dans un génome dénué de dynamique de glissement [Delgrange and Rivals, 2004, Kolpakov et al., 2003, Rose and Falush, 1998]. L'apparition des microsatellites dans un tel génome n'est censée se produire que par mutation ponctuelle aléatoire, et tout écart à cet attendu dans un génome réel suppose qu'un glissement s'est produit. Cette taille minimum de glissement introduit une propriété non plus structurelle, mais mécanique à la définition des microsatellites. La taille minimum généralement admise est de huit paires de bases, comme proposé par Rose & Falush

[Rose and Falush, 1998] suite à des analyses de distribution dans le génome de la levure (voir section 5.1.1).

La question de la taille minimum est un point central de ma thèse, car des répétitions de taille inférieure à cette limite semblent quand même être capables de glissement [Noor et al., 2001, Primmer and Ellegren, 1998]. Ce thème sera abordé plus en détail dans le chapitre 5.

Proto-microsatellites et quasi-microsatellites

Les proto-microsatellites sont des séquences répétées possédant un très petit nombre de répétitions, trop peu pour pouvoir être variables. Ils apparaissent par hasard, à la suite de mutations ponctuelles, comme proposé dans le modèle de Jarne *et al.* [Jarne et al., 1998]. Le concept de proto-microsatellite n'est valable que si l'on considère qu'une séquence répétée a besoin d'atteindre une taille minimum (en paire de bases ou en répétitions) pour devenir un microsatellite.

Les quasi-microsatellites sont des séquences non répétées, mais qui peuvent le devenir, via quelques mutations ponctuelles. Par exemple, la séquence aaACCTACTTgc est une séquence quasi-microsatellite car une substitution C→T ou T→C peut la transformer en $(ACCT)_2$ ou $(ACTT)_2$, respectivement. Une séquence telle que ttACCACCAGCta n'est pas considérée comme un quasi-microsatellite même si la transition G→C donne un $(ACC)_3$, car le proto-microsatellite $(ACC)_2$ existe déjà. Par contre, la séquence ttACCAGCACCta en est un. Le nombre de mutations n'est pas une limite exacte car il dépend de la taille du motif et de la position des mutations.

La distinction entre proto- et quasi-microsatellite peut être ambiguë, comme par exemple pour une séquence de type ttAAGAAcc. Dans ce cas là, faut-il considérer les deux AA comme des proto-microsatellites distincts, ou préférer considérer l'ensemble comme un quasi-microsatellite ? La solution est de la considérer comme étant les deux. Ce genre de cas se retrouve fréquemment dans les régions de faible complexité ou « cryptic simplicity » [Tautz et al., 1986]. Ce sont des régions de taille variable, constituées de répétitions d'un faible nombre de motifs différents, pas nécessairement adjacentes. La définition des régions de faible complexité repose là encore sur un critère statistique de sur-représentation par rapport à un attendu dans un génome aléatoire, au même titre que la question de la taille minimum des microsatellites. Elles sont relativement communes dans les génomes eucaryotes et contiennent de nombreux proto- et quasi-microsatellites.

Imperfections

Par définition, les microsatellites sont des séquences répétées en tandem, mais il arrive que ces répétitions ne soient pas parfaites. En effet, ces séquences sont soumises aux mêmes contraintes moléculaires que le reste du génome, et peuvent notamment subir des mutations ponctuelles qui briseront les répétitions. Un microsatellite ayant subi des mutations ponctuelles est qualifié d'imparfait (tableau 2.1), et les zones où la répétition est brisée sont nommées interruptions. La question des interruptions a été très peu étudiée, les seuls travaux disponibles ne concernant que des séquences très faiblement imparfaites [Taylor et al., 1999, Rolfsmeier et al., 2000, Harr et al., 2000]. Le problème de la détermination d'un taux maximum d'imperfection, au-delà duquel la séquence ne peut plus être considérée comme un microsatellite, mais comme une simple région de faible complexité, reste par exemple totalement non résolu.

D'autre part, les interruptions soulèvent la question de la cohérence du microsatellite. Par exemple, deux $(AC)_{20}$ séparés par 3 bases non répétées peuvent être considérés comme deux microsatellites parfaits distincts ou comme un seul imparfait. La plupart des analyses tolèrent des interruptions de quelques bases, mais d'autres, souvent théoriques [Bell and Jurka, 1997, Lai and Sun, 2003], préfèrent considérer deux microsatellites distincts dès qu'une interruption vient rompre la répétition. Si l'on souhaite se contenter d'étudier les microsatellites parfaits, il faut ne pas considérer les sous-parties parfaites des microsatellites imparfaits, en s'assurant que la séquence étudiée n'est pas voisine à quelques bases d'une autre séquence répétée de même motif.

Complexité

Les notions de motif et d'interruption vues précédemment ne s'appliquent qu'à des microsatellites dits simples, constitués de la répétition d'un motif unique. On peut définir deux autres types de microsatellites : composés et complexes [Chambers and MacAvoy, 2000]. Les microsatellites composés sont définis comme la concaténation de deux microsatellites de motifs distincts. Les deux sous-parties peuvent être directement adjacentes ou séparées de quelques bases non répétées, qui seront alors considérées comme une interruption (Table 2.1). Les motifs peuvent être totalement différents, tant en taille qu'en composition, mais sont en règle générale assez similaires. Il n'est par exemple pas rare d'observer des microsatellites composés $(GA)_n(GATA)_m$. Les microsatellites complexes sont une généralisation des microsatellites composés, avec plus de deux motifs distincts. Là encore, des interruptions sont possibles entre les motifs, et un motif peut se trouver à plusieurs positions différentes dans le locus (tableau 2.1). Ce genre de microsatellites est utilisé dans des études de biologie des populations, mais leur dynamique évolutive n'est encore que rarement abordée. On

suppose toutefois qu'ils proviennent de microsatellites simples ayant dégénéré [Buschiazzo and Gemmell, 2006].

La nécessité de recourir à plusieurs définitions selon la complexité du microsatellite montre la limitation de la dénomination (X)n basée sur un motif. La séquence de chaque microsatellite est en effet issue de son histoire évolutive, qui implique certes des expansions et contractions, mais aussi des mutations ponctuelles qui peuvent aboutir à la formation de ces séquences complexes. Ce problème de motif consensus pose d'ailleurs quelques difficultés pour la détection des microsatellites via des algorithmes informatiques, comme nous l'exposerons dans le chapitre 3.

Types de microsatellites

parfait	$(ACT)_n$
imparfait (subst)	$(ACT)_n CCT (ACT)_m$
imparfait (suppr)	$(ACT)_n CT (ACT)_m$
composé	$(ACT)_n (GA)_m$
complexe	$(ACT)_n (GA)_m AA (ACC)_k$

TAB. 2.1 – Les différentes classes de microsatellites, catégorisées selon leur complexité.

2.1.3 Distributions dans les génomes

Les distributions de microsatellites peuvent représenter deux informations différentes selon que l'on étudie un locus déterminé ou l'ensemble des locus d'un génome. Pour un locus, la distribution représente généralement le nombre d'allèles, un allèle étant défini par sa taille (en nombre de répétitions). Pour un génome séquencé, il s'agit de la distribution en taille de l'ensemble des locus du génome. Les distributions présentées dans cette section ne feront référence qu'à des analyses de génomes séquencés, on parlera donc toujours de distributions en nombre de locus. Une autre notion utile est la densité, à savoir le nombre de locus présents dans une séquence de taille donnée. Les densités peuvent être déterminées à un niveau global (toute la séquence) ou bien local (fragment par fragment), mais sont généralement normalisées par rapport à une taille fixe.

La présence de répétitions en tandem et de microsatellites semble être une caractéristique partagée par tous les organismes vivants, avec des densités plus ou moins fortes selon les règnes [Ellegren,

2004, Coenye and Vandamme, 2005, Trivedi, 2006]. L'analyse des distributions des microsatellites a été considérablement facilitée par les efforts de séquençage de génomes complets. A l'heure actuelle, 187 génomes eucaryotes (dont 77 animaux, 71 champignons, 11 plantes et 28 protistes) et 567 génomes bactériens sont disponibles sur le site du NCBI (http ://www.ncbi.nlm.nih.gov/Genomes/index.html). Tous ne sont pas totalement assemblés (*i.e.* les régions enchainées dans l'ordre), et les couvertures ne sont pas forcément optimales (*i.e.* le nombre de séquences pour un même fragment, qui influe sur la qualité de la séquence finale), mais les séquences des grands organismes modèles sont maintenant de bonne qualité. Nous allons ici donner les distributions et densités des microsatellites pour quelques uns de ces organismes, permettant d'évaluer l'étendue des différences entre organismes (Figure 2.1).

Les procaryotes sont réputés pour posséder très peu de microsatellites par rapport aux eucaryotes. La figure 2.1 montre en effet que *Escherichia coli* possède une densité moindre de microsatellites, mais surtout qu'ils sont de taille nettement plus réduite que chez les eucaryotes. Cela avait déjà été montré par Coenye & Vandamme [Coenye and Vandamme, 2005] pour les mononucléotides, et est une caractéristique commune à tous les procaryotes. La limitation est en général expliquée par la proportion importante de régions codantes dans les génomes procaryotes, qui contraignent le développement des microsatellites.

Il y a aussi de grosses différences de densité entre les eucaryotes unicellulaires et les eucaryotes supérieurs, comme le montre la comparaison de la levure *Saccharomyces cerevisiae* avec les autres organismes (Figure 2.1 ; [Dokholyan et al., 2000, Toth et al., 2000, Katti et al., 2001]). Les microsatellites de la levure sont plutôt courts, ce qui peut être là aussi expliqué par la forte proportion de régions codantes dans le génome (environ 70%). Une proportion inhabituelle est d'ailleurs occupée par les trinucléotides, et il a été montré qu'on les trouvait plus qu'attendu dans les régions fonctionnelles. Cette prédominance des trinucléotides dans les régions codantes de la levure soutient l'hypothèse d'une fonctionnalisation possible de certains microsatellites, l'hypermutabilité permettant de garder un haut niveau d'adaptabilité [Richard and Dujon, 1997, Young et al., 2000, Malpertuy et al., 2003], comme observé chez certaines bactéries [Moxon et al., 1994].

Parmi les eucaryotes supérieurs, les différences sont nettement moins marquées mais existent tout de même. La drosophile possède par exemple une densité de microsatellites supérieure à celle des autres génomes considérés (riz, homme et chimpanzé), et leur taille moyenne est plus importante. Ces distributions vont d'ailleurs à l'encontre d'un certain nombre d'études précédentes [Harr

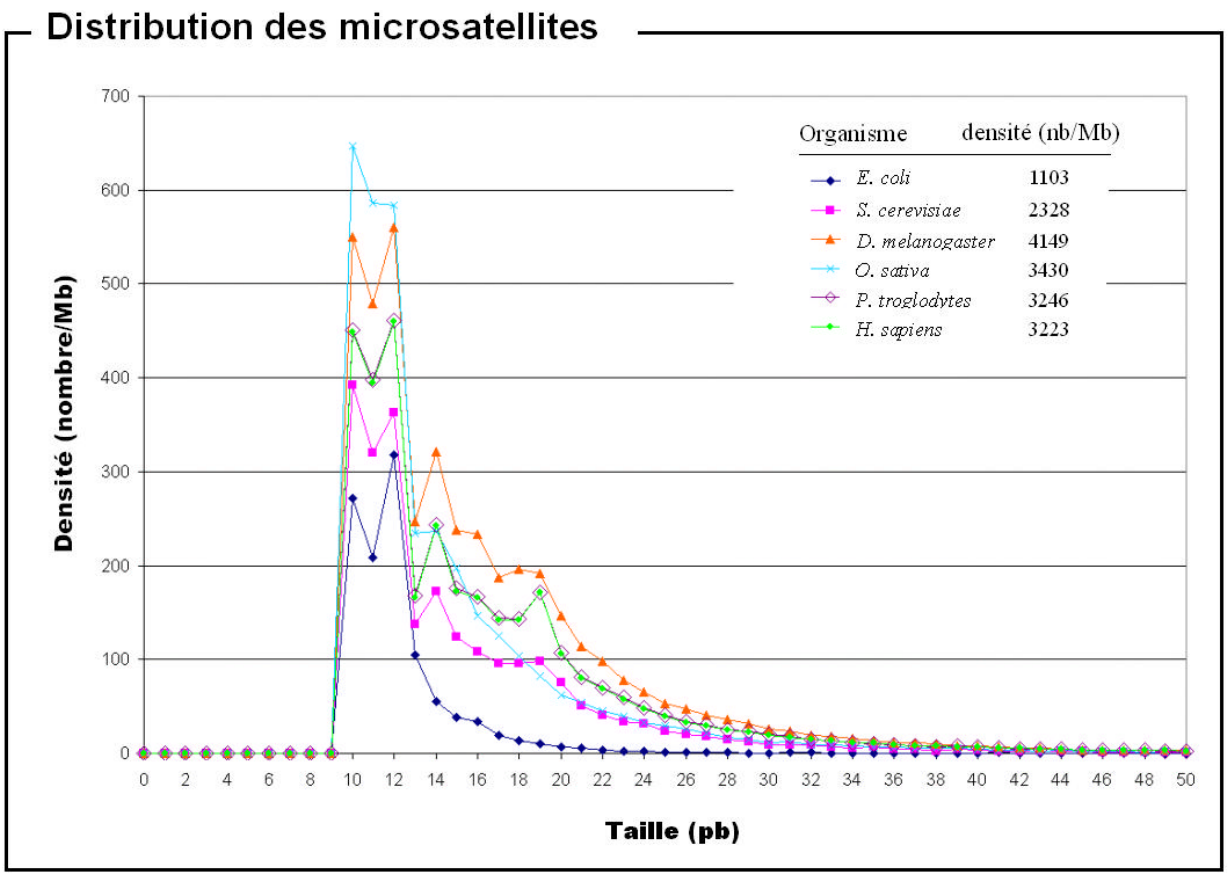


FIG. 2.1 – Distribution de la densité des microsatellites (période 1 à 6) en fonction de leur taille, pour six organismes. EC : génome complet, SC : génome complet, DM : chromosome 2L, OS : chromosome 1, PT : chromosome 1 et HS : chromosome 1. Détections réalisées avec l'algorithme TRF {+2,-5,-7 ; 20}. TRF, avec ces paramètres, ne permet pas la détection des microsatellites de taille inférieure à 10 nt.

and Schlotterer, 2000, Schug et al., 1998b, Katti et al., 2001], qui indiquaient une densité moins importante de longs microsatellites chez la mouche que chez les autres organismes. Cela dépend en fait de la méthode de détection et des microsatellites observés : la figure 2.1 montre tous les microsatellites, toutes classes confondues, alors que certaines études se contentaient des parfaits, ou uniquement des dinucléotides, etc. Cela peut aussi provenir du choix du chromosome puisque la figure 2.1 ne donne les distributions que pour un chromosome par organisme, alors qu'il existe des différences entre chromosomes [Katti et al., 2001]. On peut aussi observer que la densité pour le riz est plus élevée pour les petites tailles, ce qui est vrai de manière générale pour les plantes [Katti et al., 2001, Dieringer and Schlotterer, 2003].

Il est aussi intéressant de noter que ce ne sont pas les mêmes motifs qui sont les plus représentés dans les différents génomes. Comme nous l'avons déjà vu, les trinucléotides sont par exemple en sur-

nombre chez la levure. Chez les primates, la majorité des microsatellites sont des mono-nucléotides A/T. Les autres microsatellites prépondérants sont les AC/GT [Katti et al., 2001]. Chez la drosophile, les AC/GT sont aussi nettement sur-représentés, ainsi que les CAG/CTG, alors que la plante *Arabidopsis thaliana* contient une majorité de AG et de AAG/CTT [Katti et al., 2001].

2.2 Les différentes méthodes d'analyse

Nous avons détaillé dans le chapitre précédent les caractéristiques moléculaires des microsatellites, et les différentes définitions que l'on pouvait en donner. Qu'ils soient mono ou dinucléotidiques, parfaits ou imparfaits, courts ou longs, l'intérêt principal qui est porté à ces séquences est lié à leur forte variabilité en taille dans une population. La communauté scientifique a donc très tôt cherché à comprendre les mécanismes affectant cette variabilité, et notamment le taux de mutation. La section 2.3 sera consacrée à la présentation de ces mécanismes, mais nous allons tout d'abord présenter les principales méthodes utilisées pour les étudier. Chacune des méthodes présente des contraintes et des avantages en termes de coût financier, temporel, et technique. Toutes ne permettent pas non plus d'obtenir les mêmes types de données, ou de réaliser les mêmes analyses. Les principales méthodes sont les analyses de mutation directe, de variabilité, phylogénétique, et de génome séquencés. Seules les deux dernières méthodes peuvent servir à étudier l'apparition des microsatellites ; nous nous y attarderons donc plus longuement.

2.2.1 Les analyses de mutation directe et de variabilité

Les premiers travaux portant sur la compréhension des mécanismes de mutation des microsatellites ont été des observations directes de mutations. Levinson & Gutmann [Levinson and Gutman, 1987a] ont commencé en utilisant des techniques de manipulation génétique. L'insertion d'un microsatellite dans un gène de bactériophage de levure leur a permis de détecter des mutations par glissement en fonction de l'expression (ou non) du gène dans les cellules en culture. La technique a été utilisée dans nombre d'autres études [Strand et al., 1993, Wierdl et al., 1997], et permet de déterminer quels paramètres influent sur les taux de mutation (longueur de la séquence ou type du motif, interruptions ou non, etc.). De plus, il est possible de jouer avec certaines contraintes métaboliques des cellules, en désactivant par exemple les gènes de réparation de l'ADN ou ceux impliqués dans la recombinaison. Schlötterer et Tautz (1992) ont aussi proposé une analyse *in vitro* de la dynamique des microsatellites, qui avait pour avantage d'étudier la mécanique pure du glissement, sans se soucier des interférences avec l'environnement cellulaire.

Une autre possibilité pour observer directement des mutations est de créer des lignées d'accumulation de mutations [Schug et al., 1997]. Le principe est de faire se reproduire de multiple générations à partir d'individus possédant le même allèle microsatellite pour un certain nombre de locus, et de compter le nombre de mutations présentes dans les dernières générations. Les mutations s'étant accumulées au fil des générations, il est possible d'évaluer un taux de mutation par génération. Cette méthode demande un grand nombre de générations pour obtenir suffisamment d'événements de mutations, et ne peut donc être réalisée qu'avec des organismes à cycle de vie rapide, et dont la reproduction peut être contrôlée.

Les mutations des microsatellites d'organismes ayant un cycle de vie plus long peuvent quant à elles être observées par des analyses de pedigree. Elles reposent sur l'analyse de la transmission des allèles microsatellites dans différentes familles [Weber and Wong, 1993, Primmer et al., 1996, Xu et al., 2000]. Pour chaque famille, on compare les génotypes des parents à ceux de leur progéniture, et on compte les allèles qui ont muté chez ces derniers. Ces analyses ont été largement utilisées pour explorer la dynamique évolutive des microsatellites humains, mais elles requièrent d'avoir à disposition des liens de parenté sûrs pour un grand nombre de familles, et nécessite un contrôle rigoureux des transmissions (problème des allèles nuls, de l'origine parentale inconnue,...).

La dynamique des microsatellites peut aussi être analysée par des méthodes moins directes, telles que les analyses de variabilité. Ces méthodes sont basées sur l'analyse de la distribution en taille des allèles d'un locus donné, dans différentes populations d'une même espèce, ou dans plusieurs espèces [Rubinsztein et al., 1995, Primmer and Ellegren, 1998, Harr et al., 1998]. Ces distributions nous renseignent sur le nombre d'allèles présents dans chaque population, leur fréquence, la taille moyenne et la variance, ou encore l'hétérozygotie de la population (valeur déterminant le taux d'individus hétérozygotes dans la population). Il n'est par contre pas possible de calculer les taux de mutation directement, mais ces derniers peuvent être inférés en ajustant des modèles de mutation aux distributions (voir section 2.4.1).

2.2.2 Les analyses phylogénétiques

Les analyses phylogénétiques suivent une logique différente de celle des méthodes précédentes, et ont pour vocation d'étudier l'histoire évolutive des locus microsatellites. Elles sont réalisées via le séquençage de locus chez plusieurs individus d'espèces ou de populations dont les liens phylogénétiques sont connus. Elles peuvent être réalisées à la suite d'analyses de variabilité. Certains allèles sont choisis dans les populations étudiées, sont séquencés, et sont alignés en fonction des liens

phylogénétiques des populations. Les séquences permettent d'obtenir de nouvelles informations qui n'étaient pas détectables sur la base des distributions alléliques uniquement, comme par exemple l'effet des interruptions sur la variabilité [Richard and Dujon, 1996, Jin et al., 1996]. Les études phylogénétiques ont par ailleurs permis de dévoiler les problèmes d'homoplasie (allèles identiques en longueur, mais issus d'un ancêtre différent) et de saturation (perte du signal phylogénétique à cause de l'homoplasie) [Angers and Bernatchez, 1997, Dettman and Taylor, 2004], que nous ne détaillerons pas dans cette thèse.

Les études phylogénétiques peuvent aussi être conduites sur la base d'une ou de peu de séquences par espèce. L'étude de la variabilité ou des taux de mutation des microsatellites n'est pas possible avec ces analyses, mais elles permettent de détecter leurs apparitions ou disparitions. Si l'arbre phylogénétique utilisé est assez large (avec des espèces suffisamment éloignées phylogénétiquement), il est possible que des locus qui existent dans une espèce soient apparus dans l'une des branches uniquement, et seront donc absents chez les autres espèces (Figure 2.2). Certaines études ont ainsi permis de mettre en évidence l'apparition d'un tétranucléotide par mutation ponctuelle chez l'homme [Messier et al., 1996], ainsi que des apparitions par glissement chez les drosophiles [Noor et al., 2001]. Il est par contre à noter que ces observations ne sont qu'anecdotiques, et qu'aucune étude d'envergure n'a été réalisée pour évaluer les modes d'apparition des microsatellites sur un grand nombre de locus. La disparition des microsatellites est observable de la même manière, lorsqu'un locus est présent dans une branche complète et absent dans une sous-branche [Taylor et al., 1999] (Figure 2.2).

2.2.3 Les analyses de séquences

Toutes les méthodes précédentes, exceptées celles de manipulations génétiques, reposent sur l'utilisation de marqueurs microsatellites connus, dont la plupart ont été choisis parce qu'ils étaient polymorphes et faciles à génotyper. Or, les microsatellites polymorphes sont *a priori* ceux qui possèdent un fort taux de mutation. Il y a donc un biais inhérent à ces techniques. La mise à disposition de séquences génomiques a permis de remédier à ce problème de non-représentativité. Les séquences génomiques, qu'elles soient simplement des fragments d'ADN, ou plus récemment des séquences complètes de génomes, contiennent un grand nombre de locus microsatellites. Le principe est donc de rechercher les séquences répétées dans ces fragments, grâce à des algorithmes informatiques, puis de les analyser en tant que locus distincts [Bell and Jurka, 1997, Yeramian and Buc, 1999, Young et al., 2000, Leclercq et al., 2007].

Les données issues de cette méthode sont différentes de celles obtenues par les méthodes précé-

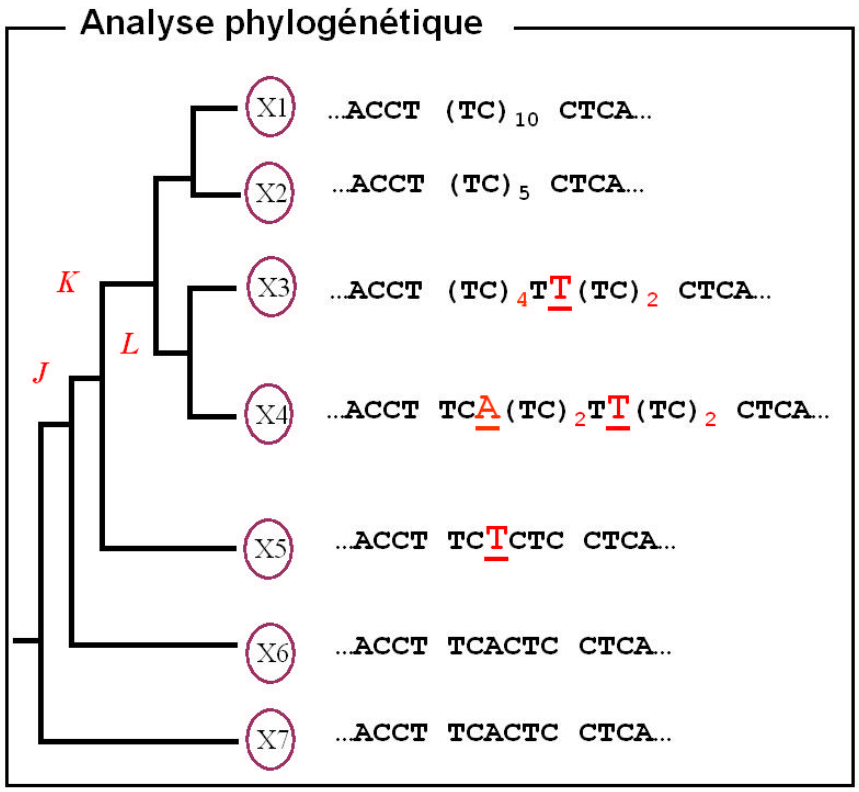


FIG. 2.2 – Méthode d'analyse phylogénétique. Les X_n représentent les individus de différentes espèces apparentées. Les bases soulignées représentent les interruptions. Une mutation ponctuelle fait apparaître un $(TC)_3$ dans la lignée J à partir de la séquence ancestrale, visible dans les individus X_6 et X_7. Le $(TC)_3$ s'est ensuite développé puis est devenu variable dans la lignée K. Enfin, dans la lignée L, il a subi des interruptions et des contractions qui ont amené à sa disparition dans l'individu X_4.

dentes. En effet, les séquences analysées ne représentent qu'un seul allèle pour chaque locus, et à un seul moment. Il est donc impossible d'en déduire directement un taux de mutation. Il permet en revanche, lorsque la proportion de séquences disponibles est suffisante, d'évaluer de manière non biaisée le nombre de locus dans les génomes, leur densité, et d'obtenir des distributions par motif ou par taille. Ces distributions peuvent alors être comparées entre organismes, comme il a été présenté dans la section 2.1.3 de ce chapitre. Une autre utilisation de ces distributions a été de permettre l'ajustement de modèles de dynamique d'évolution des microsatellites, et donc d'inférer des taux de mutation (voir section 2.4.1).

L'un des problèmes de ce genre d'étude concerne l'identification des locus, appelée extraction. Elle se fait via des programmes informatiques, qui peuvent avoir des logiques et des implémentations différentes. Nous verrons dans le chapitre 3 que ces différences engendrent de réelles divergences, les distributions obtenues pouvant être radicalement différentes selon l'algorithme de détection utilisé.

De plus, il n'est pas possible d'obtenir d'information sur l'apparition des locus. Des moyens détournés ont toutefois été mis en place, en introduisant des facteurs d'apparition dans les modèles de dynamique des microsatellites [Bell and Jurka, 1997, Jarne et al., 1998].

2.3 Les processus de mutation

Deux grandes théories ont été proposées à partir des années 80 pour expliquer la variabilité en taille des microsatellites : le glissement de polymérase ou SSM (Slipped-Strand Mispairing), et la recombinaison inégale ou UCO (Unequal Crossing Over). Chaque théorie a ses défenseurs et ses détracteurs, mais il fait peu de doutes à l'heure actuelle que le SSM est le moteur principal de la dynamique des microsatellites. Quel que soit le mécanisme réel, la variabilité est directement liée au taux de mutation de ces éléments, c'est-à-dire à la probabilité de subir une expansion ou une contraction entre deux générations. Ce taux de mutation est dépendant de plusieurs facteurs, qui sont détaillés ci-après. Cette section ne concerne pas directement l'apparition des microsatellites car la variabilité ne concerne que les locus déjà existants, mais elle est indispensable pour comprendre le modèle du cycle de vie présenté en fin de chapitre.

2.3.1 La théorie du glissement de polymérase (SSM)

Le mécanisme le plus communément admis comme facteur de la variabilité des microsatellites est le glissement de polymérase, tel qu'il est décrit dans le chapitre sur les types de mutation (voir Figure 1.7). Lors de la réplication de l'ADN, le brin synthétisé peut se dissocier du brin matrice. Si cela se produit lors de la synthèse d'un microsatellite, le ré-appariement peut être décalé à cause du caractère répété de la séquence. Ce décalage, s'il n'est pas détecté par les enzymes de réparation de l'ADN, provoque alors la perte ou le gain d'une ou plusieurs répétitions sur le nouveau brin par rapport à la matrice, selon que le décalage s'est produit en amont ou en aval de la position de désappariement.

Une seconde théorie propose que la variabilité des microsatellites est issue d'événements de crossing-over inégaux, comme c'est le cas pour les minisatellites (voir section 2.1.1). La recombinaison est la conséquence des mécanismes de réparation des cassures de l'ADN, qui utilisent la séquence homologue sur la chromatide sœur ou le chromosome homologue comme matrice (Figure 1.3). Si la cassure a lieu dans une séquence microsatellite, l'appariement sur la séquence homologue peut être décalé, et aboutir à l'ajout ou la soustraction de certaines répétitions lors de la réparation.

Plusieurs arguments plaident en défaveur de la théorie de mutations par recombinaison. Tout

d'abord, la stabilité des microsatellites n'est pas affectée lorsque l'on dégrade des gènes impliqués dans la recombinaison, ce autant pour *E. coli* que pour *S. cerevisiae* [Levinson and Gutman, 1987a, Henderson and Petes, 1992, Wierdl et al., 1997]. Les microsatellites deviennent par contre fortement instables lorsque des gènes impliqués dans la machinerie de réparation de l'ADN sont inactivés [Levinson and Gutman, 1987a, Strand et al., 1993, Wierdl et al., 1997, Rolfsmeier et al., 2000] (voir Sia [Sia et al., 1997] pour une revue). De plus, des études montrent que la variabilité des microsatellites n'augmente pas avec le taux de recombinaison local [Schug et al., 1998a, Payseur and Nachman, 2000, Huang et al., 2002]. Elle n'augmente pas non plus lors des phases de division méiotique alors que les événements de recombinaison y sont 100 à 1000 fois plus nombreux que lors des autres phases cellulaires [Strand et al., 1993].

Enfin, les distributions des tailles observées sont plus en accord avec les modèles de dynamique des microsatellites qui supposent des variations impliquant peu de répétitions [Shriver et al., 1993, Dirienzo et al., 1994]. Or, l'expansion/contraction de peu de motifs correspond mieux à la théorie du SSM, où le brin dissocié reste proche de la position de dissociation, qu'à l'hypothèse de crossing-over inégal, pour laquelle l'appariement erroné peut se faire sur n'importe quelle répétition. Cette preuve n'est pas formelle, car il arrive que le glissement puisse affecter un nombre quelconque de répétitions et la recombinaison uniquement un faible nombre, mais les contraintes moléculaires de ces deux mécanismes provoquent en général le phénomène inverse.

Les recombinaisons homologues ne semblent donc pas être le facteur primordial de la variabilité des microsatellites, mais elles y participent néanmoins [Ellegren, 2000b].

2.3.2 Taux de mutation corrélé à la longueur du microsatellite

L'un des facteurs les plus importants de la dynamique des microsatellites est la longueur de la séquence répétée. Il a été montré que le taux de mutation est corrélé positivement à la longueur du microsatellite, tant de manière expérimentale [Levinson and Gutman, 1987a, Wierdl et al., 1997, Primmer and Ellegren, 1998, Ellegren, 2000a, Webster et al., 2002] que par modélisation [Kruglyak et al., 1998, Sibly et al., 2001, Whittaker et al., 2003]. D'autres types d'études ont aussi montré que la variabilité d'un locus microsatellite au sein d'une espèce était corrélée à sa longueur [Dettman and Taylor, 2004, Primmer and Ellegren, 1998]. La variabilité étant directement liée au taux de mutation, ces résultats confortent l'idée de l'importance de la longueur sur le taux de mutation. Cette hypothèse peut s'expliquer assez intuitivement car plus un microsatellite est long, plus le nombre de positions possibles pour un mésappariement est grand, donc plus le taux de mutation pour un locus

est important.

Il est important de noter ici que le taux de mutation des microsatellites est calculé pour un locus complet, alors qu'il est généralement calculé pour un site (une base azotée) pour les mutations ponctuelles (voir section 1.2.6). Or, les locus microsatellites étudiés sont en général des dinucléotides avec une taille de 10 à 30 répétitions. Il conviendrait donc de diviser de 20 à 60 les taux de mutations obtenus, si l'on voulait les comparer au taux de mutations ponctuelles. Mais même comme cela, la variabilité des microsatellites reste nettement supérieure à celle d'une séquence quelconque, si l'on compare directement les taux de mutation. Par exemple Xu *et al.* [Xu et al., 2000] ont calculé un taux de mutation moyen de $1{,}8\text{x}10^{-3}$ mutation par locus, à partir de tétranucléotides humains (de taille de dépassant pas 50 répétitions), soit 2,5 ordres de grandeur plus élevé que le taux moyen de mutation ponctuelle pour le même organisme ($4{,}5\text{x}10^{-8}$ [Nachman and Crowell, 2000]).

2.3.3 Pas multiples

Dès les premières études sur les taux de mutations au sein des microsatellites, on s'est aperçu que les glissements pouvaient impliquer plus d'une répétition par évènement [Levinson and Gutman, 1987a, Wierdl et al., 1997, Ellegren, 2000a, Xu et al., 2000]. Ces cas sont généralement rares par rapport aux glissements d'une seule répétition (de l'ordre de un pour dix), n'impliquent souvent que de 2 à 5 répétitions, et ne semblent *a priori* pas causés par des erreurs de recombinaison ou recombinaison non-homologue [Huang et al., 2002]. Ces glissements à pas multiples sont pour une grande part des contractions, suggérant que les boucles causées par le mésappariement ont tendance à être plus grandes sur le brin matrice, ou moins bien réparées. Malgré la rareté des pas multiples, leur influence sur la dynamique globale des microsatellites n'est pas à négliger [Estoup et al., 1995, Direnzo et al., 1994, Whittaker et al., 2003].

Il est toutefois certains cas où les glissements à pas multiple sont biaisés vers les expansions, pour les locus microsatellites impliqués dans les maladies à triplet chez l'homme. Ces maladies, parmi lesquelles figurent la maladie de Huntington et le syndrome du X fragile [Mitas, 1997], ont comme caractéristique commune d'être causées par l'expansion incontrôlée de microsatellites trinucléotides au sein de l'individu, qui aboutissent à la non-expression de certains gènes. Ces trinucléotides sont pour la plupart de motif CAG ou CTG, présents soit en 5', soit dans le gène, et sont généralement parfaits. La taille moyenne de ces locus est de l'ordre de quinze répétitions, mais quelques allèles, dits à risque, peuvent comporter une centaine de répétitions. Ces allèles à risque ne semblent pas avoir

d'influence sur l'expression du gène voisin, sauf lorsqu'ils subissent des cycles d'expansion de grande ampleur. Ce phénomène se produit parfois, les allèles pouvant alors atteindre quelques milliers de répétitions. Le mécanisme d'expansion de grande ampleur n'est à ce jour pas très bien compris, mais il est possible qu'il soit causé par une déficience du système de réplication, qui créerait des boucles de manière répétée sur le brin en synthèse, lors d'une même réplication. Le motif même semble être en cause, car les boucles CAG/CTG peuvent former des structures secondaires stables, mal reconnues par les enzymes de réparation de l'ADN.

Ces cas n'ont été documentés que pour l'homme et pour certains locus très précis (ceux qui causent des dérèglements) mais il est possible que ce mécanisme d'expansion soit valable pour tous les locus capables de former des boucles stables, quelle que soit l'espèce. Cela reste toutefois anecdotique et n'a certainement que peu d'influence sur la dynamique générale des microsatellites.

2.3.4 Importance du motif

Dans les années 1990, la plupart des études portant sur la dynamique des microsatellites ont été réalisées sur des dinucléotides, généralement de type $(CA/TG)_n$, qui représentaient la majorité des marqueurs (voir section précédente). Très peu de comparaisons ont donc été réalisées entre différents motifs. Le séquençage généralisé d'ADN et de génomes complets a permis de dépasser les contraintes liées aux marqueurs et de pouvoir faire des études en fonction des motifs. De manière générale, plus la taille du motif est grande, moins le nombre de répétitions nécessaires pour être variable est important [Ellegren, 2000a]. En revanche le taux de mutation est de 1,5 à 2 fois plus fort pour les di que pour les tétranucléotides, les trinucléotides ayant un taux intermédiaire [Chakraborty et al., 1997]. De plus, le nombre de penta et d'hexanucléotides longs reste extrêmement rare, ce qui suppose que la taille du motif contraint l'expansion.

La composition du motif joue aussi un rôle dans la dynamique des microsatellites. Par exemple, les motifs TA possèdent un plus fort taux de mutation que les autres dinucléotides chez *S. cerevisiae* [Kruglyak et al., 2000]. Chez l'homme, les dinucléotides $(AC)_n$ semblent plus courts dans les régions de faible taux de GC [Calabrese and Durrett, 2003]. Comme nous l'avons précisé pour les maladies à triplets, certains motifs peuvent aussi créer des structures secondaires stables, mal reconnues par le système de réparation de l'ADN. Ces microsatellites particuliers auront donc tendance à muter plus facilement [Mitas, 1997].

2.3.5 Interruptions stabilisantes

Une autre caractéristique des microsatellites qui joue sur leur variabilité est leur taux d'imperfection. Il est maintenant démontré que les imperfections réduisent assez fortement le taux de mutation. Autrement dit, pour une taille équivalente, un microsatellite imparfait sera moins variable que son homologue parfait [Jin et al., 1996, Richard and Dujon, 1996]. L'une des explications proposées est que lorsque la séquence répétée est interrompue, même localement, elle admet moins de possibilités de réappariements erronés. Une autre possibilité est que les interruptions rendent les boucles causées par le décalage moins stables, donc plus facilement détectables par le système de réparation de l'ADN [Rolfsmeier and Lahue, 2000]. Cette hypothèse est surtout appropriée pour les motifs qui forment de larges boucles stables, c'est-à-dire principalement des locus causant des maladies à triplet.

La position de l'interruption semble aussi avoir son importance. Les taux de mutation sont nettement réduits lorsque les interruptions se trouvent éloignées du centre du microsatellite [Rolfsmeier and Lahue, 2000], positions qu'elles ont plutôt tendances à avoir [Brohede and Ellegren, 1999]. L'hypothèse proposée est que les glissements se font plutôt en fin de réplication du microsatellites, et que des erreurs seraient introduites lors de la correction des boucles par la machinerie de réparation.

2.3.6 Biais de mutation

Le modèle SSM simple suppose que les mutations sont symétriques, c'est-à-dire que la probabilité de subir une expansion est égale à celle de subir une contraction pour un microsatellite. Or, les résultats expérimentaux indiquent que la situation est plus complexe. Par exemple, un certain nombre d'auteurs montrent un biais des expansions sur les contractions, de manière globale ou uniquement sur les microsatellites les plus courts [Weber and Wong, 1993, Primmer and Ellegren, 1998, Xu et al., 2000]. D'autres montrent un biais vers les contractions pour les microsatellites les plus longs [Wierdl et al., 1997, Ellegren, 2000a, Huang et al., 2002], ou des contractions impliquant plus de répétitions [Xu et al., 2000].

Ces observations ont abouti à l'élaboration d'une théorie, dite du biais de mutation, qui envisage une taille « optimale » pour les locus microsatellites, vers laquelle les mutations seraient biaisées. Il y aurait donc plus d'expansions pour les microsatellites plus courts que cette taille, et plus de contractions (ou plus importantes) pour les plus longs [Garza et al., 1995, Xu et al., 2000, Whittaker et al., 2003, Vowles and Amos, 2006].

2.4 Le cycle de vie des microsatellites

2.4.1 Modèles théoriques

Parallèlement aux deux modèles biologiques proposés pour expliquer la dynamique des microsatellites (crossing-over inégal et glissement de polymérase), deux modèles théoriques principaux ont été développés, le IAM/KAM et le SMM. Le IAM (Infinite Allele Model) a été proposé par Kimura et Crow [Kimura and Crow, 1964]. Le principe est que chaque mutation à un locus donné produit un nouvel allèle qui n'existe pas dans la population étudiée. Le KAM (K-Allele Model) [Kimura, 1968] est la version de l'IAM pour une dimension finie. L'allèle créé par une mutation n'est pas systématiquement un nouvel allèle, mais est tiré aléatoirement dans une distribution uniforme de K allèles possibles. Ce genre de modèle s'applique assez bien pour des mutations ponctuelles, chaque nouvelle mutation ayant peu de chances d'intervenir à la même position qu'une mutation précédente. Ces modèles peuvent être adaptés à la dynamique des microsatellites si l'on considère qu'une mutation à un locus microsatellite peut engendrer un nouvel allèle d'une taille quelconque [Estoup et al., 1995].

Le SMM (pour Stepwise Mutation Model) stipule au contraire que les mutations produisent des allèles qui sont relativement proches de l'allèle original, et donc que la variabilité entre allèles ne pourrait s'obtenir que pas à pas. Ce modèle a été introduit par Ohta et Kimura [Ohta and Kimura, 1973], pour expliquer la différence de migration sur un gel entre variants d'un gène (à cause de la différence de charge que la mutation impliquait). Rapporté aux microsatellites, le SMM permet de modéliser une dynamique basée sur de petites différences de taille [Kimura and Ohta, 1978]. Le SMM ne permet que des mutations symétriques (autant de chance de grandir que de raccourcir) et à pas simple (n'impliquant qu'une répétition).

L'IAM/KAM et le SSM ont été testés, et même comparés au début des années 90, notamment avec les études de Valdes *et al.* [Valdes et al., 1993] et Shriver *et al.* [Shriver et al., 1993]. Ils ont simulé l'évolution d'un locus microsatellite pendant un temps donné dans une population fictive à l'équilibre mutation-dérive, en fonction du modèle de mutation choisi (SMM ou IAM). Ils ont ensuite comparé la distribution allélique obtenue à celle de locus réels issus d'analyses de variabilité. Dans les deux cas, les simulations développées soutiennent l'hypothèse du SMM.

Un grand nombre de déclinaisons du SMM ont ensuite été testées pour prendre en compte les divers facteurs affectant la dynamique des microsatellites (*cf.* section précédente). La modélisation des glissements à pas multiples a été introduite avec le modèle TPM (Two-Phase Model) [Dirienzo et al.,

1994] : les mutations à pas simple se font avec une certaine probabilité p, et à pas multiple avec $1-p$, la taille du pas multiple étant tirée dans une loi géométrique. Ce modèle a ensuite été simplifié en GSM (General Stepwise Model, [Fu and Chakraborty, 1998]), dans lequel la taille du pas est simplement tirée dans une loi géométrique. La mutation biaisée a été introduite en premier lieu par Walsh [Walsh, 1987], puis Tachida et Iizuka [Tachida and Iizuka, 1992]. Garza *et al.* [Garza et al., 1995] ont par la suite modélisé le biais de contraction vers une taille focale, avec plus d'expansions pour les microsatellites de taille inférieure et plus de contractions pour ceux de taille supérieure. Enfin, le taux de mutation proportionnel à la taille du locus a aussi été proposé dans les modèles de Walsh [Walsh, 1987] et Tachida et Iizuka [Tachida and Iizuka, 1992], mais a été généralisé par le modèle de Kruglyak *et al.* [Kruglyak et al., 1998]. Ces diverses implémentations sont résumées dans la figure 2.3.

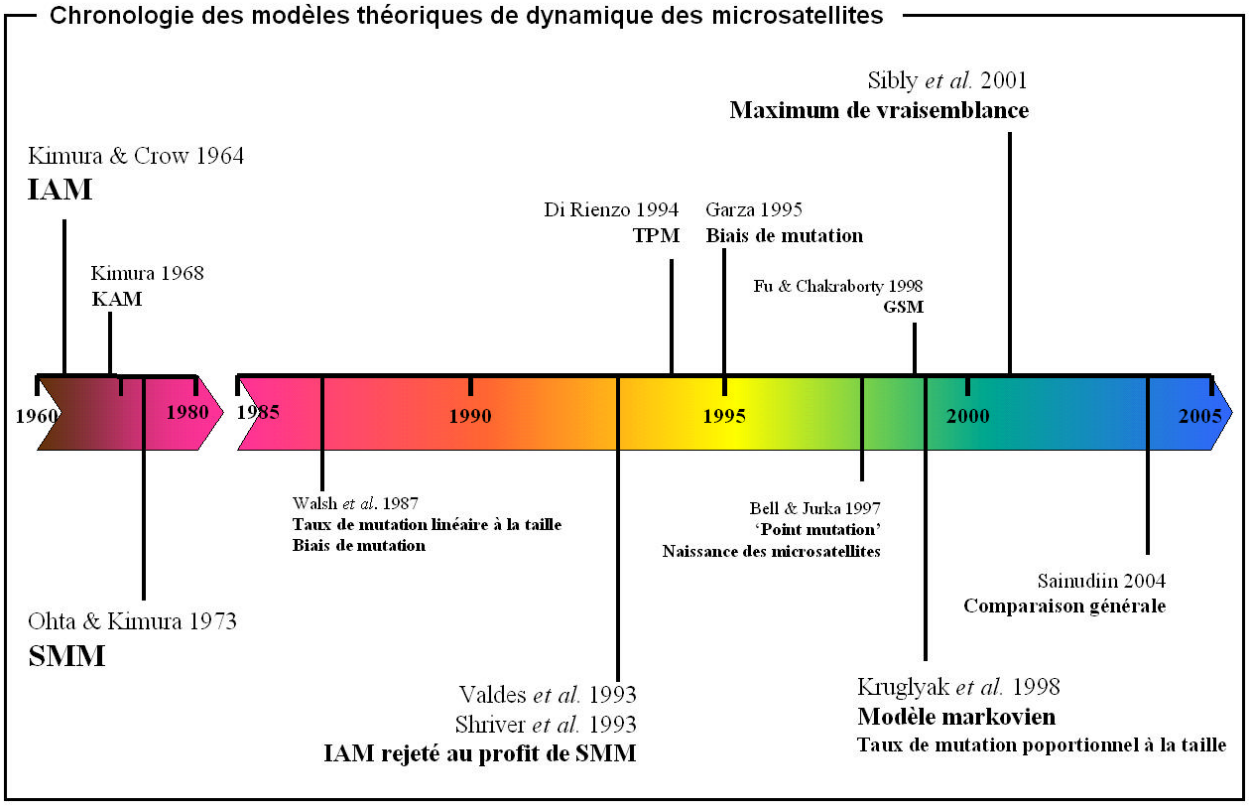


FIG. 2.3 – Chronologie des différents modèles théoriques de dynamique des microsatellites. Cette liste n'est pas exhaustive, mais indique les modèles principaux et ce qu'ils ont apporté .

Tous ces modèles ont été ajustés à des données provenant de deux sources principales : les analyses de variabilité (AV) et les analyses de séquence (AS). Or, selon le type des données, l'implémentation des modèles ne supporte pas les mêmes contraintes. Pour les modèles basés sur les AV, il s'agit de s'ajuster à une distribution d'allèles pour quelques locus déterminés. Ces modèles

proposent en général une implémentation semblable à celle utilisée par Valdes [Valdes et al., 1993] ou Shriver [Shriver et al., 1993], avec l'évolution d'un locus d'une taille donnée, dans une population fictive à l'équilibre mutation-dérive, durant un temps donné. Les modèles basés sur les AS doivent quant à eux s'ajuster à une distribution représentant un seul allèle pour une multitude de locus.

La plupart des modèles basés sur les AS ont été implémentés en utilisant un processus markovien de mutation, et sont quasiment tous inspirés du modèle de Kruglyak *et al.* [Kruglyak et al., 1998]. Le processus markovien simule l'évolution de la distribution en taille des locus en fonction de divers paramètres de mutation dans une population infinie. L'un des problèmes de ces modèles markoviens est que les mutations simulées peuvent aboutir à la disparition de tous les locus, s'ils subissent trop de contractions. Ils intègrent donc aussi un mécanisme d'apparition des microsatellites, comme proposé dans divers modèles théoriques [Walsh, 1987, Bell and Jurka, 1997, Jarne et al., 1998]. Cette remarque est importante car la plupart des modèles se contentent d'un taux d'apparition constant, qui représente le taux de mutation ponctuel. Or, les travaux présentés dans cette thèse montrent au contraire que la naissance des microsatellites répond à des phénomènes bien plus complexes que les simples mutations ponctuelles.

Nous allons conclure cette brève description des modèles de mutation des microsatellites par l'intérêt des tests par maximum de vraisemblance, introduits par Sibly *et al.* [Sibly et al., 2001]. Ces tests permettent de réaliser une comparaison des différents modèles en leur attribuant un score en fonction de leur ajustement aux données et de leur nombre de paramètres. De nombreux travaux ont comparé tous les types de modèles existants, avec plus ou moins de paramètres, sur différentes données [Calabrese and Durrett, 2003, Whittaker et al., 2003, Cornuet et al., 2006]. La comparaison de modèles la plus aboutie à ce jour est celle de Sainudiin *et al.* [Sainudiin et al., 2004] qui testent 13 modèles différents, incluant tous les types de paramètres qui ont été détaillés ici. Leurs conclusions indiquent que le modèle avec un taux de mutation proportionnel à la taille du locus, linéairement biaisé vers une taille focale, s'ajuste le mieux à leurs données. La gestion du pas multiple n'apporte en outre par de différence significative par rapport au pas simple.

2.4.2 Modèle biologique

A partir des connaissances accumulées sur les différentes propriétés des microsatellites, il a été possible de développer un modèle de leur cycle de vie. Un modèle synthétique a été proposé récemment par Buschiazzo et Gemmell [Buschiazzo and Gemmell, 2006], qui divise la vie d'un mi-

crosatellite en cinq phases majeures : la naissance, l'expansion, la dégénérescence, la contraction, et enfin la mort (figure 2.4). Ce modèle repose sur l'hypothèse originale d'un biais de mutation non pas basé sur la taille du microsatellite, mais sur son âge (même si les deux sont liés). Ainsi, les plus jeunes bénéficieraient d'un biais vers les expansions, les faisant grandir jusqu'à une taille focale, puis les contractions commenceraient à devenir plus importantes (en proportion ou en nombre de répétitions), réduisant leur taille.

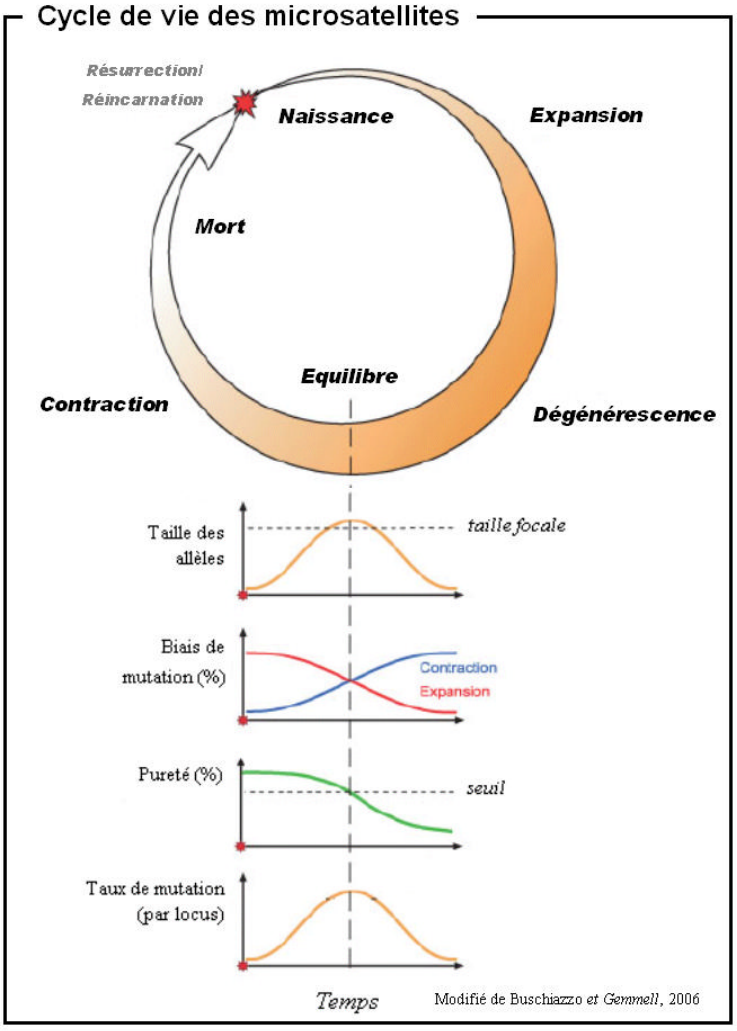


FIG. 2.4 – Modèle du cycle de vie des microsatellites.

La première phase, la naissance des microsatellites, est l'axe principal de cette thèse et sera donc détaillée plus longuement dans la section suivante, mais nous allons tout de même donner ici les notions de base. Les microsatellites apparaissent en général dans des séquences à faible complexité (voir section 2.1.2). Elles contiennent de nombreux quasi et proto-microsatellites, qui peuvent facilement donner naissance à de vrais microsatellites par l'intermédiaire de mutations ponctuelles.

Ces microsatellites sont qualifiés de *de novo*, car ils sont apparus directement à partir de la séquence génomique. A l'inverse, les microsatellites dits « adoptés » ne sont pas issus de la mutation de la séquence génomique, mais ont été insérés tels quels, généralement via les éléments transposables. L'importance relative de chacun de ces deux modes d'apparition n'est pas encore très bien connue et le travail présenté dans cette thèse a pour ambition d'y apporter quelques éclaircissements.

La naissance est alors suivie de l'expansion du jeune microsatellite. Comme nous l'avons précisé, le taux de mutation est corrélé à la taille du microsatellite, et est biaisé vers les expansions, surtout pour les petites tailles. Cela a pour conséquence une augmentation du nombre de répétitions du microsatellite, qui grandit d'autant plus vite qu'il vieillit. Cette croissance continue jusqu'à atteindre « l'âge d'or », où il est à sa taille maximum, parfait et très variable.

Mais le microsatellite finit par subir des mutations ponctuelles, créant des interruptions, et il entre dans la phase de dégénérescence. Les interruptions stabilisent la séquence (voir section 2.3.5) qui par conséquent grandit beaucoup moins vite, et accumule d'autres interruptions. De plus, il est possible que le processus de glissement puisse dupliquer les interruptions, produisant une accumulation d'autant plus rapide [Estoup et al., 1995, Rolfsmeier and Lahue, 2000]. Le passage dans la phase de dégénérescence via les interruptions peut par contre n'être que transitoire, car il a été montré que les glissements pouvaient aussi faire disparaître les interruptions [Harr et al., 2000]. L'arrivée des imperfections coïncide aussi avec le début du biais vers les contractions. Rien à l'heure actuelle ne permet de savoir si les interruptions jouent un rôle sur l'accentuation des contractions, mais il est clair que les microsatellites ayant atteint une certaine taille (et donc un certain âge) subissent des contractions, soit plus fréquentes, soit plus importantes (voir section précédente).

L'effet conjugué des interruptions et des contractions fait passer le microsatellite dans la dernière phase, celle de contraction. La quasi-totalité des mutations sont des contractions et des mutations ponctuelles, si bien que la répétition en tandem disparaît, signifiant la mort du microsatellite. Mais il est tout à fait possible qu'un nouveau microsatellite puisse renaître à partir de celui qui est décédé. En effet, la mort du microsatellite n'est autre que le passage d'une séquence répétée en tandem à une séquence de faible complexité durant la phase de contraction. Or, ces séquences sont justement un terreau favorable à l'apparition des microsatellites, comme nous l'avons indiqué plus haut. Ces réapparitions peuvent prendre deux formes : une « résurrection » lorsque le nouveau microsatellite est de même motif que l'ancien, ou une « réincarnation » lorsque le motif a changé.

L'une des questions principales concernant ce modèle est la durée du cycle. Il est bien évident que cette durée est propre à chaque locus, et dépend du motif, de la rapidité d'accumulation d'interruptions, de l'importance des contractions durant la phase de dégénérescence, etc. (*cf.* section 2.3). Toutes ces contraintes sont liées à l'environnement génomique du microsatellite (déamination des cytosines, zone de forte recombinaison pouvant entraîner des insertions, zones sous sélection) ou à l'espèce (enzymes de réparation plus ou moins performantes, contraintes de taille de génome). Il est aussi important de noter que la vision de ce modèle est centrée sur la vie d'un allèle unique. Or, la qualité intrinsèque d'un microsatellite, et qui fait son attrait principal, est justement d'être présent sous plusieurs allèles au sein des populations. Les limites entre les phases deviennent alors beaucoup plus floues, lorsque par exemple certains allèles d'un locus sont déjà engagés dans la phase de dégénérescence (à cause de mutations ponctuelles) alors que d'autres sont encore en phase d'expansion. La durée du cycle de vie est en fait contrôlée par la dérive génétique qui, par son action, finit par éliminer tous les allèles d'une phase ou d'une autre, et ainsi fait progresser ou régresser le locus dans son cycle.

La question de la durée du cycle est donc un problème ouvert, et qui ne pourra être résolu qu'en intégrant des interactions populationnelles au modèle moléculaire présenté ici. Toutefois, quelques études expérimentales laissent à penser que ce cycle pourrait être assez rapide. En effet, même entre espèces proches, il existe des différences critiques. Chez les hémiasciomycètes, la variation en densité peut aller du simple au triple, quel que soit le type de motif [Malpertuy et al., 2003] ; chez les drosophiles, les rares longs microsatellites de *D. melanogaster* ne sont apparus que très récemment [Harr and Schlotterer, 2000].

2.4.3 Apparition des microsatellites

Avec leur densité importante dans les génomes eucaryotes, et la proportion de nucléotides qu'ils peuvent représenter, les microsatellites sont des acteurs importants de l'évolution moléculaire au sens large. Nous avons jusqu'ici décrit les mécanismes contrôlant la dynamique des microsatellites, et expliqué comment ces mécanismes pouvaient s'assembler pour aboutir à un modèle de cycle de vie. Il est maintenant nécessaire de se pencher sur les conséquences du cycle de vie décrit sur la séquence génomique en son entier.

Concrètement, cela revient à se demander si la quantité de locus microsatellites croît, décroît ou est stable au fil du temps. Cette question peut être abordée de manière « démographique », en étudiant l'équilibre entre naissances et décès. Une autre possibilité, adoptée dans cette thèse, est de se concentrer sur la mécanique des phases de naissance et de contraction afin d'évaluer la capacité

des génomes à produire ou détruire des microsatellites. Notre recherche s'est focalisée uniquement sur l'apparition des microsatellites, tant le champ de recherche est large et les connaissances peu nombreuses, mais a cependant permis de clarifier et préciser l'action de certaines forces.

Deux origines majeures sont proposées pour les microsatellites, comme énoncé dans le modèle du cycle de vie : l'apparition par adoption et l'apparition *de novo*. De nombreux cas de microsatellites liés aux éléments transposables ont été relatés dans la littérature, pour toutes sortes d'organismes [Desmarais et al., 2006, Wilder and Hollocher, 2001], mais la relation la plus documentée est certainement le lien avec les séquences Alu chez les primates [Arcot et al., 1995, Jurka and Pethiyagoda, 1995, Nadir et al., 1996, Yandava et al., 1997]. Comme indiqué dans la section 1.3.1, la séquence des éléments Alu contient une queue polyA et un linker polyA. Chaque nouvel Alu importe donc avec lui deux microsatellites $(A)_n$. Ces polyA sont par ailleurs vecteurs d'apparition d'autres microsatellites tels que des GAA [Chauhan et al., 2002, Clark et al., 2004]. Néanmoins, aucune étude de grande ampleur n'a été réalisée pour évaluer l'importance que prennent les microsatellites issus de séquences Alu dans l'ensemble des microsatellites humains.

Il paraît indispensable, si l'on veut comprendre comment les microsatellites adoptés influent sur le cycle de vie des microsatellites, de savoir si la proportion de microsatellites issus des séquences Alu est négligeable ou non par rapport au total. Il est aussi primordial d'évaluer à quel point les polyA ont tendance à créer d'autres microsatellites, et de quantifier leur taux de disparition par rapport au taux d'expansion. Toutes ces interrogations seront abordées dans le chapitre 4 de ce document.

L'apparition *de novo* à partir de mutations ponctuelles dans des séquences de faible complexité est quant à elle considérée comme le facteur majeur d'apparition, même si cela n'a jamais été démontré. Certes, quelques études phylogénétiques ont fait état d'apparitions par mutations ponctuelles à partir de quasi et proto-microsatellites [Messier et al., 1996, Sokol and Williams, 2005], mais ces cas ne concernent que un ou deux locus, et ne peuvent être généralisés à l'ensemble des microsatellites. De plus, il a aussi été observé que des proto-microsatellites pouvaient subir des expansions [Primmer and Ellegren, 1998, Noor et al., 2001]. Enfin, des études récentes ont montré qu'une forte proportion des insertions de très petite taille (1-4 nt) sont en fait des duplications en tandem [Zhu et al., 2000, Messer and Arndt, 2007]. Si ces duplications s'avéraient fréquentes, elles pourraient être un facteur non négligeable d'apparition de proto-microsatellites, et même de microsatellites matures, à partir d'une séquence totalement quelconque.

Ces diverses observations laissent à penser que l'apparition par mutation ponctuelle simple peut être contestable, ou du moins à compléter par d'autres voies d'apparition. Il faut pour cela évaluer la part de chacun des mécanismes précités dans la naissance des nouveaux microsatellites. Le chapitre 5 de ma thèse s'attache donc à étudier les apparitions *de novo* des microsatellites, en évaluant l'importance relative de la mutation ponctuelle, du phénomène de duplication de petite taille, et d'un éventuel proto-glissement.

A la suite de ces deux chapitres, les différents vecteurs d'apparition des microsatellites auront été approfondis, et il s'agira d'en faire ressortir une vision globale. Le chapitre 6 mettra en évidence l'importance des diverses hypothèses les unes par rapport aux autres, leurs interactions, et leur influence sur le cycle de vie des microsatellites. Nous y discuterons aussi quelques perspectives pouvant faire suite au travail présenté, comme les implications sur les distributions des microsatellites, ou le profit que ces travaux peuvent apporter aux modèles théoriques de dynamique des microsatellites.

Chapitre 3

Les limites de la détection bio-informatique

3.1 Approche bio-informatique

3.1.1 Choix du type d'étude

Comme nous l'avons expliqué dans le chapitre 2, il existe différentes manières d'étudier les microsatellites. En ce qui concerne les apparitions, les travaux présentés sont généralement basés sur des études phylogénétiques. Ces études comportent toutefois de sévères limitations. Premièrement, elles reposent sur l'étude d'un locus existant dans une espèce, et rien ne permet *a priori* de savoir si l'on va pouvoir observer son apparition à partir des autres espèces choisies. Autrement dit, les résultats obtenus sont des observations faites par hasard et ne peuvent être systématisées. La seconde limitation, qui découle directement de la première, est le caractère anecdotique de ces résultats. En effet, les mécanismes d'apparition proposés n'étant inférés qu'à partir de quelques cas, il se peut qu'ils ne reflètent pas la généralité. Une analyse de grande envergure serait possible, en se basant sur un très grand nombre de marqueurs, tous séquencés pour un très grand nombre d'espèces, mais cela n'empêcherait pas d'avoir un biais lié aux locus et espèces choisis.

Une autre solution est de tirer parti des génomes séquencés. Il est en effet possible d'extraire tous les locus microsatellites contenus dans le génome d'un organisme donné, et ainsi obtenir la distribution exhaustive de toutes les classes de microsatellites (voir section 2.2.3). Ces distributions en elles-mêmes ne nous renseignent en rien sur l'apparition, puisque la détection ne donne justement que des locus existants. En revanche, elles permettent de produire des modèles de dynamique des microsatellites, dans lesquels il est possible d'introduire des mécanismes d'apparition [Bell and

Jurka, 1997, Kruglyak et al., 1998, Dieringer and Schlotterer, 2003]. A l'inverse des analyses phylogénétiques, les estimations obtenues par ces modèles ont un caractère général en termes de taux d'apparition, mais elles ne renseignent pas sur les mécanismes réels qui les ont produites.

Le chapitre 5 de cette thèse consistant à évaluer l'importance des divers mécanismes d'apparition sur le taux d'apparition global et sur les distributions de microsatellites observés, il était nécessaire d'avoir une vision générale, non biaisée, des mécanismes à l'étude. La méthode qui a donc été mise en place est une analyse phylogénétique sur génomes séquencés, aussi appelé analyse de génomique comparative. Elle permet dans notre cas d'observer l'apparition éventuelle de chacun des locus présents dans un organisme donné, par rapport à des organismes phylogénétiquement proches.

Pour le chapitre concernant l'apparition via les éléments transposables (chapitre 4), le recours à des méthodes de génomique comparative ne s'est pas révélé nécessaire, dans la mesure où nous nous sommes basés uniquement sur des relations de proximité entre microsatellites et séquences Alu. Cela signifie toutefois qu'il faut être en mesure de localiser exactement la position de ces deux types d'éléments dans la séquence d'ADN de l'organisme étudié, des positions qui ne peuvent être obtenues, là encore, que via l'analyse de génomes complètement séquencés et assemblés.

3.1.2 Problématique

Les génomes séquencés représentant quelques millions à quelques milliards de nucléotides, l'extraction des microsatellites doit être automatisée, via l'utilisation d'algorithmes informatiques. Etant donné le nombre important d'études des microsatellites basés sur les génomes séquencés, on pourrait penser qu'il suffit de réutiliser les méthodes de détection dont elles se sont servies.

Malheureusement, cela n'est pas aussi simple. Lorsque l'on compare les algorithmes utilisés dans la littérature, on s'aperçoit qu'ils ont en général été conçus et implémentés directement par l'équipe de recherche et que leurs critères de détection sont étonnamment variables. Ce problème concerne principalement la taille minimum de détection et les critères de perfection (tableau 3.1), et est la conséquence directe du manque de définition formelle d'un microsatellite (voir chapitre 2). De plus, certaines études ne détaillent pas précisément l'algorithme de détection utilisé (par exemple [Falush and Iwasa, 1999, Lai and Sun, 2003]), empêchant ainsi d'avoir une idée précise des microsatellites détectés. Cette disparité dans les méthodes peut aboutir à des incohérences flagrantes, comme celles présentées dans le tableau 3.2.

D'autre part, la grande majorité des travaux se sont contentés de détecter des microsatellites parfaits, et ce pour plusieurs raisons. La première est que les locus extraits sont souvent utilisés pour tester des modèles théoriques de dynamique des microsatellites. Or, les modèles développés actuellement sont incapables de gérer l'intégration d'imperfections. La seconde raison est que la détection de microsatellites imparfaits est loin d'être aussi triviale que celle des parfaits, et qu'elle demande des algorithmes plus sophistiqués.

Tous les problèmes évoqués ici sont les conséquences de la production « à la main » de ces algorithmes, sans qu'aucune normalisation n'ait été recherchée. Pourtant, des algorithmes dédiés à la détection des microsatellites ont déjà été publiés, principalement par des laboratoires de bioinformatique [Coward and Drablos, 1998, Benson, 1999, Landau et al., 2001, Castelo et al., 2002, Kolpakov et al., 2003, Delgrange and Rivals, 2004, Wexler et al., 2005]. Ces algorithmes possèdent de nombreux avantages. Tout d'abord, le fait qu'ils aient été publiés garantit leur efficacité en termes de complexité. Cela garantit de plus une transparence sur la méthode, car l'algorithme est censé être expliqué dans la publication. Une autre qualité importante est la possibilité de détection des microsatellites imparfaits, du moins pour certaines méthodes. Enfin, la majorité a fait l'objet d'une implémentation, et le programme résultant est généralement distribué librement.

Nous cherchions, pour développer les différents points de la thèse, à détecter de manière efficace, cohérente, et si possible rapide, les microsatellites de génomes complets. L'utilisation d'algorithmes de détection dédiés était donc appropriée à nos besoins. Compte tenu du nombre d'algorithmes disponibles, et surtout des différences de logique et d'implémentation de chacun, il était nécessaire de réaliser une étude comparative pour déterminer lequel répondait le mieux à nos besoins.

Les résultats présentés dans ce chapitre correspondent à la comparaison de cinq algorithmes dédiés à la détection de répétitions en tandem, ayant des logiques de détection différentes, certains étant très utilisés par la communauté scientifique, et tous étant disponibles sous forme de programmes. L'étude est décomposée en deux volets. Tout d'abord, la plupart de ces algorithmes étant paramétrables, nous avons cherché à tester l'influence de ces paramètres sur les détections. Nous avons ensuite comparé l'efficacité de ces algorithmes, et ce pour divers génomes, afin d'évaluer l'effet de la séquence sur la détection. L'ensemble des méthodes et résultats sont détaillés dans l'article « Detecting microsatellites within genomes, significant variation among algorithms », joint en Annexe 1.

auteurs		algorithme	motifs	taille minimum	perfection	remarques
Bachtrog *et al.*	1999	ns	di à tetra	5 répétitions	parfaits : séparés d'au moins 4 pb	
Bell & Jurka	1997	ns	di	2 répétitions	parfaits	*maximum de 40 répétitions*
Calabrese & Durrett	2003	ns	di et tri	5 répétitions	parfaits, séparés d'au moins 50 pb et pas d'occurrence du motif dans les 4 pb flanquants	
Dieringer & Schlotterer	2003	personnel	mono à tetra	2 répétitions	parfaits	
Dohkolyan *et al.*	2000	ns	mono et di	2 répétitions	parfaits	
Falush & Iwasa	1999	personnel	di	0	parfaits	
Field & Wills	1998	ns	mono à hexa	2 répétitions	?	
Hancock	2002	SIMPLE34	mono à tetra	?	parfaits	*paramètres de SIMPLE34 non précisés*
Jurka & Pethyagoda	1995	ns	mono à hexa	3 répétitons et 12 pb	parfaits	
Kantety *et al.*	2002	personnel	di à tetra	18 ou 20 pb	parfaits	
Katti *et al.*	2001	personnel	mono à tetra	20 ou 21pb	imparfaits : 1 erreur est permise tous les 10 pb	
Kayser, M. et al.	2006	TRF	di à penta	8 répétitions	imparfaits	*paramètres de TRF non précisés*
Kruglyak S. et al.	1998	personnel	di à tetra	5 répétitions	parfaits	
Kruglyak S. et al.	2000	personnel	di à tetra	2 répétitions	imparfaits : 5 répétitions parfaites au minimum avec une répétition du motif dans les 2 pb flanquants	
Lai & Sun	2003	ns	mono à hexa	0	parfaits : séparés d'au moins 2x la taille du motif	
Majewski & Ott	2000	TRF	di à tetra	25 puis 45 pb (à cause de TRF)	imparfaits	*paramètres de TRF : +2,-7,-7 ; 50 puis 90*
Nadir *et al.*	1996	ns	mono à hexa	16 pb	?	
Pupko and Graur	1999	ns	mono à penta	0	parfaits	
Richard and Dujon	1996	REPEAT	tri	12 pb (à cause de REPEAT)	imparfaits	*paramètres de REPEAT : +3,-6,-12 ; 36*
Richard and Dujon	1997	REPEAT	tri	12 puis 15 pb (à cause de REPEAT)	imparfaits	*paramètres de REPEAT : +3,-6,-12 ; 36 puis 45*
Rose and Falush	1998	personnel	mono, di et tetra	0	parfaits	
Sainudinn *et al.*	2004	ns	di	10 répétitions	imparfaits : une seule interruption, séparés d'au moins 50 pb	
Toth *et al.*	2000	ns	mono à hexa	12 pb	parfaits	
Vowles & Amos	2004	personnel	AC	2 répétitions	parfaits : séparés d'au moins 100 pb	
Vowles & Amos	2006	TRF	di à penta	8 répétitions ou 25 pb (à cause de TRF)	imparfaits	*paramètres de TRF : +2,-7,-7 ; 50*
Webster *et al.*	2002	SPUTNIK	mono à penta	9, 10 ou 12 bp	imparfaits	*paramètres de Sputnik : -6 pour le error match*
Yeramian & Buc	1999	TRF	all	25 pb (à cause de TRF)	imparfaits	*paramètres de TRF : +2,-7,-7 ; 50*
Young *et al.*	2000	ns	mono à tetra	15 ou 16 bp	parfaits	

TAB. 3.1 – Liste non exhaustive des méthodes de détection utilisées dans la littérature. ns : la méthode n'est pas explicitée. personnel : l'algorithme est détaillé, ou est précisé avoir été développé « à la main ». TRF est un algorithme publié [Benson, 1999], REPEAT, SPUTNIK et SIMPLE34 ne sont pas publiés mais sont (ou étaient) accessibles librement sur Internet. Les paramètres de TRF et SPUTNIK sont détaillés dans le manuscrit, ceux de REPEAT sont équivalent à ceux de TRF.

Incohérences de détection

période	taille min (nt)	Densité calculée dans chacune des études		
		Nadir, 1996	Kruglyak, 1998	Toth, 2000
2	12	-	2560	1511
	16	906	1866	-
4	12	-	-	1906
	16	1532	-	-
6	12	-	-	419
	16	940	-	-

TAB. 3.2 – Incohérences entre les densités de microsatellites humains détectés dans trois articles publiés. La densité est calculée pour 1 Mb de séquence humaine à partir des tableaux et graphes présentés dans les trois études. Les cas d'incohérence de détection sont en rouge et soulignés. La densité de dinucléotides de 16 nt minimum détectés par Kruglyak *et al.* [Kruglyak et al., 1998] est supérieure à celle de 12 nt minimum détectés par Toth *et al.* [Toth et al., 2000]. C'est la même chose entre les hexanucléotides supérieurs à 16 nt obtenus par Nadir *et al.* [Nadir et al., 1996] et ceux supérieurs à 12 nt de Toth *et al.* [Toth et al., 2000].

3.2 Méthodes et résultats

3.2.1 Description des algorithmes

Une bonne douzaine d'algorithmes ont été publiés concernant spécifiquement la détection des éléments répétés en tandem, mais ils suivent tous l'une des trois logiques majeures que sont l'approche combinatoire, l'approche statistique et l'approche par alignement. Les algorithmes combinatoires testent si chaque base de la séquence fait partie d'un microsatellite, suivant des règles strictes. Les algorithmes statistiques recherchent des régions potentiellement répétées à partir de critères statistiques et n'appliquent les règles strictes de validation qu'aux zones trouvées. On peut déjà noter que cette approche peut omettre des séquences microsatellites qui n'ont pas rempli les critères statistiques, même si elles pouvaient passer les critères de validation. Enfin, les algorithmes du troisième groupe utilisent l'alignement d'un motif ou d'une liste de motifs sur la séquence pour réaliser leurs détections. Les sous-séquences qui donnent un score d'alignement supérieur à un seuil sont validés comme microsatellites.

Les algorithmes comparés ont été choisis pour représenter au mieux ces différentes logiques, tout

en essayant de cibler ceux qui sont le plus utilisés dans la communauté scientifique. Nous avons donc choisi Mreps [Kolpakov et al., 2003] pour représenter les algorithmes combinatoires, TRF [Benson, 1999] et Sputnik [Abajian, 2004] pour les algorithmes statistiques, et STAR [Delgrange and Rivals, 2004] et RepeatMasker [Smit et al., 2004] pour les algorithmes d'alignement.

Mreps

Mreps est un programme utilisant un algorithme combinatoire basé sur les distances de Hamming [Kolpakov and Kucherov, 2003]. Cet algorithme considère que deux sous-séquences adjacentes font partie de la même répétition en tandem si elles sont identiques à k erreurs près, k étant la distance de Hamming. Mreps va donc chercher dans la séquence analysée et pour chaque période demandée, toutes les sous-séquences dont les répétitions adjacentes sont distantes de k au maximum (voir figure 3.1). Les séquences détectées sont ensuite traitées pour ne garder que la période primitive (*i.e.* qui n'est pas constituée de sous-périodes plus petites), pour exclure les erreurs situées aux extrémités (comme sur la figure 3.1), et pour supprimer celles ne répondant pas à certains critères de validation. Ces critères de validation reposent sur la probabilité d'avoir la détection donnée dans une séquence aléatoire et sont détaillés dans l'article de Kolpakov et Kucherov [Kolpakov and Kucherov, 2003]. Ils imposent notamment une taille minimum égale à la période + 9 nt. La distance k est déterminée par un paramètre nommé résolution qui sera détaillé par la suite. Il est important de noter que la distance de Hamming permet de gérer les interruptions de type substitution, mais pas les insertions-délétions. En effet, comme les indels cassent la période de répétition, la comparaison des périodes adjacentes peut donner des distances importantes (c'est le cas dans la figure 3.1). La détection dépend alors de la valeur de k, et de la composition du motif.

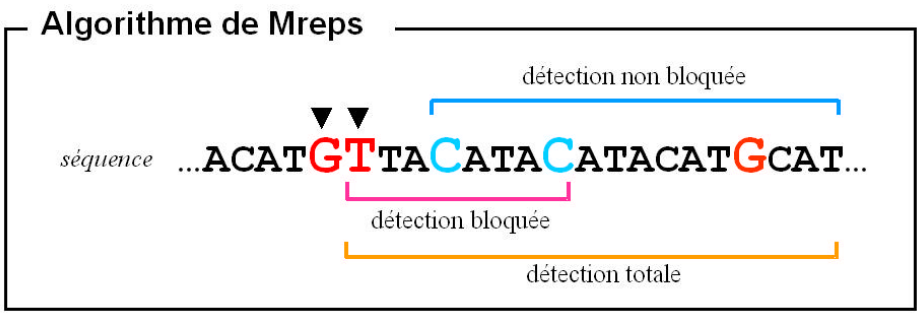


FIG. 3.1 – Détection des microsatellites par Mreps. L'algorithme est présenté pour une période de 4 et une distance de Hamming k de 1. La détection débute par la comparaison des deux motifs qui commencent aux deux C bleus. Elle s'étend ensuite en amont et aval tant qu'il y a au plus k erreurs entre les répétitions. Les flèches noires indiquent les bases provoquant un alignement incorrect. Le T et le G provoquent deux erreurs successives qui arrêtent la détection en amont. La première erreur (le C) est par contre intégrée à la détection.

TRF

TRF est le représentant le plus populaire des algorithmes de détection d'éléments répétés en tandem, et une grande partie des autres algorithmes de cette classe en sont directement dérivés. Son exécution est caractérisée par deux phases principales : une phase de validation statistique et une phase de validation combinatoire. La validation statistique est décomposée en plusieurs critères qui sont détaillés dans la publication de Benson (1999), mais dont nous allons tout de même donner les grandes lignes. TRF commence par stocker la position de tous les motifs d'une période donnée, ainsi que les distances entre les motifs identiques. Ces positions seront appelées des amorces. Des alignements locaux sont ensuite effectués pour les régions situées entre les amorces de même motif, et les alignements corrects sont comptés (figure 3.2). Cela permet d'obtenir des valeurs comme le nombre de bases alignées correctement, la distance moyenne entre ces bases alignées, leurs positions dans la région. Les régions sont alors validées statistiquement comme répétitions en tandem si les valeurs obtenues ne correspondent pas à celles attendues dans des séquences aléatoires (pré-calculées par des formules mathématiques ou par simulation).

La phase de validation combinatoire commence par une recherche du motif le plus probable pour chacune des régions validées statistiquement. Ce motif est alors aligné à la région par un algorithme de WDP (Wraparound Dynamic Programming) [Fischetti et al., 1993]. Le WDP est un alignement local de la séquence sur la série parfaite et infinie de répétitions du motif (Figure 3.2). Comme tout alignement, il nécessite des valeurs de bonus pour les alignements corrects et des pénalités pour les erreurs, et il renvoie un score maximum. La région est alors validée comme détection pour le motif calculé si le score obtenu est meilleur qu'un seuil fixé par l'utilisateur. Il faut préciser que le meilleur score peut n'être donné que par l'alignement d'une sous-partie de la région ; dans ces cas-là, seule cette sous-région sera renvoyée comme détection.

Sputnik

Sputnik est quant à lui une version moins sophistiquée de l'algorithme de TRF. Il ne possède pas de phase statistique mais teste à chaque base de la séquence analysée si un microsatellite d'une période déterminée est présent, en comparant la première période aux périodes qui suivent. L'algorithme utilise un système de score qui est incrémenté pour chaque base identique et décrémenté pour chaque erreur. Cet algorithme est récursif, et à chaque erreur détectée, trois possibilités sont explorées : la substitution, la délétion et l'insertion (voir figure 3.3). Pour chaque possibilité, l'algorithme est ré-exécuté à partir de la base erronée. Ces appels récursifs s'arrêtent lorsque le score devient inférieur ou égal à un score d'arrêt. Puis, les scores obtenus pour chaque possibilité sont comparés,

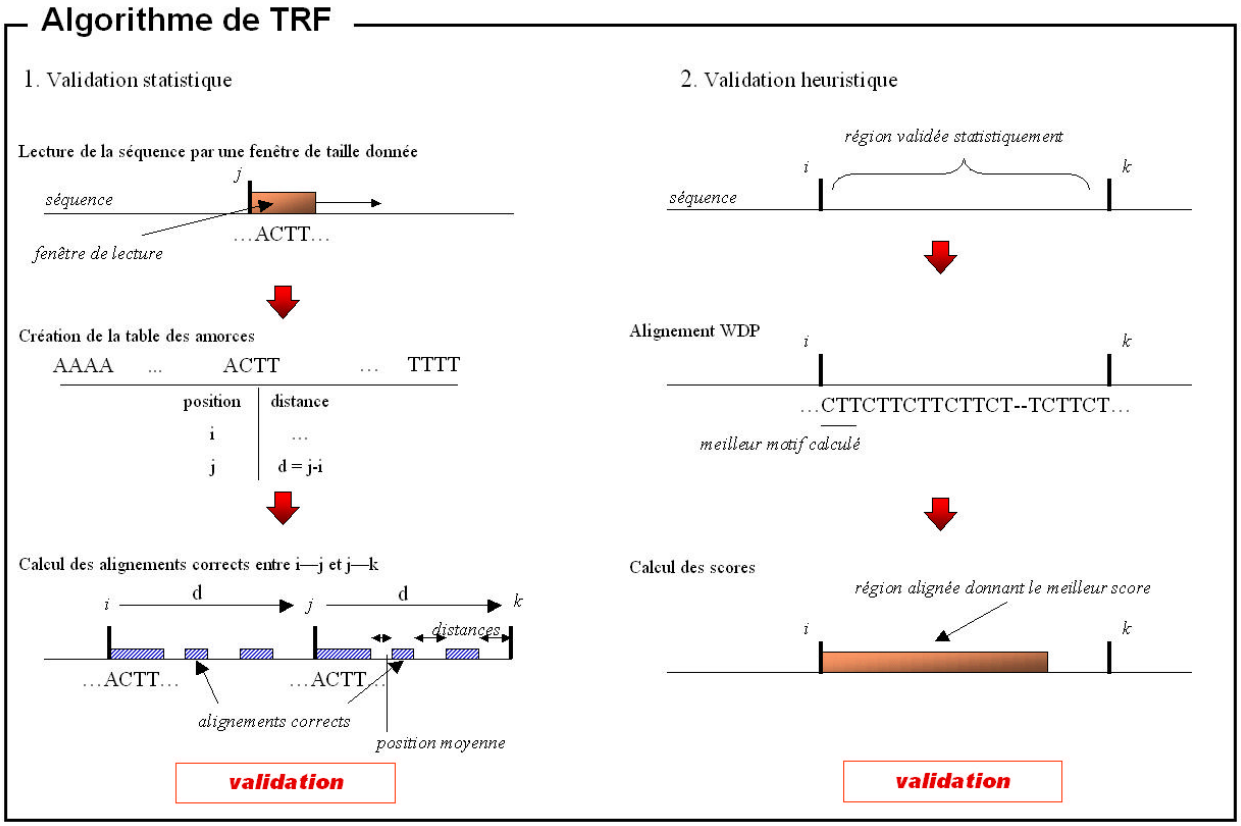


FIG. 3.2 – Détection des microsatellites par TRF. **1-** Première phase : validation statistique. Les positions de tous les motifs (ici i et j) d'une certaine taille (amorces) sont notées dans une table, ainsi que la distance entre chacune des occurrences de chaque motif (d). La séquence entre deux motifs (i–j) est alors alignée à la séquence en aval de même taille (j–k), et les zones d'alignement corrects sont déterminés. La validation dépend du nombre d'alignements corrects, de leur distance moyenne, de leur position moyenne. **2-** Deuxième phase : validation heuristique. La région est alignée par WDP au meilleur motif estimé (ici CTT), et un score d'alignement est obtenu. La région est validée comme répétition en tandem si son score dépasse un certain seuil.

et celui qui présente le score le plus haut est gardé. Il est ensuite comparé avec le score avant l'erreur et le plus haut est gardé. La position de fin de la détection est celle où ce score maximum a été obtenu. Les détections validées comme microsatellites sont ensuite celles dont le score maximum a dépassé un certain seuil. La forme canonique du motif est donnée (*i.e.* AAT au lieu de ATA) et si la taille de la détection n'est pas un multiple de sa période, elle est réduite à la taille multiple inférieure.

STAR

L'algorithme de STAR possède une approche assez originale, basée sur les propriétés de compression des séquences répétées en tandem. Pour chaque motif donné, l'algorithme WDP est exécuté

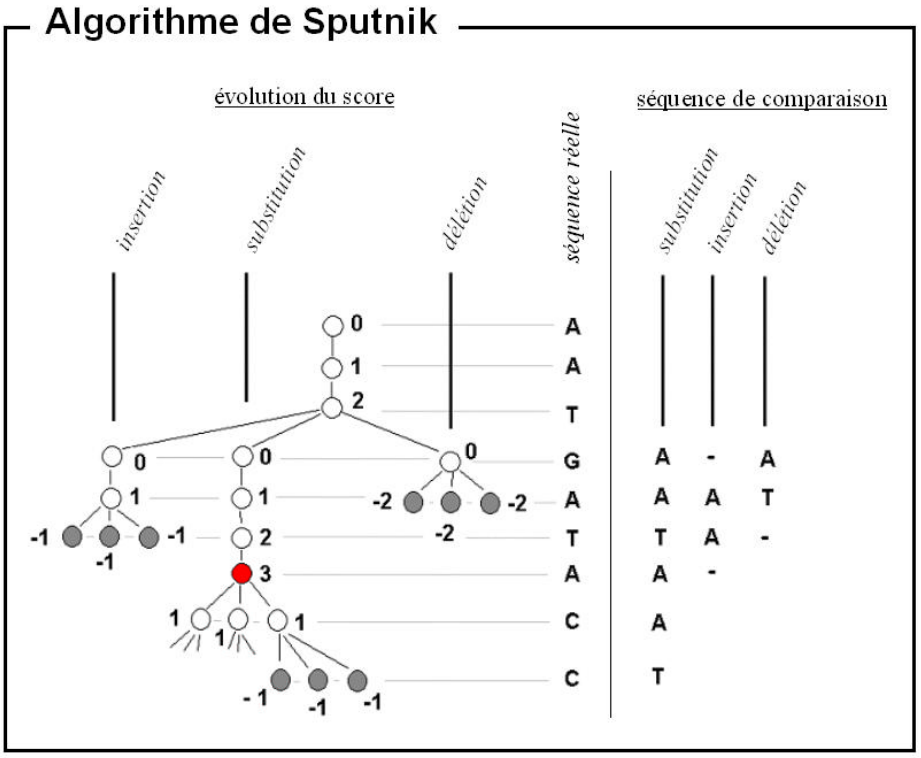


FIG. 3.3 – Détection des microsatellites par Sputnik. L'algorithme est donné pour une période de 3, avec un bonus de +1 pour les alignements corrects, une pénalité de -2 pour les erreurs et un score d'arrêt de -1. L'alignement est toujours réalisé par rapport à la première période, et à chaque erreur, trois appels récursifs sont exécutées. L'algorithme s'arrête lorsque tous les appels récursifs ont atteint le score d'arrêt. Le score final de l'alignement est le plus haut qui ait été atteint (ici 3), et la détection retournée s'arrête à la base ayant donné ce score.

sur la séquence en entier, et donne l'alignement global de la séquence par rapport au motif. L'alignement obtenu est ensuite compressé, avec comme critère de compression le motif (Figure 3.4). A la fin de la compression, les régions où le motif est répété seront fortement compressées, à l'inverse du reste de la séquence. Un gain de compression global est alors calculé pour chaque position : c'est le rapport entre la taille de la séquence compressée et celle de la séquence non compressée, la séquence étant celle comprise entre le début et la position donnée. Ce gain sera croissant pour les zones de répétition du motif et décroissant pour les autres (Figure 3.4). La dernière étape de STAR est de délimiter de manière optimale les zones de croissance du gain de compression, et de les renvoyer en tant que microsatellites. A la différence des algorithmes des deux premières classes, on est ici obligé de donner le motif des microsatellites à chercher, ou une liste de motifs.

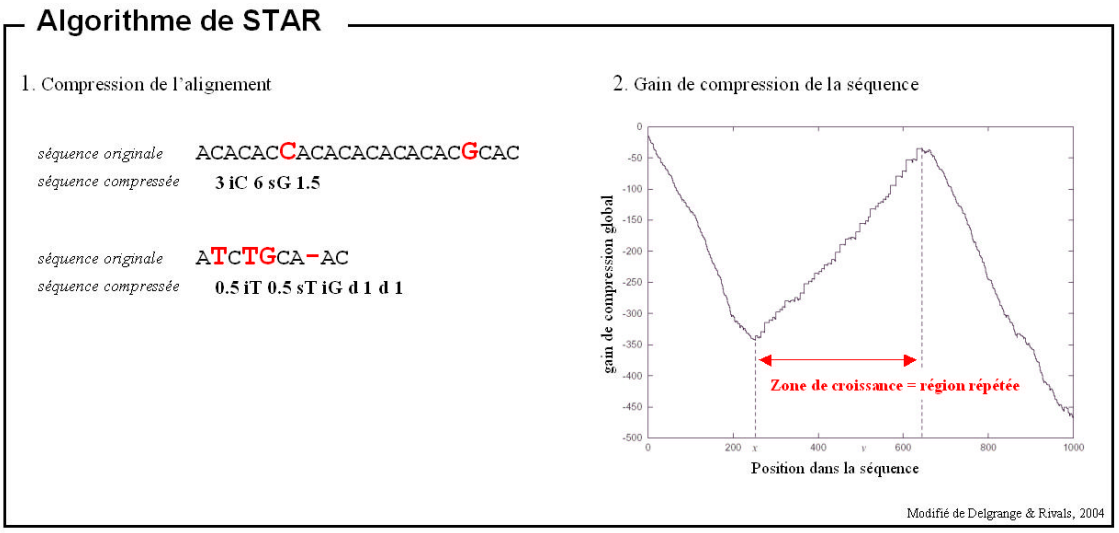


FIG. 3.4 – Détection des microsatellites par STAR. **1-** Première phase : compression de l'alignement. La compression est présentée à partir de l'alignement sur le motif AC (réalisé par WDP). Les répétitions successives parfaites du motif sont codées par leur nombre. Les insertions sont codées par sX, avec X la base erronée, les délétions par d et les insertions par iX, avec X la base insérée. **2-** Deuxième phase : détection des régions répétées à partir du gain de compression. Les régions répétées donnent un gain de compression positif (taille de la séquence compressée/non compressée), alors que les régions non répétées donnent un gain négatif.

RepeatMasker

La méthode utilisée par RepeatMasker, même si elle aussi est basée sur l'alignement, est très différente de celle de STAR. Elle utilise l'élargissement progressif de petites zones d'alignements parfaits.

L'algorithme effectue tout d'abord des alignements locaux d'une séquence répétée consensus de 180 nt d'un motif donné, sur toute la séquence, avec un algorithme de Smith-Waterman [Smith and Waterman, 1981]. Ces alignements locaux vont permettre de déterminer des ancres, qui sont des alignements parfaits de 14 nucléotides. Une région dite répétée sera alors définie autour de chaque ancre (les 14 nt flanquants), et les régions répétées se recouvrant sont fusionnées en une seule (Figure 3.5). Enfin, un score d'alignement de type Smith-Waterman est calculé pour la région répétée, par rapport à la séquence consensus du motif, et la région est détectée comme microsatellite si le score d'alignement dépasse un score de validation fixé. RepeatMasker a été à l'origine conçu pour détecter les divers éléments répétés dispersés des génomes (éléments transposables, pseudogènes, inserts bactériens) à partir de leurs séquences consensus. L'application à la détection des microsatellites est tout à fait valide, car les divers locus microsatellites d'un même motif peuvent être vus comme des éléments répétés dispersés. RepeatMasker travaille à partir d'une bibliothèque de séquences, qui contient l'ensemble des séquences consensus à rechercher. Pour notre application, la bibliothèque

contiendra des séquences représentant la répétition parfaite de chacun des motifs à étudier.

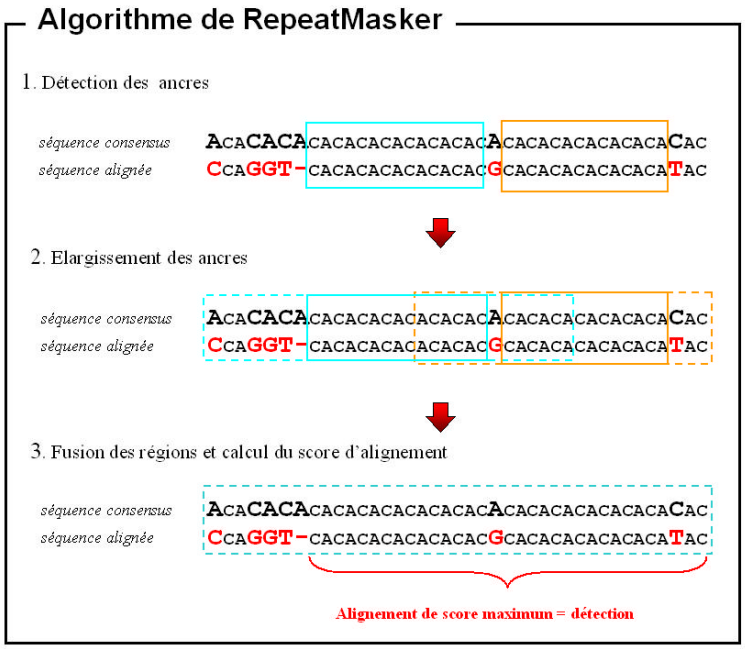


FIG. 3.5 – Détection des microsatellites par RepeatMasker, présenté pour une séquence consensus AC. **1-** Première phase : détection des ancres. Les fragments d'au moins 14 nt parfaitement alignés à la séquence consensus sont détectés. **2-** Deuxième phase : élargissement des ancres. Les fragments sont élargis à 7 nt en amont et en aval. **3-** Troisième phase : fusion et calcul du score d'alignement. Les fragments chevauchants sont fusionnés, et un score d'alignement local Smith-Watermann est calculé pour la fusion. Un microsatellite est détecté pour les régions où le score dépasse un certain seuil.

3.2.2 Influence des paramètres

Tous les programmes exposés ci-dessus, à l'exception de STAR, possèdent des paramètres qui peuvent être définis par l'utilisateur, et qui peuvent influencer la quantité et/ou la qualité des détections renvoyées. Ces paramètres sont détaillés dans l'article en Annexe 1. L'étude des paramètres ne s'est focalisée que sur certains d'entre eux, dont l'influence sur les détections n'était pas forcément triviale au premier abord, et qui sont :

- La résolution de Mreps. La résolution est la valeur qui détermine la distance de Hamming maximum permise entre deux répétitions adjacentes. Pour une résolution K donnée, l'algorithme sera exécuté pour toutes les distances k comprises entre 0 et K, sans dépasser toutefois la taille de la période -1. Les résultats des multiples exécutions sont ensuite fusionnés s'ils se chevauchent.
- Les scores de validation et les valeurs de pénalité de Sputnik et TRF. Le score de validation

est le score que doit atteindre l'alignement pour être renvoyé, et les pénalités sont les valeurs soustraites au score d'alignement pour chaque erreur. Les pénalités peuvent être différentes selon que l'erreur est une substitution ou un indel.

- Le score de validation (cutoff) de RepeatMasker, et la taille des séquences consensus. Le score de validation a la même fonction que pour Sputnik et TRF, mais pour RepeatMasker, il n'est pas possible de modifier les valeurs des pénalités. En effet, ces dernières sont issues de matrices d'alignements internes au programme, qui sont déterminées à partir du taux de GC de la séquence. La taille des séquences consensus est le nombre de répétitions qui constitue les séquences pour chaque motif dans la bibliothèque de consensus.

L'influence de ces paramètres a été analysée en exécutant chacun des programmes sur la séquence complète du chromosome X humain (version 35.1, 29 août 2004, téléchargée sur le site du NCBI : http ://ncbi.nih.gov/Genomes/), en changeant à chaque fois la valeur d'un des paramètres. Pour chaque exécution, le nombre de détections, leur taille moyenne et leur divergence ont été comparés, ainsi que les distributions en taille. Il convient de définir ici ce que l'on a nommé divergence, car chaque algorithme possède sa propre méthode de calcul (l'homologie, qui est l'inverse de la divergence, est parfois utilisée). La divergence est calculée comme le pourcentage d'erreurs présentes dans l'alignement de la séquence détectée et de la séquence consensus répétée de même taille. Ainsi une divergence de 0% désigne une séquence parfaite, et une seule erreur dans un alignement de taille 20 donne une divergence de 5%.

Influence sur le nombre de détections

Le tableau 3.6 montre le nombre de détections, la taille moyenne et la divergence moyenne obtenus pour chacune des exécutions. Les paramètres qui ont le plus d'influence sur le nombre de détections sont les scores de validation de TRF et Sputnik, avec une relation exponentielle négative entre la valeur du score et le nombre de microsatellites détectés. En effet TRF renvoie 22 fois plus de détections pour un score à 20 que pour un score à 50, et Sputnik 43 fois plus entre les scores de 7 et 20. Le nombre de détections est par contre nettement moins affecté par les valeurs de pénalités, puisque l'augmentation n'est que de 1,6 fois et 1,05 fois, respectivement, entre les valeurs les plus astreignantes et les moins astreignantes. L'influence de la résolution de Mreps est à peu près similaire avec une augmentation de 25% (soit 1,25 fois plus) entre la résolution de 1, la plus astreignante, et la résolution de 6 qui autorise plus d'erreurs entre les répétitions. La variation du score de validation et de la taille des consensus de RepeatMasker n'ont pratiquement aucune influence sur le nombre de détections, avec un écart inférieur à 10 détections par mégabase entre les différentes valeurs des paramètres.

Influence des paramètres

		densité	taille	divergence
TRF				
minimum score				
	50	110	64.44	3.96
	40	202	47.65	3.68
	30	458	32.14	3.21
	20	2425	16.07	1.60
align. weights				
	2,7,7	110	64.44	3.96
	2,5,7	125	73.62	6.01
	2,5,5	136	76.44	7.13
	2,3,5	177	83.30	11.31
Mreps				
resolution				
	1	1368	22.96	12.39
	2	1539	28.11	18.47
	3	1636	32.21	22.15
	6	1712	39.80	26.51
Sputnik				
validation score				
	20	154	34.55	1.13
	15	349	25.39	1.06
	8	4273	11.23	0.48
	7	6589	9.74	0.44
mismath penalty				
	-10	6555	9.33	0.01
	-6	6589	9.74	0.44
	-5	6818	10.12	1.19
RepeatMasker				
validation score				
	250	255	53.81	8.37
	225	256	53.97	8.40
	200	256	54.15	8.46
	150	263	54.05	8.86
consensus size				
	60	263	51.34	8.45
	90	261	51.98	8.55
	180	256	53.97	8.40

FIG. 3.6 – Densité (nb/Mb), taille moyenne et divergence moyenne des détections obtenues pour différents algorithmes, avec différents paramètres.

Influence sur la taille des détections

Comme le montre la figure 3.7, la réduction des scores de validation de TRF et Sputnik a pour principale conséquence de permettre la détection de microsatellites plus courts. On peut d'ailleurs noter que plus la taille minimum est courte, plus le nombre de nouvelles détections est important, ce qui explique l'augmentation exponentielle observée pour le nombre de détections. Cela explique

aussi la réduction de la taille moyenne entre les scores les plus hauts et les scores les plus faibles, rapportés dans le tableau 3.6. Le nombre de détections longues est aussi augmenté significativement avec la réduction du score pour TRF, mais pas pour Sputnik.

A l'inverse, l'utilisation de pénalités moins fortes pour TRF et Sputnik, et d'une résolution plus importante pour Mreps, permet d'augmenter la taille moyenne des détections (tableau 3.6). Si l'on observe la figure 3.7, on peut toutefois remarquer que l'augmentation de la taille moyenne n'est pas provoquée par la détection de nouveaux microsatellites plus longs, mais par l'élargissement de détections plus courtes. En effet, le nombre de microsatellites courts est réduit pour les pénalités les moins astreignantes, tandis que le nombre des plus longs augmente.

Les paramètres de RepeatMasker n'ont là encore que peu d'influence sur la taille des détections. On peut remarquer toutefois que la réduction de la taille des consensus provoque des pics de détection à la taille donnée (figure 3.7), mais les différences de distribution ne sont pas significatives (ANCOVA, $F_{3,200} = 0.008$, *p-value* $= 0.999$; voir article en Annexe 1 pour les détails de l'analyse statistique).

Influence sur la divergence

La divergence est affectée de deux manières différentes selon le paramètre qui est changé. Les scores de validation de TRF et Sputnik ont pour effet de réduire la divergence par un facteur 2,5 entre le plus haut et le plus bas score (tableau 3.6), pour les deux algorithmes. Le fait que l'on observe une réduction de la divergence lorsque l'on réduit le score de validation pour TRF et Sputnik signifie que toutes les nouvelles détections sont parfaites ou presque. Les valeurs de pénalité et la résolution de Mreps ont un effet inverse, avec une divergence qui augmente lorsque les paramètres deviennent moins astreignants. Ces résultats confirment ceux obtenus pour la taille moyenne, et soutiennent le fait que les détections réalisées à de faibles valeurs de pénalités sont les mêmes que celles à forte pénalité, élargie grâce à une plus grande tolérance aux interruptions. Comme pour le nombre de détections et la taille moyenne, ni le score de validation, ni la taille des séquences consensus de RepeatMasker ne semblent affecter la divergence significativement.

Résumé

Ces quelques résultats nous montrent que le choix des paramètres peut jouer un grand rôle sur la distribution des détections renvoyées, particulièrement à cause de la détection des microsatellites courts. En effet, l'abaissement des scores de validation de TRF et Sputnik permet de réduire la taille minimum de détection. Or, la figure 3.7 nous montre que les locus sont de plus en plus nombreux

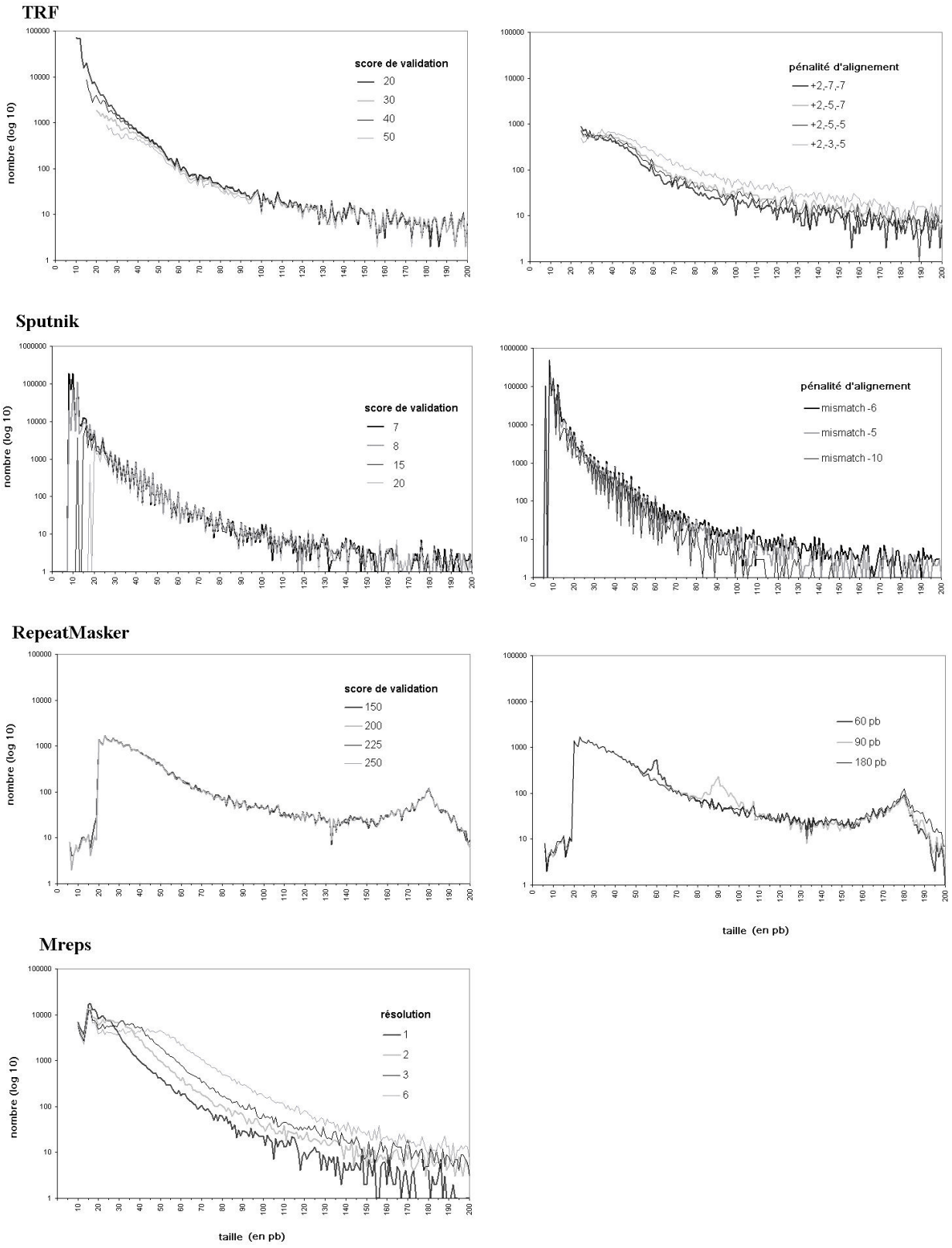


FIG. 3.7 – Distribution en taille des détections pour différents algorithmes, avec différents paramètres. Seules les détections de taille inférieure à 200 nt sont représentées.

à mesure qu'ils sont courts, leur détection faisant baisser du même coup la taille moyenne et la divergence moyenne. Ce résultat est assez inattendu, car un score de validation plus faible permet en théorie d'autoriser plus d'interruptions, et donc d'augmenter les divergence et taille moyennes. L'affaiblissement des pénalités d'erreur pour TRF et Sputnik produit par contre l'effet attendu. La taille moyenne est bien augmentée, ainsi que la divergence, tandis que le nombre de détections n'est que légèrement augmenté, signifiant que la relaxe sur les pénalités autorise plus d'interruptions dans les microsatellites. L'augmentation de la résolution de Mreps produit les mêmes effets, ce qui est aussi le comportement attendu. Tous les paramètres ne provoquent néanmoins pas de telles variations. Aucun des deux paramètres analysés pour RepeatMasker n'a par exemple eu d'influence sur les distributions des détections, tant en nombre, qu'en taille ou en divergence.

3.2.3 Comparaison des algorithmes

La comparaison des différents algorithmes a été réalisée sur la même séquence que l'analyse des paramètres, à savoir le chromosome X humain. Les paramètres des algorithmes ont été fixés au départ, pour éviter de multiplier les tests. La section précédente a démontré que les paramètres avaient une grande influence sur les distributions renvoyées ; il a donc fallu les choisir avec précaution. Après réflexion, nous avons choisi de nous contenter des paramètres donnés par défaut par les concepteurs, avec toutefois une réduction du score de 50 à 20 pour TRF, et de 10 à 7 pour Sputnik. Ce choix a été motivé par le fait que l'on a considéré pour cette analyse une taille minimum de 8 nt pour les microsatellite, comme proposé par Rose & Falush (1998), voir section 2.1.2). Or, nous l'avons constaté dans la section précédente, les scores de 50 et de 10 ne permettaient pas d'obtenir des détections de cette taille.

Nous allons résumer dans cette section les résultats détaillés dans l'article (joint en Annexe 1). La présentation a toutefois été réarrangée, permettant de mieux mettre en valeur les résultats principaux. La comparaison originale était décomposée en plusieurs étapes, qui étaient : la comparaison des microsatellites parfaits uniquement, puis l'ajout des imparfaits à l'analyse, une extension de l'analyse à 3 autres génomes (*D. melanogaster*, *Neurospora crassa* et *S. cerevisiae*), et enfin la comparaison des positions génomiques. Ici, nous nous contenterons de présenter les résultats généraux pour toutes ces analyses, l'importance des détections de petite taille, et les résultats de la comparaison des positions génomiques.

Résultats généraux

En premier lieu, les résultats généraux (Tableau 3 dans article, colonne 'All') montrent de grandes différences entre les divers algorithmes, avec plus de 25 fois plus de détections pour Sputnik que pour RepeatMasker (qui renvoient respectivement le plus et le moins de détections). Le nombre de détections est assez faible pour STAR et RepeatMasker, il est 3 à 5 fois plus important pour Mreps, encore 1,7 fois plus pour TRF, et Sputnik est seul en tête avec ses 6589 détections par mégabase. Il existe de plus une relation entre le nombre de détections et leur taille et divergence moyennes. Plus les microsatellites détectés par un algorithme sont nombreux, plus ils sont courts, et parfaits. Cette règle est valable pour tous les algorithmes sauf Mreps, dont le nombre de détections et la taille moyenne sont intermédiaires, alors que la divergence moyenne est la plus importante.

Les exécutions ont été ensuite reproduites sur les génomes de divers organismes pour évaluer l'effet de la séquence génomique sur les résultats. Les résultats obtenus montrent une relative stabilité dans les comportements de chacun des algorithmes, quel que soit l'organisme (Tableau 3 de l'article). En effet, l'ordre des algorithmes ne change pas quand on s'intéresse au nombre de détections, Sputnik retournant à chaque fois nettement plus de microsatellites que les autres, suivi de TRF puis Mreps, et STAR et RepeatMasker arrivant systématiquement à la fin. De plus, pour chaque organisme, la relation entre le nombre de détections et les tailles et divergences moyennes est respectée, signifiant qu'elles sont causées par la méthode de détection et non par la séquence génomique originale.

Cette analyse a par ailleurs permis de constater que l'homme, la drosophile et *N. crassa* possédaient une densité de microsatellites équivalente, et que seule la levure *S. cerevisiae* a une densité légèrement moins élevée. Les tailles et divergences moyennes sont quant à elles totalement similaires quel que soit l'organisme, pour un même algorithme de détection. Ces résultats ne concernent pas directement la comparaison des algorithmes, mais ils indiquent que la densité des microsatellites est moins variable entre organismes que ce qui était estimé jusqu'alors (*c.f.* section 2.1.3).

Influence de la taille des détections

Les différences en nombre de détections, et les relations avec la taille et la divergence sont directement liées à la capacité des algorithmes à détecter les microsatellites courts et parfaits, comme c'était le cas avec les variations de paramètres. En effet, l'analyse des détections imparfaites uniquement (Tableau 3 de l'article, 'Imperfect') et de la distribution en taille des parfaites (Figure 2 de l'article) pour l'homme indique que d'une part les algorithmes donnent des résultats beaucoup plus proches pour les imparfaits que pour l'ensemble des microsatellites, et d'autre part que la majorité

des détections parfaites de TRF et Sputnik sont comprises entre 8 et 12 nt (selon la classe de motif).

Or tous les algorithmes ne sont pas équivalents pour ce qui est du choix de la taille minimum. Pour RepeatMasker par exemple, une taille minimum de 20 nt non modifiable est imposée. Cette contrainte limite de fait l'efficacité de RepeatMasker en termes de nombre de détections, puisque la très grande majorité est inférieure à cette taille. STAR ne possède pas de contrainte aussi stricte, mais la figure 2 de l'article montre un nombre de détections croissant jusqu'à la taille de 20 nt (sauf pour les hexanucléotides, qui ont le pic à 10 nt), pour des raisons qui sont expliquées dans la discussion. Enfin Mreps, à cause du filtre de significativité, ne peut détecter les microsatellites de taille inférieure à la taille de leur motif + 9. Cette contrainte a assez peu d'influence pour les mono et dinucléotides, mais pénalise fortement la détection des tetra, penta et hexanucléotides dont la très grande majorité se situe en dessous du seuil.

Recouvrement inter-algorithme

Le recouvrement inter-algorithme a été calculé pour évaluer la proportion de microsatellites détectés aux mêmes positions entre les algorithmes. Deux méthodes ont été testées, se basant respectivement sur le recouvrement en nombre de locus et de nucléotides. Le recouvrement en nombre de locus montre que plus de 93% des détections de RepeatMasker et STAR sont aussi détectés par Sputnik, Mreps et TRF (Tableau 4 de l'article), tandis que le recouvrement des trois derniers par les deux premiers est nettement moins important (moins de 34% des détections de Mreps, 20% pour TRF et 10% pour Sputnik). Ces résultats sont en accord avec les nombres de détections totaux obtenus pour chacun des algorithmes. Le recouvrement en quantité de nucléotides est légèrement différent, avec un recouvrement moins prononcé de Sputnik, TRF et Mreps sur les deux autres, et inversement, un recouvrement plus prononcé de RepeatMasker et STAR sur les trois autres. Ces légères distinctions sont à corréler avec le fait que les détections de STAR et RepeatMasker sont plus longues en moyenne que celles des trois autres algorithmes, leur couverture étant donc plus efficace en termes de nucléotides.

3.3 Discussion

Les résultats présentés dans ce chapitre amènent un certain nombre de discussions concernant les capacités des diverses logiques de détection des microsatellites. Ces discussions, dont les propos sont plutôt d'ordre bio-informatique, sont largement détaillées dans l'article de l'Annexe 1 et ne seront donc pas ré-exposés ici. Nous nous permettons toutefois d'attirer l'attention sur quelques points

précis qui sont directement liés à la problématique de cette thèse.

Nous avons montré que les plus grosses différences de détection, que ce soit entre algorithmes, ou même entre diverses valeurs de paramètres pour un même algorithme, sont causées par la détection ou non de microsatellites parfaits très courts. Les tailles de ces locus sont comprises entre 8 et 12 nt, mais dépendent de la classe de motif. Même si les mono, di et trinucléotides de 8-10 nt (soit 8-10 répétitions pour les mono, 4-5 répétitions pour les di, et 3 répétitions pour les tri) sont très nombreux, la majeure partie des détections courtes est constituée des tétranucléotides de 2 à 3 répétitions, et des penta et hexanucléotides de 2 répétitions (10 et 12 nt respectivement).

Pouvoir détecter ces locus est d'une importance primordiale lorsque que l'on veut étudier l'apparition des microsatellites, car les apparitions *de novo* sont justement attendues dans cette gamme de taille (voir chapitre 2). De plus, les tétra à hexanucléotides courts semblent bien plus présents que l'attendu théorique, du moins pour certains motifs. Cette observation suggère que la sur-représentation des locus à très peu de répétitions est aussi valable pour les motifs plus petits (mono, di et tri), mais que nous n'ayons pu les détecter avec les scores de validation choisis pour Sputnik et TRF. Cette hypothèse a d'ailleurs déjà été proposée par divers auteurs [Pupko and Graur, 1999, Dieringer and Schlotterer, 2003], comme nous le verrons dans le chapitre 5.

La seconde remarque que ces travaux soulève concerne la gestion de l'imperfection. Les résultats présentés pour les diverses méthodes de détection nous ont montré que cette notion d'imperfection peut être interprétée de multiples manières. Tout d'abord les valeurs de divergence ou d'homologie données par les algorithmes ne sont pas forcément calculées de la même manière, même si elles ont toutes été normalisées dans notre comparaison. Cela signifie qu'une même séquence microsatellite n'aura pas la même valeur d'imperfection selon l'algorithme avec lequel elle a été détectée, pouvant aboutir à des incohérences si plusieurs études se focalisent sur les imperfections avec des méthodes de détection différentes.

Le vrai problème n'est pourtant pas dans la manière de calculer la divergence mais dans la manière de la concevoir. Par exemple, RepeatMasker débute sa détection par la recherche de séquences répétées parfaites de 14 nt au moins (*cf.* figure 3.5), tous les microsatellites de divergence supérieure à 7% (1 sur 14), même très longs, qui possèdent des interruptions de manière uniforme ne pourront être détectés. Cet algorithme privilégie donc les séquences avec de larges zones répétées parfaites. La conception de l'imperfection pour Mreps est encore plus éloignée de la conception classique. Comme

cet algorithme travaille uniquement avec des périodes répétées et non des motifs, les microsatellites composés (comme CACACACATATATA) seront considérés comme très peu divergents (1 seule erreur, le passage de CA à TA), alors que notre calcul donnerait une divergence importante.

Toutes ces considérations nous amènent à la conclusion que pour l'instant, l'interprétation de résultats reposant sur l'emploi d'algorithmes de détection dédiés comporte certains risques, liés à l'implémentation des méthodes. Etant donnés les problèmes que posent encore les interruptions pour la détection des microsatellites, le mieux est de se contenter d'études sur les microsatellites parfaits uniquement. Cela peut être réalisé de deux manières : soit en retirant les microsatellites imparfaits obtenus avec l'une des méthodes gérant les imperfections, soit en utilisant un algorithme simple et en contrôlant les séquences flanquantes pour s'assurer du caractère parfait des détections (comme réalisé dans l'étude de Calabrese et Durrett (2003)). Si la détection des microsatellites imparfaits est nécessaire, il vaut alors mieux utiliser l'un des algorithmes dédiés, mais se restreindre à des valeurs de paramètres assez astreignantes pour éviter d'obtenir des détections incohérentes. Dans tous les cas de figure, la nécessité absolue est de détailler le plus possible la manière dont les microsatellites sont détectés, et les valeurs des paramètres utilisées si besoin est.

Enfin, cela nous amène au choix des algorithmes que nous allons utiliser dans cette thèse. Il s'avère que chacune des sous-parties ne nécessite pas la détection du même type de microsatellites, nous allons donc utiliser deux méthodes différentes. Les travaux sur la relation entre les microsatellites et les séquences Alu (Chapitre 4) seront basés sur les détections renvoyées par TRF, avec un score de validation de 20 et des pénalités de +2,-5,-7. En effet, l'hypothèse d'apparition des microsatellites par adoption repose sur le fait que les séquences insérées sont déjà au moins dans leur phase d'expansion, donc assez longues et pourquoi pas imparfaites. La contrainte de la taille a donc été abaissée au minimum via le score, et la cohérence des microsatellites imparfaits a été contrôlée par des pénalités assez fortes. Pour les travaux sur l'apparition *de novo* (Chapitre 5), il était par contre nécessaire de détecter les proto-microsatellites, d'au moins deux répétitions, et ce quel que soit la taille du motif. Comme aucune des méthodes que nous avons comparé ici ne nous permettaient d'obtenir des séquences aussi courtes, nous avons utilisé un algorithme personnel détaillé dans le chapitre concerné.

Chapitre 4

Apparition via les séquences Alu

4.1 Relation microsatellites - séquences Alu

Comme nous l'avons expliqué dans la section 1.3, les séquences Alu sont des rétrotransposons sans LTR fortement présents dans le génome des primates (plus de 1 million de copies). Ils possèdent de plus une partie terminale constituée de 20 à 50 adénines que l'on nomme queue polyA, et d'une partie centrale $(A)_{5-6}TAC(A)_{5-6}$ appelée *linker*. Ces trois propriétés font des éléments Alu un vecteur possible d'apparition de microsatellites, et la relation entre ces deux types d'éléments a déjà fait l'objet d'études par le passé [Arcot et al., 1995, Nadir et al., 1996, Yandava et al., 1997]. Par ailleurs, le fait que le microsatellite $(GAA)_n$ responsable de la maladie à triplet Ataxie de Friedreich soit issu du linker d'un Alu a suscité quelques recherches sur la relation entre les microsatellites GAA et les éléments Alu [Chauhan et al., 2002, Clark et al., 2004].

Les études réalisées se sont généralement focalisées sur l'association entre microsatellites de tous types et la queue polyA, et sur le rapport entre familles Alu et types de microsatellites. Une synthèse de ces résultats, qui sont parfois contradictoires, est présentée dans cette section. Le nombre de travaux effectués reste néanmoins assez faible, et certains sont plutôt anciens et reposent sur un petit nombre de données. Ces divers problèmes sont détaillés dans la deuxième partie de cette section, de façon à faire ressortir la problématique ayant conduit au travail présenté dans ce chapitre.

4.1.1 Etat de l'art

Relation de proximité

Deux études principalement se sont intéressées à l'association entre microsatellites et séquences Alu [Jurka and Pethiyagoda, 1995, Nadir et al., 1996]. Les deux sont basées sur des microsatellites

parfaits et des Alu détectés dans des séquences Genbank (6 Mb de génome de primate pour Jurka et Pethiyagoda (1995) et 2,84 Mb de séquence humaine pour Nadir *et al.* (1996)). Ils relèvent que respectivement 76 et 85% des mononucléotides $(A)_n$ sont situés à proximité de séquences Alu (tableau 4.1). La différence peut tenir au fait que les données génomiques d'origine ne sont pas les mêmes, tout comme la distance maximum utilisée pour définir la proximité entre les éléments. En effet, Nadir *et al.* (1996) autorisent une distance maximum de 100 nt, alors qu'elle est limitée à 50 nt pour Jurka et Pethiyagoda (1995). Quoi qu'il en soit, ces taux très importants suggèrent que les Alu sont les principaux vecteurs des microsatellites polyA. De même, la grande majorité (entre 60 et 94% selon les études et les motifs) des microsatellites très riches en A ($(AAAX)_n$, $(AAAAX)_n$ et $(AAAAAX)_n$, avec X égal à C, G ou T) sont aussi à proximité des Alu. Enfin, tous ces microsatellites proches des Alu sont à plus de 80% situés directement en 3' de l'élément transposable. Ces divers résultats semblent donc indiquer que les éléments Alu, via leur queue polyA, sont des vecteurs significatifs d'apparition de microsatellites, au moins pour ceux riches en A.

Pour les di et trinucléotides, le constat est plus nuancé (tableau 4.1). Les dinucléotides ne sont associés qu'au maximum à 35% aux Alu, pour les $(AT)_n$, et une étude complémentaire confirme que seuls 17% des $(AC)_n$ sont à proximité de Alu [Arcot et al., 1995]. 60-75% des trinucléotides AAT et AAC sont à proximité de Alu, et le taux est presque deux fois plus faible pour les AAG (36-44%). Deux études plus récentes viennent toutefois corriger le taux pour les AAG à 62-63% [Chauhan et al., 2002, Clark et al., 2004]. En revanche, ces di et trinucléotides sont plus présents dans le linker des Alu que ne le sont les tetra, penta et hexanucléotides.

En résumé, ces résultats indiquent que les rétrotransposons Alu ont une forte influence sur la présence des séquences répétées, principalement celles riches en A. La queue polyA insérée avec chaque Alu en est le principal vecteur, même si le linker semble capable de favoriser l'apparition de trinucléotides. Les motifs non-riches en A sont par contre très peu associés aux séquences Alu, comme le montre les tests réalisés pour les microsatellites $(CGG)_n$, $(AGGC)_n$, $(AGGG)_n$ et $(AGAGGG)_n$ (tableau 4.1).

Influence des familles Alu

Comme nous l'avons expliqué dans l'introduction sur les séquences Alu (section 1.3), ces dernières sont classées en familles, selon l'ancienneté de leur introduction dans le génome. Les trois grandes familles sont les AluJ (les plus vieux), les AluS, et les AluY (les plus récents), chacune étant divisée en sous-familles. Il est dès lors possible de s'intéresser à l'association Alu-microsatellite en

Relation microsatellites-Alu dans la littérature

motif	proportion liée aux Alus	position (% de ceux liés) 3'	position (% de ceux liés) linker	référence
A	76 - 85	91	9	*[1] [3]*
AC	11 - 17 - 26	75 - 85	15 - 25	*[1] [2] [3]*
AG	21 - 27	67	33	*[1] [3]*
AT	27 - 35	67	33	*[1] [3]*
AAC	64 - 75 - 100	54 - 92	8 - 46	*[1] [3] [5]*
AAG	36 - 44 - 62 -63	75 - 90 - 94	25 - 10 - 6	*[1] [3] [4] [5]*
AAT	60 - 74 - 87	77 - 66	23 - 33	*[1] [3] [5]*
AGC	0			*[3]*
AGG	5 - 28 - 29	25 - 100	0 - 75	*[1] [3] [5]*
CGG	0			*[3]*
AAAC	67 - 85	93	6	*[1] [3]*
AAAG	64 - 94	85	14	*[1] [3]*
AAAT	79 - 89	94	5	*[1] [3]*
AAGG	51 - 89	75	25	*[1] [3]*
AATC	51 - 75	100	0	*[1] [3]*
AATG	12 - 25	100	0	*[1] [3]*
ACAT	50 - 57	50	50	*[1] [3]*
AGAT	25 - 36	100	0	*[1] [3]*
AGGC	23			*[1]*
AGGG	10			*[1]*
AAAAC	62 - 77	100	0	*[1] [3]*
AAAAG	77 - 82	93	7	*[1] [3]*
AAAAT	76 - 80	100	0	*[1] [3]*
AAATT	79			*[1]*
AAAAAC	71 - 85	96	4	*[1] [3]*
AAAAAG	82 - 83	95	5	*[1] [3]*
AAAAAT	71 - 87	81	19	*[1] [3]*
ACCCCC	0			*[3]*
AGAGGG	20			*[3]*

[1] Jurka & Pethiyagoda 1995 **[2]** Arcot *et al.* 1995 **[3]** Nadir *et al.* 1996
[4] Chauhan *et al.* 2002 **[5]** Clark *et al.* 2004

TAB. 4.1 – Taux de proximité entre microsatellites et séquences Alu (références en pied de tableau), pour quelques motifs microsatellites. Les proportions de position sont basées sur le total des microsatellites associés aux Alu.

fonction de la famille des Alu.

Cela a été réalisé pour les AC [Arcot et al., 1995, Yandava et al., 1997], les GAA [Chauhan et al., 2002, Clark et al., 2004], et quelques autres motifs [Yandava et al., 1997, Clark et al., 2004]. Les résultats résumés dans le tableau 4.2 indiquent que les microsatellites sont associés aux AluS pour plus de 50% d'entre eux, entre 16 et 30% aux AluJ, et le reste aux AluY. Ces résultats, bien que variables selon le motif du microsatellite, respectent la répartition globale des familles Alu dans le génome. Les AluS représentent en effet approximativement 60% du total des éléments Alu, les AluJ, 25%, et les AluY, 15% (selon Chauhan *et al.* 2004).

Toutefois, on relève un certain nombre de contradictions entre ces études. Clark *et al.* (2004) donnent par exemple un taux d'association de 60% entre les trinucléotides AGG et les AluJ, alors qu'il est limité à 18% pour Yandava *et al.* (1997). Là encore, la méthode de calcul de l'association peut expliquer cette incohérence, Yandava *et al.* (1997) ayant utilisé un logiciel de détection des séquences Alu plus ancien (PHYTIA) que celui utilisé par Clark *et al.* (2004)(RepeatMasker avec la librairie RepBase Update), et pas sur le même type de données (uniquement des marqueurs microsatellites existants pour les premiers, et l'ensemble des locus du génome pour les seconds). Une autre incohérence est à noter pour les $(AC)_n$ que Yandava *et al.* (1997) estiment à 47% associés aux AluY, proportion réduite à seulement 5% pour Arcot *et al.* (1995).

Yandava *et al.* (1997) ont aussi conduit une analyse statistique de représentativité en fonction de la famille Alu, mais nous n'allons pas la présenter ici. En effet, les quelques incohérences de détection déjà citées, plus la grande différence entre les proportions par familles Alu données par Chauhan *et al.* (2002) et Yandava *et al.* (1997), semblent indiquer qu'il existe un biais dans les données de ces derniers. Leurs tests statistiques portent par conséquent sur des données non représentatives de l'ensemble des associations, et sont donc sans réelle valeur. Chauhan *et al.* (2002) et Clark *et al.* (2004) ont par contre montré que l'association entre $(AAG)_n$ et AluY est inférieure à celle attendue par rapport au nombre total de Alu. Cela semble aussi vrai pour les $(AC)_n$, dont 5% seulement sont proches des AluY, au lieu des 15% attendus.

Lien entre microsatellites et familles Alu dans la littérature

famille de Alu	proportion par famille	proportion par famille de Alus pour chaque motif microsatellite (%)								
		AC	AAC	AAT	AAG	AGG	AAAG	AAAT	AGAT	AAGG
AluJ	23 - **25**	**36** - 21	**25**	**16** - 21	30 - **35**	18 - **60**	35	20	23	18
AluS (dont Sx)	50 - **60**	**59** - 32 (2 - **21**)	**58**	70 - **73** (42)	**60** - 65	**40** - 55 (18)	50 (21)	67 (44)	36 (13)	56 (28)
AluY	**15** - 27	**5** - 47	**17**	9 - **11**	**5** - 5	**0** - 27	15	13	41	26

TAB. 4.2 – Proportion des microsatellites associés aux Alu, selon la famille Alu. Les proportions de Alu proviennent de Yandava *et al.* (1997) (écriture normale) et de Chauhan *et al.* (2002) (en gras). Les proportions en gras pour les AC proviennent de Arcot *et al.* (1995). Celles en gras pour les trinucléotides ont été estimées à partir de la figure 2B de Clark *et al.* (2004). Toutes les autres sont extraites de Yandava *et al.* (1997), à l'exception des proportions pour les GAA, tirées de Chauhan *et al.* (2002).

Taille des microsatellites associés aux séquences Alu

Les travaux présentés dans la littérature se sont principalement intéressés à la taille des polyA, et des $(AAG)_n$. Les polyA ont une taille moyenne qui se réduit avec l'âge des Alu. Ceux associés aux AluY ont en effet une taille moyenne de 21,5 nt, qui se réduit à 19,8 et 17 nt pour ceux associés

aux AluS et AluJ, respectivement [Chauhan et al., 2002]. A l'inverse, les $(AAG)_n$ ont tendance à grandir avec l'âge, puisque ceux associés aux AluJ font en moyenne quatre répétitions de plus que ceux associés aux AluS [Clark et al., 2004]. Il y a de plus une corrélation négative entre la taille du polyA et la taille du AAG, lorsque ceux-ci sont détectés côte à côte en région 3' du même élément Alu [Chauhan et al., 2002]. Ces derniers résultats soutiennent l'hypothèse d'apparition des GAA par dégradation des polyA originaux, créant ainsi des microsatellites composés (voir section 2.1.2).

Les polyA ont donc une taille moyenne de 21-22 nt dans les AluY, mais une étude plus détaillée sur les différentes sous-familles a montré que la taille à l'insertion est nettement plus importante [Roy-Engel et al., 2002]. En effet, certains AluYa5a2, l'une des rares familles encore actives actuellement, possèdent des polyA pouvant atteindre 50 nt. L'étude du polymorphisme de longueur de ces locus polyA indique par ailleurs que beaucoup sont bimodaux, avec une série d'allèles autour de 40/50 nt, et une autre série autour de 20/30 nt [Roy-Engel et al., 2002].

4.1.2 Problématique

La littérature nous renseigne donc sur la relation entre microsatellites et séquences Alu. L'association est très forte avec les polyA, et est de manière générale proportionnelle à la richesse en A du motif, les dinucléotides étant par exemple assez peu associés aux Alu. Les comparaisons entre les diverses familles Alu indiquent que l'apparition des microsatellites est causée par la dégradation de la queue polyA des éléments insérés, dont la taille se réduit avec le temps.

Certaines critiques sont toutefois à apporter à ces diverses études. La première est le manque de données dans les travaux les plus anciens. Arcot *et al.* (1995) ne se sont basés que sur 365 marqueurs AC, et les deux études les plus générales (Jurka et Pethiyagoda (1995), Nadir *et al.* (1996)) ont extrait les microsatellites d'une sous-partie limitée du génome humain/primate. Ils n'ont de plus extrait que les microsatellites parfaits, certains motifs sont donc présents en très petit nombre (par exemple, 15 représentants ou moins pour chaque motif trinucléotide chez Nadir *et al.* (1996)). A l'inverse, les articles plus récents travaillant sur des chromosomes complets se sont limités à l'analyse des trinucléotides, plus particulièrement des GAA. Enfin, toutes ces études (excepté Jurka et Pethiyagoda (1995)) ont imposé une taille minimum de détection assez importante (15, 16, et 24 nt respectivement pour Chauhan *et al.* (2002), Clark *et al.* (2004) et Nadir *et al.* (1996)). Nous avons vu au chapitre 3 que la taille minimum de détection avait des conséquences importantes sur le nombre de microsatellites renvoyés, les plus courts constituant la majorité des locus. La représentativité peut donc être biaisée si l'on se contente d'analyser la proximité des Alu avec les microsatellites les plus

longs.

Il n'y a pour l'instant aucune étude présentant la relation microsatellites-Alu pour l'ensemble des motifs microsatellites et à l'échelle de génomes complets. Une telle analyse paraît pourtant nécessaire afin de confirmer et préciser les résultats précédents. Ce constat nous a donc poussé à reprendre un certain nombre des analyses présentées ci-dessus, en les généralisant à un jeu de données bien plus conséquent, qu'est la séquence complète du génome humain. La suite de ce chapitre est consacrée à la présentation des résultats obtenus, que ce soit les proportions de microsatellites associés aux Alu en fonction de leur motif, de leur position dans la séquence Alu, de la famille Alu, et en fonction de leur taille.

D'autre part, la grande majorité des articles cités ici avaient pour stratégie de rechercher les Alu uniquement dans les 100 ou 500 nt flanquants des microsatellites [Arcot et al., 1995, Chauhan et al., 2002, Clark et al., 2004], ou même de n'analyser que ceux déjà associés aux Alu [Yandava et al., 1997, Roy-Engel et al., 2002]. Etant donné le nombre très important de séquences Alu présentes dans le génome humain (plus de 1 million), il n'est pourtant pas exclu que la proximité entre certains microsatellites et les Alu ne soit que le fait du hasard. Nadir *et al.* (1996) ont proposé une analyse statistique pour évaluer l'influence réelle des Alu dans l'apparition des microsatellites, mais leur manque de données pour certains motifs a affaibli la puissance des tests. Nous reprenons donc ce test pour évaluer la significativité des associations microsatellites-Alu à partir de nos données.

Enfin, les analyses des relations entre microsatellites et séquences Alu s'est essentiellement focalisée sur l'apport potentiel de la queue polyA, et dans une moindre mesure sur celui du linker. Une séquence Alu n'est pourtant pas restreinte à ces deux zones particulières, et les 260 paires de bases des deux monomères pourraient aussi être vecteurs de microsatellites. En effet, de nombreuses zones ont un fort taux de GC, créant des régions de faible complexité. La possibilité d'apparition de microsatellites à l'intérieur des Alu a donc aussi été explorée durant ma thèse.

4.2 Méthodes et résultats

4.2.1 Extraction des données

Détection des microsatellites

Fort de notre expérience concernant les différents programmes de détection de microsatellites disponibles dans la littérature (voir chapitre 3), nous avons pu choisir celui qui était le plus appro-

prié à nos besoins. Notre choix s'est tourné vers TRF [Benson, 1999] pour plusieurs raisons :

- Nécessité de prendre en compte les microsatellites imparfaits. Les grandes vagues d'amplification des éléments Alu ayant eu lieu il y a plus de 30 Ma, les microsatellites qui leur sont associés ont nécessairement accumulé des mutations ponctuelles. Ne pas tenir compte des microsatellites imparfaits aurait donc largement biaisé les relations avec les Alu les plus anciens.
- Besoin de catégoriser les microsatellites en fonction de leur motif. L'apparition des microsatellites à partir des queues polyA passant *a priori* par une phase de microsatellites composés, il faut pouvoir distinguer les deux sous-parties de ces locus. Des logiciels comme Mreps n'auraient par exemple détecté qu'un seul locus composé au lieu de deux locus adjacents comme peut le faire TRF (voir chapitre 3).
- Ne pas limiter la détection aux locus les plus longs. La version de TRF utilisée permet en effet de récupérer les locus d'une taille minimum de 10 nt, quelle que soit la classe de motif (à part les hexanucléotides où deux répétitions sont nécessaires, donc 12 nt), alors que presque tous les autres algorithmes sont limités à des tailles supérieures. Sputnik autorise aussi une taille minimum très faible (on peut même aller jusqu'à 8 nt), mais son implémentation limite fortement le nombre de microsatellites imparfaits détectés.

Le choix des valeurs paramétrables de TRF joue aussi un rôle important sur les détections. Nous avons choisi un score minimum de détection de 20 (qui fixe la taille minimum de détection à 10 nt), et des valeurs de pénalité d'alignement à {+2,-5,-7}. Ces valeurs signifient que chaque base correcte par rapport au microsatellite parfait rapporte deux points, alors qu'une substitution en fait perdre cinq et un indel sept. Ces valeurs sont un bon compromis entre nombre de détections imparfaites et leur taux d'imperfection. L'extraction a été réalisée sur tous les chromosomes du génome humain (version 35.1, 29 août 2004, téléchargée sur le site du NCBI : http ://ncbi.nih.gov/Genomes/). Les données ont ensuite été nettoyées (élimination de la redondance) selon la procédure détaillée dans l'article concernant la comparaison des algorithmes de détection (Annexe 1).

La densité de microsatellites obtenus par classe de motifs est résumée dans le tableau 4.3. La taille moyenne et la proportion de génome occupée sont aussi indiquées. Ces valeurs sont légèrement variables selon les chromosomes, mais des tests de χ^2 ne donnent aucune différence significative entre chromosomes pour ces caractéristiques.

Densité des microsatellites dans le génome humain

	mono	di	tri	tetra	penta	hexa	**total**
densité (nb/Mb)	419	189	227	564	587	638	**2624**
taille moyenne (pb)	18.8	30.3	15.8	17.2	14.5	15.1	**18.6**
proportion du génome	0.79%	0.57%	0.36%	0.97%	0.85%	0.97%	**4.89%**

TAB. 4.3 – Densité moyenne, taille moyenne des microsatellites, et proportion de séquence qu'ils occupent, selon la classe de motifs, pour l'ensemble du génome humain. Les séquences microsatellites ont été extraites avec TRF, paramètres{2,5,7 ; 20}.

Détection des séquences Alu

La détection des éléments transposables dans les génomes est, comme celle des éléments répétés en tandem, un problème qui a suscité et suscite encore l'intérêt de bon nombre d'algorithmiciens [Volfovsky et al., 2001, Bao and Eddy, 2002, Edgar and Myers, 2005, Morgulis et al., 2006]. Nous n'avons pu conduire de comparaison des différents algorithmes publiés comme celle réalisée pour les algorithmes de détection des microsatellites, et nous nous sommes contentés d'utiliser le logiciel de référence, RepeatMasker (v 3.1.0 ; Smit *et al.* 1999).

Ce logiciel aligne de manière locale des régions de la séquence d'entrée avec des séquences de référence (consensus) regroupées dans une bibliothèque de séquences (voir section 3.2.1). La bibliothèque de séquences consensus utilisée est la bibliothèque de référence Repbase Update (version 9.11, accessible sur le site du Giri, http ://www.girinst.org/). Elle contient une grande partie des éléments répétés décrits dans la littérature (y compris les microsatellites), quel que soit l'organisme. Pour nos analyses, nous avons préalablement retiré tous les consensus d'éléments répétés autres que ceux définis comme Alu, et nous avons supprimé les queues polyA présentes à la fin de chaque séquence consensus. Il était en effet plus simple pour nous de calculer la proximité entre les microsatellites et les Alu à partir du 3' non polyA des Alu.

La densité des Alu est donnée par famille dans le tableau 4.4. Là encore, des tests de χ^2 n'indiquent aucune différence significative entre les chromosomes, malgré une densité bien plus importante sur le chromosome 19 (de l'ordre de 1000 Alu par mégabase ; données non montrées). On peut remarquer que les sous-familles disponibles dans la bibliothèque RepBase ne correspondent pas forcément aux familles que nous avons données en introduction (section 1.3.3). Les familles concordantes sont marquées d'une étoile dans le tableau 4.4. Nous avons regroupé les diverses sous-familles en

familles génériques lors de nos analyses, afin de disposer de suffisamment de membres par famille. Ces familles génériques représentent chacune une phase d'amplification distincte, délimitée dans le temps. Les familles FLA, FLAM_A et FLAM_C sont regroupées sous l'appellation FAM (pour Fossil Alu Monomer), mais ne seront pas utilisées par la suite, car elles ne possèdent pas de linker. Les sous-familles de AluY non référencées ne seront pas non plus utilisées, car nous ne disposions pas de leur âge, même approximatif.

Densité des séquences Alu dans le génome humain

famille	densité (nb/Mb)	groupe	âge estimé
FLA	2.26	FAM	> 100 Ma
FLAM_A	4.99		
FLAM_C	1.69	non analysé	
* AluJo	75.11	J	60 - 80 Ma
* AluJb	52.88		
* AluSz	82.22	S	30 - 50 Ma
* AluSx	62.73		
* AluSq	36.41		
* AluSg	33.93		
* AluSc	17.96		
* AluSp	22.84		
* AluY	39.14	Y	10 - 25 Ma
* AluYe5	0.48		
* AluYd3	0.98		
* AluYi6	0.44	Y+	< 5 Ma
* AluYg6	0.33		
* AluYd8	0.12		
* AluYb8	1.32		
* AluYa5	1.18		
* AluYa8	0.17		
* AluYb9	0.17		
* AluYc1	2.05		
AluYd3a1	0.01	non analysé	
AluYc2	1.08		
AluYa1	2.68		
AluYa4	0.53		
AluYb3a1	3.27		
AluYb3a2	0.84		
AluYf1	0.92		
AluYh9	0.16		
AluYbc3a	0.56		
AluYe2	0.95		
AluYf2	3.42		
total	**453.82**		

TAB. 4.4 – Densité moyenne des séquences Alu selon leur famille/sous-famille, pour l'ensemble du génome humain. Les séquences ont été extraites avec RepeatMasker, en utilisant les consensus disponibles dans Repbase 9.11. Les familles concordantes avec celles présentées dans le tableau 1.1 ont été marquées d'une étoile. L'âge donné correspond à l'ordre de grandeur de l'âge de la famille. Les sous-familles pour lesquelles nous ne disposions pas d'estimation d'âge n'ont pas été analysées. Les FAM, ne possédant pas de linker, n'ont pas non plus été analysés.

4.2.2 Méthodes de calcul de la proximité

Distances entre microsatellites et éléments Alu

Tant TRF que RepeatMasker permettent d'obtenir la position génomique exacte de chacun des éléments détectés, la distance entre les différents éléments a donc été calculée à partir de leurs positions de début et de fin. Nous définirons pour la suite la distance Alu-microsatellite comme la distance entre le début de la séquence Alu et le début de la séquence microsatellite. Les microsatellites n'étant pas orientés sur la séquence génomique, leur début sera systématiquement l'extrémité la plus proche du début de l'élément Alu. Par contre, les séquences Alu étant orientées, deux cas de figures se sont présentés. Si l'élément Alu est orienté dans le sens de la séquence génomique, la distance est calculée entre la première base de Alu et la base de début du microsatellite. Si l'élément Alu est orienté dans le sens complémentaire, la position de début de Alu sera sa dernière base. Si un microsatellite est à proximité de deux Alu (de part et d'autre de ses extrémités), les deux associations seront comptées indépendamment. De même, les associations entre plusieurs microsatellites et un même élément Alu seront comptées indépendamment.

D'autre part, un certain nombre de Alu ont été tronqués, soit dès leur insertion, soit par des mutations postérieures. Pour ces derniers, la position de début (ou de fin) détectée ne correspond pas à la position théorique s'ils avaient été complets. Il a donc fallu corriger les distances pour prendre en compte ces Alu tronqués. Le calcul de la distance se fera donc à partir de la première position théorique de l'élément Alu, obtenue grâce à son consensus, comme expliqué dans la figure 4.1.

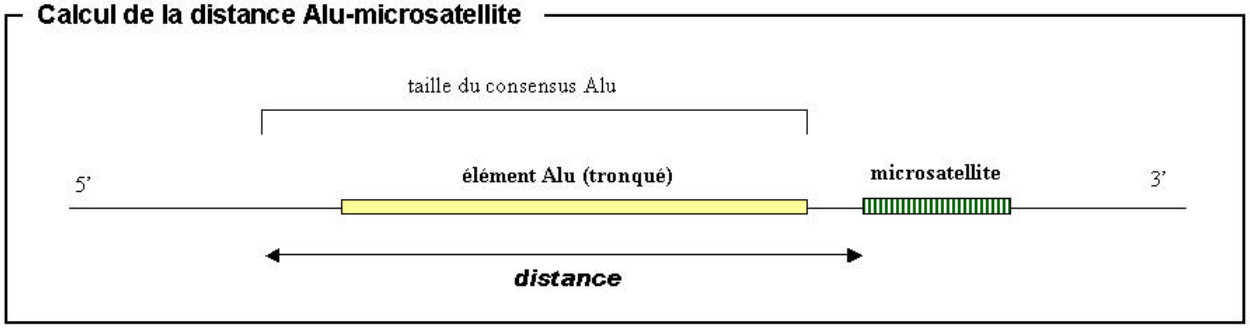


FIG. 4.1 – Calcul de la distance entre élément Alu et microsatellite. Le calcul se fait à partir de la première base de chaque élément. Le microsatellite n'étant pas orienté, la première base est celle qui se trouve la plus proche du début de la séquence Alu. La première base de Alu est calculée à partir de la taille de sa séquence consensus. La position théorique de ce consensus est obtenue à partir des données d'alignement fournies par RepeatMasker.

Analyse statistique

Le calcul des distances entre microsatellites et éléments Alu va nous permettre de savoir combien de microsatellites sont associés à une séquence Alu. Ces proportions ne signifient pas pour autant que les Alu soient vecteurs de microsatellites. En effet, ces derniers peuvent *a priori* apparaître à partir de n'importe quelle séquence (voir chapitre 5), y compris dans les séquences Alu. Etant donné le nombre important de ces éléments dans le génome humain, il est possible que les associations observées entre les deux éléments ne soient que le fruit d'apparitions aléatoires. Pour vérifier que les séquences Alu sont significativement vecteurs de microsatellites, il faut s'assurer que le nombre d'associations est supérieur à l'attendu aléatoire, atteint lorsque les microsatellites ont autant de chance d'apparaître proches d'un Alu qu'ailleurs.

Supposons une séquence génomique de taille S, contenant M locus microsatellites et A locus Alu. Selon l'équation proposée par Nadir *et al.* (1996), la probabilité de trouver au hasard un microsatellite dans les d bases proches de la fin d'un Alu est égale à :

$$P_{prox,d} = A \times \frac{d}{S} \tag{4.1}$$

Le nombre d'associations entre microsatellites et 3' de Alu attendues au hasard à au plus une distance d est donc $N_d = M \times P_{prox,d}$. La formule est exactement la même pour calculer le nombre de microsatellites attendus dans le linker sous l'hypothèse nulle, en prenant la taille du linker pour d ; et par extension pour n'importe quelle zone interne, avec d la taille de la zone.

Nous considérerons ici que les microsatellites et les éléments Alu sont à peu près distribués uniformément dans le génome humain [Lander et al., 2001]. Dans ce cas-là, les valeurs obtenues de manière théorique (N_d) suivent une loi binomiale, qui est la probabilité d'avoir un succès (*i.e.*, d'avoir un microsatellite dans un périmètre d) multiplié par le nombre d'essais (*i.e.*, le nombre de Alu). On peut donc effectuer un test binomial pour savoir si les valeurs observées suivent les valeurs attendues théoriques, autrement dit si elles sont conformes à l'hypothèse nulle d'association aléatoire entre les éléments. Pour cela, on calcule :

$$z = \frac{(N_o \pm 0.5 - N_e)}{\sqrt{MP_{prox,d}(1 - P_{prox,d})}} \tag{4.2}$$

avec N_o le nombre d'associations observées. Le ± 0.5 est une correction due à la non-continuité des tests binomiaux, alors que la significativité va se calculer à partir d'une loi Normale continue. La

valeur de z permet ensuite de déterminer la *p-value* à partir de la loi Normale, une *p-value* inférieure à 0.05 permettant de rejeter l'hypothèse nulle avec moins de 5% de risque. On considère dans ces cas-là que le nombre d'associations observées est significativement différent de celui attendu, donc que les distributions des Alu et des microsatellites ne sont pas indépendantes.

4.2.3 Proximité entre microsatellites et éléments Alu

Association entre microsatellites et éléments Alu

La figure 4.2 donne la distribution des distances entre les microsatellites et les séquences Alu. Elle montre très nettement une position préférentielle (280-281) correspondant au 3' de Alu, confirmant l'importance primordiale de la queue polyA dans les relations entre Alu et microsatellites. Les $(A)_n$ représentent les 3/4 de ces associations, mais même lorsqu'on les retire de l'analyse, la position en 3' de Alu reste proéminente. En comparaison, le linker est associé à près de 10 fois moins de microsatellites que la queue polyA (distance 117), les 3/4 de ces derniers étant là encore des $(A)_n$. Un certain nombre de microsatellites sont aussi à proximité directe (moins de 5 nt) du 5' des Alu.

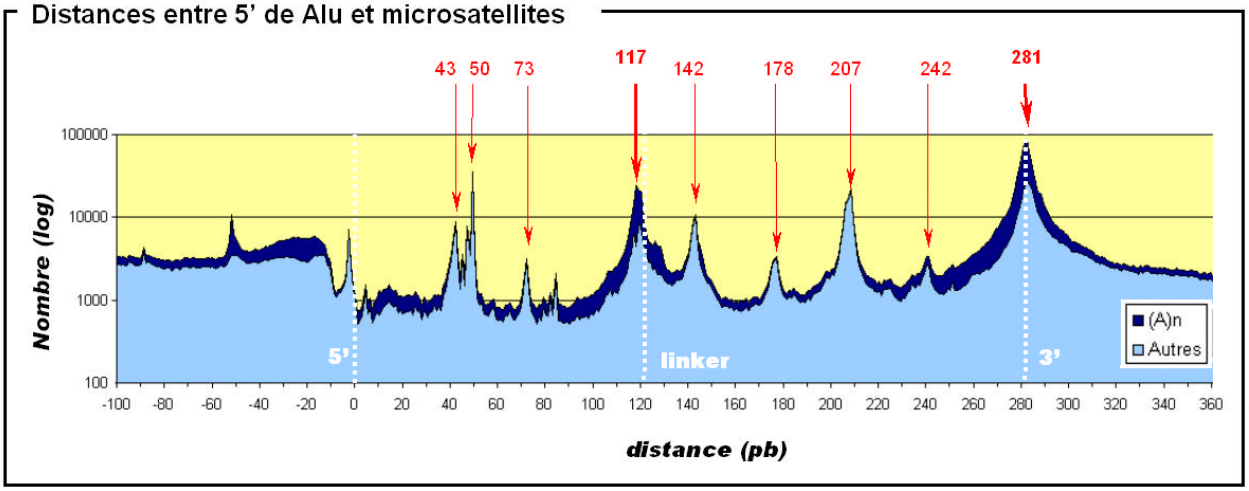


FIG. 4.2 – Distribution cumulative des distances entre $(A)_n$ et autres microsatellites, et séquences Alu pour l'ensemble du génome humain, toutes familles Alu confondues. Le nombre de détections par distance est présenté selon une échelle logarithmique. Les positions des éléments Alu particulièrement associées à des microsatellites sont indiquées d'une flèche rouge.

Nous avons ensuite évalué la proportion de ces microsatellites associés aux séquences Alu par rapport à la totalité. Nous avons compté le nombre de microsatellites commençant à moins de 10 nt du 3', du 5', ou dans le linker d'un élément Alu. Les microsatellites chevauchant les Alu sur 5 nt au maximum ont aussi été comptabilisés, pour prendre en compte ceux qui commencent ou finissent

dans la séquence Alu. De même, comme les Alu sont légèrement variables en taille (selon les familles), les microsatellites considérés dans le linker sont ceux compris entre les positions 110 et 130. Ces calculs ont été effectués pour diverses classes de microsatellites, et les résultats sont présentés dans le tableau 4.5.

Plus de 10% de l'ensemble des microsatellites sont à proximité directe du 3' d'un Alu, et presque 3% sont présents dans le linker, ces proportions étant hautement significatives. Cette association est en grande partie causée par les séquences $(A)_n$, dont environ 50% (soit plus de 500 000 locus) sont proches d'éléments Alu. De manière générale, tous les motifs riches en A (*i.e.*, qui possèdent strictement plus de A que d'autres bases) sont plus proches des 3' et linker de Alu qu'attendu aléatoirement. Les AAT et AAC sont particulièrement associés aux séquences Alu, tant dans la queue polyA (21,3 et 16,3% respectivement) que dans le linker (5 et 3,4%), alors que la proportion de GAA associés est beaucoup plus faible (2,7 et 0,46% dans le polyA et le linker respectivement). Les microsatellites pauvres en A présentent la tendance inverse, avec moins d'associations qu'attendu aléatoirement, tant en 3' que dans le linker. Les polyG/C représentent toutefois une exception avec presque 4,5% d'association avec le 3' ou le linker d'un élément Alu. Les dinucléotides AT, AC ou AG ne sont ni riches, ni pauvres en A, et sont positivement associés aux 3' et linker des séquences Alu dans une proportion de 5,15%, 2,23% et 4,04%, respectivement.

Enfin, les microsatellites associés aux 5' des Alu sont là aussi significativement plus nombreux que l'attendu aléatoire, même s'ils ne représentent que 1,12% de l'ensemble des microsatellites. Cette sur-représentation est causée uniquement par les motifs riches en A, plus les $(AT)_n$.

Présence de microsatellites à l'intérieur des Alu

Les résultats les plus inattendus que nous montre le graphique 4.2 sont toutefois les multiples pics à diverses positions internes des Alu (positions 43, 50, 73, 142, 178, 207 et 242). L'analyse des distances a donc été reconduite motif par motif pour déterminer si ces pics étaient produits par des motifs particuliers. Les résultats sont présentés dans la figure 4.3.1. La position 43 des séquences Alu est associée à trois motifs de manière équivalente ($(GAGGCX)_n$, avec X égal à C, G ou T), tandis que les autres positions sont préférentiellement associées à un seul motif.

La séquence consensus Alu a été analysée aux positions données pour déterminer si ces microsatellites sont présents par défaut dans les éléments Alu. Aucun microsatellite n'a été détecté, mais la séquence présente presque à chaque fois un quasi-microsatellite favorisant la création d'un mi-

Proportion de microsatellites liés à des éléments Alu

motif	nombre	proportion 3'		proportion linker		proportion 5'		proportion interne	
total	7416090	11.33	+++	2.85	+++	1.12	+++	-	
polyA/T	1129119	44.73	+++	10.23	+++	3.11	+++	-	
polyC/G	26920	2.81	+++	1.59	+++	0.31	---	-	
dinucléotides	539033	2.60	+++	1.50	+++	0.45	---	-	
AT	159096	3.30	+++	1.85	+++	0.80	+++	-	
AG/TC	137480	2.23	+++	0.79	n.s	0.41	---	-	
AC/TG	240013	2.36	+++	1.68	+++	0.24	---	-	
GC	2444	0.53	n.s	0.20	--	0	--	-	
trinucléotides	639988	5.69	+++	1.37	+++	0.76	+++	-	
AAT/ATT	93516	21.33	+++	4.92	+++	3.16	+++	-	
AAG/TTC	77723	2.17	+++	0.46	+++	0.82	+++	-	
AAC/TTG	86497	16.33	+++	3.36	+++	1.00	+++	-	
GGC/GCC	12271	0.05	---	0.06	---	0.03	---	-	
autres tri	639988	0.10	---	0.14	---	0.06	---	-	
tetranucléotides	1600349	8.19	+++	1.19	+++	0.77	+++	-	
tetra riches en A	681486	17.88	+++	2.39	+++	1.45	+++	-	
autres tetra	918863	1.00	+++	0.29	---	0.27	---	-	
pentanucléotides	1673577	4.41	+++	0.94	+++	0.75	+++	-	
penta riches en A	959062	7.60	+++	1.46	+++	1.15	+++	-	
autres penta	714515	0.13	---	0.24	---	0.22	---	-	
hexanucléotides	1807104	2.95	+++	1.73	+++	0.70	+++	-	
hexa riches en A	623877	8.16	+++	4.61	+++	1.55	+++	-	
autres hexa	1183227	0.21	---	0.21	---	0.25	---	-	
AGGC/TGCC	103983	0.02	---	0.10	---	0.09	---	47.80	+++
GTG/CAC	83032	0.26	---	0.74	n.s.	0.08	---	48.80	+++
GAGGCT/AGCCTC	25663	0.004	---	0.16	---	0.26	---	39.06	+++
GAGGTG/CACCTC	101673	0.02	---	0.18	---	0.01	---	89.25	+++
CACTG/CAGTG	21037	0.03	---	0.46	---	0.05	---	50.12	+++

TAB. 4.5 – Nombre total de microsatellites mono à hexanucléotides dans le génome humain, pour différentes classes de motifs, et la proportion qui est associée au 3', au linker, au 5', ou à une position interne d'un Alu (à lire par ligne). L'association est limitée aux microsatellites présents au maximum à 10 bases de l'extrémité du Alu, ou à 5 bases à l'intérieur, pour les régions 3' et 5'. Le linker est délimité par les positions 110 et 130 du Alu, et les diverses zones internes sont dépendantes du motif : zone 36-53 pour les AGGC, 124-155 pour les TGG, 167-183 pour les AGGCTG, 173-228 pour les AGGTGG et 229-248 pour les ACTGC. Les motifs riches en A sont ceux possédant strictement plus de A que de n'importe quelle autre base. La différence entre la proportion observée et celle attendue aléatoirement a été calculée à partir des formules 4.1 et 4.2, pour chaque type de motifs. $^{+++}$: proportion plus importante que l'attendu aléatoire avec $p\text{-}value < 0.00001$, $^{---}$: proportion moins importante avec $p\text{-}value < 0.00001$, $^{--}$: proportion moins importante avec $p\text{-}value < 0.001$, n.s. : pas de différence significative.

crosatellite (figure 4.3.2). L'apparition du motif répété ne nécessite en général qu'une substitution, sauf pour les $(AGGC)n$ en position 50 qui ont besoin de deux mutations. On peut remarquer que les associations les plus importantes sont créées par des mutations CpG (aux positions 50, 142, et 207). La position 73 est principalement associée aux microsatellites $(GAGGTG)n$, mais la séquence consensus à cette position est peu susceptible de produire ce type de motif. La suite des analyses présentées dans ce chapitre ne concerneront que les positions avec les cinq plus fortes associations microsatellite-Alu (positions 50, 142, 178, 207 et 242).

Ces locus présents dans les séquences Alu représentent en moyenne 50% du total de ces motifs dans le génome, à l'exception de $(GAGGTG)_n$, dont presque 90% sont associés aux séquences Alu (tableau 4.5) ; toutes les associations sont significatives.

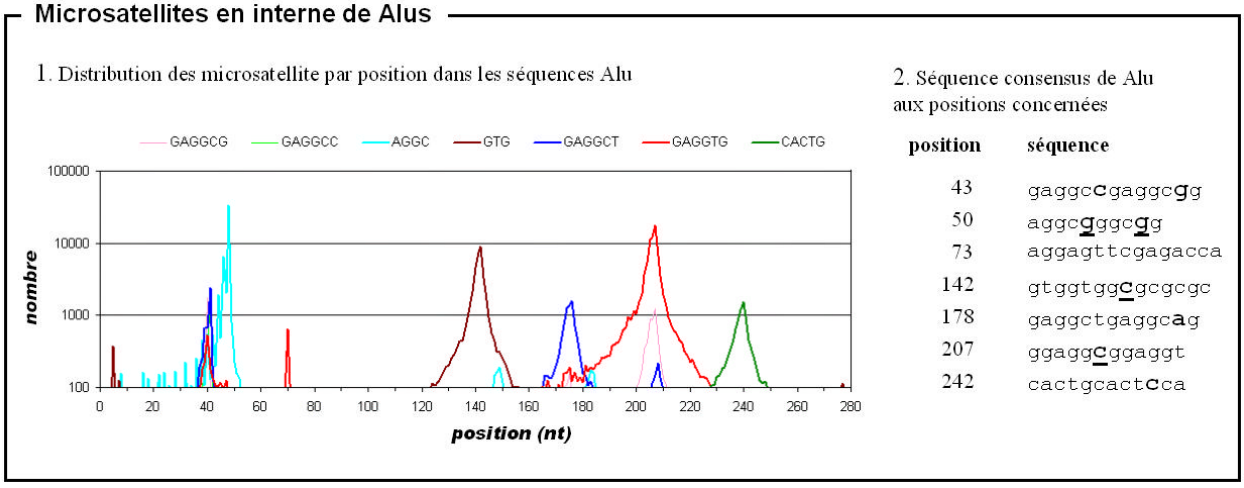


FIG. 4.3 – **1-** Distribution de certains microsatellites en fonction de leur position dans les séquences Alu. Le nombre de microsatellites est présenté sur une échelle logarithmique. **2-** Séquence consensus Alu pour les différentes positions d'association caractéristique avec des microsatellites. L'apparition des locus microsatellites est provoquée par la mutation aux bases représentées en gras, les mutations aux sites CpG étant de plus soulignées.

4.2.4 Influence de la famille Alu

Association des microsatellites selon la famille Alu

Les éléments Alu peuvent être répartis en familles et sous-familles, correspondant approximativement à des phases d'amplification génomique (*i.e.* d'intégration). Nous disposons d'un calibrage temporel pour ces phases d'amplification, qui permettent d'estimer un âge approximatif pour les familles Alu (voir tableau 4.4). Les AluJ sont les plus anciens, suivi des AluS, puis des AluY. Enfin, les

AluY+ représentent les sous-familles de AluY les plus récentes. Nous avons analysé les associations entre microsatellites et éléments Alu, famille par famille, pour les différentes positions d'association préférentielles déterminées précédemment. Les résultats sont présentés dans le tableau 4.6.

Les familles J, S, et Y présentent une association en 3' significative avec tous les microsatellites considérés, à l'exception des $(AG)_n$, pour qui elle n'est significative que pour les AluJ et AluS. En revanche, l'association avec les AluY+ n'est pas systématique. Elle ne concerne en effet que les microsatellites avec un motif riche en A, sauf les $(AAG)_n$. Les $(AT)_n$ sont eux-aussi associés aux 3' des AluY+, avec toutefois une valeur de significativité moins élevée.

L'association des microsatellites avec le linker des séquences Alu montre plus de disparités entre les types de microsatellites. Les $(A)n$ et les motifs riches en A sont globalement associés à toutes les familles, tandis que les non-riches en A ne sont associés significativement qu'aux éléments Alu les plus anciens (J et S). Deux exceptions sont toutefois à relever : les $(AC)_n$ sont très significativement associés aux AluY+, et les $(AAG)_n$ ne sont associés qu'aux AluJ.

L'analyse des associations avec la région 5' des éléments Alu montrent que les $(A)_n$, $(AAT)_n$ et les microsatellites riches en A (tétranucléotides et plus) sont significativement associés à tous les éléments Alu, quel que soit leur âge (tableau 4.6). Les motifs microsatellites non-riches en A ne présentent quant à eux aucune association significative, à l'exception des $(AT)_n$, avec la région 5' des AluS. Enfin, les $(AAG)_n$ sont faiblement associés aux trois familles J, S et Y, et les $(AAC)_n$ aux deux familles les plus anciennes (J et S).

Les mêmes analyses pour les motifs microsatellites présents aux positions internes des éléments Alu montrent une association significative quelle que soit la famille Alu (J, S, Y ou Y+ ; tableau 4.6.2). Le calcul a donc été réalisé pour toutes les sous-familles de chaque famille. Toutes les sous-familles de AluJ, AluS et AluY sont significativement associées à tous ces microsatellites internes. Les associations sont en revanche nettement plus hétérogènes entre les sous-familles de AluY+. Les éléments des sous-familles Yi6, Yg6, Yd8, Ya8 et Yc1 sont associés à quasiment tous les motifs, tandis que ceux des sous-familles Yb8/b9 et Ya5 ne sont associés qu'aux $(AGGC)_n$ (et $(GTG)_n$ pour les Ya5). Il est intéressant de remarquer que les sous-familles possédant très peu d'associations avec ces motifs internes sont celles connues pour être encore très actives actuellement [Carroll et al., 2001, Roy-Engel et al., 2001].

Association microsatellites-Alu par famille

1. Pour les microsatellites associés aux régions 3', linker, et 5' des éléments Alu

motif	nombre	3'				linker				5'			
		J	S	Y	Y+	J	S	Y	Y+	J	S	Y	Y+
polyA/T	1129119	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++
polyC/G	26920	+++	+++	+++	n.s.	+++	n.s.	-	n.s.	--	--	n.s.	n.s.
AT	159096	+++	+++	+++	+	+++	+++	-	n.s.	n.s.	+++	n.s.	n.s.
AG/TC	137480	+++	+++	n.s.	n.s.	+++	---	---	--	---	---	-	n.s.
AC/TG	240013	+++	+++	+++	n.s.	+++	+++	-	+++	---	---	---	--
AAT/ATT	93516	+++	+++	+++	+++	+++	+++	n.s.	++	+++	+++	+++	+++
AAG/TTC	77723	+++	+++	+++	n.s.	++	---	---	-	+	++	+	n.s.
AAC/TTG	86497	+++	+++	+++	+++	+++	+++	++	n.s.	++	+++	n.s.	n.s.
autres riches en A	2264425	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++

2. Pour les microsatellites associés à des positions internes

						Y+							
		J	S	Y	Y+	Yi6	Yg6	Yd8	Yb8	Ya5	Ya8	Yb9	Yc1
AGGC/TGCC	103983	+++	+++	+++	+++	+++	+++	++	+	+	+++	++	+++
GTG/CAC	83032	+++	+++	+++	+++	+++	n.s.	++	n.s.	++	+++	n.s.	-
GAGGCT/AGCCTC	25663	+++	+++	+++	+++	+++	++	+	n.s.	n.s.	+++	n.s.	+++
GAGGTG/CACCTC	101673	+++	+++	+++	+++	+++	+++	++	-	-	n.s.	n.s.	+++
CACTG/CAGTG	21037	+++	+++	+++	+++	+++	n.s.	+++	n.s.	n.s.	+++	n.s.	+++

TAB. 4.6 – Valeur de l'association entre microsatellites et les familles Alu (J, S, Y, Y+), en fonction de la position des microsatellites par rapport aux Alu. Seuls les motifs exhibant une association significative globale (voir tableau 4.5) ont été analysés. La détermination des familles et leurs fourchettes d'âge sont données dans le tableau 4.4. Significativité de la sur-représentation des associations microsatellite-Alu par rapport à l'attendu aléatoire : $^{+++}$: *p-value* < 0.00001, $^{++}$: < 0.001, $^{+}$: < 0.05. Significativité de la sous-représentation des associations microsatellite-Alu par rapport à l'attendu aléatoire : $^{---}$: *p-value* < 0.00001, $^{--}$: < 0.001, $^{-}$: < 0.05. n.s. : pas de différence significative avec l'attendu aléatoire. **1-** Association avec les régions 3', linker, et 5' des éléments Alu. **2-** Association avec les positions internes des éléments Alu, en fonction du motif. L'association avec les sous-familles de Y+ a été détaillée.

Proportion d'éléments Alu associés à des microsatellites

Nous nous sommes intéressés jusqu'alors à la proportion de microsatellites associés aux éléments Alu. Le Tableau 4.7 présente cette fois la proportion d'éléments Alu associés aux microsatellites, famille par famille.

La proportion d'éléments Alu possédant un microsatellite en région 3' est extrêmement élevée pour les familles jeunes (96,2% des AluY et 92,6% des AluY+), mais se réduit nettement avec leur âge (54,6% pour les AluJ). Ces associations élevées sont principalement causées par la queue polyA, puisque environ 75% des AluY+ en possèdent une. Leur présence devient moins importante avec

l'âge, et seuls 23% des AluJ possèdent encore une queue polyA. L'association avec les microsatellites riches en A (tétranucléotides et plus) est aussi assez importante, et croît avec l'âge, pour les éléments Alu plus récents que les AluJ. Les éléments de cette dernière famille montrent une proportion d'association plus faible que celle des AluS, pour ce type de microsatellites. La proportion d'association en 3' avec les autres types de microsatellites est nettement plus réduite, quelle que soit la famille Alu. Elle augmente toutefois là-aussi avec l'âge jusqu'aux AluS, et diminue pour les AluJ (sauf avec les microsatellites $(AG)_n$, $(AC)_n$, et $(AAG)_n$, pour lesquels elle augmente constamment).

La proportion d'éléments Alu associés à un microsatellite dans leur linker est aussi relativement élevée, et elle croît avec l'âge des éléments (de 25% des AluY+ à 50% des AluJ). Par contre, contrairement à la région 3', l'association dans le linker est majoritairement causée par des microsatellites riches en A, autres que polyA. Ainsi, 17% des AluJ possèdent un polyA dans leur linker, mais 30% possèdent un autre microsatellite riche en A, ce ratio de 1 pour 2 étant valable pour toutes les familles (sauf les AluS, avec un ratio de 1 pour 4). L'association avec les di et trinucléotides est en revanche très faible pour toutes les familles (inférieure à 1% des éléments Alu), même si elle augmente légèrement pour les AluJ.

Enfin, l'association avec des microsatellites en région 5' est similaire pour toutes les familles de Alu, et concerne entre 11 et 15% des éléments Alu. Les associations se font là encore principalement avec des microsatellites riches en A autre que trinucléotides, et dans une moindre mesure avec des polyA. Il n'y a pas de variation entre les familles, quel que soit le motif, à l'exception des AluJ, qui possèdent moins de polyA en 3' que les autres familles (1,9% des AluJ, et 3,2-4,5% des autres familles).

4.2.5 Taille des microsatellites associés aux séquences Alu

Microsatellites associés au 3' de séquence Alu

La distribution des tailles des polyA en 3' de Alu (figure 4.4) montre que la majorité des éléments Alu intégrés récemment (52% des AluY+) possède une queue polyA comprise entre 15 et 30 nucléotides, avec le maximum vers 22-26 nt. En comparaison, les $(A)_n$ non associés aux régions 3' des éléments Alu sont en très grande majorité de taille courte (10 nt, et des pointes à 14 et 17 nt). Lorsque l'on s'intéresse à la distribution pour les autres familles Alu, on distingue une dynamique d'évolution des polyA en deux phases. La première est la réduction brusque de la taille. En effet, la majorité des AluY conservent leur queue polyA (67% des éléments la possède encore ; voir tableau 4.7), mais ces dernières font en majorité entre 12 et 19 nt. La taille pour les polyA associés aux AluS

Proportion de Alu associés à des microsatellites

microsatellite	en région 3'				dans le linker				en région 5'			
	J	S	Y	Y+	J	S	Y	Y+	J	S	Y	Y+
polyA/T	22.87	48.18	66.67	74.52	17.06	5.44	10.90	8.85	1.88	3.24	4.49	3.42
polyC/G	0.04	0.08	0.07	0.01	0.07	0.02	0.01	0.02	0	0.01	0.01	0.01
AT	0.43	0.47	0.34	0.15	0.5	0.14	0.09	0.09	0.09	0.12	0.1	0.11
AG/TC	0.44	0.18	0.06	0.06	0.17	0.03	0.01	0.01	0.04	0.05	0.05	0.04
AC/TG	0.56	0.48	0.26	0.16	0.58	0.22	0.20	0.38	0.05	0.05	0.03	0.06
AAT/ATT	1.55	1.86	1.44	0.74	0.95	0.14	0.07	0.15	0.21	0.27	0.26	0.17
AAG/TTC	0.18	0.14	0.08	0.06	0.07	0.01	0	0.01	0.05	0.05	0.06	0.06
AAC/TTG	1.07	1.33	1.03	0.52	0.54	0.11	0.10	0.1	0.07	0.08	0.06	0.05
autres riches en A	27.49	30.61	26.27	16.37	30.35	23.28	20.46	15.41	8.69	10	10.1	9.37
total	**54.63**	**83.32**	**96.22**	**92.58**	**50.3**	**29.38**	**31.84**	**25**	**11.08**	**13.88**	**15.16**	**13.29**

TAB. 4.7 – Proportion d'éléments Alu possédant un microsatellite dans la région 3', linker, ou 5', en fonction du type de microsatellite, et de la famille de l'élément Alu. La détermination des familles et leurs fourchettes d'âge sont données dans le tableau 4.4.

et AluJ reste ensuite assez constante, avec un maximum vers 12-14 nt, tandis que la proportion d'association se réduit.

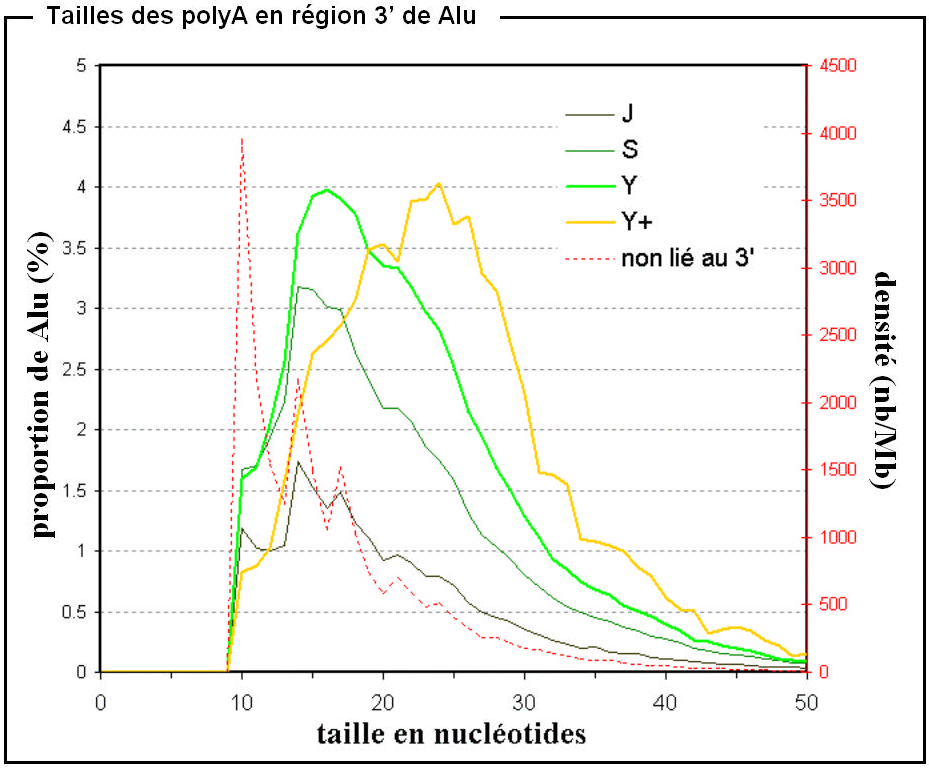


FIG. 4.4 – Proportion d'éléments Alu associés à des $(A)_n$ dans leur région 3', en fonction de la taille de ces derniers, et de la famille Alu. La densité des locus $(A)_n$ non liés aux régions 3' de Alu est donnée sur l'échelle de droite.

Les distributions en taille des microsatellites riches en A (autres que polyA) présents en 3' de Alu nous montrent une évolution un peu différente (Figure 4.5). Les microsatellites associés aux AluY+, AluY et AluS ont une distribution en taille à peu près équivalente, pour les trois familles. Leur taille est plutôt longue (entre 15 et 25-30 nt) par rapport à celle des microsatellites de même type présents hors des régions 3' de Alu (qui font en majorité 10 nt, ou 12 pour les hexanucléotides). Ces observations sont valables pour tous les types de motifs riches en A (tri à hexanucléotides). La proportion d'éléments Alu associés à ces microsatellites augmente par contre avec leur âge, comme cela avait été remarqué à la section précédente. Les microsatellites associés aux AluJ sont quant à eux plus courts que ceux associés aux autres familles, pour tous les types de microsatellites, et tendent vers la distribution donnée pour des locus hors Alu. Cette réduction de taille est donc synchronisée avec la réduction d'association observée dans le tableau 4.7 pour cette famille Alu.

Nous n'avons pas étudié la distribution en taille des dinucléotides associés aux AluY et AluY+, car ils étaient trop peu nombreux. Les distributions pour les AluS et AluJ sont relativement proches de celles des dinucléotides externes aux régions 3' de Alu (4.6), quel que soit le motif. Le pic à 10 nt, qui représente la majeure partie des locus hors Alu, est toutefois absent de ces distributions.

Une analyse complémentaire a été effectuée sur les polyC/G présents en région 3' de Alu. En effet, le tableau 4.6 donnait une sur-représentation de ces microsatellites à cette position, alors que le polyA ne devrait *a priori* pas être vecteur de ce genre de motifs. La figure 4.6 indique que les locus sont en majorité très courts (entre 10 et 17 nt), et sont moins présents dans les AluJ. On peut observer, dans la distribution des polyC/G non associés aux 3' de Alu, que plus de 25% de ces derniers ont une taille de 22 nt, alors que les plus courts sont en proportion moins importante. Il est peu envisageable que les polyC/G apparaissent spontanément à cette taille, ces locus pourraient donc être associés à un autre type d'éléments répétés du génome. Si l'on retire ces microsatellites de 22 nt de la distribution, les tailles des polyC/G associés aux 3' de Alu deviennent similaires à celles des polyC/G hors des éléments Alu.

Taille des microsatellites associés aux autres régions de séquences Alu

Nous avons ensuite analysé la distribution des tailles des microsatellites présents dans le linker des séquences Alu pour les différentes familles. Les résultats pour les $(A)_n$ montrent là encore une évolution en deux étapes (figure 4.7). Pour les familles récentes (*i.e.* AluY et AluY+), les polyA font en moyenne 17 nt, et sont présents pour 8 à 10% des Alu. Le taux de présence se réduit alors à 5,4%

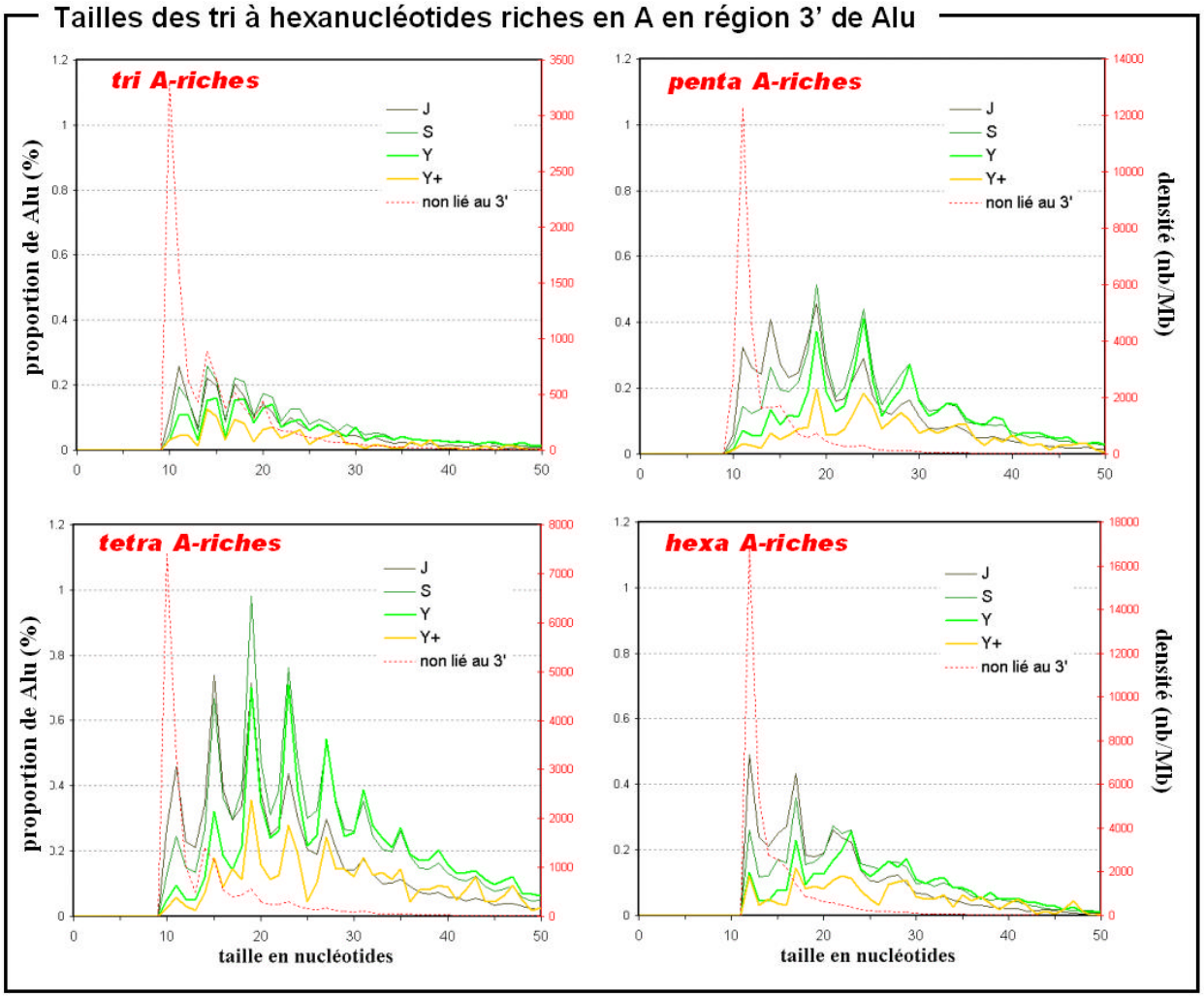


FIG. 4.5 – Proportion d'éléments Alu associés à des tri, tétra, penta et hexanucléotides riches en A dans leur région 3', en fonction de la taille de ces derniers, et de la famille Alu. La densité des locus non liés aux régions 3' de Alu, pour chacun des motifs, est donnée sur l'échelle de droite. Les pics observés dans chacune des distributions correspondent à la période des motifs concernés.

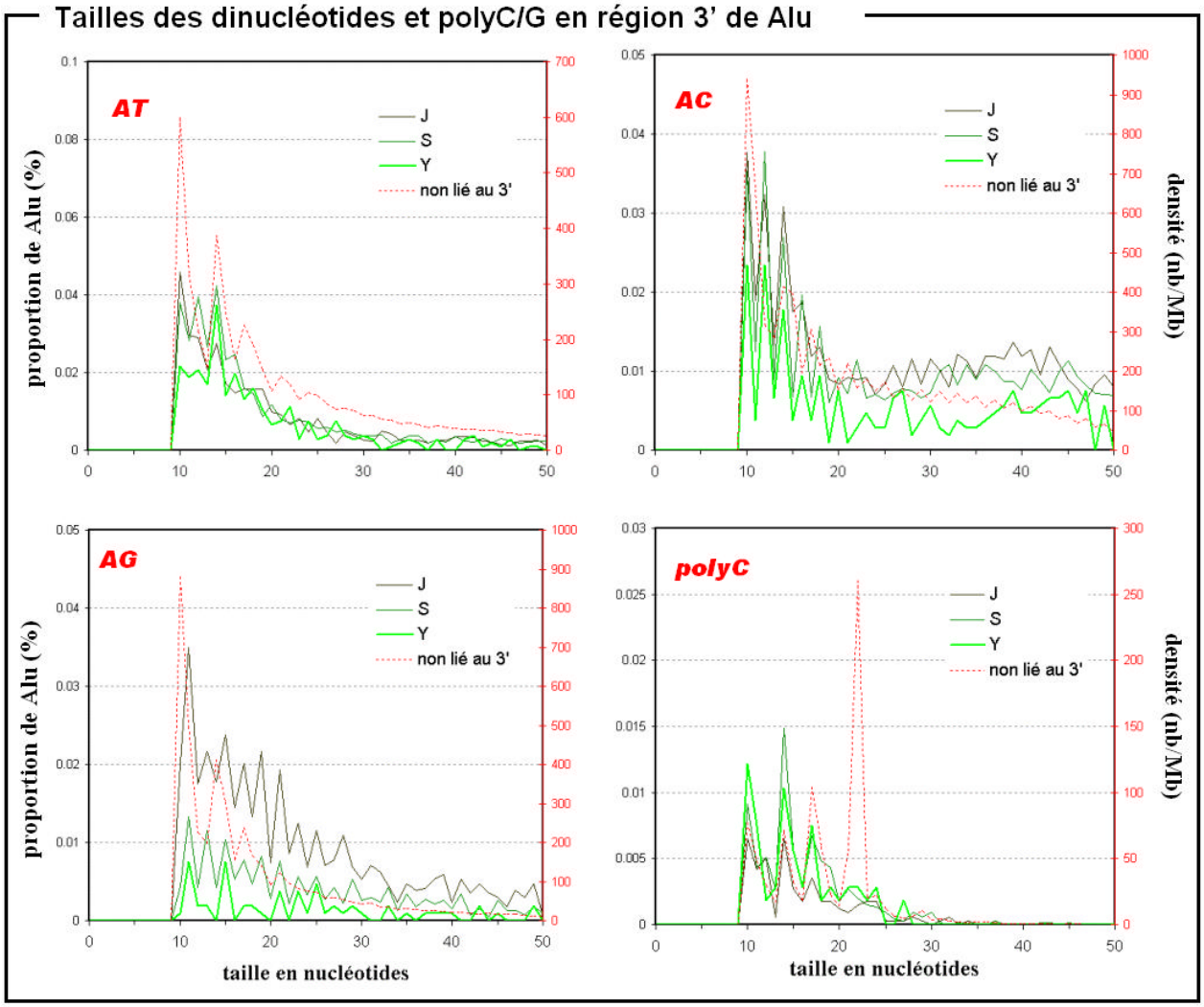


FIG. 4.6 – Proportion d'éléments Alu associés à des dinucléotides ou à des polyC/G dans leur région 3', en fonction de la taille de ces derniers, et de la famille Alu. La densité des locus non liés aux régions 3' de Alu, pour chacun des motifs, est donnée sur l'échelle de droite.

des AluS, sans que la taille ne soit affectée. Enfin, ils sont de nouveau plus présents dans le linker des AluJ (17% des éléments), mais avec une taille réduite (entre 10 et 14 nt pour la plupart). La distribution des di et trinucléotides est par contre similaire à celle des locus non associés au linker, quelle que soit la famille Alu (figure 4.7). Les distributions des autres microsatellites riches en A associés au linker sont elles aussi conformes à la distribution hors du linker, avec toutefois un pic à 16 nt pour les microsatellites associés aux AluY+ et surtout aux AluY, qui pourrait correspondre à une taille d'apparition privilégiée.

Les distributions des tailles pour les microsatellites présents en région 5' de Alu sont présentées dans la figure 4.8. Il apparaît que quels que soient le motif ($(A)_n$, $(AAT)_n$, ou autre riche en A) et la famille Alu, les tailles sont similaires à celles des microsatellites non associés aux 5' des éléments Alu. Les proportions de Alu associés au 5' sont similaires pour toutes les familles, sauf pour les AluJ qui possèdent moins de $(A)_n$. Ces distributions reflètent les proportions d'associations données dans le tableau 4.7.

Enfin, nous avons analysé les tailles des microsatellites trouvés aux positions internes des éléments Alu. Les résultats présentés en figure 4.9 indiquent que la taille des locus ne varie pas en fonction de l'âge des Alu auxquels ils sont associés, presque tous ayant gardé la taille à laquelle ils sont apparus (entre 10 et 12 nt). Les GAGGCT et les CACTG possèdent une seconde taille caractéristique à 16 et 14 nt respectivement, pour toutes les familles Alu, mais elles ne sont pas compatibles avec des événements de glissement. En revanche, la proportion d'éléments Alu possédant ces microsatellites internes varie fortement en fonction de la famille de Alu. La proportion de AluY+ possédant un microsatellite est systématiquement plus faible que celle des AluY et AluS, tandis que celle des AluJ est de nouveau plus réduite par rapport à ces deux familles, pour tous les motifs. Les distributions de taille pour les microsatellites présents dans le génome en général montrent que ces motifs, même s'ils sont en très grande majorité de taille très réduite (10 à 12 nt), sont capables de se développer, puisque l'on observe des locus de taille plus importante.

4.3 Discussion

4.3.1 Ré-évaluation de l'association entre microsatellites et séquences Alu

L'association entre microsatellites et rétrotransposons Alu a déjà été caractérisée à plusieurs reprises, mais les études étaient généralement basées sur un nombre limité de séquences génomiques ou de motifs microsatellites [Jurka and Pethiyagoda, 1995, Nadir et al., 1996, Clark et al., 2004].

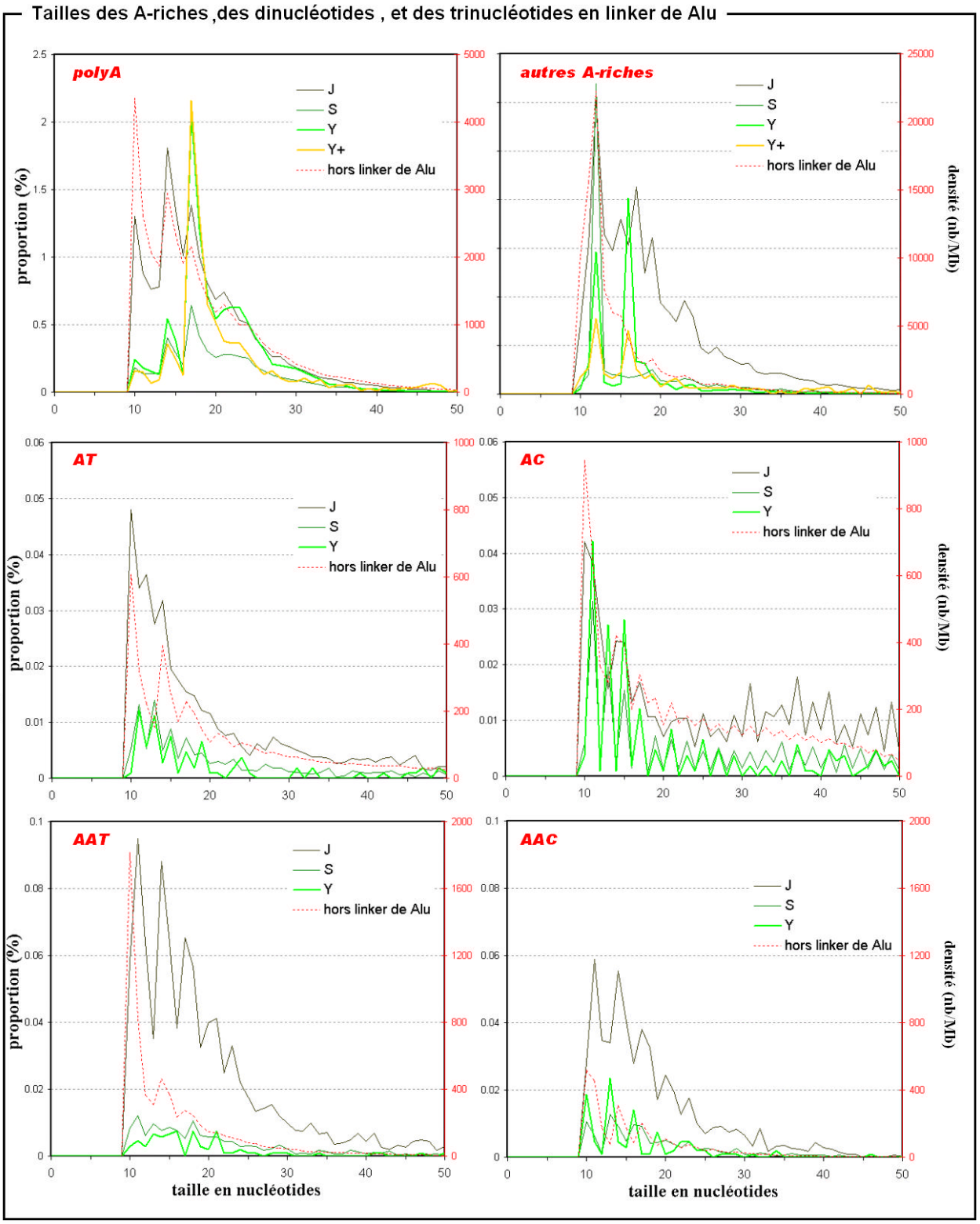


FIG. 4.7 – Proportion d'éléments Alu associés à des microsatellites dans leur linker, en fonction de la taille de ces derniers, et de la famille Alu. La densité des locus non liés aux linkers de Alu, pour chacun des motifs, est donnée sur l'échelle de droite.

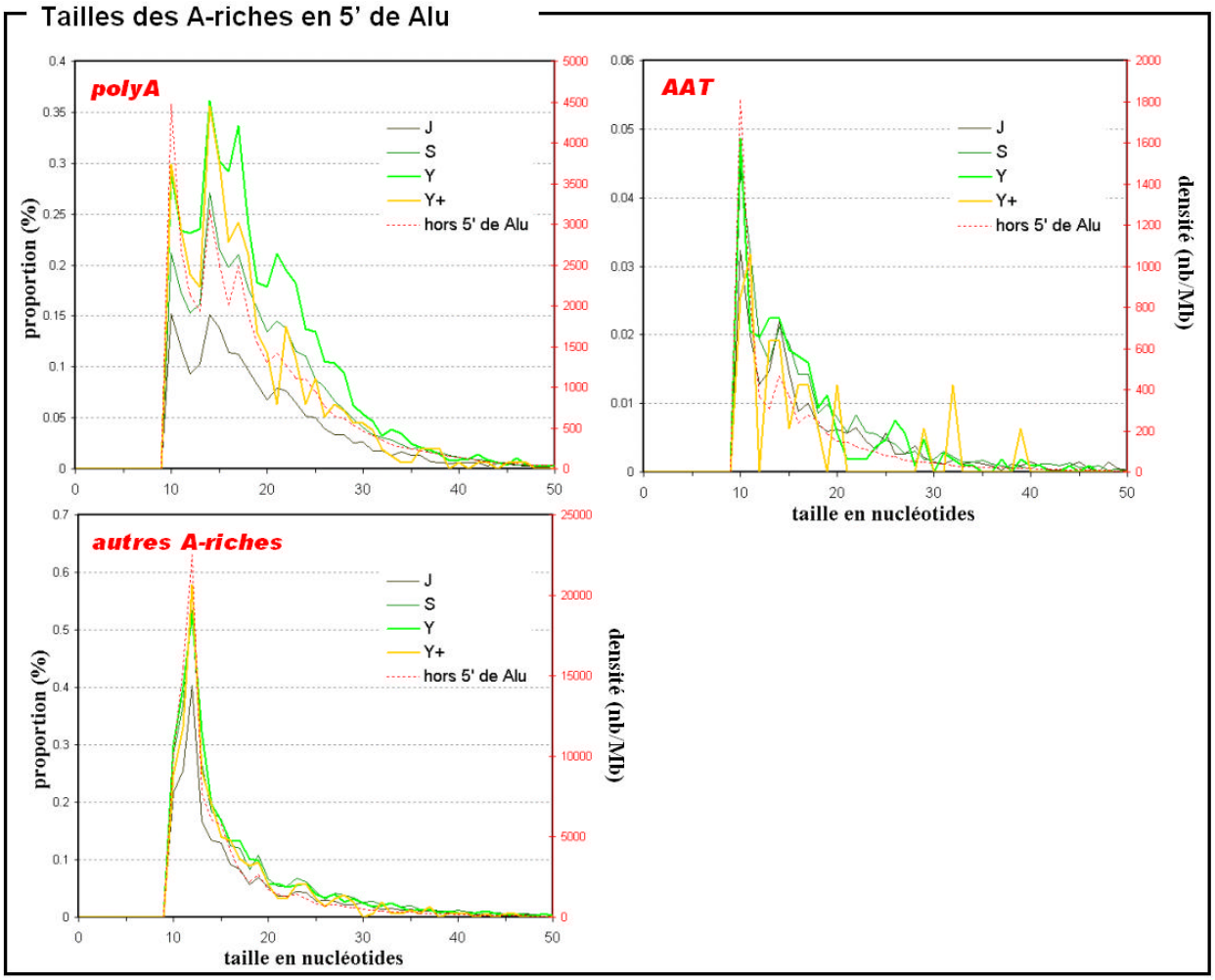


FIG. 4.8 – Proportion d'éléments Alu associés à des microsatellites dans leur région 5', en fonction de la taille de ces derniers, et de la famille Alu. La densité des locus non liés aux linkers de Alu, pour chacun des motifs, est donnée sur l'échelle de droite.

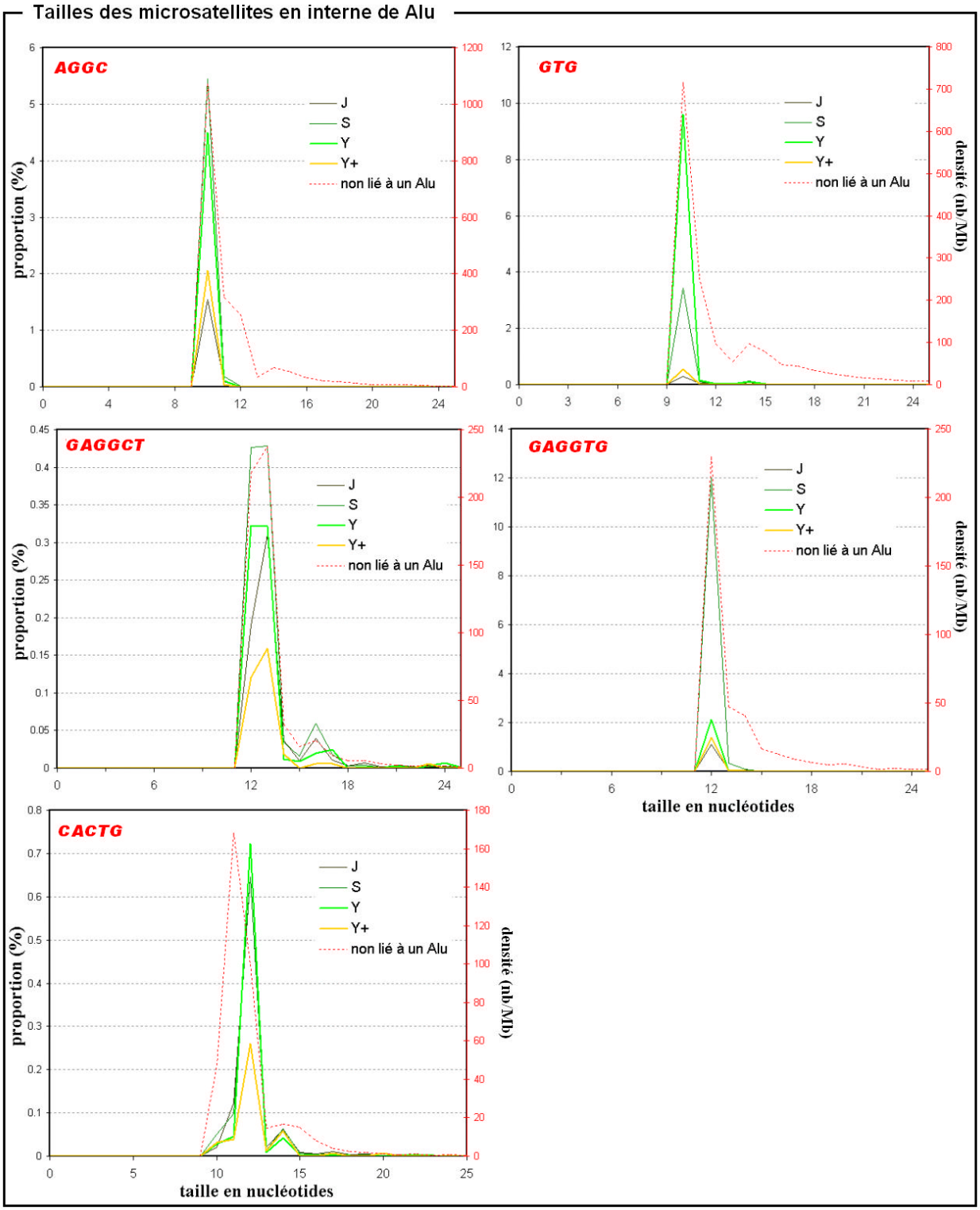


FIG. 4.9 – Proportion d'éléments Alu associés à des microsatellites en interne, en fonction de la taille de ces derniers, et de la famille Alu. Les motifs analysés et leur position sont définis à partir de la figure 4.3.1. La catégorie non lié à un Alu contient tous les locus microsatellites hors des positions définies, et est représentée par leur densité dans le génome (nombre par mégabase).

Nous avons ici étendu les analyses à l'ensemble du génome humain et des microsatellites de période 1 à 6 nt. Nos résultats confirment l'association entre ces deux types d'éléments génomiques, plus particulièrement avec la queue polyA et le linker des séquences Alu.

Nos observations diffèrent toutefois des estimations précédentes, par des proportions d'association nettement moins importantes. Par exemple, entre 75 et 85% des $(A)_n$ étaient estimés être associés aux Alu, alors que nous n'en trouvons que 50%. De même, les $(AAT)_n$ et $(AAC)_n$ qui ont pourtant une proportion d'association particulièrement élevée dans notre étude, à hauteur de 21,33 et 16,33% respectivement, étaient estimés être associés aux séquences Alu entre 60 et 100% du temps. Cette sur-estimation de l'association dans les études précédentes est valable pour tous les motifs, et est causée par la méthode de détection des microsatellites. En effet, Jurka et Pethiyagoda (1995) ont utilisé une limite inférieure de détection à 12 nt et 3 répétitions, et Nadir *et al.* (1996) à 16 nt. Or, nous avons remarqué dans le chapitre sur les algorithmes de détection (chapitre 3) que l'essentiel des microsatellites ont une taille inférieure ou égale à 12 nt, signifiant que les deux études suscitées n'ont pas pris en compte la majeure partie des locus présents dans leur séquence génomique. De plus, les figures 4.4 et 4.5 indiquent que les microsatellites riches en A associés aux 3' des éléments Alu sont plus longs que ceux du reste du génome. Jurka et Pethiyagoda (1995) et Nadir *et al.* (1996) ont donc sous-estimé la fraction de microsatellites non associés aux éléments Alu.

Le problème de détection joue aussi sur le rapport entre les microsatellites associés aux régions 3' et ceux associés aux linker de Alu. De manière générale, nos résultats sont en accord avec ceux précédemment obtenus, sauf pour les $(A)_n$ et les penta-hexanucléotides riches en A. Respectivement 16 et 50% des penta et des hexanucléotides associés aux Alu sont dans le linker, alors que ces taux n'étaient précédemment estimées qu'au maximum à 7 et 20%. La cause de cette sous-estimation est là-encore la non-détection des microsatellites très courts (2 répétitions, soit 10-12 nt) qui peuvent apparaître dans le linker (voir figure 4.7). De même, près de 20% des polyA associés aux Alu se trouvent dans le linker, alors que ce rapport était estimé à 10% chez Nadir *et al.* (1996). Le linker est constitué de la séquence $(A)_{5-6}TAC(A)_{5-6}$, qui ne peut être détectée telle quelle qu'en tant que polyA imparfait. Nadir *et al.* (1996) n'ayant extrait que les microsatellites parfaits, ils n'ont pu détecter que les linkers ayant subi une expansion d'au moins 6-7 répétitions (pour arriver à 12 nt) dans au moins l'une des deux sous-parties. En revanche, TRF peut détecter les microsatellites imparfaits,et c'est pourquoi nous avons obtenu beaucoup plus de locus à cette position.

Cette propriété de TRF nous amène par contre à nous demander pourquoi nous n'avons pas détecté de polyA dans le linker de chaque Alu (cela n'arrive que pour 10% d'entre eux environ). De

plus, les AluJ (les plus anciens) sont ceux qui possèdent le plus de polyA dans leur linker, avec 17% des éléments de cette famille. Il est donc peu probable que leur absence dans le linker soit causée par leur dégradation. La véritable explication est tout autre, et tient dans les paramètres de TRF. Nous avons fixé les pénalités d'alignement telles qu'une substitution baisse le score de 5 points, et qu'une base correcte l'augmente de 2 points. Le linker possédant deux substitutions par rapport à un polyA parfait, son score d'alignement est au maximum de 16 points. Le score minimum de détection étant fixé à 20 points, les linkers ne peuvent *a priori* pas être détectés par TRF avec nos paramètres. Pour atteindre 20 points et être détectés, les locus ont besoin de deux répétitions correctes supplémentaires, c'est-à-dire avoir une taille minimum de 17 nt, comme le montre justement la figure 4.7. Nos résultats sont par conséquent aussi certainement en dessous de la réalité concernant la proportion de polyA associés aux linkers de Alu, puisque l'on ne détecte que ceux qui ont déjà subi des expansions. Cela signifie par ailleurs que la proportion de polyA plus courts (10-15 nt) détectés pour les AluJ ne peuvent être issus de la réduction de ceux de taille 17 nt. Ils seraient donc plutôt apparus par l'expansion de l'une des deux sous parties $(A)_{5-6}$.

Les artefacts causés par la méthode de détection n'expliquent pourtant pas toutes les distributions que nous observons. La distribution en taille des microsatellites riches en A associés au linker des éléments Alu possède un pic très prononcé pour la taille 16 nt pour les AluY. L'analyse des séquences de ces microsatellites a montré que ce pic est représenté à 90% par la séquence AAATACAAAAACAAAA, détectée en tant que $(AAAAAC)_{2,66}$. Cette séquence particulière est issue de la contraction de la partie gauche du linker (deux A), de l'extension de la partie droite (quatre A), et de la substitution de A vers C à la 12e position. Aboutir à cette séquence à partir d'un linker sain est extrêmement complexe et a peu de chance d'arriver deux fois indépendamment. Deux possibilités sont alors envisageables : soit tous les AluY associés à ce microsatellite sont issus d'un même gène maître qui possédait déjà cette séquence, soit des événements de conversion génique (causés par de la recombinaison ectopique ; voir section 1.2.3) ont dupliqué les mutations dans des éléments déjà intégrés. La conversion génique entre éléments transposables ne convertit généralement pas la séquence en son entier [Roy et al., 2000], et une partie de l'élément original devrait être encore détectable. Or, très peu de AluS et AluJ ne possèdent de $(AAAAAC)_{2,66}$ à cette position, malgré leur prédominance dans le génome. En revanche, les AluY sont très nombreux à le posséder, et il est peu probable que les conversions géniques n'aient concerné que cette famille. L'hypothèse de la présence du microsatellite présent dans le gène maître nous semble donc plus réaliste. Cette association pourrait *a priori* permettre de caractériser une nouvelle sous-famille de AluY, le linker modifié étant la mutation diagnostique. Cette possibilité ouvre de larges perspectives pour la compréhension de

la dynamique des microsatellites hexanucléotides. En effet, le devenir de chacun des microsatellites insérés devrait pouvoir être déterminé, grâce à l'étude exhaustive des éléments de cette sous-famille.

4.3.2 Réduction de la taille des microsatellites

Une grande partie des microsatellites riches en A est donc associée aux séquences Alu, via la queue polyA ou le linker. L'analyse des associations en fonction de la famille des éléments Alu, complétée par l'analyse de la distribution en taille des microsatellites associés, nous a donné de précieux indices sur leur dynamique.

De manière générale, plus les éléments Alu sont anciens, plus leur association avec des microsatellites autres que polyA est importante. L'association est significative dès les AluY pour les microsatellites associés à la région 3'. Pour les éléments Alu les plus récents ($<$ 5 ma, les AluY+), l'association est significative pour tous les microsatellites riches en A à l'exception des $(AAG)_n$, mais pas pour les dinucléotides ni les polyC/G. Les polyA sont quant à eux toujours très significativement associés à la région 3' des éléments Alu, mais leur proportion baisse de manière importante avec l'âge de ces derniers. Nous avons par ailleurs montré qu'ils subissaient aussi une forte réduction en taille avec l'âge.

Cette réduction drastique de la taille des polyA avait déjà été observée par Roy-Engel *et al.* (2002), et Chauhan *et al.* (2002), ces derniers ayant proposé que cette réduction soit causée par l'apparition et l'expansion de microsatellites $(GAA)_n$ en région 3'. Etant donné le nombre de polyA par rapport à celui des $(GAA)_n$, la seule influence de ces derniers ne peut expliquer la réduction en taille, et il est nécessaire d'étendre l'hypothèse à l'ensemble des motifs riches en A. Plusieurs de nos résultats semblent toutefois en désaccord avec cette hypothèse. En effet, nous avons constaté que la réduction en taille des polyA est la plus forte entre les AluY+ et les AluY, alors que les microsatellites riches en A associés aux régions 3' sont déjà présents dans les Alu les plus récents, avec des tailles importantes. Certes, la proportion d'éléments Alu possédant un microsatellite riche en A en 3' augmente entre la famille Y+ et la famille Y, mais nous devrions aussi observer une augmentation de la taille de ces microsatellites associés. Il est donc peu probable que l'apparition des microsatellites dans les queues polyA soit la cause principale de leur réduction en taille. Ces résultats soutiennent en revanche l'hypothèse d'un biais de contraction des microsatellites les plus longs, comme proposé dans le modèle du cycle de vie des microsatellites (voir section 2.4.2).

Les résultats présentés ici ont par ailleurs permis d'obtenir une évaluation, certes très grossière,

de la rapidité à laquelle cette inéluctable réduction de taille peut avoir lieu. Pour les polyA, cette réduction est très rapide, comme nous l'avons déjà expliqué. Globalement, ils perdent en moyenne entre 10 et 15 nucléotides dans les 10 Ma suivant l'intégration de leur Alu associé. Cette réduction ne peut être causée par des interruptions, car l'algorithme utilisé pour la détection prend en charge les imperfections. L'effet des mutations se fait par contre sentir dans les 50 Ma qui suivent, avec une disparition progressive des polyA pour les AluS et AluJ, sans que leur taille ne soit affectée. La réduction de taille peut aussi être évaluée pour les autres microsatellites riches en A présents en 3' d'éléments Alu. Contrairement aux polyA, ces microsatellites ne semblent se contracter que pour les Alu les plus anciens, âgés de 60 à 80 Ma. Toutefois, de nouveaux locus apparaissent au cours du temps (la proportion d'association augmente avec l'âge), il est donc possible que les locus apparus en premier aient subi une réduction de taille assez tôt, mais que l'effet ne soit pas visible dans les distributions à cause des nouveaux locus, qui eux sont de taille importante. Le fait que la proportion de Alu associés à un microsatellite non polyA en 3' soit plus faible pour les AluJ que pour les AluS peut expliquer pourquoi la réduction de taille est visible pour les AluJ.

Ces interprétations supposent toutefois que les séquences Alu aient été intégrées avec un polyA d'approximativement la même taille, quelle que soit leur famille. Cette hypothèse semble néanmoins raisonnable si l'on suppose que le modèle d'intégration des éléments Alu proposé actuellement est correct (voir section 1.3.2), même si l'on ne connaît pas encore avec certitude cette taille d'intégration.

4.3.3 Apparition des microsatellites à partir du polyA

L'une des questions principales que soulèvent nos travaux est la manière dont apparaissent les microsatellites associés aux Alu. En effet, nous avons observé que les microsatellites riches en A situés en 3' des Alu les plus récents, même s'ils sont assez peu présents, sont déjà de taille importante, contredisant le modèle d'apparition par mutation ponctuelle, suivie d'une expansion. Dans leur article de 1995, Arcot *et al.* ont émis trois hypothèses pour expliquer l'association entre les microsatellites et la région 3' des éléments Alu (figure 4.10).

La première des hypothèses, qui est aussi évoquée par Nadir *et al.* (1996), propose que l'intégration des séquences Alu se fasse préférentiellement à l'intérieur d'un microsatellite existant. Comme nous l'avons expliqué, l'intégration produit des répétitions directes, dont l'une est adjacente au 3' du polyA du Alu. Si cette répétition directe est un microsatellite, cela créé de fait une association avec le Alu. Cette hypothèse est soutenue par le fait que les sites d'intégration sont probablement de type TTAAAA [Jurka, 1997], qui, sans être en soi des microsatellites, ont de fortes chances d'être

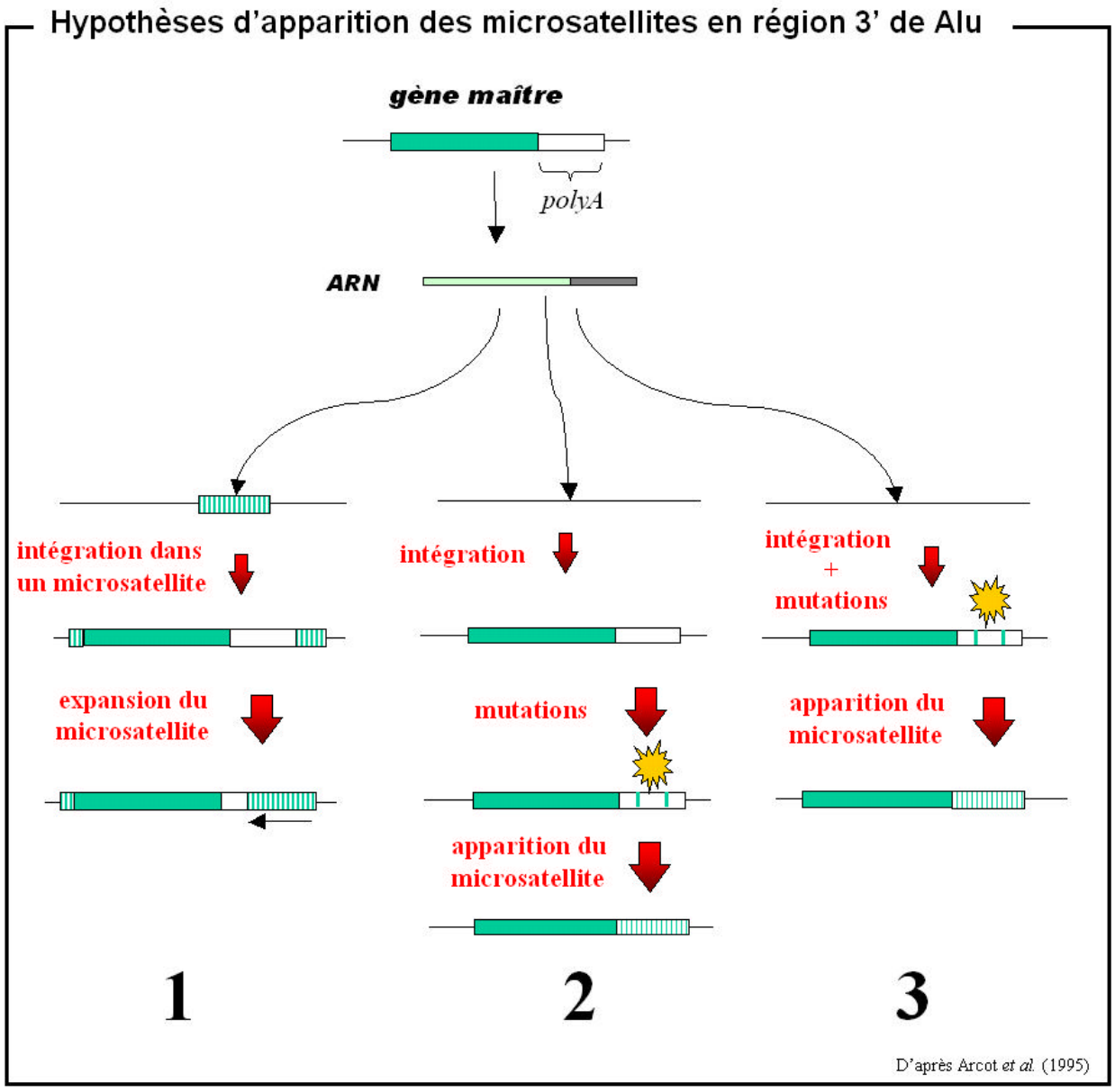


FIG. 4.10 – Les différentes hypothèses proposées par Arcot *et al.* (1995) pour expliquer l'association entre les microsatellites et la région 3' des éléments Alu. Dans l'hypothèse 1, la copie de l'élément Alu est intégrée à l'intérieur d'un microsatellite. Ce dernier se développe ensuite dans la queue polyA. Dans l'hypothèse 2, l'élément Alu est intégré avec sa queue polyA saine, puis celle-ci accumule des mutations jusqu'à l'apparition d'un microsatellite. Dans l'hypothèse 3, la queue polyA subit des mutations ponctuelles avant l'intégration de l'élément Alu, favorisant l'apparition d'un microsatellite.

présents dans des séquences de type $(A_{3-5}T)_n$. Nous avons par ailleurs observé une association significative entre les régions 5' de Alu et certains microsatellites riches en A (et notamment les $(AT)_n$ et $(AAT)_n$), même pour les intégrations les plus récentes. Les distributions en taille de ces

microsatellites étant de plus conformes à celles de leurs homologues non associés aux Alu, ils étaient probablement déjà présents dans le génome avant l'intégration des Alu. L'association observée résulte donc d'une intégration ciblée des Alu, qui a pu promouvoir l'apparition des microsatellites en 3' de Alu via la répétition directe. Les proportions d'association en 5' sont toutefois relativement faibles par rapport à celles en 3', et l'intégration ne semble ciblée que sur les microsatellites à base de A et T. La première hypothèse de Arcot *et al.* (1995) ne peut donc expliquer qu'une fraction minime de l'ensemble des apparitions observées dans le polyA des Alu.

La seconde hypothèse suppose que la queue polyA accumule des mutations après l'intégration de Alu, qui peuvent de temps en temps aboutir à un proto-microsatellite capable de se développer par la suite. La troisième hypothèse suppose que les mutations apparaissent, non pas après l'intégration mais durant celle-ci, à cause de la faible qualité de la polymérase chargée de la synthèse du brin complémentaire. Les mutations produites serviront ensuite de support à l'apparition d'un microsatellite comme dans la seconde hypothèse. Il est très difficile de distinguer les effets des deux hypothèses, sinon que la seconde permet de produire des microsatellites plus rapidement que la première. Elles impliquent toutes deux en revanche une augmentation progressive du nombre de microsatellites en 3' de Alu au cours du temps, et de plus en plus développés.

Nos résultats montrent en effet une telle augmentation, mais les microsatellites semblent développés quel que soit l'âge des Alu. Nous ne connaissons pas la vitesse à laquelle un microsatellite peut se développer à partir d'un proto-microsatellite, et il est donc possible que les locus observés en 3' de AluY+ soient issus de l'expansion de microsatellites apparus avec des mutations contractées durant l'intégration (hypothèse 3). Les microsatellites associés aux AluY et autres Alu plus anciens seraient apparus plutôt à partir de mutations contractées après l'intégration (hypothèse 2).

La rapidité de production des microsatellites dépendrait par ailleurs de leur motif. Nous avons remarqué que les dinucléotides ne présentent pas d'association particulière avec les AluY+. Les motifs dinucléotides n'étant pas riches en A, il est logique qu'ils soient plus difficiles à créer à partir d'une séquence polyA. Ils sont par contre significativement associés aux AluY, signifiant qu'ils sont apparus dans les 10-20 derniers millions d'années, ce qui est assez rapide. Cette production de nouveaux locus en région 3' de Alu serait ensuite continue, réduisant par la même occasion la proportion de polyA associés à cette région des Alu, jusqu'à ce que la majeure partie de ces polyA aient été dégradés. Nos distributions montrent que cet effet d'épuisement des « ressources polyA » pourrait être atteint pour les AluJ, pour lesquels la proportion d'association avec des microsatellites (donc le nombre qui sont apparus) a diminué. Il semblerait donc que les différents mécanismes proposés

par Arcot *et al.* (1995) aient tous une influence sur l'association entre microsatellites et éléments Alu.

Nous avons caractérisé la réduction de la taille des microsatellites dans la section précédente, et leur apparition dans cette section. A partir des différentes hypothèses développées, et de l'âge des familles Alu, il est possible de proposer un modèle de cycle de vie des microsatellites situés dans la région polyA des séquences Alu (4.11).

Tout d'abord, une partie des polyA se transforment en microsatellites riches en A, à partir de mutations ponctuelles contractées durant l'intégration. Ces microsatellites sont assez peu nombreux et se développent rapidement (en moins de 5 millions d'années). Les polyA qui ne se sont pas transformés subissent, eux, une réduction de taille, et accumulent des mutations ponctuelles, qui aboutiront là encore à l'apparition de quelques microsatellites riches en A. Cette phase se déroule en une dizaine de millions d'années environ. Les mutations s'accumulent ensuite de manière constante, produisant régulièrement de nouveaux microsatellites tandis que la proportion de polyA se réduit. Parallèlement, les microsatellites apparus en premier subissent eux-aussi une réduction de taille qui mène probablement à leur disparition. Le temps nécessaire à la réduction et à la disparition de ces microsatellites ne peut par contre pas être évalué à partir de nos données. Enfin, tous les polyA ont été dégradés ou transformés, et les microsatellites restants sont eux-aussi sur le déclin. Le début de cette phase semble être visible pour les AluJ, c'est-à-dire pour des âges entre 60 et 80 Ma.

4.3.4 Apparition en interne

L'une des observations les plus intéressantes des analyses présentées est l'association très forte de certains motifs avec des positions particulières des séquences Alu. Les microsatellites $(AGGC)_n$, $(GTG)_n$, $(GAGGCT)_n$, $(GAGGTG)_n$ et $(CACTG)_n$ sont associés aux positions 50, 142, 178, 207 et 242 dans la séquence consensus du AluS, respectivement. Ces positions sont légèrement variables selon la famille Alu. En général, près de 50% de ces motifs sont associés aux éléments Alu, proportion atteignant 90% pour les GAGGTG. Ces microsatellites sont en général très courts (10 à 13 nt) et apparaissent par mutation ponctuelle à partir d'un quasi-microsatellite présent dans la séquence Alu. On peut d'ailleurs remarquer un effet majoritaire des mutations aux sites CpG (environ 10 fois plus de locus créés par ce biais qu'à partir d'autres sites ; voir figure 4.3).

Ces microsatellites sont significativement associés à l'ensemble des familles Alu, mais l'étude détaillée des sous-familles des éléments Alu les plus récents montre une répartition hétérogène des motifs. En effet, les sous-familles Alu connues pour être encore actives de nos jours (Ya5 et Yb8/b9) ne sont associées à aucun des microsatellites internes (sauf les AGGC), alors que les autres sous-

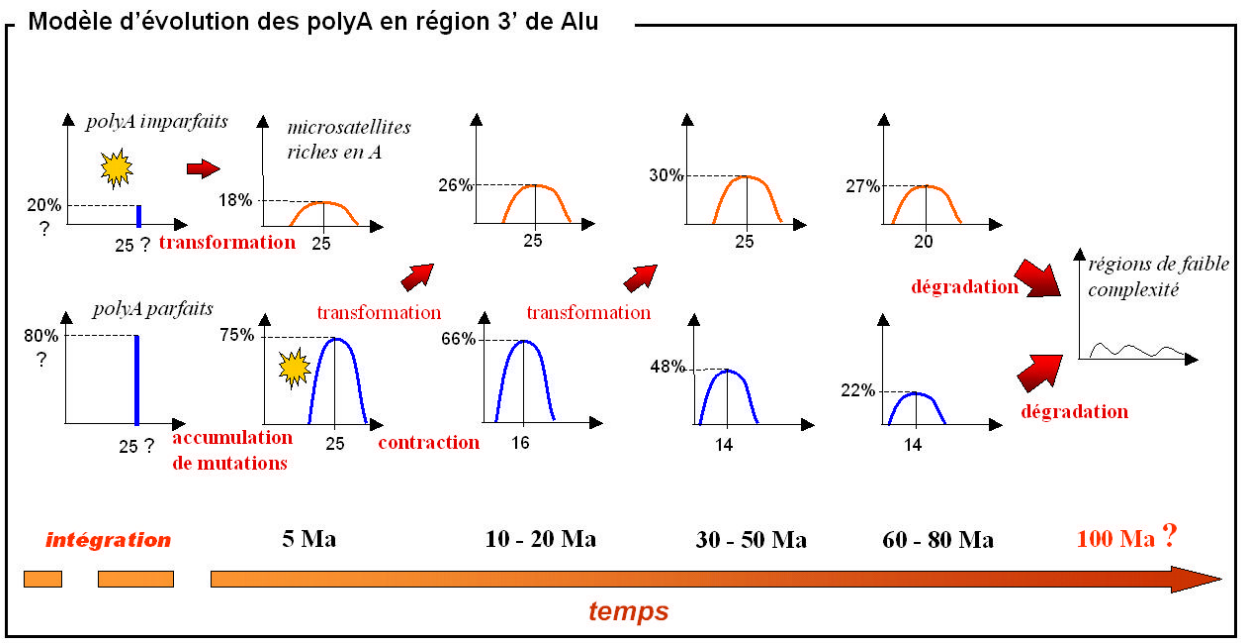


FIG. 4.11 – Modèle du cycle de vie des polyA associés aux régions 3' de séquences Alu. L'échelle temporelle est indiquée en millions d'années, en bas de la figure. Chaque diagramme correspond à la distribution de taille des microsatellites associés aux régions 3' des éléments Alu. l'abscisse est la taille en nucléotides, et l'ordonnée est la proportion d'éléments Alu associés. L'intégration provoque deux distributions, suivant les hypothèses 2 et 3 du modèle présenté figure 4.10. Une majorité des polyA intégrés sont de grande taille, et parfait. Une plus petite quantité est intégrée à la même taille, mais sont imparfaits. Ces derniers se transforment en microsatellites riches en A, tandis que les parfaits accumulent des mutations puis se contractent. Certains peuvent ensuite se transformer en microsatellites riches en A. La proportion de polyA associés aux éléments Alu se réduit au cours du temps, à cause des disparitions par mutation ponctuelle, et des transformations. La proportion de microsatellites riches en A augmente grâce à l'apport des polyA. Enfin, les polyA deviennent trop peu nombreux et trop dégradés pour produire de nouveaux microsatellites riches en A, la proportion de ces derniers se réduit donc, jusqu'à leur disparition. Et les derniers polyA finissent par disparaître. La taille d'intégration et les proportions de départ sont inconnues, ainsi que le temps de dégradation final. Les autres valeurs sont estimées à partir des résulats présentés dans le texte.

familles en possèdent déjà la plupart. Une explication possible est que la mutation ait eu lieu dans le gène maître de certaines des sous-familles, et que le microsatellite ait été intégré de fait avec toutes les nouvelles copies. Toutefois, cela supposerait que les gènes maîtres des cinq sous-familles fortement associées aient tous subi au moins trois des cinq mutations nécessaires à l'apparition de ces motifs, sans que les AluYa5/b8/b9 n'en subissent aucune, malgré leurs âges assez similaires. En revanche, le fait que les AluYa5/b8/b9 soient toujours en activité signifie que de nombreuses copies ont été intégrées très récemment dans le génome. Ces copies n'ont par conséquent pas encore pu accumuler de mutations par rapport au gène maître, et font donc baisser le taux d'association de leur sous-famille. L'association entre ces microsatellites et les éléments Alu est donc plus probablement

causée par une apparition rapide suite à l'intégration.

Quel que soit l'âge d'intégration de l'élément Alu dans lequel ils sont apparus, ces microsatellites internes ne semblent pas être capables d'expansion. En effet, même pour les éléments Alu les plus anciens, les locus de plus de 12 nt sont quasiment inexistants. Les $(GAGGCT)_n$ et $(CACTG)_n$ présentent bien des pics à des tailles de 16 et 14 nt respectivement, c'est-à-dire à 4 et 2 nt de plus que leur taille à la création. Etant donné la période de ces motifs, ces accroissements ne peuvent être dus à des événements de glissement, et sont probablement causés par de nouvelles mutations ponctuelles. Dès lors, il serait plus juste de qualifier ces détections de proto-microsatellites, conformément à notre modèle du cycle de vie présenté en section 2.4.2. Pourquoi ces proto-microsatellites sont-ils incapables d'entamer un cycle de vie normal alors que leur apparition est largement favorisée à ces sites ? Quatre des cinq motifs sont des penta et hexanucléotides, pour lesquels 10 à 12 nt ne représentent que deux répétitions. Il est donc possible que cela soit insuffisant pour autoriser le glissement, et ces proto-séquences seraient amenées à disparaître sans avoir la possibilité de se développer. D'autre part, ces motifs sont très riches en G/C, et il est possible que leur composition soit un frein à leur expansion.

Des cas d'apparition de microsatellites à l'intérieur d'éléments transposables ont déjà été documentés pour le blé [Ramsay et al., 1999], la souris [Desmarais et al., 2006], et la drosophile [Wilder and Hollocher, 2001]. Le cas des éléments mobiles *mini-me* de la drosophile est particulièrement intéressant. Deux types de microsatellites sont associés à cet élément : des $(TA)_n$ et des $(GTCY)_n$ (avec Y égal à C ou T). Les premiers sont constitués de quatre répétitions dans le consensus de l'élément mobile, et les cas d'expansion/contraction semblent assez courants. Le deuxième type de microsatellites apparaît via des mutations ponctuelles à un site quasi-microsatellite, comme c'est le cas pour nous. Malgré les diverses mutations possibles, les auteurs observent eux-aussi préférentiellement des transitions C vers T. Mais contrairement à ce que nous avons trouvé, les microsatellites produits sont capables d'expansion, puisque la taille moyenne est de 5,1 répétitions pour les 23 locus qu'ils ont observés (le plus long possédant 37 répétitions). Le motif (GTCY) est relativement riche en G/C, et la période est de quatre nucléotides.

Une différence majeure avec la plupart de nos apparitions est que les mutations ponctuelles dans les *mini-me* produisent directement des $(GTCT)_3$ et des $(GTCC)_{2,75}$. On pourrait donc penser que la taille des proto-microsatellites est déterminante pour pouvoir initier le glissement. Pourtant, les $(GTG)_n$ que nous observons apparaissent directement avec 3,3 répétitions, certains CACTG et GAGGCT atteignent aussi une taille de 2,8 et 2,7 répétitions grâce à des mutations ponctuelles

supplémentaires, mais aucun ne semble avoir pu se développer. D'autres contraintes sont donc certainement à l'œuvre. On peut par exemple supposer que le développement de ce type de microsatellites est plus aisé chez la drosophile que chez l'homme. De plus, il ne faut pas oublier que ces microsatellites courts sont apparus dans des séquences particulières, qui, même si elles ne sont probablement pas actives, possèdent peut-être des propriétés structurelles défavorisant des variations de taille.

4.3.5 Conclusion

Ce chapitre indique que les éléments transposables jouent un rôle majeur dans la densité de microsatellites dans les génomes, du moins pour ce qui est des primates. 50% des $(A)_n$, le motif le plus représenté dans le génome, sont en effet directement issus d'un polyA ou d'un linker de Alu. Ces polyA sont aussi un vecteur très favorable d'apparition d'autres microsatellites riches en A. De plus, nous avons montré que ces polyA et microsatellites sont plus grands que ceux présents de manière générale dans le génome, ce qui pourrait avoir une influence sur leur dynamique évolutive.

La transformation des polyA en un autre motif est par ailleurs extrêmement rapide, laissant supposer une accumulation étonnante de mutations dans la queue polyA de chaque Alu nouvellement intégré. Une hypothèse alternative a été proposée par Arcot *et al.* (1995), selon laquelle les mutations seraient contractées durant l'intégration des éléments Alu. Nos résultats semblent confirmer ce phénomène, qui serait un complément à l'apparition par mutation aléatoire classique. Nous avons alors proposé un modèle d'évolution des polyA des séquences Alu, qui suppose une transformation des polyA en autres microsatellites de manière continue, jusqu'à l'épuisement de toutes les séquences polyA intégrées. Cet épuisement pourrait avoir lieu en 60 millions d'années environ.

Enfin, nous avons constaté l'apparition de multiples proto-microsatellites en des sites internes de Alu. Ces séquences, pour la plupart des tetra, penta et hexanucléotides de deux ou trois répétitions et riches en G/C ne semblent pas pouvoir se développer en microsatellites plus longs. Ces résultats sont d'une grande importance pour comprendre la dynamique d'apparition de ce type de microsatellites, qui ont été très peu étudiés. Enfin nous avons observé dans le linker d'une partie des AluY, un proto-microsatellite $(AAAAAC)_n$ probablement directement intégré avec les éléments Alu. Tous ces éléments d'origine commune pourraient permettre la caractérisation d'une nouvelle sous-famille de AluY, suivie d'une étude systématique du devenir de ces séquences.

Chapitre 5

Apparition *de novo*

5.1 Une taille minimum pour les microsatellites ?

Le chapitre précédent a été consacré à l'influence des séquences Alu, un élément transposable des primates, sur l'apparition des microsatellites. Les microsatellites étant présents chez tous les organismes, et tous les motifs étant représentés (certes à des degrés variables), il est évident qu'il existe d'autres mécanismes à l'origine des répétitions en tandem. Ce chapitre se focalise sur l'apparition *de novo*, c'est-à-dire à partir de séquences quelconques d'ADN.

Pour étudier correctement l'apparition des microsatellites, il convient tout d'abord de se poser la question de savoir ce qu'est exactement un microsatellite. Nous avons expliqué dans le chapitre 2 qu'une définition serait de considérer un microsatellite comme n'importe quelle séquence contenant au moins deux motifs identiques adjacents. Cette définition a l'avantage considérable d'être simple et formelle, donc aisément programmable si l'on a besoin de détecter des microsatellites dans de l'ADN séquencé (*cf.* chapitre 3). Toutefois, la réalité biologique est nettement plus complexe, car une seconde propriété de ces séquences est leur dynamique particulière d'expansion/contraction par glissement. D'un point de vue biologique, une séquence répétée ne pourrait donc être définie comme microsatellite qu'à partir du moment où elle est capable de glissement, les autres étant considérées comme proto-microsatellites. Il a été démontré, tant d'un point de vue expérimental [Levinson and Gutman, 1987a] que par modélisation [Kruglyak et al., 1998], que le taux de glissement était corrélé à la taille du locus. Il est donc naturel de se demander s'il existe une taille minimum à partir de laquelle le glissement est possible, donnant une limite de taille pour la définition d'un microsatellite.

Cette question de la taille minimum de glissement est le point central de la plupart des travaux

réalisés sur l'apparition des microsatellites. Nous présenterons dans cette section une synthèse de la littérature concernant ce sujet, et nous proposerons une méthode pour estimer si une taille minimum de glissement existe réellement. La réponse à cette question nous permettra alors de déterminer quels sont les mécanismes à l'origine de l'apparition des microsatellites.

5.1.1 Une sur-représentation des locus courts

Un certain nombre d'études ont testé l'hypothèse d'une taille minimum pour le glissement en évaluant la sur-représentation des locus microsatellites dans un génome en fonction de leur taille. Cette méthode est basée sur le principe que la dynamique de glissement créé un surplus de locus de longue taille par rapport à ce qu'on attend dans une séquence où cette dynamique ne s'exercerait pas. Cette hypothèse est justifiée par le fait que le glissement semble biaisé vers les expansions, au moins pour les locus courts [Primmer and Ellegren, 1998, Xu et al., 2000, Huang et al., 2002]. Par simplicité, le génome dénué de dynamique de glissement est généralement représenté par une séquence avec une répartition aléatoire des nucléotides. Le nombre de locus attendus dans ce type de séquence est alors qualifié d'aléatoire. Une taille minimum pour le glissement serait donc une taille au delà de laquelle le nombre de locus observés dans une séquence est supérieur à celui d'attendus aléatoirement. Plusieurs études se sont basées sur cette hypothèse pour évaluer la présence d'une telle limite, mais leurs résultats et leur conclusions sont assez contradictoires.

Rose et Falush [Rose and Falush, 1998] ont réalisé le calcul sur les mono, di et certains tétranucléotides présents dans le génome de *S. cerevisiae*. Ils ont observé un écart entre nombre de locus observés et nombre attendus à partir de 8 nt, quelle que soit la taille du motif (figure 5.1.1). Ils en ont donc conclu que la taille minimum pour le glissement ne se comptait pas en nombre de répétitions, mais en nombre de nucléotides, et était de 8 nt. Ces résultats ont très largement fait référence par la suite, particulièrement lors de l'implémentation des modèles théoriques de dynamique des microsatellites (voir section 2.4.1). Pupko et Graur [Pupko and Graur, 1999] ont effectué le même travail pour les mono à pentanucléotides, sur la même séquence (celle de la levure), mais ont paradoxalement obtenu des écarts observés/attendus strictement supérieurs à un, pour toutes les tailles, quel que soit le motif (figure 5.1.2). Il n'y a donc pas, selon eux, de limite inférieure pour le glissement, mais simplement une rareté des événements liée à la petite taille, comme cela avait déjà été remarqué pour les mononucléotides [Levinson and Gutman, 1987a]. Enfin, Dieringer et Schlottërer [Dieringer and Schlotterer, 2003] ont eux aussi évalué la sur-représentation des mono à tétranucléotides très courts dans le génome de la levure et leurs résultats montrent une limite dépendante de la période du motif (figure 5.1.3). Ces derniers ont par ailleurs effectué les analyses

sur huit autres génomes, et obtiennent des résultats différents pour tous les organismes. Ils estiment ensuite eux aussi une taille minimum de 8 nt, à partir d'un modèle de dynamique des microsatellites basé sur les mononucléotides.

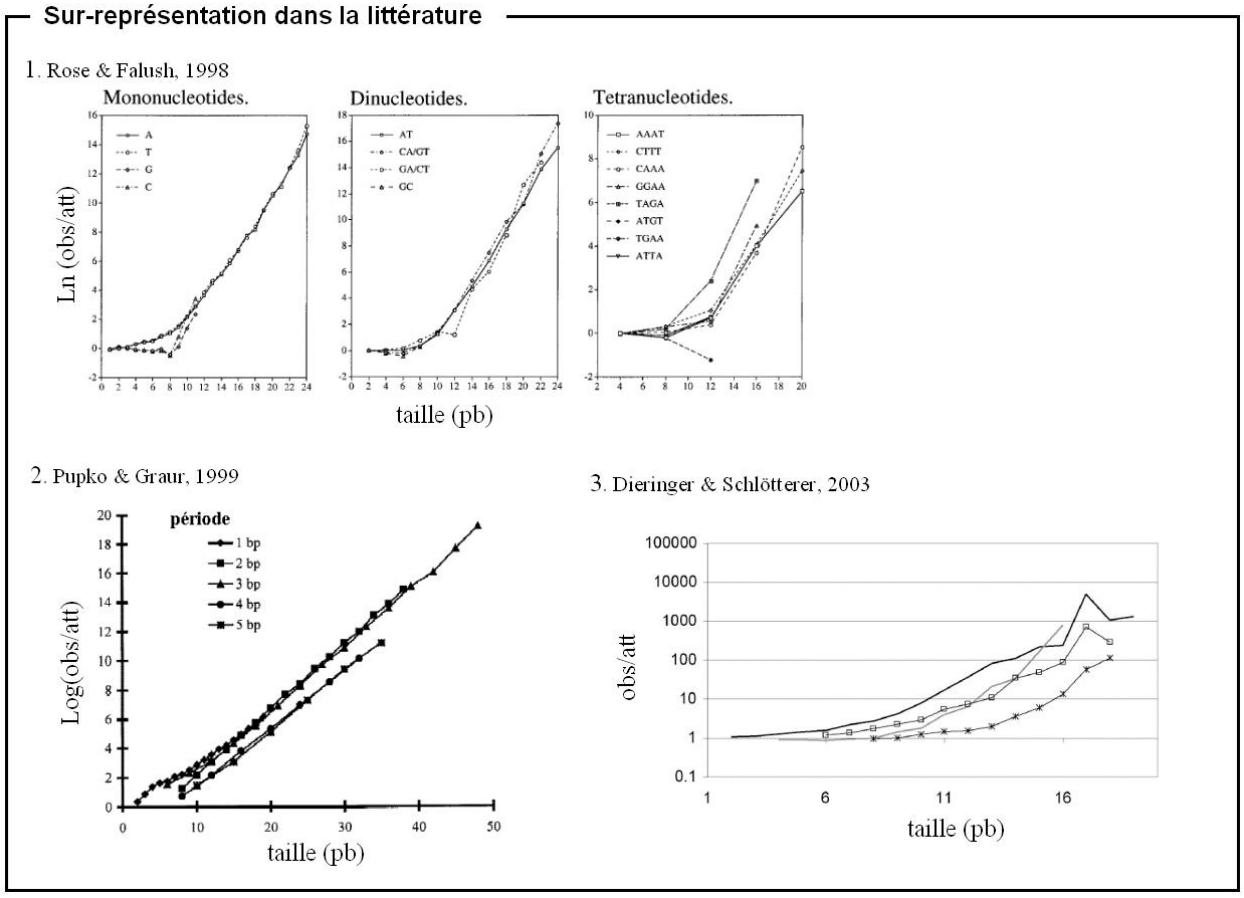


FIG. 5.1 – Ratio entre le nombre observé de microsatellites et le nombre attendu dans un génome aléatoire, par taille, pour plusieurs motifs, chez *S. cerevisiae*. **1-** Rose & Falush (1998). L'attendu est calculé théoriquement à partir du nombre d'occurrences des motifs présents dans la séquence. **2-** Pupko & Graur (1999). L'attendu est calculé théoriquement à partir du taux de GC global de la séquence. **3-** Dieringer & Schlötterer (2003). L'attendu est calculé par simulation à partir du taux de GC local de la séquence.

Les trois études étant basées sur la même séquence génomique, la différence de résultats ne peut s'expliquer que de deux manières. Soit la méthode d'extraction des répétitions ne donne pas le même nombre de locus observés, soit la méthode de calcul de l'attendu aléatoire ne donne pas le même nombre théorique, voire les deux. Seul Dieringer et Schlötterer (2003) expliquent leur méthode d'extraction, or nous avons vu dans le chapitre 3 que la méthode choisie pouvait avoir une influence critique sur le nombre de détections. Nous ne pouvons donc pas nous assurer que la méthode d'extraction donne des données cohérentes entre les études. Néanmoins, aucune des analyses n'a géré

les microsatellites imparfaits, et ne se s'est pas donnée de limite inférieure de taille de détection, les deux facteurs principaux de la variation entre algorithmes de détection. Il est donc probable que les locus utilisés par les trois méthodes aient été sensiblement les mêmes.

Les calculs du nombre théorique de microsatellites attendus sont par contre détaillés dans chacun des articles, et présentent des différences qui ont à l'évidence influencé les résultats. Pupko et Graur (1999) et Dieringer et Schlötterer (2003) se sont basés sur une séquence aléatoire qui respecte le taux de GC global de la séquence de la levure. Ils ont calculé le nombre d'attendus pour chaque taille et pour chaque motif à partir de la proportion de chaque base du motif dans la séquence complète. A titre d'exemple, le nombre de $(AT)_2$ attendus est calculé par :

$$Ne_{ATAT} = S \times (1 - p_A \times p_T)(p_A^2 \times p_T^2)(1 - p_A \times p_T) \tag{5.1}$$

et

$$Ne_{ATAT} = S \times (1 - p_T)(p_A^2 \times p_T^2)(1 - p_A) \tag{5.2}$$

respectivement pour Pupko et Graur (1999) et Dieringer et Schlötterer. S est la taille de la séquence génomique, et p_A et p_T sont les proportions de A et de T dans la séquence. La différence entre les deux méthodes est que la première ne considère un AT à deux répétitions uniquement si il n'est pas encadré par deux autres AT, alors que la seconde restreint à ceux encadrés ni par un T à gauche, ni par un A à droite. Cela signifie que la formule 5.1 compte une séquence de type CCATATAC comme un $(AT)_2$, mais pas la formule 5.2. Il faut en revanche préciser que Dieringer et Schlötterer (2003) calculent aussi le nombre d'attendus pour les locus de taille non multiple de leur période, comme par exemple pour les $(AT)_{2,5}$:

$$Ne_{ATATA} = S \times (1 - p_A)(p_A^{2,5} \times p_T^{1,5})(1 - p_A) \tag{5.3}$$

Cela signifie que dans les calculs théoriques, la séquence CCATATAC est comptée deux fois (pour $(AT)_2$, et $(TA)_2$) avec la première formule, et une fois seulement (pour $(AT)_{2,5}$) dans la seconde. L'extraction des microsatellites suivant les mêmes règles (nous l'espérons !), les locus à nombre de répétitions non entier prennent une importance considérable dans le compte de la valeur observée. Et ce d'autant plus que la taille du motif est grande, puisque un tétranucléotide $(ATGC)_{2,75}$ sera compté quatre fois avec la formule 5.1.

Rose et Falush (1998) se basent, quant à eux, non pas sur une séquence aléatoire respectant le taux de GC de la séquence de la levure, mais respectant le taux du motif considéré. Ils ont en effet calculé le nombre d'occurrences de chaque motif dans la séquence, et ont calculé le nombre attendu en fonction de ce dernier. La formule pour $(AT)_2$ est par exemple :

$$Ne_{ATATA} = S \times (1 - p_{AT}) \times p_{AT}^2 \times (1 - p_{AT}) \tag{5.4}$$

où p_{AT} est la probabilité d'avoir le dinucléotide AT à chaque position de la séquence, calculé par $p_{AT}=\frac{N_{AT}}{S}$, avec N_{AT} le nombre d'occurrences de AT dans la séquence de la levure. Cette formule, à la différence des deux premières, permet de se libérer des contraintes d'adjacence entre bases, qui peuvent jouer sur la composition en motifs, tout en étant indépendant du phénomène de glissement. Le cas le plus évident est la sur-représentation des motifs TG et CA dans les génomes causée par la déamination de sites CpG méthylés (voir section 1.2.2). Par contre, par cette méthode, la séquence CCATATAC sera aussi comptée deux fois, pour les mêmes raisons que dans la formule 5.1. Il est à noter que ces trois formules sont équivalentes concernant les mononucléotides, or les courbes de ratio observés/attendus sont différentes, même pour cette classe de microsatellites (figure 5.1).

Enfin, les articles de Rose et Falush (1998) et de Pupko et Graur (1999) ne proposent qu'un ratio de nombre observé sur nombre attendu, sans évaluer si ce ratio est significativement différent de 1, qui représente l'hypothèse nulle. Dans la figure 5.1.1, on observe très clairement une augmentation du ratio à partir de 8 nt pour les $(C)_n$ et $(G)_n$, mais cette valeur est nettement plus arbitraire concernant les $(A)_n$ et $(T)_n$. Or la séquence de *S. cerevisiae* est composée à environ 60% de A/T [Goffeau et al., 1996], ce qui joue un rôle important sur la probabilité d'obtenir une séquence polyA ou polyT d'une taille donnée par hasard. Dieringer et Schlötterer (2003) ont en revanche calculé un intervalle de confiance pour leurs ratios, et montrent justement que les $(A)_n$ sont sur-représentés à partir de 3 nt, alors que les $(G)_n$ ne le sont qu'à partir de 11 nt. Ils obtiennent aussi une sur-représentation des dinucléotides à partir de 12 nt, et des tri et tétranucléotides à partir de 6-8 nt, 12-13 nt ou 15-16 nt, en fonction du motif.

La synthèse de ces résultats montre que le calcul de la sur-représentation des motifs dans les génomes par rapport à un attendu aléatoire dépend de plusieurs paramètres, et cela a abouti à des incohérences entre les divers travaux. Dès lors, les estimations d'une taille minimum pour le glissement à partir de ces sur-représentations sont à prendre avec précaution, et la limite à 8 nt proposée par Rose et Falush (1998) est probablement bien moins évidente qu'elle n'a été supposée.

5.1.2 Cas d'apparition de répétitions en tandem

L'évaluation de la sur-représentation des locus par rapport à un attendu aléatoire permet éventuellement d'estimer une taille minimum à partir de laquelle un proto-microsatellite deviendrait un locus mature, capable de glissement, mais ne renseigne en rien sur les mécanismes qui permettent d'atteindre cette taille. L'hypothèse la plus communément admise (et sur laquelle se basent implicitement les tests de sur-représentation énoncés plus haut), est que les séquences répétées apparaissent par mutation ponctuelle à partir de séquences de faible complexité, ou régions de « cryptic simplicity », en anglais [Levinson and Gutman, 1987a, Tautz et al., 1986].

En 1996, Messier, Li et Stewart [Messier et al., 1996] ont présenté le premier exemple d'apparition de microsatellite par mutation ponctuelle. Ils ont comparé les séquences d'un pseudogène de $\eta - globin$ chez plusieurs primates, et ont caractérisé la création de deux microsatellites à partir d'une séquence quasi-microsatellite (figure 5.2.1). Chez la plupart des primates, la séquence est TTGTATGTGTGTGA, mais dans la branche commune à l'homme et au gorille, ils observent la séquence $\mathrm{TTGT(ATGT)_{4-5}GA}$. Ils supposent donc qu'une mutation G vers A à la position 9 a produit un $(\mathrm{ATGT})_2$ qui s'est ensuite développé par glissement. Ils montrent de plus que dans un autre singe, une mutation de A vers G à la cinquième position a produit un $(\mathrm{TG})_6$ qui a ensuite gagné une répétition.

Quelques autres études ont par la suite relaté d'autres cas d'apparition de répétitions, par exemple chez les conifères [Sokol and Williams, 2005] ou chez les drosophiles [Noor et al., 2001]. L'étude de Noor *et al.* (2001) est particulièrement intéressante, car elle propose une apparition par glissement. Ils ont observé le passage d'un $(\mathrm{CA})_3$ à un $(\mathrm{CA})_5$ dans l'une des branches du groupe *obscura*, ce $(\mathrm{CA})_5$ étant ensuite devenu polymorphe. Ils déduisent de cette observation qu'il n'existe pas de limite de taille au glissement, en tout cas pas à 8 nt. L'expansion d'un $(\mathrm{GA})_2$ à un $(\mathrm{GA})_3$ a aussi été observé chez des hirondelles [Primmer and Ellegren, 1998]. Enfin, Gordon a publié un article en 1997 [Gordon, 1997] qui proposait une interprétation alternative aux conclusions de Messier, Li et Stewart (1996). Il se proposait d'expliquer les séquences observées grâce à des événements de glissement plus que par mutation ponctuelle (figure 5.2.2). Les apparitions présentées ci-dessus ne sont que des cas particuliers, ne représentant que l'histoire évolutive du locus concerné, et ne peuvent être pris pour des généralités. Ils montrent néanmoins que la mutation ponctuelle et les processus de glissement peuvent être à l'origine de locus microsatellites.

Au delà de l'apparition par mutation ponctuelle et du glissement même pour les petites tailles,

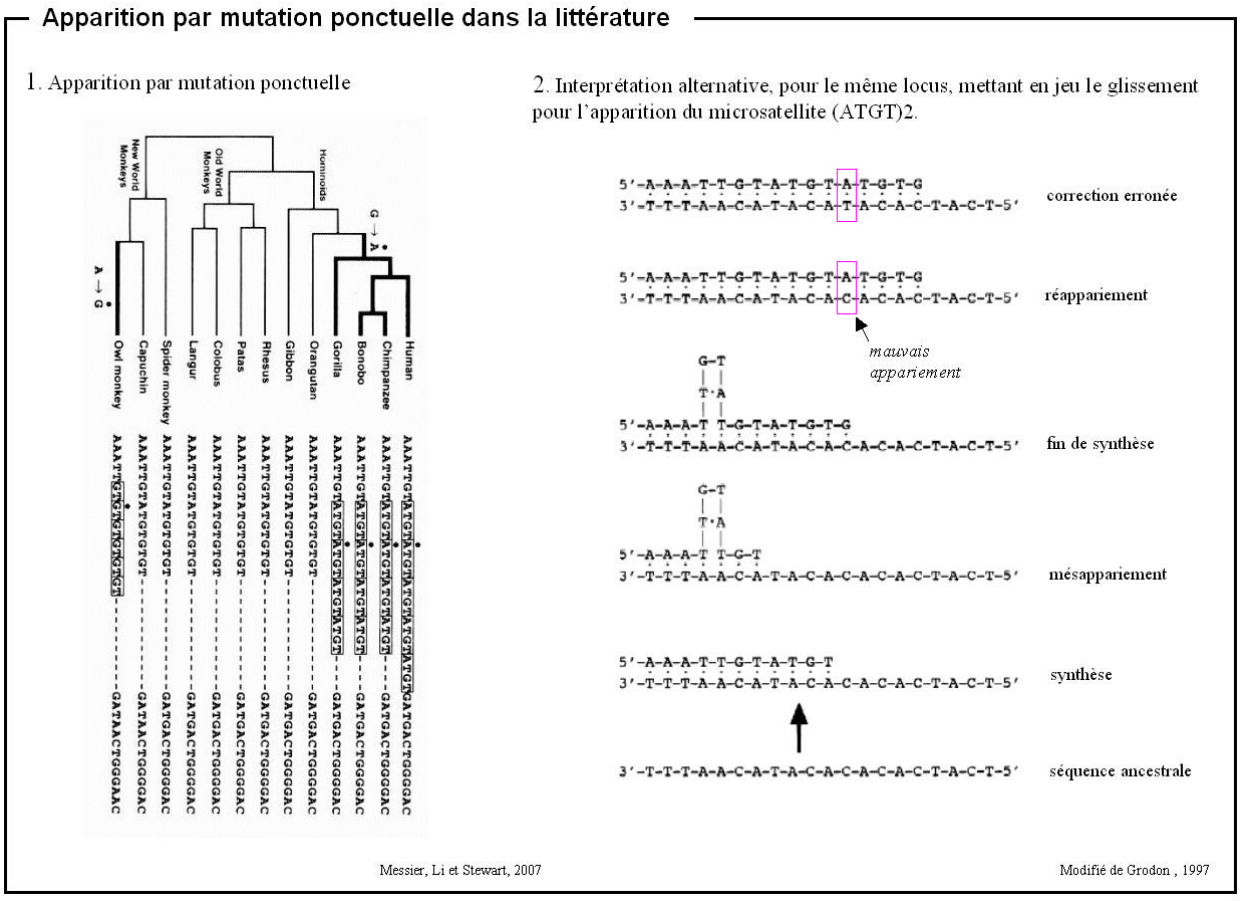


FIG. 5.2 – **1-** Apparition d'un tétranucléotide et d'un dinucléotide par mutations ponctuelles dans la séquence d'un pseudogène de $\eta - globin$ de primate. Les positions de mutation sont indiquées par un point au dessus de la séquence. **2-** Hypothèse alternative, expliquant l'apparition du $(ATGT)_n$ de l'exemple précédent par un événement de glissement. La succession des événements se lit de bas en haut.

il semblerait qu'il existe un troisième mécanisme pouvant expliquer l'apparition des microsatellites. Zhu *et al.* [Zhu et al., 2000], puis Messer et Arndt [Messer and Arndt, 2007], ont en effet observé que les insertions de très petites tailles étaient causées majoritairement par des duplications du motif adjacent. Zhu *et al.* (2000) montrent par exemple que c'est le cas de plus de 70% des insertions causant des maladies génétiques chez l'homme. Messer et Arndt (2007), à partir de l'alignement global homme-chimpanzé-macaque, ont trouvé un taux de 84% pour les insertions de 1 à 30 nt. Ces valeurs sont toutefois à revoir à la baisse car une partie des insertions semble causée par des événements de glissement. Respectivement 26% et 32% des insertions de 2 et 3 nt ont ainsi eu lieu à des positions où il existait déjà de 2 à 4 répétitions du motif, dans les données de Zhu *et al.* (2000). Par contre ce n'est le cas que pour 4% des insertions de 4 nt. De même, pour Messer et Arndt (2007), il existait aussi une petite répétition en tandem pour 25% de leurs duplications.

Néanmoins, il reste 75% des duplications qui ne peuvent être expliquées par du glissement, ce qui suppose qu'il existe un autre phénomène favorisant l'apparition de duplications en tandem à partir d'une séquence quelconque. Zhu *et al.* (2000) ont toutefois estimé que seulement 2-3% des di et trinucléotides de deux répétitions sont apparus par ce biais, les autres étant apparus par substitution. Ce rapport passe à presque 20% pour les tétranucléotides.

5.1.3 Estimer l'importance du glissement pour les petites tailles

La synthèse de ces résultats ne nous a finalement pas permis de déterminer s'il existait une taille minimum pour le glissement, qui permettrait de donner une borne inférieure pour la définition d'un microsatellite. En effet, les calculs de sur-représentation se contredisent, tout comme les études phylogénétiques. De plus, la découverte du phénomène de duplication en tandem nécessite une ré-interprétation des résultats précédents. Ce mécanisme, que nous nommerons *micro-duplication* par la suite, a les mêmes conséquences que le glissement, à savoir l'insertion adjacente d'une répétition d'un motif. Si le motif n'est pas répété, cela créé un simple doublon, mais si le motif est déjà répété en tandem, cela produit une expansion. Le cas d'expansion interprété par Noor *et al.* (2001) comme du glissement pourrait n'être que la conséquence d'une micro-duplication du motif CACA. De même, la sur-représentation des motifs répétés observée par Pupko et Graur (1999), pourrait n'être causée que par les événements de micro-duplication. C'est d'ailleurs l'interprétation proposée par Dieringer et Schlötterer (2003), qui intègrent un paramètre de *indel-slippage* à leur modèle théorique de dynamique des microsatellites pour expliquer la sur-représentation des locus de faible taille.

Il semble donc que trois mécanismes existent pour la création de microsatellites. Nous nous proposons, dans le travail présenté dans la suite, d'évaluer l'importance de chacun d'eux, à partir d'événements d'apparition recensés dans le génome humain (méthode détaillée dans la section suivante). L'effet de la mutation ponctuelle est aisément reconnaissable, car l'apparition est causée par une transformation aléatoire (substitution ou indel) de la séquence flanquant la répétition. L'effet des deux autres mécanismes est plus difficile à discerner en ce qu'ils provoquent tous deux la duplication adjacente d'un motif complet. Si l'on suppose que ces deux phénomènes sont indépendants, leurs effets devraient être additifs. Autrement dit, s'il est possible d'estimer le nombre de duplications causées par l'un des deux mécanismes, l'effet de l'autre mécanisme pourra être déduit du nombre total d'apparition par duplication (par une simple différence).

Notre premier travail a donc été de distinguer les conséquences d'événements de micro-duplication

de ceux de glissement. Nous nous sommes focalisés sur l'apparition des répétitions les plus petites possible, constituées de la répétition simple d'un motif donné, que nous nommerons doublon. L'apparition de ces éléments dans le génome n'est possible que par la mutation ponctuelle ou la micro-duplication, car le glissement ne peut s'effectuer s'il n'existe pas de répétition préalable (figure 5.3). Grâce à ce modèle d'étude, l'importance de la micro-duplication dans l'apparition des séquences répétées a pu être déterminé, sans influence du glissement.

L'estimation du taux de micro-duplication nous a ensuite permis d'évaluer l'importance du glissement dans l'apparition des microsatellites, à partir de l'apparition des motifs répétés trois fois, appelés triplets. Ces éléments peuvent en théorie être créés par mutation ponctuelle à partir d'un quasi-microsatellite, ou par micro-duplication/glissement à partir d'un doublon (figure 5.3). Les locus créés par duplication seront donc la conséquence d'événements de micro-duplications d'une part, et d'événements de glissement d'autre part. Si le taux de duplication observé est le même que pour les doublons, cela signifiera que seule la micro-duplication entre en jeu (en supposant que le taux de micro-duplication est le même pour les doublons et les triplets), indiquant une taille minimum pour le glissement.

De plus, il est possible que si une telle limite existe, elle ne soit pas la même pour toutes les tailles de motif, comme le suggère les taux de sur-représentation de Dieringer et Schlötterer (2003). Nous avons donc étudié l'influence de la micro-duplication et du glissement sur l'ensemble des classes de motifs (période de 1 à 6 nt). L'analyse sera de plus étendue aux heptanucléotides, afin d'évaluer si les résultats obtenus peuvent être généralisés aux motifs d'une taille supérieure à celles généralement considérées pour les microsatellites.

5.2 Méthodes et résultats

5.2.1 Méthode d'alignement

Un outil, la génomique comparative

Nous avons vu que pour caractériser des événements d'apparition de locus microsatellites, il était possible d'utiliser une approche phylogénétique (section 2.2). Malheureusement, observer un nombre d'apparitions suffisant pour pouvoir généraliser les résultats suppose de savoir à l'avance que l'apparition des locus choisis sera visible à partir des espèces sélectionnées pour réaliser l'analyse. La durée du cycle de vie des microsatellites étant inconnue et probablement variable pour chaque locus, il semble peu réaliste d'espérer obtenir suffisamment de données à partir d'une étude phylogénétique

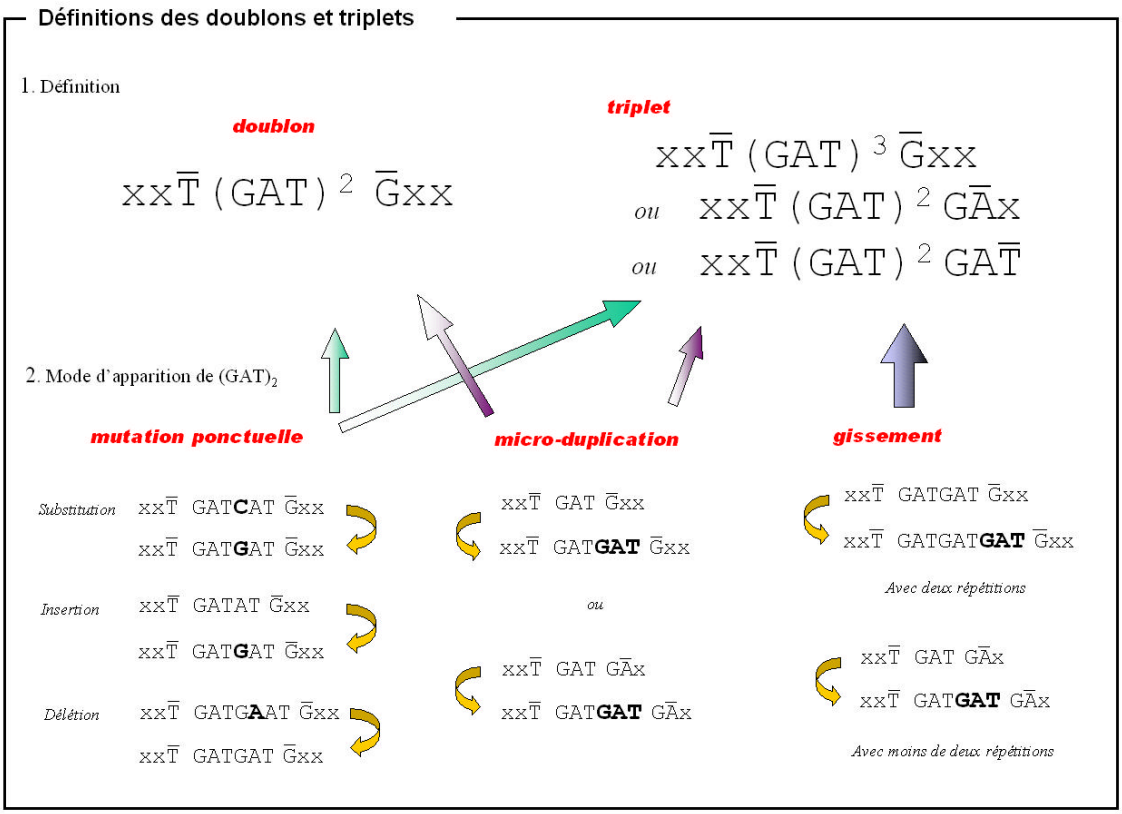


FIG. 5.3 – **1-** Définition des doublons et triplets. Les définitions sont données à titre d'exemple pour les doublons et tripets de motif (GAT), mais sont valables pour tous les motifs de toutes les tailles. ($xx\bar{T}$) représente n'importe quel motif trinucléotide ne finissant pas par un T, ($\bar{G}xx$) n'importe quel motif ne commençant pas par un G, ($G\bar{A}x$) n'importe quel motif commençant par un G, mais sans A à la deuxième position, et ($GA\bar{T}$) les motifs commençant par GA et ne finissant pas par T. Les locus possédant un nombre de répétitions compris entre 2 et 3 ont été intégrés aux triplets, car nous avons considéré que le glissement était possible, même avec une répétition incomplète. **2-** Mécanismes possibles pour l'apparition de ces doublons et triplets. Ces mécanismes n'ont été représentés que pour les doublons $(GAT)_2$, mais sont valables pour tous les motifs de toutes les tailles. La symbolique est la même que pour les définitions. Les doublons ne peuvent apparaître que par mutation ponctuelle ou micro-duplication, tandis que les triplets peuvent apparaître par mutation ponctuelle, micro-duplication et glissement.

classique.

Nous allons donc utiliser une méthode dite de génomique comparative. Durant les dix dernières années, le séquençage de génomes complets a été effectué sur un grand nombre d'organismes divers. Ces génomes complets peuvent être alignés, permettant de déterminer les différences au niveau moléculaire entre les organismes. En alignant la séquence de trois organismes, il est de plus possible de savoir dans quel génome les mutations se sont produites, et leur type (substitutions, duplication segmentaire, etc.), comme cela a été expliqué pour les méthodes phylogénétiques (voir section 2.2.2).

Mais à la différence des techniques phylogénétiques classiques, les techniques de génomique comparative autorisent l'analyse des mutations sur l'ensemble du génome, et non plus à un ou quelques locus déterminés.

Cette technique va nous permettre de caractériser un grand nombre d'événements d'apparition de doublons et triplets, et ainsi d'évaluer l'importance relative des trois mécanismes proposés. Ce type d'approche nécessite toutefois de comparer les génomes d'organismes n'ayant pas divergé depuis trop longtemps, pour éviter les cas d'apparition à un même site, mais dans deux génomes différents. C'est par exemple le cas du triplet homme-chimpanzé-macaque, dont les divergences sont estimées au minimum à 4,2 Ma [Kumar et al., 2005] entre l'homme et le chimpanzé et 23,8 Ma pour celle homme/chimp-macaque [Kumar et al., 2005]. L'Université de Californie-Santa Cruz (UCSC) propose sur son site l'alignement multiple de génomes de 16 vertébrés, dont ceux de l'homme (hg18, mars 2006), du chimpanzé (panTro1, novembre 2003) et du macaque (rheMac2, janvier 2006). Nous avons extrait les alignements de ces trois espèces et considéré le macaque comme groupe externe pour évaluer les apparitions.

Alignement

Tous les doublons présents dans la séquence humaine ont d'abord été détectés, en respectant la définition donnée dans la figure 5.3. La séquence homologue chez les deux autres singes a ensuite été extraite pour chacun d'entre eux, à partir de l'alignement multiple. Les 10 nt flanquants en 3' et en 5' ont aussi été extraits, et elles serviront à évaluer la taille des indels. Les alignements obtenus ont ensuite été traités pour rendre les données plus facilement analysables. En effet, certains des alignements sont situés dans des régions chromosomiques insérées ou supprimées dans l'une des espèces, par exemple à l'intérieur d'éléments transposables. Les apparitions (ou disparitions) relevant de ce genre d'événements ne nous intéressent pas, car elles sont considérées comme adoptées, et non comme apparues à partir de la séquence d'origine. Ces insertions-délétions sont visibles dans l'alignement multiple par des gaps (*i.e.* des bases non alignées) importants dans au moins une des séquences (voir figure 5.4.1). Nous avons donc retiré de l'analyse tous les locus dont l'alignement total (doublons plus flanquant) de l'une des séquences possédait plus de 10 gaps. Cette valeur a été choisie de manière à ce que tout alignement avec l'une des deux régions flanquantes non alignée soit éliminé.

Les alignements ont ensuite été homogénéisés aux sites où plusieurs alignements étaient possibles. En effet, l'insertion (ou la délétion) de quelques bases peut produire une incertitude au niveau de

l'alignement, quand la ou les bases précédant l'indel sont les mêmes que la ou les dernières bases du fragment inséré ou supprimé (figure 5.4.2). Dans ces cas particuliers, l'alignement a été modifié pour reconstruire au maximum des doublons complets à partir du flanquant, comme présenté dans la figure 5.4.2. Par extension, les doublons créés par contraction d'un triplet (où toutes les bases supprimées correspondent aux bases adjacentes) sont aussi systématiquement reconstruits.

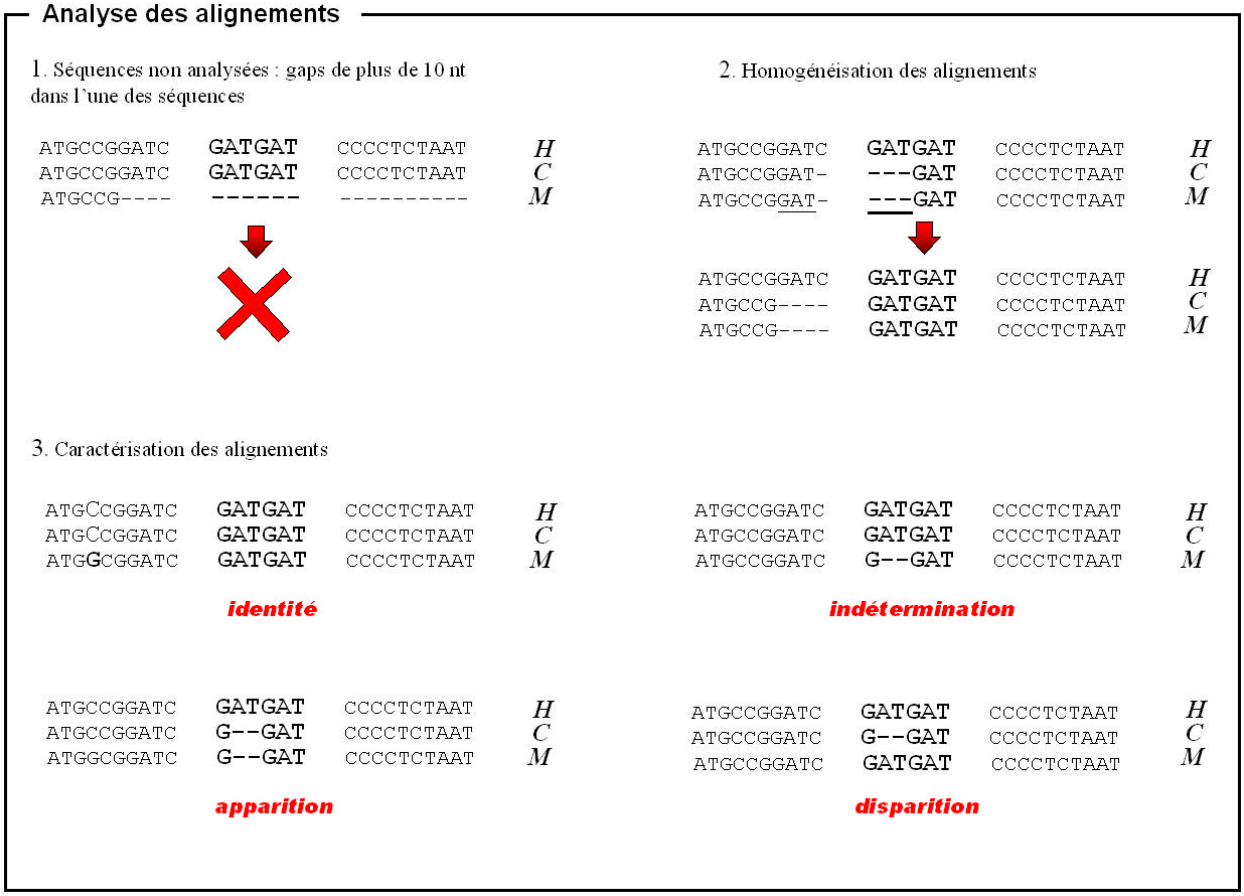


FIG. 5.4 – Les différents traitements effectués sur les alignements. *H*, *C* et *M* représentent les séquences humaine, de chimpanzé et de macaque, respectivement. **1-** Filtre des alignements. Les alignements dont au moins l'une des séquences possède un gap de plus de 10 nt sont retirés de l'analyse. **2-** Homogénéisation des alignements. Les alignements équivalents sont résolus pour favoriser la formation de doublons ou triplets. Les bases soulignées représentent des alignements équivalents. **3-** Les quatre cas d'alignements possibles. Identité, lorsque l'alignement du doublon/triplet est identique pour les trois séquences. Indétermination, lorsque les séquences de l'homme et du chimpanzé sont identiques, mais que celle du macaque est différente, au niveau du doublon/triplet. Apparition, lorsque les séquences du macaque et du chimpanzé sont identiques, mais que celle de l'homme est différente, au niveau du doublon/triplet. Disparition, lorsque les séquences du macaque et de l'homme sont identiques, mais que celle du chimpanzé est différente, au niveau du doublon/triplet.

Les alignements homogénéisés ont ensuite été analysés, et quatre cas de figure sont alors observables : le doublon est présent chez les trois organismes, il est apparu chez l'homme, il a disparu chez le chimpanzé, ou bien il y a une indétermination. Puisque nous travaillons sur les doublons existants chez l'homme, nous ne pouvons détecter les apparitions chez le chimpanzé, ou les disparitions chez l'homme.

La présence chez les trois espèces est caractérisée par le doublon identique sur les trois séquences, indépendamment de la séquence flanquante (figure 5.4.3). Ceci implique que tous les doublons apparus par contraction d'un triplet seront considérés dans notre analyse comme identiques chez les trois espèces. En effet, à cause de notre phase d'homogénéisation, le doublon sera aligné sur deux des trois répétitions du triplet, la dernière étant alignée à un gap dans la région flanquante (voir figure 5.4.2). Cela était nécessaire, car l'estimation de la micro-duplication repose sur des apparitions issues d'une séquence génomique quelconque, donc non répétée. Les cas d'apparition sont détectés lorsque le doublon existe dans l'alignement humain, mais est absent pour les deux autres primates. Il faut en outre que les séquences de ces derniers soient identiques. Cela signifie sans ambiguïté que la séquence originale ne possédait pas le doublon, et qu'un événement de mutation l'a fait apparaître (figure 5.4.3). De la même manière, la disparition est caractérisée par l'absence du doublon dans la séquence du chimpanzé, alors qu'il est présent chez l'homme et le macaque. Enfin, les cas indéterminés sont ceux où le doublon est présent chez l'homme et le chimpanzé, mais absent chez le macaque (figure 5.4.3). En effet, on ne peut déterminer si cet alignement a été produit par l'apparition du doublon dans l'ancêtre commun de l'homme et du chimpanzé, ou par sa disparition chez le macaque.

Notre travail s'est ensuite focalisé sur l'analyse de l'apparition des doublons. Là encore, plusieurs cas sont à distinguer. Une apparition par substitution est détectée lorsque l'alignement du doublon entre l'homme et les deux autres séquences ne diffère que par des substitutions, indépendamment des séquences flanquantes (figure 5.5). Il n'y a en général qu'une seule substitution visible, mais les rares cas de substitutions multiples sont considérés comme une seule apparition. L'apparition par insertion est caractérisée par la présence d'un gap dans les séquences du chimpanzé et du macaque, situé dans l'alignement du doublon (figure 5.5). Parfois l'insertion est à cheval entre le doublon et la région flanquante, la taille complète de l'insertion sera dans ces cas-là prise en compte. Dans les rares cas où les séquences diffèrent par une insertion plus une substitution, seule l'insertion sera comptabilisée. Dans les cas encore plus rares où deux insertions indépendantes se sont produites, seule l'insertion la plus grande sera comptabilisée. Enfin, l'apparition par délétion est caractérisée par un gap dans l'alignement du doublon humain, absent dans les deux autres séquences (figure 5.5). Les mêmes règles que pour l'insertion sont alors appliquées.

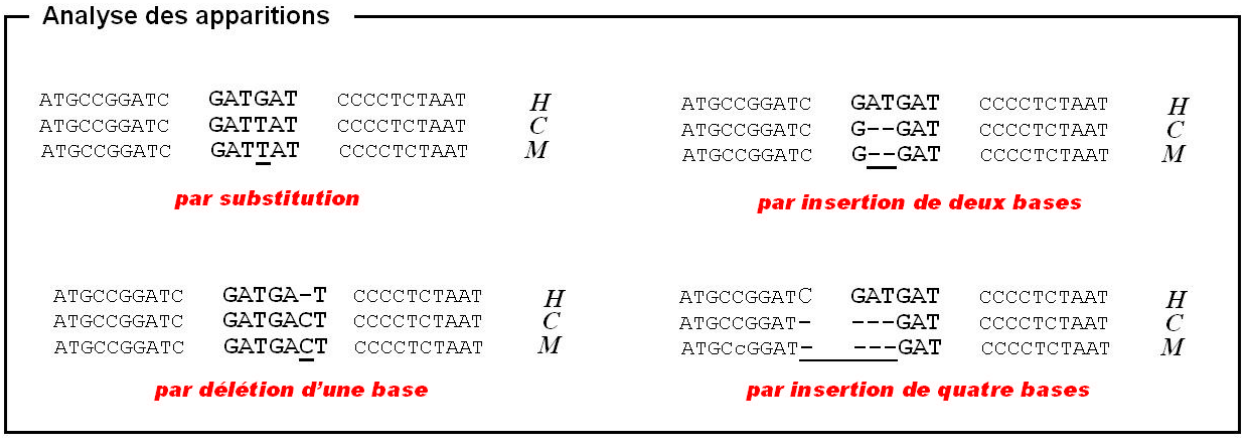


FIG. 5.5 – Les différents cas d'apparition de doublons et triplets. *H*, *C* et *M* représentent les séquences humaine, de chimpanzé et de macaque, respectivement. Les bases soulignées représentent les sites de mutation.

L'analyse de l'apparition des triplets suit exactement la même procédure, la taille des locus étant simplement plus longue. Nous avons par ailleurs choisi de ne pas considérer l'apparition de triplets causée par la contraction éventuelle de quadruplets, car nos travaux sont basés sur la comparaison d'effets similaires entre micro-duplication et glissement, c'est-à-dire l'expansion. Enfin, le protocole aurait pu être appliqué pour évaluer l'apparition des doublons et triplets chez le chimpanzé, mais il semblerait que la séquence disponible pour cette espèce contienne encore de nombreuses erreurs de séquençage (L. Duret, communication personnelle). Une petite portion des mutations observables sur la séquence du chimpanzé sont donc artéfactuelles. Ne connaissant pas exactement le taux d'erreur, les conclusions tirées d'une analyse de l'apparition chez le chimpanzé pourrait être faussées. En revanche, ces erreurs ne gênent pas l'analyse des apparitions chez l'homme, qui implique l'identité des séquences du chimpanzé et du macaque aux sites étudiés, et il est très peu probable qu'une même erreur soit présente dans les deux séquences.

5.2.2 Calcul de la sur-représentation

L'analyse de l'alignement des doublons et triplets va nous permettre d'évaluer les modes d'apparition des doublons et des triplets, mais ne renseigne en rien sur leur influence sur le nombre de locus présents dans la séquence. Pour déterminer si les phénomènes de micro-duplication et glissement éventuellement observés jouent un rôle dans la densité de doublons et triplets, nous allons réaliser une analyse de sur-représentation par rapport à une densité attendue dans un génome aléatoire.

Comme nous l'avons vu en section 5.1.1, il est possible de calculer un nombre de locus attendus

aléatoirement par des formules théoriques, et ainsi déterminer un écart par rapport au nombre de locus observés. Le problème majeur de ces calculs théoriques est qu'ils ne permettent pas de calculer d'intervalle de confiance, il est donc impossible de déterminer si l'écart observé est significatif ou non. Or nous avons vu que les interprétations des résultats peuvent être radicalement différentes selon de degré de significativité obtenu. Dieringer et Schlötterer (2003) ont calculé un nombre de locus attendus en simulant un génome aléatoire et en comptant le nombre de microsatellites de chaque taille présent dans cette séquence. En répétant la procédure un certain nombre de fois, on obtient une distribution des nombres attendus, qui permet de calculer un nombre d'attendus moyen et un intervalle de confiance. Il ont effectué 250 simulations, pour avoir des intervalles de confiance à 5% très précis.

Notre analyse de la sur-représentation des doublons-triplets dans le génome humain sera basée sur la même méthode. Nous avons simulé 100 séquences aléatoires de la taille du chromosome 22, et calculé les nombres de doublons et triplets de chaque motif présents dans ces séquences. Le nombre moyen pour chaque motif sera utilisé comme attendu aléatoire et comparé au nombre effectivement détecté sur le chromosome 22. La répétition de 100 simulations s'est révélée suffisante pour obtenir de bons intervalles de confiance. Nous nous sommes restreints à l'analyse du chromosome 22 car les simulations et les analyses sont très gourmandes en temps de calcul, et l'analyse du génome complet n'aurait pu être réalisée dans le cadre de ma thèse. Néanmoins, la densité de microsatellites est équivalente pour tous les chromosomes humains (voir section 4.2.1), nous avons donc supposé que c'était le cas aussi pour les doublons et les triplets.

Il a ensuite fallu déterminer comment construire nos séquences aléatoires. En effet, une séquence aléatoire pure est une séquence dont la probabilité de chaque base azotée est de 1/4 pour chaque position, correspondant à un taux de GC de 50%. Or le taux de GC du génome humain est de 41%, signifiant que la probabilité d'obtenir deux A ou deux T adjacents est plus forte qu'avec un taux de GC de 50%. C'est pourquoi la majorité des analyses utilisant des attendus aléatoires prennent en compte la composition en GC de la séquence étudiée. Cette suggestion est en fait valable pour n'importe quel motif, les chances d'obtenir deux AC adjacents aléatoirement étant par exemple plus grandes si la proportion de AC dans le génome est plus importante que celle des autres dinucléotides. Rose et Falush (1998) ont pris en compte ce biais possible en calculant un attendu théorique à partir d'une formule qui tient compte de la proportion de chaque motif d'une taille donnée dans le génome.

Nous avons mené une analyse préliminaire pour évaluer si la prise en compte de la proportion de

chaque motif était nécessaire, ou si le taux de GC suffisait à obtenir des génomes aléatoires respectant la composition en motif. Les résultats de ces simulations montrent sans équivoque que ne gérer que la composition en GC ne permet pas de construire des génomes qui respectent les proportions de chaque motif (figure 5.6). Les nombres attendus de doublons et de triplets seront donc calculés par classe de motifs, à partir de génomes simulés respectant la proportion de chaque motif. Il faut remarquer que ces constructions sont dépendantes de la taille de motif analysée. Ainsi, les génomes aléatoires construits pour calculer de nombre de tétranucléotides ne respectera pas forcément la composition en motifs des trinucléotides, ce qui explique que l'on ait dû réaliser les analyses séparément pour chaque classe.

Nous avons montré dans le chapitre précédent que les éléments transposables, et plus particulièrement les séquences Alu chez l'homme, ont une influence non négligeable sur l'apparition des microsatellites. Etant donné leur nombre important dans le génome, et le caractère non aléatoire de leur séquence, les éléments transposables ont été retirés des analyses de sur-représentation. RepeatMasker a été exécuté sur la séquence 22 du chromosome humain, avec les paramètres par défaut, et la bibliothèque Repbase Update (version 9.11) complète. La taille des séquences simulées est donc égale au nombre de bases non masquées, et le nombre d'occurrences de chaque motif est celui obtenu à partir des régions non masquées. De la même manière, le compte des doublons et des triplets observés a été réalisé à partir des régions non masquées.

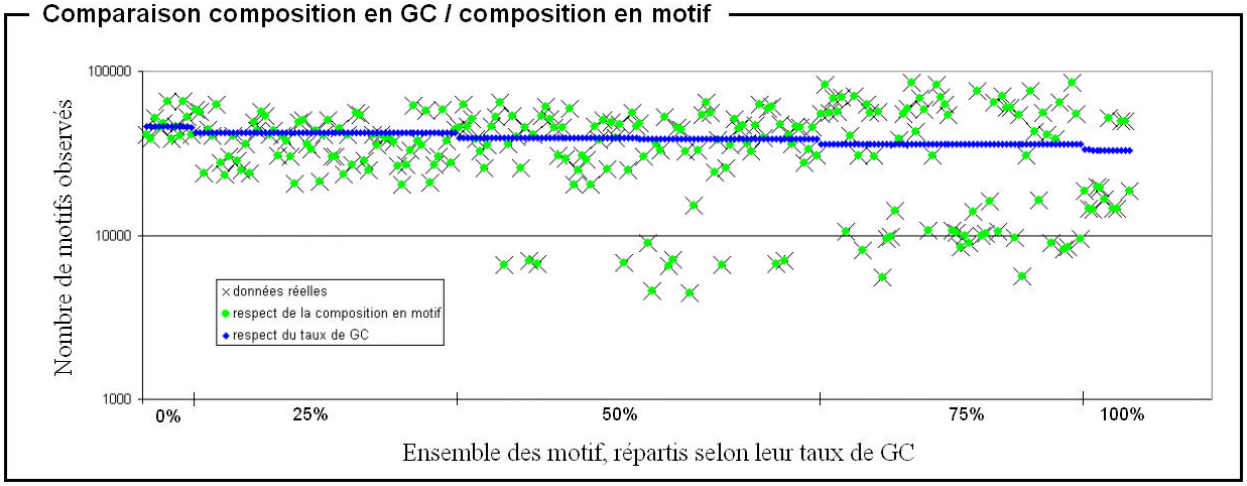


FIG. 5.6 – Densité moyenne des motifs tétranucléotides calculée à partir de 100 génomes aléatoires construits soit en respectant la composition en GC, soit en respectant la composition en motifs du chromosome 22 humain. Les intervalles de confiance sont donnés mais ne sont pas visibles sur la figure, car sont très réduits. La densité réelle est donnée à titre de comparaison.

5.2.3 Etude de la micro-duplication

Caractérisation de la micro-duplication

Nous avons en premier lieu étudié les événements de mutation qui ont concouru à l'apparition des doublons dans le génome humain. A partir de la comparaison des chromosomes 1 humain, de chimpanzé et de macaque, nous avons calculé que 82,9 % des doublons sont présents chez les trois primates, 1,2% ont disparu chez le chimpanzé, 1,6% sont apparus chez l'homme, et 14,3% ont pu disparaître chez le macaque ou bien apparaître chez l'ancêtre de l'homme et du chimpanzé (tableau 5.1). La similarité entre l'homme et le chimpanzé est donc de 97,2%, valeur conforme à celle généralement calculée pour des régions non soumises à sélection [Mikkelsen et al., 2005].

Les doublons sont apparus chez l'homme à 95,5% par substitution, à 2,8% par insertion et à 1,6% par délétion (tableau 5.1). Ces proportions sont toutefois différentes entre les classes de motifs, et le taux d'apparition par insertion est légèrement plus important pour les grandes tailles de motif, par rapport aux petites tailles (excepté les mononucléotides). Elle ne dépasse toutefois pas 5% du total des apparitions, sauf pour les doublons heptanucléotides, qui apparaissent à 12,6% par insertion. Le taux d'apparition par délétion augmente quant à lui régulièrement avec la taille du motif (de 1,1% pour les mononucléotides à 7,4% pour les heptanucléotides).

Résultats de l'alignement des doublons

classe de motif	nombre	proportion de l'alignement (%)				nombre d'apparitions	proportion des apparitions (%)		
		identique	indéterminé	disparition	apparition		substitution	insertion	délétion
Mono	21272257	86.5	11.2	1	1.3	207868	95.5	3.4	1.1
Di	4044283	73.9	22	1.8	2.3	73937	96.8	1.2	2.1
Tri	1434330	66.7	28.4	2.2	2.8	31009	95	2	3
Tetra	410799	59	35.5	2.4	3.1	9909	91.6	5	3.4
Penta	134003	52.9	41.1	2.5	3.5	3346	91.3	3.9	4.8
Hexa	54301	45.1	48.5	2.7	3.7	1447	90.7	3.9	5.4
Hepta	16145	40.8	53.1	2.6	3.5	420	80	12.6	7.4
Total	**27366118**	**82.9**	**14.3**	**1.2**	**1.6**	**327936**	**95.5**	**2.8**	**1.6**

TAB. 5.1 – Résultats de l'alignement des doublons du chromosome 1 humain sur les séquences de chimpanzé et de macaque.

Nous avons ensuite analysé plus en détail les insertions qui ont conduit à l'apparition de ces doublons, pour déterminer la proportion qui aurait pu être provoquée par une micro-duplication. Les résultats sont présentés dans la figure 5.7.1. Quelle que soit la classe de motif, la création d'un doublon par une insertion dont la longueur coïncide avec la taille du motif, que nous appellerons *insertion focale*, est toujours nettement majoritaire. Elle représente au minimum 31,5% des dinucléotides, au maximum 71,1% des tétranucléotides, en en moyenne 40,7% des apparitions par insertion. En comparaison, l'insertion d'une base unique cause en moyenne 19,6% des apparitions (mononu-

cléotides exclus). La proportion est réduite à 5,3% pour les insertions de 2 nt (dinucléotides exclus), et diminue jusqu'à 3% pour les insertions de 7 nt (tableau 5.2.1). On peut remarquer que les apparitions causées par l'insertion de plus de 7 nt sont aussi relativement importantes, de l'ordre de 12,5% du total des apparitions par insertion. Néanmoins, cette classe représente les insertions de taille 8, 9 et 10 nt (voir section 5.2.1), chacune contribue donc à environ 4%. Ce taux est à peu près équivalent à celui des autres tailles d'insertion.

Les insertions focales représentent les insertions de la taille du motif, donc une duplication de ce dernier. La proportion d'apparition par insertion focale est donc *a priori* la proportion d'apparition par micro-duplication. Toutefois, il n'est pas exclu que des insertions aléatoires puissent être (par chance !) de même motif que la région flanquant l'insertion. Ce type de mutation ponctuelle créerait donc une duplication du motif, comme dans le cas de la micro-duplication. Ces apparitions seront qualifiées d'apparition par *insertion basale*. La proportion par insertion focale (de la taille du motif) sera donc la proportion d'apparition par insertion basale (insertion aléatoire du motif) plus la proportion d'apparition par micro-duplication. Nous avons supposé que la probabilité d'apparition par insertion aléatoire d'une taille donnée est la même pour tous les motifs. Cela signifie que la proportion d'apparition par insertion basale pour une taille de motif est égale à la proportion d'apparition par insertion de cette même taille dans toutes les autres classes de motifs. Il devient donc possible d'évaluer la proportion d'apparition par micro-duplication en faisant la différence entre la proportion d'apparition par insertion focale et la proportion d'apparition par insertion basale (égale à la moyenne des apparitions de cette taille pour les autres classes de motifs ; voir tableau 5.2.1).

Les taux de micro-duplications calculés par cette méthode sont compris entre 20% des apparitions par insertion pour les micro-duplications de 1 base, et 65,8% pour les micro-duplications de 4 bases (tableau 5.2.1). Comme nous connaissons la contribution totale des insertions à l'apparition des doublons, nous sommes à même d'évaluer celle des micro-duplications. La micro-duplication est donc à l'origine de 1% du total des apparitions de doublons, et cette valeur est inférieure à 2% pour toutes les classes de motif, sauf pour les tétranucléotides (3,3%) et les heptanucléotides (7,7%).

A titre informatif, nous avons aussi analysé les délétions qui ont conduit à l'apparition de doublons. Notre méthode d'analyse des alignements permet d'exclure les doublons apparus par contraction d'une séquence répétée plus longue (voir section 5.2.1). Cela nous assure que les apparitions analysées sont bien issues de séquences non répétées. Les résultats montrent qu'à l'inverse des apparitions par insertion, la délétion focale (de la taille du motif) n'a pas d'importance particulière

(figure 5.7.2). Par contre les tailles de délétions ne sont pas équitablement réparties. Les délétions d'un nucléotide représentent entre 40 et 50% du total des apparitions par délétion, alors que les délétions de deux nucléotides ne représentent que 10 à 20%, et les suppressions plus de 5 nt, moins de 10% en tout.

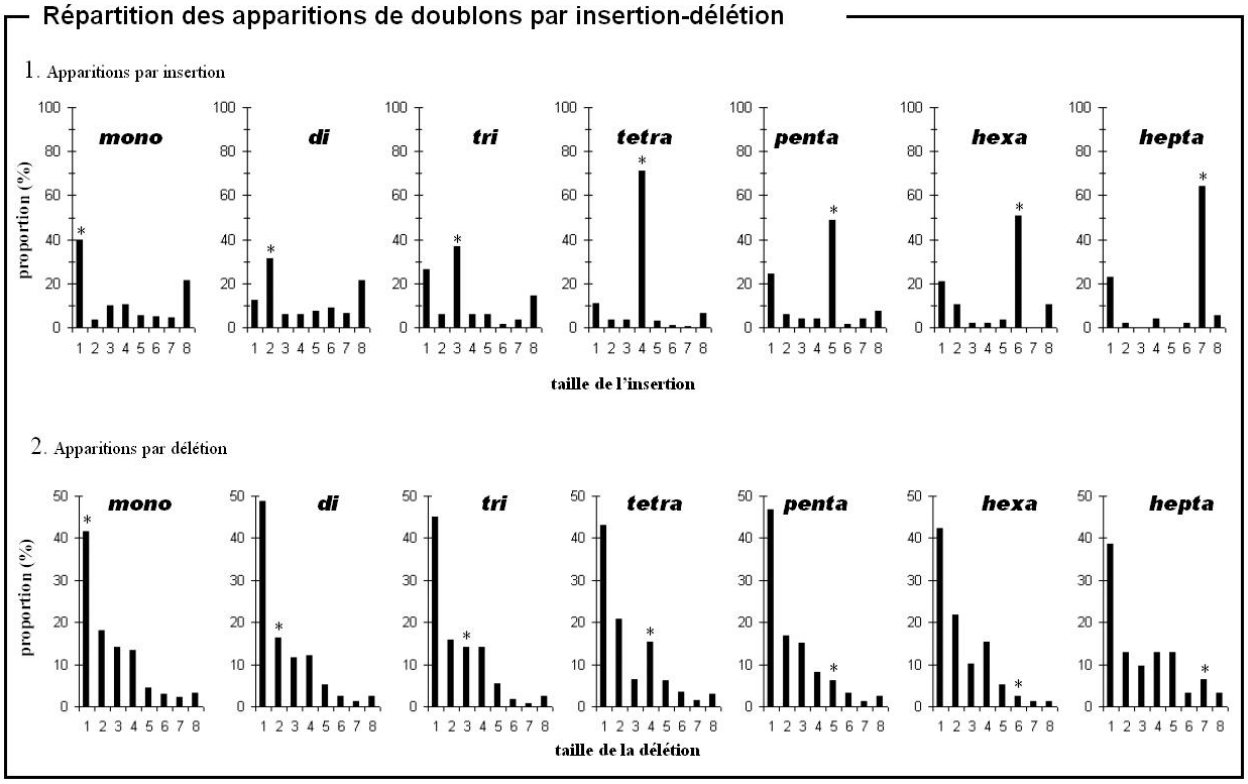


FIG. 5.7 – **1-** Répartition des apparitions par insertion pour chaque classe de motifs, en fonction du nombre de bases insérées. La classe 8+ représente les insertions entre 8 et 10 nt. **2-** Répartition des apparitions par délétion pour chaque classe de motifs, en fonction du nombre de bases supprimées. La classe 8+ représente les délétions entre 8 et 10 nt. Les insertions/délétions focales sont marquées d'une étoile.

Effet de la micro-duplication sur la sur-représentation des doublons

Les résultats des analyses d'alignement montrent que les micro-duplications contribuent à 1% du total des apparitions, même si cette proportion est plus importante pour les motifs les plus longs (7,8% des doublons heptanucléotides). L'effet de la micro-duplication reste donc mineur par rapport à celui de la mutation ponctuelle, mais il est néanmoins possible qu'elle soit suffisante pour créer une sur-représentation des doublons dans le génome. Nous avons testé cette possibilité en calculant le nombre attendu de doublons dans une séquence aléatoire, pour chaque motif, et en le comparant au nombre de doublons observés dans le génome humain.

Calcul des taux de micro-duplication et de glissement

1. Estimation de la proportion d'apparition par micro-duplication

	taille de l'insertion							totalité
	1	2	3	4	5	6	7	
doublons								
(1) insertion focale	39.6	31.5	36.6	71.1	48.9	50.9	64.2	40.7
(2) **insertion basale**	19.6	5.3	4.1	5.3	4.2	3.3	3	5.7
(3) **micro-dupl parmi insertions**	**20**	**26.2**	**32.5**	**65.8**	**44.6**	**47.5**	**61.1**	**35**
(4) apparitions par insertion	3.4	1.2	2	5	3.9	3.9	12.6	2.8
(5) **apparition par micro-duplication**	**0.7**	**0.3**	**0.6**	**3.3**	**1.7**	**1.9**	**7.7**	**1**

2. Estimation de la proportion d'apparition par glissement

triplets								
(6) insertion focale	39.2	52.7	70.4	78.7	69.1	66.3	79.3	50.2
(7) **glissement parmi insertions**	**-0.5**	**21.2**	**33.8**	**7.6**	**20.3**	**15.4**	**15.2**	**9.5**
(8) apparitions par insertion	3.9	2.1	3.9	8.4	9.7	12.2	24	4.4
(9) **apparition par glissement**	**0**	**0.4**	**1.3**	**0.6**	**2**	**1.9**	**3.6**	**0.4**

TAB. 5.2 – **1-** Estimation de la proportion d'apparition de doublons par micro-duplication. Le taux d'insertion focale est la proportion d'apparition de doublons ayant un motif de la taille insérée. Le taux d'insertion basale est la moyenne des proportions d'apparition de doublons ayant un motif de taille différente de la taille insérée. Le taux de micro-duplication parmi les insertions (ligne 3) est la différence entre les lignes 1 et 2, pour une taille d'insertion donnée. Le taux global d'apparition par micro-duplication (ligne 5) est le produit des lignes 3 et 4. La colonne totalité est la moyenne des autres colonnes, pondérée par leur nombre d'apparition. **2-** Estimation de la proportion d'apparition de triplets par glissement. Le taux de micro-duplication parmi les insertions (ligne 7) est la différence entre les lignes 6 et 1. Le taux global d'apparition par glissement (ligne 9) est le produit des lignes 8 et 7.

Les résultats pour les sept classes de motifs analysés sont présentés dans le tableau 5.3. Le chapitre 4 nous a montré que les éléments transposables pouvaient être vecteurs d'apparition de proto-microsatellites, et notamment de doublons. Or ces séquences ayant pour certaines un grand nombre de copies dans le génome, leur présence pourrait augmenter artificiellement le compte des doublons observés. C'est pourquoi la séquence a été masquée avant l'analyse, les résultats présentés sont donc exempts du biais éventuel causé par les éléments transposables.

Les rapports entre nombre de doublons observés et nombre d'attendus peuvent être répartis en deux classes, selon la taille du motif. Les motifs de petite taille accusent un déficit significatif de doublons observés d'environ 5% pour les mono et trinucléotides, mais de 30% pour les dinucléotide. Les motifs tétranucléotides et plus montrent au contraire une sur-représentation significative du nombre de doublons observés par rapport à ceux attendus. Il est à noter que la sur-représentation est d'autant plus forte que la taille du motif est grande, avec un ratio maximum de 2,5 pour les heptanucléotides.

Sur-représentation des doublons

	observé/attendu	nombre de motifs +	nombre de motifs normal	nombre de motifs -	taux de GC des motifs sur-représentés
Mono	0.94 $^{-}$	2	0	2	100
Di	0.71 $^{-}$	11	0	1	45.4
Tri	0.95 $^{-}$	47	11	2	47
Tetra	1.10 $^{+}$	156	80	4	44.3
Penta	1.48 $^{+}$	171	846	3	55.9
Hexa	1.87 $^{+}$	239	3780	1	51.9
Hepta	2.55 $^{+}$	230	16150	0	47.7
Total	-	**856**	**20867**	**13**	**48.7**

TAB. 5.3 – Rapport entre nombre de doublons observés dans le chromosome 22 humain et nombre d'attendus dans un génome aléatoire, nombre de motifs dont les doublons sont sur-représentés (+), normalement représentés (normal), sous-représentés (-), et composition en GC pour l'ensemble des motifs sur-représentés, par classe de motif. La significativité du rapport a été calculée à partir de 100 génomes aléatoires. $^{+}$: sur-représentation significative des doublons observés. $^{-}$: sous-représentation significative des doublons observés.

La sur-représentation motif par motif a ensuite été calculée, pour savoir si tous les motifs étaient concernés de la même manière. Assez peu de motifs sont en fait sur-représentés (table 5.3), et cela dépend de leur taille. La grande majorité des motifs di et trinucléotides sont par exemple sur-représentés, ce qui est assez inattendu puisque ces classes de motif sont sous-représentés dans leur globalité. Les motifs tétranucléotides sont eux-aussi pour la plupart sur-représentés, mais la proportion décroît rapidement avec la taille des motifs, pour ne représenter que 1,4% des heptanucléotides. Les mononucléotides représentent un cas particulier, puisque 50% (les A et les T) sont sous-représentés dans le génome, et les deux autres motifs sont sur-représentés. D'autres motifs sont aussi sous-représentés, mais leur nombre est très faible, et ils ne jouent pas sur le taux global de leur classe de motifs. La composition en GC des motifs sur-représentés à été évaluée (tableau 5.3), et donne des taux très proches de 50% de GC pour toutes les classes de motifs (excepté les mononucléotides). Il semblerait donc que la sur-représentation ne soit pas dépendante de la composition en GC des motifs.

Le rapport nombre observé/nombre attendu motif par motif indique de plus que la sur-représentation est généralement assez faible pour chacun des motifs (figures 5.9 à 5.12). En effet, la densité

de doublons pour les motifs sur-représentés atteint très rarement le double de la densité attendue lorsque cette dernière est inférieure à 50 locus par 10 Mb. Par contre, si la densité attendue est très faible (par exemple pour les motifs hexa et heptanucléotides), le rapport observé/attendu peut être très important. Ces fortes sur-représentations ne sont toutefois causées que par un petit nombre de locus supplémentaires. On peut noter, pour les pentanucléotides et heptanucléotides, que certains motifs ont une sur-représentation nettement plus importante que les autres. Ces motifs sont les ATATA/TATAT et CACAC/GTGTG pour les pentanucléotides, et les ATATATA/TATATAT et ACACACA/TGTGTGT. Ils ont probablement été créés par la délétion d'une base unique dans des microsatellites de type $(xA)_{5,5}$ et $(xA)_{7,5}$, et ne sont par conséquent pas issus du processus de micro-duplication. La sur-représentation importante des doublons TGGGG/CCCCA n'est par contre pas expliquée.

Pour résumer, nous observons une sur-représentation à deux niveaux dans le génome humain, en fonction de la taille du motif. Les motifs trinucléotides et inférieurs sont sous-représentés de manière générale, bien que les motifs soient un à un sur-représentés. Ce phénomène n'est pas expliqué, mais pourrait provenir de la méthode de test de significativité par simulation. Les motifs tétranucléotides et supérieurs sont quant à eux sur-représentés, même si cette sur-représentation ne concerne que quelques motifs. Selon nos hypothèses, cela signifie que la micro-duplication participe de manière non négligeable au nombre de locus microsatellites de période égale ou supérieure à 4 présents dans le génome humain, mais n'a que peu d'effet sur le nombre de ceux de période plus petite.

5.2.4 Etude du glissement pour les répétitions de petite taille

Estimation du taux de glissement

Les estimations du taux de micro-duplication calculées dans la section précédente vont nous servir à déterminer si les répétitions de très petite taille peuvent subir des événements de glissement. Nous allons pour cela analyser l'apparition des triplets, qui peut être provoquée par des mutations ponctuelles, par des micro-duplications, ou par glissement (figure 5.3).

Les résultats de l'alignement des triplets (figure 5.4) donnent un taux de locus identiques entre les trois espèces légèrement inférieur à celui obtenu pour les doublons (78,4% contre 82,9%). Les locus triplets étant par définition plus longs que les doublons, ils ont une probabilité plus forte de subir des mutations depuis la divergence homme/chimpanzé-macaque. Cette proportion moindre d'identité était donc attendue. De plus, le surplus de mutation a eu lieu en majorité avant la spéciation entre homme et chimpanzé (donnant une proportion importante de cas indéterminés). Cette

observation est là encore conforme avec le fait que l'âge de spéciation entre ces deux espèces (5 Ma) est beaucoup plus faible que l'âge de spéciation d'avec le macaque (30 Ma). Les événements de disparition et d'apparition sont quant à eux dans des proportions équivalentes à celles observées pour les doublons, mais le taux d'apparition par substitution est légèrement moins important (93,8%). Les apparitions par insertion et par délétion ont augmenté de manière équivalente pour les motifs de petite taille, mais les motifs de grande taille accusent un surplus d'apparition par insertion, par rapport aux doublons (entre 2 et 3 fois plus).

Résultats de l'alignement des triplets

		proportion de l'alignement (%)				nombre d'apparitions	proportion des apparitions (%)		
classe de motif	nombre	identique	indéterminé	disparition	apparition		substitution	insertion	délétion
Mono	6948280	83.5	13.8	1	1.6	71692	94.4	3.9	1.7
Di	1533809	67.7	27.5	2	2.7	31277	95.1	2.1	2.9
Tri	598540	60.8	33.7	2.3	3.3	13687	92.5	3.9	3.6
Tetra	206167	53	41.2	2.4	3.4	4957	87.1	8.4	4.5
Penta	67702	47.2	46.6	2.5	3.7	1674	84.5	9.7	5.9
Hexa	30780	39.6	54.4	2.3	3.7	706	82.3	12.2	5.5
Hepta	10582	32	62.4	2.3	3.3	242	70.2	24	5.8
Total	**9395860**	**78.4**	**18.3**	**1.3**	**2**	**124235**	**93.8**	**3.8**	**2.4**

TAB. 5.4 – Résultats de l'alignement des triplets du chromosome 1 humain sur les séquences de chimpanzé et de macaque.

Nous avons ensuite examiné les apparitions par insertion, pour déterminer si ce surplus pouvait être causé par du glissement. Les insertions focales sont là encore très majoritaires et représentent 50,2% du total des apparitions par insertion. Cette proportion est variable selon la classe de motifs, mais est à chaque fois plus importante qu'elle ne l'était pour les doublons. Seuls les mononucléotides possèdent un taux d'apparition par insertion focale similaire pour les doublons et les triplets (entre 39 et 40% des apparitions par insertion). Ce taux plus important signifie, selon nos hypothèses, que la micro-duplication n'est pas la seule à favoriser les apparitions par insertion focales, donc que le glissement est à l'origine d'une partie des apparitions de triplets.

Si l'on suppose que le le taux de micro-duplication est le même pour les doublons et les triplets, la contribution du glissement au total des apparitions peut être calculée, à partir du taux de micro-duplication précédemment estimé. Les résultats, présentés dans le tableau 5.2.2, indiquent que 9,5% des apparitions par insertion sont causées par le glissement, soit 0,4% des apparitions totales. Ce taux est variable entre les classes de motifs, allant de 0,4% pour les dinucléotides à 3,6% pour les heptanucléotides. Il est de plus intéressant de remarquer que les mononucléotides ne semblent pas affectés par les événements de glissement (taux d'apparition par glissement à 0%).

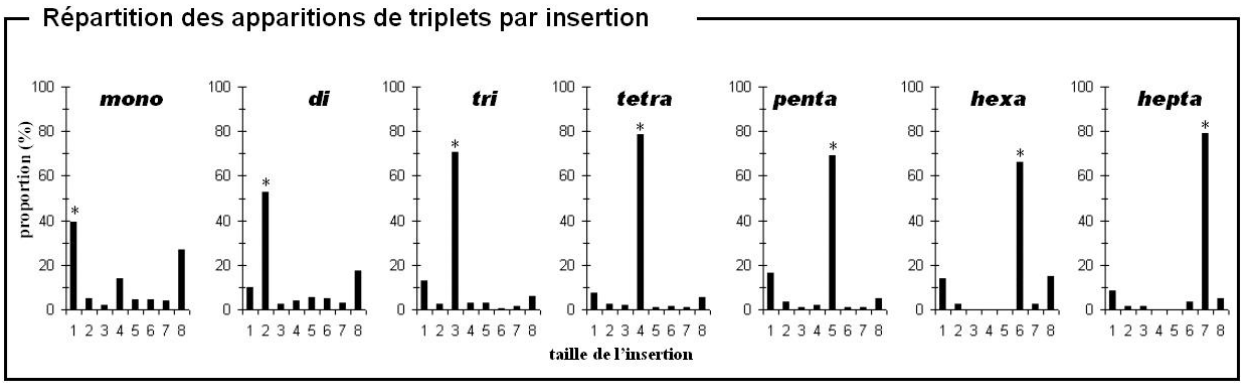


FIG. 5.8 – Répartition des apparitions par insertion pour chaque classe de motif, en fonction du nombre de bases insérées. La classe 8+ représente les insertions entre 8 et 10 nt. Les insertions focales sont marquées d'une étoile.

Sur-représentation causée par le glissement

Nous avons finalement évalué de taux de sur-représentation des triplets dans le génome humain, pour déterminer si le glissement observé avait un impact sur le nombre de locus. Par rapport à la présence des doublons, nous pouvons déjà remarquer que toutes les classes de motifs sont sur-représentées, à l'exception des dinucléotides (tableau 5.5). Le clivage entre petites et grandes tailles de motifs a donc disparu. L'excédent est de plus assez important, notamment pour les mononucléotides qui sont 2,24 fois plus nombreux que l'attendu, et les pentanucléotides ou supérieurs tous au moins deux fois plus nombreux. Le nombre de motifs sur-représentés a lui aussi nettement augmenté par rapport aux doublons, et ce pour toutes les classes. Quasiment tous les motifs mono à tétranucléotides sont sur-représentés (à l'exception des GC, CGAA et des TCGA), ainsi qu'une grande majorité des pentanucléotides. Le ratio observé/attendu motif par motif n'est pas très informatif, sinon qu'il montre un plus grand nombre de motifs sur-représentés, sans pour autant donner des valeurs de sur-représentation plus importantes que pour les doublons (données non montrées). On peut toutefois constater que le seul motif sous-représenté, le motif GC, accuse un déficit très faible par rapport à l'attendu aléatoire.

Il aurait été intéressant d'évaluer la part réelle du glissement dans ces sur-représentations, c'est-à-dire en corrigeant l'effet de la micro-duplication, mais nous n'avons pu mettre la méthode au point, par manque de temps. En effet, les triplets étant des séquences plus longues que les doublons, leur probabilité d'apparition aléatoire est plus faible. Dans ces cas là, il est tout à fait possible que le taux de micro-duplication (que nous avons supposé constant entre les doublons et les triplets) puisse suffire à produire l'excédent de sur-représentation observé.

Sur-représentation des triplets

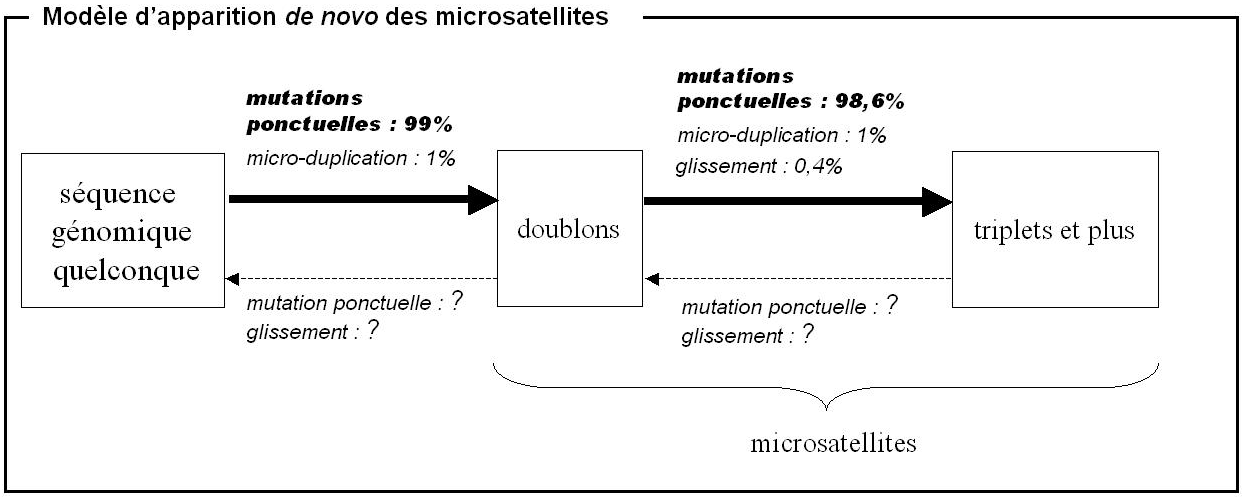
	observé/attendu	nombre de motifs +	normal	-	taux de GC des motifs sur-représentés
Mono	2.24 $^{+}$	4	0	0	50
Di	0.83 $^{-}$	11	0	1	45.4
Tri	1.15 $^{+}$	60	0	0	50
Tetra	1.58 $^{+}$	238	2	0	50
Penta	1.98 $^{+}$	829	191	0	49
Hexa	2.5 $^{+}$	1613	2407	0	49.2
Hepta	3.07 $^{+}$	1275	15105	0	49.5
Total	-	**4030**	**17705**	**1**	**49**

TAB. 5.5 – Rapport entre le nombre de triplets observés dans le chromosome 22 humain et le nombre d'attendus dans un génome aléatoire, nombre de motifs dont les triplets sont sur-représentés (+), normalement représentés (normal), sous-représentés (-), et composition en GC pour l'ensemble des motifs sur-représentés, par classe de motif. La significativité du rapport a été calculée à partir de 100 génomes aléatoires. $^{+}$: sur-représentation significative des triplets observés. $^{-}$: sous-représentation significative des triplets observés.

5.3 Discussion

Nous avons démontré, dans ce chapitre, que des événements de glissement pouvaient produire des triplets à partir de doublons (à part pour les mononucléotides), donc que ce mécanisme n'était pas contraint par une taille limite. Si l'on considère la capacité de glissement comme une condition suffisante, nous pouvons donc définir un microsatellite comme toute séquence répétée au minimum deux fois. Ne pas poser de limite de taille minimum à la définition des microsatellites permet de grandement simplifier leur dynamique d'apparition. La naissance d'un microsatellite est dans ces cas là impérativement restreinte à la création du doublon original, qui ensuite peut se développer et donner un microsatellite de taille plus importante. Les travaux que nous avons réalisés présentent deux forces distinctes pour la création de doublons, donc de microsatellites. La plus importante est la mutation ponctuelle, qui créé aléatoirement des répétitions en tandem, le plus souvent de petite taille. Cette force a toujours été considérée comme le processus majeur d'apparition *de novo* mais nous avons ici pu quantifier la proportion de doublons créés de cette manière, qui s'élève à 99%. Les 1% d'apparitions restantes sont la conséquence d'un mécanisme bien moins connu, que nous avons nommé micro-duplication (figure 5.6).

Le doublon créé va ensuite pouvoir se développer pour atteindre un nombre de répétitions plus important. Nous nous sommes focalisés ici sur l'expansion de deux à trois répétitions, et avons observé que trois forces étaient en jeu : la mutation ponctuelle, la micro-duplication et le glissement de polymérase. La mutation ponctuelle reste encore le facteur le plus important, causant 98,6% des triplets. Nous avons supposé que la micro-duplication gardait un taux constant à 1%, nous permettant d'évaluer le taux de glissement à 0,4% (figure 5.6). Ces proportions ne sont bien sûr valables que pour l'apparition des triplets, et il est probable que l'importance de la mutation ponctuelle se réduise avec le nombre de répétitions, tandis que celle du glissement augmente.



TAB. 5.6 – Modèle d'apparition et de développement des doublons, à partir d'une séquence génomique quelconque. Le principal facteur d'apparition est la mutation ponctuelle, la micro-duplication ne jouant que faiblement. Le principal facteur d'expansion en triplet est aussi la mutation ponctuelle. Les cas de contraction et de disparition des doublons et triplets (flèches tirées) n'ont pas été étudiés dans ces travaux, mais pourraient être provoqués par glissement et par mutation ponctuelle.

Zhu *et al.* (2000) avaient déjà estimé la proportion de doublons produits par micro-duplication, mais leurs valeurs étaient nettement plus élevées que les nôtres. En effet, pour eux, 1,7% des di, 2,5% des tri et 18,5% des doublons tétranucléotides avaient cette origine, alors que nous trouvons des taux respectifs de 0,3%, 0,6% et 3,3%. Zhu *et al.* (2000) ont estimé ces valeurs à partir de mutations extraites de la base de données Human Gene Mutation Database, qui recense les mutations causant des maladies génétiques chez l'homme [Krawczak and Cooper, 1997]. L'avantage d'utiliser cette base est que la séquence non mutée peut être déterminée sans ambiguïté (c'est la séquence du gène sain). Le désavantage est qu'elle ne concerne que des gènes, et tous les types de mutations n'ont pas la même valeur dans ces régions. Les indels de taille non multiple de 3 produisent par

exemple un décalage du cadre de lecture, amenant à un dysfonctionnement systématique du gène. En revanche, les substitutions ne provoquent le changement que d'un acide aminé au maximum, ou même zéro dans le cas des mutations synonymes [Duret, 2002]. Il est donc évident que l'impact des substitutions sur le fonctionnement des gènes est bien moindre que celle des indels, signifiant que le nombre d'événements de substitutions observables dans la base HGMD est inférieur à celui qui a réellement lieu. Les données de Zhu *et al.* sont donc biaisées en faveur des indels. Notre méthode de génomique comparative donne par contre une représentation non biaisée des événements de mutation, puisque basée sur l'ensemble du chromosome 1, ce qui suppose que les taux d'apparition par micro-duplication calculés ici sont probablement plus proches de la réalité.

Nos calculs de sur-représentation indiquent que le nombre de doublons de période supérieure à trois est plus important que l'attendu aléatoire. Nous avons expliqué cela par l'effet de la micro-duplication qui produit systématiquement des doublons. Par contre, pour les motifs de plus petite taille, nous observons une sous-représentation générale des doublons. Ce phénomène avait déjà été remarqué par Dieringer et Schlötterer (2003), pour les mono à tétranucléotides. Le modèle théorique qu'ils ont mis en place leur a permis d'expliquer cette sous-représentation grâce à un paramètre nommé *indel-slippage*, correspondant à notre micro-duplication. Leur paramètre autorise toutefois autant le gain que la perte d'une répétition, copiant en cela le mécanisme du glissement. La sous-représentation des petits locus obtenue avec leur modèle est donc la conséquence de la disparition par contraction de leurs doublons.

Nous avons pour notre part considéré que la micro-duplication, à l'inverse du glissement, ne peut qu'insérer de nouvelles répétitions. Nous avons par contre montré que le glissement est capable de produire des triplets à partir des doublons, et bien que le biais vers les expansions soit avéré, il est possible qu'il puisse provoquer la disparition de certains doublons par contraction. Il est toutefois assez peu probable, étant donné les taux de glissement estimés, que ce processus soit assez puissant pour causer la sous-représentation des motifs courts, contrecarrant l'effet de la micro-duplication.

De manière surprenante, nous avons observé que le rapport observés/attendus était significativement supérieur à 1 pour la grande majorité des motifs de ces classes sous-représentées. De plus, le peu de motifs sous-représentés ne le sont que très faiblement (voir les figures 5.9 à 5.12), ils n'ont donc pu à eux seuls provoquer la sous-représentation de la classe de motif en son entier. Cette incohérence n'a *a priori* pas de signification biologique, et il est probable qu'elle soit la conséquence d'un problème dans les tests statistiques. En effet, nous avons supposé que les proportions de doublons de chaque motif sont indépendantes dans nos génomes aléatoires. Or la probabilité d'avoir des dou-

blons, disons riches en A/T, a nécessairement une influence sur la probabilité d'avoir des doublons riches en G/C. La construction des génomes aléatoires respectant la proportion de chaque motif est censée corriger ce genre de biais, mais il est possible que d'autres facteurs non pris en compte jouent sur ces rapports entre motifs. Une plus ample réflexion sera nécessaire pour déterminer la cause de cette incohérence.

Quelle est l'origine de ces micro-duplications ? Nous ne pouvons pour l'instant pas répondre à cette question, car le protocole d'étude utilisé n'est pas expérimental et ne permet pas de tester des hypothèses sur les processus moléculaires en jeu. A partir de nos résultats de sur-représentation nous pouvons néanmoins dresser un profil général du mécanisme. Tout d'abord, toutes les classes de motif sont affectées, et la composition en GC des motifs ne semble pas jouer de rôle, étant donné que la moyenne des compositions en GC des motifs sur-représentés est de 50% environ, quelle que soit leur taille. L'ensemble des motifs dinucléotides sont sur-représentés, mais les motifs de grande taille sont au contraire assez rarement sur-représentés. Lorsque peu de motifs sont sur-représentés, ceux-ci n'ont pas de rapport entre eux (sauf les cas exceptionnels des CACAC, TATAT et CACACAC). Le processus n'est donc *a priori* pas spécifique à certains types de motifs particuliers, ce qui confirme les résultats de Dieringer et Schlötterer (2003). Ces différentes caractéristiques semblent donc indiquer que la micro-duplication est un phénomène aléatoire, pouvant créer des microsatellites à partir de n'importe quelle séquence génomique.

Seuls Zhu *et al.* (2000) ont émis une hypothèse sur le mécanisme moléculaire provoquant la micro-duplication. Ils proposent que le glissement soit possible, même sans répétition préalable, et que les premières répétitions en tandem en soit issues. Le glissement nécessitant un ré-appariement décalé, de tels événements dans une séquence dénuée de répétition doivent être relativement rares. Par contre, dès qu'une répétition est présente, les chances de glissement devraient nettement augmenter. Nos résultats montrent qu'à l'inverse, la différence du nombre d'apparitions par insertion focale entre les doublons (sans appariement erroné possible) et les triplets (avec appariement erroné possible) est très faible. C'est pourquoi nous pensons que les deux mécanismes dont distincts, et des recherches complémentaires seront nécessaires pour déterminer la nature de la micro-duplication.

Nos travaux sur le glissement des petites répétitions ne sont que préliminaires, mais démontrent que des événements de glissement sont à l'origine de l'apparition de nouveaux triplets. En revanche, nous n'avons pas pu déterminer l'influence de ce phénomène sur la sur-représentation des triplets. Notre définition des triplets comprend les locus possédant entre deux et trois répétitions, donc issus de l'éventuel glissement de locus ayant moins de deux répétitions. Or nous ne savons dans quelle mesure ce type d'événement est possible. Cela revient encore à se demander quelle est la taille mini-

mum pour le glissement ; il aurait peut-être été préférable de séparer les triplets en deux catégories, l'une avec trois répétitions et l'autre avec moins de trois. De manière plus générale, il pourrait être intéressant d'évaluer ce taux de glissement pour les locus avec un nombre de répétitions plus important, qui permettrait de déterminer si le glissement dépend de la taille du locus, même pour les très petites tailles. Enfin, le glissement peut aussi produire des contractions. Ces événements, même s'ils n'ont pas été traités ici, font partie intégrante du processus, et pourraient jouer sur les taux calculés, de même que sur les taux de sur-représentation (figure 5.6). L'analyse des disparitions de doublons chez le chimpanzé pourrait *a priori* nous informer sur ce taux de contraction, mais ces études n'ont à l'heure actuelle pas été réalisées.

Pour conclure, nous pouvons remarquer que notre étude repose sur les doublons et triplets déjà existants dans le génome humain. Cette méthode a l'avantage d'être simple à mettre en œuvre et nous a déjà permis d'évaluer la faible importance de la micro-duplication et du glissement par rapport à la mutation ponctuelle. Elle souffre néanmoins d'une lacune majeure qui est qu'elle ne permet pas de calculer des taux de mutation pour ces mécanismes. Nous avons en effet accès uniquement aux événements qui ont effectivement donné lieu à la création d'un doublon et d'un triplet, et non à tous les événements qui auraient pu avoir lieu. Or le taux de mutation est justement le rapport entre le nombre d'événements observés et le nombre potentiel.

Une méthode pour calculer ces taux de mutation sera donc mise en place prochainement, afin de compléter les travaux présentés ici. Elle reposera là encore sur l'alignement des trois génomes utilisés ici, dans lesquels tous les sites possibles d'apparition de doublons/triplets dans le génome humain seront comptés, selon le type d'apparition (substitution, insertion, délétion, duplication). Ces sites nous donnerons alors le nombre d'apparitions potentielles, nous permettant de calculer des taux de mutation pour l'apparition des doublons. Nous pourrons dès lors faire des comparaisons entre le taux de mutation ponctuelle créant des doublons par rapport à celui calculé pour l'ensemble du génome, ou déterminer si les taux de micro-duplications et glissements sont significativement plus importants que les taux normaux d'insertion.

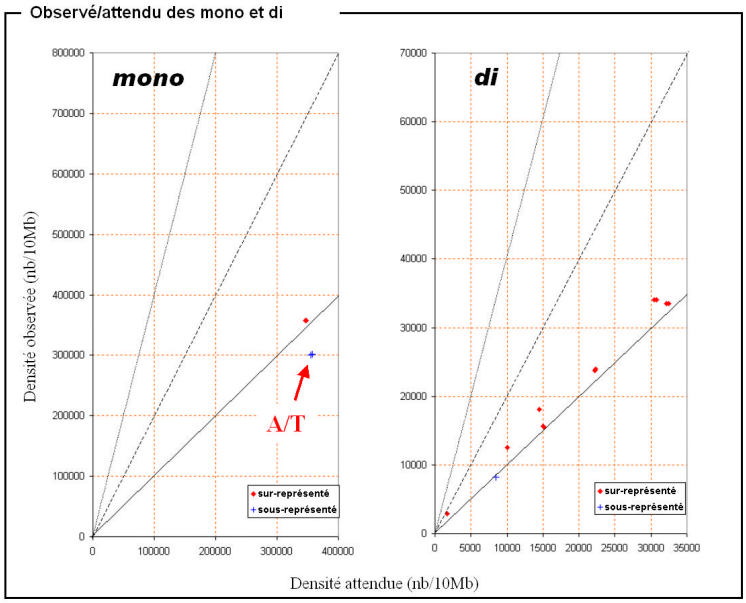


FIG. 5.9 – Densité de doublons observés par rapport à la densité de doublons attendus en nombre de locus par 10 Mb, pour les mono et dinucléotides. points : motifs sur-représentés. cercles : motifs normalement représentés. croix : motifs sous-représentés. Ligne pleine : ratio 1 :1, ligne tirée : ratio 2 :1, ligne pointillée : ratio 4 :1.

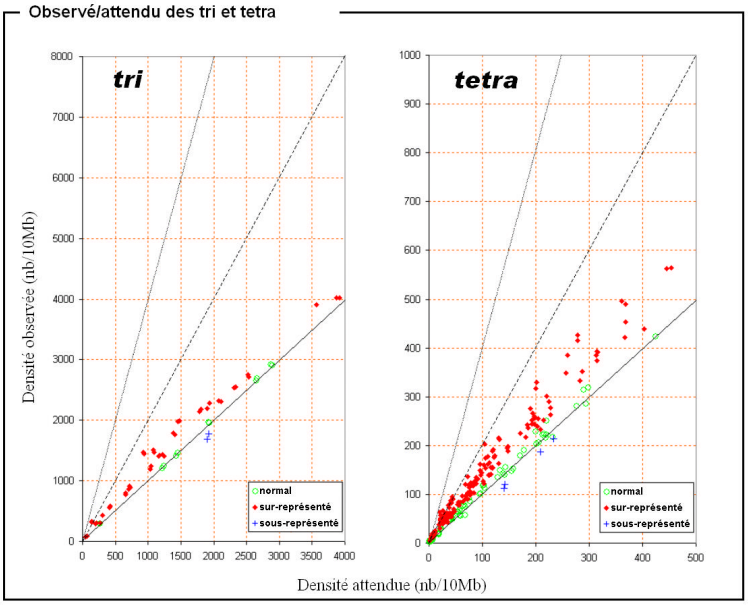


FIG. 5.10 – Densité de doublons observés par rapport à la densité de doublons attendus en nombre de locus par 10 Mb, pour les tri et tétranucléotides. points : motifs sur-représentés. cercles : motifs normalement représentés. croix : motifs sous-représentés. Ligne pleine : ratio 1 :1, ligne tirée : ratio 2 :1, ligne pointillée : ratio 4 :1.

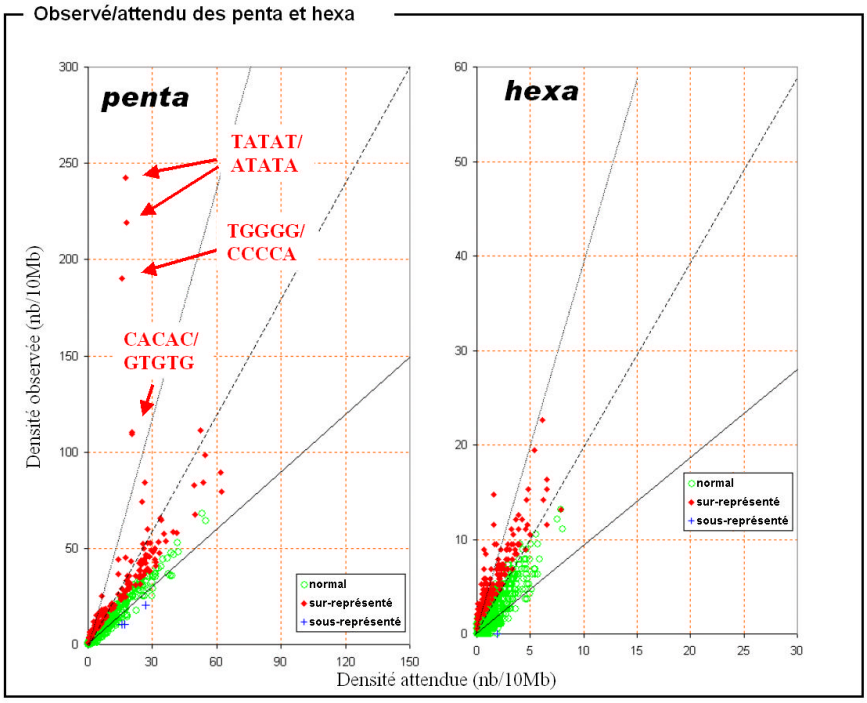


FIG. 5.11 – Densité de doublons observés par rapport à la densité de doublons attendus en nombre de locus par 10 Mb, pour les penta et hexanucléotides. points : motifs sur-représentés. cercles : motifs normalement représentés. croix : motifs sous-représentés. Ligne pleine : ratio 1 :1, ligne tirée : ratio 2 :1, ligne pointillée : ratio 4 :1.

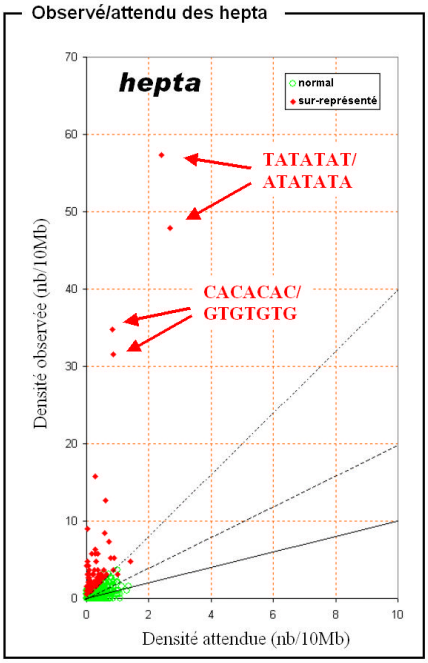


FIG. 5.12 – Densité de doublons observés par rapport à la densité de doublons attendus en nombre de locus par 10 Mb, pour les heptanucléotides. points : motifs sur-représentés. cercles : motifs normalement représentés. croix : motifs sous-représentés. Ligne pleine : ratio 1 :1, ligne tirée : ratio 2 :1, ligne pointillée : ratio 4 :1.

Chapitre 6

Dissertation et conclusion

A la suite des travaux présentés dans les chapitres 4 et 5, un modèle général d'apparition des microsatellites peut être proposé. Il est détaillé dans la section suivante et présenté dans la figure 6.1. Nous y décrivons l'inclusion d'un nouveau mécanisme d'apparition nommé micro-duplication, la possibilité d'un glissement de polymérase quel que soit le nombre de répétitions, et la manière dont les éléments transposables peuvent participer à la création de nouveaux microsatellites. Les implications de ce modèle seront ensuite débattues, tant au niveau de la distribution des microsatellites dans les génomes, qu'au niveau de la gestion de l'apparition dans les modèles théoriques de dynamique des microsatellites.

6.1 Un modèle d'apparition des microsatellites

Le modèle présenté ici concerne les deux premières phases du cycle de vie des microsatellites, qui sont la naissance et l'expansion (que nous appellerons développement), proposés dans le modèle de Buschiazzo et Gemmel (2006 ; voir figure 2.4), en y apportant des données quantitatives. Quatre forces distinctes agissent sur l'apparition et le développement des microsatellites, avec plus ou moins d'importance selon la phase. La première est la mutation ponctuelle, qui génère des motifs via des substitutions ou des indels aléatoires. Si le motif créé est adjacent à un motif identique, elle produit une répétition. La deuxième est la micro-duplication, qui copie en tandem un fragment de séquence de manière aléatoire. Ce mécanisme produit de fait une répétition. La troisième est le glissement de polymérase, qui permet d'ajouter une répétition lorsque des motifs adjacents sont déjà présents. Enfin, l'insertion d'éléments transposables génère, pour chaque copie, toutes les répétitions déjà présentes dans la séquence insérée. On parlera alors de microsatellites adoptés, qui peuvent être des doublons ou des microsatellites de taille plus importante.

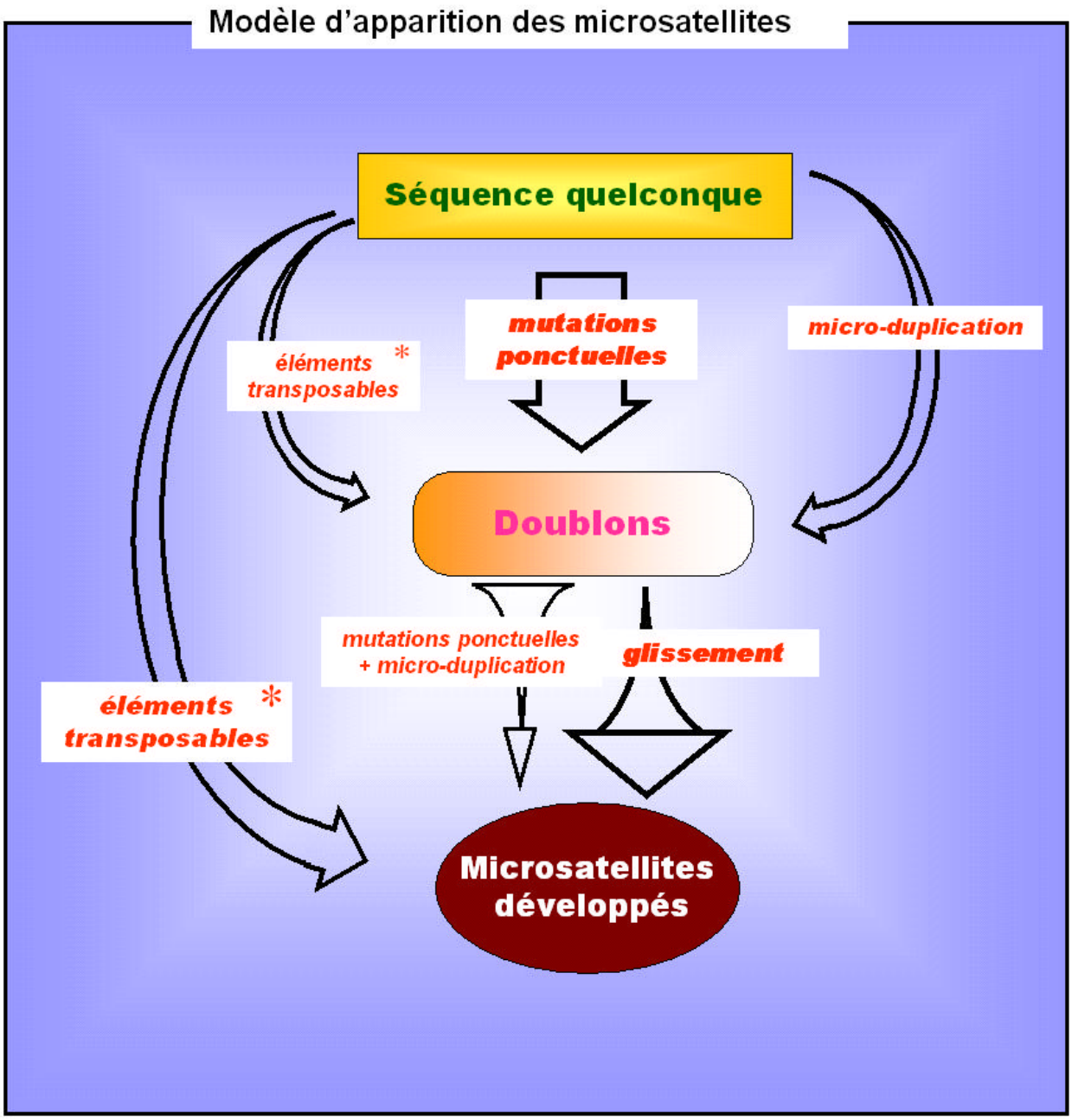


FIG. 6.1 – Modèle général d'apparition des microsatellites. La largeur des flèches symbolise l'importance relative du mécanisme. (*) : processus qui ne permet l'apparition que de certains motifs. La mutation ponctuelle est la force principale pour créer des doublons, la micro-duplication jouant un rôle moindre. Les doublons apparus via les éléments transposables sont limités à ceux présents dans la séquence de l'élément. Les microsatellites développés possèdent plus de deux répétitions, et apparaissent à partir de doublons, soit par mutation ponctuelle, soit par micro-duplication, soit par glissement. Le glissement est d'autant plus important que le microsatellite comporte de répétitions, et c'est l'inverse pour les mutations ponctuelles et la micro-duplication. Les éléments transposables permettent l'apparition à une taille donnée, mais pour quelques classes de motifs seulement, en fonction de la séquence de l'élément. Ce modèle représente uniquement les mécanismes qui créent et développent les microsatellites, sans intégrer ceux qui les réduisent et les détruisent (contraction, mutation ponctuelle, délétion, etc.).

6.1.1 Apparition des doublons

La phase de naissance correspond à la création d'un doublon (deux répétitions adjacentes) à partir d'une séquence dénuée de répétition. Les mécanismes qui entrent en jeu dans cette phase sont la mutation ponctuelle, la micro-duplication (voir chapitre 5) et l'insertion d'éléments transposables (voir chapitre 4). Ils possèdent tous des contraintes diverses que nous nous proposons de détailler ici.

Les doublons produits par la mutation ponctuelle sont dépendants de la séquence d'origine. En effet, la substitution ne provoque le changement que d'une base (ou plus rarement de deux [Podlutsky et al., 1998]) à la fois. Les autres bases du doublon doivent donc être déjà présentes. L'insertion nécessite aussi d'avoir un fragment du doublon déjà présent. Dans le cas d'une apparition par délétion ponctuelle, toutes les bases du doublon existent déjà, mais sont séparées par le fragment supprimé. Les différents doublons qui peuvent apparaître par la mutation ponctuelle à une position donnée sont donc très limités, et dépendent de la séquence. De plus, la probabilité d'apparition est relative à la taille du motif, les motifs longs nécessitant plus de bases préexistantes. L'apparition par micro-duplication dépend aussi de la séquence d'origine, car elle ne crée un doublon qu'à partir d'un motif déjà présent. En revanche, elle peut produire un doublon de n'importe quelle période à n'importe quelle position.

Ces deux types de mutations étant contraints par la séquence au niveau local, elles ne peuvent produire que certains types de doublons à une position donnée. Elles sont par contre génératrices de l'ensemble des types de motifs si l'on considère la séquence génomique en son entier.

Les doublons issus d'éléments transposables, quant à eux, ne dépendent en théorie pas du tout de la séquence qui existait à la position de l'insertion. La situation est plus complexe dans la réalité puisque nous avons vu que l'insertion ne peut se faire qu'à certaines positions précises, du moins pour les éléments Alu. En revanche, les doublons générés sont ceux existants dans la séquence de la copie insérée, et dépendent donc de la séquence du gène maître.

La distribution des doublons dans le génome en termes de motifs dépend donc énormément des taux relatifs de chacun de ces processus de mutation. La mutation ponctuelle produit une distribution conforme à celle attendue dans un génome aléatoire, pour une composition nucléotidique donnée. La proportion de doublons de motifs courts sera notamment supérieure à celle de doublons de motifs longs. Le taux de micro-duplication crée quant à lui un surplus de doublons, sans *a priori* de motif. Enfin, l'insertion d'éléments transposables crée un surplus de certains motifs, sans *a priori* de période.

Les résultats du chapitre 5 indiquent que la mutation ponctuelle est le processus majeur de création de doublons, puisqu'elle est à l'origine de 99% des apparitions, si l'on ne considère pas les doublons adoptés. Le 1% restant est produit par micro-duplication. Par contre, le calcul de la sur-représentation nous a montré que la micro-duplication participe tout de même significativement à la quantité de doublons présents dans le génome humain, du moins pour les motifs de taille importante (supérieur à trois). Nous n'avons pas calculé le nombre de doublons adoptés, ce qui nous empêche d'évaluer dans quelle mesure les éléments transposables participent à la création de doublons par rapport aux deux autres mécanismes.

6.1.2 Développement des microsatellites

Il est généralement admis que les microsatellites ont besoin d'une taille minimum, aux environs de huit nucléotides, pour pouvoir entrer en phase de développement [Rose and Falush, 1998, Sibly et al., 2001, Buschiazzo and Gemmell, 2006], les séquences de taille plus petite étant considérées comme proto-microsatellites. Notre modèle propose au contraire que le développement débute dès l'état de doublon, qui est en cela considéré comme un microsatellite à part entière. Les quatre forces précitées peuvent contribuer au développement des microsatellites, et là encore à des degrés divers, comme détaillé ci-dessous.

Comme pour la création de doublons, le développement des microsatellites par mutation ponctuelle dépend de la séquence flanquante, car la création d'une répétition supplémentaire nécessite un certain nombre de bases préexistantes. Il est donc plus difficile à un motif de période importante de se développer de cette manière. Le nombre de répétitions déjà présentes pourrait aussi avoir une influence sur la probabilité qu'un tel événement se produise, car il semblerait que les séquences flanquantes dépendent de la taille du microsatellite [Vowles and Amos, 2004]. Cette taille a aussi une grande importance pour le développement par micro-duplication. En effet, si l'on considère que la probabilité de micro-duplication est la même à chaque position du génome, un grand microsatellite a plus de chance de subir un tel événement de mutation. La micro-duplication devra par contre être impérativement de la taille du motif, sous peine de briser le microsatellite au lieu de le développer. Elle sera par conséquent moins efficace que pour l'apparition de doublons. Le glissement de polymérase dépend lui-aussi du nombre de répétitions du microsatellite, pour des raisons similaires à la micro-duplication. Le mécanisme de glissement produit en revanche systématiquement une duplication de la taille du motif, ce qui lui permet de participer plus efficacement au développement du microsatellite (sous l'hypothèse d'un biais vers les expansions). Le rôle de l'insertion d'éléments

transposables dans le développement des microsatellites est légèrement différent de celui des trois autres mécanismes de mutation. En effet, elle ne permet pas le développement de locus déjà existants, mais en fait apparaître de nouveaux déjà développés. Elle n'aura donc d'influence que sur la distribution de certains microsatellites, et uniquement pour certaines tailles.

La contribution relative de ces processus au développement des microsatellites, et surtout à la distribution en taille de ces derniers dans les génomes, dépend donc de multiple facteurs. Nous n'avons évalué l'importance de la mutation ponctuelle, de la micro-duplication et du glissement que pour l'expansion des locus les plus courts (de deux répétitions). La mutation ponctuelle est encore le mécanisme prédominant à ce stade de développement, avec plus de 98% des créations de triplets (exceptés ceux adoptés). Nos calculs rudimentaires indiquent que le glissement ne concerne encore que 0,4% des apparitions, mais paraît plus fort pour les grands motifs. Pour les locus possédant plus de répétitions, le rôle de la mutation ponctuelle s'amoindrit nettement par rapport à celui des deux autres processus, comme l'indique la sur-représentation des microsatellites de tailles plus importantes [Rose and Falush, 1998, Pupko and Graur, 1999, Dieringer and Schlotterer, 2003]. Nos résultats montrent enfin que les éléments transposables ont une influence certaine sur le nombre de microsatellites développés, au moins chez l'homme, puisque 12% du total des microsatellites en sont issus (section 4.2.3). Une certaine partie est toutefois apparue dans l'élément après qu'il ait été inséré, et ne peuvent en cela être considérés comme des microsatellites adoptés.

6.2 Les diverses implications du modèle

6.2.1 Sur la distribution des microsatellites

Le modèle d'apparition et de développement présenté ci-dessus suggère que les types de microsatellites développés et leurs positions reposent en partie sur l'apparition des doublons. Les doublons étant majoritairement issus de la mutation ponctuelle, les contraintes inhérentes à ce processus ont une importance déterminante dans la production de microsatellites.

Composition génomique

La composition génomique est la première de ces contraintes. Dans un génome de composition non biaisée, tous les nucléotides sont en proportion équivalente. Si le génome est parfaitement aléatoire, tous les doublons d'une même période ont de plus la même probabilité d'apparaître. Cependant, tous les génomes des organismes vivants présentent des contraintes compositionnelles. Dieringer et

Schlötterer (2003) ont montré que le biais compositionnel (surplus d'un type de base par rapport aux autres) avait un effet non seulement sur les types de microsatellites produits, mais aussi sur leur nombre (figure 6.2). En effet, plus la composition est biaisée vers les C/G (ou A/T), plus le nombre de microsatellites présents au hasard est important, l'absence de biais donnant la plus faible densité. Ils ne présentent que les résultats pour les mononucléotides supérieurs à 9 nt, mais il est vraisemblable que l'influence sur les locus de taille inférieure (dont les doublons) soit la même. Le biais, en augmentant la proportion de certains nucléotides, augmente aussi la probabilité d'apparition des motifs de taille 2 (et plus) composés de ces nucléotides, et donc les chances de les trouver adjacentes au hasard. Les motifs les plus favorisés par un biais compositionnel sont ceux dont la composition se rapproche le plus de celle de l'ensemble du génome.

De plus, le biais compositionnel n'est pas homogène tout le long des séquences génomiques. Tous les génomes comportent en effet des zones plus ou moins riches en G/C [Lynch, 2007], cette tendance étant amplifiée chez les vertébrés, avec la présence des isochores [Duret et al., 2002]. La variation de composition peut d'ailleurs être très importante et très brusque. Les probabilités d'apparition des doublons répondent donc, à un niveau local, aux contraintes de biais compositionnel. Concrètement, les variations de biais compositionnel participent au développement de zones de faible complexité, elles-mêmes vecteurs de microsatellites.

Plus les zones de faible complexité sont nombreuses, plus le nombre de microsatellites créés est important. En théorie, cela signifie que les génomes très compartimentés (hétérogènes) devraient avoir des densités de microsatellites plus importantes. Les comparaisons que nous avons effectué dans le chapitre 3, ou des bases de données telles que TRDB [Gelfand et al., 2007] pourraient être de bons supports pour évaluer cette hypothèse. Il faut toutefois prendre en compte les autres forces contrôlant le développement des microsatellites, particulièrement la sélection dans les génomes compacts (par exemple, le génome de *S. cerevisiae* est constitué de plus de 70% de région codante).

Mutations CpG

Nous avons montré dans la section 5.2.2 que la prise en compte de la composition génomique ne suffisait pas à estimer correctement le nombre de motifs autres que mononucléotides dans un génome. Ne pas avoir tenu compte de l'hétérogénéité compositionnelle dans nos simulations a peut-être faussé nos résultats, même si des tests supplémentaires que nous avons réalisés semblent démontrer le contraire. Les autres facteurs possibles sont les biais qui s'exercent directement sur la mutation ponctuelle, en particulier la déamination des sites CpG méthylés (vois section 1.2.2). Ce type de

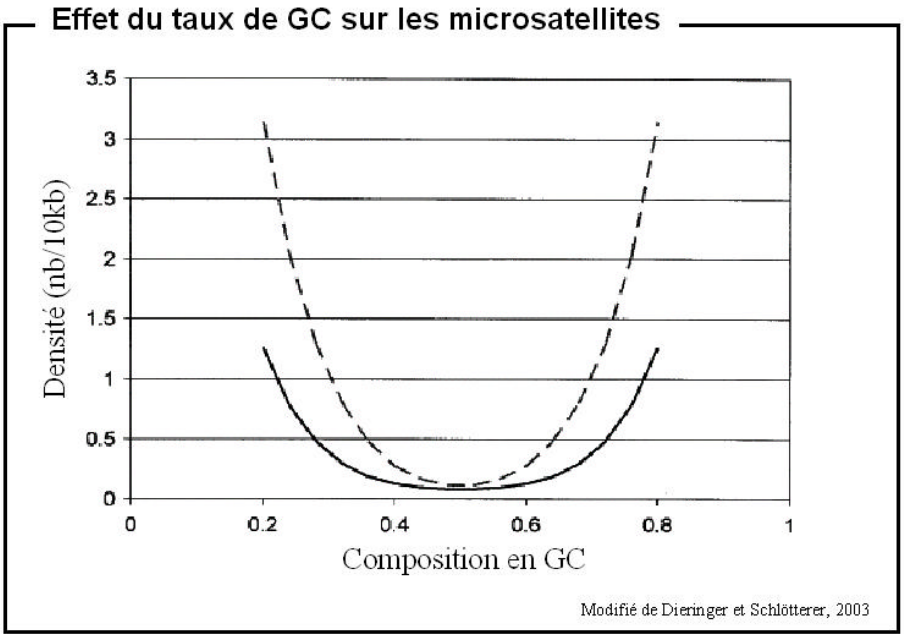


FIG. 6.2 – Densité de mono (ligne pointillée) et dinucléotides (ligne pleine) attendus dans une séquence génomique aléatoire, en fonction du taux de G/C de cette dernière. Seuls les microsatellites de taille supérieure à 9 nt sont considérés.

mutation est deux à quatre fois plus commun que l'ensemble des autres mutations ponctuelles. Elle produit par ailleurs systématiquement un motif TG/CA à partir d'un motif CG. L'apparition de doublons CACA est donc favorisée dans les génomes ayant une méthylation active. Selon notre modèle de développement des microsatellites, la densité de $(CA)_n$ devrait être plus élevée dans ces génomes. C'est effectivement le cas dans le génome humain, dont la méthylation est très active et qui possède un nombre anormalement élevé de dinucléotides AC. La généralisation à d'autres organismes est là-encore nécessaire pour confirmer cette hypothèse.

Plus généralement, la déamination des sites CpG méthylés favorise l'apparition de motifs comprenant la suite CA/TG. Ce sont donc l'ensemble de ces motifs qui devraient être sur-représentés dans les génomes à méthylation active, par rapport aux autres motifs. Toutefois, les mutations CpG étant des mutations ponctuelles, elles répondent aux mêmes contraintes de composition de la séquence. En particulier, l'apparition de doublons de motifs plus grands que CA nécessite d'avoir plus de bases préexistantes, limitant l'effet des mutations CpG pour la sur-représentation de ces motifs.

Nous avons reporté divers cas possibles de créations de microsatellites via des mutations CpG, à partir de la séquence interne des éléments Alu. Ces microsatellites sont des doublons ou des triplets, de période assez importante, et représentent 50 à 90% de l'ensemble des locus de mêmes motifs (GTG,

AGGC, et GAGGTG) détectés dans le génome humain (voir section 4.2.3). On pourrait donc penser que leur densité à un stade développé (plus de deux ou trois répétitions) est plus importante que celle des autres motifs de même taille. A titre d'exemple, la densité des GTG développés (6 répétitions ou plus) a été comparée à celle des autres trinucléotides (tableau 6.1). Les résultats montrent que ces microsatellites ne présentent pas de sur-densité. Cette observation est aussi valable pour les autres motifs créés en interne de séquences Alu ; la création des doublons n'est par conséquent pas la seule force influençant la distribution générale des microsatellites, en termes de motifs.

Densité des tri-nucléotides

taux de GC	0%	33%		66%		100%
motif, *densité (nb/Mb)*	AAT *7.16* ATT *6.89*	TTG *4.46* AAC *4.37* TTC *4.09* AAG *3.81*	ATC *1.12* ATG *1.06* ACT *0.34* AGT *0.29*	TCC *1.93* AGG *1.86* **ACC *0.78*** **TGG *0.67***	AGC *0.67* TGC *0.61* TCG *0.02* ACG *0.01*	GCC *0.41* GGC *0.40*

TAB. 6.1 – Densités de trinucléotides de taille supérieure ou égale à 18 nt dans le chromosome X humain, réparties par motifs. Les ACC/TGG (en gras), dont l'apparition est favorisée via les mutations CpG dans les séquences Alu, ne présentent pas de densité anormale.

Eléments transposables

Le troisième facteur qui joue sur la distribution génomique des microsatellites est l'influence des éléments transposables. Nous avons spécifié dans notre modèle que ce mode d'apparition ne concernait que certains motifs, ceux portés par les éléments intégrés. Par exemple, les séquences Alu sont à l'origine de 50% de l'ensemble des séquences $(A)_n$ présentent dans le génome humain, via leur queue polyA (voir section 4.2.3). Or, ce type de motif représente à lui seul plus d'un septième de l'ensemble des microsatellites de dix nucléotides et plus. Il est donc indéniable que ces rétrotransposons jouent un rôle particulier dans la distribution des microsatellites dans le génome des primates.

Nous n'avons étudié, dans ce manuscrit, que l'influence des éléments Alu sur la présence des microsatellites, et il serait intéressant d'évaluer l'apport d'autres éléments transposables, éventuellement dans d'autres espèces. Les rétrotransposons L1, par exemple, suivent le même mécanisme de rétrotransposition, nécessitant une queue polyA pour s'intégrer [Feng et al., 1996]. Ils participent donc probablement à biaiser la composition du génome des primates en faveur des $(A)n$, même s'ils sont deux fois moins nombreux que les éléments Alu [Lander et al., 2001].

L'influence des éléments Alu sur le type de microsatellites présents dans le génome humain ne s'arrête pas à l'importation de polyA. Ils augmentent aussi la densité des autres microsatellites riches en A, via la dégradation de ces polyA. Cette dégradation est principalement « naturelle », avec une accumulation de mutations au cours du temps, mais nous avons aussi montré qu'elle était favorisée. En effet, il est probable que des mutations apparaissent dans le polyA dès l'intégration de l'élément Alu, pouvant créer de fait des doublons ou des triplets de motifs riches en A. En d'autres termes, ces éléments transposables agissent comme des générateurs de régions de faible complexité, qui, comme nous l'avons expliqué, jouent un rôle important dans la distribution des microsatellites.

Enfin, les éléments transposables peuvent aussi participer à l'apparition de doublons. La séquence consensus Alu contient par exemple 46 doublons (tableau 6.2). Compte tenu du nombre de séquences Alu présentes dans le génome, cela nous donne plus de 6,5 million de doublons importés (sur les 500 millions que comprend le génome humain), avec une composition génomique très biaisée vers les C/G. De même, nous avons vu que certains sites des séquences Alu étaient des quasi-microsatellites, pouvant aboutir à des doublons ou des triplets via une seule mutation. Toutefois, nos résultats montrent que les doublons créés ne se développent pas, et n'ont par conséquent pas d'influence sur la distribution générale des microsatellites, en contradiction avec les résultats obtenus chez les drosophiles[Wilder and Hollocher, 2001]. Il existe dans le génome de ces insectes un élément transposable appelé *mini-me*, dont la séquence favorise l'apparition de triplets de motif GTCT ou GTCC. Les microsatellites issus de ces éléments transposables représentent les deux tiers du total dans l'ensemble du génome, pour les motifs considérés. Par contre, les auteurs ne disent rien sur la proportion de ces motifs par rapport aux autres dans le génome de la drosophile, mais il n'en reste pas moins que l'élément *mini-me* joue un rôle dans la distribution générale des microsatellites de ces espèces.

6.2.2 Sur la construction des modèles théoriques

Les résultats que nous avons présentés, et le modèle d'apparition des microsatellites qui en découle, ont donc certaines implications sur la distribution des microsatellites dans les génomes. Or, un certain nombre de modèles théoriques de dynamique des microsatellites sont ajustés sur ce type de distributions (voir section 2.4.1). Il était donc intéressant de réfléchir à comment nos résultats peuvent être pris en compte par ces modèles, et dans quelle mesure ils peuvent avoir une influence.

Doublons dans la séquence Alu

motif	nombre de doublons
A	5
C	14
G	17
T	4
AG	1
GA	4
CTA	1
Total	**46**

TAB. 6.2 – Nombre de doublons présents dans la séquence consensus de Alu, motif par motif.

Pas de taille minimum de glissement

La question la plus importante est certainement celle de la taille minimum de glissement. Nous avons remarqué que le glissement était effectif dès les doublons, même si le taux était très faible par rapport aux mutations ponctuelles. Or, dans la plupart des modèles actuels, le taux de glissement est automatiquement contraint à 0 en deçà d'une certaine taille, généralement entre 4 et 5 répétitions [Calabrese and Durrett, 2003, Dieringer and Schlotterer, 2003, Lai and Sun, 2003, Sainudiin et al., 2004]. Ces auteurs justifient cette taille minimum soit à partir de l'estimation de sur-représentation donnée par Rose & Falush (1998 ; voir chapitre 5), soit à partir de leur propres estimations de sur-représentation. Sibly *et al.* (2001) ont testé l'effet de la taille minimum sur l'ajustement du modèle aux données. Ils ont comparé, entre autres, un modèle avec une taille minimum et un autre sans, tous deux avec un taux de glissement proportionnel à la taille. Le meilleur ajustement aux données est obtenu pour le premier modèle, avec une taille minimum de glissement à 4 répétitions (soit 8 nucléotides dans leur cas). Ces résultats semblent donc contredire les nôtres.

La solution pourrait être de ne pas considérer le glissement comme proportionnel à la taille, mais avec une relation quadratique ou exponentielle. Dans ce modèle, le glissement serait très faible pour les microsatellites de petite taille, augmentant faiblement jusqu'à une certaine taille, à partir de laquelle l'augmentation de son taux serait plus prononcée. Ce type de modèle a déjà été implémenté par Calabrese *et al.* (2003) et a donné de meilleurs résultats que les modèles avec un taux proportionnel. Ces auteurs n'ont toutefois pas testé leur modèle sur l'ensemble de la distribution, car ils ont

imposé un seuil minimum de glissement. Bell et Jurka (1997) avaient aussi eu cette intuition, puisque qu'ils avaient remarqué qu'en réduisant le taux de glissement pour les petites tailles, l'ajustement aux données était meilleur qu'avec le simple taux proportionnel. Ils avaient toutefois fixé des taux manuellement pour chaque taille, sans réellement tester de relation quadratique (ou exponentielle).

La relation proportionnelle entre glissement et taille des locus donne des taux de glissement trop importants pour les locus de faible taille, si l'on veut que l'ajustement soit correct pour ceux de plus grande taille. C'est pourquoi les modèles ayant été ajustés sur des distributions de locus assez longs [Kruglyak et al., 1998, Sibly et al., 2001, Sainudiin et al., 2004] donnaient des résultats corrects avec une relation proportionnelle entre glissement et taille des locus. Et les modèles ajustés sur des distributions complètes ont dû se contraindre à une taille minimum pour le glissement, rendant invisible le mauvais ajustement aux petites tailles.

Prise en compte des micro-duplications

Le phénomène de micro-duplication que nous avons caractérisé devrait aussi être pris en compte dans les modèles. Pratiquement tous les modèles proposés jusqu'à maintenant supposent une apparition à une certaine taille par mutation ponctuelle, cette dernière n'ayant pas d'effet sur les autres tailles. Or, si l'on suppose une relation quadratique entre glissement et taille du locus, quelle que soit la taille, la micro-duplication pourrait avoir une influence sur la distribution des petits locus. C'est justement ce que montrent Dieringer & Schlötterer (2003) dans leur modèle, qui incorporent certaines caractéristiques des distributions aux petites tailles grâce à l'intégration d'un paramètre de *indel-slippage* (équivalent à notre micro-duplication). Ils n'ont toutefois pas essayé d'ajuster leurs résultats à des données réelles, ni de les comparer à d'autres modèles.

De plus, selon notre modèle d'apparition et de développement des microsatellites, intégrer uniquement l'effet des micro-duplications ne suffit pas à capturer l'ensemble des phénomènes affectant les petites tailles. En effet, l'expansion par mutation ponctuelle reste prédominante, au moins pour les doublons. Un paramètre de mutation ponctuelle devrait donc être ajouté aux modèles de dynamique des microsatellites, avec un taux inversement relié à la taille. Bell et Jurka (1995), ainsi que Dieringer et Schlötterer (2003), ont chacun modélisé ce mécanisme, mais en tant que modèle nul (modèle aléatoire où aucune dynamique particulière ne favorise l'apparition de microsatellites). L'intégration dans un modèle général aurait par exemple probablement permis à Bell et Jurka (1995) de mieux ajuster les données pour les petites tailles (*cf.* figure 1 dans leur article).

Prise en compte des éléments transposables

La dernière force que nous avons décrite dans le modèle d'apparition des microsatellites, et qui pourrait avoir sa place dans un modèle théorique de dynamique des microsatellites, est l'action des éléments transposables. A la différence des phénomènes précédents, l'importation de microsatellites avec les éléments transposables n'est gérée par aucun modèle, à l'exception de celui de Jarne *et al.* (1998). Ces derniers ont proposé d'ajouter un paramètre qui crée de nouveaux microsatellites à taux constant. Leur travail ne s'est toutefois focalisé que sur la densité et le polymorphisme des microsatellites du chromosome X humain par rapport à ceux des autosomes, et n'a pas pris en compte la dynamique de taille des locus.

Nos résultats montrent que les éléments transposables peuvent influencer l'apparition de microsatellites de deux manières : soit ils sont intégrés directement avec un locus répété développé de taille déterminée, soit ils contiennent des régions de faible complexité qui favorisent la genèse de certains motifs. Les modèles de dynamique des microsatellites ne sont *a priori* pas affectés par le second mécanisme, puisqu'il s'agit d'apparitions *de novo*, à partir d'une séquence quelconque. En effet, seule la densité de microsatellites est affectée par ce type d'apparition, et les locus créés suivent normalement la même dynamique que dans le reste du génome. Par contre, l'importation de microsatellites déjà développés peut avoir une grande influence sur la construction des modèles théoriques. Dans le modèle SMM (dont tous les modèles actuels sont dérivés), le nombre de locus à une taille donnée est généralement calculé à partir du nombre de locus aux tailles voisines, et de la probabilité qu'un glissement les fassent changer de taille. Si un phénomène, tel que l'intégration d'éléments transposables, vient modifier le nombre de microsatellites à une taille déterminée, c'est la dynamique complète qui sera changée. Ne pas prendre en compte ce genre d'événements peut donc biaiser l'estimation des paramètres de dynamique de taille.

Deux solutions sont possibles pour éviter le biais dû aux éléments transposables. La première est de retirer de l'analyse tous les locus associés à ces éléments, en masquant la séquence d'origine, comme nous l'avons fait pour le calcul de sur-représentation des doublons et des triplets (voir chapitre 5). La seconde solution est d'intégrer un paramètre d'apparition à taux constant, tel que celui utilisé par Jarne *et al.* (1998), à une taille déterminée. La gestion des éléments Alu se ferait par exemple avec une apparition à une taille de 25 nt, dans les modèles de dynamique des polyA. Actuellement, seuls Lai & Sun (2003) et Dieringer & Schlötterer (2003) ont inclus les mononucléotides humains (à 90% des polyA) dans leurs estimations de paramètres, mais sans prendre en compte l'influence des éléments transposables. La plupart des autres modèles sont basés sur des distributions de

dinucléotides [Kruglyak et al., 1998, Sibly et al., 2001, Calabrese and Durrett, 2003], pour lesquels les éléments Alu ont peu d'influence (voir chapitre 4). Toutefois, nos travaux n'ayant porté que sur l'association des microsatellites avec les éléments Alu, nous ne sommes pas en mesure d'évaluer à quel point d'autres éléments transposables pourrait influencer la distribution d'autres motifs microsatellites.

L'intégration de l'effet des éléments transposables telle qu'elle est présentée ici présente toutefois quelques lacunes. La principale est qu'elle suppose une apparition constante des éléments transposables dans le génome. Or, nous avons vu que les intégrations se faisaient plutôt par vagues d'amplification (section 1.3.3). Malheureusement, la plupart des modèles ajustés à des distributions génomiques sont basés sur des chaînes de Markov, et supposent une distribution à l'équilibre des tailles des locus. Ils ne peuvent donc gérer des événements temporels, telles que des vagues d'apparition de locus à une taille donnée pendant un temps donné.

Biais de mutation

Au-delà du modèle d'apparition des microsatellites, nos travaux ont aussi permis de caractériser la réduction en taille qui affectait les microsatellites les plus longs. Ce phénomène est surtout visible pour les $(A)_n$ intégrés au génome humain avec les séquences Alu, mais semble aussi s'appliquer aux microsatellites riches en A qui se sont développés par la suite dans ces séquences. Ces observations supportent fortement l'hypothèse d'un biais de mutation, qui favorise les contractions pour les locus de taille importante. Ce biais de contraction a été incorporé de nombreuses fois dans les modèles de dynamique des microsatellites [Garza et al., 1995, Calabrese and Durrett, 2003, Sainudiin et al., 2004], et permet à chaque fois un meilleur ajustement aux données. Il permet en outre d'expliquer l'absence de microsatellites très longs, alors que le modèle SMM original [Ohta and Kimura, 1973] prédit une expansion non contrainte pour les grandes tailles.

Comme nous l'avons proposé dans le modèle de la figure 4.11, les polyA en queue d'éléments Alu perdent en moyenne une dizaine de répétitions en environ 5-15 Ma, alors que l'accumulation de mutations semble être plus lente (voir section 4.2.5). Cette contraction rapide a déjà été reportée pour quelques locus Alu récemment intégrés [Roy-Engel et al., 2002], mais c'est la première fois qu'elle est observée à grande échelle. Le biais de mutation est généralement modélisé par une probabilité plus forte de subir des contractions que des expansions, sans toutefois changer la valeur du pas de mutation (le nombre de répétitions impliquées à chaque événement). Même avec un biais de contraction très fort, il semble peu envisageable de subir une réduction de taille aussi importante en

si peu de temps. Un modèle comme celui de Xu *et al.* (2000), dans lequel les locus longs subissent plus de contractions à pas multiples qu'à pas simples, provoquant ainsi une réduction rapide de leur taille, serait plus en accord avec nos données.

6.3 Conclusion

6.3.1 Synthèse

La caractéristique la plus étonnante des microsatellites est leur très forte variabilité en taille. Cette variabilité a suscité une réelle dynamique de recherche pour essayer d'en comprendre les tenants et aboutissants. Un modèle moléculaire (le glissement de polymérase) a été proposé pour expliquer cette variabilité [Levinson and Gutman, 1987b], et un modèle théorique (le SMM) a pu y être appliqué [Kimura and Ohta, 1978]. Les diverses utilisations des microsatellites (en biologie des populations, pour les empreintes ADN et les tests de paternité, etc.) reposent eux aussi sur cette variabilité [Jarne and Lagoda, 1996, Balding, 1999]. Néanmoins, au-delà de leur variabilité, les microsatellites sont dignes d'intérêt à bien d'autres égards. La densité des ces éléments est par exemple plus élevée que celle à laquelle on peut s'attendre, et leur distribution en termes de motifs est différente selon les organismes. La densité élevée est généralement attribuée au mécanisme de glissement, qui, par le biais des expansions, augmente significativement le nombre de séquences répétées longues. Il faut toutefois pour cela que les microsatellites existent déjà, en l'occurrence sous une forme moins développée appelée proto-microsatellite.

Le travail présenté dans cette thèse s'est focalisé sur les divers mécanismes pouvant aboutir à l'apparition des microsatellites. Mais comme pour tout modèle d'étude, le premier travail à été de choisir une méthode pour obtenir les données. Notre choix s'est porté sur l'analyse de séquence, qui permet d'obtenir un très grand nombre de locus, et de s'émanciper des contraintes populationnelles. Ce type d'analyse nécessite d'extraire les locus microsatellites de la séquence analysée, par le biais d'algorithmes informatiques. Nous avons commencé par réaliser une étude comparative de différents algorithmes dédiés à la détection de séquences répétées, et observé des différences très importantes entre les divers résultats. La disparité est principalement causée par deux facteurs principaux que sont la mise en place d'une taille minimum de détection et la gestion de l'imperfection des microsatellites. La question de la taille minimum est critique, car nous avons observé que plus les microsatellites sont courts, plus ils sont nombreux, quel que soit l'organisme. Ces travaux nous ont fait conclure que l'utilisation des algorithmes dans la littérature devraient être mieux documentée et qu'une définition des microsatellites plus appropriée que le simple $(X)_n$ (avec X le motif et n le nombre de répétitions) serait à envisager.

La question de l'apparition des microsatellites a ensuite été abordée par le biais de leur association avec les rétrotransposons Alu chez l'homme. Ces éléments sont très nombreux dans le génome des primates et possèdent une séquence polyA dans leur région 3', ainsi qu'une seconde zone riche en A en leur centre, toutes deux propices à l'apparition de microsatellites. Nos travaux démontrent que l'association avec les éléments Alu est significative pas uniquement pour les polyA, mais pour tous les microsatellites riches en A et pour les dinucléotides. Ces associations sont valables quel que soit l'âge de l'élément Alu, à part pour les plus récents. De plus, une réduction très rapide de la taille des polyA a été observée. L'ensemble de ces résultats nous ont servi à développer un modèle de cycle de vie des polyA associés aux 3' de séquences Alu. D'autre part, un certain nombre de microsatellites de très petites tailles (2 à 3 répétitions) ont aussi été caractérisés à l'intérieur des éléments Alu, mais leur capacité de développement s'est avérée nulle. Enfin, un microsatellite $(AAAAAC)n$ a été détecté dans le linker de certains éléments Alu, et pourraient servir à caractériser une nouvelle sous-famille Alu.

La dernière phase de ma thèse a été consacrée à l'apparition *de novo* des microsatellites, c'est-à-dire à partir d'une séquence génomique quelconque. Une méthode de génomique comparative a été mise en place, grâce à laquelle les apparitions intervenues dans le génome humain depuis la spéciation avec le chimpanzé ont pu être détectées et analysées. Nous nous sommes restreints à l'apparition des séquences de deux et trois répétitions uniquement, sous l'hypothèse de trois mécanismes possibles : la mutation ponctuelle, la micro-duplication (définie comme la duplication en tandem d'un fragment d'ADN très petit), et le glissement de polymérase. Nous sommes parvenus à la conclusion que la mutation ponctuelle était le mécanisme principal pour l'apparition de ces répétitions très courtes (99%). La micro-duplication provoque toutefois une sur-représentation des doublons (séquences à deux répétitions) dans le génome, au moins pour les tétra à hexanucléotides. Le glissement de polymérase semble quant à lui actif dès que deux répétitions sont présentes, signifiant que donner une taille minimum [Rose and Falush, 1998, Sibly et al., 2001, Lai and Sun, 2003] pour le glissement n'a pas lieu d'être.

La synthèse de ces résultats a abouti à l'élaboration d'un modèle général d'apparition et de développement des microsatellites. Dans ce modèle, l'apparition correspond soit à la création d'un doublon par mutation ponctuelle ou micro-duplication, qui va ensuite pouvoir se développer, soit à l'adoption (intégration) par le biais d'un élément transposable. Le développement sera ensuite contrôlé principalement par les mutations ponctuelles pour les petites tailles, puis de plus en plus

par le glissement avec l'augmentation du nombre de répétitions. L'apparition des microsatellites dans les génomes n'est donc pas le seul fait de la mutation ponctuelle, même si elle y contribue pour une grande part. Les autres mécanismes sont aussi à prendre en compte si l'on veut pouvoir comprendre la distribution actuelle de ces éléments répétés dans les génomes, et si l'on veut pouvoir modéliser correctement leur dynamique.

6.3.2 Vers un envahissement du génome ?

La question de l'apparition des microsatellites amène à celle de leur développement, et finalement, à celle de leur disparition. Les microsatellites sont considérés, comme la plupart des éléments répétés des génomes, comme des régions non-codantes, *a priori* non fonctionnelles, qui peuvent être regroupées sous la notion d' *ADN égoïste* [Doolittle and Sapienza, 1980, Orgel and Crick, 1980]. Cette notion a toujours excité les évolutionnistes, car elle a permis de résoudre le paradoxe de la *C-value*, pour lequel la complexité d'un organisme devait être en relation avec la masse de son ADN (censée être représentative de son nombre de gènes), alors que ce n'était manifestement pas le cas [Gregory, 2005]. L'existence d'ADN égoïste a rendu caduc le rapport entre nombre de gènes et masse du génome. L'ADN égoïste étant constitué pour une grande part d'éléments transposables, de gènes dupliqués ou de pseudogènes, on a longtemps estimé qu'il ne pouvait que s'accumuler au cours de l'évolution des espèces. Or, la comparaison des tailles des génomes de diverses espèces à partir d'une phylogénie a montré que ces derniers pouvaient aussi se contracter [Leitch et al., 2005], supposant une disparition de cet ADN égoïste. Les mécanismes proposés pour cette disparition sont essentiellement la recombinaison inégale entre éléments transposables et la réparation de CDB par SSA (voir section 1.2.3).

La disparition des microsatellites semble quant à elle décomposée en deux phases [Taylor et al., 1999]. Elle commence par une accumulation de mutations, qui réduit la variabilité du locus [Jin et al., 1996, Richard and Dujon, 1996], puis est suivie par une série de contractions. La période ne sera au final plus reconnaissable, signifiant la « mort » du microsatellite. A la place se trouvera par contre une région de faible complexité, favorisant l'apparition d'un nouveau microsatellite. L'apparition étant aussi favorisée dans des séquences 'normales' (pas de faible complexité) grâce à l'effet de la micro-duplication, il est envisageable que les régions non fonctionnelles des génomes soient à terme envahies de microsatellites et de régions de faible complexité. Un tel scénario pourrait avoir des conséquences dramatiques sur l'organisation des génomes, à moins que d'autres mécanismes de disparition ne soient à l'œuvre. Ce qui ne pourra être déterminé que par une étude approfondie des disparitions des microsatellites, telle que celle que nous avons conduite pour l'apparition.

Bibliographie

[Abajian, 2004] Abajian, C. (1994-2004). Sputnik.

[Andolfatto, 2005] Andolfatto, P. (2005). Adaptive evolution of non-coding dna in drosophila. *Nature*, 437(7062) :1149–52.

[Angers and Bernatchez, 1997] Angers, B. and Bernatchez, L. (1997). Complex evolution of a salmonid microsatellite locus and its consequences in inferring allelic divergence from size information. *Molecular Biology and Evolution*, 14(3) :230–238.

[Arcot et al., 1995] Arcot, S. S., Wang, Z., Weber, J. L., Deininger, P. L., and Batzer, M. A. (1995). Alu repeats : a source for the genesis of primate microsatellites. *Genomics*, 29(1) :136–44.

[Baer et al., 2007] Baer, C. F., Miyamoto, M. M., and Denver, D. R. (2007). Mutation rate variation in multicellular eukaryotes : causes and consequences. *Nat Rev Genet*, 8(8) :619–31.

[Balding, 1999] Balding, D. (1999). Forensic applications of microsatellite markers. In Goldstein, D. B. and Schlötterer, C., editors, *Microsatellites evolution and applications*, pages 198–210. Oxford University Press, Oxford.

[Bao and Eddy, 2002] Bao, Z. and Eddy, S. R. (2002). Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res*, 12(8) :1269–76.

[Batzer and Deininger, 2002] Batzer, M. A. and Deininger, P. L. (2002). Alu repeats and human genomic diversity. *Nat Rev Genet*, 3(5) :370–9.

[Batzer et al., 1996] Batzer, M. A., Deininger, P. L., Hellmann-Blumberg, U., Jurka, J., Labuda, D., Rubin, C. M., Schmid, C. W., Zietkiewicz, E., and Zuckerkandl, E. (1996). Standardized nomenclature for alu repeats. *J Mol Evol*, 42(1) :3–6.

[Batzer et al., 1995] Batzer, M. A., Rubin, C. M., Hellmann-Blumberg, U., Alegria-Hartman, M., Leeflang, E. P., Stern, J. D., Bazan, H. A., Shaikh, T. H., Deininger, P. L., and Schmid, C. W. (1995). Dispersion and insertion polymorphism in two small subfamilies of recently amplified human alu repeats. *J Mol Biol*, 247(3) :418–27.

[Bell and Jurka, 1997] Bell, G. I. and Jurka, J. (1997). The length distribution of perfect dimer repetitive dna is consistent with its evolution by an unbiased single-step mutation process. *J Mol Evol*, 44(4) :414–21.

[Benson, 1999] Benson, G. (1999). Tandem repeats finder : a program to analyze dna sequences. *Nucleic Acids Res*, 27(2) :573–80.

[Blanquart, 2007] Blanquart, S. (2007). *Reconstruction phylogénétique par analyse bayésienne des séquences moléculaires*. PhD thesis, Université Montpellier II.

[Boeke, 1997] Boeke, J. D. (1997). Lines and alus–the polya connection. *Nat Genet*, 16(1) :6–7.

[Brohede and Ellegren, 1999] Brohede, J. and Ellegren, H. (1999). Microsatellite evolution : polarity of substitutions within repeats and neutrality of flanking sequences. *Proc Biol Sci*, 266(1421) :825–33.

[Brookfield, 1993] Brookfield, J. F. Y. (1993). The generation of sequence similarity in sines and lines. *Trends in Genetics*, 9(2) :38–38.

[Buard and Jeffreys, 1997] Buard, J. and Jeffreys, A. J. (1997). Big, bad minisatellites. *Nat Genet*, 15(4) :327–8.

[Buschiazzo and Gemmell, 2006] Buschiazzo, E. and Gemmell, N. J. (2006). The rise, fall and renaissance of microsatellites in eukaryotic genomes. *Bioessays*, 28(10) :1040–1050.

[Calabrese and Durrett, 2003] Calabrese, P. and Durrett, R. (2003). Dinucleotide repeats in the drosophila and human genomes have complex, length-dependent mutation processes. *Molecular Biology and Evolution*, 20(5) :715–725.

[Carroll et al., 2001] Carroll, M. L., Roy-Engel, A. M., Nguyen, S. V., Salem, A. H., Vogel, E., Vincent, B., Myers, J., Ahmad, Z., Nguyen, L., Sammarco, M., Watkins, W. S., Henke, J., Makalowski, W., Jorde, L. B., Deininger, P. L., and Batzer, M. A. (2001). Large-scale analysis of the alu ya5 and yb8 subfamilies and their contribution to human genomic diversity. *J Mol Biol*, 311(1) :17–40.

[Castelo et al., 2002] Castelo, A. T., Martins, W., and Gao, G. R. (2002). Troll-tandem repeat occurrence locator. *Bioinformatics*, 18(4) :634–636.

[Chakraborty et al., 1997] Chakraborty, R., Kimmel, M., Stivers, D. N., Davison, L. J., and Deka, R. (1997). Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. *Proc Natl Acad Sci U S A*, 94(3) :1041–6.

[Chambers and MacAvoy, 2000] Chambers, G. K. and MacAvoy, E. S. (2000). Microsatellites : consensus and controversy. *Comp Biochem Physiol B Biochem Mol Biol*, 126(4) :455–76.

[Charlesworth et al., 1994] Charlesworth, B., Sniegowski, P., and Stephan, W. (1994). The evolutionary dynamics of repetitive dna in eukaryotes. *Nature*, 371(6494) :215–20.

[Chauhan et al., 2002] Chauhan, C., Dash, D., Grover, D., Rajamani, J., and Mukerji, M. (2002). Origin and instability of gaa repeats : insights from alu elements. *J Biomol Struct Dyn*, 20(2) :253–63.

[Clark et al., 2004] Clark, R. M., Dalgliesh, G. L., Endres, D., Gomez, M., Taylor, J., and Bidichandani, S. I. (2004). Expansion of gaa triplet repeats in the human genome : unique origin of the frda mutation at the center of an alu. *Genomics*, 83(3) :373–83.

[Coenye and Vandamme, 2005] Coenye, T. and Vandamme, P. (2005). Characterization of mononucleotide repeats in sequenced prokaryotic genomes. *DNA Research*, 12(4) :221–233.

[Cordaux et al., 2004] Cordaux, R., Hedges, D. J., and Batzer, M. A. (2004). Retrotransposition of alu elements : how many sources ? *Trends Genet*, 20(10) :464–7.

[Cornuet et al., 2006] Cornuet, J. M., Beaumont, M. A., Estoup, A., and Solignac, M. (2006). Inference on microsatellite mutation processes in the invasive mite, varroa destructor, using reversible jump markov chain monte carlo. *Theor Popul Biol*, 69(2) :129–44.

[Coward and Drablos, 1998] Coward, E. and Drablos, F. (1998). Detecting periodic patterns in biological sequences. *Bioinformatics*, 14(6) :498–507.

[Deininger et al., 1992] Deininger, P. L., Batzer, M. A., Hutchison, C. A., r., and Edgell, M. H. (1992). Master genes in mammalian repetitive dna amplification. *Trends Genet*, 8(9) :307–11.

[Delgrange and Rivals, 2004] Delgrange, O. and Rivals, E. (2004). Star : an algorithm to search for tandem approximate repeats. *Bioinformatics*, 20(16) :2812–20.

[Desmarais et al., 2006] Desmarais, E., Belkhir, K., Garza, J. C., and Bonhomme, F. (2006). Local mutagenic impact of insertions of ltr retrotransposons on the mouse genome. *J Mol Evol*, 63(5) :662–75.

[Dettman and Taylor, 2004] Dettman, J. R. and Taylor, J. W. (2004). Mutation and evolution of microsatellite loci in neurospora. *Genetics*, 168(3) :1231–48.

[Dewannieux and Heidmann, 2005] Dewannieux, M. and Heidmann, T. (2005). Role of poly(a) tail length in alu retrotransposition. *Genomics*, 86(3) :378–81.

[Dieringer and Schlotterer, 2003] Dieringer, D. and Schlotterer, C. (2003). Two distinct modes of microsatellite mutation processes : evidence from the complete genomic sequences of nine species. *Genome Res*, 13(10) :2242–51.

[Dirienzo et al., 1994] Dirienzo, A., Peterson, A. C., Garza, J. C., Valdes, A. M., Slatkin, M., and Freimer, N. B. (1994). Mutational processes of simple-sequence repeat loci in human-populations.

Proceedings of the National Academy of Sciences of the United States of America, 91(8) :3166–3170.

[Dokholyan et al., 2000] Dokholyan, N. V., Buldyrev, S. V., Havlin, S., and Stanley, H. E. (2000). Distributions of dimeric tandem repeats in non-coding and coding dna sequences. *J Theor Biol*, 202(4) :273–82.

[Doolittle and Sapienza, 1980] Doolittle, W. F. and Sapienza, C. (1980). Selfish genes, the phenotype paradigm and genome evolution. *Nature*, 284(5757) :601–3.

[Duret, 2002] Duret, L. (2002). Evolution of synonymous codon usage in metazoans. *Current Opinion in Genetics and Development*, 12(6) :640–649.

[Duret et al., 2002] Duret, L., Semon, M., Piganeau, G., Mouchiroud, D., and Galtier, N. (2002). Vanishing gc-rich isochores in mammalian genomes. *Genetics*, 162(4) :1837–1847.

[Edgar and Myers, 2005] Edgar, R. C. and Myers, E. W. (2005). Piler : identification and classification of genomic repeats. *Bioinformatics*, 21 :I152–I158.

[Ehrlich and Wang, 1981] Ehrlich, M. and Wang, R. Y. (1981). 5-methylcytosine in eukaryotic dna. *Science*, 212(4501) :1350–7.

[Eickbush and Eickbush, 2005] Eickbush, T. and Eickbush, D. (2005). Transposable elements : Evolution. *Encyclopedia of Life Science*.

[Eickbush and Malik, 2001] Eickbush, T. and Malik, H. (2001). Origins and evolution of retrotransposons. In al., N. C. e., editor, *Mobile DNA*, pages 1111–1144. ASM Press, Washington, D.C., second edition.

[Ellegren, 2000a] Ellegren, H. (2000a). Heterogeneous mutation processes in human microsatellite dna sequences. *Nat Genet*, 24(4) :400–2.

[Ellegren, 2000b] Ellegren, H. (2000b). Microsatellite mutations in the germline : implications for evolutionary inference. *Trends Genet*, 16(12) :551–8.

[Ellegren, 2004] Ellegren, H. (2004). Microsatellites : simple sequences with complex evolution. *Nat Rev Genet*, 5(6) :435–45.

[Estoup et al., 1995] Estoup, A., Garnery, L., Solignac, M., and Cornuet, J. M. (1995). Microsatellite variation in honey bee (apis mellifera l.) populations : hierarchical genetic structure and test of the infinite allele and stepwise mutation models. *Genetics*, 140(2) :679–95.

[Falush and Iwasa, 1999] Falush, D. and Iwasa, Y. (1999). Size-dependent mutability and microsatellite constraints. *Molecular Biology and Evolution*, 16(7) :960–966.

[Feng et al., 1996] Feng, Q. H., Moran, J. V., Kazazian, H. H., and Boeke, J. D. (1996). Human l1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell*, 87(5) :905–916.

[Ferguson and Holloman, 1996] Ferguson, D. O. and Holloman, W. K. (1996). Recombinational repair of gaps in dna is asymmetric in ustilago maydis and can be explained by a migrating d-loop model. *Proceedings of the National Academy of Sciences of the United States of America*, 93(11) :5419–5424.

[Fischetti et al., 1993] Fischetti, V. A., Landau, G. M., Sellers, P. H., and Schmidt, J. P. (1993). Identifying periodic occurrences of a template with applications to protein-structure. *Information Processing Letters*, 45(1) :11–18.

[Flavell, 1995] Flavell, A. J. (1995). Retroelements, reverse transcriptase and evolution. *Comp Biochem Physiol B Biochem Mol Biol*, 110(1) :3–15.

[Formosa and Alberts, 1986] Formosa, T. and Alberts, B. M. (1986). Dna-synthesis dependent on genetic-recombination - characterization of a reaction catalyzed by purified bacteriophage-t4 proteins. *Cell*, 47(5) :793–806.

[Fu and Chakraborty, 1998] Fu, Y. X. and Chakraborty, R. (1998). Simultaneous estimation of all the parameters of a stepwise mutation model. *Genetics*, 150(1) :487–97.

[Garza et al., 1995] Garza, J. C., Slatkin, M., and Freimer, N. B. (1995). Microsatellite allele frequencies in humans and chimpanzees, with implications for constraints on allele size. *Mol Biol Evol*, 12(4) :594–603.

[Gelfand et al., 2007] Gelfand, Y., Rodriguez, A., and Benson, G. (2007). Trdb - the tandem repeats database. *Nucleic Acids Research*, 35 :D80–D87.

[Goffeau et al., 1996] Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H., and Oliver, S. G. (1996). Life with 6000 genes. *Science*, 274(5287) :546, 563–7.

[Gordon, 1997] Gordon, A. J. E. (1997). Microsatellite birth register. *Journal of Molecular Evolution*, 45(3) :337–338.

[Gregory, 2005] Gregory, T. (2005). Genome size evolution in animals. In Gregory, T., editor, *The evolution of the genome*. Elsevier Academic Press, Burlington, MA.

[Haag-Liautard et al., 2007] Haag-Liautard, C., Dorris, M., Maside, X., Macaskill, S., Halligan, D. L., Charlesworth, B., and Keightley, P. D. (2007). Direct estimation of per nucleotide and genomic deleterious mutation rates in drosophila. *Nature*, 445(7123) :82–5.

[Harr and Schlotterer, 2000] Harr, B. and Schlotterer, C. (2000). Long microsatellite alleles in drosophila melanogaster have a downward mutation bias and short persistence times, which cause their genome-wide underrepresentation. *Genetics*, 155(3) :1213–20.

[Harr et al., 1998] Harr, B., Zangerl, B., Brem, G., and Schlotterer, C. (1998). Conservation of locus-specific microsatellite variability across species : a comparison of two drosophila sibling species, d. melanogaster and d. simulans. *Mol Biol Evol*, 15(2) :176–84.

[Harr et al., 2000] Harr, B., Zangerl, B., and Schlotterer, C. (2000). Removal of microsatellite interruptions by dna replication slippage : phylogenetic evidence from drosophila. *Mol Biol Evol*, 17(7) :1001–9.

[Hartl and Clark, 1997] Hartl, D. and Clark, A. (1997). *Principles of Population Genetics*. Sinauer Associates, Inc., Sunderland, Massachusetts, third edition.

[Hastings, 1988] Hastings, P. J. (1988). Recombination in the eukaryotic nucleus. *Bioessays*, 9(2-3) :61–64.

[Henderson and Petes, 1992] Henderson, S. T. and Petes, T. D. (1992). Instability of simple sequence dna in saccharomyces cerevisiae. *Mol Cell Biol*, 12(6) :2749–57.

[Huang et al., 2002] Huang, Q. Y., Xu, F. H., Shen, H., Deng, H. Y., Liu, Y. J., Liu, Y. Z., Li, J. L., Recker, R. R., and Deng, H. W. (2002). Mutation patterns at dinucleotide microsatellite loci in humans. *Am J Hum Genet*, 70(3) :625–34.

[Jarne et al., 1998] Jarne, P., David, P., and Viard, F. (1998). Microsatellites, transposable elements and the x chromosome. *Mol Biol Evol*, 15(1) :28–34.

[Jarne and Lagoda, 1996] Jarne, P. and Lagoda, P. J. L. (1996). Microsatellites, from molecules to populations and back. *Trends in Ecology and Evolution*, 11(10) :424–429.

[Jin et al., 1996] Jin, L., Macaubas, C., Hallmayer, J., Kimura, A., and Mignot, E. (1996). Mutation rate varies among alleles at a microsatellite locus : phylogenetic evidence. *Proc Natl Acad Sci U S A*, 93(26) :15285–8.

[Jurka, 1997] Jurka, J. (1997). Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proc Natl Acad Sci U S A*, 94(5) :1872–7.

[Jurka and Pethiyagoda, 1995] Jurka, J. and Pethiyagoda, C. (1995). Simple repetitive dna sequences from primates : compilation and analysis. *J Mol Evol*, 40(2) :120–6.

[Kapitonov and Jurka, 1996] Kapitonov, V. and Jurka, J. (1996). The age of alu subfamilies. *J Mol Evol*, 42(1) :59–65.

[Katti et al., 2001] Katti, M. V., Ranjekar, P. K., and Gupta, V. S. (2001). Differential distribution of simple sequence repeats in eukaryotic genome sequences. *Mol Biol Evol*, 18(7) :1161–7.

[Kimura, 1968] Kimura, M. (1968). Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. *Genet Res*, 11(3) :247–69.

[Kimura, 1980] Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol*, 16(2) :111–20.

[Kimura and Crow, 1964] Kimura, M. and Crow, J. F. (1964). The number of alleles that can be maintained in a finite population. *Genetics*, 49 :725–38.

[Kimura and Ohta, 1978] Kimura, M. and Ohta, T. (1978). Stepwise mutation model and distribution of allelic frequencies in a finite population. *Proc Natl Acad Sci U S A*, 75(6) :2868–72.

[Klinman et al., 1999] Klinman, D. M., Verthelyi, D., Takeshita, F., and Ishii, K. J. (1999). Immune recognition of foreign dna : a cure for bioterrorism ? *Immunity*, 11(2) :123–9.

[Kolpakov et al., 2003] Kolpakov, R., Bana, G., and Kucherov, G. (2003). mreps : Efficient and flexible detection of tandem repeats in dna. *Nucleic Acids Res*, 31(13) :3672–8.

[Kolpakov and Kucherov, 2003] Kolpakov, R. and Kucherov, G. (2003). Finding approximate repetitions under hamming distance. *Theoretical Computer Science*, 303(1) :135–156.

[Krakauer and Plotkin, 2002] Krakauer, D. C. and Plotkin, J. B. (2002). Redundancy, antiredundancy, and the robustness of genomes. *Proc Natl Acad Sci U S A*, 99(3) :1405–9.

[Krawczak and Cooper, 1997] Krawczak, M. and Cooper, D. N. (1997). The human gene mutation database. *Trends in Genetics*, 13(3) :121–122.

[Kruglyak et al., 2000] Kruglyak, S., Durrett, R., Schug, M. D., and Aquadro, C. F. (2000). Distribution and abundance of microsatellites in the yeast genome can be explained by a balance between slippage events and point mutations. *Mol Biol Evol*, 17(8) :1210–9.

[Kruglyak et al., 1998] Kruglyak, S., Durrett, R. T., Schug, M. D., and Aquadro, C. F. (1998). Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proc Natl Acad Sci U S A*, 95(18) :10774–8.

[Kumar et al., 2005] Kumar, S., Filipski, A., Swarna, V., Walker, A., and Hedges, S. B. (2005). Placing confidence limits on the molecular age of the human-chimpanzee divergence. *Proceedings of the National Academy of Sciences of the United States of America*, 102(52) :18842–18847.

[Lai and Sun, 2003] Lai, Y. L. and Sun, F. Z. (2003). The relationship between microsatellite slippage mutation rate and the number of repeat units. *Molecular Biology and Evolution*, 20(12) :2123–2131.

[Landau et al., 2001] Landau, G. M., Schmidt, J. P., and Sokol, D. (2001). An algorithm for approximate tandem repeats. *J Comput Biol*, 8(1) :1–18.

[Lander et al., 2001] Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822) :860–921.

[Leclercq et al., 2007] Leclercq, S., Rivals, E., and Jarne, P. (2007). Detecting microsatellites within genomes : significant variation among algorithms. *BMC Bioinformatics*, 8 :125.

[Leitch et al., 2005] Leitch, I. J., Soltis, D. E., Soltis, P. S., and Bennett, M. D. (2005). Evolution of dna amounts across land plants (embryophyta). *Annals of Botany*, 95(1) :207–217.

[Levinson and Gutman, 1987a] Levinson, G. and Gutman, G. A. (1987a). High frequencies of short frameshifts in poly-ca/tg tandem repeats borne by bacteriophage m13 in escherichia coli k-12. *Nucleic Acids Res*, 15(13) :5323–38.

[Levinson and Gutman, 1987b] Levinson, G. and Gutman, G. A. (1987b). Slipped-strand mispairing : a major mechanism for dna sequence evolution. *Mol Biol Evol*, 4(3) :203–21.

[Li, 1997] Li, W. (1997). *Molecular evolution*. Sinauer Associates, Inc., SunderLand.

[Lynch, 2007] Lynch, M. (2007). *The origins of genome architecture*. Sinauer Associates, Inc., Sunderland.

[Malpertuy et al., 2003] Malpertuy, A., Dujon, B., and Richard, G. F. (2003). Analysis of microsatellites in 13 hemiascomycetous yeast species : mechanisms involved in genome dynamics. *J Mol Evol*, 56(6) :730–41.

[Marais, 2002] Marais, G. (2002). *Les effets pervers du sexe sur l'évolution des génomes*. PhD thesis, Claude Bernard - Lyon 1.

[Mcgill et al., 1989] Mcgill, C., Shafer, B., and Strathern, J. (1989). Coconversion of flanking sequences with homothallic switching. *Cell*, 57(3) :459–467.

[Messer and Arndt, 2007] Messer, P. W. and Arndt, P. F. (2007). The majority of recent short dna insertions in the human genome are tandem duplications. *Mol Biol Evol*, 24(5) :1190–7.

[Messier et al., 1996] Messier, W., Li, S. H., and Stewart, C. B. (1996). The birth of microsatellites. *Nature*, 381(6582) :483.

[Mikkelsen et al., 2005] Mikkelsen, T. S., Hillier, L. W., Eichler, E. E., Zody, M. C., Jaffe, D. B., Yang, S. P., Enard, W., Hellmann, I., Lindblad-Toh, K., Altheide, T. K., Archidiacono, N., Bork, P., Butler, J., Chang, J. L., Cheng, Z., Chinwalla, A. T., deJong, P., Delehaunty, K. D., Fronick, C. C., Fulton, L. L., Gilad, Y., Glusman, G., Gnerre, S., Graves, T. A., Hayakawa, T., Hayden, K. E., Huang, X. Q., Ji, H. K., Kent, W. J., King, M. C., Kulbokas, E. J., Lee, M. K., Liu, G., Lopez-Otin, C., Makova, K. D., Man, O., Mardis, E. R., Mauceli, E., Miner, T. L., Nash, W. E., Nelson, J. O., Paabo, S., Patterson, N. J., Pohl, C. S., Pollard, K. S., Prufer, K., Puente, X. S., Reich, D., Rocchi, M., Rosenbloom, K., Ruvolo, M., Richter, D. J., Schaffner, S. F., Smit, A. F. A., Smith, S. M., Suyama, M., Taylor, J., Torrents, D., Tuzun, E., Varki, A., Velasco, G., Ventura, M., Wallis, J. W., Wendl, M. C., Wilson, R. K., Lander, E. S., Waterston, R. H., and Consortium, C. S. A. (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 437(7055) :69–87.

[Mitas, 1997] Mitas, M. (1997). Trinucleotide repeats associated with human disease. *Nucleic Acids Res*, 25(12) :2245–54.

[Morgulis et al., 2006] Morgulis, A., Gertz, E. M., Schaffer, A. A., and Agarwala, R. (2006). Windowmasker : window-based masker for sequenced genomes. *Bioinformatics*, 22(2) :134–141.

[Moxon et al., 1994] Moxon, E. R., Rainey, P. B., Nowak, M. A., and Lenski, R. E. (1994). Adaptive evolution of highly mutable loci in pathogenic bacteria. *Current Biology*, 4(1) :24–33.

[Nachman and Crowell, 2000] Nachman, M. W. and Crowell, S. L. (2000). Estimate of the mutation rate per nucleotide in humans. *Genetics*, 156(1) :297–304.

[Nadir et al., 1996] Nadir, E., Margalit, H., Gallily, T., and Ben-Sasson, S. A. (1996). Microsatellite spreading in the human genome : evolutionary mechanisms and structural implications. *Proc Natl Acad Sci U S A*, 93(13) :6470–5.

[Noor et al., 2001] Noor, M. A., Kliman, R. M., and Machado, C. A. (2001). Evolutionary history of microsatellites in the obscura group of drosophila. *Mol Biol Evol*, 18(4) :551–6.

[Oakes et al., 2007] Oakes, C. C., La Salle, S., Smiraglia, D. J., Robaire, B., and Trasler, J. M. (2007). A unique configuration of genome-wide dna methylation patterns in the testis. *Proc Natl Acad Sci U S A*, 104(1) :228–33.

[Ohta and Kimura, 1973] Ohta, T. and Kimura, M. (1973). Model of mutation appropriate to estimate number of electrophoretically detectable alleles in a finite population. *Genetical Research*, 22(2) :201–204.

[Orgel and Crick, 1980] Orgel, L. E. and Crick, F. H. C. (1980). Selfish dna - the ultimate parasite. *Nature*, 284(5757) :604–607.

[Paques and Haber, 1999] Paques, F. and Haber, J. E. (1999). Multiple pathways of recombination induced by double-strand breaks in saccharomyces cerevisiae. *Microbiology and Molecular Biology Reviews*, 63(2) :349–+.

[Payseur and Nachman, 2000] Payseur, B. A. and Nachman, M. W. (2000). Microsatellite variation and recombination rate in the human genome. *Genetics*, 156(3) :1285–1298.

[Petes, 2001] Petes, T. D. (2001). Meiotic recombination hot spots and cold spots. *Nat Rev Genet*, 2(5) :360–9.

[Podlutsky et al., 1998] Podlutsky, A., Osterholm, A. M., Hou, S. M., Hofmaier, A., and Lambert, B. (1998). Spectrum of point mutations in the coding region of the hypoxanthine-guanine phosphoribosyltransferase (hprt) gene in human t-lymphocytes in vivo. *Carcinogenesis*, 19(4) :557–66.

[Price et al., 2004] Price, A. L., Eskin, E., and Pevzner, P. A. (2004). Whole-genome analysis of alu repeat elements reveals complex evolutionary history. *Genome Res*, 14(11) :2245–52.

[Primmer and Ellegren, 1998] Primmer, C. R. and Ellegren, H. (1998). Patterns of molecular evolution in avian microsatellites. *Mol Biol Evol*, 15(8) :997–1008.

[Primmer et al., 1996] Primmer, C. R., Saino, N., Moller, A. P., and Ellegren, H. (1996). Directional evolution in germline microsatellite mutations. *Nat Genet*, 13(4) :391–3.

[Pupko and Graur, 1999] Pupko, T. and Graur, D. (1999). Evolution of microsatellites in the yeast saccharomyces cerevisiae : role of length and number of repeated units. *J Mol Evol*, 48(3) :313–6.

[Ramsay et al., 1999] Ramsay, L., Macaulay, M., Cardle, L., Morgante, M., degli Ivanissevich, S., Maestri, E., Powell, W., and Waugh, R. (1999). Intimate association of microsatellite repeats with retrotransposons and other dispersed repetitive elements in barley. *Plant J*, 17(4) :415–25.

[Richard and Dujon, 1996] Richard, G. F. and Dujon, B. (1996). Distribution and variability of trinucleotide repeats in the genome of the yeast saccharomyces cerevisiae. *Gene*, 174(1) :165–74.

[Richard and Dujon, 1997] Richard, G. F. and Dujon, B. (1997). Trinucleotide repeats in yeast. *Res Microbiol*, 148(9) :731–44.

[Richard and Paques, 2000] Richard, G. F. and Paques, F. (2000). Mini- and microsatellite expansions : the recombination connection. *EMBO Rep*, 1(2) :122–6.

[Rolfsmeier et al., 2000] Rolfsmeier, M. L., Dixon, M. J., and Lahue, R. S. (2000). Mismatch repair blocks expansions of interrupted trinucleotide repeats in yeast. *Mol Cell*, 6(6) :1501–7.

[Rolfsmeier and Lahue, 2000] Rolfsmeier, M. L. and Lahue, R. S. (2000). Stabilizing effects of interruptions on trinucleotide repeat expansions in saccharomyces cerevisiae. *Mol Cell Biol*, 20(1) :173–80.

[Rose and Falush, 1998] Rose, O. and Falush, D. (1998). A threshold size for microsatellite expansion. *Mol Biol Evol*, 15(5) :613–5.

[Rossi et al., 1990] Rossi, M. S., Reig, O. A., and Zorzopulos, J. (1990). Evidence for rolling-circle replication in a major satellite dna from the south american rodents of the genus ctenomys. *Mol Biol Evol*, 7(4) :340–50.

[Roy et al., 2000] Roy, A. M., Carroll, M. L., Nguyen, S. V., Salem, A. H., Oldridge, M., Wilkie, A. O., Batzer, M. A., and Deininger, P. L. (2000). Potential gene conversion and source genes for recently integrated alu elements. *Genome Res*, 10(10) :1485–95.

[Roy-Engel et al., 2001] Roy-Engel, A. M., Carroll, M. L., Vogel, E., Garber, R. K., Nguyen, S. V., Salem, A. H., Batzer, M. A., and Deininger, P. L. (2001). Alu insertion polymorphisms for the study of human genomic diversity. *Genetics*, 159(1) :279–90.

[Roy-Engel et al., 2002] Roy-Engel, A. M., Salem, A. H., Oyeniran, O. O., Deininger, L., Hedges, D. J., Kilroy, G. E., Batzer, M. A., and Deininger, P. L. (2002). Active alu element "a-tails" : size does matter. *Genome Res*, 12(9) :1333–44.

[Rubinsztein et al., 1995] Rubinsztein, D. C., Amos, W., Leggo, J., Goodburn, S., Jain, S., Li, S. H., Margolis, R. L., Ross, C. A., and Ferguson-Smith, M. A. (1995). Microsatellite evolution–evidence for directionality and variation in rate between species. *Nat Genet*, 10(3) :337–43.

[Sainudiin et al., 2004] Sainudiin, R., Durrett, R. T., Aquadro, C. F., and Nielsen, R. (2004). Microsatellite mutation models : insights from a comparison of humans and chimpanzees. *Genetics*, 168(1) :383–95.

[Schug et al., 1998a] Schug, M. D., Hutter, C. M., Noor, M. A., and Aquadro, C. F. (1998a). Mutation and evolution of microsatellites in drosophila melanogaster. *Genetica*, 102-103(1-6) :359–67.

[Schug et al., 1998b] Schug, M. D., Hutter, C. M., Wetterstrand, K. A., Gaudette, M. S., Mackay, T. F., and Aquadro, C. F. (1998b). The mutation rates of di-, tri- and tetranucleotide repeats in drosophila melanogaster. *Mol Biol Evol*, 15(12) :1751–60.

[Schug et al., 1997] Schug, M. D., Mackay, T. F., and Aquadro, C. F. (1997). Low mutation rates of microsatellite loci in drosophila melanogaster. *Nat Genet*, 15(1) :99–102.

[Shriver et al., 1993] Shriver, M. D., Jin, L., Chakraborty, R., and Boerwinkle, E. (1993). Vntr allele frequency distributions under the stepwise mutation model : a computer simulation approach. *Genetics*, 134(3) :983–93.

[Sia et al., 1997] Sia, E. A., Kokoska, R. J., Dominska, M., Greenwell, P., and Petes, T. D. (1997). Microsatellite instability in yeast : dependence on repeat unit size and dna mismatch repair genes. *Mol Cell Biol*, 17(5) :2851–8.

[Sibly et al., 2001] Sibly, R. M., Whittaker, J. C., and Talbot, M. (2001). A maximum-likelihood approach to fitting equilibrium models of microsatellite evolution. *Molecular Biology and Evolution*, 18(3) :413–417.

[Smit et al., 2004] Smit, A. F., Hubley, R., and Green, P. (1996-2004). Repeatmasker.

[Smith and Nicolas, 1998] Smith, K. N. and Nicolas, A. (1998). Recombination at work for meiosis. *Curr Opin Genet Dev*, 8(2) :200–11.

[Smith and Waterman, 1981] Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *J Mol Biol*, 147(1) :195–7.

[Sokol and Williams, 2005] Sokol, K. A. and Williams, C. G. (2005). Evolution of a triplet repeat in a conifer. *Genome*, 48(3) :417–26.

[Stephan, 1986] Stephan, W. (1986). Recombination and the evolution of satellite dna. *Genet Res*, 47(3) :167–74.

[Strand et al., 1993] Strand, M., Prolla, T. A., Liskay, R. M., and Petes, T. D. (1993). Destabilization of tracts of simple repetitive dna in yeast by mutations affecting dna mismatch repair. *Nature*, 365(6443) :274–6.

[Szostak et al., 1983] Szostak, J. W., Orr-Weaver, T. L., Rothstein, R. J., and Stahl, F. W. (1983). The double-strand-break repair model for recombination. *Cell*, 33(1) :25–35.

[Tachida and Iizuka, 1992] Tachida, H. and Iizuka, M. (1992). Persistence of repeated sequences that evolve by replication slippage. *Genetics*, 131(2) :471–8.

[Tautz, 1993] Tautz, D. (1993). Notes on the definition and nomenclature of tandemly repetitive dna sequences. *Exs*, 67 :21–8.

[Tautz et al., 1986] Tautz, D., Trick, M., and Dover, G. A. (1986). Cryptic simplicity in dna is a major source of genetic variation. *Nature*, 322(6080) :652–6.

[Taylor et al., 1999] Taylor, J. S., Durkin, J. M., and Breden, F. (1999). The death of a microsatellite : a phylogenetic perspective on microsatellite interruptions. *Mol Biol Evol*, 16(4) :567–72.

[Toth et al., 2000] Toth, G., Gaspari, Z., and Jurka, J. (2000). Microsatellites in different eukaryotic genomes : survey and analysis. *Genome Res*, 10(7) :967–81.

[Trivedi, 2006] Trivedi, S. (2006). Comparison of simple sequence repeats in 19 archaea. *Genet Mol Res*, 5(4) :741–72.

[Valdes et al., 1993] Valdes, A. M., Slatkin, M., and Freimer, N. B. (1993). Allele frequencies at microsatellite loci : the stepwise mutation model revisited. *Genetics*, 133(3) :737–49.

[Volfovsky et al., 2001] Volfovsky, N., Haas, B. J., and Salzberg, S. L. (2001). A clustering method for repeat analysis in dna sequences. *Genome Biol*, 2(8) :RESEARCH0027.

[Vowles and Amos, 2004] Vowles, E. J. and Amos, W. (2004). Evidence for widespread convergent evolution around human microsatellites. *PLoS Biol*, 2(8) :E199.

[Vowles and Amos, 2006] Vowles, E. J. and Amos, W. (2006). Quantifying ascertainment bias and species-specific length differences in human and chimpanzee microsatellites using genome sequences. *Mol Biol Evol*, 23(3) :598–607.

[Walsh, 1987] Walsh, J. B. (1987). Persistence of tandem arrays : implications for satellite and simple-sequence dnas. *Genetics*, 115(3) :553–67.

[Wang et al., 1980] Wang, R. Y., Gehrke, C. W., and Ehrlich, M. (1980). Comparison of bisulfite modification of 5-methyldeoxycytidine and deoxycytidine residues. *Nucleic Acids Res*, 8(20) :4777–90.

[Watson and Crick, 1953] Watson, J. D. and Crick, F. H. (1953). Molecular structure of nucleic acids ; a structure for deoxyribose nucleic acid. *Nature*, 171(4356) :737–8.

[Weber and Wong, 1993] Weber, J. L. and Wong, C. (1993). Mutation of human short tandem repeats. *Human Molecular Genetics*, 2(8) :1123–1128.

[Webster et al., 2002] Webster, M. T., Smith, N. G. C., and Ellegren, H. (2002). Microsatellite evolution inferred from human-chimpanzee genomic sequence alignments. *Proceedings of the National Academy of Sciences of the United States of America*, 99(13) :8748–8753.

[Wexler et al., 2005] Wexler, Y., Yakhini, Z., Kashi, Y., and Geiger, D. (2005). Finding approximate tandem repeats in genomic sequences. *Journal of Computational Biology*, 12(7) :928–942.

[Whittaker et al., 2003] Whittaker, J. C., Harbord, R. M., Boxall, N., Mackay, I., Dawson, G., and Sibly, R. M. (2003). Likelihood-based estimation of microsatellite mutation rates. *Genetics*, 164(2) :781–787.

[Wierdl et al., 1997] Wierdl, M., Dominska, M., and Petes, T. D. (1997). Microsatellite instability in yeast : dependence on the length of the microsatellite. *Genetics*, 146(3) :769–79.

[Wilder and Hollocher, 2001] Wilder, J. and Hollocher, H. (2001). Mobile elements and the genesis of microsatellites in dipterans. *Molecular Biology and Evolution*, 18(3) :384–392.

[Xu et al., 2000] Xu, X., Peng, M., and Fang, Z. (2000). The direction of microsatellite mutations is dependent upon allele length. *Nat Genet*, 24(4) :396–9.

[Yandava et al., 1997] Yandava, C. N., Gastier, J. M., Pulido, J. C., Brody, T., Sheffield, V., Murray, J., Buetow, K., and Duyk, G. M. (1997). Characterization of alu repeats that are associated with trinucleotide and tetranucleotide repeat microsatellites. *Genome Res*, 7(7) :716–24.

[Yeramian and Buc, 1999] Yeramian, E. and Buc, H. (1999). Tandem repeats in complete bacterial genome sequences : sequence and structural analyses for comparative studies. *Res Microbiol*, 150(9-10) :745–54.

[Yoder et al., 1996] Yoder, A. D., Cartmill, M., Ruvolo, M., Smith, K., and Vilgalys, R. (1996). Ancient single origin for malagasy primates. *Proc Natl Acad Sci U S A*, 93(10) :5122–6.

[Young et al., 2000] Young, E. T., Sloan, J. S., and Van Riper, K. (2000). Trinucleotide repeats are clustered in regulatory genes in saccharomyces cerevisiae. *Genetics*, 154(3) :1053–68.

[Zhu et al., 2000] Zhu, Y., Strassmann, J. E., and Queller, D. C. (2000). Insertions, substitutions, and the origin of microsatellites. *Genetical Research*, 76(3) :227–236.

[Zietkiewicz et al., 1998] Zietkiewicz, E., Richer, C., Sinnett, D., and Labuda, D. (1998). Monophyletic origin of alu elements in primates. *J Mol Evol*, 47(2) :172–82.

Annexe 1

BMC Bioinformatics

BioMed Central

Research article

Open Access

Detecting microsatellites within genomes: significant variation among algorithms

Sébastien Leclercq*[1,2], Eric Rivals[1] and Philippe Jarne[2]

Address: [1]LIRMM, UMR 5506 CNRS – Université de Montpellier II, 161 rue Ada, Montpellier, France and [2]CEFE, UMR 5175 CNRS – Université de Montpellier II, 1919 route de Mende, Montpellier, France

Email: Sébastien Leclercq* - sebastien.leclercq@cefe.cnrs.fr; Eric Rivals - rivals@lirmm.fr; Philippe Jarne - philippe.jarne@cefe.cnrs.fr

* Corresponding author

Published: 18 April 2007

BMC Bioinformatics 2007, **8**:125 doi:10.1186/1471-2105-8-125

Received: 7 December 2006

Accepted: 18 April 2007

This article is available from: http://www.biomedcentral.com/1471-2105/8/125

Abstract

Background: Microsatellites are short, tandemly-repeated DNA sequences which are widely distributed among genomes. Their structure, role and evolution can be analyzed based on exhaustive extraction from sequenced genomes. Several dedicated algorithms have been developed for this purpose. Here, we compared the detection efficiency of five of them (TRF, Mreps, Sputnik, STAR, and RepeatMasker).

Results: Our analysis was first conducted on the human X chromosome, and microsatellite distributions were characterized by microsatellite number, length, and divergence from a pure motif. The algorithms work with user-defined parameters, and we demonstrate that the parameter values chosen can strongly influence microsatellite distributions. The five algorithms were then compared by fixing parameters settings, and the analysis was extended to three other genomes (*Saccharomyces cerevisiae*, *Neurospora crassa* and *Drosophila melanogaster*) spanning a wide range of size and structure. Significant differences for all characteristics of microsatellites were observed among algorithms, but not among genomes, for both perfect and imperfect microsatellites. Striking differences were detected for short microsatellites (below 20 bp), regardless of motif.

Conclusion: Since the algorithm used strongly influences empirical distributions, studies analyzing microsatellite evolution based on a comparison between empirical and theoretical size distributions should therefore be considered with caution. We also discuss why a typological definition of microsatellites limits our capacity to capture their genomic distributions.

Background

Microsatellites are genomic sequences comprised of tandem repeats of short nucleotide motifs (1 to 6 bp). They occur in all eukaryotic organisms and to a limited extent in prokaryotes, mostly in intergenic regions. Indeed, they may represent a significant part of genomes, for example about 3% of genome size (*i.e.*, millions of loci) in humans [1]. Microsatellite loci vary in length due to insertions or deletions (*i.e.*, indels) of one or more repeats, which are caused by a not-fully-understood molecular phenomenon, referred to as polymerase slippage [2,3]. A peculiarity of some loci, and the main reason for their wide use in biology, is hypermutability, with a slippage mutation rate of approximatively 0.001 mutation per locus par generation in humans [3]. Biologists have been interested in studying microsatellites for at least two reasons. First, some microsatellites are involved in molecular functions, such as recombination [4] or regulation of transcription

factors [5,6]. Others, present in coding regions, are involved in neurodegenerative disorders, including Fragile X Syndrome and Huntington's disease [7], and in some forms of cancer [8]. Second, they have been widely used as molecular markers in population biology [2,9]. High mutation rates result in extensive polymorphism within populations, and most microsatellites are selectively neutral. Therefore, understanding their evolutionary dynamics, especially the effect of mutation, is important [2]. These dynamics have been studied directly by analyzing the rate and nature of mutations in pedigrees [3,10]. An alternative approach uses distributions of microsatellites extracted from large stretches of DNA or fully sequenced genomes [11-13]. Theoretical distributions based on specified models of mutation can be fitted to these empirical distributions in order to infer the most appropriate model [14-17]. Hence, by understanding the evolutionary dynamics of microsatellites, we can gain both pure and applied knowledge about molecular evolution.

Given the size of sequenced genomes, microsatellite detection requires computer programs. Moreover, microsatellites may exhibit more or less complex nucleotide sequence, since stretches of tandem repeats may be interrupted by point mutations or indels and the detection of these is not trivial. A comparison of studies based on the genomic distribution of microsatellites reveals a surprising variability in the criteria used to detect microsatellites. For example, these criteria include the minimum or maximum repeat number [14-16,18], the motif type (e.g., AC) [17], or the minimum distance between successive microsatellites [16,17]. Another aspect of this variability is the method used to detect microsatellites: either it is not mentioned or it relies on home-made, poorly explained algorithms [19,20]. This variability is likely to affect empirical distributions of microsatellites, and therefore might affect the inferred mutation parameters. In addition, this comparison also reveals that imperfections (termed interruptions), are managed differently. Such imperfections are of a few types, including single mismatches in a locus, multiple mismatches at consecutive or non-consecutive positions, the succession of different motifs (compound microsatellites), and perfect microsatellites separated by several nucleotides (interrupted microsatellites) [21]. Imperfect microsatellites are generally excluded from studies, either by decomposing imperfect loci into perfect independant subparts, or by taking into account only perfect isolated loci. Both solutions provide a biased view of reality, because imperfections result from the evolutionary process, and influence the evolutionary dynamics by restricting the slippage rate [22-24]. A more integrated view on microsatellites requires more sophisticated and dedicaded algorithms.

At least a dozen detection algorithms have been described in the literature over the last ten years and they are based on three main approaches. First, combinatorial algorithms [25-27] scan genomic sequences linearly and detect tandem repeats as sub-sequences following specific construction rules. Various rules have been proposed, but these methods guarantee exhaustive detection of all subsequences corresponding to the rules. The second group of methods [28-30] uses algorithms that first scan genomic sequences to detect regions that may be microsatellites under given statistical rules. These regions are then submitted to validation tests that sieve out desired sequences. This pool of sequences may not be exhaustive because some sub-sequences that could pass validation tests may not be detected by statistical tests. However, these algorithms are time-efficient, and appropriate statistical criteria insure relevant results. In the third approach, algorithms align a given motif, or library of motifs, along genomic sequences [31,32]. Regions detected as microsatellites are those whose alignment score is higher than a given threshold.

The rules leading to microsatellite detection are clearly defined for all these algorithms. However, it is likely that because they are based on different mechanisms they will detect different sets of microsatellites. Moreover, the rules upon which some of these algorithms rely are defined by parameters whose value can be set by the user (this is not true of all algorithms). Detections can also be affected by the genomic sequence under consideration because of differences among the genomes (e.g., structure, GC content, and gene composition). As far as we know, no study has been conducted to compare the relative efficiency of these approaches and to evaluate how the parameter settings of given algorithms can affect empirical microsatellite distributions. Here, we analyze the distributions of mono- to hexanucleotide microsatellites using five algorithms representative of the different classes of methods, namely Mreps [27], Sputnik [33] (first approach), TRF [29] (second approach), RepeatMasker [31], and STAR [32] (third approach). Three of them (Sputnik, TRF, and RepeatMasker) are rather widely used by biologists. These distributions were characterized by microsatellite number and size, divergence from pure microsatellites (*i.e.*, imperfection level), and genomic position. Most of the analyses were conducted using the genomic sequence of the human X chromosome, but some analyses were also conducted in three other genomes of very different size and structure (*Saccharomyces cerevisiae*, *Neurospora crassa*, and *Drosophila melanogaster*). For three algorithms (Sputnik, TRF, and Mreps), we first evaluated the influence of variable parameter settings, and then we compared the five algorithms with fixed parameter values of Sputnik, TRF, and Mreps.

Results

Parameter influence

The number of detections with TRF increases exponentially as the alignment score decreases from 50 to 20 (default alignment weights {2,7,7}; Table 1). This increase is paralleled by an important reduction of the average length, and a more limited reduction in divergence. The variation in detection number is mainly due to the minimum size of detections, which is correlated to the score (Figure 1a). However, for microsatellites larger than 25 bp, which are not affected by the minimum size constraint, the number of detections is still significantly larger at lower score (ANCOVA on distributions in the range 25–70 bp, $F_{3,180} = 65.2$, $P < 0.0001$). Also note in Figure 1a the approximatively exponential decrease in detection number with length regardless of score, at least for lengths of less than 50. Modifying alignment weight also affects the number of detections, though to a more limited extent (Table 1; 61% increase between {2,7,7} and {2,3,5}). Interestingly, this is related to the detection of longer (larger than 30 bp [see Additional file 1]), more divergent microsatellites. For example, the average divergence grows from about 4% to 11.3% (Table 1). Decreasing alignment penalties for different minimum scores (20 to 40) reveals the same tendancy, with an increase in average detection length and divergence [see Additional file 2]. The validation score and mismatch penalty of Sputnik have the same effect as the alignment score and weights of TRF (Table 1). The number of detections increases exponentially as the validation score decreases because the minimum size of detections decreased. However, contrary to TRF, the validation score does not affect distributions of detections that are larger than the threshold size (Figure 1b) (ANCOVA on distributions in the range 20–70 bp, $F_{3,200} = 0.749$, $P = 0.524$). Smaller values of mismatch penalty greatly increase the average divergence (from 0,01% with a -10 penalty to 1.19% with a -5 penalty) and slightly increase the number of detections and average length (8.5% and 4% respectively). This means that microsatellites detected with a -5 penalty are essentially a set of enlarged microsatellites detected with a -10 penalty, due to better tolerance to imperfections. The influence of Mreps resolution parameter parallels that of alignment weights in TRF and mismatch penalty of Sputnik. Indeed, larger resolution values lead to larger and more divergent detections (Table 1). Between resolutions 1 and 6, the number of detections is 25% higher, while the corresponding increase for average length and average divergence are 73.4% and 114%. Again, this means that greater values of resolution essentially enlarge existing detections by allowing more errors. Examples of detections for different parameter settings of TRF, Sputnik, and Mreps are provided in Table 2.

Comparison of algorithms for perfect detection

Algorithms were first executed on the human X chromosome with TRF theshold score set to 20, TRF alignment weights to {2,7,7}, Mreps resolution to 1, and Sputnik mismatch penalty and validation score to -6 and 7 respectively (as explained in the *Methods* section). The distribution of perfect detections was studied first. The absolute numbers of detections are critically different, with a 80-fold ratio between the two extreme values, returned by Sputnik and RepeatMasker (6228 and 76 detections per megabase respectively). TRF (1913 detections/Mb) is three times less efficient than Sputnik, while STAR and Mreps return 135 and 285 detections/Mb respectively.

The comparison of length distributions revealed that the differences among algorithms depend mainly on the minimum detection length (Figure 2). For detections larger than 20 bp, the number of detections by Mreps and STAR are smaller than those of Sputnik, TRF, and RepeatMasker, for all motif classes except di- and trinucleotides (where Mreps was much less efficient). These differences are highly significant for all motif classes (ANCOVA on distributions in the range 20–70 bp, all $P \leq 0.01$), except for penta- and hexanucleotides due to a lack of power($F_{4,50} = 1.08$, $P = 0.376$, $F_{4,35} = 0.223$, $P = 0.923$). It could be noticed that the 'humps' in the di- and tetranucleodite distributions previously reported [16,20] are equally detected by all algorithms. For small sizes (less than 20 bp), striking differences are observed among algorithms. First, RepeatMasker is highly constrained by its internal minimum-size threshold, which prevents detection of microsatellites that are smaller than 20 bp. On the other hand, TRF and Sputnik essentially detect microsatellites that are smaller than 15 bp for all motif classes, especially tetra- to hexanucleotides. Indeed, very short (8–12 bp) tetra- to hexanucleotides, representing detections with 2 to 2.5 repeats, are about 3.7-fold more numerous than mono- to trinucleotides of 8–12 bp (4 to 12 repeats) for TRF, and 2-fold for Sputnik. The minimum-size effect is also clearly visible with Mreps. Detection starts at 11 bp for dinucleotides, 12 bp for trinucleotides, and up to 15 bp for hexanucleotides. This explains why Mreps detects far fewer microsatellites than TRF and Sputnik. STAR distributions are very different from those returned by the three other algorithms under 20 bp, with the number of detections increasing rather than decreasing. The maximum number of detections of STAR is generally reached around 20 bp, except for dinucleotides for which the number of detections starts to decrease beyond 15 bp. Microsatellites below these sizes are at the limit to yield a local increase in compression gain. In such cases, only regions that are near enough from the previous detection are reported (see Delgrange and Rivals [32], for details).

Table 1: Number of detections per megabase, average length (bp), and average divergence (%) of detections for combinations of parameters in the human X chromosome.

		number	length	divergence
TRF				
minimum score				
	50	110	64.44	3.96
	40	202	47.65	3.68
	30	458	32.14	3.21
	20	2425	16.07	1.60
align. weights				
	2,7,7	110	64.44	3.96
	2,7,5	125	73.62	6.01
	2,5,5	136	76.44	7.13
	2,3,5	177	83.30	11.31
Mreps				
resolution				
	1	1368	22.96	12.39
	2	1539	28.11	18.47
	3	1636	32.21	22.15
	6	1712	39.80	26.51
Sputnik				
minimum score				
	20	154	34.55	1.13
	15	349	25.39	1.06
	8	4273	11.23	0.48
	7	6589	9.74	0.44
Sputnik				
mismatch penalty				
	-10	6555	9.33	0.01
	-6	6589	9.74	0.44
	-5	6818	10.12	1.19

TRF alignment weights were set to {2,7,7} when varying the minimum threshold score, and the minimum threshold score to 50 when alignment weights varied. Mreps resolution was 1, 2, 3, and 6. Sputnik mismatch penalty was set to -6 when varying the minimum threshold score, and the minimum threshold score to 7 when varying the mismatch penalty. Match bonus and fail score were always fixed to 1 and -1, respectively. Divergence is deduced from the alignment of the detected sequence with the perfectly repeated corresponding sequence of focal consensus motif:

divergence = (*substitutions* + *insertions* + *deletions*)/*alignment length*).

Statistical tests were not performed for distributions of short detections (under 20 bp) because detection levels ensure critical differences.

Comparison of algorithms for imperfect detections

Differences among algorithms for the detection of imperfect microsatellites in the human genome do not follow those observed for perfect ones. Sputnik (resp. TRF) detects only 2-fold (2.9-fold) more imperfect microsatellites than RepeatMasker, compared to the 80-fold (25-fold) ratio for perfect detections (Table 3). Moreover, Sputnik and TRF detect respectively almost 17- and 4-times less imperfect microsatellites than perfect ones, while the other algorithms detect about 2- to 4-times more imperfect than perfect microsatellites. The average length and divergence are negatively related to the number of detections for TRF, Sputnik, STAR, and RepeatMasker. For example, the highest average length and divergence are obtained for RepeatMasker, which also exhibits the lowest number of detections. The average length and number of detections are directly linked to the minimum detection length (20 bp), which prevents detection of many short microsatellites, but also increases the average divergence level (because longer microsatellites are proportionnally more imperfect; see Discussion). Similarly, high average length and divergence, and low detection number for STAR are explained by its limited capacity to detect short microsatellites (Figure 2). Interestingly, Mreps shows the reverse pattern, with the largest number of detections (1084 detections/Mb, 6-fold more than RepeatMasker) obtained for the shortest, more divergent loci.

When perfect and imperfect microsatellites are considered at once (Table 3), Sputnik is the most efficient in terms of the number of detections, followed by TRF and Mreps, while STAR and RepeatMasker still yield a much lower number of detections. Note also that the average size of imperfect detections is larger than the average of all detections, for all algorithms except Mreps. This confirms that imperfect and perfect microsatellites detected by Mreps have about the same length.

An important issue is whether the detections returned by the five algorithms occur at the same physical locations in genomes. This was evaluated through the 'coverage' parameter. More than 93.5% of RepeatMasker and STAR detections are also detected by Sputnik, TRF, and Mreps, with a full coverage of RepeatMasker by Sputnik (Table 4). On the other hand, the coverage of Sputnik, TRF, and Mreps by STAR and RepeatMasker is much lower (< 34% for Mreps, < 20% for TRF, and < 10% for Sputnik; Table 4). This is consistent with the fact that the latter algorithms detect more microsatellites than the former. Notably, the coverage between algorithms is also consistent with the number of detections (e.g., STAR detected 16% fewer microsatellites than TRF and 17% of the sequences detected by TRF were also detected by STAR). This suggests that detections common to the five algorithms are generally located at the same positions.

The coverage can also be estimated in nucleotide numbers. This method yields a slightly different answer than the one provided by the number of detections (Table 4). On the whole, frequent detections are associated with small microsatellites (Table 4; under the diagonal). The reverse pattern is observed above the diagonal of Table 4.

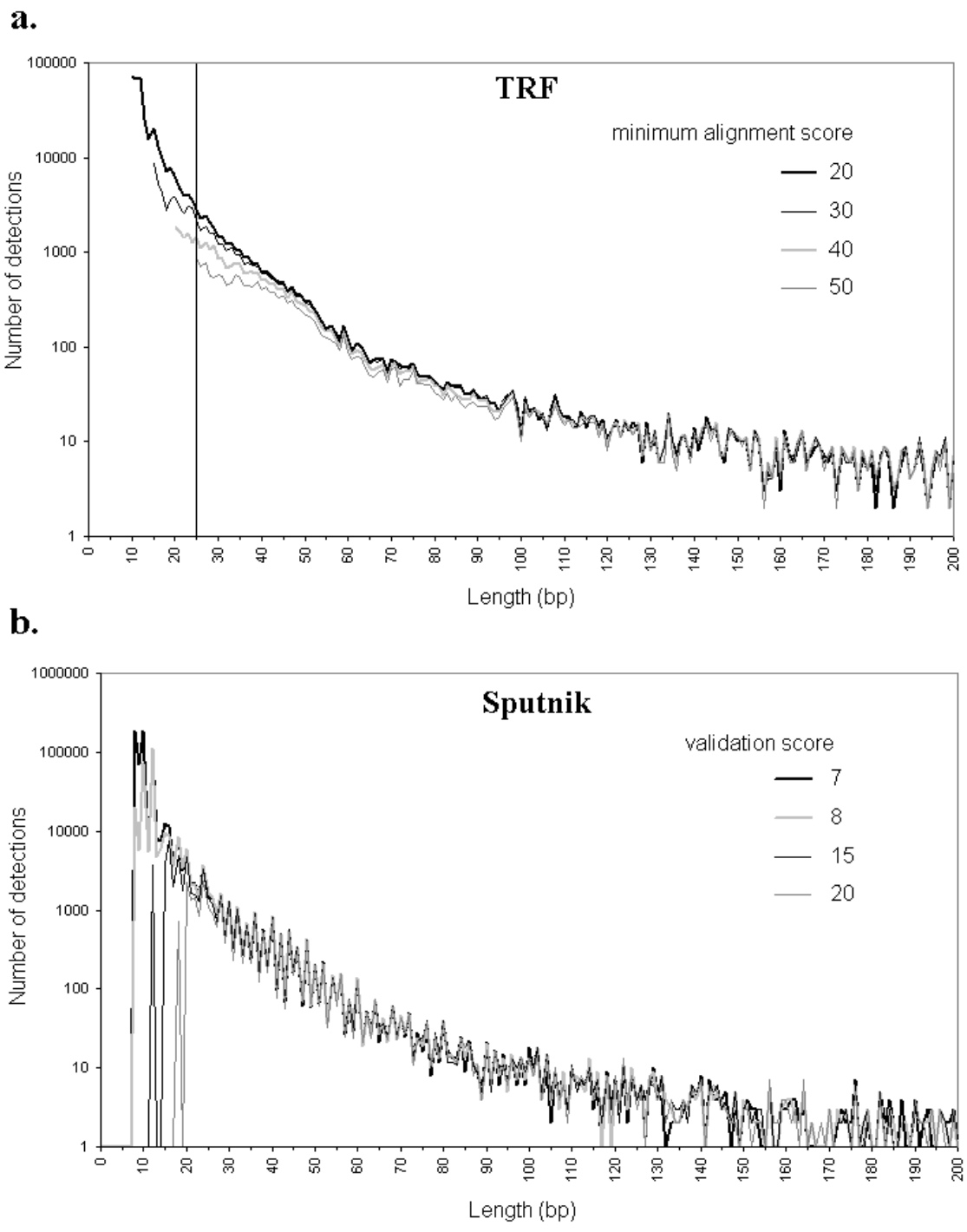


Figure 1

Length distributions for different minimum threshold scores of TRF. **a-** Number of detections (log scale) with TRF in the human X chromosome as a function of length (in bp) for minimum threshold score between 20 and 50. The alignment weights were {2,7,7}, and the few detections larger than 200 bp were discarded. The solid vertical line represents the minimum length not affected by the threshold score constraint. **b-** Number of detections (log scale) with Sputnik in the human X chromosome as a function of length (in bp) for validation score set to 7, 8, 15, and 20. The mismatch penalty was -6, and the few detections larger than 200 bp were discarded.

Table 2: Detection sample obtained with TRF with different alignment weights, Sputnik with different mismatch penalty, and Mreps with different resolution, in the human X chromosome.

		start	end	divergence	motif	sequence
TRF alignment scores						
2,7,7		304646	304658	0	CTCTC	CTCTCCTCTCCTC
		304696	304713	5.55	TCCTC	TCCTCTTCTCTCCTCTCC
		305863	305872	0	CCTTC	CCTTCCCTTC
2,5,7	**c**	304646	304713	18.3099	TCTCC	CTCTCCTCTCCTCCTTCTCCGCTCCCTGCACTGCCCTCCGCTCCCTCCGGTCCTCTTCT CTCCTCTCC
		305863	305872	0	TTCCC	CCTTCCCTTC
2,5,5		304646	304713	18.0556	TCTCC	CTCTCCTCTCCTCCTTCTCCGCTCCCTGCACTGCCCTCCGCTCCCTCCGGTCCTCTTCT CTCCTCTCC
	e	305836	305872	17.9487	TTCCC	CCCTCTCCACTTCCTTCTCTTCC**AC**C**T**CCTTCCCTTC
2,3,5	**e**	304643	304713	18.9189	CTCCT	CTGCTCTCCTCTCCTCCTTCTCCGCTCCCTGCACTGCCCTCCGCTCCCTCCGGTCCTC TTCTCTCCTCTCC
	n	305765	305800	25.641	CCA	CCACACCACCTCTGACGCCCACCACAGCCCCCCACC
		305836	305872	17.9487	CCCTT	CCCTCTCCACTTCCTTCTCTTCCACCTCCTTCCCTTC
Sputnik mismatch penalty						
-10		552928	552935	0	AG	GAGAGAGA
		552939	552948	0	AG	GAGAGAGAGA
		552954	552963	0	AAGAG	AAGAGAAGAG
		552964	552975	0	AG	AGAGAGAGAGAG
-6		552928	552935	0	AG	GAGAGAGA
		552939	552948	0	AG	GAGAGAGAGA
	c	552954	552975	9.09	AAGAG	AAGAGAAGAGAGAGAGAGAGAG
-5	**c**	552928	552948	9.52	AG	GAGAGAGAAAGGAGAGAGAGA
		552954	552975	9.09	AAGAG	AAGAGAAGAGAGAGAGAGAGAG
Mreps resolution						
1		119591	119610	20	AAT	ACAAAAAATAATAATTATAA
		119611	119628	5.56	AAAAA T	ATAAATAAAAATAAAAAT
2	**e**	119591	119615	24	AAT	ACAAAAAATAATAATTATAAATAAA
		119611	119628	5.56	AAAAA T	ATAAATAAAAATAAAAAT
3	**c**	119591	119638	33.33	A	ACAAAAAATAATAATTATAAATAAATAAAAATAAAAATTCAACTGTAA
6	**e**	119590	119638	34.69	A	TACAAAAAATAATAATTATAAATAAATAAAAATAAAAATTCAACTGTAA

Threshold alignment score of TRF was set to 20 and alignment weights varied from {2,7,7} to {2,3,5}. Sputnik mismatch penalty was set to -10, -6, and -5. Mreps resolution value varied from 1 to 6. For each detection, we report the start/end positions, divergence from a pure repeat, motif and actual sequence. Variation of detection when reducing weights is as follows: **n**: newly detected sequence; **e**: enlargement of a previous sequence; **c**: concatenation of previous sequences. New nucleotides detected by enlarging or concatenating previous sequences are underlined. The sequence at position 305765 is an example of a microsatellite detected at low values of alignment weights of TRF. It cannot be detected with alignment weights down to {2,3,5} because correct match bonuses cannot compensate for imperfection penalties. Reducing alignment weights may also enlarge detections, as shown for alignment weights {2,5,5} at position 305836. A succession of close errors (in boldface) decreases the alignment score, which falls under the threshold score for weight values larger than {2,5,5}. Reducing alignment weights also provokes concatenation, when an enlarged tandem repeat overlaps with one of its neighbors. At position 304696, two substitutions (in boldface), stops detection when alignment weights are set to {2,7,7}. With a smaller substitution penalty (5 or less), the detection is enlarged up to position 304646 and overlaps with the other detection. Reducing Sputnik mismatch penalty allows detection of larger microsatellites, by concatenating shorter, perfect ones. The two detections at position 552928 and 552939 are concatenated with a mismatch penalty of -5, because the penalty induced by two errors at position 552936 and 552938 are compensated by the second detection. A second concatenation occurs at position 552964 with a mismatch of -6. The two merged detections are not of the same motif, but the two errors induced by this difference are compensated by the matching bases with low values of mismatch penalty.
A larger resolution value for Mreps enlarges already-detected tandem repeats. In the first part of the tandem repeat at position 119591, adjacent repeats are separated by at most one error, and this part is detected at resolution 1; however repeats TAT and AAA are separated by two errors, so the second part can only be found at resolution 2 or higher. Finally, increasing resolution provokes concatenation. Detections for resolution 2 at positions 119591 and 199611 are enlarged when resolution is 3; both periods are reduced to 1 (see explanations in Methods), and the two sequences are merged.

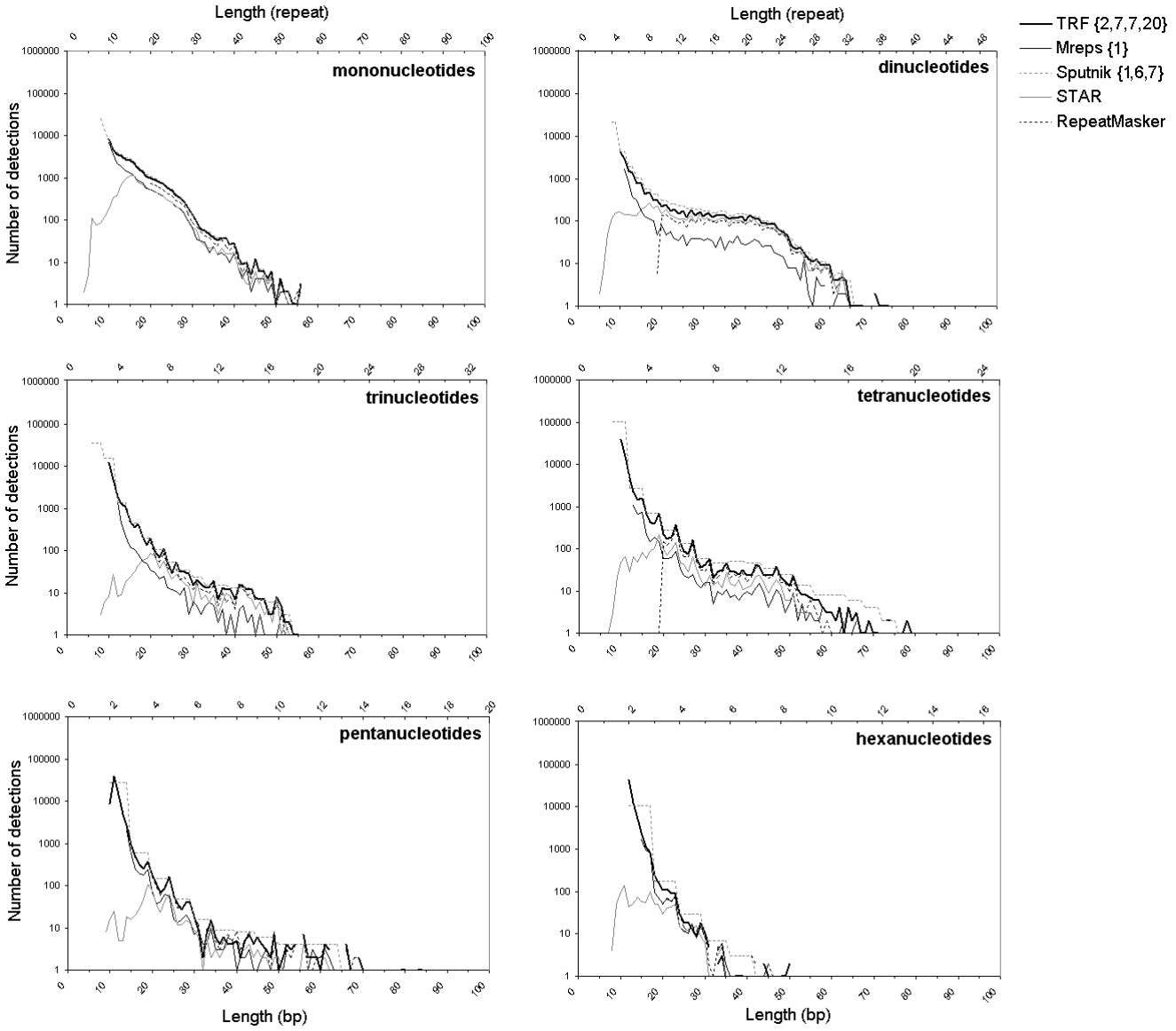


Figure 2

Length distributions of perfect detections for each algorithms. Number of perfect detections (log scale) in the human X chromosome as a function of length (in bp) for the six motif classes and for each algorithm. Sputnik groups all detections with a decimal number of repeats into the previous integer number of repeat class. The numbers of detections were averaged by motif size to display values for lengths representing a decimal number of repeats.

This is again likely due to the difference in average detection sizes for the five algorithms: for example, TRF detections covered by STAR and RepeatMasker are the longest ones.

Comparison of organisms

The algorithms were executed on three other genomic sequences and the results are presented in Table 3. The number of detections per Mbp was larger for *N. crassa*, *H. sapiens*, and *D. melanogaster* than for *S. cerevisiae*, although the difference is not significant (Kruskal-Wallis test, $H_{observed} = 0.85$, $d.f. = 3$, $P = 0.837$). On the other hand, a lot of variation was detected among algorithms for a given genome, as previously observed for the human X chromosome (Kruskal-Wallis test, $H_{observed} = 17.7$, $d.f. = 4$, $P = 0.001$). Interestingly, algorithms rank exactly in the same order for the four species with regard to the number of detections.

Table 3: Number of detections per Mbp, average length, and average divergence for TRF, Mreps, Sputnik, STAR, and RepeatMasker, in the genome of four species.

		HS		DM	NC	SC
		All	Imperfect			
detection number						
	TRF {2,7,7;20}	2425	512	3119	2902	1822
	Mreps 1	1368	1084	1653	1371	879
	Sputnik {1,-6,7}	6589	361	7475	7665	5712
	STAR	395	260	311	343	182
	RepeatMasker	256	179	207	230	104
average length						
	TRF {2,7,7;20}	16.07	28.84	14.24	14.61	13.85
	Mreps 1	22.96	24.99	20.04	20.93	20.28
	Sputnik {1,-6,7}	9.74	19.83	9.39	9.35	8.98
	STAR	39.89	49.80	31.07	32.86	33.12
	RepeatMasker	53.97	64.93	48.52	45.80	54.88
average divergence						
	TRF {2,7,7;20}	1.60	7.59	1.61	1.47	1.35
	Mreps 1	12.39	15.65	11.46	10.10	11.71
	Sputnik {1,-6,7}	0.44	7.96	0.46	0.38	0.32
	STAR	7.45	11.33	7.98	6.44	7.59
	RepeatMasker	8.40	11.97	13.42	9.31	13.14

Both imperfect and all (perfect plus imperfect) detections are provided for the human genome while all detections only are reported for the other genomes. HS = *Homo sapiens*, SC = *Saccharomyces cerevisiae*, DM = *Drosophila melanogaster*, NC = *Neurospora crassa*. Divergence is deduced from the alignment of the detected sequence with the perfectly repeated corresponding sequence of focal consensus motif:

divergence = (*substitutions* + *insertions* + *deletions*)/*alignment length*).

Comparing length and divergence also provides similar values among species for a given algorithm when considering all microsatellites (Table 3; Kruskal-Wallis tests, $H_{observed} = 0.337$, $d.f. = 3$, $P = 0.953$ and $H_{observed} = 0.577$, $d.f. = 3$, $P = 0.902$ for average length and divergence, respectively). Length distributions of perfect microsatellites in *S. cerevisiae*, *N. crassa*, and *D. melanogaster* show patterns similar to those observed in humans [see Additional file 3, 4, 5]. As for the number of detections, extensive variation is observed among algorithms for a given genome (Table 3; Kruskal-Wallis tests, $H_{observed} = 18.29$, $d.f. = 4$, $P = 0.001$ and $H_{observed} = 17.37$, $d.f. = 4$, $P = 0.002$ for average lengths and divergences, respectively). The rank order of algorithms was the same as described previously, the only exception being in *D. melanogaster* and *S. cerevisiae* where Mreps divergence is lower than that of RepeatMasker.

Discussion and conclusion

We compared the performance of five algorithms, four of which have been developed for detecting tandem repeats. The logic underlying microsatellite detection by these five algorithms is representative of the three main approaches that are currently available (see Introduction). In order to analyze the performance of these algorithms as fully as possible, we considered several parameters (number of loci detected, length, divergence, and redundancy), the six motif lengths corresponding to the classical definition of microsatellites (mono- to hexanucleotides), and four dif-

Table 4: Loci and nucleotide coverage between algorithms

		B				
		Sputnik {1,-6,7}	**TRF {2,7,7;20}**	**Mreps**	**STAR**	**RepeatMasker**
A	**Sputnik {1,-6,7}**	-	34.94 *(58.81)*	20.4 *(47.9)*	9.51 *(39.02)*	7.37 *(36.98)*
	TRF {2,7,7;20}	85.61 *(72.82)*	-	45.3 *(54.72)*	17.26 *(32.69)*	12.6 *(27.08)*
	Mreps	82.63 *(59.82)*	80.85 *(67.73)*	-	33.34 *(39.03)*	24.63 *(32.37)*
	STAR	95.29 *(69.56)*	93.92 *(80.03)*	93.61 *(77.31)*	-	57.98 *(66.83)*
	RepeatMasker	100 *(66.39)*	97.89 *(75.43)*	97.64 *(73)*	82.13 *(76.2)*	-

Proportion of the total number of detections (perfect and imperfect) of algorithm A also detected (*i.e.*, covered) by algorithm B in the human X chromosome. The value in brackets is the proportion of nucleotides detected by A and covered by B.

ferent genomes. Our first conclusion is that in algorithms where parameter values can be modified by the user, the settings of these parameters is critical. For example, increasing TRF minimum score and Sputnik validation score allows detection of 20- to 40-times more microsatellites, especially those that are smaller and more perfect. Conversely longer and more imperfect microsatellites were detected by decreasing TRF weights, Sputnik mismatch penalty, and increasing Mreps resolution. Therefore, modifying parameter settings has important consequences.

Interestingly, this variation was not reported in the original articles [27,29] in which detection efficiency was evaluated with respect to execution time (e.g., between resolution 1 and 20 for Mreps). Delgrange and Rivals [32] noticed though the large variation in results associated with parameters setting in TRF, but were not concerned with size or divergence level. Extending our comparison to five algorithms provides generally similar results. On the whole, RepeatMasker and STAR detect fewer and longer microsatellites than TRF and Mreps (both perfect and imperfect microsatellites). Divergence is also larger for RepeatMasker and STAR than for TRF. Sputnik results are similar to those of TRF, despite a different algorithmic approach. The microsatellite sets detected by the five algorithms are also very different: on the whole, most microsatellites detected by RepeatMasker and STAR are also detected by TRF, Sputnik, and Mreps, while the reverse is far from true. Such conclusions are likely generalizable because similar results were obtained in four genomes of different sizes and GC contents. Although RepeatMasker and STAR were classified in the third approach (see Methods) while Mreps, Sputnik, and TRF are representatives of the first and second approaches, respectively, we do not conclude that the third approach generally differs in efficiency from the other two approaches.

These results require some explanation. First, the striking difference among algorithms (or even for different parameters of the same algorithm) are mainly due to differential detection of short microsatellites, especially perfect ones. The bulk of microsatellites in genomes are short (*i.e.*, less than 12 bp). More precisely, microsatellites (at least perfect ones) exhibit a negative exponential size distribution within genomes [15,16,20,34]. Large threshold sizes (e.g., with RepeatMasker, or TRF with score sets to 50) or sharp constraints on size imposed by the significance threshold (the compression gain in STAR) therefore prevents detection of the majority of microsatellites. A noteworthy contribution to these short detections by Sputnik and TRF is made by tetra-, penta-, and hexanucleotides. These microsatellites with two-to-three repeats make almost one half of the total number of microsatellites detected by TRF and they are much more numerous than expected. For example, $(ACTGGT)_2$ roughly has a probability of $0.59^6 \times 0.41^6$ of occurence in the human genome, corresponding to about 7.3 detections on the X chromosome. TRF returned 826 detections, more than 100 times the expected value. Interestingly, the same patterns were detected in the four genomes studied. We cannot offer any clear explanation to the occurrence of these short repeats. However, even when short microsatellites are not taken into account, the five algorithms do not return the same sets of detections, therefore exhibiting different efficiencies. One reason is that the same repeat region might be interpreted differently by the five algorithms. These differences in detections are illustrated in Table 5 where some long, imperfect detections reported by RepeatMasker and STAR are decomposed into much smaller detections by Mreps (resolution 1), TRF (parameters setting {2,7,7;20}), and Sputnik (parameters setting {1,-6,7}).

Second, Mreps detected more divergent microsatellites than the other four algorithms. This might partly be due to compound microsatellites, *i.e*, succession of motifs such as $(AT)_6(AG)_5$. Based on our definition, which considers only one motif per detection, such detections are ascribed to a single motif, here $(AT)_{11}$.

The right part of the sequence is read as $(AT)_5$ with five errors, giving a 20% divergence. Such a compound microsatellite is erroneously counted as one short imperfect detection, and would be better counted as two shorter perfect detections of different motifs. The wide average divergence is also induced by the absence of a validation score in Mreps. Such a score imposes a minimum number of correct repeats for detections to be validated. Because increasing the proportion of wrong repeats reduces the number of correct ones, detections must be longer to reach a given score. The absence of such a constraint in Mreps results in short detections that can be as divergent as long ones.

Third, the limited differences detected among the four genomes studied were not fully unexpected, though smaller than those that have been previously reported [13,15]. This result suggests that the evolution of microsatellites is related to forces that are little affected by local processes or characteristics, either genomic (e.g., GC rate, density of transposable elements) or populational (e.g., effective population size). Microsatellites are affected by two types of mutations, *i.e.*, slippage and point mutations. It might be that the net outcome of their action does not vary among genomes larger than a few tens of millions base pairs, as are those studied here.

Our results have some practical implications. First, it has become common practice, when genomes are newly sequenced, to evaluate the relative size of genomic frac-

Table 5: RepeatMasker, STAR, Mreps, TRF and Sputnik detections between starting positions 532800 and 53500 in the human X chromosome.

	start	end	divergence	motif	sequence
RepeatMasker					
	531688	531713	0	AAT	AATAATAATAATAATAATAATAATAA
	532355	532540	15.05	TTCC	TTCCTTCCTCCCTTCCTTCCTTCCTTTCTTCTTTCTTTCTTTCCTTCCTTCC TGCTTTCCTTCCTTCC TTTCTTTTCTTTCTTTCCTTCCTTCCTTGCTTCCTTCCTTCCATCTTTCTCT TTCTCTTTTTCTTTCT TTCTCTCCTTCCTTCTTTCCTTCCTTCCTTCCCTTCCCTTCCTTCCTTCC
	532704	532891	15.87	TTCC	CCTTCCTTCCTTTCTTCTTTCTTTCCTTCCTTCCTTGCTTCCTTCCTTCCAT CTTTCTTTCTTTCTTT CTTCCTCTCCTTCCTTCTTTCCTTCCTTCCTTCCCTTCCCTCCTTCCTTTT TCTTCTTCTCTTTCTTT CTTTCTCTTTCCTTCCTTCCTTCCTTCTTTCTCCTTCCTTCCTTCTTTCCTT
STAR					
	531688	531713	0	AAT	AATAATAATAATAATAATAATAATAA
	532537	532731	25.38	TTTTTC	TTCCTTTTTCTTCTTCTCTTTCTTTCTTTCTTTTTCTTTCCTTCCTTCCTTCT TTCTCCTTCCTTCCT TCCATTTTTCTTTCTTTCTTTCTTTCTCTCTCTCTCTTTCTTTCTTTCT CTCTCTCTCTTCTTC CTTCCTTCCTTCCATTCTTCTTTCTTTCTTTCCTTCCTTCCTTTCTTCTTTC TTTCCTT
Mreps					
	531688	531715	3.45	AAT	AATAATAATAATAATAATAATAATAAAA
	532330	532429	15.84	TTCC	TTTCCTTCTTTCTTTCTTACTTTCTTTCCTTCCTCCCTTCCTTCCTTCCTTT CTTCTTTCTTTCTTTC CTTCCTTCCTGCTTTCCTTCCTTCCTTTCTTT
	532428	532467	12.5	TTCC	TTTCTTTCTTTCCTTCCTTCCTTGCTTCCTTCCTTCCATC
	532466	532490	4	TTTCTC	TCTTTCTCTTTCTCTTTTTCTTTCT
	532491	532524	11.76	TTCC	TTCTCTCCTTCCTTCTTTCCTTCCTTCCTTCCCT
	532525	532542	5.56	TTCC	TCCCTTCCTTCCTTCCTT
	532551	532593	13.95	TTTC	TCTCTTTCTTTCTTTCTTTTTCTTTCCTTCCTTCCTTCTTTCT
	532593	532609	5.88	TTCC	TCCTTCCTTCCTTCCAT
	532609	532667	16.95	TC	TTTTTCTTTCTTTCTTTCTTTCTCTCTCTCTCTTTCTTTCTTTCTCTCT CTCTCT
	532667	532689	8.7	TTCC	TTCTTCCTTCCTTCCTTCCATTC
	532690	532756	11.94	TTCC	TTCTTTCTTTCTTTCCTTCCTTCCTTTCTTCTTTCTTTCCTTCCTTCCTTGC TTCCTTCCTTCCATC
	532755	532777	4.35	TTTC	TCTTTCTTTCTTTCTTTCTTCCT
	532776	532820	8.89	TTCC	CTCTCCTTCCTTCTTTCCTTCCTTCCTTCCCTTCCCTCCTTCCTT
TRF {2,7,7;20}					
	531688	531713	0	AAT	AATAATAATAATAATAATAATAATAA
	532313	532330	5.26	TTTTC	TTTTCTTTTCTTTCTTTT
	532423	532438	5.88	TTTTC	TTTCTTTTCTTTCTTT
	532466	532490	4	TTTCTC	TCTTTCTCTTTCTCTTTTTCTTTCT
	532544	532553	0	TTC	TTCTTCTTCT
	532550	532576	13.79	TTTCTC	TTCTCTTTCTTTCTTTCTTTTTCTTTC
	532633	532667	8.57	TC	TCTCTCTCTCTCTTTCTTTCTTTCTCTCTCTCTCT
Sputnik {1,-6,7}					
	531568	531576	0	ACC	ACCACCACC
	531688	531711	0	AAT	AATAATAATAATAATAATAATAAT
	531849	531856	0	TTGC	CTTGCTTG
	531893	531900	0	TG	TGTGTGTG
	531927	531934	0	ATGC	TGCATGCA
	532078	532085	0	AGGC	GCAGGCAG
	532266	532273	0	ATGC	TGCATGCA
	532313	532322	0	TTTTC	TTTTCTTTTC

Table 5: RepeatMasker, STAR, Mreps, TRF and Sputnik detections between starting positions 532800 and 53500 in the human X chromosome. *(Continued)*

532335	532354	5	TTTC	TTCTTTCTTTCTTACTTTCT
532355	532422	10.29	TTCC	TTCCTTCCTCCCTTCCTTCCTTCCTTTCTTCTTTCTTTCTTTCCTTCCTTCC TGCTTTCCTTCCTTCC
532423	532439	5.88	TTTC	TTTCTTTTCTTTCTTTC
532440	532463	4.17	TTCC	CTTCCTTCCTTGCTTCCTTCCTTC
532466	532489	4.17	TTTCTC	TCTTTCTCTTTCTCTTTTTCTTTC
532500	532541	7.14	TTCC	TCCTTCTTTCCTTCCTTCCTTCCCTTCCCTTCCTTCCTTCCT
532544	532552	0	TTC	TTCTTCTTC
532553	532568	0	TTTC	TCTTTCTTTCTTTCTT
532569	532576	0	TTTC	TTTCTTTC
532577	532588	0	TTCC	CTTCCTTCCTTC
532596	532607	0	TTCC	TTCCTTCCTTCC
532615	532656	7.14	TTTC	TTTCTTTCTTTCTTTCTTTCTCTCTCTCTCTTTCTTTCTTTC
532657	532666	0	TC	TCTCTCTCTC
532669	532684	0	TTCC	CTTCCTTCCTTCCTTC
532687	532692	0	TTC	TTCTTC
532693	532704	0	TTTC	TTTCTTTCTTTC
532705	532752	8.33	TTCC	CTTCCTTCCTTTCTTCTTTCTTTCCTTCCTTCCTTGCTTCCTTCCTTC
532755	532774	0	TTTC	TCTTTCTTTCTTTCTTTCTT
532780	532820	7.32	TTCC	CCTTCCTTCTTTCCTTCCTTCCTTCCCTTCCCTCCTTCCTT

Resolution of Mreps was set to 1, threshold alignment score of TRF to 20 and alignment weights of TRF to {2,7,7}. Sputnik mismatch penalty and validation score were set to -6 and 7, respectively. The number of detections varies with algorithms (from 3 to 18). Moreover, the sequence information is dealt with in different ways; an example is the region of cryptic simplicity between positions 532815 and 533080. RepeatMasker and STAR decompose it into large, distant and highly imperfect detections, though not the same for the two algorithms. Mreps returns a succession of shorter detections, overlapping the whole region. TRF detects only short, not much divergent, subregions, which do not completely overlap with the whole region. Sputnik detections are very numerous, short and slightly divergent, but overlap the whole region. Detection of compound microsatellites by Mreps is illustrated at position 533706, where other algorithms detect only a perfect polyA strech. The detection at position 534186 is returned as two detections by Mreps, because the two consecutive errors (insertions of G and C) stop the detection when resolution is set to 1. Very short hexanucleotides (12 bp) are detected by both TRF and Sputnik at positions 533138 and 534112. Most detections of Sputnik are two-repeat tetranucleotides, or three-repeat trinucleotides, which cannot be detected by other algorithms.

tions (e.g., coding *versus* non-coding sequences). Although microsatellites do not constitutes a large parts of genomes (a few percent), estimates of their density will depend on the algorithm used and on the algorithm parameters. As an example, the International Human Genome Sequencing Consortium [1] estimated that mono- to hexanucleotides constitute 1.5% of the human genome using TRF with weight {2,3,5} and score 50. This value rises to 3.9% using weight {2,7,7} and score 20, as we did here. Second, and less trivial, our results might interfere with conclusions on microsatellite evolution drawn from genomic approaches. As mentioned in Introduction, microsatellite evolution can be derived by fitting models specifying mutation parameters (e.g., slippage and point mutation rates) to actual length distributions of microsatellites. These distributions are generally built using home-made, often poorly described algorithms. For example, out of a representative set of studies following this approach [14-20,35], only two describe the algorithm used. It is therefore not clear how one can compare these approaches to the approach defined in Introduction and whether the fitted distributions are representative of the actual set of microsatellites. For those studies fitting length distributions, the extensive variation in the number of detections among algorithms might be less of an issue than the shape of the length distributions (negative exponential). However, it might be interesting to fit distributions with outcomes from different algorithms to evaluate the amount of variation in the inferred mutation processes. Another problem is that these studies focus completely on perfect microsatellites. Detecting such tandem repeats is algorithmically less complex than detecting imperfect ones, for example by scanning specific motifs and extending the search to neighbouring positions, or searching adjacent identical subsequences of the same length, whatever the motif (both approaches are combinatorial). Perfect microsatellites can also be detected using the other approaches, and filtering out *a posteriori* imperfect microsatellites. Whatsoever, we have seen that the set of returned perfect microsatellites depends on the algorithm. This, again, might affect the inference of mutational processes acting on microsatellites. Moreover, restricting the analysis to perfect microsatellites does not render the full complexity of microsatellite evolution. Note that, although some studies included imperfect microsatellites [36-38], the chosen parameters were so stringent that detection was close to perfect. These examples show an attempt to integrate imperfection into microsatellite studies, but none discuss the implications of their parameter choices, despite the non-negligible influence of imperfections on microsatellite evolutionary dynamics [22-24].

One reason why different sets of perfect microsatellites are detected by different algorithms relies on the choice of different minimum distances separating two successive microsatellites. From an algorithmic point of view, two tandemly-repeated stretches, each of the same motif, and separated by a single (or a few) nucleotide(s) (e.g., $(CA)_{10}G(CA)_{10}$) can be considered as two perfect microsatellites. From an evolutionary point of view, such a sequence is best viewed as a single imperfect microsatellite resulting from an insertion within a perfect microsatellite. A less rhetorical example can be drawn from the literature. Dieringer *et al.*, Calabrese and Durrett, and Lai and Sun [15,16,20] all looked for dinucleotides in the human genome, but used different definitions. For Lai and Sun, a detection was considered as perfect when none of the four bases on its left side were included in another detection. For Calabrese and Durrett, perfect detections must be separated by at least 50 bp and should not include a repeat of the focal motif within the 4 bp flanking sequences. Divergently, Dieringer *et al*. considered all perfect subparts as independent microsatellite detections. Counting only those detections equal to 10 repeats (from Tables and Figures in these references), the detection numbers are about 100000, 4500, and 163000 for Dieringer *et al.*, Calabrese and Durrett, and Lai and Sun respectively.

More generally, our results highlight the problem of defining a microsatellite. The simple widely-used definition is the one given in Introduction (tandem repeats of short nucleotide motifs; perfect if the same motif is repeated without interruptions, imperfect or compound otherwise [21]). However, these definitions are not precise enough to aid in decidions regarding which nucleotide regions are microsatellites. Indeed, they do not characterize the minimal required length, nor the level of imperfection. For example, compound microsatellites set specific challenges to detection methods, as mentioned above. Some attempts have been done to generalise the definition of microsatellites, for example by introducing wildcarded motifs [25]. In this case, a compound microsatellite *ATATATATACACACAC* is defined as a $(A^*)_8$, where * can be replaced by any nucleotide. Other authors [39] provided a first attempt to distinguish between complex and compound microsatellites, and to return them in a comprehensive way (e.g.,$(AT)_4(AC)_4$).

This typological definitions are those retained in the combinatorial algorithms used above. An important line of research would be to design new algorithms that couple microsatellite detection and the inference of the most parsimonious history of duplications and point mutations for the region being analysed [40,41]. The tandem repeat detected would then be described by both its sequence and history of duplications. In a duplication history, different motifs may be duplicated and such an approach would authorize several motifs to be involved in the formation of a single tandem repeat, as in compound microsatellites. The duplication history would help in both delimiting the tandem repeat and producing an explicit consensus sequence.

Methods

Algorithms

The comparative analysis was conducted using Mreps (version 2.11), Sputnik (modified version from M. Morgante 06-2001), TRF (version 3.21 for Windows), RepeatMasker (version 13-07-2002), and STAR. We will first describe at some length the logic and algorithm of these programs, because this is instrumental for understanding variation among returned microsatellite sets. In what follows, 'microsatellite' refers to those sequences we searched for, under the definitions given below. Their number, exact sequence, and positions in the genomic sequence are not known. 'Detections' are those sequences returned by algorithms. Their number is exactly known, as well as their sequence and position.

We first used Mreps which is based on the combinatorial Hamming distance algorithm for the detection of approximate repeats [42]. This algorithm considers that two adjacent sub-sequences, or repeats, with the same period (*i.e.*, repeat size) in a given sequence are part of the same tandem repeat if they differ by at most k mismatches. The process progresses along the sequence by comparing successive repeats and stops when two adjacent repeats differ by $k + 1$ errors. The whole detected region is called a k-tandem repeat, and it is of distance k. For example, a perfect tandem repeat is defined by $k = 0$. Mreps searches for all possible k-tandem repeats with all possible periods (up to half the length of the sequence analyzed) and k lying between 0 and a parameter value called esolution. As the Hamming Distance method stops when $k + 1$ errors are detected, both extremities of detected regions are artefactually lengthened by erroneous nucleotides. These nucleotides are deleted by Mreps during a phase called *edge trimming*. Mreps then computes the best shortest period minimizing the average error rate of detected repeated areas (for example, transforming a periodically repeated tetranucleotide *ATAT* into a dinucleotide *AT*). The error rate is calculated as *error number*/(*length* – *period*), where *length* is the length of the repeated region, *period* the repeat period, and *error number* the sum of distances between all adjacent repeats. Repeats with the same best period and overlapping over at least two periods are assembled as a unique detection. Detections are filtered out in order to eliminate those expected in a random sequence, based on two filters. The *length filter* eliminates all detections smaller than *period* + 9 bases (e.g., 11 bp for dinucleotides). The *quality filter* removes detections whose error

rate and length do not satisfy internal conditions of significance. These conditions are pre-calculated by analyzing results obtained with Mreps in a pseudo-random genome, but are not detailed in Mreps documentation. Note that Mreps does not work with motifs, but with periods, so that results correspond to motif length, not to given motifs. Moreover, the Hamming distance method cannot handle indels, but this can be accounted for by using large k values. Indeed, indels disrupt the repeat phase, but not the repeat period. Consequently, if the distance between the two phases is smaller or equal to k, the two subsequences with different phases are considered as the same microsatellite.

The second sofware is Sputnik, which is based on a combinatorial method. Scanning the sequence from left to right, Sputnik considers that adjacent similar subsequences with the same period as part of the same tandem repeat. Adjacent sub-sequences are compared with the first sub-sequence of the detection. Matches increase the global score, while mismatches decrease it, and a detection is validated when reaching a threshold score. When an error decreases the score below a fail score set by the user, the comparison stops and the score is returned. Errors can be substitutions, insertions or deletions. In order to discriminate between these three possibilities, the comparison is recursively performed three times from the erroneous base. The three resulting scores are compared to the score before the erroneous position and the highest is returned. The starting position of validated detections is the first base of the first subsequence and the last position corresponds to the base associated to the best score. The algorithm resumes the procedure after this last position, for periods two to five. A post-treatment is finally applied to reduce the size of each detection to a multiple of its period.

TRF is probably the most popular algorithm for detecting tandem repeats. It was, for example, used by the International Human Genome Sequencing Consortium to detect microsatellites in the human genome sequence [1]. TRF scans sequences in order to determine regions where motifs are periodically repeated, though not necessarily tandemly repeated, based on a set of statistical rules detailed in the TRF article [29]. The most appropriate motif is then determined for each region, and this motif is aligned along the region using a Wraparound Dynamic Programming (WDP) algorithm [43]. The WDP procedure takes as input a motif and a sequence; it yields an optimal global alignment between the sequence and a perfect tandem repeat of the motif. WDP optimizes both the alignment score and the number of repeats of the motif. A score is computed from this alignment by attributing a positive weight to each correctly aligned nucleotide (*matches*), and a negative weight to substitutions (*mismatches*) and insertions-deletions (*indels or gaps*). Alignment weights can be adjusted by users, but only to a limited extend in the Windows version. When the alignment score is higher than a threshold (that can also be adjusted by the user), the alignment is returned as detection with the corresponding consensus motif. Different motifs can be aligned along a single region, in which case the three best detections only are returned. Note that the best alignment(s) might be shorter than the initially detected region.

The fourth algorithm used in this study is RepeatMasker. It was initially developed for both extracting and masking interspersed repeats from DNA sequences. As microsatellites potentially occur anywhere in genomes, they can also be considered as interspersed repeats and are searched for by RepeatMasker. However, it should be noted that RepeatMasker was not primarily developed for such a task.

RepeatMasker works with a library of reference sequences of 180 bp, each one representing the perfect repeated sequence of a given motif (e.g., $(CA)_{90}$ or $(GATA)_{45}$). RepeatMasker cuts the analyzed sequence in 40 Kbp pieces, overlapping over 1 Kbp. Alignment with the target sequences is based on perfect match over at least 14 bp based on the Smith-Waterman method [44]. Perfect matches separated by less than 14 bp are grouped together to constitute a single repeated region. This is conducted using the cross match program. A Smith-Waterman score is computed for the region based on predefined weights for perfect matches, substitutions, and indels. Weights are given by RepeatMasker and depend on the GC content of the 40 Kbp analyzed subsequence. The regions retained as detections are those with a Smith-Waterman score higher than a threshold (cutoff score; which can be modulated by the user). Overlapping detections are managed as follows: a detection covered over 80% of its length (or more) by another detection with a better score is not returned. RepeatMasker uses the Repbase Update Library [45] as default reference library. As some simple repeated sequences were found to be rare in the human genome, they were not included in Repbase [11]. Some penta- and hexanucleotide sequences are also missing. We therefore created a custom library containing all 501 possible reference sequences of mono- to hexanucleotide microsatellites (964 motifs with complementary ones) with sequence length set to 180 bp.

The last software we considered is STAR, which is based on a sequence-compression method, and uses the informativity of tandem repeats compared to non-repeated sequences. More specifically, STAR takes a motif as parameter, and uses a WPD algorithm [43] in order to align this motif all along the query sequence. The aligned sequence

is then encoded using a lossless compression method: the encoded sequence is a succession of numbers of perfectly aligned bases (e.g., AAAAAAAAAA is encoded as 10 for motif A), and separated by encoded mismatches and indels. Good alignments lead to small encoded sequence, while the encoded sequence can be larger than the original one when the fraction of mismatches or indels is high. STAR computes a compression gain for each sequence position, as the ratio between original and encoded sequence sizes from sequence origin to this base. The gain increases in repeated regions and decreases in others. STAR uses an optimization procedure that detects the boundaries of these regions. A detection must start and end at matching positions, and series of non-matching positions could be interpreted as a non-repeated sequence between two detections, or as errors in a single detection. STAR chooses the best altenative to maximize the compression gain over the whole sequence. Algorithmic Information Theory ensures that compressible regions are significant repeated regions, which cannot be found in random sequences [46]. There is currently no statistical theory that enables one to compute the significance of an approximate tandem repeat. Thus, the rationale followed in STAR is to use the compression gain for testing the significance of a detected tandem repeat (facilitated by the Algorithmic Information Theory, also known as the Kolmogorov Complexity Theory.) and optimizing this gain globally for a set of detected tandem repeats [32]. STAR aims at finding all and only significant approximate tandem repeats of a given motif according to this criterion. STAR does not report overlapping detections because a given run focuses on one motif only, and two overlapping regions with the same motif form the same tandem repeat.

Parameters

For these five softwares, except STAR, some input parameters are left to the user and we describe here their functions and implementations. Mreps parameters are the minimum and maximum lengths of detections (in bp), the minimum number of repeats, and the minimum and maximum motif lengths. These parameters do not affect algorithm execution, but are used to filter out final results returned to the user. Recall though that detections with a length shorter than *period* + 9 are automatically removed (see above).

The Hamming Distance algorithm used by Mreps runs for *k*-values that are independent of the tandem repeat period. When *k* is small, large periods are penalized because few errors are allowed between adjacent repeats. Therefore, almost all sequences detected are perfect tandem repeats. On the other hand, small periods are not detected for high *k* values because only periods up to *k*+1 are searched for. The resolution parameter was implemented to bypass this problem, by running the algorithm from all values between 1 and the resolution value. This may produce overlapping detections of same periods, which are merged when they overlap over at least two periods. As a consequence, this merging step may return larger repeat regions.

Sputnik has seven standard parameters, which can be set by the user, namely the match bonus and mismatch penalty, the validation score, the fail score, the maximum number of recursions, the minimum percentage of perfection, and the period size. As for TRF, high penalty values define more stringent conditions, and the minimum detection length is directly linked to the match bonus and the validation score. Too many close errors in a row drive the score below the fail score which stops the recursion. Setting a low fail score allows merging close microsatellites with the same motif, depending on their length. The maximum number of recursions can be considered as an absolute maximum number, which stops the recursion. The minimum percentage of perfection is used, in a post-treatment filter, to discard detections not reaching this threshold. The last parameter is the period size to be searched for. We used a version of Sputnik that allowed us to search for mono- to pentanucleotides [47]. Moreover, we modified the source code to take hexanucleotides into account. In addition to these standard parameters, we used the '-j' option. By default, the first period of a detection is not counted in the score, meaning that a pentanucleotide needs to be 15 bp long to reach a score of 10, while a mononucleotide only needs to be 11 bp long. The '-j' option allows inclusion of the first repetition into the score.

In the Windows version of TRF, three parameters can be adjusted, namely the maximum motif size, alignment weights, and minimum alignment (threshold) score. The first one is a post-treament filter removing all detections with a consensus motif size larger than a given size, and takes value between 1 and 2000 bp. Alignment weights and threshold score both influence the capacity of a detected region to be validated during the scoring phase. Alignment weights include a scoring bonus (*match*) and two scoring penalties (*mismatch*, *indel*). Weights with high penalty values define more stringent conditions, because errors are more penalized during the WDP scoring computation. For example, weights list {2,7,7} is more stringent than {2,3,5} and will detect fewer imperfect microsatellites. The threshold score is the minimal score that a repeated region should reach to be validated. A high score is therefore more stringent because detections of given length must have more matching positions. Note that both the weights and threshold influence the length of detected sequences. For example, if the match bonus is +2, a score of 20 will be reached for 10 correct matches,

while 25 correct matches are required to reach a score of 50. More generally, the minimum length is given by the ratio of the threshold to the match bonus.

For Repeatmasker, the cutoff value only is implemented when searching for microsatellites. It determines the minimum alignment score used by cross match to validate detections. This parameter has the same effect as the threshold score parameter of TRF: a smaller cutoff allows detecting more imperfect and/or shorter repeats, because imperfections decrease the score. However, the same cutoff value may select different sets of repeated sequences, because the scoring matrices, which are automatically chosen by RepeatMasker, depends on local GC content (see above). It is therefore difficult to evaluate how detection varies with the cutoff value. A final point is that Repeatmasker does not return detections smaller than 20 bp, independent of the cutoff value and scoring matrices.

Detection in STAR is based on differences between tandem repeats and their surrounding regions, and the complete set of information needed to run the algorithm is contained in the query sequence itself. The only information required is the type of tandem repeat, characterized by its motif. STAR does not use integrated filters based on minimum or maximum length, number of repetitions or imperfection level, and users must implement their own filters if needed.

Redundancy

The algorithms used may detect a given tandem repeat more than once for example, when two motifs with a valid detection value represent the same sequence or when two tandem repeats overlap. Redundancy has no biological meaning and essentially results from the methods implemented by the algorithms. However, from a biological point of view, a given base in a sequence belongs to a single microsatellite. Repeatmasker partly manages redundancy by returning the detection with the highest score (see above). TRF provides detections with the three best scores. Mreps and STAR do not manage redundancy. There is no redundancy in Sputnik detections, because a new search is always initiated after the end of the previous detection. To homogenize redundancy among results, we filtered out redundant repeated areas for the four algorithms using two rules. When the shortest detection of a pair of detections overlapped the longest one by 80% or more, we kept the detection with the lowest divergence from a pure motif (defined below). In case of equal divergence, or when overlap was less than 80%, the shortest detection was discarded. When two detections overlapped over less than five nucleotides, we always kept both detections.

Characterizing microsatellite distributions

Algorithms were compared based on five microsatellite characteristics, namely number, length, divergence compared to the consensus motif, motif class, and genomic position. As each algorithm idiosyncratically computes length and divergence depending on the detection method, we normalized definitions in order to compare algorithms. Length was defined as *end position – start position* + 1 in bp. This was preferred to *motif length* × *repeat number* in order to avoid difficulties when counting indels. Divergence was defined as the number of differences between a detection and the perfectly repeated corresponding sequence of the same alignment length for the consensus motif of the detection.

divergence = error number/alignment length with *error number = substitutions + insertions + deletions*, and *alignment length = substitutions + insertions + deletions + matching bases*. The algorithms used provide output values which are more or less related to divergence. Homology in TRF is the average rate of matches between adjacent repeats, based on local alignments only. Divergence could therefore not be computed from homology, and we scanned output alignment files to count both mismatches and indels. The definition of *div* in RepeatMasker differs from ours, since it provides *substitutions*/(*substitutions + matching bases*). However, RepeatMasker also returns three values (*ins*, *del*, and *length*) which are defined respectively as *ins = insertions*/(*insertions + substitutions + matching bases*), *del = deletions*/(*deletions + substitutions + matching bases*), and *length = substitutions + matching bases + insertions – deletions*. Numbers of matches, substitutions, and indels were deduced from these four values. Mreps error rate (see *Algorithm* section) cannot be used to estimate divergence. A WDP algorithm [43] (see description above) was applied to Mreps detections to get number of matches, substitutions, and indels. This algorithm uses a motif as input. However, Mreps detections are returned as a succession of same period repeat units, without any consensus motif. The consensus motif was defined as the most common repeated motif in the detection. Sputnik returns a percentage of perfection as 100 × (*reference sequence length -error number*)/*reference sequence length*. This value is not compatible with our definition of divergence, so the WDP algorithm was applied to

Sputnik detections as well. STAR gives directly the number of matches, substitutions, and indels per detection. Motif classes represent the different motif sizes of microsatellites. Six classes are defined for mono- to hexanucleotides. Detections are counted in the class of its shortest period only (e.g., $(AT)_{12}$ is counted only in class 2, and not in classes 4 or 6).

Execution

Genome sequences depend on the evolutionary history of organisms and specific genomes may therefore vary with regard to microsatellite distribution and structure. In order to provide a general picture of the efficiency of algorithms to detect microsatellites, our study was conducted using four fully sequenced genomes spanning a range of sizes and representing very different organisms. These are the unicellular fungi *Saccharomyces cerevisiae* [48] (version Jul 26, 2004) and *Neurospora crassa OR74A* [49] (version Feb 17, 2005), the arthropod *Drosophila melanogaster* [50] (build version 4.1 Jul 21, 2005), and the vertebrate *Homo sapiens* [1] (build version 35.1, Aug 29, 2004). Genome sizes are 12 Mb (*S. cerevisiae*), 43 Mb (*N. crassa*), 110 Mb (*D. melanogaster*), and 3200 Mb (*H. sapiens*), and their average GC-content is 38%, 50%, 35%, and 41% respectively. All sequences were downloaded from the NCBI genome page [51]. Our analysis was conducted on the whole fungi sequences, but restricted to the 2L and X chromosomes of *D. melanogaster* and *H. sapiens*, respectively. Their sizes are 22 Mb and 153 Mb, but the microsatellite distributions along these chromosomes are representative of that of their whole genome (data not shown). Note also that the human, fruit fly, and *N. crassa* genomes are not fully assembled, leaving some gaps in the sequences, represented as 'N' streches. Mreps replaces gaps with random series of nucleotides, which may create artificial tandem repeats. Tandem repeats detected within gaps were excluded from the analysis.

The five programs used have default parameter values, but changing parameters may critically change length and divergence distributions as explained above. The influence of parameters on detections were first analyzed for each algorithm independently using distributions of detections from the human X chromosome. TRF default values are 500 for the maximum motif length, {2,7,7} for alignment weights, and 50 for the minimum threshold score. Microsatellites have, by definition, a motif length between 1 and 6. However, the maximum motif length was set to 10, because size 6 is not proposed in the TRF version we used. All repeats with motif size larger than 6 were discarded from the analysis. The first analysis were performed using four threshold scores (20, 30, 40, and 50) with alignment weights fixed to default. The threshold score was then fixed to default, and alignment weights to {2,7,7}, {2,5,7}, {2,5,5}, and {2,3,5}.

The default cutoff value of Repeatmasker is 225, and Smith et al. [31] suggest using values in the range 200–250 to avoid detection errors (for lower values) and underdetection (for higher values). Results obtained with different values in this window were not significantly different (data not shown), so that 225 was the cutoff value in all results reported here. Minimum and maximum motif lengths were fixed at 1 and 6 when using Mreps, as for TRF, and the minimum number of repeats was fixed at 2, representing a single tandem repeat. Mreps was run with resolution value set to 1, 2, 3, and 6. Sputnik has default parameters 1, -6, 8, and -1 for the match bonus, mismatch penalty, validation, and fail scores, respectively. The program was first executed with the validation score set to 7, 8, 10, 15, and 20. It was then set to 7, and a second analysis was performed with mismatch penalty set to -5, -6, and -10. The minimum percentage of perfection is a post-treatment filter only and does not influences the algorithm itself, so it was not investigated. The maximum-recursion parameter was evaluated, but had no influence on results for values other than 0 (which returns only perfect microsatellites). Minimum and maximum motif lengths were fixed at 1 and 6. The only input parameter in STAR is the microsatellite motif, and it was run using all 501 non-redundant, non-cyclically equivalent motifs of 1 to 6 bp long already used to construct our RepeatMasker exhaustive library.

The performance of the five algorithms was then compared. However, parameters must be adjusted for TRF and Mreps before the comparison. Rose and Falush [34] showed that the number of perfect microsatellite loci is significantly higher than expected under a random (Bernoulli) model for lengths larger or equal to 8 bp. Parameters were fixed to return microsatellites larger than 8 bp. TRF and Sputnik do not have minimal length parameter, but the threshold score restricts the minimal size of detection. A minimal length of 8 bp requires a minimum threshold score of 16 for TRF and 7 for Sputnik (because the score must be stricly larger than the threshold for the detection to be validated); as 16 is not available in the Windows version of TRF, we used 20. The minimal length of Mreps was set to 8 bp, but the length filter eliminates all detections smaller than *period* + 9 bp, which *de facto* gives a minimal detection size of 10 bp for mononucleotides, 11 bp for dinucleotides, etc. As very long microsatellites are rare, though not absent, no maximum size was fixed in Mreps options. TRF alignment weights, Sputnik mismatch penalty, and Mreps resolution principally affect the divergence level, but this criteria is still largely unknown and no consensus or limit can be proposed at this time. We kept values advocated by developers, *i.e.*, {2,7,7} for the alignment weights of TRF, -6 for the mismatch penalty of Sputnik, and 1 for the resolution of Mreps (as resolution 0 provides only perfect detections).

Statistical methods

The variation in length distributions between different TRF threshold score parameters was analyzed using analyses of covariance (ANCOVA) under a linear regression model [52] Type III. Detection numbers were the dependant variable (in log10), length was a covariate, and the

parameter settings were included as a factor. As the distributions roughly follow a negative exponential in the window of 25–70 bp, the use of a linear regression model on the variable in log10 is appropriate. The variation in length distributions between Sputnik' validation scores was analysed using the same ANCOVA test in the range of 20–70 bp. Length comparisons between algorithms were also performed using ANCOVA tests, taking algorithms as a qualitative factor and using linear regressions. Distributions were normalized prior to analysis. Indeed, Sputnik shrinks all detections to the largest size multiple of the motif size, by discarding the incomplete end repeat. This means that all non-multiple lengths are lacking from the distributions, while multiple lengths are artificially increased. Linear regressions were performed on integer parts of the detection numbers, for the five algorithms. This critically decreases the power of the tests, especially for penta- and hexanucleotides, with regressions based on ten and eight points respectively. Comparison among species were conducted using Kruskal-Wallis tests on algorithms, for detection numbers, average lengths, and average divergences.

Authors' contributions

S.L. collected the data, ran the comparisons and formated the results. All authors conceived the study and contributed to the discussion. All authors were equally involved in writing the manuscript.

Additional material

Additional file 1

Number of detections (log scale) with TRF in the human X chromosome as a function of length (in bp) for different alignment weights.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-8-125-S1.pdf]

Additional file 2

Number of detections per megabase, average length (bp), and average divergence (%) of detections for combinations of parameters in the human X chromosome.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-8-125-S2.pdf]

Additional file 3

Length distributions of perfect detections (log scale) for the six motif classes and the five algortihms, on the 2L chromosome of Drosophila melanogaster.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-8-125-S3.pdf]

Additional file 4

Length distributions of perfect detections (log scale) for the six motif classes and the five algortihms, on the whole genome of Neurospora crassa.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-8-125-S4.pdf]

Additional file 5

Length distributions of perfect detections (log scale) for the six motif classes and the five algortihms, on the whole genome of Saccharomyces cerevisiae.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-8-125-S5.pdf]

Acknowledgements

We thank the CNRS department of information and engineering sciences for providing us with a computer cluster and the MAB team for his technical help, G. Benson and N. Galtier for helpful discussions, F. Massol for help with statistical analysis, Josh Auld for significantly improving english and three anonymous reviewers for comments on the manuscript. The authors are supported by research grants from the "Action Concertée Incitative – Informatique, Mathématiques, Physique pour la Biologie" and from the BioSTIC-LR program. S.L. is supported by a fellowship from the Ministère Fran cais de la Recherche.

References

1. Consortium IHGS: **Initial sequencing and analysis of the Human Genome.** *Nature* 2001, **409(6822):**860-921.
2. Goldstein D, Schlötterer C: *Microsatellites Evolution and Applications* Oxford University Press; 1999.
3. Ellegren H: **Microsatellites: simple sequences with complex evolution.** *Nat Rev Genet* 2004, **5(6):**435-45.
4. Benet A, Molla G, Azorin F: **d(GA × TC)(n) microsatellite DNA sequences enhance homologous DNA recombination in SV40 minichromosomes.** *Nucleic Acids Res* 2000, **28(23):**4617-22.
5. Martin P, Makepeace K, Hill S, Hood D, Moxon E: **Microsatellite instability regulates transcription factor binding and gene expression.** *Proc Natl Acad Sci USA* 2005, **102(10):**3800-4.
6. Moxon ER, Rainey PB, Nowak MA, Lenski RE: **Adaptive evolution of highly mutable loci in pathogenic bacteria.** *Current Biology* 1994, **4:**24-33.
7. Mitas M: **Trinucleotide repeats associated with human disease.** *Nucleic Acids Res* 1997, **25(12):**2245-54.
8. Arzimanoglou I, Gilbert F, Barber H: **Microsatellite instability in human solid tumors.** *Cancer* 1998, **82(10):**1808-20.
9. Jarne P, Lagoda PJL: **Microsatellites, from molecules to populations and back.** *Trends Ecol Evol* 1996, **11(10):**424-9.
10. Harr B, Schlötterer C: **Long microsatellite alleles in *Drosophila melanogaster* have a downward mutation bias and short persistence times, which cause their genome-wide underrepresentation.** *Genetics* 2000, **155(3):**1213-20.
11. Jurka J, Pethiyagoda C: **Simple repetitive DNA sequences from primates: compilation and analysis.** *J Mol Evol* 1995, **40(2):**120-6.
12. Pupko T, Graur D: **Evolution of microsatellites in the yeast *Saccharomyces cerevisiae*: role of length and number of repeated units.** *J Mol Evol* 1999, **48(3):**313-6.
13. Katti M, Ranjekar P, Gupta V: **Differential distribution of simple sequence repeats in eukaryotic genome sequences.** *Mol Biol Evol* 2001, **18(7):**1161-7.
14. Kruglyak S, Durrett R, Schug M, Aquadro C: **Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations.** *Proc Natl Acad Sci USA* 1998, **95(18):**10774-8.
15. Dieringer D, Schlötterer C: **Two distinct modes of microsatellite mutation processes: evidence from the complete**

genomic sequences of nine species. *Genome Res* 2003, **13(10):**2242-51.
16. Calabrese P, Durrett R: **Dinucleotide repeats in the Drosophila and human genomes have complex, length-dependent mutation processes.** *Mol Biol Evol* 2003, **20(5):**715-25.
17. Sainudiin R, Durrett R, Aquadro C, Nielsen R: **Microsatellite mutation models: insights from a comparison of humans and chimpanzees.** *Genetics* 2004, **168:**383-95.
18. Bell GI, Jurka J: **The length distribution of perfect dimer repetitive DNA is consistent with its evolution by an unbiased single-step mutation process.** *J Mol Evol* 1997, **44(4):**414-21.
19. Falush D, Iwasa Y: **Size-dependent mutability and microsatellite constraints.** *Mol Biol Evol* 1999, **16(7):**960-966.
20. Lai YL, Sun FZ: **The relationship between microsatellite slippage mutation rate and the number of repeat units.** *Mol Biol Evol* 2003, **20(12):**2123-31.
21. Chambers G, MacAvoy E: **Microsatellites : consensus and controversy.** *Comp Biochem Physiol B Biochem Mol Biol* 2000, **126(4):**455-476.
22. Jin L, Macaubas C, Hallmayer J, Kimura A, Mignot E: **Mutation rate varies among alleles at a microsatellite locus: phylogenetic evidence.** *Proc Natl Acad Sci USA* 1996, **93(26):**15285-8.
23. Petes TD, Greenwell PW, Dominska M: **Stabilization of microsatellite sequences by variant repeats in the yeast *Saccharomyces cerevisiae*.** *Genetics* 2000, **146(2):**491-8.
24. Taylor J, Durkin J, Breden F: **The death of a microsatellite : a phylogenetic perspective on microsatellite interruptions.** *Mol Biol Evol* 1999, **16(4):**567-72.
25. Landau GM, Schmidt JP, Sokol D: **An algorithm for approximate tandem repeats.** *J Comput Biol* 2001, **8:**1-18.
26. Castelo AT, Martins W, Gao GR: **TROLL-Tandem Repeat Occurrence Locator.** *Bioinformatics* 2002, **18(4):**634-6.
27. Kolpakov R, Bana G, Kucherov G: **mreps: Efficient and flexible detection of tandem repeats in DNA.** *Nucleic Acids Res* 2003, **31(13):**3672-8.
28. Coward E, Drablos F: **Detecting periodic patterns in biological sequences.** *Bioinformatics* 1998, **14(6):**498-507.
29. Benson G: **Tandem repeats finder: a program to analyze DNA sequences.** *Nucleic Acids Res* 1999, **27(2**573-80 [Http://tandem.bu.edu/trf/trf.html].
30. Wexler Y, Yakhini Z, Kashi Y, Geiger D: **Finding approximate tandem repeats in genomic sequences.** *J of Comput Biol* 2005, **12(7):**928-42.
31. Smit A, Hubley R, Green P: **RepeatMasker.** 1996 [Http://repeatmasker.org].
32. Delgrange O, Rivals E: **STAR : an algorithm to search to Tandem Approximate Repeats.** *Bioinformatics* 2004, **20(16**2812-20 [Http://atgc.lirmm.fr/star/].
33. Abajian C: **Sputnik.** 1994 [Http://espressosoftware.com/pages/sputnik.jsp].
34. Rose O, Falush D: **A threshold size for microsatellite expansion.** *Mol Biol Evol* 1998, **15(5):**613-5.
35. Kruglyak S, Durrett R, Schug MD, Aquadro CF: **Distribution and abundance of microsatellites in the yeast genome can be explained by a balance between slippage events and point mutations.** *Mol Biol Evol* 2000, **17(8):**1210-9.
36. Majewski J, Ott J: **GT repeats are associated with recombination on human chromosome 22.** *Genome Res* 2000, **10(8):**1108-14.
37. Kayser M, Vowles EJ, Kappei D, Amos W: **Microsatellite length differences between humans and chimpanzees at autosomal loci are not found at equivalent haploid Y chromosomal loci.** *Genetics* 2006, **173(4):**2179-86.
38. Webster MT, Smith NGC, Ellegren H: **Microsatellite evolution inferred from human-chimpanzee genomic sequence alignments.** *Proc Natl Acad Sci USA* 2002, **99(13):**8748-53.
39. Hauth A, Joseph D: **Beyond tandem repeats: complex pattern structures and distant regions of similarity.** *Bioinformatics* 2002, **18(Suppl 1:S):**31-7.
40. Rivals E: **A Survey on Algorithmic Aspects of Tandem Repeats Evolution.** *International J of Foundations of Computer Science* 2004, **15(2):**225-257.
41. Rivals E: **Algorithmes d'analyse de séquences en bioinformatique. Périodicité et répétitions.** Université Montpellier II. Montpellier, France; 2005.
42. Kolpakov R, Kucherov G: **Finding approximate repetitions under Hamming distance.** *Theor Comp* 2003, **303:**135-56 [Http://bioinfo.lifl.fr/mreps/].
43. Fischetti V, Landau G, Sellers P, Schmidt J: **Identifying periodic occurences of a template with applications to protein structure.** *Inf Proc Letters* 1993, **45:**11-18.
44. Smith TF, Waterman MS: **Identification of common molecular subsequences.** *J Mol Biol* 1981, **147:**195-7.
45. Jurka J: **Repbase update: a database and an electronic journal of repetitive elements.** *Trends Genet* 2000, **16(9**418-20 [Http://www.girinst.org].
46. Rivals E, Dauchet M, Delahaye JP, Delgrange O: **Compression and genetic sequence analysis.** *Biochimie* 1996, **78(5):**315-22.
47. Morgante M, Hanafey M, Powell W: **Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes.** *Nat Genet* 2002, **30(2):**194-200.
48. Goffeau A, Barrell B, Bussey H, Davis R, Dujon B, Feldmann H, Galibert F, Hoheisel J, Jacq C, Johnston M, *et al.*: **Life with 6000 genes.** *Science* 1997, **275(5303):**1051-2.
49. Galagan JE, *et al.*: **The genome sequence of the filamentous fungus *Neurospora crassa*.** *Nature* 2003, **422(6934):**859-68.
50. Adams M, Celniker S, Holt R, Evans C, Gocayne Jea: **The genome sequence of *Drosophila melanogaster*.** *Science* 2000, **287(5303):**2185-95.
51. **NCBI Genome Biology** [Http://ncbi.nih.gov/Genomes/]
52. Sokal R, Rohlf F: *Biometry : the principles and practice of statistics in biological research* W.H. Freeman; 1995.

TITRE : Origines des séquences microsatellites dans les génomes eucaryotes

---

RESUME

Les microsatellites, séquences répétées en tandem de période une à six paires de bases, sont des entités génomiques présents dans tous les organismes qu'ils soient animaux, végétaux ou microbiens. Ils présentent un cycle de vie caractérisé par trois phases principales : une apparition et une maturation, une dynamique à l'état mature, puis une dégénérescence. Nous nous intéressons dans cette thèse à la première phase, l'apparition des microsatellites.
Pour traiter cette question, nous nous sommes basés sur l'analyse de la séquence du génome humain. L'une des lacunes de ce type d'analyse est qu'il faut d'abord extraire les microsatellites du génome, et qu'il existe plusieurs algorithmes de nature et fonctionnement différents. La première partie de cette thèse se concentre donc sur la comparaison de quelques-uns des principaux algorithmes de recherche de répétitions en tandem, et dresse un portait des différentes qualités et limitations de chacun d'eux.
Deux possibilités majeures sont détaillées, l'apparition par l'intermédiaire d'éléments transposables (ETs), et l'apparition spontanée à partir d'une séquence quelconque. Dans le premier cas, l'étude est focalisée sur le rôle des queues polyA des séquences Alu chez les primates. La question de l'apparition à partir d'une séquence quelconque cherche à établir l'impact de trois mécanismes mutationnels différents sur la création et le développement primordial des microsatellites : la mutation ponctuelle, le glissement de polymérase et la micro-duplication adjacente de quelques nucléotides. Un modèle général d'apparition des microsatellites est aussi proposé, suggérant une dynamique d'apparition plus complexe que ce qui était précédemment supposé.

TITLE : Origins of microsatellites within eukaryotic genomes

ABSTRACT

Microsatellites, tandemly repeated sequences of period one to six basepairs, are genomic elements found in the genomes of all living species, from bacteria to humans. They exhibit a life cycle which can be decomposed into three major phases: an apparition and maturation phase, a mature dynamic phase, and a degeneration phase. This thesis focuses on the first phase, the microsatellite apparition.
This issue was investigated by analysing the complete human genome sequence. This requires extracting microsatellite loci from the sequence, using specific algorithms. However these available algorithms differs in detection method and efficiency. The first part of this thesis is dedicated to the comparison of some of the major algorithms of tandem repeats detection, and present an overview of their qualities and limitations.
Microsatellite birth essentially derived from transposable elements (TEs) or from mutation from any DNA sequence. The analysis of TE-mediated birth focuses on the role of Alu elements, using a comparative approach in three primate genomes. Mutation mediated birth is analysed though three different mechanisms: point mutation, DNA slippage and adjacent micro-duplication of a small number of nucleotides. A general model of microsatellite apparition is then proposed, suggesting a more complex apparition dynamic than was previously thought.

DISCIPLINES : *Evolution moléculaire, Bioinformatique, Génomique comparative*

MOTS-CLES : Microsatellite, Séquence répétée, Mutation, Micro-duplication, Glissement de polymérase, Elément Alu, Algorithme de détection, Humain

---

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

CEFE UMR 5175
Campus CNRS
1919 route de Mende
34293 Montpellier cedex 5