HESE DE DOCTORAT DE L'UNIVERSITE PIERRE & MARIE CURIE (PARIS VI)

SPECIALITE : GENETIQUE

Présentée par

Guillaume ACHAZ

Pour obtenir le titre de
Docteur de l'Université Pierre & Marie Curie (Paris VI)

Sujet de la thèse :

Etude de la dynamique des génomes : les répétitions intrachromosomiques.

Soutenue le 28 mai 2002
Devant le jury composé de :

Pr Dominique ANXOLABEHERE	Président
Dr Francis FABRE	Rapporteur
Dr Laurent DURET	Rapporteur
Dr Marie-France SAGOT	Examinatrice
Pr André GOFFEAU	Examinateur
Pr Pierre NETTER	Directeur de thèse
Dr Eric COISSAC	Codirecteur de thèse

Résumé

Les séquences répétées, ou répétitions sont une des clefs de voûte de l'évolution des génomes. Elles sont la marque d'un processus continu de duplications et de réarrangements des chromosomes de toutes les espèces. Par une méthode bioinformatique, nous avons recherché toutes les répétitions intrachromosomiques dans les séquences de 6 génomes eucaryotes et 53 génomes bactériens. Au cours de l'analyse de ces répétitions, nous avons pu montré que les répétitions directes (dont les deux copies sont dans la même orientation) présentent des caractéristiques différentes des répétitions inversées.

Tout d'abord, dans tous ces génomes, il n'existe quasiment pas de répétitions inversées dont les deux copies soient proches alors qu'il y a une grande abondance de répétitions directes proches.

Pour ces répétitions directes proches, nous avons observé *i*) une corrélation négative entre l'identité partagée par leurs deux copies et la distance qui les séparent et *ii*) une corrélation positive entre la longueur des copies et la distance qui les séparent.

A la suite de ces observations, nous avons proposé un modèle dynamique de l'évolution des répétitions intrachromosomiques qui semble s'appliquer à presque tous les génomes étudiés. Ce modèle propose qu'une grande partie de ces répétitions sont créées en tandem et sont remaniées postérieurement à leur duplication. Comme la plupart des répétitions des organismes analysés partagent des caractéristiques communes, on peut penser que l'apparition des mécanismes de duplication précède la divergence entre Eucaryotes, Bactéries et Archées.

Abréviations utilisées

ADN : Acide DésoxyriboNucléique.
ARN : Acide RiboNucléique.
ARNm : ARN messager
ARNr : ARN ribosomique
ARNt : ARN de transfert

pb : paires de bases
kb : kilobases (1000 pb)
Mb : Mégabases (1000 kb)
cM : centiMorgan
Mhz : Méga Hertz
Go : Giga octet

ORF : *Open Reading Frame*
ORI : ORIgine de la réplication bactérienne
TER :TERminaison de la réplication bactérienne

CDR : *Close Direct Repeat*
IS : *Insertion Sequence*
LINE : *Long INterspersed Element*
LTR : *Long Terminal Repeat*
ROS : Répétition d'Ordre Supérieur
RT : *Reverse Transcriptase*
SINE : *Short INterspersed Element*
SSR : *Simple Sequence Repeat*
T_{ADN} : Transposon à ADN
VNTR : *Variable Number of Tandem Repeat*

NHEJ : *Non Homologous End Joining*
SDSA : *Simple Dependant Strand Annealing*

CMH : Complexe Majeur d'Histocompatibilité
DHFR : Di-Hydro-Folate Reductase
Ig : Immunoglobuline
MTX : Méthotréxate
TCR : *T Cell Receptor*

DDC : *Duplication, Degeneration and Complementation*

Sommaire

Tables

Illustrations

Avant-propos

Une des barrières à la compréhension du monde vivant est son opulente biodiversité, présente tant chez nos plus proches 'cousins' les animaux que chez les plantes. Pourtant, les animaux comme les plantes ne représentent qu'une petite partie de la biodiversité des règnes vivants : les Bactéries, les Archées et les Eucaryotes. Si la diversité du règne eucaryote nous est apparue plus rapidement à l'œil de l'homme, celle des deux autres règnes, où se côtoient des organismes comme, par exemple, la Bactérie *Escherichia coli* (vivant dans le tube digestif humain) et l'Archée *Thermoplasma acidophilum* (vivant à 59°C, à pH 2 et respirant du souffre), est encore plus importante. Les premières formes des sciences naturelles, comme la célèbre *Histoire des animaux* (Aristote, VI^{ème} siècle avant J.C.), se sont attachées à décrire et analyser soigneusement ce foisonnement de formes de vie. Ces études systématiques, loin d'être achevées, définissent le niveau *organisme* de la Biologie.

La complexité apparente devient plus grande pour la compréhension d'un ensemble d'organismes, qu'ils soient de la même espèce ou d'espèces différentes. Et pourtant, il semble intéressant et nécessaire, pour comprendre le fonctionnement d'un organisme, de prendre en compte aussi celui de son environnement. Réciproquement, il paraît risqué de décrire le comportement d'un ensemble d'organismes sans rien connaître des fonctionnements intrinsèques de chacun d'entre eux. Ces niveaux plus intégrés de la *population* (voire de l'*espèce*) et de l'*écologie* n'ont été abordés que tardivement dans l'histoire de la Biologie, en s'appuyant sur des outils statistiques complexes. Ils constituent aujourd'hui les niveaux d'organisation les plus intégrés parmi les problématiques étudiées dans les sciences de la vie.

L'étude du comportement animal défini le niveau de l'*éthologie*, frontière avec les sciences dites humaines (type psychologie, etc.). Celle-ci peut être classée comme intermédiaire entre l'*organisme* et la *population*.

De même qu'il existe des niveaux d'organisation supérieurs à celui de l'*organisme*, il en existe des inférieurs. Chez les animaux et les plantes, les fonctions au sein d'un organisme

sont assurées par des organes, entités pluricellulaires qui œuvrent pour une fonction spécialisée, comme par exemple l'apport en énergie, la motricité, la reproduction, la défense de l'organisme, la sensation de l'environnement et la coordination entre ces unités. Ce niveau de l'*organe*, absent de la plupart des êtres vivants, est une clef de voûte de la compréhension des comportements animaux et végétaux. Son étude, la physiologie, fut amorcée par les grecs (par exemple Galien) et poursuivie par les Arabes (par exemple Avicenne). La compréhension des règles empiriques qui ont servi à bâtir la physiologie se fit à travers l'étude du niveau inférieur d'organisation, celui de la *cellule*.

Découverte au XIX^{éme} siècle par Schwann et Schleiten, la cellule apparue longtemps comme l'unité fondamentale des êtres vivants. Ils démontrèrent que tous les êtres vivants sont formés d'une ou plusieurs cellules. En observant la *cellule*, bien que d'importantes différences subsistent entre les divers êtres vivants, des structures communes transparaissent, laissant entrevoir les limites de la complexité du monde vivant. Par exemple, les cellules sont limitées par une membrane déterminant la matrice intérieure et le milieu extérieur, ce qui permet de limiter les variations de milieu interne. L'intérieur de la cellule eucaryote est divisé en compartiments souvent spécialisés : apport et transformation de l'énergie (mitochondries et chloroplastes), synthèse et mise en œuvre des "ouvriers" cellulaires (réticulum endoplasmique et appareil de Golgi), conservation et réplication de la "mémoire" cellulaire (noyau), etc. La communauté scientifique s'accorde à penser que certains des compartiments de la cellule eucaryote, tels que les mitochondries, les chloroplastes et le noyau, sont les reliquats d'endosymbioses abouties entre un pré-eucaryote et des bactéries ou/et archées. Cette communauté semble d'accord sur les dates d'apparition des divers organismes (Table 1) mais les causes et les mécanismes de ces transitions restent encore inexpliqués.

Date approximative (millions d'années)	***Avènement***
4 000	Terre
3 800	Bactéries ou/et Archées
1 800	Eucaryotes
700	Métazoaires
530	Vertébrés

Tableau 1 : Abrégé des dates marquant l'histoire de la vie.

Si la *cellule* a été longtemps posée comme unité élémentaire de la vie, l'étude de la *molécule* laisse entendre qu'il n'en est rien. A ce niveau d'organisation, les différences entre les organismes vivants restent présentes bien qu'elles soient encore amoindries. Les *molécules* forment actuellement la limite inférieure des investigations biologiques et peuvent être aujourd'hui assimilées aux briques élémentaires du vivant. Au sein de ces briques, cohabitent des molécules de petites tailles et des macromolécules formées par l'assemblage de petites entités moléculaires. Les macromolécules biologiques, qui peuvent être considérées comme intermédiaires entre les *molécules* et les *cellules*, sont de quatre types. Il peut être attribué à ces macromolécules des rôles principaux. (1) Les lipides sont dédiés à l'alimentation et à la formation des membranes. (2) Les glucides sont à la fois utilisés comme aliment et comme supports aux fonctions chimiques. (3) Les protéines constituent les principales molécules ouvrières de la cellule. (4) Les acides nucléiques sont chargés de véhiculer et d'exprimer la mémoire de la cellule.

L'ensemble des séquences d'ADN (macromolécules "mémoire") d'un organisme forme son génome. Chez les Eucaryotes, sont distingués les génomes mitochondriaux, chloroplastiques et nucléaire ; le mot génome seul faisant souvent référence au génome nucléaire. De plus, le génome est fractionné en plusieurs chromosomes chez les Eucaryotes et chez quelques Bactéries. Ceux-ci sont visibles, dans leur forme condensée, à l'observation microscopique au cours de la division cellulaire. Chez les organismes pluricellulaires, le génome est quasi identique dans toutes les cellules (exception faite des mutations somatiques

et des réarrangements dans les lymphocytes). Par ailleurs, il n'est transmis que par un petit nombre de cellules, les cellules germinales, qui sont mise en place précocement au cours du développement. Si le génome peut être vu comme une macromolécule codant pour des "fonctions" assurant sa propre réplication (Dawkins 1976), il peut également être assimilé à une "mémoire" que les "fonctions" utilisent pour se perpétuer (Morange 1998). Malgré les discordances que peuvent engendrer ces deux points de vue, pourtant non exclusifs, le génome est une entité macromoléculaire animée d'une dynamique intrinsèque qui le rend fluide et qui façonne sa structure.

Des contraintes et de l'évolution du vivant.

L'étude d'un niveau donné doit susciter trois questions clefs. *i*) Quelles sont les contraintes structurales qui lui sont imposées ? *ii*) Quelles sont les contraintes sélectives qu'il subit ? *iii*) Quelle est la liberté laissée par la réunion de ces deux types de contraintes ?

Les contraintes structurales sont celles imposées par un niveau d'organisation inférieur. Ainsi, par exemple, il faut, pour répliquer une macromolécule, assembler des molécules, et donc créer des liaisons chimiques *de novo*. La formation de ces liaisons covalentes requiert un coût énergétique élevé, qui est obtenu par cassure d'une autre liaison riche en énergie (le plus souvent ATP -> ADP + Pi). Conjointement, la réplication de cette macromolécule va imposer à son tour des contraintes à la cellule. En effet, la cellule doit, pour assembler des macromolécules couramment, mettre en place un système de synthèse d'énergie permettant ces assemblages. Si l'organisme est unicellulaire, il doit alors puiser cette énergie directement dans le milieu extérieur, mais s'il est pluricellulaire, la cellule fait appel à un organe de stockage qui lui fournira cette énergie. Cet organe influera à son tour sur l'organisme pour qu'il puise l'énergie dans le milieu externe. L'organisme se nourrit donc du milieu extérieur, soit d'éléments inertes, comme l'énergie solaire (pour les organismes autotrophes), soit de matière organique (pour les organismes hétérotrophes), soit d'autres êtres vivants (pour les organismes prédateurs hétérotrophes). Il est à noter que l'existence d'organismes prédateurs impose des contraintes sélectives sur les populations.

Les contraintes sélectives sont, à l'inverse des précédentes, celles imposées par les niveaux supérieurs. Elles s'établissent dans le temps et sont, dans une certaine mesure, assimilables à des forces de sélection naturelle. En effet, seul ce qui a perduré est aujourd'hui encore observable. Par exemple, la perpétuation par multiplication, qui semble avoir été adoptée par tous les êtres vivants, impose aux populations des contraintes issues (parfois indirectement) de la taille fixe de l'espace matériel (contrainte émanant d'un niveau non biologique). Par ailleurs, des contraintes sélectives peuvent naître de contraintes structurales. Pour reprendre l'exemple utilisé ci-dessus, la prédation a conduit à la sélection, au sein des populations, des organismes permettant la persistance de ces populations. Or, pour que ces organismes émergent, il est bien souvent nécessaire de modifier certains éléments des niveaux d'organisations inférieurs. Ainsi, l'*écologie* impose, de proche en proche, des contraintes sélectives sur la *molécule*.

De ce point de vue, l'évolution desdits niveaux de la Biologie résulterait d'un compromis entre les contraintes structurales des niveaux inférieurs et des contraintes sélectives des niveaux supérieurs. Cependant, il semble qu'à chacun d'entre eux, il existe une « marge de manœuvre » non soumise aux contraintes structurales et sélectives. Cette liberté d'évolution, a été mathématiquement décrite pour les *molécules* (Kimura 1983). Ce concept d'évolution neutraliste suggère que l'évolution de chaque niveau ne s'établit pas uniquement par la réunion des contraintes structurales et sélectives, mais qu'il existe une certaine liberté laissé au gré du hasard ; les variations aléatoires ne pouvant s'effectuer que dans l'espace défini par les deux types de contraintes. Ainsi, par exemple, il peut exister des variations dans les formes de certains organes qui n'affectent en rien leurs fonctions ou leur élaboration.

L'étude de la dynamique des génomes, objet de cette thèse, sera abordée ici en tentant de comprendre les contraintes structurales imposées par la chimie du vivant (à travers les mécanismes moléculaires qui la sous-tendent), de voir quelles contraintes sélectives sont imposées par les niveaux supérieurs (en intégrant les conséquences des changements dans les génomes) et enfin d'établir le champ de liberté de l'évolution des génomes (en analysant les mouvements observés dans les génomes).

De l'impossibilité d'observer directement la dynamique des génomes.

L'étude de l'évolution des génomes se ferait idéalement en observant (au microscope par exemple) les changements de ceux-ci pendant un temps très long. Or, l'observation des macromolécules est encore difficile (voire bien souvent impossible) et plus encore si le fonctionnement de l'organisme ne doit pas être perturbé. Néanmoins, il est plus aisé de rechercher et d'étudier les empreintes durables que laisse cette dynamique sur les structures des génomes.

Certaines de ces empreintes sont liées aux compositions en nucléotides qui peuvent être réparties inéquitablement le long des séquences. C'est, par exemple, le cas de la composition en nucléotides des génomes (Karlin *et al.* 1998), celui de l'organisation des chromosomes de vertébrés en isochores (zones relativement homogènes en nucléotides G+C) (Bernardi 2000) ou encore celui du biais de composition entre les deux brins de réplication (tardif et précoce) dans les Bactéries et les Archées (Rocha *et al.* 1999c).

D'autres empreintes sont liées à des structures particulières dans les chromosomes : la localisation des gènes et des unités fonctionnelles (comme les ARNr et les ARNt) dans les chromosomes, la répartition et les caractéristiques des séquences répétées (au sens large) ou la localisation chromosomique des centromères dans les génomes eucaryotes.

Enfin le dernier type d'empreinte, plus difficile à observer, est lié à des facteurs non détectés dans les séquences primaires. Ces facteurs sont soit réellement absents de la séquence primaire, soit encore non détectés aujourd'hui. C'est le cas, par exemple, de la présence et la localisation d'hétérochromatine[1], de la variation du taux de recombinaison le long des chromosomes ou de la présence de bases méthylées dans certaines régions des chromosomes.

La problématique que l'on se propose d'étudier, en abordant la dynamique des génomes par l'étude des séquences primaires, consiste à reconstituer une histoire (celle de l'évolution des génomes) à partir de ces quelques clichés. La séquence de plusieurs dizaines

[1] On distingue souvent deux type de chromatine : l'« euchromatine », peu condensée et exprimée et l'« hétérochromatine », très condensée et peu exprimée.

de génomes (pour le moment quelques Eucaryotes, une dizaine d'Archées et une cinquantaine de Bactéries) est entièrement déterminée et rendue publique. Si l'achèvement de ces séquences donne un nouvel essor à la compréhension des fonctions d'un organisme vivant et laisse entrevoir un espoir lointain d'application médicale, il apporte un grand avancement dans l'étude de la dynamique des génomes, donnant pour la première fois un paysage figé mais complet de quelques unes de ces macromolécules. Les empreintes que nous avons choisies de détecter et d'analyser dans l'ensemble de ces séquences primaires sont les répétitions intrachromosomiques.

I. Introduction

L'introduction de cette thèse est articulée autour de trois chapitres qui aborderont successivement les trois questions exposées dans l'avant propos. Le premier chapitre, intitulé Mécanismes moléculaires des répétitions, s'attache, à travers des exemples, à décrire les mécanismes de duplication et ceux amorcés par les répétitions. Le second chapitre, intitulé Conséquences des répétitions, tente de montrer quelles peuvent être les conséquences des séquences dupliquées (et en particulier des duplications de gènes) sur les organismes et les populations. Enfin le troisième chapitre, nommé Dynamique des répétitions, a pour ambition de replacer les répétitions dans un contexte évolutionniste et de montrer qu'elles doivent être considérées comme des marqueurs très informatifs de la dynamique des génomes.

A. Mécanismes moléculaires des répétitions.

A.1. Définitions.

Une *répétition* sera définie, dans ce manuscrit, comme une séquence d'ADN présente sous des formes similaires au moins deux fois dans un génome. Cette similarité est caractérisée par un pourcentage d'identité entre les *copies*. L'ADN n'étant constitué que de quatre bases différentes (Adénine, Cytosine, Guanine et Thymine), il faut envisager que de nombreuses répétitions apparaissent par de simples mutations ponctuelles. Ces répétitions seront appelées *répétitions fortuites*.

Il existe approximativement une probabilité de $1/4^n$ de trouver une copie exacte d'une séquence de taille n à une position donnée (si la fréquence de A, C, G et T est équiprobable et indépendante). Le nombre de copies moyen d'un mot de taille n, dans une séquence de taille L, est $L/4^n$. La probabilité de trouver k copies d'un mot de taille n dans une séquence de taille L est donnée par la loi de poisson P(k), dont la moyenne est $L/4^n$. La probabilité de trouver au moins deux copies d'un mot est donc 1-P(0)-P(1), soit $1-e^{-\lambda}(1-\lambda)$, avec $\lambda = L/4^n$. Le nombre de mots de taille n est 4^n. On a donc environ $4^n(1-e^{-\lambda}(1+\lambda))$ mots répétés de taille n.

Le plus petit génome séquencé, celui de *Mycoplasma genitalium*, étant composé d'environ 0,58 millions de paires de bases, aucune répétition exacte (totalement identique) de 20 paires de bases n'est attendue dans un génome de cette taille. Cependant, dans cette

estimation tous les nucléotides sont considérés comme ayant la même fréquence (probabilité de 1/4). Or, le génome de *M. genitalium* est très pauvre en nucléotides G+C (32%). Un calcul plus exact, qui tient compte de la fréquence de chacun des nucléotides (Karlin and Ost 1985), montre qu'il y a une chance sur mille de trouver une répétition de taille supérieure à 23 bases. Pourtant, devant l'abondance de répétitions exactes de taille supérieure à cette longueur, il est nécessaire de postuler l'existence de mécanismes qui créent des grandes répétitions non fortuites. Une *duplication* sera donc définie comme un évènement associé à un mécanisme qui crée une répétition (non fortuite). La diversité des types de répétitions rencontrées dans les séquences génomiques laisse supposer l'existence de plusieurs mécanismes de duplications. Nous poursuivrons donc cet exposé par un bref inventaire des différentes répétitions et des mécanismes qui les génèrent.

A.2. Les différentes duplications.

Le premier type de répétitions auquel nous nous attachons est celui des répétitions multicopie en tandem (répétitions satellites). Elles forment une super-famille de répétitions où peuvent se dessiner au moins trois classes. Ces dernières se distinguent par la taille de l'unité répétée en tandem plusieurs fois. Si cette classification peut paraître totalement arbitraire, il semble au contraire que, malgré un certain recouvrement, ces catégories ne sont pas créées par les mêmes mécanismes.

A.2.1. Les répétitions satellites.

A.2.1.1. Les microsatellites ou SSR (Simple Sequence Repeats).

Les plus simples répétitions rencontrées dans les génomes sont constitués de séquences de faible complexité, nommées microsatellites dans les génomes eucaryotes et SSR dans les génomes bactériens. Dans ce manuscrit, le terme SSR désignera ce type de répétitions. Elles sont constituées d'un motif très simple (souvent d'une longueur inférieure à six bases) et sont répétées en tandem des dizaines (voire des centaines) de fois. Ces répétitions sont extrêmement répandues dans les génomes animaux et végétaux. A titre

d'exemple, il existe une SSR tous les 2 kb (en moyenne) dans le génome humain (TIHGSC 2001). Une des principales caractéristiques de ces SSR est leur taux élevé de modifications, que l'on nommera instabilité dans la suite du manuscrit. Cette instabilité se manifeste principalement par la perte ou le gain d'unité(s) répétée(s) (expansion et contraction de la SSR). On définira la fréquence d'instabilité comme le nombre d'instabilités observées pour une SSR par génération. La fréquence d'instabilité des SSR est en moyenne de 10^{-4} chez *Homo sapiens* (Payseur and Nachman 2000), 6.10^{-6} chez *Drosophila melanogaster* (Schug *et al.* 1997) et 10^{-4} chez *Escherichia coli* (Levy and Cebula 2001). Ces valeurs ne sont données qu'à titre indicatif car elles peuvent varier considérablement en fonction de nombreux paramètres comme la taille et la composition de la SSR. Il a été montré, chez la levure *Saccharomyces cerevisiae*, que l'instabilité d'une SSR croît avec la taille totale de la répétition (Pupko and Graur 1999; Wierdl *et al.* 1997). De plus, la taille modifie également la nature de ces instabilités : dans les grandes SSR, les délétions sont plus fréquentes que dans les petites SSR (Wierdl *et al.* 1997). D'autre part, la composition de la SSR module également cette fréquence, des répétitions composées uniquement de G et C ayant une stabilité accrue *in vitro* (Schlotterer and Tautz 1992).

La densité des SSR n'est pas équivalente dans tous les génomes. De façon générale, elle est plus élevée dans le règne eucaryote (Richard *et al.* 1999). Cette observation pourrait être liée aux pressions sélectives tendant à conserver un génome compact chez les Bactéries et les Archées. Au sein des génomes eucaryotes, les densités en SSR de mono-, di-, tri- et tétranucléotides sont très variables d'une espèce à une autre. A l'exception du chromosome X de *D. melanogaster*, les chromosomes d'une même espèce présentent des densités en SSR similaires (Katti *et al.* 2001). De même, la composition de l'unité des SSR les plus fréquentes est spécifique à chaque génome mais n'est pas une simple conséquence du biais de composition des chromosomes. Dans les génomes bactériens, la densité est également très variable selon les espèces (Le Fleche *et al.* 2001). Cette variation n'est pas liée au caractère pathogène de l'organisme : les densités les plus élevées étant rapportées chez *Buchnera sp.*, endosymbionte non pathogène d'alphides (insectes) et chez *M. genitalium*, un agent pathogène de l'homme. (Le Fleche *et al.* 2001).

Les deux principaux mécanismes proposés généralement pour expliquer les expansions et contractions des SSR sont : (*i*) la recombinaison inégale entre chromosomes homologues pendant la méiose ou entre chromatides sœurs (figure 1a) et (*ii*) les erreurs au cours de la réplication (figure 1b).

Chez les bactéries, le modèle par dérapage de la polymérase pendant la réplication est plus souvent mis en avant (Levinson and Gutman 1987). Ce modèle suppose que lors de la polymérisation du néo-brin d'ADN, l'enzyme polymérase effectue des pauses. Durant ces pauses, le néo-brin peut se désapparier se rapparier de manière décalée, laissant une boucle non appariée. Selon le décalage produit, la fin de la réplication produit une addition ou une perte d'unité(s) de la SSR. Des études *in vitro* montrent que ce phénomène est observable pour la plupart des polymérases (Schlotterer and Tautz 1992). La boucle non appariée est la cible des enzymes de réparation des mésappariements. La mutation d'un gène codant pour une de ces enzymes entraîne une augmentation des fréquences d'instabilité des SSR, d'un facteur 20 pour *E. coli* (gènes MutS ou MutL) (Levy and Cebula 2001) (Levinson and Gutman 1987) et d'un facteur 100 chez *S. cerevisiae* (gène Msh2) (Wierdl *et al.* 1997). Chez *S. cerevisiae*, une étude plus fine montre que la fréquence des instabilités est sous la dépendance de deux voies de réparation des mésappariements :

- celle du complexe 1 – ou complexe -1 (formé par les gènes Msh2 et Msh6) pour les SSR dont la longueur de l'unité est petite (1 ou 2 pb),
- celle du complexe 2 – ou complexe -2 (formé par les gènes Msh3 et Msh6) pour celles dont la longueur de l'unité est plus grande (jusqu'à 8 bp) (Sia *et al.* 1997).

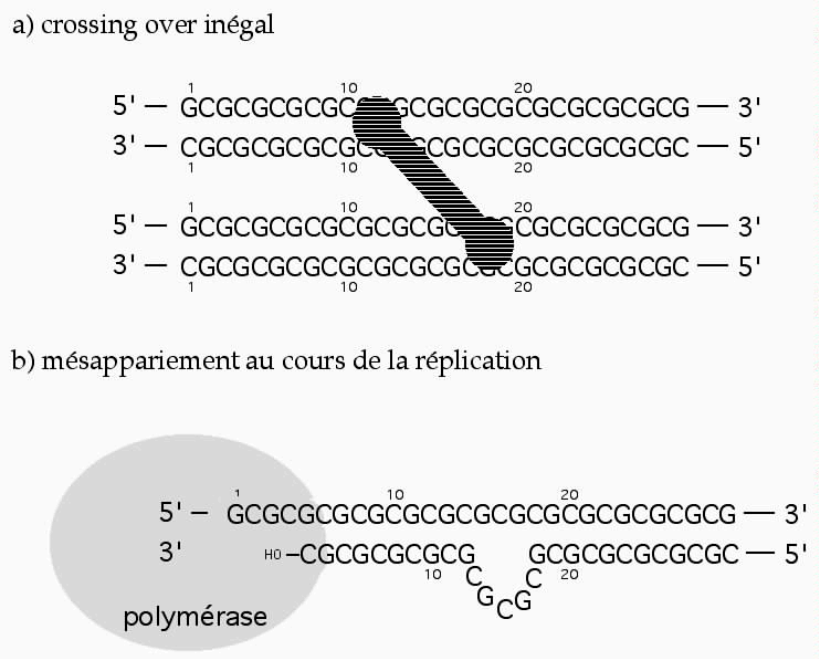

Figure 1 : Deux mécanismes proposés pour les instabilités des microsatellites.
Les instabilités les plus fréquentes sont les changements (perte ou gain) du nombre de copie de ces répétitions. a) Crossing-over inégal entre deux chromosomes homologues (ou entre deux chromatides sœurs). Il résulte de ce crossing-over une insertion sur un des chromosomes et une délétion sur l'autre. b) Mésappariement au cours de la réplication. Au cours de la réplication, s'il survient une pause, le brin néo-formé peut s'ouvrir et mal se réapparié. Dans le schéma, cela crée une insertion, mais si la boucle avait été sur l'autre brin, cela aurait créé une délétion.

Chez les eucaryotes, les accidents de réplication sont généralement considérés comme un mécanisme d'instabilité des SSR. Toutefois, certains auteurs suggèrent que les recombinaisons inégales peuvent également être un mécanisme important de ces instabilités. Dans cette hypothèse, ils ont vainement cherché des liens entre les localisations des SSR et les taux de recombinaison le long des chromosomes chez *H. sapiens* (Payseur and Nachman 2000) et chez *D. melanogaster* (Bachtrog *et al.* 1999). Chez *S. cerevisiae*, deux arguments indiquent que l'effet de la recombinaison sur la fréquence d'instabilité des SSR est faible, voire inexistant (Wierdl *et al.* 1997):

- l'absence d'une enzyme pilote de la recombinaison (codée par le gène RAD52) n'affecte pas la fréquence d'instabilité.
- La fréquence d'instabilité est identique à la méiose et à la mitose, alors que le taux de recombinaison est accrue au cours de la méiose.

Ainsi, l'instabilité des SSR paraît liée aux « accidents » survenant au cours de la réplication (figure 1.b), aussi bien dans les génomes eucaryotes que bactériens. Cependant, si la recombinaison inégale ne semble pas massivement impliquée dans l'instabilité des SSR, des mécanismes de conversion (échanges non réciproques) ont été récemment mis en avant (Richard and Paques 2000).

A.2.1.2. Les minisatellites ou VNTR (Variable Number of Tandem Repeats).

Les répétitions minisatellites sont définies comme des répétitions tête-à-queue dont la taille unitaire est d'une ou de quelques dizaines de nucléotides. Elles n'ont été étudiées que dans le règne eucaryote et en particulier dans les génomes des mammifères où elles sont les plus représentées. La recombinaison, qui ne semble pas impliquée dans les instabilités des SSR, paraît avoir un rôle prépondérant dans celles des minisatellites (Richard and Paques 2000; Vergnaud and Denoeud 2000). Deux points distinguent les mécanismes impliqués dans les instabilités des minisatellites de ceux impliqués dans les instabilités des SSR, établissant une frontière imprécise entre SSR et minisatellites :

- Leur instabilité est lié à la machinerie de recombinaison.
- Ils ne sont pas affectés par les dysfonctionnements du système de réparation des mésappariements (Sia *et al.* 1997).

Le premier locus contenant un minisatellite fut mis en évidence par Wyman et White qui recherchaient des loci, chez *H. sapiens*, présentant une variabilité suffisamment grande pour être utilisée en cartographie génétique (Wyman and White 1980). Ils se focalisèrent sur un fragment de chromosome qui présentait une importante variabilité de longueur dans la population. Ce n'est que quelques années plus tard que fut révélée la présence d'un minisatellite (nommé MS32) dans ce fragment (Jeffreys *et al.* 1985). Si la fréquence des instabilités des SSR peut paraître élevée (au maximum 0.001), celle des minisatellites l'est encore plus puisqu'elle peut atteindre 0.13 (pour CEB1 en lignée germinale mâle) (Vergnaud *et al.* 1991). Chez *S. cerevisiae*, un minisatellite inséré artificiellement dans le génome présente plus d'instabilité durant la méiose que durant la mitose (Debrauwere *et al.* 1999).

Vraisemblablement, ceci explique pourquoi, chez *H. sapiens*, les instabilités des minisatellites sont beaucoup plus fréquentes (de plusieurs ordres de grandeur) dans les lignées germinales que dans les lignées somatiques (Jeffreys *et al.* 1988; Jeffreys *et al.* 1994). Les instabilités somatiques ayant des caractéristiques très différentes de celles observées dans les lignées germinales, la plupart des auteurs proposent que les deux types d'instabilités dépendent de mécanismes différents (Jeffreys and Neumann 1997).

Les instabilités les plus étudiées sont celles observées dans les lignées germinales. Ces dernières sont modulées par la taille de la répétition et il existe une faible corrélation positive entre la taille d'un minisatellite et son instabilité (Buard *et al.* 1998). D'autres facteurs peuvent influencer leur instabilité. Pour MS32, la présence d'un allèle en *cis* du minisatelitte, nommé O1C (car associé à un polymorphisme G/C en 5'), (Monckton *et al.* 1994) abolit presque totalement son instabilité. Enfin, à l'instar des évènements de recombinaison, les instabilités présentent des différences qualitatives (localisation) (Jeffreys *et al.* 1988) et quantitatives (taux d'instabilité) (Vergnaud *et al.* 1991) entre les deux sexes.

Le taux d'instabilité des minisatellites varie d'un génome à l'autre. Par exemple, pour les mammifères, l'instabilité moyenne est moins élevée dans le génome de *Mus musculus* que dans celui de *H. sapiens* (Bois *et al.* 1998). La localisation des minisatellites varie également en fonction de l'espèce (Amarger *et al.* 1998). Chez *H. sapiens*, la plupart sont situées dans les régions subtélomèriques. Cette singularité est moins évidente dans le génome de *Sus scrofa* (le cochon) et n'est détectable ni dans le génome de *Ratus norvegicus* (Amarger *et al.* 1998) ni dans celui de *M. musculus* (Bois *et al.* 1998). Néanmoins, la mise en relation de l'histoire des remaniements chromosomiques avec la localisation actuelle des minisatellites suggère une origine subtélomèrique pour la plupart d'entre eux (Amarger *et al.* 1998).

Les minisatellites ne sont pas des répétitions strictes d'une même unité. Les multiples copies de l'unité de base constituant le minisatellite présentent des différences de séquence. Afin de mieux comprendre les mécanismes impliqués dans les instabilités méiotiques, une technique d'analyse fine des variants des minisatellites fut mise au point (Jeffreys *et al.* 1991; Jeffreys *et al.* 1990). Cette technique consiste en une PCR qui utilise des oligonucléotides

amorces spécifiques de chaque variant. Plusieurs amorces, dont la partie 5' diffère, sont utilisées pour détecter le même variant. Cela évite que seul le produit le plus court soit observé à la fin de la réaction. Grâce à elle, trois caractéristiques importantes ont été notamment décrites :

- les remaniements impliquent aussi bien des évènements intrachromosomiques qu'interchromosomiques (Jeffreys *et al.* 1991),
- la plupart des évènements de recombinaison sont complexes (mélangeant conversion et crossing-over) (Jeffreys *et al.* 1998; Jeffreys *et al.* 1994),
- les instabilités sont localisées préférentiellement sur un seul des bords du minisatellite (rarement au milieu) (Jeffreys *et al.* 1994).

Ces résultats ont conduit la plupart des auteurs à proposer que les minisatellites sont localisés à proximité de *hot-spot* de recombinaison (un *hot-spot*, ou point chaud, est une région où les coupures double-brins, initiatrices de la recombinaison méiotique, sont très fréquentes). Cette hypothèse fut étayée par l'étude d'un minisatellite inséré à plusieurs localisations choisies dans le chromosome III de *S.□erevisiae* (Debrauwere *et al.* 1999). En effet, Pour ce chromosome, une carte précise des fréquences des cassures double-brins a été établie (Baudat and Nicolas 1997). Cela permet de mettre en relation aisément ces fréquences avec celles d'instabilités du minisatellite. Par ailleurs, en absence de Spo11, endonucléase initiatrice de la coupure double-brins, les fréquences d'instabilités sont fortement réduites (Debrauwere *et al.* 1999). Enfin, les rayons gamma, générateurs de coupure double-brins, augmentent la fréquence des instabilités (Dubrova *et al.* 1993).

Les minisatellites sont donc des répétitions dont la grande instabilité durant la méiose est la conséquence de leur localisation à proximité d'un *hot-spot* de recombinaison. Une étude des régions adjacentes aux minisatellites CEB1, MS32 et MS31 n'a, néanmoins, pu mettre en lumière des signaux propres à ces régions (Murray *et al.* 1999).

A.2.1.3. Les répétitions centromériques ou satellites.

Les facteurs impliqués dans la fonction biologique du centromère restent encore mal compris (Pennisi 2001) . Une vaste controverse opposant une conception génétique à une conception épigénétique conduit à s'interroger sur le rôle fonctionnel des séquences associées aux constrictions primaires des centromères natifs : des répétitions satellites.

Rosenberg et ses collaborateurs ont montré que 7% du génome de *Cercopithecus aethiops* (le singe vert d'Afrique) est constitué d'une séquence répétée en tandem de 172 paires de bases localisées au niveau des centromères (Rosenberg *et al.* 1978). Les imperfections des réactions de séquence amenèrent les auteurs à conclure, à raison, que les copies n'étaient pas toujours strictement répétées (Rosenberg *et al.* 1978). Depuis cette découverte, plusieurs types de satellites, comme les satellites (Waye and Willard 1989) ou (Lee *et al.* 2000) ont été décrits dans les régions centromériques des chromosomes humains. Le satellite le plus courant (et le plus étudié) est l'homologue de celui découvert chez le singe vert, le satellite .

Chez *H. sapiens*, les satellites sont des répétitions (avec un nombre de copies variable) d'un monomère de 171 bp. A l'instar de ceux détectés chez *C. aethiops*, il existe plusieurs séquences divergentes des monomères des satellites (nommés variants), à partir desquelles fut établie une séquence consensus. Les monomères « voisins » se ressemblent peu (80% d'identité), mais il existe une périodicité où les monomères sont très identiques (98% d'identité) (figure 2). Cette périodicité définie des Répétitions d'Ordre Supérieur (ROS) (Willard and Waye 1987). Le nombre de ROS est polymorphe entre les satellites.

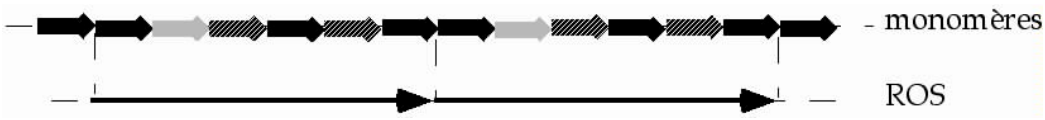

Figure 2 : Structure des répétitions satellites des centromères de primates.
Les petites flèches du haut représentent les monomères. Les monomères de même couleur sont identiques. Les grandes flèches sont les ROS (Répétitions d'Ordre Supérieure). Ces ROS définissent une périodicité d'identité dans les monomères. Dans le schéma, la périodicité est de six, il s'agit donc d'une ROS hexamérique.

L'étude précise du centromère du chromosome 17 de *H. sapiens* montre que si la ROS la plus courante dans ce chromosome est un héxadécamère (16 monomères) de 2,7 kb, il en

existe d'autres constituées de quinze monomères et de douze monomères (Waye and Willard 1986). L'analyse de la structure de ces trois ROS montre qu'elles dérivent toutes les unes des autres par des insertions et des délétions de monomères. Les structures des ROS sont, en général, spécifiques à chaque chromosome (Willard and Waye 1987), mais certaines sont présentes sur plusieurs centromères. Par exemple, la ROS endécamèrique (de 11 monomères) du centromère du chromosome 13 est également présente sur le chromosome 21. Cependant, l'analyse fine montre qu'il existe de petites variations de séquences entre les ROS des deux chromosomes. (Greig *et al.* 1993). Ceci suggère qu'après la duplication de cette ROS (d'un centromère vers l'autre), les deux centromères ont évolué indépendamment.

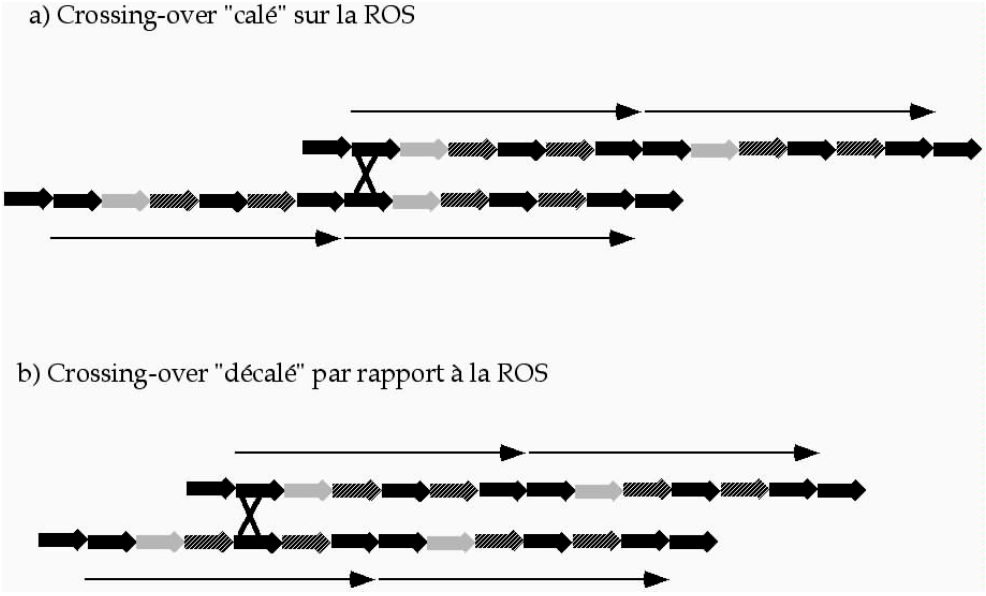

Figure 3 : Les deux types de crossing-over inégaux pouvant expliquer les variations dans les séquences satellites.
a) Le crossing-over est "calé" sur la ROS (Répétitions d'Ordre Supérieure, voir figure 2). Dans ce cas, seul le nombre de ROS est changé après le crossing-over. b) Le crossing-over est décalé par rapport à la ROS. Dans ce cas, le crossing-over crée une nouvelle ROS dans les deux séquences satellites.

Le mécanisme le plus souvent envisagé pour expliquer les polymorphismes (nombre de ROS) des satellites est le crossing-over inégal. Si le crossing-over est calé sur les ROS, alors il y a un changement du nombre de ROS (figure 3a), sinon il induit la formation d'une nouvelle ROS par insertion ou délétion de un ou plusieurs monomères (figure 3b). Ce mécanisme fut initialement proposé par une étude théorique (Smith 1976), qui prédit l'homogénéisation des séquences et l'apparition de ROS par ce type de mécanisme. L'observation précise des sites de réarrangements internes aux monomères a mis en évidence

une forte similarité de séquence à leur niveau (Warburton *et al.* 1993). Cette caractéristique est nécessaire pour un évènement de recombinaison.

Le centromère du chromosome X de *H. sapiens* est constitué principalement d'une ROS dodécamèrique (de 12 monomères) située au cœur du centromère(Willard and Waye 1987). La zone de transition d'environ 1Mb entre le centromère (constitué du dodécamère) et le bras p du chromosome X a été minutieusement étudié (Schueler *et al.* 2001). Elle possède également des satellites mais plus épars. Les satellites les plus proches des satellites centromériques sont organisés aussi en dodécamères et présentent un fort taux d'identité avec la ROS centrale (98%). Cette identité décroît au fur et à mesure que l'on s'éloigne du centromère (jusqu'à 70 % d'identité pour les ROS situées à 40 kb des ROS centrales). Cette conservation de structure associée à la perte d'identité aux extrémités des ROS centrales est également attendue dans l'hypothèse du crossing-over inégal.

Chez *H. sapiens*, la fréquence moyenne de recombinaison le long des chromosomes varie entre 1 et 2 cM/Mb. Alors que le taux de recombinaison des régions centromériques est quasi nul (TIHGSC 2001). Une étude fine des taux de recombinaison (pour le chromosome 5) montre que :

- dans la région de 5 Mb située entre le centromère et le bras q du chromosome 5, la fréquence de recombinaison est réduite à 0,64 cM/Mb.
- dans la région de 5,5 Mb située entre le centromère et le bras p du chromosome 5, la fréquence de recombinaison est inférieure à 0,3 cM/Mb.

Cette réduction de la recombinaison pourrait mettre en doute l'hypothèse selon laquelle les centromères évoluent majoritairement par crossing-over inégaux. Donc, si ces derniers sont réellement à l'origine de la variabilité centromérique, il faut admettre que la persistance des quelques rares évènements de recombinaison suffit à rendre compte du polymorphisme centromérique observé.

Si les satellites sont l'apanage des primates, les centromères de la plupart des espèces eucaryotes présentent des structures répétées en tandem en plusieurs copies

(Tableau 2). Seule *S. cerevisiae* ne présente des petits centromères d'environ 125 pb non répétés (Clarke 1990).

Espèces	Taille	Références
Homo sapiens	171 bp	(Willard and Waye 1987)
Cercopithecus aethiops	172 bp	(Rosenberg *et al.* 1978)
Mus musculus	120 pb et 234 pb	(Willard 1990)
Drosophila melanogaster	5 bp	(Sun *et al.* 1997)
Arabidopsis thaliana	180 pb	(Copenhaver *et al.* 1999)
Zea mays	180 pb	(Alfenito and Birchler 1993)
Beta sp.	160 pb	(Gindullis *et al.* 2001)
Schizosaccharomyces pombe	<1kb – 6.4 kb	(Clarke 1990)

Tableau 2 : Répétitions centromériques dans les différentes espèces.

Les répétitions centromériques sont la plupart du temps des répétitions en tandem (tête à queue), mais une étude récente s'est focalisée sur un autre type de séquences répétées localisées préférentiellement dans certains satellites : les éléments transposables (Mashkova *et al.* 2001). L'existence de sites préférentiels pour la localisation de ces éléments mobiles montre qu'ils obéissent principalement à des mécanismes de duplication différents de ceux créant les répétitions tête-à-queue (recombinaison inégale, conversion et accident de réplication).

A.2.2. Les éléments mobiles.

Les éléments mobiles forment une famille particulière de séquences répétées. Ils ne seront ici que très brièvement décrit mais un aperçu rapide de ces séquences montre qu'ils constituent des éléments produisant des duplications à longue distance. Ces évènements de duplications très élaborés, nommés transpositions, ne sont possibles que pour des séquences très particulières, les transposons.

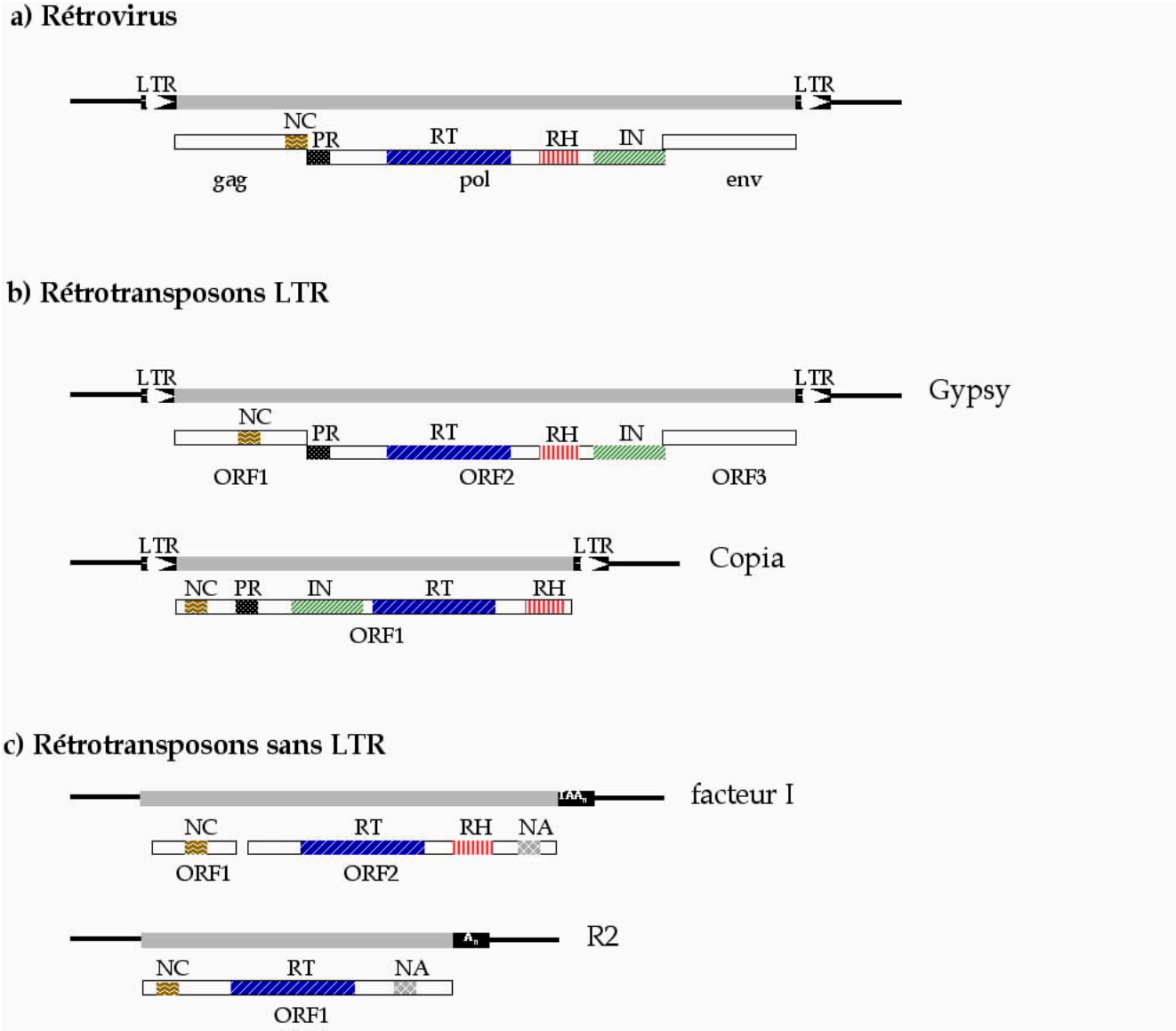

Figure 4 : Structure des principaux groupes de rétroïdes (d'après (Eickbush 1994)).
Dans tous les schémas, les extrémités sont en noires et les phases codantes (ORF pour Open Reading Frame) en gris. Le détail de ces ORF est donné en dessous. Chaque domaine identifié est indiqué par un sigle et une couleur. NC : Nucléocapside ; PR : Protéase ; RT : Reverse Transcriptase ; RH : RNAse H ; IN : Intégrase ; NA : Nucleic Acid binding site. a) Structure typique d'un rétrovirus. b) Deux exemples de rétrotransposons avec LTR (Long Terminal Repeat). c) Deux exemples de rétrotransposons sans LTR.

Il existe parfois des fuites permettant la duplication ponctuelle de séquences quelconques. Les éléments mobiles ont un taux de transposition élevé, qui représente jusqu'à 6 % des mutations dans le génome de *M. musculus*, où ils sont considérés comme très actifs (TIHGSC 2001). Les éléments transposables sont classiquement divisés en deux groupes qui se distinguent par leurs mécanismes de transposition : les rétroïdes et les transposons ADN (T_{ADN}).

A.2.2.1. Les rétroïdes.

On définit par rétroïde tout élément génétique utilisant une reverse transcriptase (RT), qui est une ADN-polymérase ARN-dépendante. Celle-ci a mis à bas un des dogmes de la biologie (le flux d'information va de l'ADN vers les protéines). Depuis sa découverte simultanée par deux équipes (Baltimore 1970; Temin and Mizutani 1970), de nombreux éléments furent rattachés aux rétroïdes. Les rétroïdes peuvent être divisés en plusieurs classes, mais tous utilisent le même mécanisme de duplication. Il est assimilable à celui d'un "copier-coller". Ce mécanisme s'effectue en trois étapes : (1) la transcription complète de l'élément, (2) la rétrotranscription de cet élément et (3) l'insertion de cet élément dans le génome. Il existe bien sûr de nombreuses variations qui définissent les groupes de rétroïdes (figure 4). La plus grande division du monde des rétroïdes est établie sur la présence (ou l'absence) de grandes séquences répétées aux extrémités de l'élément (Long Terminal Repeat – LTR).

A.2.2.1.1. Les rétroïdes à LTR.

Les éléments possédant des LTR peuvent être à nouveau subdivisés en trois grandes classes. Tous les rétroïdes à LTR comportent dans leur cycle de multiplication une phase cytoplasmique où ils sont visibles en microscopie sous forme de condensat (Eickbush 1994). D'autre part, tous ces éléments dupliquent leurs sites d'intégration sur quelques bases, le nombre étant variable selon les trois catégories.

La première catégorie constitue la célèbre famille des rétrovirus. Ils sont le plus souvent formés de 3 ORF principales encadrées par les deux LTR. Les ORF, nommées *gag*, *pol* et *env* codent respectivement pour les fonctions d'encapsidation, de réplication et d'enveloppe du virus. Les fonctions de *gag* et *pol* sont partagées par tous les rétroïdes à LTR mais le gène *env* est spécifique des rétrovirus, puisqu'il qui permet la propagation infectieuse du virus. Ces rétrovirus coexistent sous deux états. Le premier état est dit "exogène" et correspond aux phases infectieuses du virus. Il se transmet alors principalement horizontalement entre les organismes et utilise la RT à chaque génération. Deux exemples célèbres appartiennent à ce groupe : HIV et Influenzavirus (grippe). Le second état est dit

"endogène" et correspond à une phase au cours de laquelle le virus est intégré en une ou plusieurs copies dans le génome mais ne se réplique que par l'intermédiaire de réplication du chromosome. Il ne se transmet alors que verticalement entre les organismes. Comme exemple, on peut citer le HERV-K (Human Endogenous RetroVirus-K) qui est similaire au MMTV (Mouse Mammalian Tumor Virus), virus murin exogène. Cette alternance d'états endogènes et exogènes pose des problèmes pour la filiation de ces virus, puisqu'elle fait varier les vitesses d'évolution de ces éléments entre celle du chromosome (lente) et celle induite par une RT très peu fidèle (très rapide) (Doolittle *et al.* 1989; Doolittle *et al.* 1990). Seuls les rétrovirus endogènes peuvent être assimilés à des séquences dupliquées. Il existe, par ailleurs, des pararétrovirus, comme les Hepadnavirus (virus à hépatite) (Seeger *et al.* 1986) et les Caulimovirus (virus du chou-fleur) (Pfeiffer and Hohn 1983), qui ne s'intègrent jamais dans le génome.

La seconde catégorie est constituée par des rétrotransposons à LTR qui sont globalement organisés comme les rétrovirus, mais qui ne sont pas infectieux (pas d'équivalent du gène *env*). Ils possèdent deux ou trois ORF qui codent pour les fonctions de *gag* et *pol* : une nucléocapside qui enveloppe le rétroïde dans le cytoplasme, une protéase chargée de cliver les chaînes multimèriques en plusieurs protéines, une intégrase qui assure l'intégration de l'élément dans le génome, la RT et une RNAse H, dégradant le complexe ARN/ADN produit par la rétrotranscription. Trois membres célèbres de cette catégorie sont l'élément *Gypsy* de *D. melanogaster* (Marlor *et al.* 1986), l'élément *17.6* de *D. melanogaster* (Saigo *et al.* 1984) et l'élément *Ty3* de *S. cerevisiae* (Hansen *et al.* 1988).

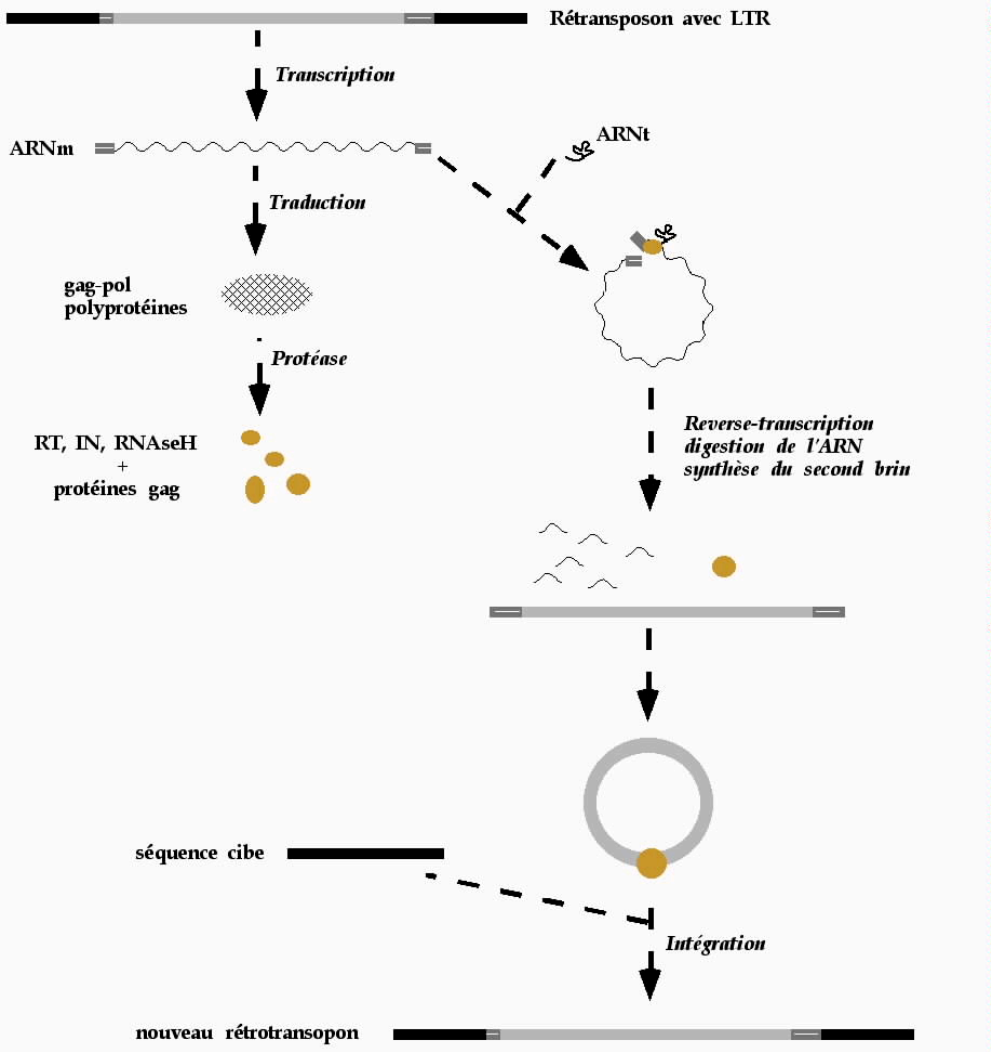

Figure 5 : Processus de transposition des rétrotransposons avec LTR (d'après (Eickbush 1994)).
Au cours de ce processus, le rétroïde est transcrit et traduit en une polyprotéine, qui est clivée en plusieurs protéines, dont la Reverse transcripase (RT), l'intégrase (IN) et la RNAseH. Ces dernières vont se charger respectivement de la rétrotranscription de l'élément, de la dégradation de l'ARN dans le complexe ARN-ADN et de l'intégration de l'élément sur le chromosome.

Enfin, la troisième catégorie comporte les rétroïdes à LTR les plus élémentaires. Ils sont constitués d'une ou deux ORF encadrées par les LTR. S'il y a deux ORF, il est avancé qu'elles sont traduites en une seule et même polyprotéine par un décalage du cadre de lecture au cours de la traduction. La polyprotéine est maturée en plusieurs protéines qui possèdent les fonctions susdécrites du groupe Gipsy-like. Des représentants célèbres sont l'élément *Copia* de *D. melanogaster* (Mount and Rubin 1985), *Ty1* et *Ty2* de *S. cerevisiae* (Clare and Farabaugh 1985) et *Ta1* de *A. thaliana* (Voytas and Ausubel 1988).

L'organisation commune des ORF de ces éléments, ainsi que la phylogénie de toutes les RT, suggèrent une ascendance commune à tous ces rétroïdes (Eickbush 1994) (McClure 1993). Les rétroïdes à LTR ne sont pas répartis également dans tous les génomes eucaryotes puisqu'ils abondent en formes et en nombres dans le génome de *H. sapiens* (100 familles couvrant 8,1 % du génome), sont plus modestes dans celui de *D. melanogaster* et quasi absents de celui de *C. elegans* et de *A. thaliana* (TIHGSC 2001). Les mécanismes de transposition de ces trois catégories d'éléments sont globalement similaires (figure 5). La transcription de l'élément est initiée au milieu du LTR situé en amont et se termine au milieu du LTR situé en aval. Les protéines sont traduites et maturées en unités fonctionnelles. Puis la reverse transcription procède jusqu'à la synthèse d'un ADN double-brins, initié par un ARNt (comme amorce). Celui-ci est alors intégré dans le génome recréant un autre élément complet à une distance plus ou moins grande de son ancêtre.

A.2.2.1.2. Les rétroïdes sans LTR.

Un autre groupe de rétroïdes est celui dont les éléments sont dépourvus de LTR. Cette famille paraphylétique est d'une grande diversité, bien plus importante que celle des rétroïdes à LTR et est composée d'éléments souvent moins élaborés. Cette famille possède des éléments qui ne peuvent être assimilés à des répétitions comme les plasmides mitochondriaux de Mauriceville (chez *Neurospora crassa*) (Nargang *et al.* 1984) et les rétrons de *Escherichia coli* et de *Myxococcus xanthus* (Lampson *et al.* 1991; Temin 1989). Une autre classe est constituée par les répétitions terminales des télomères synthétisés, pour la plupart, par une RT, la télomèrase (Eickbush 1997; Singer 1995). Par la suite, ne seront approfondies les caractéristiques que de trois types d'éléments : les « *LINE-like* », les « *SINE-like* » et les rétropseudogènes.

Les « *LINE-like* » (Long Interspersed Elements) sont des éléments que l'on peut considérer comme « complets » puisqu'ils codent pour toutes les fonctions nécessaires à leur transposition. A titre d'exemple, 868 000 LINE sont dénombrés dans le génome humain représentant 20,4% du génome. Tous les *LINE-like* présentent une queue poly-A en 3' et une petite duplication du site d'intégration. Les mécanismes de transposition de ces éléments ne

sont pas encore très clairs. Néanmoins, le modèle le plus mis en avant est présenté sur la figure 6 (Eickbush 1994).

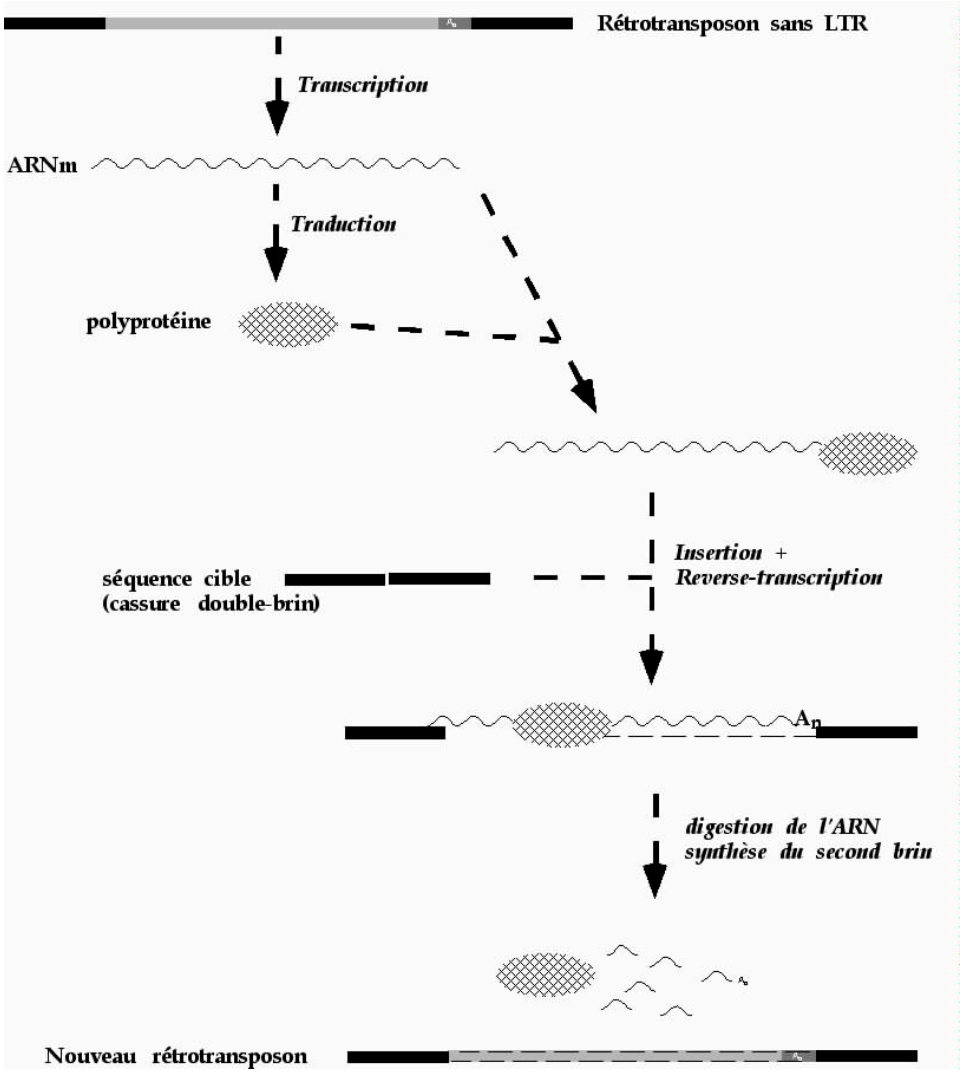

Figure 6 : Hypothétique processus de transposition des rétrotransposons sans LTR (d'après (Eickbush 1994)).
Au cours de ce processus, le rétroïde est transcrit et traduit en une polyprotéine. Cette dernière reconnaît, son messager et permet l'insertion de celui-ci au niveau d'une cassure double-brins et sa rétrotranscription.

Dans ce modèle, la transcription de ces éléments est assurée par la RNA-polymérase II (Singer and Skowronski 1985) dont un promoteur est contenu dans la partie 5' des éléments. La reverse transcription est initiée par l'insertion de l'élément. Ce dernier est lié en 3' au site d'insertion, l'ADN chromosomique servant ainsi d'amorce pour la réaction. Les mécanismes de réparation de la cellule achèvent la transposition de l'élément. La rétrotranscription de ces éléments génère de nombreux éléments incomplets tronqués en 5' par des rétrotranscriptions avortées. Il est très difficile d'établir des familles de ces éléments

car ils présentent des formes variées. Par exemple, l'élément *R2* de *D. melanogaster* possède une ORF et se localise souvent dans l'ADN ribosomique (ARNr 28S) (Jakubczak *et al.* 1990). A l'inverse, les éléments *I* de *D. melanogaster* (Fawcett *et al.* 1986) ou *L1* de *H. sapiens* (Singer 1982) ne possèdent qu'une seule ORF et semblent répartis dans tous les chromosomes.

Les « *SINE-like* » (Short Interspersed Elements) sont des éléments « incomplets », dans le sens ou ils ne codent pour aucune protéine. Ils utilisent les enzymes synthétisées par la classe des *LINE-like* pour leurs propres transpositions. A l'instar des LINE, ils présentent une queue poly-A en 3' et une petite duplication du site d'intégration. Ils possèdent un promoteur interne qui recrute l'ARN polymérase III (TIHGSC 2001). L'exemple le plus connu de SINE est l'élément *Alu*, qui a envahi le génome de *H. sapiens*. Il existe 1,5 million de copies de ces éléments qui couvrent 13,1% du génome (TIHGSC 2001). Les séquences *Alu* existent dans les génomes de la plupart des mammifères. Comme ceux de *H. sapiens*, la plupart de ceux des autres primates sont constitués d'une répétition de deux monomères de 130 bp alors que les éléments *B1* de *M. musculus* ne sont constitués que d'un monomère (Singer 1982).

Les rétropseudogènes sont des ARN messagers qui ont été pris comme substrat par une RT. Ils sont exempts d'intron, possèdent une queue poly-A en 3' et leur intégration provoque une petite duplication (quelques bases) de leur site d'insertion. La plupart de ces rétropseudogènes ne codent plus pour une protéine, mais l'on rapporte quelques cas de rétropseudogènes fonctionnels dans le génome de *H. sapiens* (TIHGSC 2001). Il a été montré expérimentalement qu'ils sont générés par la RT des *LINE-like* (Maestre *et al.* 1995). Les gènes dont ils dérivent sont courts, fortement exprimés et très conservés entre les espèces (Goncalves *et al.* 2000). Ils peuvent donc être considérés, au même titre que les SINE, comme des produits secondaires de l'activité des *LINE-like*. Néanmoins, à la différence des SINE, ils ne transposent plus quand ils sont intégrés. Ils constituent des répétitions plus "génériques" que les éléments transposables.

Parmi les caractéristiques notables de ces rétroïdes sans LTR, on peut noter leur répartition singulière dans les domaines GC. En effet, ils ne sont pas répartis aléatoirement

dans le génome, les SINE étant plutôt localisés dans les régions riches en GC, les LINE et les rétropseudogènes plutôt dans les régions plus pauvres en GC (Goncalves *et al.* 2000; Mazzarella and Schlessinger 1997). Ce biais de localisation peut avoir deux explications : soit ces éléments ont des sites préférentiels d'intégration (TIHGSC 2001), soit les éléments sont intégrés aléatoirement et sont préférentiellement perdus dans un type de domaine GC (Pavlicek *et al.* 2001). Les deux explications ont été proposées, mais le problème reste entier puisqu'il semble que les *Alu* récents soit préférentiellement situés dans les domaines plus pauvres en GC (TIHGSC 2001).

Bien que les éléments mobiles les plus répandus dans le génome humain soient des rétroïdes (plus de 3 000 000 (Li *et al.* 2001)), il existe un certain nombre d'éléments mobiles (plus de 300 000 (Li *et al.* 2001) répartis en 60 familles couvrant 2-3% du génome (Li *et al.* 2001)) qui ne transposent pas par un intermédiaire ARN.

A.2.2.2. Les transposons à ADN (T_{ADN}).

Les T_{ADN}, sont présents dans les trois règnes du vivant. Le mécanisme qui procède à la duplication des T_{ADN} est comparable à celui d'un "couper-coller". Bien que certains d'entre eux présentent des structures complexes, la plupart possèdent une structure très simple et une taille variant entre 800 pb et 3 kb (voir quelques exemples sur la figure 7).

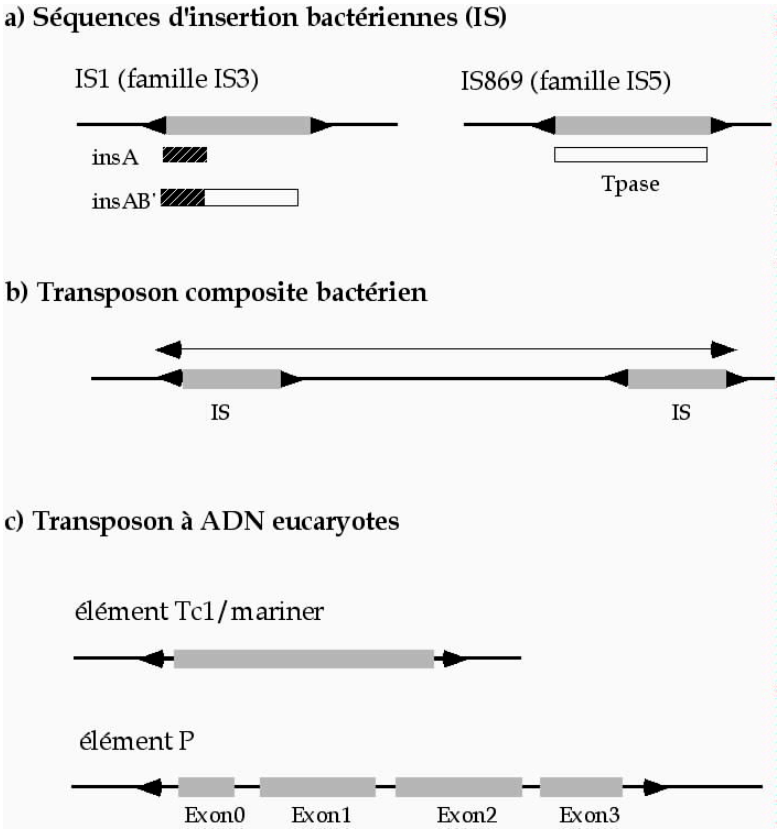

Figure 7 : Exemples de transposons à ADN bactériens et eucaryotes.
a) transposons bactériens, contenant un gène codant pour une transposase, encadrée par deux répétitions inversées (les "pieds"). Ces deux répétitions inversées sont les sites de reconnaissance des transposases. b) Le transposon composite est composé de deux IS qui sont mobilisées en même temps, aboutissant ainsi à la transposition de la séquence entre les deux IS. c) Deux exemples de transposons à ADN eucaryotes. la famille de Tc1/mariner *est dispersée dans tous les eucaryotes.*

Il existe des T_{ADN} complexes dans les génomes bactériens, comme les transposons composites (une séquence encadrée par deux IS -*Insertion Sequence*- simples), des transposons type Tn7, les phages ou les intégrons (pour revue (Merlin and Toussaint 1999)), mais ils ne seront pas détaillés ici. Les T_{ADN} bactériens les plus simples sont les IS. Elles possèdent une ou deux ORF codant pour une transposase ; un décalage du cadre de lecture étant souvent nécessaire pour former la transposase complète. Ce gène est encadré par deux répétitions inversées terminales de taille variable (10 à 500 bp). Les IS provoquent par leur intégration une duplication du site cible. La taille de cette répétition directe est variable d'une famille d'IS à une autre (2 à 20 pb). Il est possible de classer tous ces éléments en familles sur la base (1) des similarités de séquence et d'organisation entre les transposases, (2) de la similarité de leurs répétitions inversées terminales et (3) de la taille des duplications de leur site d'intégration.

Chez les eucaryotes les T_{ADN} se regroupent en quelques familles dont la plus grande est *Tc1/mariner*, présente dans tous les phylums (Plasterk *et al.* 1999). L'élément *Tc1*, découvert en 1983, chez *C. elegans* (Emmons *et al.* 1983), fut ensuite rapproché de l'élément *mariner* de *Drosophila mauritiana* (Jacobson *et al.* 1986), puis d'autres familles comme les éléments pogo de *D. melanogaster* (Tudor *et al.* 1992). Une seconde famille bien connue des T_{ADN} eucaryotes est celle qui englobe l'élément P (Bingham *et al.* 1982). Cet élément a envahi les génomes de *D. melanogaster* il y a quelques décennies (Engels 1996). Il a été, par la suite, décrit chez de nombreuses espèces de drosophiles, et plus récemment dans d'autres espèces, y compris *H. sapiens* (Hagemann and Pinsker 2001). Il existe bien sûr d'autres familles, mais elles semblent moins variées que les nombreuses IS bactériennes.

Chez les bactéries, 500 IS distinctes ont été récemment classées en 17 familles (Mahillon and Chandler 1998)). Les familles les plus nombreuses sont celle des IS3 (50 membres) et celle des IS5 (47 membres); les familles les moins nombreuses présentent seulement quelques membres. Ces familles présentent des particularités de structure mais également de répartition dans les populations. Par exemple, plusieurs auteurs (Hall *et al.* 1989; Sawyer *et al.* 1987) ont étudié la répartition des IS dans plus de soixante dix isolats naturels de *E. coli*. Ils ont montré que cette répartition est complètement hétérogène pour toutes les IS, tant au niveau du chromosome qu'à celui des plasmides. Il est d'ailleurs proposé que les plasmides conjugatifs soient les vecteurs principaux des transferts horizontaux des IS.

Les mécanismes de transposition sont relativement similaires pour les différents T_{ADN} eucaryotes et bactériens. Néanmoins, certains, comme la famille IS91, présentent une transposition très différente (semblable à la réplication du phage ϕX-174) (Mendiola and de la Cruz 1992). La plupart des transposases ont des similarités marquées notamment par la présence d'un trio d'acides aminés, DDE, à des positions précises dans la séquence primaire; cette triade formant le site catalytique de l'enzyme. Curieusement, on peut noter la présence de ce motif dans les intégrases du rétrovirus HIV, du phage Mu ainsi que dans RuvC, enzyme résolvant les structures de Holliday. La structure tridimensionnelle de l'intégrase du

phage Mu (Rice and Mizuuchi 1995) confirme les similarités entre ces protéines, suggérant que toutes ces intégrations procèdent via un mécanisme commun (figure 8).

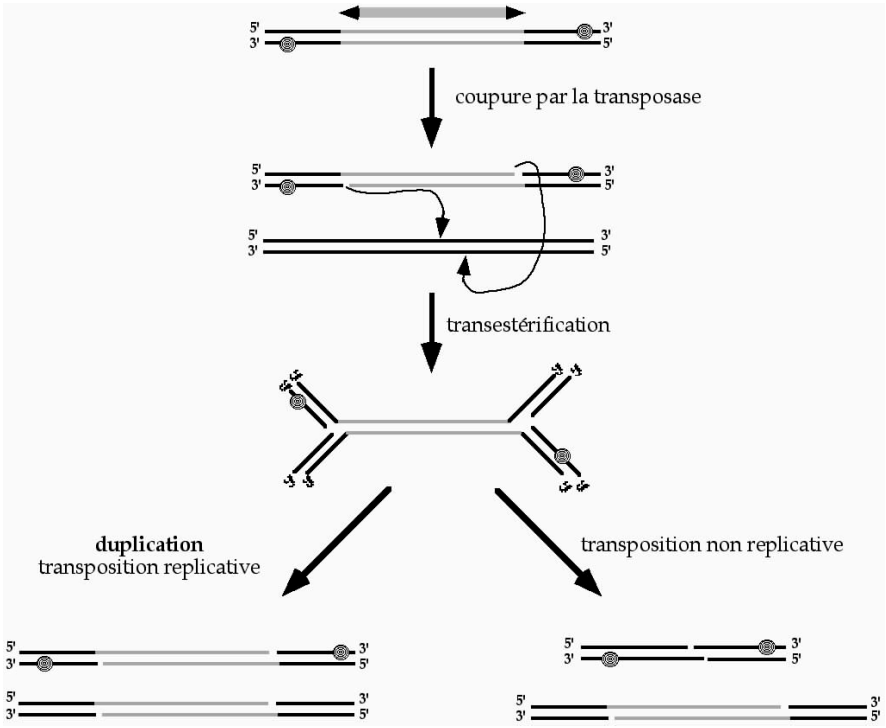

Figure 8 : Schéma du mécanisme de transposition des T_{ADN} (d'après (Mahillon and Chandler 1998)).
La transposition est initiée par la coupure simple brin des deux pieds du transposons. Puis, l'extrémité OH libre, issue de la coupure, "attaque" le site d'insertion et se lie avec elle par transestérification. La résolution de cet intermédiaire peut mener à la duplication du transposon ou à sa simple transposition.

Ce dernier est initié par l'hydrolyse simple brin des extrémités de l'élément. L'extrémité libre 3'OH se lie à un site d'intégration aléatoire ou spécifique (selon les familles). Mizuuchi propose que le transfert de brin se fasse en une seule étape de transesterification (Mizuuchi 1992). La transposition laisse une cassure double-brins au site d'excision qui est réparée par les mécanismes de l'hôte. Ainsi, ce mécanisme mène à la transposition non réplicative des T_{ADN}.

Cependant, il existe dans les génomes bactériens et les génomes eucaryotes des mécanismes pouvant expliquer la duplication de ces éléments. Dans les bactéries, si la réplication est initiée avant la résolution, alors il y a formation d'un cointégrat, qui après résolution mène à la duplication de l'élément. Chez les eucaryotes, la coupure double-brins peut être réparée par copie de la chromatide sœur (85% des cas pour l'élément P (Engels 1996)) ou du chromosome homologue ce qui conduit à la duplication du transposon.

Cependant, si la réparation par copie est avortée, seules les extrémités du T_{ADN} sont copiées et raboutées. Les MITEs (Miniature Inverted repeat Transposable Elements), très répandus dans les génomes eucaryotes (Surzycki and Belknap 1999) (Surzycki and Belknap 2000), présentent ce type de structure. Il a d'ailleurs été montré chez *A. thaliana* que certains MITE sont des T_{ADN} tronqués mobilisables en *trans* par des transposases d'éléments complets (Feschotte and Mouches 2000).

La régulation de la transposition des T_{ADN} est très complexe. Néanmoins, l'un de ces mécanismes nécessite la synthèse d'une transposase incomplète (dont la partie Cter est tronquée). En effet, la transposase est divisée en deux domaines distincts, chargés d'une part de la reconnaissance du site (partie Nter) et d'autre part de l'activité d'endonucléase (partie Cter). Ainsi, une transposase ne possédant que la partie Nter se fixe sur son ligand, empêchant d'autres transposases de se fixer, mais n'effectue pas la coupure. Dans les IS bactériennes, le début et la fin des gènes de transposases ne codent pas dans la même phase ; ainsi, la synthèse d'une transposase complète nécessite souvent un décalage de cadre de lecture. Si ce dernier n'a pas lieu, une protéine tronquée est produite (Mahillon and Chandler 1998). Pour l'élément P eucaryote, c'est la non-épissure de l'intron 2-3 qui est responsable de la synthèse de cette protéine tronquée (Engels 1996).

Les facteurs de l'hôte régulent souvent la transposition des T_{ADN}, mais n'interviennent pas souvent directement dans le mécanisme d'excision ou d'insertion. Par exemple, les éléments de la famille *Tc1-mariner* semblent autosuffisants pour leur activité puisqu'ils sont présents dans un très large spectre d'hôte. Une protéine synthétisée chez *E. coli* suffit à les mobiliser *in vitro* (Lampe *et al.* 1996). Par ailleurs, une équipe a construit un élément artificiel de cette famille (nommé Sleeping Beauty) en récupérant des fragments dans plusieurs génomes de poissons (Ivics *et al.* 1997). Cet élément est fonctionnel dans les cellules de mammifères (Horie *et al.* 2001).

Les éléments mobiles, aussi bien les rétroïdes que les T_{ADN}, sont donc des structures spécialisées dans des duplications à longue distance. Toutefois le « prix » pour ces duplications efficaces est une structure complexe ; bien que cette dernière puisse recouvrir

plusieurs formes très différentes, les duplications d'éléments mobiles ne sont pas « génériques ». Cependant, la présence de ces éléments conduit parfois à des accidents de transposition pouvant mener à la duplication de séquences plus génériques. C'est par exemple le cas des rétropseudogènes et des transposons bactériens composites. Par ailleurs, la présence de nombreux éléments mobiles entraîne des réarrangements qui peuvent conduire à créer des répétitions de grande taille, les répétitions géantes.

A.2.3. Les répétitions géantes.

Il existe, notamment dans les génomes eucaryotes, un certain nombre de répétitions de très grande taille. Ces répétitions peuvent couvrir de quelques dizaines à quelques centaines de kilobases (duplications segmentaires) , voire des chromosomes entiers (hyperploïdie) ou des génomes entiers (polyploïdie). Le premier type de répétitions géantes que nous nous attacherons à décrire est celui des répétitions segmentaires.

A.2.3.1. Les répétitions segmentaires.

Une analyse récente des génomes de *D. melanogaster*, *C. elegans* et *S. cerevisiae* montre que les deux derniers possèdent dans leur génome de grandes régions dupliquées (Friedman and Hughes 2001a). Dans le génome de *C. elegans*, les auteurs décrivent cinq répétitions géantes intrachromosomiques (dont une comprenant 21 gènes sur le chromosome 5) et dans celui de *S. cerevisiae*, ils en décrivent trente-neuf (Friedman and Hughes 2001a).

La description de la première répétition géante dans le génome de levure de boulanger comprend les centromères des chromosomes III et XIV (Lalo *et al.* 1993). Elle est décrite comme une région synténique entre ces deux chromosomes. La synténie n'est cependant pas parfaite puisque s'insèrent dans ces régions des gènes uniques. Deux autres répétitions géantes imparfaites furent décrites au moment du séquençage complet du génome de la levure entre les chromosomes II et VII ainsi qu'entre les chromosomes I et VII (Feuermann *et al.* 1997). L'avènement de la séquence complète de la levure fit ressortir dans plusieurs études concomitantes l'existence d'une cinquantaine de répétitions géantes partagées entre les différents chromosomes de *S. cerevisiae* (Coissac *et al.* 1997; Mewes *et al.*

1997; Wolfe and Shields 1997) ; aucune de ces répétitions n'est intrachromosomique. Les bornes de ces répétitions ont été légèrement précisées dans une étude postérieure (Seoighe and Wolfe 1999). Ces répétitions présentent plusieurs caractéristiques remarquables :

- l'orientation des deux copies est la même par rapport aux télomères (Coissac *et al.* 1997),
- la distance au télomère de chacune des copies de ces répétitions est similaire (Coissac *et al.* 1997),
- elles ne sont jamais en plus de deux copies (Wolfe and Shields 1997),
- bien qu'imparfaites (figure 9), elles couvrent quasiment 50% des chromosomes (Coissac *et al.* 1997; Wolfe and Shields 1997).

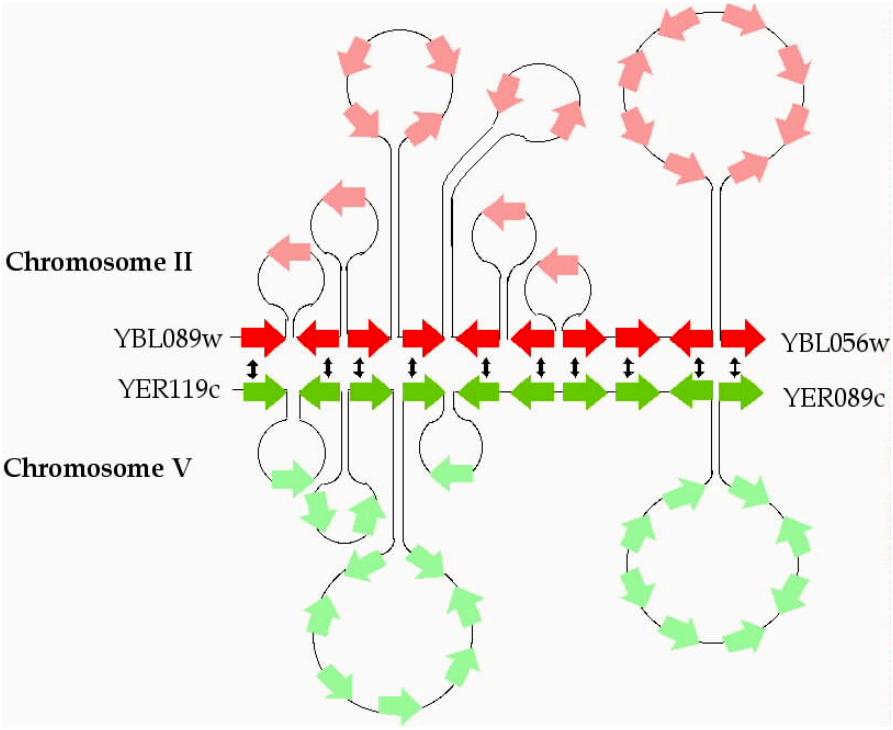

Figure 9 : Répétition segmentaire chez *S. cerevisiae* partagée par le chromosome II et le V (d'après (Coissac *et al.* 1997)).
Chaque flèche représente une ORF (Open Reading Frame). Si l'ORF possède une copie sur l'autre chromosome, elle est représentée sur la "tige" centrale, sinon elle représentée sur les boucles externes. Cette répétition segmentaire est donc imparfaite puisqu'elle possède des ORF sans paralogue. On peut supposer que ces paralogues ont été perdus après la duplication et que des ORF se sont insérées après la duplication.

Ce type de répétitions géantes a également été décrit chez *A. thaliana* (Terryn *et al.* 1999). Comme chez *S. cerevisiae*, une analyse plus exhaustive des séquences révèle de nombreuses répétitions géantes dans ce génome (Blanc *et al.* 2000; TAGI 2000; Vision *et al.* 2000). Ces répétitions présentent des caractéristiques différentes de celles du génome de la

levure . D'une part, plusieurs d'entre elles sont intrachromosomiques ; d'autre part, il semble exister des répétitions géantes de trois copies (Ku *et al.* 2000). Les orientations et localisations par rapport aux télomères de ces répétitions ne semblent pas avoir été analysées.

Dans les génomes de vertébrés, on trouve également des répétitions géantes comparables à celles présentes dans les génomes de *S. cerevisiae* et *A. thaliana*. Trois de ces répétitions géantes ont été plus particulièrement caractérisées (Skrabanek and Wolfe 1998): celle comprenant les gènes du CMH (Complexe Majeur d'Histocompatibilité) (Kasahara *et al.* 1996), celle comprenant les gènes FGFR (Fibroblast Growth Factor Receptor) (Pebusque *et al.* 1998) et celle comprenant les gènes Hox (Bailey *et al.* 1997). De façon remarquable, ces trois répétitions géantes présentent 4 copies chez les vertébrés tétrapodes. A cette règle, il semble que les poissons téléostéens constituent une exception, puisqu'il existe sept copies de la région contenant les gènes Hox dans le génome de *Danio rerio* (Zebra fish ou poisson zèbre) (Amores *et al.* 1998).

Les mécanismes qui ont forgé les répétitions géantes de *S. cerevisiae*, *A. thaliana*, et celles des vertébrés restent encore mal déterminés. Pourtant, ils sont au cœur d'un vaste débat opposant une partie de la communauté qui soutient que certaines répétitions géantes sont la trace d'évènement de polyploïdisation contre une autre partie soutenant qu'elles sont apparues par des duplications successives. Les détails de ce débat seront abordés dans le troisième chapitre. Toutefois, il est nécessaire de mentionner ici quelques répétitions géantes dont il est certain qu'elles ne sont pas la relique de polyploïdies.

Chez la levure, les tailles des chromosomes de plusieurs souches utilisées dans la vinification varient beaucoup au point que l'on trouve un chromosome dont la taille est accrue de 450 kb (Bidenne *et al.* 1992). De telles duplications sont également observées dans les souches de laboratoire. En effet, la souche FL100trp possède une translocation réciproque entre 80 kb du chromosome III et 45 kb du chromosome I. Si, dans cette souche, il n'y a ni gain ni perte d'ADN, cette translocation explique une répétition géante d'une partie du chromosome III dans la souche FL100 (Casaregola *et al.* 1998). Par ailleurs, il a pu être mis en évidence des duplications « expérimentales ». En recherchant des mutants de réversion pour

plusieurs mutations dans le gène URA2, Roealants et ses collaborateurs ont découvert des duplications de plusieurs dizaines de kilobases (voire de centaines) (Roelants *et al.* 1995). L'étude particulière d'une de ces duplications a conduit à décrire une répétition géante interchromosomique localisée à l'intérieur des deux chromosomes (Bach *et al.* 1995).

Dans les génomes des grands singes (et des hominidés), plusieurs répétitions de grandes tailles ont été décrites entre des régions péricentromériques (impliquant parfois des régions subtélomèriques). A titre d'exemple, dans le génome de *H. sapiens*, on observe une grande répétition (250 kilobases) entre les deux péricentromères du chromosome 10 (au-delà des satellites) (Jackson *et al.* 1999). Les auteurs observent que cette région a été soumise à de nombreux réarrangements dans les génomes des hominoïdes (inversion, délétion, duplication). Un autre exemple est celui d'une région de 30 kb dupliquée très récemment (95% d'identité) entre le subtélomère du chromosome X et le péricentromère du chromosome 16 de *H. sapiens* (Eichler *et al.* 1996). L'étude des génomes d'autres hominoïdes et de *M. musculus* permet de montrer que l'original est situé sur le chromosome X. D'autre part, il semble que la répétition soit bordée de copies plus ou moins exactes du motif CAGGG. L'étude des bordures des répétitions géantes péricentromériques et subtélomèriques montre que ces motifs CAGGG semblent présents presque tout les cas (Eichler *et al.* 1999). Les auteurs proposent que ces motifs soient impliqués dans la genèse de ces répétitions géantes. La publication de la séquence complète du génome humain a permis de montrer que ces répétitions géantes sont nombreuses. Les répétitions de longueur supérieure à 10kb et d'identité supérieure à 98% représentent 2,5% de la séquence (TIHGSC 2001). Une étude des répétitions d'une taille supérieure à 1kb montre que ces dernières sont principalement intrachromosomiques et localisées dans les régions péricentromériques et subtélomèriques (Bailey *et al.* 2001).

Il existe des répétitions encore plus grandes que les répétitions segmentaires ; il s'agit des répétitions de chromosomes, à savoir les hyperploïdies (chromosomes surnuméraires) et les polyploïdies (duplication de génomes).

A.2.3.2. Hyperploïdie et polyploïdie.

Comment aborder l'hyperploïdie sans mentionner l'une des anomalies génétiques de *H. sapiens* les plus connues : la trisomie 21 ? De la fréquence d'apparition de trisomies 21 viable chez *H. sapiens* (~ 1/700 naissances), on doit estimer que la fréquence totale de ces accidents de ségrégation méiotique des chromosomes est élevée, comme le montre la figure 10 (Suzuki *et al.* 1989). Les évènements d'hyperploïdie ont de telles conséquences sur les organismes eucaryotes complexes qu'elles ne sont pas conservées au cours des générations.

Chez *S. cerevisiae*, on observe parfois des levures trisomiques. Néanmoins toutes les souches utilisées dans la vinification sont diploïdes (Mortimer *et al.* 1994) bien qu'elle n'ait pas une forte parenté (Johnston *et al.* 2000; Mortimer 2000). Les quelques rares levures trisomiques présentent des défauts de sporulation (Johnston *et al.* 2000). Curieusement, la délétion de certains gènes entraîne une augmentation de la fréquence de trisomie (le chromosome surnuméraire pouvant varier) (Hughes *et al.* 2000).

Dans certains cas, une mauvaise ségrégation peut conduire à une duplication complète d'un ou de deux lots de chromosomes menant respectivement à des tri et des tétraploïdes. Ces phénomènes de polyploïdisation sont bien connus chez les plantes (Suzuki *et al.* 1989). Il existe deux types de polyploïdisation : l'allopolyploïdisation, qui résulte de la fusion de deux génomes différents et l'autopolyploïdisation qui consiste en la duplication interne de tous les chromosomes.

Un des exemples documentés de polyploïdie chez les plantes est celui de *Zea mays*. La comparaison des cartes génétiques de *Sorghum bicolor* et de *Z. mays* montre que la plupart des régions du génome du Sorgho existe en double exemplaire dans le génome du maïs (Whitkus *et al.* 1992). Il semble aujourd'hui que *Z. mays* est issu d'une polyploïdisation par hybridation entre deux souches ou deux espèces, bien que les deux espèces ayant fondé cet hybride n'existent plus aujourd'hui (Gaut *et al.* 2000).

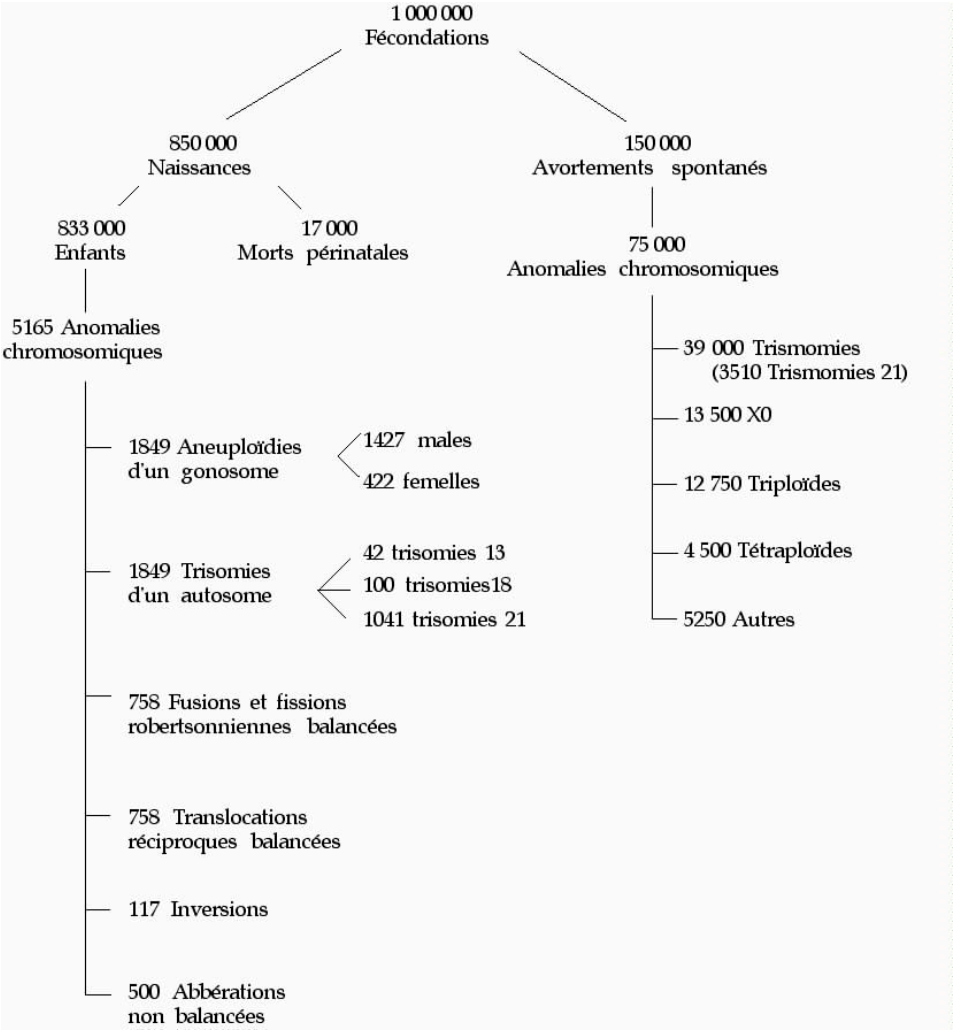

Figure 10 : Devenir d'un million de zygotes humains avec une attention particulière aux anomalies chromosomiques majeures (d'après (Suzuki *et al.* 1989)).

Chez les vertébrés, un des exemples de polyploïdie le plus spectaculaire est celui de la famille des xénopes. L'observation des caryotypes de diverses espèces de xénopes fait ressortir que si la plupart d'entre elles présentent un caryotype de 36 chromosomes, comme *Xenopus laevis laevis*, d'autres en ont 72, comme *Xenopus vestitus* et *Xenopus sp. n.*, voire 108 dans le cas de *Xenopus ruwenzoriensis*. La mesure de la quantité d'ADN contenue dans les noyaux de chacune de ces espèces confirme qu'il s'agit bien de tétraploïdes et d'héxaploïdes (Thiebaud and Fischberg 1977).

A.2.4. Les répétitions génériques.

Les répétitions décrites ci-dessus ont toutes des particularités qui permettent de les regrouper en classes : elles sont répétées en tandem et sont multi-copie (satellites), elles ont une structure permettant leur transposition (éléments mobiles) ou elles sont de très grandes tailles (les répétitions géantes). Néanmoins, la plupart des répétitions n'appartiennent pas à ces catégories et seront nommées les répétitions "génériques". Elles peuvent être classées selon qu'elles existent en (1) deux-copie ou multi-copie, (2) qu'elles présentent des localisations interchromosomiques ou intrachromosomiques et (3) que, dans le cas des intrachromosomiques, elles ont des copies proches ou distantes et (4) directes ou inversées. Seuls quelques mécanismes sont documentés pour leur genèse.

A.2.4.1. les répétitions intrachromosomiques directes proches.

Le premier type de répétitions génériques est constitué par les répétitions intrachromosomiques directes proches. Les descriptions de ce type de répétitions abondent dans tous les génomes (eucaryotes et bactériens). Une structure probable pour la genèse de la plupart de ces répétitions directes proches est l'amplicon. Celui-ci, bien caractérisé chez les bactéries (Romero and Palacios 1997), est composé d'un corps central encadré par deux répétitions directes, nommées pieds (voir figure 11).

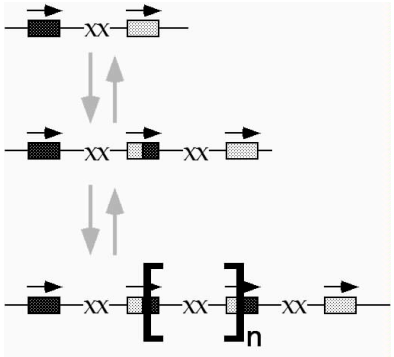

Figure 11 : Amplification d'un amplicon.
Un amplicon est un séquence quelconque (XX), bordée de deux répétitions directes. Les évènements de recombinaison entre les deux répétitions peuvent conduire à la perte de la séquence (délétion) ou à la duplication réversible de cette séquence.

Un évènement de recombinaison entre les pieds crée potentiellement une duplication ou une délétion de l'amplicon. Trois principaux mécanismes non exclusifs ont été avancés pour expliquer les duplications d'amplicons.

Le premier mécanisme est un crossing-over inégal entre un pied situé sur une chromatide (ou un chromosome) et l'autre pied situé sur l'autre chromatide (ou chromosome) au cours de la réplication (voir figure 12.a) (Anderson and Roth 1981). Dans ce mécanisme, l'augmentation du nombre de copies ne peut pas être supérieur au nombre initial de copies. Il suppose donc de multiples étapes pour obtenir un nombre élevé de copies à partir d'un amplicon. De plus, cet échange réciproque entraîne, pour chaque duplication, une délétion. L'observation de plasmides dimériques portant une duplication et une délétion suggère que les crossing-over inégaux sont à l'origine de certaines duplications d'amplicons (Dianov *et al.* 1991).

Le second mécanisme proposé est l'excision de l'amplicon par recombinaison entre les pieds et sa réinsertion sur un autre chromosome par une seconde recombinaison (voir figure 12.b). Ainsi, dans ce mécanisme, chaque duplication est également associée à une délétion. La transposition de l'amplicon peut s'effectuer entre deux chromosomes (ou chromatides au cours de la réplication), mais peut également s'effectuer entre deux loci distincts du même chromosome. La détection d'intermédiaires circulaires libres suggère que ce second mécanisme peut également expliquer certains évènements de duplication (Flores *et al.* 1993).

Le troisième mécanisme est la réplication circulaire de l'amplicon (Young and Cullum 1987) (figure 12.c). Il permet, à l'inverse des deux précédents, de créer un amplicon multi-copie en une seule étape. Comme le précédent mécanisme, celui-ci nécessite deux évènements de recombinaisons entre les pieds de l'amplicon. Le résultat d'un tel mécanisme est une amplification rapide de l'amplicon non associée à une délétion. L'observation de structures complexes prédites par le modèle suggère que ce mécanisme peut aussi participer aux duplications d'amplicons (Petit *et al.* 1992).

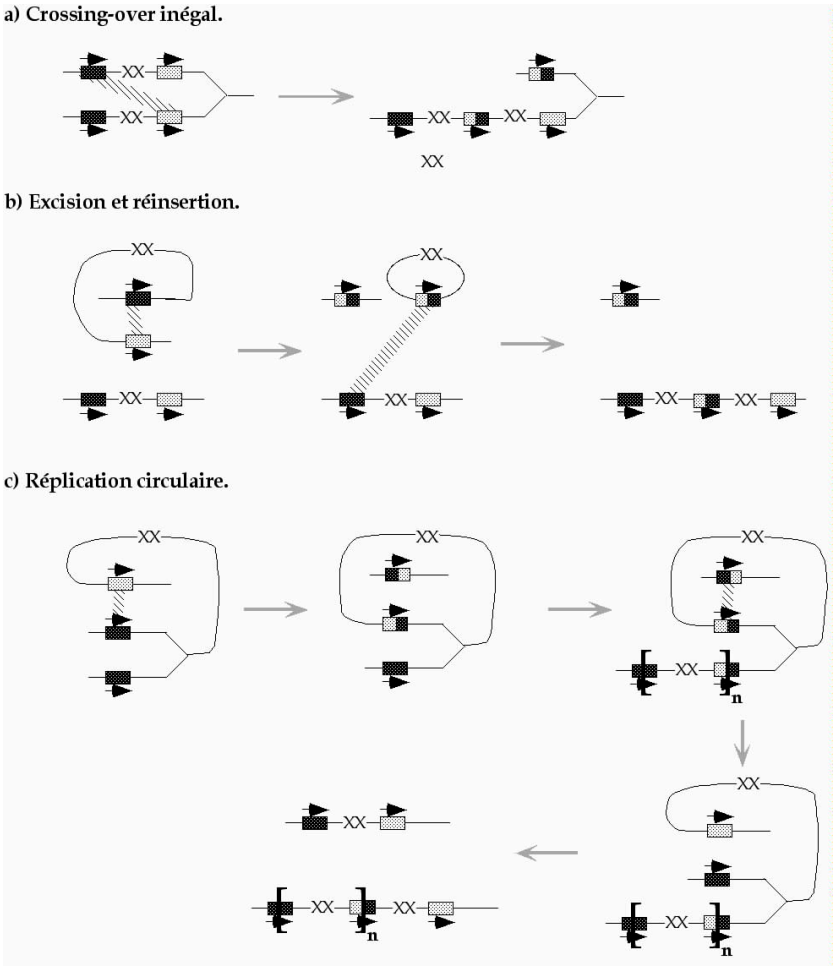

Figure 12 : Mécanismes pour l'amplification d'un amplicon (d'après (Romero *et al.* 1999)).
a) crossing-over inégal entre deux chromosomes en cours de réplication. b) excision et réinsertion dans le génome. La réinsertion peut avoir lieu à un locus différent de l'excision. c) amplification par réplication circulaire. Ce mécanisme permet de créer en une seule étape de nombreuses copies de l'amplicon.

Pour conclure, ces trois mécanismes semblent coexister dans les génomes. Ils requièrent tous au moins un évènement de recombinaison entre les deux pieds de l'amplicon. Aussi, la recombinaison homologue semble indispensable pour ces évènements de duplication d'amplicon et son implication semble confirmée par la nécessité des gènes de la famille RecA dans ce processus. Ce dernier code pour une protéine pilote de la recombinaison homologue. La modification de l'activité de cette protéine change la fréquence de duplication d'amplicons : chez *Vibrio cholerae*, une délétion du gène l'abaisse d'un facteur 20 (Goldberg and Mekalanos 1986) alors que, chez *E. coli*, son expression constitutive entraîne l'augmentation d'un facteur 10 (Dimpfl and Echols 1989).

Dans le génome de *H. sapiens*, un amplicon est particulièrement bien étudié. Celui-ci, présent sur le bras court du chromosome 17 contient en son corps le gène PMP22 (Peripheral

Myelin Protein 22), dont la perte ou la duplication entraîne de graves conséquences pathologiques. Chez un individu sain, l'amplicon n'est présent qu'à une copie ; sa duplication provoquant la maladie de Charcot-Marie-Tooth de type 1 (CMT1) et sa délétion une "Hereditary Neuropathy with liability to Pressure Palsies" (HNPP). La duplication de l'amplicon est le produit d'une recombinaison entre les pieds de l'amplicon (Lupski *et al.* 1991). Une analyse fine des recombinants montre que les caractéristiques de la recombinaison sont dépendantes du sexe. La recombinaison s'effectue préférentiellement entre chromosomes homologues au cours de la spermatogenèse et préférentiellement entre chromatides sœurs au cours de l'ovogenèse (Lopes *et al.* 1997). Par ailleurs, il a été proposé que ces duplications et ces délétions sont fréquentes à cause de la localisation proche d'un *hot-spot* de recombinaison (Lopes *et al.* 1996; Lopes *et al.* 1999).

Dans les répétitions génériques, la plupart des évènements de duplications intrachromosomiques observés forment des répétitions directes proches. Néanmoins, au moins un mécanisme de duplication a été proposé pour expliquer des répétitions intrachromosomiques directes mais plus distantes (espacées d'une dizaine de kilobases).

A.2.4.2. Les répétitions intrachromosomiques directes distantes.

Il a été montré que l'insertion d'un fragment d'ADN circulaire dans le chromosome peut être un mécanisme de duplication (figure 13). Ce mécanisme est celui évoqué pour l'intégration d'un fragment portant une résistance à la tétracycline chez *E. coli* (Lin *et al.* 1984) et plus généralement pour toutes les transformations intégratives, en particulier chez *Bacillus subtilis* (Dubnau 1993). L'intégration d'un fragment exogène dépend de plusieurs étapes : état de compétence de l'organisme, conservation de l'ADN en milieu externe, absorption du fragment dans la cellule, etc. Peu d'organismes présentent une transformation intégrative naturelle efficace (Lorenz and Wackernagel 1994). L'intégration du fragment circulaire nécessite une recombinaison entre deux séquences similaires, une contenue dans le fragment, l'autre dans le chromosome.

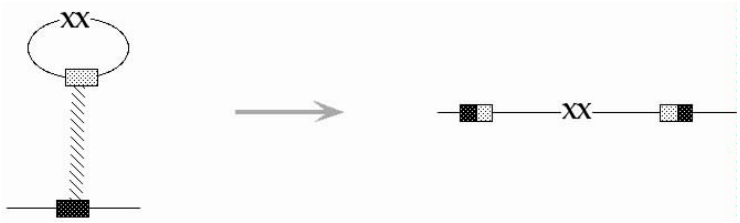

Figure 13 : Insertion d'un ADN circulaire.
Les deux régions homologues (marquées par des rectangles) recombinent et créent une répétition directe dont les deux copies sont espacées par la séquence insérée non homologue.

Chez *B. subtilis*, l'efficacité de cette recombinaison dépend de la similarité entre les deux séquences : par exemple, une divergence de 14,2% abaisse la fréquence d'un facteur 100 (par comparaison avec une divergence nulle) (Zawadzki and Cohan 1995). Bien que des petits fragments (547 bp) soient intégrés (Zawadzki and Cohan 1995), la taille moyenne des intégrations est d'environ 8-10 kb (Lorenz and Wackernagel 1994). Ce mécanisme a été proposé pour expliquer une vingtaine de répétitions directes espacées d'une dizaine de kb chez la bactérie *B. subtilis* (Rocha *et al.* 1999a). Cette proposition est étayée par l'analyse de la séquence située entre les deux copies. Celle-ci ne présente pas la même composition nucléotidique que le reste du chromosome. Ce type d'intégration crée une répétition directe espacée de la taille du fragment. Il est intéressant de noter que cette structure correspond à celle d'un amplicon. Il paraît donc vraisemblable que les organismes naturellement transformables créent des répétitions de cette façon, mais ceci ne semble pas envisageable pour tous les organismes non transformables.

Les mécanismes précédents ne permettent que la genèse de répétitions intrachromosomiques, mais il existe, outre les translocations créant surtout des répétitions géantes, au moins un mécanisme pouvant expliquer la genèse de répétitions dispersées : l'insertion.

A.2.4.3. Les répétitions dispersées.

Dans ce paragraphe, seules les insertions de séquences quelconques (dans des chromosomes parfois différents) seront considérées, celles des éléments mobiles ayant déjà été exposées ci-dessus. Plusieurs mécanismes peuvent être envisagés pour expliquer ces

insertions, mais les plus documentés sont en relation avec la réparation des cassures doubles brins. En effet, dans plusieurs chromosomes eucaryotes, les réparations de cassures doubles brins semblent être associées à des insertions de séquences quelconques. Pour étudier ces réparations dans plusieurs génomes, des coupures double-brins artificielles ont été induites par des endonucléases (souvent l'enzyme *HO* ou *I-sceI*). Il faut noter que, chez *S. cerevisiae*, ces expériences sont réalisées dans des souches haploïdes, car, dans les souches diploïdes, la plupart des cassure doubles brins sont réparées par copie du chromosome homologue.

On peut distinguer au moins trois types de situations dans lesquelles les réparations de ces coupures double-brins ont été testées (figure 14) : (1) les deux extrémités libres de part et d'autre de la coupure double-brins présentent des similarités avec une autre séquence, (2) seule une des deux extrémités est similaire à une autre séquence et (3) aucune des extrémités ne présente de similarité étendue avec une autre séquence. Il est important de soulever que dans le premier cas, des mécanismes intervenant dans les deux autres cas peuvent également intervenir. En matière de recombinaison, "qui peut le plus peu le moins".

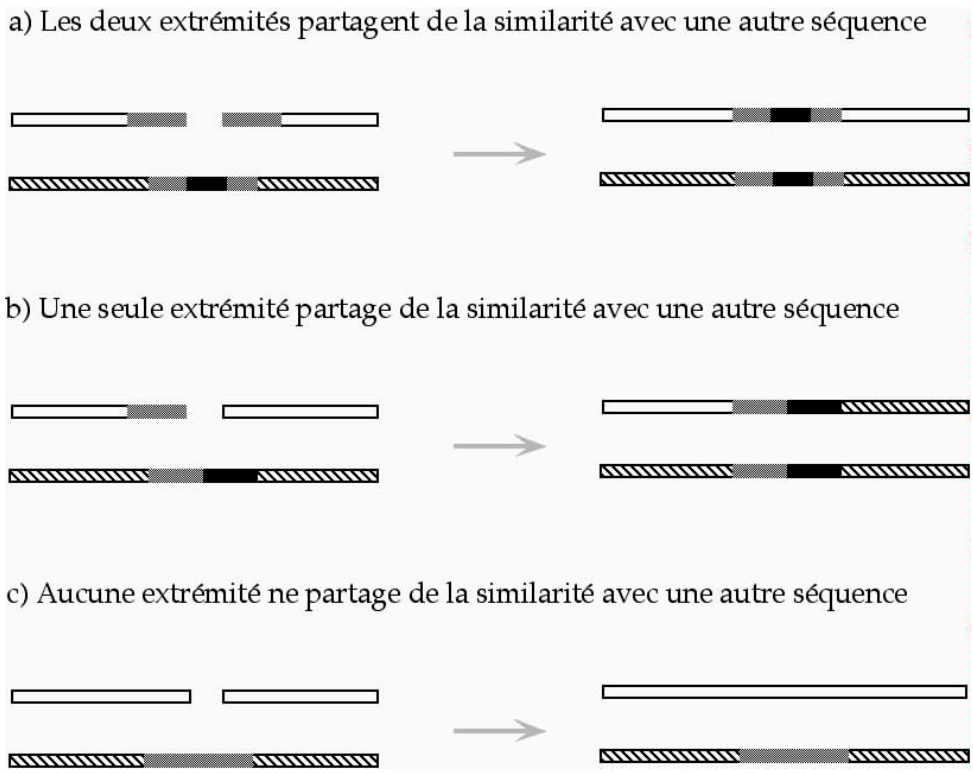

Figure 14 : Duplication par réparation des cassures double-brins.
a) Si les deux extrémités laissées par la cassure sont similaires à une autre séquence, la réparation peut être faite par copie de la séquence similaire (Gap Repair). b) Si une seule des deux extrémités est similaire à une autre

séquence, la réparation peut être faite par copie d'une grande partie de la séquence similaire (Break Induced Replication, BIR). c) Si aucune des deux extrémités n'est similaire à une autre séquence, la réparation peut-être faite par jointure des extrémités (Non Homologous End Joining, NHEJ). Il faut noter que dans le cas (a), la réparation peut également être faite par BIR et dans les cas (a) et (b) par NHEJ.

(1) la coupure laisse deux extrémités présentant de la similarité avec une autre séquence dans le chromosome (figure 14.a). Chez *S. cerevisiae*, où la recombinaison homologue est très efficace, on observe également des évènements de « réparation par copie » (forme de conversion nommée aussi gap repair). Une étude récente montre que cette conversion s'effectue mieux si la séquence similaire est portée par un autre chromosome que si elle est portée par un plasmide (Paques *et al.* 1998). Cette situation produit des répétitions « quelconques », seules les extrémités libres issues de la coupure double-brins devant être similaires à une autre séquence. Le mécanisme avancé pour expliquer ce type d'évènement est le SDSA (Simple-Dependent Strand Annealing, pour revue voir (Paques and Haber 1999)).

(2) seule une des deux extrémités présente de la similarité avec une autre séquence (figure 14.b). Chez *S. cerevisiae*, cette extrémité envahit la séquence similaire et initie une réplication d'un fragment de chromosome. Cette réplication peut s'étendre sur une très grande distance (Voelkel-Meiman and Roeder 1990), voire à un bras de chromosome entier (Bosco and Haber 1998). Ce mécanisme est aujourd'hui documenté sous le nom de BIR (Break Induced Replication) (Kraus *et al.* 2001). Ce type d'évènement peut être également observé dans des cellules ES (Embryonic Stem, de *M. musculus*) (Richardson and Jasin 2000). Cependant, seules de petites conversions (inférieures à 1kb) ont été observées dans ces cellules. En outre, la partie ne présentant pas de similarité avec le chromosome est raboutée à une séquence présentant de la microhomologie avec elle. Cette dernière est supposée, dans les cellules de mammifères, aider à rabouter des extrémités double-brins libres (voir ci-dessous).

(3) aucune des extrémités libres ne présente de similarité étendue avec une autre séquence du génome (figure 14.c). Dans ce cas, ces extrémités sont simplement raboutées l'une avec l'autre par un mécanisme et documenté sous le nom de Non Homologous End Joining (NHEJ) (pour revue (Critchlow and Jackson 1998)). Ce dernier est réalisé par un

grand complexe protéique (dont Ku70 et Ku80). Le NHEJ est facilité par la présence de microhomologie (quelques bases) entre les deux brins. Ces jointures de doubles brins peuvent être associées à des délétions ou à des insertions de séquence. Ainsi, chez *S. cerevisiae*, la jointure d'une coupure double-brins d'un chromosome est accompagnée de délétions (dans 1% des levures survivantes) ou d'insertions (2%) (Ricchetti *et al.* 1999) ; d'autres auteurs observant des fréquences d'insertion avoisinantes (0,4% (Yu and Gabriel 1999) ou de 1% (Moore and Haber 1996)). Les insertions semblent exclusivement de deux types :

- des insertions d'ADN complémentaires d'élément mobiles (des rétroïdes Ty1) (Moore and Haber 1996) (Yu and Gabriel 1999) ou
- des insertions d'ADN mitochondrial (Ricchetti *et al.* 1999; Yu and Gabriel 1999). Curieusement, aucune insertion d'ADN nucléaire n'a été observée chez *S. cerevisiae*.

Le même type d'expérience mené dans des cellules CHO (Chinese Hamster Ovary) (Sargent *et al.* 1997) ou de *Nicotiana tabacum* (Gorbunova and Levy 1997) conduit également à des insertions et/ou des délétions. Les insertions ont une taille maximum de 2,1 kb pour les cellules CHO et de 1,2 kb pour les cellules de *N. tabacum*. Elles sont constituées par un patchwork plus ou moins complexe de rétroïdes (*Alu* et *L1*), de plasmides et de séquences nucléaires quelconques. Aussi bien chez *S. cerevisiae* que dans les cellules d'eucaryotes pluricellulaires, des similarités de quelques bases entre les séquences insérées et les extrémités libres sont observées dans la majorité des insertions.

Cinq grands mécanismes sont donc à considérer pour la création des répétitions : (1) le dérapage lors de la réplication, (2) la recombinaison homologue, (3) la rétrotransposition, (4) les duplications de chromosomes et (5) la jointure de séquences non homologues. La description de ces mécanismes de duplication s'appuie sur des exemples de répétitions bien documentés.

Cependant, cette liste ne doit pas être considérée comme exhaustive, car il existe beaucoup d'autre répétitions, comme par exemple :

- les ARN non codants tels que les ARNr, souvent disposés dans les génomes en tandem multi-copie, les ARNt (86 copies chez *E. coli*, 275 chez *S. cerevisiae* et 497 chez *H. sapiens*) et les petits RNA (snRNA, tmRNA),
- les éléments répétés bactériens tels que les BIME (Bacterial Interspersed Mosaic Element), les REP (REPetitive element), les Rhs ou les séquences , dont la plupart ont des fonctions bien caractérisées (exception faite des BIME).

La partie suivante décrit non plus les mécanismes de duplication, mais les mécanismes qui agissent après la duplication, ceux qui ciblent les répétitions.

A.3. Mécanismes amorcés par les répétitions.

Parmi les cinq mécanismes présentés dans le chapitre précédent, au moins trois d'entre eux utilisent des séquences similaires préexistantes : la recombinaison homologue, le dérapage lors de la réplication et la jointure de séquences quelconques (quoique, dans ce dernier cas, la similarité ne paraît pas indispensable). Ainsi, ces mécanismes peuvent être considérés comme générateur de répétitions mais également comme mécanismes ciblant les répétitions. Dans cette seconde partie, seront donc abordés les mécanismes de la recombinaison homologue et ceux de la recombinaison non homologue (dérapage lors de la réplication et jointure de séquences non homologues). La méthylation, un mécanisme épigénétique inhibant la recombinaision entre les séquences répétées, sera ensuite exposé.

A.3.1. La recombinaison homologue.

La recombinaison homologue est définie ici comme un évènement d'échange (réciproque ou non) entre deux séquences similaires et faisant intervenir l'appareillage enzymatique de recombinaison. Ces évènements peuvent avoir lieu au cours de la mitose ou au cours de la méiose, bien qu'ils semblent plus fréquents dans le second cas : il se produit une centaine de recombinaisons au cours de la méïose I chez *S. cerevisiae* (Paques and Haber

1999). Cette recombinaison homologue est souvent divisée en deux classes : la recombinaison allélique, caractérisée par les échanges qui ont lieu entre des chromosomes homologues sur des mêmes loci (c'est la recombinaison "classique"), et la recombinaison ectopique qui concerne tous les autres évènements de recombinaison homologue (c'est celle qui est susceptible de cibler les répétitions).

A.3.1.1. Modèle de recombinaison homologue.

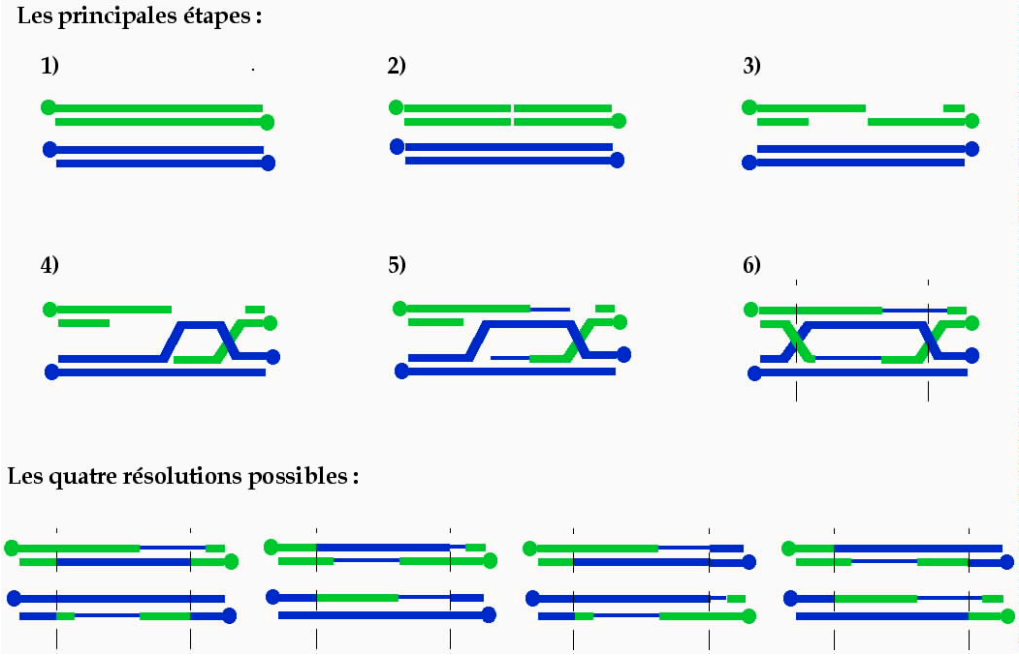

Figure 15 : Modèle de recombinaison méiotique.
Dans ce modèle, l'évènement déclenchant la recombinaison méiotique est une cassure double-brins (2). Cette cassure entraîne la dégradation 5'->3' des extrémités libres (3), puis l'envahissement de la séquence homologue par les extrémités 3' libres (4). Une courte synthèse (5) mène à deux jonctions de Holliday (6), dont la résolution donne quatre produits possibles (dont seulement deux crossing-over). Dans le schéma, les extrémités 5' sont marquées par des boules, les séquences synthétisées par un trait fin et la localisation les jonctions de Holliday par des pointillés.

Le modèle aujourd'hui le plus souvent admis pour expliquer la recombinaison homologue, et en particulier la recombinaison durant la méiose, est celui proposé par Szotack et ses collaborateurs (Szostak *et al.* 1983). Ce modèle, présenté sur la figure 15, permet de rendre compte du lien fort qu'il peut exister entre la conversion (échange non réciproque) et le crossing-over (échange réciproque). Dans ce modèle, l'évènement initiateur de la recombinaison est une cassure double-brins. Celle-ci va entraîner la dégradation (de 5' vers 3') des deux extrémités 5' libérées. Puis, les extrémités 3' saillantes vont envahir une

séquence homologue et initier une courte synthèse. La ligature de ces néo-brins avec les brins 5' laissés libres va produire une double jonction de Holliday. La résolution de cette structure va mener ou non à l'échange réciproque des séquences adjacentes, et la correction des mésappariements entre les deux jonctions de Holliday vont conduire (ou non) à une conversion. Ce modèle n'est pas incompatible avec des évènements de conversion provoqués par des mécanismes de type SDSA (Simple-Dependent Strand Annealing) ou BIR (Break Induced Replication) décris dans la partie précédente. Les relations complexes existant entre tous ces mécanismes ne seront pas abordées ici, mais ont été bien disséquées par F. Pacques dans sa revue (Paques and Haber 1999). Ainsi, la recombinaison homologue est ici définie comme les évènements d'échange (crossing-over et conversions) gérés par la machinerie enzymatique de recombinaison (le groupe épistatique de RAD52 de *S. cerevisiae* ou de RecA de *E. coli*).

A.3.1.2. Caractéristiques de la recombinaison homologue.

Quatre principaux paramètres semblent influencer le taux de recombinaison homologue : la longueur des séquences répétées, leur degré de similarité, leur localisation sur le chromosome et leur localisation relative.

Ainsi, il a été suggéré l'existence d'une longueur minimale en deçà de laquelle la recombinaison homologue ne peut avoir lieu (MEPS - Minimal Efficient Processing Segment). Il faut noter que les plus petites longueurs pour lesquelles ont été observées des évènements de recombinaison homologue sont d'environ 25 pb pour *E. coli* (Shen and Huang 1986), pour *S. cerevisiae* (Ahn *et al.* 1988) et pour les cellules EJ (*H. sapiens*) (Ayares *et al.* 1986). Néanmoins, chez *E. coli*, des expériences de recombinaison entre un phage et un plasmide montrent qu'une similarité de 20 paires de bases est suffisante pour obtenir un taux de recombinaison supérieur au taux basal (obtenu quand aucune similarité n'est introduite) (Watt *et al.* 1985). Par ailleurs, il a été montré que la fréquence de recombinaison est corrélée à la longueur des répétitions chez *E. coli* (Shen and Huang 1986), chez *S. cerevisiae* (Ahn *et al.* 1988; Jinks-Robertson *et al.* 1993), dans les cellules CHO (Scheerer and Adair 1994) et dans les cellules L (*M. musculus*)(Liskay *et al.* 1987).

Le second paramètre influençant la fréquence de la recombinaison homologue est la similarité partagée par les deux répétitions. Dans des cellules L (*M. musculus*), 19% de divergence entre deux séquences réparties sur la globalité de la séquence (ne laissant pas plus de 30 bases consécutives sans mésappariement) ne permettent plus d'observer des évènements de recombinaison (Waldman and Liskay 1987). Chez *E. coli*, une divergence de 10% entre les deux répétitions réduit les évènements de recombinaison d'un facteur 40 (Shen and Huang 1986). Chez *S. cerevisiae*, une divergence de 15% réduit de 75 fois les évènements de recombinaison entre deux gènes paralogues (Harris *et al.* 1993). La diminution de la similarité entre les deux séquences entraîne une baisse de la longueur de la répétition : en deçà d'un certain seuil, une partie trop divergente n'est vraisemblablement plus considérée comme similaire par les enzymes de recombinaison ; les deux paramètres longueur et identité sont donc fortement liés.

Le troisième paramètre analysé est la localisation chromosomique des copies de la répétition. En méiose, le taux de recombinaison allélique est très variable entre les différentes régions chromosomiques. Cette recombinaison étant initiée par une cassure double-brins, il est proposé que la fréquence de cassure double-brins soit corrélée au taux de recombinaison. La mesure les fréquences de coupure double-brins (en méiose I) le long du chromosome III de *S. cerevisiae* (Baudat and Nicolas 1997) a mis en évidence une hétérogénéité très importante entre les fréquences de recombinaison allélique à la méiose (une variation d'un facteur 50).

Le taux de recombinaison ectopique entre deux copies d'une répétition est limité par le taux de recombinaison allélique de chacune des copies (Lichten *et al.* 1987). Le taux de recombinaison ectopique entre deux copies est égal au taux de recombinaison allélique le plus faible des deux. Le taux de recombinaison ectopique est donc, comme le taux de recombinaison allélique, très variable (d'un facteur 40) en fonction de la localisation des deux copies.

Le dernier paramètre analysé est la localisation relative des deux copies. Ce paramètre semble influencer, dans certains cas, la fréquence de recombinaison en mitose.

Toutefois, chez *S. cerevisiae*, la fréquence de recombinaison méiotique varie peu en fonction de la localisation des copies de la répétition (Lichten and Haber 1989). Ceci suggère, ainsi qu'une expérience ultérieure (Haber and Leung 1996), qu'au cours de la mitose, une séquence donnée explore avec une même probabilité tout le génome à la recherche d'un partenaire potentiel pour la recombinaison homologue. Chez *B. subtilis*, certains auteurs ont mesuré la fréquence de recombinaison homologue entre un plasmide et des loci variables du chromosome et ont pu montrer que cette fréquence ne varie que faiblement (au maximum d'un facteur 3) (Biswas *et al.* 1992). Cependant, si les deux copies sont situées sur le même chromosome et proches l'une de l'autre, la fréquence de recombinaison est sensiblement augmentée. Chez *S. cerevisiae*, elle est accrue d'un facteur 10 si les deux copies sont situés à moins de 20 kilobases l'une de l'autre (Lichten and Haber 1989). Chez *E. coli*, ce seuil est estimé à moins de 1kb (Bi and Liu 1994).

Lorsque deux copies sont très proches, des mécanismes de recombinaison non homologue se mettent en place.

A.3.2. Les évènements de recombinaison non homologue.

Les mécanismes des évènements de recombinaison non homologue proposés sont le plus souvent le dérapage pendant la réplication et/ou la jointure de séquences non homologues (NHEJ, Non Homologous End Joining). Le premier mécanisme est plutôt avancé pour les bactéries et le second plutôt pour les eucaryotes. La question de savoir quel mécanisme semble le plus probable pour la recombinaison non homologue dans les différents organismes ne sera pas traitée ; les deux mécanismes paraissant impliqués à des degrés divers suivant les espèces. La plupart des résultats sur les paramètres qui les modulent proviennent d'études réalisées chez *E. coli* et *B. subtilis* et plus particulièrement sur des plasmides. Cependant, quelques études ont été réalisées dans les génomes eucaryotes (pour revue, voir (Klein 1995)).

La premier point important pour ces mécanismes est qu'ils ne concernent que des répétitions intrachromosomiques proches (séparées au plus de quelques kilobases). Ces

répétitions peuvent être directes (dans la même orientation) ou inversées (figure 16), bien que dans ce dernier cas, le taux de recombinaison soit souvent plus faible (Klein 1995).

Pour étudier les recombinaisons entre répétitions proches, ces dernières sont insérées dans des plasmides. Pour les répétitions directes, une recombinaison homologue produit une conversion et/ou une délétion alors qu'une recombinaison non homologue génère une délétion (Bi and Liu 1996b). Pour les répétitions inversées, une recombinaison homologue produit une conversion et/ou une inversion alors qu'une recombinaison non homologue génère un plasmide dimérique (Lyu *et al.* 1999). En outre, une répétition inversée bordée d'une répétition directe stimule fortement la recombinaison non homologue entre les deux répétitions directes (Peeters *et al.* 1988).

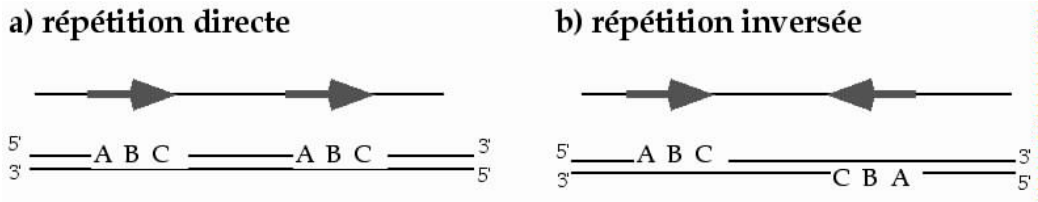

Figure 16 : Répétition directe et répétition inversée.
Une répétition directe est définie lorsque deux copies du même mot sont sur le même brin. Pour une répétition inversée, chaque copie est située sur un des brins.

Quelle que soit l'orientation relative des deux copies de la répétition, le taux de recombinaison non homologue est négativement corrélé à la distance séparant les deux copies, le "spacer" (cette propriété semble également vraie pour la recombinaison homologue). Chez *E. coli*, ceci a été observé pour des répétitions directes (Lovett *et al.* 1994) et inversées (Bi and Liu 1996a) et, chez *B. subtilis*, uniquement pour les répétitions directes (Chedin *et al.* 1994). Cette relation entre taille du *spacer* et le taux de recombinaison non homologue suit, dans tous les cas, une exponentielle décroissante.

Chez *E. coli*, la fréquence de recombinaison non homologue est positivement corrélée à la longueur de la répétition (Bi and Liu 1994). Néanmoins, cette augmentation de la fréquence atteint un plateau lorsque la longueur de la répétition dépasse 100 pb. De plus, il semble que la longueur minimum pour ce type de recombinaison soit en dessous de celle requise pour la recombinaison homologue. En effet, il a été observé de la recombinaison non

homologue entre des copies d'une répétition directe de 14 pb chez *E. coli* (Bi and Liu 1994) et 18 pb chez *B. subtilis* (Chedin *et al.* 1994).

Pour conclure, il est intéressant de regarder quel type de recombinaison (homologue ou non homologue) semble prédominant dans plusieurs conditions. Pour cela, Bi et Liu ont comparé les fréquences de délétion entre deux répétitions directes situées sur un plasmide dans plusieurs contextes (Bi and Liu 1994). Notamment, ils ont comparé systématiquement tous leur résultats obtenus dans une souche *RecA*$^+$ à ceux obtenus dans une souche *recA*$^-$. Le gène RecA pilote une grande part de la recombinaison homologue et est l'orthologue du gène RAD52 de *S. cerevisiae*. De ces expériences, il apparaît que :

- la délétion est plus fréquente dans la souche RecA$^+$, ce qui montre que les deux recombinaisons sont additives,
- si les deux copies sont très proches, les différences de fréquence de délétion entre les deux souches sont amoindries, et donc la recombinaison non homologue est majoritaire,
- si les répétitions sont de petite taille, les différences sont également amoindries, et donc la recombinaison non homologue est prépondérante.
- si les répétitions sont petites et très proches, il n'y a pas de différence entre les deux souches et donc seule la recombinaison non homologue est efficace.

A.3.3. La méthylation.

Les répétitions peuvent être la cible de mécanismes de recombinaison (homologue ou non homologue) qui produisent parfois d'importants remaniements chromosomiques. Une des questions soulevées par ce constat est comment des génomes composés pour moitié de répétitions (comme ceux des mammifères) peuvent rester stables ? On peut proposer, pour répondre partiellement à cette question, l'existence d'un mécanisme épigénétique faisant une barrière à la recombinaison entre les séquence répétées. Or, il semble que la méthylation de la cytosine soit un bon candidat car l'état de méthylation d'une répétition change sensiblement sa fréquence de recombinaison.

Dans les génomes eucaryotes, la plupart des répétitions sont méthylées (Yoder *et al.* 1997). Cette méthylation, qui semble se mettre en place très tôt au cours du développement (entre le cinquième et le septième jour du développement de *M. musculus*), concerne au moins les éléments mobiles, les satellites centromériques, les doublets CpG et les séquences soumises à l'empreinte parentale (Yoder *et al.* 1997). Ces dernières sont des séquences dont l'expression diffère selon qu'elles proviennent de la mère ou du père. Paldi et ses collaborateurs ont comparé, dans les régions connues pour être soumises à l'empreinte parentale, les taux de recombinaison méiotique (cM/Mb) de l'homme avec ceux de la femme (Paldi *et al.* 1995). Ils ont observé que l'état inactif de ces régions (état méthylé) est associé à une diminution de la fréquence de recombinaison. Ceci suggère donc que l'empreinte parentale (et donc la méthylation) peut être un des mécanismes responsables des différences de distances génétiques existant entre l'homme et la femme (Thomas and Rothstein 1991).

Dans le génome du champignon *Ascobolus immersus,* la relation entre la méthylation et la fréquence de recombinaison a été analysée plus en détail (Colot and Rossignol 1999). En effet, la présence d'une répétition à deux copies entraîne la méthylation rapide des deux copies. Comme pour la recombinaison, cette méthylation n'est plus observée en dessous d'une taille minimale (Goyon *et al.* 1996a). Cette taille minimale est d'environ 300 pb si les deux copies sont localisées en tandem et d'environ 600 pb si les deux copies sont dispersées dans le génome (Goyon *et al.* 1996a). La méthylation agit au niveau des éléments mobiles, de certaines régions du rDNA, mais pas sur les ARNt et les ARNr 5S. Ces derniers sont probablement des répétitions trop petites (respectivement 74 pb et 119 pb) (Goyon *et al.* 1996b). La fréquence de recombinaison entre deux séquences est abaissée d'au moins 50 fois si ces séquences sont méthylées (Maloisel and Rossignol 1998). En outre, seule la méthylation d'une des deux séquences suffit pour constituer cette barrière efficace contre la recombinaison.

Curieusement, chez *E. coli*, la méthylation produite par la Dam méthylase, entraîne une augmentation de la fréquence de recombinaison non homologue entre deux copies d'une répétition proche portant la séquence GATC (site reconnu par la Dam méthylase). Lovett et Feschenko observent que cet effet ne se produit que si les copies de la répétition sont peu

similaires. Ils proposent que cet effet de la Dam sur l'instabilité des répétitions en tandem soit lié à l'implication de la Dam dans les mécanismes de réparation des mésappariements (Lovett and Feschenko 1996). D'autres auteurs observent également la stabilisation d'une répétition multi-copie en tandem par la délétion du gène codant pour la Dam méthylase. Ils montrent, par cela, que cet effet n'est pas relié à la réparation des mésappariements (Troester *et al.* 2000).

La méthylation est, en conclusion, un facteur épigénétique important pour la stabilité des séquences répétées dans les génomes eucaryotes. Pourtant, ce mécanisme de stabilisation des répétitions ne semble pas s'étendre aux génomes bactériens (tout du moins pas à *E. coli*).

Au cours du premier chapitre, j'ai présenté certaines répétitions des génomes (eucaryotes et bactériens) ainsi que les mécanismes proposés pour leur genèse : dérapage lors de la réplication, recombinaison homologue, rétrotransposition, hyperploïdie et jointure d'extrémités libres. J'ai également présenté deux mécanismes remaniant les répétitions (la recombinaison homologue et la recombinaison non-homologue) ainsi que le mécanisme présumé qui protége les répétitions d'être remaniées. Ces mécanismes nous renseignent sur les contraintes structurales moléculaires pouvant s'appliquer sur les répétitions.

B. Conséquences des répétitions.

Ce chapitre est dévolu aux « fonctions » des répétitions et illustre les contraintes sélectives qui s'exercent sur celles-ci. La description des « petites répétitions » (dont la taille est inférieure à un gène), des répétitions de gènes et des répétitions géantes délimitent trois parties de ce chapitre.

B.1. Petites répétitions.

Les répétitions dont la taille est inférieure à celle d'un gène seront nommées « petites répétitions ». Parmi ces dernières, les plus étudiées sont les répétitions directes proches

(voire en tandem) ; c'est donc sur cette catégorie de répétition que porte la majorité de cette partie.

B.1.1. De la neutralité vers l'implication de la sélection.

Les premiers modèles simulant l'évolution des séquences répétées en tandem postulent que les répétitions ne sont quasiment pas soumises à des pressions sélectives (seule une très grande longueur étant délétère) (Charlesworth *et al.* 1994; Smith 1976; Stephan 1989). Ces modèles prédisent que les répétitions en tandem peuvent apparaître dans une séquence non répétée par crossing-over inégaux (Smith 1976) ou par une combinaison de crossing-over inégaux et de dérapages de la polymérase au cours de la réplication (Stephan 1989). Ces résultats montrent que plus le taux de crossing-over inégal est élevé, plus les répétitions ont des copies petites et similaires et qu'à l'inverse, si le taux de crossing-over est faible, les copies sont grandes et divergentes. De plus, à un faible taux de crossing-over, on observe l'apparition de répétitions d'ordre supérieur semblables à celles des satellites centromériques (figure 2). W Stephan propose donc que les différences de structure entre satellites, minisatellites et microsatellites ne soient que le simple reflet d'un taux de crossing-over différent (Stephan 1989). Cependant, dans une étude postérieure (Stephan and Cho 1994), l'auteur se ravise et propose que les pressions sélectives ont également un rôle dans la modulation de la longueur et de l'identité de ces répétitions.

Au vu des données biologiques, les pressions sélectives ne peuvent pas être reléguées au second plan et leur rôle dans la structuration de ces petites répétitions (au moins pour certaines d'entre elles) est certainement important. A travers deux exemples choisis, il sera montré que ces contraintes sélectives sont parfois positives, parfois négatives.

B.1.2. Un exemple de contrainte sélective positive : le rôle des satellites α.

Bien que la communauté scientifique s'accorde à penser qu'aucune protéine n'est codée par les satellites α, ces derniers font l'objet d'une controverse autour de leur possible implication dans les fonctions centromériques chez les primates.

Un des arguments prêchant contre un rôle fonctionnel des satellites α dans les centromères est l'absence de conservation, voire la grande divergence, de leur séquence et leur organisation entre espèces proches (Kato *et al.* 1999). D'autre part, les néocentromères, qui sont des centromères fonctionnels apparus récemment permettant notamment la ségrégation d'un chromosome artificiel, ne possèdent pas de satellites (Lo *et al.* 2001b; Saffery *et al.* 2001). Ces deux observations rendent compte du caractère dispensable des satellites dans les fonctions centromériques des primates. Des satellites de *H. sapiens* insérés dans un chromosome de *C. aethiops* (le singe vert d'Afrique), acquièrent une pleine fonction de centromère natif (Haaf *et al.* 1992). Si le centromère natif n'est pas rendu non fonctionnel, les deux centromères (natif et inséré) rentrent en conflit et provoquent des ségrégations anormales. Les satellites ne sont donc pas indispensables mais suffisants au bon fonctionnement d'un centromère de primate.

Plusieurs hypothèses, non exclusives, ont été émises quant aux mécanismes expliquant le rôle des satellites. Un premier mécanisme propose que les repliements ordonnés, qu'ils peuvent adopter *in vitro*, jouent un rôle dans la structuration en hétérochromatine (Gilbert and Allan 2001). La taille des satellites (171 pb) est supposée proche de celle nécessaire pour interagir avec un histone. Une seconde hypothèse propose que les satellites fixent diverses protéines importantes pour les centromères. C'est ainsi que des protéines associées aux centromères ont été découvertes et nommées CENP-A, B, C, D, G, H pour les protéines continuellement localisées aux centromères et CENP-E, F pour celles qui n'y sont localisées que ponctuellement. Par la suite, il ne sera détaillé que les fonctions de CENP-A et CENP-B, pressenties comme deux des acteurs majeurs (avec CENP-C) des fonctions centromériques.

La protéine CENP-A a été proposée récemment comme le moteur évolutif des satellites (Henikoff *et al.* 2001). C'est une protéine essentielle : la délétion du gène étant létale au stade embryonnaire chez *M. musculus* (Howman *et al.* 2000). L'absence de la protéine provoque la délocalisation des protéines CENP-B et CENP-C, ce qui confirme le rôle essentiel de CENP-A dans la structuration des centromères. CENP-A est une petite protéine homologue de l'histone H3, qui remplace cette dernière aux niveaux des centromères (Sullivan *et al.* 1994). Curieusement, ce remplacement s'effectue également dans les néocentromères, bien que ces derniers soient dépourvus de satellites (Lo *et al.* 2001a; Lo *et al.* 2001b). Tout ceci suggère que CENP-A est une protéine clef des fonctions centromériques, mais ne donne pas d'informations sur les structures satellites (la fixation de CENP-A ne semblant pas spécifique aux satellites).

La protéine CENP-B, dont la structure tridimensionnelle a été décrite récemment (Tanaka *et al.* 2001), est formée d'un domaine Hélice-Tour-Hélice de liaison à l'ADN et d'un domaine de dimérisation. Au niveau de l'ADN, elle se fixe sur une séquence de 17 paires de bases, nommée *CENP-B box*, située dans l'un des variants des satellites (Masumoto *et al.* 1989) ; les séquences adjacentes à ces 17 paires de bases semblent également augmenter l'affinité de la protéine pour son substrat (Muro *et al.* 1992). En outre, le variant de satellite où se fixe CENP-B est présent sur tous les centromères sauf sur celui du chromosome Y (seul chromosome qui ne possède pas de réel homologue) (Earnshaw *et al.* 1987). L'étude précise des sites correspondant aux évènements de recombinaison entre répétitions satellites, montre que ces sites sont systématiquement localisés à proximité des *CENP-B box* (Warburton *et al.* 1993). Cette observation a suggéré aux auteurs que CENP-B pourrait sous forme de dimère maintenir deux centromères proches l'un de l'autre (figure 17), favorisant les échanges aux abords des sites d'attache. D'autre part, la similarité frappante entre CENP-B et la transposase de *Tigger*, un transposon à ADN de la famille *pogo*, a fait naître l'hypothèse que CENP-B serait une relique dégénérée d'une transposase (Kipling and Warburton 1997). Les auteurs suggèrent donc que CENP-B pourrait avoir conservé l'activité endonucléasique de la transposase, qui stimulerait les évènements de recombinaison. Enfin, CENP-B possède un orthologue murin très conservé (Sullivan and Glass 1991). Tous ces arguments suggèrent que

CENP-B possède un rôle important dans les fonctions centromériques. Toutefois, la délétion du gène codant pour CENP-B chez *M. musculus* n'entraîne aucun phénotype visible, comme par exemple, des ségrégations anormales (Tomascik-Cheeseman *et al.* 2002).

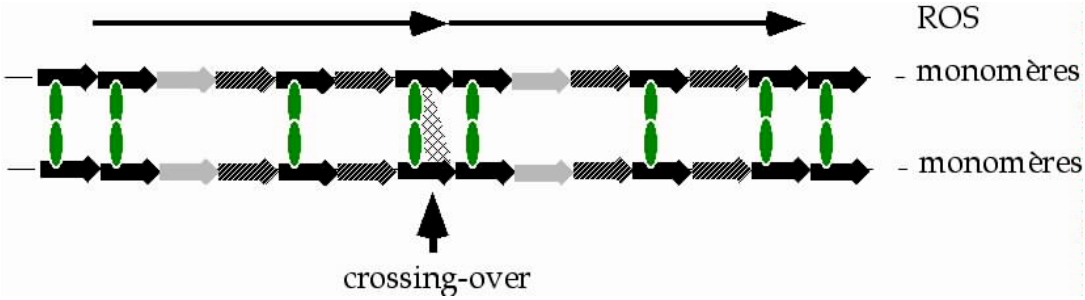

Figure 17 : Rapprochement de deux séquences satellites par CENP-B (d'après (Warburton *et al.* 1993)).
CENP-B se fixe sur un motif de 17 pb présent sur l'un des variants des satellites . CENP-B présente un domaine de fixation à l'ADN et un domaine de dimérisation. Le modèle exposé ci-dessus propose que CENP-B, en rapprochant deux satellites, permet d'accroître le taux de crossing-over. Les CENP-B sont représentées par les ovales verts, les variants des satellites par des flèches pleines et le crossing-over par des hachures.

Le modèle d'évolution des satellites par l'intervention de CENP-A (Henikoff *et al.* 2001), qui pourrait s'appliquer aux autres protéines se fixant sur les satellites (comme CENP-B), se fonde sur la sélection positive de certains satellites. Dans ce modèle, CENP-A aurait une affinité différente pour chacun des variants de satellites et cette affinité aurait changée au cours de l'évolution de CENP-A. La méiose ayant lieu lors de l'ovogenèse est asymétrique : seul un lot de chromosomes forme le gamète, les autres sont expulsés dans les globules polaires. Les auteurs proposent qu'une fixation de CENP-A pourrait entraîner un biais dans le positionnement des chromosomes : un centromère ayant beaucoup de CENP-A serait positionné préférentiellement pour migrer vers le gamète. Ainsi, un chromosome ayant une majorité de variants de satellites fixant CENP-A serait positivement sélectionné lors de l'ovogenèse. Un changement d'affinité de CENP-A pour un autre variant entraînerait la sélection positive d'un autre satellite.

Ainsi, les répétitions des satellites centromériques, suffisantes mais pas nécessaires pour former un centromère fonctionnel, constituent un exemple intéressant qui éclaire les fonctions de petites répétitions. Ainsi, grâce à l'exposé ci-dessus, il est possible de percevoir quels types de contraintes sélectives positives peuvent s'exercer sur de telles répétitions.

B.1.3. Un exemple de contrainte sélective négative : les pathologies liées aux SSR.

Un exemple bien connu des conséquences des petites répétitions en tandem est celui des pathologies associées aux SSR (Simple Sequence Repeats, voir le premier chapitre). Ces pathologies sont, pour la plupart, des maladies neurodégénératives causées par l'expansion d'une SSR dans un gène. De plus, ce sont surtout des SSR d'un motif de trois nucléotides qui paraissent impliquées dans ce type de pathologie. En général, il existe un polymorphisme dans le nombre de copie de la SSR, qui n'est pas associé à une pathologie. Cette dernière ne se déclare que si longueur de la SSR dépasse une certaine taille (spécifique à chaque SSR). Le tableau 3 dresse un aperçu non exhaustif de certaines des maladies associées à des expansions de SSR de trinucléotides.

Maladie	Gène		SSR		Nombre d'unités	
	Locus	Nom	Séquence	Localisation	Normal	Pathologique
X fragile	Xq27.3	FMR-1	CGG	5'UTR[b]	6-54	200-4000
Dystrophie myotonique de Steinert	19q13.3	?	GTG	3'UTR[b]	5-30	45-3000
Ataxie de Friedreich	9q13	X25[a]	GAA	Intron	7-22	200-900
Atrophie musculaire spino-bulbaire	Xq11.12	RA	CAG	Codant	17-26	40-62
Ataxie spinocérébelleuse dominante de type 1	6p22.23	Ataxine 1[a]	CAG	Codant	6-39	41-81
Ataxie spinocérébelleuse dominante de type 2	14q24.3-32	MJD1 [a]	CAG	Codant	13-36	68-79
Atrophie dentalo-rubro-pallidoluysienne	12p12.ter	Atrophine 1[a]	CAG	Codant	7-23	49-75
Maladie de Hungtington	4p16.3	IT15[a]	CAG	Codant	11-34	37-121

Tableau 3 : Pathologies liées aux expansions de SSR (d'après (Neri *et al.* 1996)).
a: gènes de fonction inconnue, b: UTR = UnTranslated Region (région non traduite)

Il est intéressant de mentionner que dans le cas de l'expansion liée à l'X fragile, on trouve des répétitions de 6 à 54 copies dans la population saine mais la pathologie n'est associée qu'à des répétitions de plus de 200 copies. Les individus présentant une SSR ayant entre 50 et 200 unités sont qualifiés de « prémutés ». En effet, ils présentent un phénotype « sain » mais leur descendance peut être atteinte. Chez *H. sapiens*, l'analyse du génotype des descendants atteints montre qu'ils possèdent une SSR avec plus de 200 copies.

La plupart des maladies présentées dans le tableau 3 sont associées à une expansion dans la partie codante du gène, ce qui produit une expansion dans la séquence protéique. Toutefois, d'autres expansions ont lieu dans les parties non codantes (5'UTR, 3'UTR ou intron) (Mitas 1997). L'analyse systématique de la répartition des SSR , ayant au moins 8 copies, dans les gènes a permis d'établir que :

- Dans la partie codante, la répétition de CAG est la plus souvent observée (Stallings 1994). Ceci corrobore le nombre important de pathologies liées aux expansions de Glutamine (Lebre and Brice 2001). Cette abondance de maladies dites «□à polyglutamine□» est surprenante. Un auteur propose qu'une petite répétition polyglutamine troublerait peu la fonction de la protéine et serait donc peu (ou pas) sélectionnée négativement (Katti *et al.* 2001).
- La répétition de CAG n'est jamais présente dans les introns (Stallings 1994). Les auteurs proposent que, due à sa ressemblance avec les signaux d'épissage, sa présence perturberait le bon épissage du messager.
- Le nombre de copies des SSR est plus important dans les gènes humains que dans leur orthologues chez les primates, et bien plus important que dans leur orthologues chez les rongeurs (Djian *et al.* 1996). Cette différence du nombre de copies suggère que les SSR serait plus stables chez les rongeurs. La SSR présente dans le gène humain RA (voir tableau 3), insérée dans l'orthologue murin de ce gène, ne présente plus aucune instabilité. Les auteurs proposent donc que l'instabilité des SSR soit un phénomène développé surtout chez les primates (Bingham *et al.* 1995).

Plusieurs conséquences ont été proposées pour expliquer comment les expansions de SSR peuvent aboutir à une pathologie. Pour les maladies associées aux expansions de glutamine, l'hypothèse souvent émise est une aggrégation des protéines portant l'expansion de glutamine. En effet, il semble que ces domaines polyglutamines soient impliqués dans des interactions entre protéines (Lebre and Brice 2001). Pour l'X fragile, le mécanisme responsable de la maladie n'est pas encore très clair. D'une part, l'expansion supprime la fonction du gène FMR1, dont la délétion entraîne, chez la souris, le phénotype pathologique (retard mental, hyperactivité et macroorchidisme[1]) (Consortium 1994) . D'autre part, la présence d'une grande région répétée fragilise le chromosome qui se casse et pourrait être la cause réelle de la pathologie (Sutherland *et al.* 1998). Pour d'autres pathologies, la présence de grandes SSR entraîne des repliements dans les séquences nucléiques et en particulier dans

[1] grosses testicules.

l'ARN messager (Mitas 1997). Ces repliements pourrait empêcher la traduction, abaissant ainsi la quantité de protéine.

Les pathologies exposées ci-dessus sont dues à l'expansion de SSR de trois nucléotides. Cependant, il existe d'autres types de SSR pouvant entraîner une pathologie humaine. Par exemple, dans l'épilepsie myoclonique progressive, on retrouve l'expansion d'un dodécamère dans le gène codant pour la cystatine B (Lalioti *et al.* 1999). Ainsi, à travers ces quelques exemples, il apparaît clairement que les petites répétitions peuvent être soumises à des pressions de sélection négatives fortes.

B.1.4. Les gènes immortels.

Dans les années 80, S. Ohno proposa que les petites répétitions puissent être à l'origine d'un certain nombre de gènes (Ohno 1985). Ceux-ci, qu'il qualifia de « gènes immortels », possèdent une structure telle qu'ils deviennent polymorphes et résistants aux mutations entraînant un changement de cadre de lecture. Cette structure est formée de petites répétitions en tandem (figure 18).

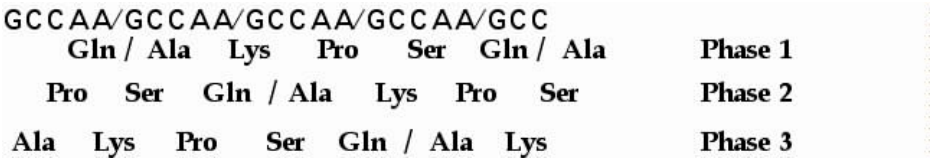

Figure 18 : Structure d'un « gène immortel » (d'après (Ohno 1987b)).
Ces gènes sont formés par la répétition en tandem d'un motif non multiple de trois (ici cinq). Le gène code ainsi dans les trois phases, sur une grande longueur et contient des répétitions peptidiques.

Plusieurs arguments supportent cette hypothèse (Ohno 1984). En effet, les petites répétitions en tandem :

- permettent de créer une phase ouverte de lecture, qu'il serait improbable d'obtenir par simple juxtaposition aléatoire de nucléotides,
- établissent une périodicité dans le gène (et donc dans la protéine) telle que l'on peut en observer dans les structures protéiques (hélices et feuillets),
- abolissent les effets des insertions et des délétions si la taille de la copie n'est pas un multiple de trois. En effet, dans ce cas, aucune des trois phases ne contient de

codon STOP et une insertion (ou une délétion) ne change le cadre de lecture que localement.

D'autre part, en replaçant ces répétitions à l'origine même des séquences nucléotidiques (en chimie prébiotique), Ohno ajouta à la liste des "avantages" de ces répétitions de quelques nucléotides la propriété, si elles sont simples brins, de s'autoassembler en grandes séquences double-brins (figure 19) (Ohno 1987a).

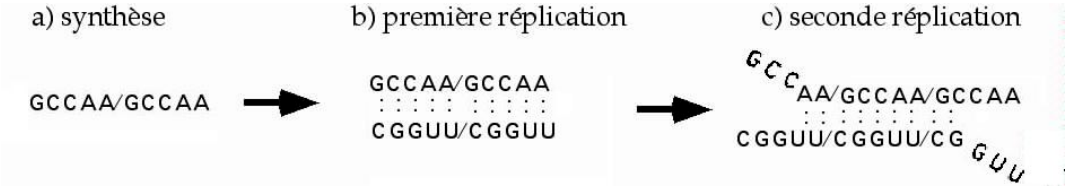

Figure 19 : Autoassemblage d'une répétition pentanucléotidique (d'après (Ohno 1987b)).
Ce schéma présente comment une répétition de cinq nucléotides peut s'autoassembler (dans des conditions de chimie prébiotique). Seule la première répétition doit être obtenue par synthèse sans matrice. a) synthèse de novo *de deux copies. b) première réplication créant le complémentaire de la répétition. c) dissociation et réappariement en décalé. Nouvelle réplication agrandissant la répétition.*

L'analyse de quelques séquences codantes révèle que cette idée originale pourrait être une des voies de construction de certains gènes. Un premier exemple est celui du gène codant pour la 6-AHA LOH (6-AminoHexadoic Acid Linear Oligomer Hydrolase), une protéine impliquée dans la dégradation du nylon chez la bactérie *Acromobacter vobacterium* (Ohno 1984). La phase ouverte de lecture codant pour cette enzyme est incluse dans une autre phase ouverte de lecture, plus grande et dont la phase est décalée d'un nucléotide. Cette seconde grande phase, nommée PRC (Preexisting Coding Sequence), est cependant interrompue par un codon STOP. Ohno propose que le gène codant pour la 6-AHA LOH est issu de PRC par un récent décalage de phase de lecture. D'autre part, due à la rareté des codons stop dans les trois phases, la propension de cette séquence à coder pour une protéine indépendamment du cadre de lecture est importante. La présence de nombreuses répétitions internes, dont le très abondant décamère CGACGCCGCT , suggère que ce gène est la relique d'un assemblage de répétitions en tandem. Dans un second exemple , Ohno analyse la séquence qui code pour l'histone H1 (Ohno 1987b). Dans cette dernière, il décrit un pentamère (CCAAG) présent à 25 copies, codant pour 20% de la protéine. Enfin, chez *S. scrofa*, la séquence du gène codant pour le récepteur muscarinique à l'acéthylcholine constitue un troisième exemple (Ohno 1987a). Ce gène est constitué pour 70% de trois

heptamères dispersés dans la séquence et les 30% restant sont probablement des reliquats dégénérés de ces heptamères.

Ces exemples illustrent les fonctions de ces « petites répétitions » en tandem. Ainsi, ces répétitions peuvent parfois être à l'origine de certains gènes et donc de certaines fonctions. Cependant, la plupart des gènes ne semblent pas dériver de répétitions internes, mais souvent ces répétitions peuvent plus modestement être impliquées dans des répétitions de motifs protéiques.

B.1.5. Répétitions de motifs protéiques : le cas de l'involucrine.

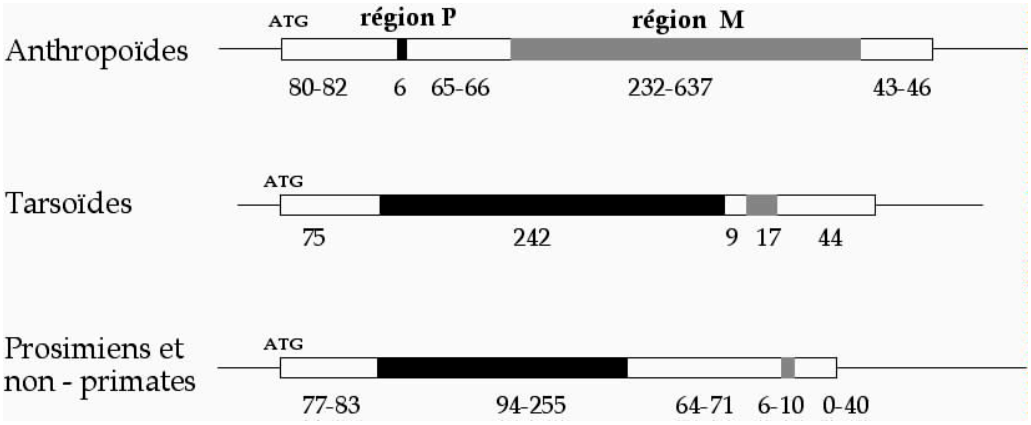

Figure 20 : Structure de l'involucrine dans différents phylums.
L'involucrine présente deux régions répétées M (pour Moderne), répétées chez les primates à un grand nombre de copies et la région P, répétée chez les non-primates (et les prosimiens). Les tarsoïdes appartiennent à un phylum intermédiaire.

L'involucrine est une protéine synthétisée par les kératinocytes, intervenant dans la formation d'une enveloppe. Elle est la cible de la transglutaminase, une enzyme créant une liaison entre une glutamine et une lysine. La composition particulière en acides aminés de l'involucrine humaine (plus de 45% de glutamine et glutamate) en font un substrat de choix pour la synthèse de cette enveloppe. La séquence du gène codant pour cette protéine a été déterminée chez *H. sapiens* en 1986 (Eckert and Green 1986). Cette séquence révéla une structure interne singulière : au milieu du gène est située une répétition d'une trentaine de copies de 30 nucléotides (région M, figure 20). Les copies de cette région ne sont pas toujours exactement répétées, ce qui rappelle la structure des minisatellites (voir le chapitre *Mécanismes moléculaires*). La fréquence extrêmement élevée du codon CAG dans le gène et

des codons proches (possédant deux des trois nucléotides, comme TAG, CTG, etc.) a permis de proposer que cette protéine dérive de la répétition d'un antique motif CAG. La détermination de la séquence de ce gène chez *Lemur catta* (un lémurien, prosimien) et *Gorilla gorilla* a permis de relever au moins deux caractères importants de l'évolution de cette protéine : (1) Chez *L. catta*, la protéine présente une répétition d'un motif de 13 à 16 acides aminés et est située en amont de la répétition présente chez *H. sapiens* (Tseng and Green 1988). (2) Chez *G. gorilla*, la répétition est localisée au même endroit que celle de *H. sapiens* (Teumer and Green 1989).

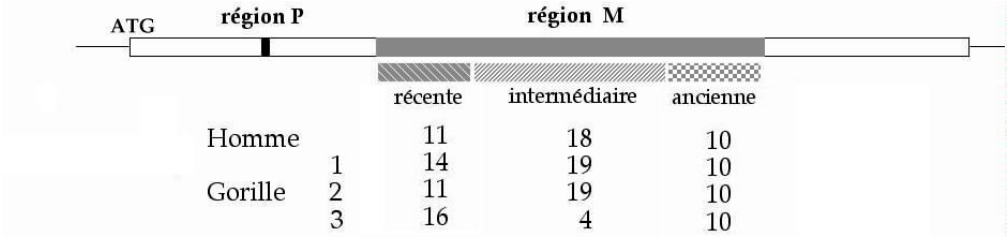

Figure 21 : Les trois zones de la région M : ancienne, intermédiaire et récente.
La partie ancienne est quasi invariante entre H. sapiens *et* G. gorilla*, et la partie récente est différente dans les quatre allèles de l'involucrine. Le nombre de copies pour chaque partie est donnée dans le bas du dessin.*

D'autre part, trois allèles ayant été séquencés chez *G. gorilla*, il fut possible, en comparant les quatre répétitions (3 de *G. gorilla* et 1 de *H. sapiens*), de définir, dans la région répétée, trois zones (Teumer and Green 1989) (figure 21) :

- une partie dite « ancienne », du coté 3' de la région répétée, qui ne présente pas de variations majeures entre les quatre répétitions orthologues,
- une partie dite « intermédiaire », où les copies sont présentes au moins sur deux des quatre répétitions,
- une partie « récente », située du coté 5' et propre à chaque répétition.

La détermination de la séquence chez d'autres singes, présentée sur le tableau 4, a permis de confirmer ces résultats préliminaires. Notamment, les trois zones définies par les comparaisons *H. sapiens - G. gorilla*, s'avèrent pertinentes. De ces résultats, il apparaît ainsi que l'évolution de l'involucrine dans les génomes des singes du nouveau et de l'ancien monde se fait principalement dans cette répétition décapeptidique. De plus, bien que les changements soient surtout des additions (ou des délétions) de copies dans la partie « récente » de la répétition, certaines délétions sont observées dans les autres régions.

Espèce	Phylum	Région M			Référence
		Récente	Intermédiaire	Ancienne	
Aotus trivirgatus	Platyrhiniens	7	13-18	10	(Tseng and Green 1989)
Cebus albifrons	Platyrhiniens	0	13	11	(Phillips *et al.* 1990)
Saguinus oedipus	Platyrhiniens	2	15	10	(Phillips *et al.* 1990)
Macaca fascicularis	Cercopithécoïdes	0	27	9	(Djian and Green 1992)
Macaca mulatta	Cercopithécoïdes	0	19	9	(Djian and Green 1992)
Cercophithecus hamlyni	Cercopithécoïdes	0	16	7	(Djian and Green 1992)
C. aethiops	Cercopithécoïdes	0	17	7	(Djian and Green 1992)
Hylobates lar	Hylobatoïdes	5	17	10	(Djian and Green 1990)
P. pygmaeus	Hominoïdes	31	18	10	(Djian and Green 1989a)
G. gorilla	Hominoïdes	11-16	4-19	10	(Teumer and Green 1989)
P. paniscus/ P. troglodytes	Hominoïdes	0	14	10	(Djian and Green 1989b)
H. sapiens	Hominoïdes	11	18	10	(Teumer and Green 1989)

Tableau 4 : Taille des zones de la région M des Simiiformes.
Pour chaque grand phylum des simiiformes (Platyrhiniens, Cercopithécoïdes, Hybatoïdes et Hominoïdes) sont présentées quelques espèces et le nombre de répétitions dans les segments récents, intermédiaires et anciens de la région M du gène de l'involucrine.

Les gènes codant pour les involucrines de mammifères non-simiiformes, présentent aussi une répétition de 10 à 16 codons dans une autre partie du gène. Ceci a amené à définir deux loci répétés dans le gène de l'involucrine : le locus P, répété chez tous les mammifères sauf les simiiformes et le locus M répété spécifiquement chez les simiiformes. L'analyse des séquences de la répétition P montre, comme pour la répétition M, un nombre de copies très polymorphe entre les espèces (de 13 à 20 codons). L'étude de sept allèles de ce gène, chez

M. musculus, montre que le nombre de copie varie entre 20 et 38 selon les allèles (Delhomme and Djian 2000). Bien que la séquence du motif répété varie beaucoup entre les espèces, elle est similaire entre les différentes copies de la même répétition (Phillips *et al.* 1990). Cette dernière observation suggère qu'un mécanisme, peut-être apparenté à la conversion, homogénéise les différentes copies de la même répétition ; comme les copies proches se ressemblent plus, il faut inférer que ce mécanisme agit de proche en proche.

La répétition M est absente de la plupart des mammifères, on peut en conclure que cette répétition est apparue dans le phylum des simiiformes. A l'inverse, l'absence de la répétition P suggère sa disparition dans ce phylum. Afin d'obtenir des précisions sur la transition entre les deux états du gène (répété en M ou répété en P), les séquences des gènes codant pour l'involucrine de *Tarsius bancanus*, très proche phylogénétiquement des simiiformes, ont été déterminées. Dans cette espèce, le gène présente les deux répétitions, la répétition P avec 17-18 copies et la répétition M avec 2 copies (Djian and Green 1991). La disparition de la répétition P semble donc rapide et postérieure à l'apparition de la répétition M.

L'influence de cette transition sur la fonction de la protéine est difficile à analyser, d'autant plus que la fonction de l'involucrine n'est pas complètement claire. En effet, la délétion de ce gène chez la souris n'entraîne pas de phénotype visible (Djian *et al.* 2000). Les différentes formes d'involucrine ont des compositions moyennes en acides aminés différentes. Les proportions moyennes des différents acides aminés en fonction du site et du phylum (non-primates, haplorrhiniens-et-tarsiformes ou simiiformes) sont représentées sur la figure 22. Il faut noter au vu de cette figure que les proportions de Lysine et de Glutamine, substrats de l'enzyme créant les liaisons covalentes entre involucrines, ne changent pas beaucoup entre les trois catégories. Une étude du ratio des taux de substitutions synonymes et non-synonymes montre que l'involucrine est peu contrainte par la sélection (Green and Djian 1992). Cependant, il est intéressant de noter que, malgré la grande variabilité de cette protéine, les cystéines 44 et 75 sont conservées dans quasiment toutes les espèces (la cystéine 75 pouvant disparaître si beaucoup d'autre cystéines sont présentes dans la séquence). Ceci suggère un rôle important pour ces cystéines dans la fonction de la protéine.

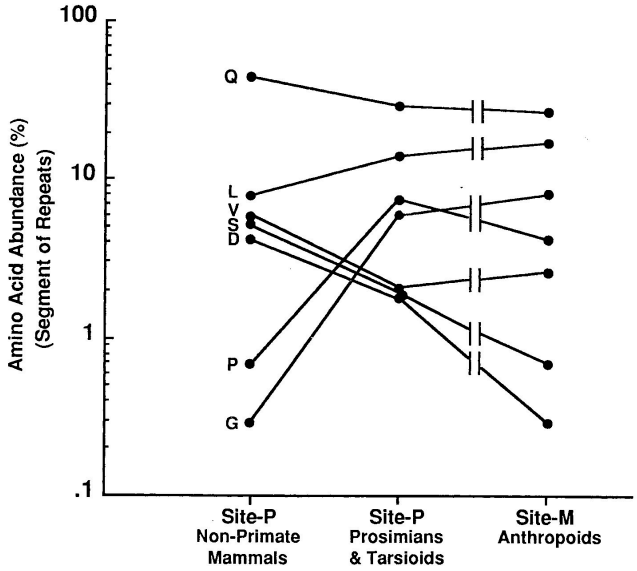

Figure 22 : Composition des répétitions de l'involucrine (d'après (Green and Djian 1992)).
Composition relative de certains acides aminés dans les régions P et M de l'involucrine. Si certains acides aminés ont une abondance très variable, la lysine et l'acide glutamique, qui sont les réels acteurs de la fonction de l'involucrine, ne présentent pas de fort changement de composition.

Pour conclure, cette protéine est un cas intéressant car les répétitions qu'elle contient présentent un fort taux de variation qui entraîne un changement fort dans sa teneur en acides aminés. L'involucrine illustre élégamment quelles conséquences peuvent avoir ces « petites répétitions » sur les protéines. Ce type de répétitions est également décrit dans le gène codant pour le collagène (Yamada *et al.* 1980). Dans ce dernier, chaque copie, qui a une taille de 54 bases est localisée dans un exon différent. Les auteurs proposent que l'amplification d'un putatif exon ancestral de 54 bases serait la clef de l'évolution de ce gène. Enfin, la répétition de motif protéique est très courante dans les protéines, puisqu'une étude récente montre que 14% des protéines sont constituées en partie d'une répétition (Marcotte *et al.* 1999).

B.2. Les répétitions de gènes.

Dans la partie précédente, je vous ai présentés quelques exemples permettant d'illustrer les conséquences et donc les contraintes sélectives qui peuvent s'appliquer sur des répétitions dont la taille ne suffit pas à coder un gène entier. Par la suite, nous étudierons les conséquences de la répétition de gènes entiers. En premier lieu, seront exposées, à travers

quelques exemples, les conséquences « immédiates » d'une duplication génique, puis, dans un second temps, les conséquences « à long terme » de ces duplications de gènes.

B.2.1. Conséquence immédiate : l'amplification d'une fonction.

La duplication se définie par la "copie" d'une séquence préexistante au sein du même génome. Elle entraîne, par essence, une redondance exacte de la séquence dupliquée. Si cette séquence code pour un gène et si aucun rétrocontrôle n'est impliqué dans l'expression de ce gène, alors la quantité de protéine est doublée. Même si l'on sait cette simplification osée, il faut admettre, devant les réalités biologiques, que, dans un certain nombre de cas, la duplication d'un gène aboutit à l'augmentation du nombre de protéines (ou d'ARN). C'est par exemple le cas des ARN dits « de ménage », comme les ARN ribosomiques et les ARN de transfert (Ohno 1970). Il est communément admis que les nombreuses copies de ces gènes reflètent leur fort taux d'expression. D'autres gènes sont dupliqués dans des situations plus ponctuelles. C'est notamment le cas du gène CUP1 et ACP1 (renommé PHO3) de *S. cerevisiae* et du gène codant pour la DHFR dans les cellules de mammifères.

B.2.1.1. Les duplications de CUP1 et la résistance aux métaux lourds.

Le gène CUP1 code pour une métallothionéine impliquée dans la détoxication des métaux lourds, comme le cuivre, chez *S. cerevisiae*. Le locus associé à cette résistance au cuivre contient en réalité plusieurs copies répétées en tandem du gène CUP1 (Fogel and Welch 1982). De façon remarquable, le nombre de copies du gène CUP1 détermine la concentration maximum de cuivre supportée par cet organisme : une levure dite sensible (mourant à une concentration de cuivre supérieure à 0,3 mM) ne possède qu'une copie du gène, tandis qu'une levure résistante (jusqu'à 2mM de cuivre) en possède une quinzaine et une levure hyper-résistante (jusqu'à 12 mM de cuivre) une vingtaine (Fogel and Welch 1982). Dans les souches de *S. cerevisiae* utilisées dans l'industrie, une étude ultérieure des longueurs variables du locus CUP1 a renforcé cette corrélation entre le nombre de copies du gène CUP1 et le taux de résistance au cuivre de l'organisme (Welch *et al.* 1983).

La variabilité entre les souches résistantes au cuivre n'est pas seulement présente au niveau du nombre de copies du gène CUP1, mais également au niveau de sa structure et de sa localisation chromosomique (Welch *et al.* 1983). En effet, il a été décrit, dans une souche résistante, un gène CUP1 de plus petite taille et dans d'autres souches, des loci secondaires contenant les répétitions de CUP1.

Le nombre de copies des gènes CUP1 est variable au sein d'une même souche. En effet, 10% des méioses sont associées à un changement dans le nombre de gènes CUP1 (Welch *et al.* 1990). Ce taux important de variation dans le nombre de copies n'est pas sans rappeler les expansions et contractions de minisatellites chez *H. sapiens*. Ainsi, il a été montré que ce locus est souvent impliqué dans un chiasma au cours de la méiose. Les chiasmas étant la marque cytologique des crossing-over, cela suggère que le taux de recombinaison à ce locus doit être probablement élevé.

D'autre part, le gène CUP1 possède également une fonction dans la résistance au Cadmium (Jeyaprakash *et al.* 1991). Cette propriété est intéressante car elle illustre la plurifonctionnalité de certaines enzymes *a priori* spécialisées. Bien que l'action de détoxication du cuivre, utilisé en vinification comme fongicide, soit vraisemblablement la cause de la sélection de ce gène, l'enzyme conserve sa fonction généraliste. Comme pour la résistance au cuivre, la résistance au cadmium est accrue avec le nombre de copies du gène CUP1.

Ainsi, CUP1 est un exemple intéressant de conséquence à cours terme d'évènements de duplication, qui permet ainsi aux levures de devenir résistantes aux métaux lourds dans certaines conditions.

B.2.1.2. Les duplications du gène de la DHFR et la résistance au méthotréxate.

Un autre exemple intéressant est le cas des amplifications géniques dans la résistance au méthotréxate (MTX) dans les cellules de mammifères CHO (hamster) et S-180 (souris). Le gène amplifié code pour la Di Hydro Folate Réductase (DHFR), une enzyme impliquée dans

la synthèse des purines, du dTMP et de la glycine. Le MTX est un inhibiteur compétitif de cet enzyme, et est donc un poison pour la cellule, qui ne survit pas à l'absence des métabolites décrits ci-dessus. Cependant, il est possible d'obtenir des cellules CHO ainsi que des cellule S-180 résistantes au MTX.

- Dans les cellules CHO, le caractère de résistance au MTX se conserve au cours des générations même si les cellules ne croissent plus en présence du poison (Nunberg *et al.* 1978).
- A l'inverse, les cellules S-180 perdent le caractère de résistance si les cellules ne croissent plus en présence de MTX (Kaufman *et al.* 1979).

Bien que les mécanismes de résistance paraissent différents dans les deux lignées de cellules, le nombre de copies du gène DHFR est très augmenté dans les cellules résistantes des deux lignées ; dans les cellules sensibles, il n'existe qu'une copie du gène et il en existe jusqu'à 200 dans les cellules résistantes. Ainsi, la résistance est due à une multiplication importante de la cible du poison et donc à la persistance de l'activité enzymatique malgré la présence de MTX.

Dans les cellules CHO, les gènes DHFR sont répétés en tandem sur le chromosome ; cette répétition en tandem étant relativement stable, le caractère de résistance se maintient même si les cellules ne croissent plus en présence de MTX (Nunberg *et al.* 1978).

Dans les cellules S-180, les amplifications sont portées par de petits fragments d'ADN extrachromosomiques nommés doubles-minutes (Kaufman *et al.* 1979). Ces doubles-minutes ne contiennent pas de centromère et sont donc perdus, dilués au cours des générations. Cependant, en plus de l'effet de dilution, des effets sélectifs se surajoutent pour la perte de ces doubles-minutes. En effet, en absence de MTX, les cellules croissent d'autant mieux qu'elles n'ont pas (ou peu) de doubles-minutes (Kaufman *et al.* 1981). Ainsi, au cours des générations (sans MTX), les cellules sans double-minute sont avantagées et envahissent la population. Ceci aboutit à la perte de la résistance pour ces cellules S-180.

Ces deux mécanismes (amplification extra et intrachromosomique) ne sont bien sûr pas exclusifs, voire complémentaires .

- Si des cellules CHO sont mises en présence de MTX pendant quelques générations seulement, les cellules résistantes n'ont quasiment que des amplifications de doubles-minutes (Kaufman and Schimke 1981).
- Des cellules S-180 qui ont poussé plusieurs années en présence de MTX présentent des amplifications stables (des répétitions en tandem) (Kaufman *et al.* 1981).

Les deux mécanismes se mettent donc en place à des phases différentes de l'adaptation au poison. Les doubles-minutes constituent un moyen rapide de parer à une élévation brutale de MTX mais sont désavantageux pour la croissance des cellules. Après les doubles-minutes, les duplications en tandem se mettent en place et remplacent dans la population les doubles-minutes. La différence entre les cellules CHO et S-180 réside donc dans le temps nécessaire pour passer du premier au second mécanisme.

B.2.1.3. La duplication de ACP1 chez S. cerevisiae.

Un troisième exemple de conséquence « immédiate » issue de la duplication d'un gène, est celui de la duplication du gène ACP1, qui code pour une phosphatase.

Hansche et ses collaborateurs ont mis en place une expérience de compétition entre des levures dans un chémostat. Dans ce dernier, la seule source de carbone était le β-glycérophosphate (qui nécessite l'action d'une phosphatase pour produire du glycérol) et le pH était à 6 (ou la phosphatase n'a que 40% de son activité). Avant 1000 générations, sur deux expériences parallèles, les levures qui avaient colonisé le chémostat étaient (1) des mutants de croissance, (2) des mutants d'agrégation (moins sensibles aux dilutions) et (3) des mutants de la phosphatase (codée par le gène ACP1) ayant une activité accrue à pH 6 (Francis and Hansche 1972; Francis and Hansche 1973). Cependant, après 1000 générations, ils observèrent également une souche portant une duplication d'une des phosphatases mutées (avec une activité accrue) (Hansche 1975). Cette souche présentait une activité très accrue pour cette phosphatase par rapport à la souche sans duplication. L'étude génétique de la seconde copie du gène ACP1 montra qu'elle n'était pas liée à une hyperploïdie et qu'elle n'était pas située sur le même chromosome. Dans une seconde étude, ils générèrent d'autres

duplications du gène (cinq au total, mais dont trois furent obtenues par irradiation aux UV) (Hansche *et al.* 1978). L'analyse génétique de ces évènements montra qu'il s'agissait de cinq duplications géantes vers un autre chromosome (qui resta non identifié). Ces évènements de translocations non-réciproques pourraient être sans doute aujourd'hui mieux analysés, mais les expériences n'ont pas été poursuivies.

Dans ce dernier exemple, il est intéressant de voir que des changements d'environnement (ici de source de carbone), associés à une compétition entre organismes peuvent induire la sélection de changements moléculaires (comme des duplications de gènes).

Les duplications permettent donc de répondre à des changements de milieu de plusieurs types : résistance à un poison par amplification du gène responsable de la détoxication (cas de CUP1), résistance à un inhibiteur spécifique par amplification de la quantité d'enzyme potentiellement inhibée (cas de la DHFR), meilleure assimilation d'un substrat rare par amplification d'une enzyme impliquée dans sa dégradation (cas de ACP1). Par ailleurs, ces mécanismes d'adaptation à un nouvel environnement génèrent des duplications le plus souvent instables. C'est le cas de certaines duplications observées chez les bactéries, dont les copies ne persistent pas au cours des génération (Jackson and Yanofsky 1973). C'est ainsi que AL Koch montra à l'aide de simulations informatiques que, pour des répétitions en tandem, les copies ont tendance à disparaître dès que les pressions sélectives qui les ont créées disparaissent (Koch 1979).

B.2.2. Conséquence à long terme : l'émergence de nouvelles fonctions.

La partie précédente s'est focalisée sur les conséquences « immédiates » d'évènements de duplications géniques. Elle donne donc une clef importante pour comprendre comment des grandes familles de gènes ont pu s'établir dans les génomes. Toutefois, les duplications géniques jouent également un rôle important dans l'émergence de nouvelles fonctions. Le nombre de familles de gènes dont les fonctions sont divergentes est

tellement important qu'il n'est pas envisageable d'en faire ici un catalogue. A titre d'exemple, il a été estimé la proportion de gènes non uniques (présentant au moins un paralogue) dans différents organismes. Cette proportion est de 31% chez *E. coli* (Blattner *et al.* 1997), 51% chez *Mycobacterium tuberculosis*, 30% chez *S. cerevisiae* (Coissac *et al.* 1997), 35% chez *A. thaliana* (TAGI 2000) et 38% chez *H. sapiens* (Li *et al.* 2001). Les descriptions précises de familles de gènes abondent dans la littérature, qu'elles soient petites ou très grandes. Parmi les plus grandes familles, on trouve les ABC-transporteurs (ABC : ATP-Binding Cassette) (Dean *et al.* 2001) , les canaux ioniques (Lopreato *et al.* 2001) et les récepteurs nucléaires (Laudet *et al.* 1992). Par ailleurs, certaines familles de gènes, comme les célèbrissimes gènes *Hox* (Gellon and McGinnis 1998), illustrent la diversité apportée par les duplications de gènes.

Après un bref coup d'œil sur une étude récente d'une famille comprenant trois paralogues chez *S. cerevisiae*, seront décrits deux exemples de famille de gènes, peut-être parmi les plus anciens : les globines et les gènes du système immunitaire (et en particulier ceux du CMH).

B.2.2.1. Une divergence récente : des gènes impliqués dans la synthèse de la Thiamine.

Une question soulevée par la découverte de nombreux gènes paralogues chez *S. cerevisiae* est la relation entre gène dupliqué et redondance fonctionnelle. En d'autres termes, quel est le rapport entre la similarité de séquence et la similarité fonctionnelle ? Une récente étude s'est penchée sur cette question grâce à l'étude de trois gènes paralogues : YOL055c, YPL258c et YPR121c (Llorente *et al.* 1999). Ces trois gènes semblent impliqués dans la synthèse de la Thiamine PyroPhosphate, coenzyme indispensable à la réaction transformant le pyruvate en acétyl-coA (réaction clef pour le catabolisme du pyruvate, produit de la glycolyse).

Les similarités de séquence protéique entre ces gènes sont les suivantes : 80% d'identité entre YOL055c et YPL258c, 76% d'identité entre YOL055c et YPR121c et 78% d'identité entre YPL258c et YPR121c.

La délétion de chacun de ces gènes n'est associée à aucun phénotype, mais la délétion conjointe de YOL055c et YPL258c entraîne une auxotrophie pour la Thiamine. Ainsi, les deux gènes les plus similaires semblent partager une fonction similaire, et le gène YPR121c, certainement dupliqué avant la divergence YOL055c-YPL258c, ne partage plus la même fonction.

Par ailleurs, l'étude du profil d'expression de ces trois gènes montre qu'ils sont encore interconnectés. En effet, leur expression est sous le contrôle de deux autres gènes impliqués dans le métabolisme de la Thiamine : THI2 et THI3. Ainsi, malgré les divergences de séquence et de fonction, YPR121c partage encore un profil d'expression comparable à ces deux paralogues. Les auteurs postulent donc que ce gène soit encore impliqué dans le métabolisme de la Thiamine.

B.2.2.2. Les divergences de protéine et de promoteur : le cas des globines.

Un second exemple de divergence entre des gènes dupliqués est celui de la famille des globines. Les globines sont des protéines fixant un hème lié à un atome de fer et fixant temporairement de l'oxygène. Les globines sont présentes dans la plupart des organismes (aucune globine n'ayant encore été documentée chez les Archées) (Hardison 1998). Chez les métazoaires, la famille des globines est divisée en deux grand types : (1) les hémoglobines, présentes dans les érythrocytes, chargées de transporter l'oxygène dans le sang et (2) les myoglobines, chargées du transport de l'oxygène dans les muscles, grands consommateurs d'énergie. Chez les plantes, il existe également deux types de globines : (1) les hémoglobines, transportant l'oxygène nécessaire à la respiration ou produit lors de la photosynthèse et (2) les leghémoglobines, présentes uniquement dans les nodules des racines de légumes, où la réduction de l'azote en ammonium consomme beaucoup d'énergie. Dans tous les autres organismes, ces protéines présentent des fonctions variables, comme par exemple la flavhémoglobine, fusion entre une globine et un domaine de liaison au FAD (Flavine Adénine Dinucléotide), intervenant dans les réponses aux stress oxydant chez *S. cerevisiae* ((Zhao *et al.* 1996)) et chez *E. coli* ((Vasudevan *et al.* 1991)).

L'origine des globines dans le monde vivant reste encore indéterminée, mais il est vraisemblable que ces protéines fassent partie d'une super famille regroupant les cytochromes et les autres hémoprotéines (Hardison 1998). Dans de nombreux organismes, l'évolution des globines a procédé par duplication et divergence. En effet, des évènements de duplication indépendants dans plusieurs phylums, comme les mollusques et les arthropodes, ont donné naissance à de nombreux gènes différents (Goodman *et al.* 1988), comme, par exemple, celui de la globine de *Ascaris lumbricoides* (un nématode) qui est formé par l'assemblage de deux gènes répétés en tandem (Goldberg 1995).

La famille des globines fut particulièrement étudiée au sein des vertébrés. Après l'apparition des gnathostomes (vertébrés à mâchoires), les globines se divisent en myoglobine, hémoglobine α et hémoglobine β (Goodman *et al.* 1975). L'holoprotéine d'hémoglobine est un tétramère formé de deux sous-unités α et deux sous-unités β. Seule la sous-unité α est capable de former un homotétramère et le gène codant pour cette sous-unité est donc pressenti comme le gène ancestral. Cette diversification entre les trois grands types de globines de gnathostomes illustre les changements fonctionnels des gènes dupliqués. Toutefois, la diversification des globines s'est poursuivie à travers l'évolution du phylum de gnathostomes. Par exemple, dans la plupart d'entre eux, les gènes codant pour les hémoglobines sont localisés à un seul locus chromosomique ; cependant, dans les génomes des oiseaux et des mammifères, ils sont répartis sur deux loci contenant respectivement les hémoglobines α ou β (Hardison 1998).

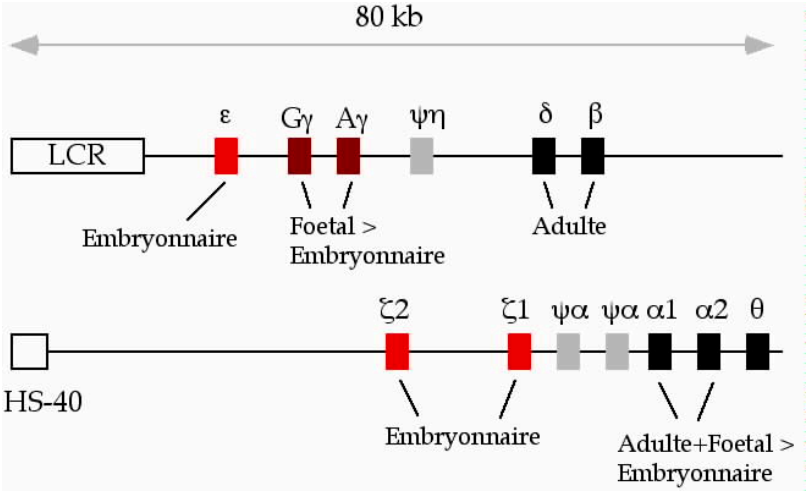

Figure 23 : Structure des locus des gènes des hémoglobines chez *H. sapiens* (d'après (Hardison 1998)).
Les régions des hémoglobines et avec les fonctions de chacun des gènes d'hémoglobine. LCR : Locus Control Region. : pseudogène.

Sur la figure 23, sont représentées la répartition et la fonction des gènes et des pseudogènes d'hémoglobines chez *H. sapiens*. Le locus des hémoglobines a été particulièrement étudié. Deux types de régions régulatrices ont été définie : (1) le LCR (Locus Control Region) situé en amont de tout le locus et (2) les séquences intergéniques qui contiennent les séquences régulatrices de chacun des gènes de ce locus (Hardison *et al.* 1997). Ce locus contient également les hémoglobines et , qui remplacent l'hémoglobine respectivement dans le sang embryonnaire et dans le sang fœtal.

Chez les vertébrés non-primates, la fonction de l'hémoglobine n'est pas la même. Elle est, dans ces organismes, exprimée comme l'hémoglobine (dans l'embryon). On peut donc inférer que l'hémoglobine ait acquis sa fonction fœtale chez les primates anthropoïdes (platyrrhiniens et catarrhiniens). La recherche d'un mécanisme pouvant expliquer ce changement fonctionnel de l'hémoglobine a mis en évidence que le promoteur de ce gène a subit un nombre de mutation élevé (Chiu *et al.* 1997; Gumucio *et al.* 1994). C'est donc vraisemblablement par un changement de promoteur que le gène a acquis sa nouvelle fonction.

Pour conclure, l'exemple de la famille des globines est riche en enseignements sur les changements possibles faisant suite à une duplication génique. Ainsi, plusieurs types de

divergences semblent pouvoir s'opérer après un évènement de duplication. Dans l'exemple des globines, sont présentés un changement radical de fonction dû à une longue période évolutive (illustré, par exemple, par les différences entre vertébrés et bactéries) et un changement dans le profil d'expression du gène (illustré par la globine chez les primates anthropoïdes).

B.2.2.3. Divergence au cœur d'un réseau fonctionnel : le système immunitaire.

Le troisième exemple choisi est celui de trois familles de gènes impliqués dans le système immunitaire des vertébrés. Ces familles sont les gènes codant pour le CMH (Complexe Majeur d'Histocompatibilité), les TCR (T-Cell Receptor) et les Ig (Immuno-globulines). Ces trois types de protéines sont trois acteurs principaux du fonctionnement du système immunitaire.

B.2.2.3.1. Les gènes du CMH.

Les gènes du CMH codent pour des glycoprotéines de surfaces qui présentent des antigènes aux lymphocytes T. Ces antigènes sont des peptides de petite taille (de 10 à 20 acides aminés) et proviennent de la dégradation des protéines par le protéasome. Ces protéines sont soit des protéines endogènes (du soi), soit des protéines exogènes, dont la présentation provoque une réaction immunitaire. Chez *H. sapiens*, les gènes codant pour le CMH sont de deux types : le CMH I[1] et le CMH II [1]. Le locus contenant les gènes codant pour le CMH est une région d'environ 4 Mb, portant les gènes du CMH I, ceux du CMH II et d'autres impliqués dans la réponse immunitaire (figure 25) (Kasahara *et al.* 1996) (Trowsdale 1993). Seul les gènes HLA-A, B et C sont impliqués dans la présentation antigénique du CMH I (et les gènes HLA-DP, DR, DQ pour le CMH II). Les autres paralogues codent pour d'autres fonctions ou sont des pseudogènes (Trowsdale 1993).

[1] les CMH de type I sont formés de deux chaînes protéiques. La première, nommée est constituée de trois domaines (1, 2 et 3) et est codée par un gène du locus du CMH ; la seconde, 2m, est une petite chaîne annexe et n'est pas codée par les loci du CMH (Hughes and Yeager, 1997). Ce sont les domaines 1 et 2 qui possèdent les sites de fixation des antigènes (Bjorkman et al. 1987).

L'étude du polymorphisme associé aux gènes des CMH I (Hughes and Nei 1988) montre que le taux d'hétérozygotie chez *H. sapiens* et chez *M. musculus* est supérieur à 80%. L'étude fine des substitutions au niveau de la séquence d'ADN montre que les site de fixation de l'antigène ont un taux de substitution non synonyme plus important que le reste de la protéine. Une étude similaire a également été réalisée sur les gènes du CMH II et a abouti à des conclusions similaires (Hughes and Nei 1989). De ces deux études, les auteurs ont proposé pour expliquer la propension à une forte hétérozygotie et une grande variabilité des site de fixtation de l'antigène, que les gènes du CMH soit soumis à des contraintes sélectives de type : « avantage de l'hétérozygote ».

[1] Les CMH II sont formés de deux chaînes α et β. Toutes deux sont formées de deux domaines, respectivement $\alpha 1$, $\alpha 2$, $\beta 1$ et $\beta 2$. Les sites de fixation sont localisés dans les domaines $\alpha 1$ et $\beta 1$ (Stern *et al*. 1994).

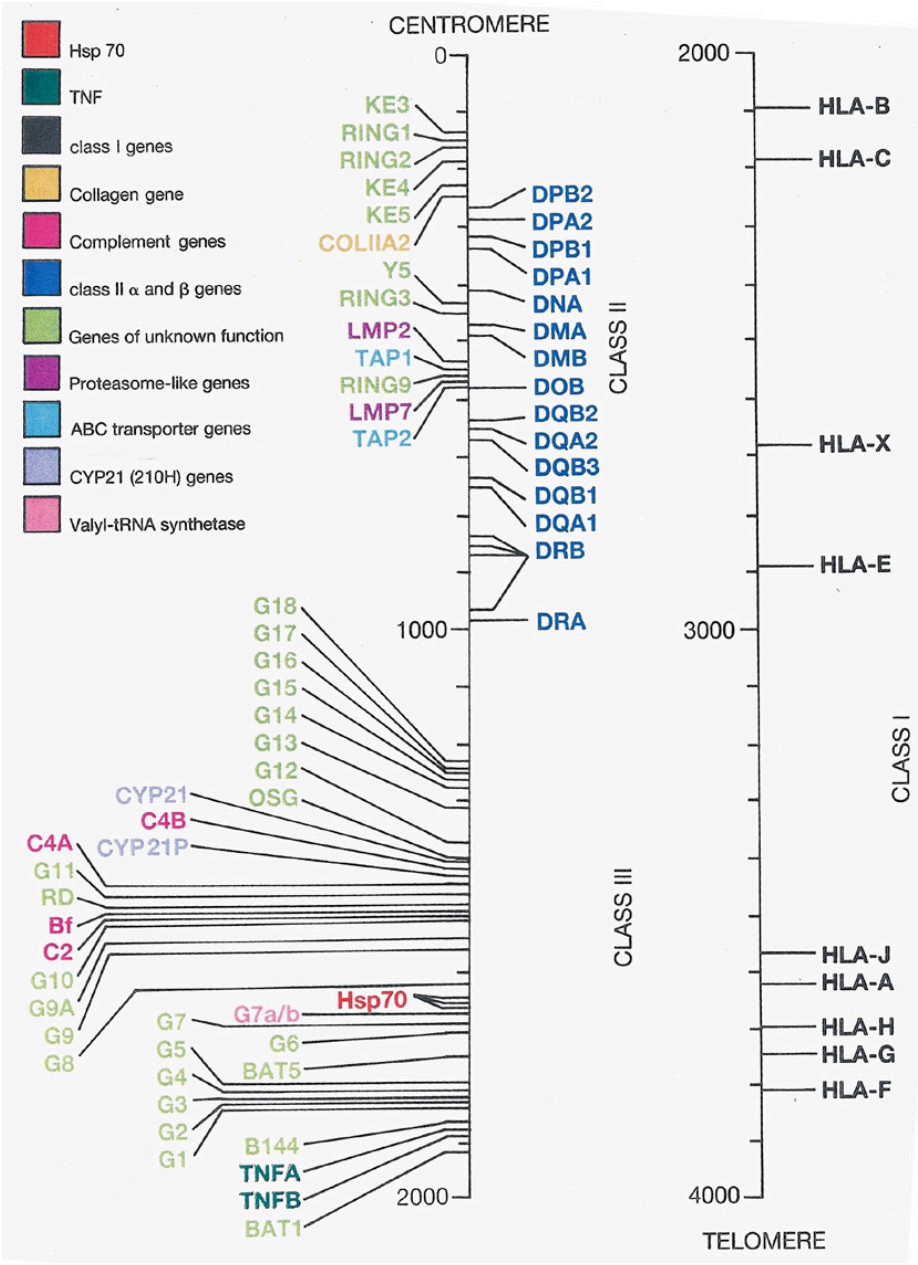

Figure 24 : Structure de la région chromosomique contenant les gènes des CMH chez *H. sapiens* (d'après (Trowsdale 1993)).

B.2.2.3.2. Les Ig et les TCR.

Les Ig sont des tétramères exprimés à la surface des lymphocytes B, composés de deux chaînes légères, codées par les loci et et deux chaînes lourdes codées par un seul locus. Les chaînes lourdes sont composées de trois domaines C et d'un domaine V, et les chaînes légères ne sont composées que d'un seul domaine C et d'un domaine V.

Les TCR sont des protéines dimériques[1] exprimés à la surface des lymphocytes T. Chaque monomère est constitué d'un domaine dit « constant » (domaine C) et d'un domaine dit « variable » (domaine V). Le domaine V présente un grand polymorphisme et, en particulier, trois régions dites « hypervariables » (Patten *et al.* 1984).

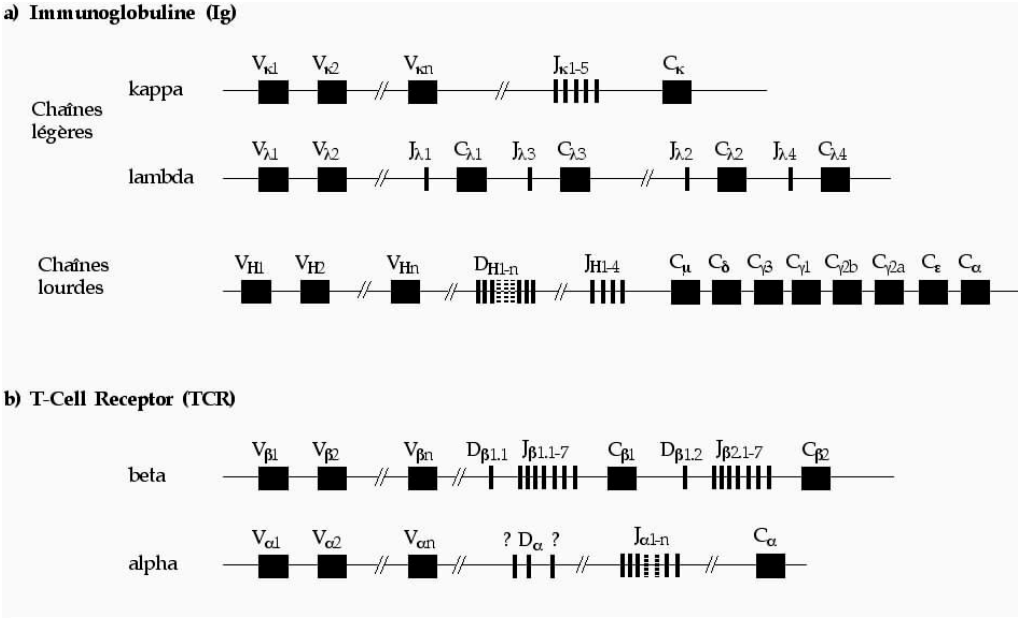

Figure 25 : Structure des régions chromosomiques contenant les gènes des Ig et des TCR (d'après (Hood *et al.* 1985)).
Les Ig sont formées de deux chaînes légères (codées par le locus kappa ou gamma) et de deux chaînes lourdes. Les TCR sont formés d'une chaîne β et d'une chaîne α (ou par une chaîne γ et une chaîne δ non présentées ici) .

Pour les TCR, comme pour les Ig, les chaînes V sont composées de plusieurs parties nommées V, (D) et J. L'assemblage de ces parties avec C, dans le génome des lymphocytes, forme la monomère complet[2]. Les organisations schématiques des loci d'Ig ou de TCR sont présentées sur la figure 25. L'examen des loci non réarrangés révèle l'importance du phénomène de duplication dans la richesse du répertoire immunitaire des vertébrés à mâchoires. A l'inverse du cas de l'involucrine, seule une copie de chaque motif est traduite en protéine : après le réarrangement une seule copie de chaque partie est exprimée.

[1] Les dimères sont constitués de α (Chien *et al.* 1984) et β (Hedrick *et al.* 1984) ou de γ (Saito *et al.* 1984) et δ (Elliot *et al.* 1988), bien que le second dimère ne soit présent que dans 5 à 10% des cellules T (Elliot *et al.* 1988).
[2] Ces assemblages ont lieu au niveau de l'ADN chromosomique par un mécanisme « recombinomorphe » qui raboute les parties V,(D),J et C indispensables à une protéine complète. Ce mécanisme est appelé recombinaison V(D)J (Agrawal, Eastman and Schatz 1998).

B.2.2.3.3. La super-famille des Immunoglobulines.

Comme nous l'avons vu, les répétitions jouent un rôle clef dans les fonctions des CMH, des TCR et des Ig. De plus, d'autres gènes du système immunitaire sont dupliqués, comme, notamment, les protéines du complément (Kasahara *et al.* 1996). Mais il y a plus ! Les gènes des CMH, des TCR et des Ig appartiennent vraisemblablement à une même famille de paralogues : la superfamille des immunoglobulines (Hood *et al.* 1985). En effet, les domaines C des TCR et des Ig sont homologues aux domaines α2 et β2 du CMH I ainsi qu'aux domaines α3 et β2m du CMH II. Les autres domaines du CMH (présentateurs des antigènes) sont homologues entre eux, mais ne partagent pas d'homologie visible avec cette super-famille. Un modèle de la phylogénie des gènes du système immunitaire est présenté sur la figure 26. Il semble donc que le système immunitaire soit né d'un seul gène ancestral qui par duplication et divergence a aboutit à différentes copies qui ont les fonctions complexes et spécialisées connues aujourd'hui pour le système immunitaire.

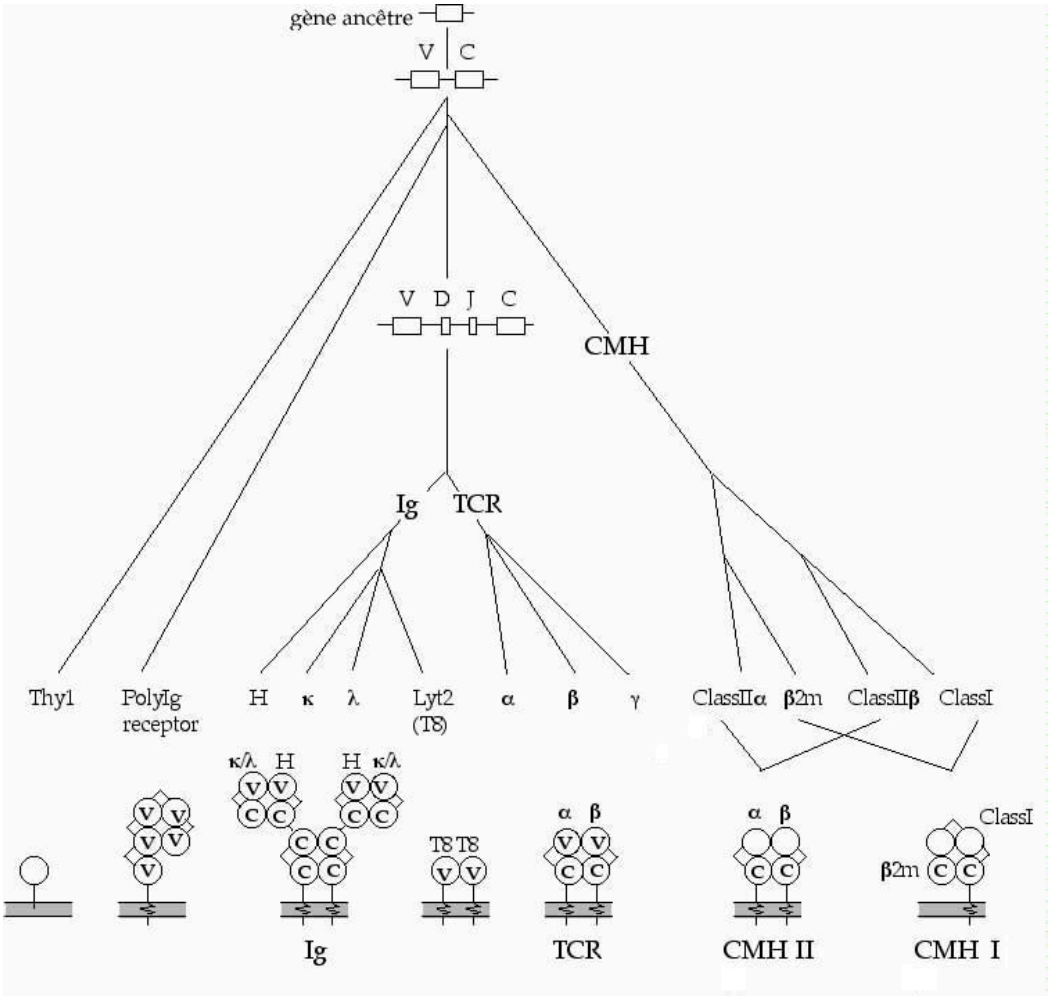

Figure 26 : Phylogénie hypothétique des acteurs majeurs du système immunitaire (d'après (Hood *et al.* 1985)).
Dans cette hypothèse, les protéines majeures du système immunitaire (Ig, TCR et CMH) dériveraient d'un seul gène ancestral commun. L'assemblage de ses hypothétiques domaines ancestraux sont présentés en dessous de cette phylogénie.

En conclusion de cette partie, nous avons vu que les duplications de gènes pouvaient avoir des conséquences majeures à long terme dans l'émergence de nouvelles fonctions. Les trois exemples choisis illustrent trois niveaux de relation entre « répétition de gènes » et « redondance de fonctions ». Des trois exemples, le premier illustre un début de divergence, le second une divergence plus grande et le troisième laisse entrevoir à travers la mise en place du système immunitaire, comment des gènes répétés peuvent peu à peu occuper des fonctions divergentes.

B.2.3. Des modèles théoriques simulant l'évolution des répétitions.

B.2.3.1.1. Modèle de duplication, relaxation et divergence.

De ce qui précède, il faut être convaincu que les duplications de gènes sont monnaie courante dans l'histoire évolutive des génomes. C'est ainsi que S. Ohno, releva dès 1970 le rôle clef des duplications dans l'évolution des génomes (Ohno 1970). Dans son raisonnement, la plupart des nouveaux gènes sont créés par duplication, et rarement par genèse *de novo*. L'évolution des gènes est souvent contrainte par les fonctions qu'exercent la protéine codée, et donc la plupart des mutations sont « interdites ». *A contrario*, lorsqu'une duplication de gène se produit, seule une des deux copies suffit à assurer les fonctions initiales et la seconde copie est « libre » d'évoluer vers une nouvelle fonction.

Dans une étude récente, il a été montré que suite à la duplication d'un gène, la divergence entre les copies s'établie sans grande contrainte sélective (Lynch and Conery 2000). La majorité des gènes dupliqués sont perdus mais ceux qui sont conservés adoptent une nouvelle fonction et leur évolution devient contrainte à nouveau.

Dans cette idée de relaxation des contraintes sélectives, JB Walsh a tenté d'évaluer les probabilités relatives pour un gène dupliqué d'évoluer vers une nouvelle fonction ou vers une perte de fonction (de devenir un pseudogène) (Walsh 1995). Le modèle postule qu'après duplication, les deux copies accumulent des mutations délétères, neutres ou avantageuses (la fraction de mutations avantageuses étant très inférieure à 1). Le modèle néglige les délétions et les fontionnalisations de pseudogènes (retour vers un gène fonctionnel). Dans ces conditions, un gène dupliqué a une probabilité minime d'évoluer vers une nouvelle fonction, sauf si la population d'allèles est très grande.

Dans une autre d'analyse, le même auteur étudie la probabilité, pour un gène qui vient d'être dupliqué, de pouvoir échapper à la conversion (qui empêche la divergence) (Walsh 1987). Dans ce modèle, la conversion est impossible en dessous d'un certain seuil d'identité entre les copies. Ce modèle montre que si le rapport taux de mutation/taux de conversion est supérieur à 0,1, la probabilité d'échapper à la conversion est forte. Ainsi, les

variations du taux de conversion en fonction de la localisation des répétitions sont très importantes pour le devenir d'un gène dupliqué.

Une autre série d'étude, menée par T. Ohta, a analysé la mise en place des familles multigéniques dans un génome (Ohta 1987a; Ohta 1987b; Ohta 1988; Ohta 1991). L'auteur postule que la sélection défavorise un individu qui possède un nombre de copies de gènes inférieur à la moyenne. Dans ces modèles, les duplications sont créées par crossing-over inégaux et les mutations avantageuses apparaissent dix fois moins souvent que les autres. Des simulations informatiques font ressortir plusieurs points importants :

- La formation d'une famille multigénique fonctionnelle requiert l'action d'une sélection positive (Ohta 1987b).
- L'émergence de nouveaux gènes (divergents de l'original) est accompagnée d'un « fardeau génétique », mélange de gènes redondants (identiques à l'original) et de pseudogènes (Ohta 1987a).
- Comparées aux populations haploïdes, les populations diploïdes forment plus de nouveaux gènes fonctionnels, mais ont une charge génétique plus importante (Ohta 1988).

B.2.3.1.2. Modèle de duplication, divergence et complémentation.

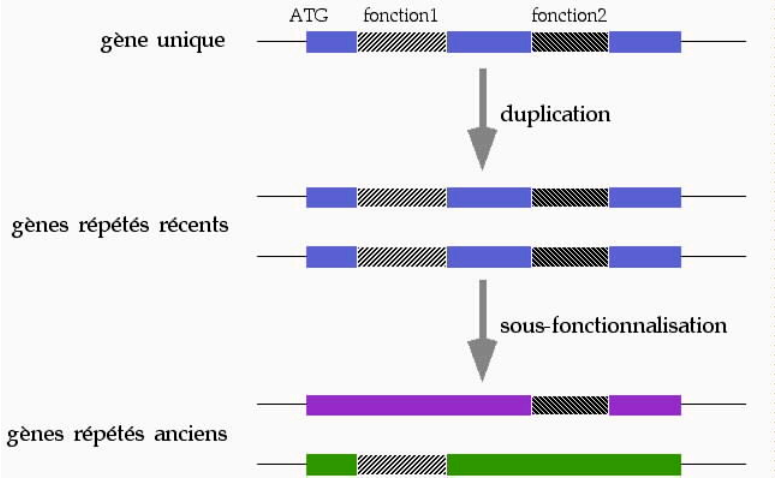

Figure 27 : Evolution des gènes dupliqués par sous-fonctionnalisation.
Ce modèle propose que tous les gènes ont des fonctions multiples (dans ce schéma deux fonctions). Après la duplication chacune des copies possède les deux fonctions. La sous-fonctionnalisation est la perte asymétrique des fonctions dans chacune des copies.

D'autres auteurs ont remis en cause le modèle proposé par Ohno et ont eux-mêmes proposé une hypothèse alternative pour expliquer le devenir de gènes nouvellement dupliqués : la spécialisation (ou sous-fonctionnalisation) (figure 27). Ce modèle postule que les gènes ont de multiples fonctions. Ainsi, lorsqu'un gène est dupliqué, aucune des deux copies n'évolue réellement vers une nouvelle fonction, mais chacune des copies se spécialise dans une des fonctions originelles. Par exemple, un gène ubiquitaire se duplique et chaque copie devient tissu-spécifique en conservant la fonction originelle. Cette idée fut proposée par AL Hughes (Hughes 1994), qui s'appuya sur des exemples biologiques, comme notamment celui de l'hémoglobine décrit ci-dessus. Cette idée fut ensuite reprise et théorisée dans une étude ultérieure, sous le nom de DDC (*Duplication, Degeneration and Complementation*) (Force *et al.* 1999). Dans ce modèle, la conservation par DDC des deux copies d'un gène s'établit lorsque chacune des copies a perdu l'une des fonctions du gène original. Chaque copie est donc « indispensable » à l'organisme. Les auteurs considèrent un modèle théorique dans lequel un gène est récemment dupliqué et les deux copies ne peuvent subir aucune conversion, délétion ou nouvelle duplication. Chaque copie est divisée en sous-parties (représentant les sous-fonctions) qui peuvent accumuler des mutations avantageuses ou délétères. Dans ces conditions, la probabilité de sous-fonctionnalisation est positivement corrélée au nombre de sous-fonctions et le temps nécessaire à la réalisation de cette sous-fonctionnalisation est négativement corrélé à ce nombre (Force *et al.* 1999). Dans une seconde étude, les auteurs montrent que la probabilité de sous-fonctionnalisation devient faible si la population effective est très grande ($>10^4$) (Lynch and Force 2000).

Pour conclure sur les répétitions de gènes, il apparaît que les répétitions ont principalement deux types de conséquences : (1) des conséquences « immédiates », où la quantité de protéine exprimée par un gène répété est accrue et peut changer le phénotype d'un organisme et (2) des conséquences « à long terme », où chaque copie du gène répété est susceptible d'évoluer vers une nouvelle fonction ou vers une spécialisation de la fonction originale.

B.3. Hyperploïdie et Polyploïdie.

Comme nous l'avons évoqué dans la première partie, les duplications de chromosomes entiers, liées aux accidents de division cellulaire, sont assez fréquentes. A l'encontre de ce que l'intuition première laisse penser, il semble en observant les exemples dans la nature que la duplication de génomes entiers (polyploïdies) soit plus viable à long terme que les duplications d'un seul chromosome (hyperploïdies).

En effet, les exemples d'hyperploïdies documentés correspondent souvent à des phénotypes « pathologiques ». Les trisomies chez *H. sapiens* sont pour la plupart associées à une mort embryonnaire, et ont pour les autres des conséquences pathologiques très fortes sur l'organisme (Suzuki *et al.* 1989). Par ailleurs, si les trisomies semblent être viables chez *S. cerevisiae*, et apporter des avantages de croissance chez certains mutants (Hughes *et al.* 2000), elles sont généralement associées à des défauts de sporulation (Johnston *et al.* 2000), ce qui doit expliquer leur rareté dans les populations naturelles (Mortimer *et al.* 1994).

A l'inverse, dans un certain nombre de phylums, les polyploïdes ne semblent pas générer des organismes infertiles. Chez les plantes, les xénopes, les poissons téléostéens et les levures, de nombreux cas d'organismes polyploïdes ont été décrits. Une explication a été proposée (Shimeld 1999) : les gènes sont presque toujours impliqués dans un réseau de gènes qui requiert une coopération de ceux-ci (voies de métabolisme, voies de signalisation, enzyme multimérique, etc.).

La duplication d'une partie seule d'un réseau pourrait perturber son fonctionnement alors que la duplication de tout le réseau (comme dans le cas des polyploïdes) ne le perturberait que peu ou pas. Une des conséquences de la polyploïdisation est la production d'organismes souvent de plus grosse taille.

Chez *H. sapiens*, les tétraploïdes et les triploïdes n'aboutissent pas au développement d'un organisme viable (Suzuki *et al.* 1989). Cependant, de nombreux exemples de polyploïdes sont décrits dans les organismes eucaryotes :

Chez les plantes, ce phénomène est extrêmement répandu. Une étude sur la taille des cellules semble montrer que 70% des angiospermes sont de nouveaux ou d'anciens

polyploïdes (Masterson 1994). Des exemples de plantes polyploïdes bien connus sont *Z. mays* (le maïs - tétraploïde), celui de *Triticum aestivum*, (le blé - hexaploïde) et celui de la famille des *Brassica* (chou, navet, moutarde, etc. - polyploïdies variées) (Suzuki *et al.* 1989).

Chez les xénopes, la polyploïdisation est également un phénomène courant. *Xenopus laevis laevis* posséde 36 chromosomes (voir première partie), mais il existe des espèces de xénope possédant 20, 36, 72 et 108 chromosomes. Sur la base de l'éloignement phylogénétique entre *Xenopus tropicalis*, qui ne possède que 20 chromosomes, et les autres xénopes (avec un nombre supérieur de chromosomes), certains auteurs ont proposé que le diploïde originel n'avait que 20 chromosomes et que les xénopes à 36 chromosomes sont le résultat d'une ancienne poyploïdisation (Bisbee *et al.* 1977).

Chez les levures, des tétraploïdes peuvent être aisément obtenus par diverses techniques, dont la culture sous haute pression (Hamada *et al.* 1992). Par ailleurs, une étude de la ségrégation méiotique des levures tétraploïdes montre que ces dernières ont un taux de ségrégation correct similaire à celui des levures diploïdes (Loidl 1995). De telles levures semblent conservée la polyploïdie au cours des générations. De façon remarquable, J Loidl rapporte que des levures triploïdes sont viables, mais présentent un défaut de division méiotique fort, dû aux nombreux aneuploïdes générés (Loidl 1995). Chez *S. cerevisiae*, les triploïdes ne semblent pas pouvoir former une population importante et viable à long terme.

Il est supposé que, parmi les poissons téléostéens, plusieurs phylums sont d'anciens polyploïdes. C'est le cas notamment de la famille des *Catostomides* (carpes) (Uyeno and Smith 1972) et de celle des *Salmonoïdes* (saumons) (Ohno 1970). L'évènement de polyploïdisation pour les *Catostomides* date probablement de 50 millions d'années (Uyeno and Smith 1972) et celui des *Salmonoïdes* est du même ordre de grandeur (Schmidtke and Kandt 1981). Ainsi, ces poissons constituent d'excellents modèles pour analyser la « diploïdisation » d'anciens tétraploïdes. La diploïdisation est définie comme le retour à un état diploïde à partir d'un état tétraploïde. L'analyse des caryotypes de plusieurs espèces de *Salmonoïdes* a montré une grande variabilité dans leur nombre de chromosomes (voire dans leur nombre de bras chromosomiques) laissant imaginer que ces remaniements sont le reflet du processus de

diploïdisation (Hartley and Horne 1984). En outre, une autre conséquence de ce processus est la perte de nombreuses répétitions de gènes. L'analyse des polymorphismes protéiques, chez les *Catostomides*, montre que la fréquence de gènes répétés varie entre 35% et 65% (Ferris and Whitt 1977). Il a par ailleurs été montré, chez un *Salmonoïde* que la perte d'une des copies d'un gène était accompagnée d'une augmentation du taux d'hétérozygotie (Allendorf 1978). Ainsi, on peut penser que la duplication permet l'existence de quatre « allèles » au sein d'un même organisme et que la perte de cette variabilité est compensée par l'augmentation du taux d'hétérozygotie.

Ainsi, il semble clair que contrairement aux hyperploïdies et aux polyploïdes impairs (comme les triploïdes), les polyploïdies sont courantes dans l'histoire évolutive des organismes du règne eucaryote.

Dans les chapitres précédents, deux des trois questions posées dans l'avant-propos ont été étudiées. La première visait à déterminer les contraintes structurales que subissent les génomes, en particulier les séquences répétées. Ceci fut abordé à travers les mécanismes moléculaires liés aux répétitions (ceux qui les créent et ceux qui les ciblent). La seconde question était de déterminer les contraintes sélectives s'exerçant sur les séquences répétées dans les génomes. Dû à la difficulté de formalisation de ces contraintes, quelques exemples éclairant les conséquences des séquences répétées ont été choisis.

C. Dynamique des répétitions.

Le troisième et dernier chapitre de cette introduction est dévolu à replacer les répétitions dans la dynamique des génomes, à analyser quelle « liberté » est laissée aux répétitions par les pressions structurales et sélectives. Le travail exposé dans la partie « Résultats » abordant plus particulièrement cette vue évolutive des répétitions, les deux axes qui composent ce chapitre mettent l'accent sur deux points discutés dans cette thèse, à savoir les « relations entre répétitions et remaniements » et la « transformation des répétitions ».

C.1. Relations entre répétitions et remaniements.

Dans le second chapitre sont exposés quelques exemples de conséquences d'évènements de duplication. Cependant, les conséquences de la présence des répétitions sur la stabilité des génomes ne sont pas abordées. Les répétitions peuvent être la cible de la recombinaison homologue et de la recombinaison non-homologue. Si la recombinaison a lieu entre deux séquences répétées dispersées, cela entraîne des remaniements chromosomiques, tels que des inversions, délétions, duplications et des translocations.

De nombreux remaniements ont eu lieu dans l'histoire évolutive des génomes de primates et plus généralement des mammifères (Dutrillaux 1997; Dutrillaux and Richard 1997). L'importance de ces évènements dans les processus comme la spéciation semble aujourd'hui acquise, bien qu'il serait réductionniste d'associer simplement spéciation et remaniement. Cependant, les remaniements chromosomiques observés chez *H. sapiens* sont souvent associés à des pathologies ((Ji *et al.* 2000; Lupski 1998; Mazzarella and Schlessinger 1997)). Il est donc séduisant de penser que des mécanismes comme la méthylation (voir le premier chapitre) se sont mis en place pour abaisser le taux de recombinaison entre les séquences répétées, permettant une stabilisation des chromosomes. Ce type de mécanismes pourrait expliquer que la recombinaison méiotique allélique chez *H. sapiens* est 1000 fois plus faible que chez *S. cerevisiae*. En effet, 1 cM correspond à quelques kilobases chez *S. cerevisiae* (Baudat and Nicolas 1997) et à quelques Mb chez *H. sapiens* (Dunham *et al.* 1999).

Chez *S. cerevisiae*, organisme plus « tolérant » aux remaniements chromosomiques plusieurs études ont été effectuées. Ainsi, il fut observé que les translocations ont souvent pour origine des recombinaisons entre des séquences répétées (gène, Ty ou autres) dispersées dans le génome (Jinks-Robertson and Petes 1986; Loidl and Nairz 1997). L'analyse des translocations chromosomiques dans diverses espèces de levures appartenant au genre *Saccharomyces sensu stricto* (Fischer *et al.* 2000) a révélé que le nombre de remaniements n'est pas corrélé à l'éloignement phylogénétique des espèces. On observe entre certaines espèces un nombre de remaniements important associé à une faible distance phylogénétique (et vice-versa). Ainsi la densité de remaniements n'est pas un bon marqueur de la distance évolutive

entre les espèces (au moins chez les levures). Par ailleurs, la comparaison de longs fragments de séquences de *S. cerevisiae* et de *Saccharomyces bayanus* (*Saccharomyces uvarum*), a permis de montrer que la plupart des ruptures de synténie ne sont pas dues à des évènements de remaniements mais à l'accumulation de mutations ponctuelles (Fischer *et al.* 2001). Ainsi, chez les levures, les remaniements chromosomiques ne sont pas si fréquents que l'on aurait pu le penser.

Chez les bactéries, certaines répétitions sont également à l'origine de remaniements chromosomiques importants. Un des exemples les mieux connus est certainement celui de l'inversion d'un énorme fragment chromosomique ayant eu lieu après la séparation de *E. coli* et *Salmonella thyphimurium* (Hughes 1999). Cette inversion n'a pas changé les places respectives de ORI et TER, ce qui explique qu'elle n'ait pas bouleversé l'organisation générale du chromosome. Par ailleurs, la stabilité des chromosomes a été négativement corrélée à la densité de répétitions, montrant que plus le génome est répété, moins il est « stable » (Rocha *et al.* 1999b).

C.2. La transformation des répétitions : des polyploïdes aux répétitions segmentaires.

Les évènements de polyploïdie ne sont pas rares dans l'histoire des organismes eucaryotes. Néanmoins, s'il est aisé de les repérer lorsqu'ils sont récents, il l'est moins lorsque l'évènement est ancien. En effet, comme nous l'avons vu au second chapitre, la polyploïdie est généralement suivie d'une diploïdisation, un retour à l'état diploïde. Cette diploïdisation s'accompagne de nombreux remaniements chromosomiques menant à la perte des répétitions de chromosomes. Néanmoins, les remaniements n'effacent pas totalement les traces de la polyploïdie et des modèles proposent que, si les chromosomes sont principalement remaniés par des recombinaisons, les répétitions de chromosomes se transforment en duplications segmentaires (Nadeau and Sankoff 1997; Seoighe and Wolfe 1998).

Ainsi les répétitions segmentaires observées dans les génomes peuvent être la trace d'anciens évènements de polyploïdie, mais comment trancher entre duplications de

segments et duplication de génome ? C'est ce sujet brûlant qui agite une part de la communauté scientifique : peut-on inférer des polyploïdies parce que l'on observe des répétitions segmentaires ? Ce type de question a été débattu au sujet des génomes de *S. cerevisiae* (Wolfe and Shields 1997) et de *A. thaliana* (Blanc *et al.* 2000) . Deux évènements de polyploïdisation se seraient produits dans le génome d'un ancêtre des vertébrés (l'hypothèse des 2R – 2 Rounds of polyploïdisation) (Ohno 1970) et un troisième dans le génome d'un ancêtre des poissons téléostéens (l'hypothèse des 3R– 3 Rounds of polyploïdisation) (Amores *et al.* 1998) . Par la suite, seront discutés les arguments en faveur et en défaveur d'une ancienne polyploïdie pour les vertébrés et pour *S. cerevisiae*.

C.2.1. *S. cerevisiae* est- il un polyploïde ancestral ?

La séquence du génome de *S. cerevisiae* (Goffeau and authors) 1997) a révélé une cinquantaine de répétitions segmentaires couvrant près de la moitié du génome (voir le premier chapitre). KH Wolfe et DC Shields ont proposé que ces répétitions géantes soient la trace d'une ancienne polyploïdie (Wolfe and Shields 1997). Il est bien sûr difficile de se prononcer sur la pertinence de cette proposition, cependant l'examen des arguments en faveur et en défaveur de cette hypothèse nous renseigne sur la complexité de la question.

Les arguments en faveur de l'ancienne polyploïdie sont :

- les répétitions segmentaires n'existent qu'en deux copies non chevauchantes (Wolfe and Shields 1997).
- les deux copies des répétitions sont colocalisées et coorientées par rapport aux télomères (Coissac *et al.* 1997). Cette agencement est attendu si les remaniements ayant suivi la polyploïdisation sont principalement des crossing-over.
- Les tétraploïdes de *S. cerevisiae* sont parfaitement viables et sporulent correctement (Loidl 1995).
- Le chromosome I de *Ashbya gossypii*, ascomycète ayant un petit génome (et supposé non tétraploïde) présente une synténie entremêlée avec les régions de

deux chromosomes de *S. cerevisiae* (Dietrich *et al.* 1999). Ce chromosome serait une forme ancêtre de ces deux régions.

Les arguments en défaveur de l'ancienne polyploïdie sont :

- Comme la plupart des organismes (et en particulier les bactéries), *S. cerevisiae* montre une densité de gènes répétés ne dépendant que du nombre de gènes (Coissac *et al.* 1997). Ceci montre que *S. cerevisiae* ne semble pas plus répété que d'autres génomes.
- *Kluyveromyces lactis*, pressenti comme un génome n'ayant pas subi de tétraploïdisation (Keogh *et al.* 1998) possède un nombre de gènes répétés (estimé) comparable à celui de *S. cerevisiae* (Llorente *et al.* 2000).
- Certaines nouvelles répétitions segmentaires sont observées dans les souches de laboratoire (Bach *et al.* 1995).
- L'estimation de l'âge relatif de ces répétitions segmentaires montre que trois d'entre elles se ressemblent plus que les autres (Friedman and Hughes 2001a), ce qui ne serait pas attendu si toutes les répétitions avaient été créées en même temps.

Le mystère sur cette question reste donc encore entier ! Malgré les réponses qu'apporte l'hypothèse d'une ancienne polyploïdisation, certains éléments restent encore en contradiction avec ce scénario. Des résultats récents (Wolfe 2002) montrent que les fragments de séquences des levures étudiées par le programme Génolevure (Souciet *et al.* 2000) partagent des synténies avec, en général, deux régions du génome de *S. cerevisiae*. Ces régions synténiques en deux exemplaires ne sont pas chevauchantes et couvrent 82% du génome de *S. cerevisiae*. Cela renforce très fortement l'hypothèse selon laquelle *S. cerevisiae* est un « paléopoplyploïde ».

C.2.2. L'hypothèse des 2R : Les vertébrés dérivent-t-ils de deux « Round » de polyploïdie ?

S. Ohno en 1970 propose pour expliquer la présence de plusieurs groupes de gènes (région chromosomique) en quatre exemplaires dans les génomes des vertébrés que ces derniers résultent de deux polyploïdisations qui se serait succédées rapidement (Ohno 1970). Cette hypothèse sera adoptée par de nombreux biologistes (pour revue voir (Skrabanek and Wolfe 1998)). L'argument principal qui plaide en faveur d'une telle hypothèse est la présence de quelques régions en quatre exemplaires dans les génomes des vertébrés : les régions comprenant les gènes du CMH (Kasahara *et al.* 1997), celle comprenant les gènes Hox (Bailey *et al.* 1997) et celle comprenant les FGFR (Fibroblast Growth Factor Receptor) (Pebusque *et al.* 1998). Aujourd'hui, la plupart des auteurs s'accordent à penser que cette hypothèse est valide et seuls quelques scientifiques, tel que AL Hughes, s'efforcent de la démanteler. Bien sûr, l'accord tacite des scientifiques ne transforme pas cette hypothèse en fait. Les deux principaux arguments en défaveur de cette hypothèse des 2R sont (Friedman and Hughes 2001b; Hughes 1998; Hughes *et al.* 2001) les suivants :

- Le premier concerne la quantité de gènes répétés observée dans les génomes des vertébrés. En effet, les vertébrés possèdent une proportion de gènes répétés identique à celle des organismes invertébrés. Un excédent aurait été attendu si les vertébrés étaient issus des 2R (deux polyploïdisations).
- Le second argument concerne la topologie des répétitions. Les gènes issus des 2R devraient être en quatre copies et présenter deux paires de gènes ayant des similarité plus fortes (topologie [AB][CD], regroupant A-B avec C-D). Or, si l'on sélectionne les gènes présents à quatre copies chez les vertébrés et à une copie chez les invertébrés, la topologie [A[B[CD]]], regroupant C-D avec B puis avec A est trop souvent retrouvée.

Il est donc très difficile de répondre avec certitude à ce problème pour le moins épineux. La problématique posée par les questions de polyploïdisations des génomes de vertébrés a été très élégamment écrite : « Take four, or maybe eight, decks of 52 playing

cards. Shuffle them all together and then throw some cards away. Pick 20 cards at random and drop the rest on the floor. Give 20 cards to some evolutionary biologists and ask them to figure out what you've done. For encouragement, tell them they can have the cards on the floor in 2005.» (Skrabanek and Wolfe 1998).

L'année 2001 a vu aboutir une grande partie de la séquence du génome humain (TIHGSC 2001). De cette séquence attendue par tous (scientifiques et non scientifiques), on aurait pu espérer entrevoir quelques pistes concernant la question des 2R. Les conclusions de l'analyse des gènes répétés chez *H. sapiens* montrent qu'avec les données du génome humain, il est impossible de répondre à la question des 2R, car même si les évènements avaient eu lieu, les traces serait trop effacées (TIHGSC 2001).

Pour conclure sur ces hypothèses concernant les polyploïdisations, il faut également mentionner que les propositions d'ancêtres polyploïdes pour *A. thaliana* (Vision *et al.* 2000) et pour les poissons téléostéens (Robinson-Rechavi *et al.* 2001) font également l'objet de controverses. Il faut donc considérer que ces questions restent pour le moment souvent sans réponse, en espérant que certains indices, encore manquants, pourront nous donner de plus amples informations. Néanmoins, l'idée clef que nous retiendrons de ces hypothèses est que les répétitions que nous observons aujourd'hui dérivent potentiellement de répétitions très différentes. Des répétitions classées aujourd'hui comme segmentaires sont tout à fait susceptibles d'être des reliquats de duplications de génomes et peuvent être des précurseurs de répétitions dispersées.

II. Matériel & Méthodes

Le matériel et les méthodes utilisés au cours de ma thèse étant principalement décrits dans les articles, je ne ferai pas un exposé détaillé sur ce sujet, mais j'expliquerai plutôt les raisons des choix que nous avons été amenés à faire. Après une partie succincte sur le matériel, j'expliquerai les différentes stratégies que nous avons adoptées pour la détection des répétitions.

A. Le matériel.

Les données concernant les génomes étudiés proviennent pour la plupart du site ftp de GenBank (ftp.ncbi.nlm.nih.gov) dans le répertoire contenant les génomes (/genbank/genomes). Les données de deux génomes proviennent d'autres sites :

- le génome de *S. cerevisiae*, qui provient du site ftp de SGD (*Saccharomyces Genome Database*) à l'adresse : genome-ftp.stanford.edu dans le répertoire /pub/yeast/genome_seq
- le génome de l'Archée *Pyrococcus furiosus*, accessible à partir de la page http://www.genome.utah.edu/sequence.html

Les calculs ont été réalisés sur plusieurs ordinateurs, mais les gros calculs ont été réalisées sur une station de travail Silicon, ayant deux processeurs R10000 et dotée de 1,2 Go de mémoire vive. La plupart des analyses ont été réalisées sur un PC-linux, équipé d'un processeur AMD-Athlon cadencé à 750 MHz et doté de 264 Mo de mémoire vive.

B. Les méthodes.

Idéalement, la recherche de répétitions dans une séquence d'ADN s'effectue en alignant ces séquences sur elles même par des sous-alignements optimaux. Après chaque alignement, la séquence jugée « répétée » est marquée et ne peut plus être redétectée comme la même répétition. Au fur et à mesure les meilleurs sous-alignements sont déterminés et les répétitions caractérisées. Cette méthode est adaptée pour des petites séquences, mais ne peut pas être appliquée à des séquences très longues, car le coût de temps de calcul et de mémoire est trop important.

Pour rechercher des répétitions dans des chromosomes, nous avons mis en place une méthode heuristique[1]. Comme la plupart des méthodes de recherche de répétitions non-exactes dans des séquences de grande taille (Leung *et al.* 1991) (Vincens *et al.* 1998), notre méthode est fondée sur la détection de répétitions strictes (graines), qui sont étendues par alignement local par programmation dynamique (Smith and Waterman 1981).

L'étape « limitante » de notre méthode est le temps nécessaire à la réalisation des nombreux alignements locaux. En effet, le nombre d'alignements est égal au nombre de graines conservées. Donc, plus le nombre de graines est élevé plus le temps de calcul total est important. A titre d'exemple, la détection des répétitions dans le génome de *E. coli* est de l'ordre de quelques heures. De ces quelques heures, seules quelques minutes sont nécessaires à la détection des graines.

Le problème posé par ce type de méthodologie est de déterminer la limite à partir de laquelle une répétition est significative. Aucune statistique exacte ne permettant de calculer la probabilité d'observer une répétition non-stricte dans une séquence donnée, nous avons choisi deux stratégies pour conserver les répétitions significatives uniquement : a) la méthode « eucaryote » et b) la méthode « bactérienne ».

[1] 1. heuristique : méthode permettant d'accélérer la recherche mais ne garantissant pas de conserver le meilleur résultat.

Parti pris (biais) qui permet d'élaguer un espace de recherche. L'élagage étant nécessaire pour rendre cette recherche possible. Plus une heuristique réduit l'espace de recherche (tout en conservant dans cet espace la solution du problème), plus l'heuristique est bonne.

(d'après P. Brezellec)

B.1.1. Nos deux stratégies pour détecter les répétitions.

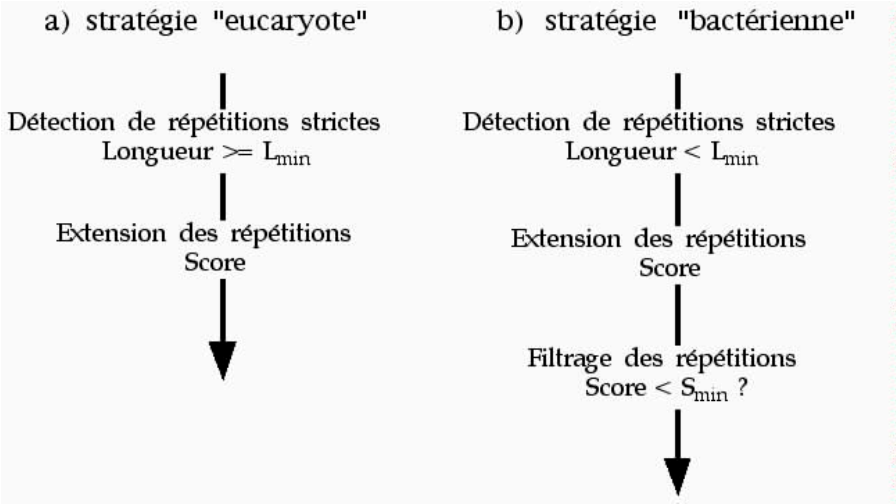

Figure 28 : Les deux stratégies utilisées pour détecter les répétitions.

Dans la méthode « eucaryote » (figure 28), nous ne détectons que des graines significatives. Toutes les graines sont donc étendues en répétitions et aucun filtrage n'est nécessaire. Pour déterminer un seuil minimum correspondant aux graines significatives, nous avons utilisé une formule qui calcule une longueur minimum (L_{min}). Cette dernière correspond à une probabilité de 1/1000 de trouver une répétition plus grande ou égale à L_{min} dans une séquence de même taille et de même composition en mononucléotides (Karlin and Ost 1985). Cette L_{min} varie entre 20 et 25 paires de bases pour les chromosomes bactériens et entre 21 et 28 paires de bases pour les chromosomes eucaryotes. La stratégie « eucaryote » minimise le nombre de graines à aligner et permet donc de détecter des répétitions dans des chromosomes aussi grands et répétés que les chromosomes 21 et 22 de *H. sapiens*.

Dans la méthode « bactérienne » (figure 28), nous cherchons à étendre même des graines non significatives (dont la longueur est inférieure à L_{min}). Ce choix est motivé par l'idée qu'il existe un grand nombre de répétitions significatives qui ne possèdent pas de graines significatives. Toutes les graines sont étendues, mais les répétitions doivent être filtrées afin de retirer celles qui ne sont pas significatives. La stratégie « bactérienne » recherche le maximum de répétitions significatives. Elle est donc bien adaptée aux petits chromosomes peu répétés (comme ceux de *S. cerevisiae* ou ceux des Bactéries et des Archées).

Pour détecter les répétitions dans le génome de *S. cerevisiae* (premier article), nous avons choisi d'utiliser la stratégie « bactérienne ». Ce choix a été motivé par la petite taille des chromosomes de *S. cerevisiae* (tous les chromosomes ont une taille inférieure à 1,5 Mb). Nous voulions donc conserver un maximum de répétitions significatives. Les détails de la méthode sont les suivants. Nous avons conservé des graines dont la taille est supérieure à 15-17 paires de bases (L_{min} étant entre 21 et 24 bases). Après l'extension, les répétitions ayant une longueur inférieure à 30 paires de bases ou ayant une identité inférieure à 50% ont été retirées. Nous avons également retiré les répétitions de basse complexité (SSR) en calculant l'entropie de chaque répétition (Schneider *et al.* 1986), ainsi que certaines répétitions connues pour avoir une dynamique spéciale (transposons, ARNr, ARNt, et répétitions des régions subtélomériques).

Pour détecter les répétitions dans les chromosomes eucaryotes (second article), nous avons utilisé la stratégie « eucaryote ». En effet, si la stratégie « bactérienne » aurait pu être utilisée pour les petits chromosomes (ceux de *S. cerevisiae* ou ceux de *P. falciparum*), elle n'est pas envisageable pour les grands chromosomes (nous avions estimé le temps de calcul à plusieurs années). Nous avons donc choisi d'utiliser la stratégie « eucaryote » qui nous permettait de détecter les répétitions dans un temps raisonnable. Nous n'avons considéré que les graines de longueur supérieure ou égale à L_{min}. Devant l'abondance des graines (le temps d'extension était encore trop important), nous avons décidé de retirer les SSR (répétitions de faible complexité) et les répétitions multi-copies. En ne conservant que les répétitions à deux copies, nous avons ainsi retiré quasiment tous les éléments répétés en un grand nombre de copies (comme les éléments mobiles) sans nous baser sur les annotations.

Pour détecter les répétitions dans les génomes bactériens (troisième article), nous avons pu, à nouveau, utiliser la stratégie « bactérienne ». Pour cela, nous avons détecté des graines de 15 paires de bases (L_{min} varient entre 21 et 25 bases pour ces génomes). Toutes les graines détectées ont été étendues. Pour filtrer les répétitions nous avons mis au point une méthode empirique fondée sur le score d'alignement. Nous avons fixé un score minimum d'alignement (S_{min}) en dessous duquel une répétition est considérée comme non significative. Pour cela, nous avons construit 10 génomes aléatoires pour chaque génome analysé. Dans

ces génomes aléatoires, nous avons détecté les répétitions et établi la distribution de leurs scores d'alignement. Nous avons calculé S_{min} comme le quantile 1/1000 de cette distribution. Cette dernière méthode est certainement la plus « aboutie » des méthodes que nous avons utilisées, mais elle n'est pas applicable pour des grands génomes très répétés, car le temps de calcul deviendrait trop important.

B.1.2. Les algorithmes utilisés.

Dans les deux stratégies que nous avons utilisées, la détection des répétitions nécessite deux étapes clefs : (1) la détection des graines et (2) l'extension de ces graines.

Pour la détection des graines, nous avons utilisé deux algorithmes : KMR et un arbre de suffixe. Pour l'analyse des chromosomes de *S. cerevisiae* (premier article), nous avons choisi d'utiliser KMR, très performant pour la détection des graines. Cependant, cet algorithme crée un léger biais dans les graines détectées (voir ci-dessous). Après cette analyse, une implémentation très performante (et sans biais) d'un algorithme d'arbre de suffixes a été publiée (Kurtz and Schleiermacher 1999). Nous avons donc depuis, utilisé ce second algorithme pour détecter les graines.

L'extension des graines est réalisée en utilisant un programme d'alignement local basée sur un algorithme de programmation dynamique (Smith and Waterman 1981). Nous avons utilisé une version un peu différente de l'algorithme original afin de pouvoir aligner de très grandes séquences.

B.1.2.1. L'algorithme dérivé de KMR.

Cet algorithme est utilisé pour détecter des répétitions dans plusieurs types de séquences (Soldano *et al.* 1995) (Sagot *et al.* 1995) (Rocha *et al.* 1999a). Il est fondé sur l'algorithme de KMR (Karp, Miller et Rosenberg) (Karp *et al.* 1972). Ce denier repose sur l'idée intuitive que si deux répétitions strictes de taille k sont espacée d'une distance $\leq k$, il existe une répétition de taille k+ α (figure 29).

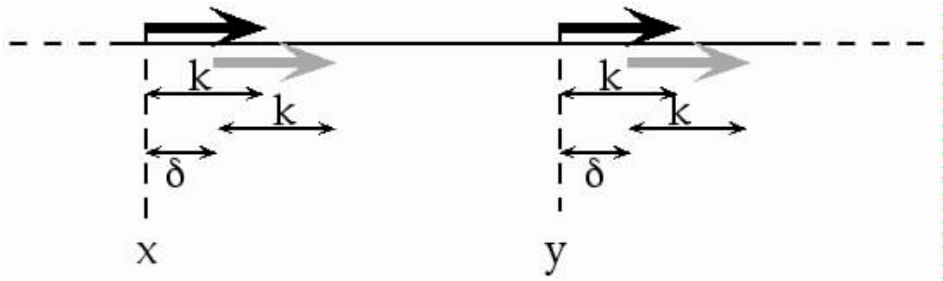

Figure 29 : Schéma représentant le principe sur lequel est fondé l'algorithme KMR.
Soit deux répétitions, dont les positions de la première sont x et y. k représente la longueur des deux répétitions et δ l'espacement entre ces répétitions. Si δ < k, alors il existe une répétition plus grande de taille k+ δ.

On peut ainsi facilement, à partir des mots répétés d'une taille 1, établir les mots répétés de taille 2, puis les répétitions exactes de taille supérieure. La version de KMR que nous avons utilisée recherche la graine la plus grande, remplace la séquence de cette graine par des "X", puis recherche la nouvelle répétition la plus grande (les "X" ne sont pas pris en compte pour la recherche de répétitions). Ce remplacement est nécessaire pour ne conserver que des graines de taille maximale. Cependant, son utilisation crée un léger biais. Si une position est déjà localisée dans une répétition, elle n'est plus prise en compte pour les autres recherches. Ceci peut conduire à manquer certaines graines. Par exemple, s'il existe trois copies d'une graine, mais deux d'entre elles sont plus grandes, la troisième ne sera jamais détectée.

B.1.2.2. Arbres de suffixes et reputer.

Nous avons utilisé une implémentation récente et performante d'arbre de suffixe, le programme *reputer* (Kurtz and Schleiermacher 1999). L'utilisation des arbres de suffixes pour rechercher des séquences répétées n'est pas récente, mais les implémentations précédentes étaient trop peu performantes pour être réellement utilisables (Kurtz and Schleiermacher 1999).

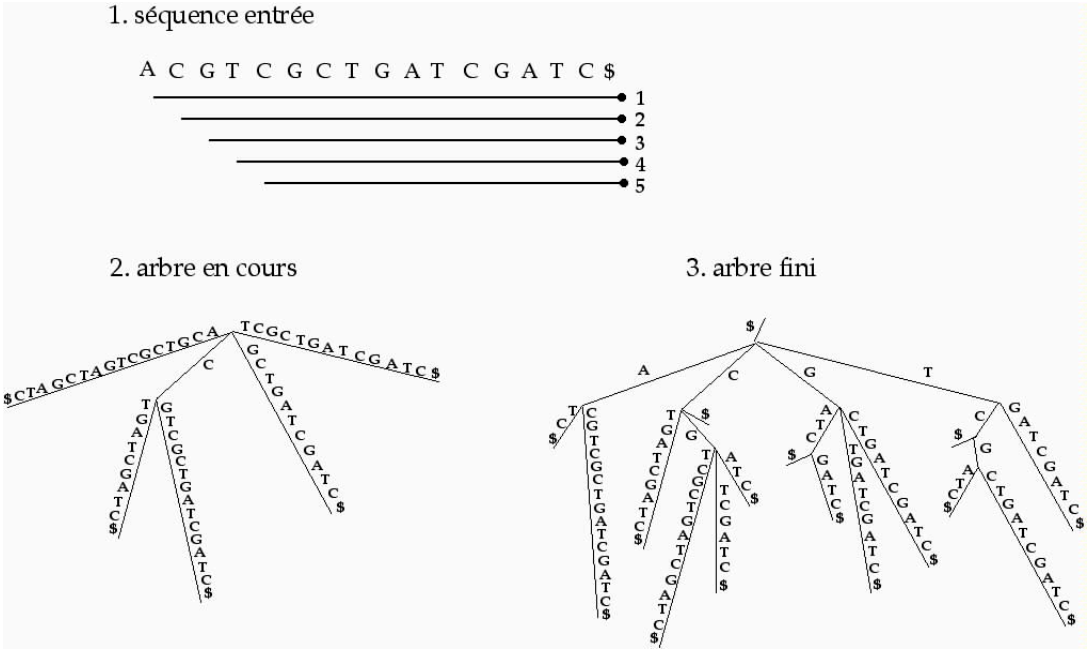

Figure 30 : Construction d'arbre de suffixes.
Ces quelques étapes indiquent la méthode générale à suivre pour cette construction.

Un arbre de suffixes est une structure arborescente qui permet, entre autre, la détection rapide des répétitions (figure 30). La construction se fait en ajoutant à l'arbre des suffixes (séquences) de plus en plus courts. Après chaque insertion dans l'arbre, le nouveau suffixe considéré est le suffixe précédent moins sa première lettre. Si un suffixe partage un préfixe commun avec un autre suffixe, les deux suffixes sont branchés sur leur partir commune. La fin est indiquée par un symbole non utilisé dans la séquence (dans le schéma un $). Quand toutes les séquences suffixes sont intégrées à l'arbre, la recherche de répétition se fait aisément en parcourant l'arbre et en recherchant les nœuds de l'arbre.

Reputer permet de détecter les répétitions directes et inversées dans une même recherche. Pour cela, il construit une séquence contenant la séquence entrée et son complémentaire. Les deux séquences sont séparées par des lettres non utilisées dans la séquence. Par exemple, à partir de la séquence entrée de la figure 30, on établit la séquence $xACGTCGCTGATCGATCyGATCGATCAGCGACGTz$. On construit avec cette dernière un arbre de suffixes, et l'on recherche les répétitions. Toutes les répétitions étant présentes deux fois dans cet arbre, seule une occurrence est considérée. Pour des raisons

d'optimisation, *reputer* ne considère que les quatre lettres ATCG et recherche uniquement des couples de copies de taille maximale.

Le programme *reputer* est un outil bien plus performant que le programme dérivé de l'algorithme de KMR pour la détection des couples de répétitions strictes dans une séquence d'ADN sans indécision dans la séquence (séquence ne contenant que des A, C, G ou T). Cependant, l'algorithme dérivé de KMR présente des fonctionnalités intéressantes non implémentées dans *reputer*. Comme on ne dispose pas des « sources » du programme, aucune modification ne peut être faite. Les fonctionnalités sont, par exemple, la recherche de répétitions à n copies, la détermination rapide de la séquence répétée la plus grande ou la recherche dans une séquence composée de plus de quatre lettres (protéine, ADN dégénéré, etc.).

B.1.2.3. Alignement local par programmation dynamique.

L'algorithme d'alignement local par programmation dynamique (Smith and Waterman 1981) dérive de l'algorithme d'alignement global (Needleman and Wunsch 1970). Ce dernier utilise la notion de score d'alignement. Un score d'alignement marque la ressemblance entre deux séquences, plus ce score est important, plus les deux séquences sont semblables.

L'idée de l'algorithme d'alignement global est de trouver un arrangement entre deux séquences qui maximise le score d'alignement (noté désormais S). Ce score est la somme de toutes les opérations élémentaires d'édition permettant l'arrangement de la séquence : décalages, insertions et délétions. Pour maximiser S, on utilise un algorithme de type « *divide and conquer* ».

$$S_{i,j} = \max \begin{cases} S_{i-1,j-1} + w(a_i, a_j) \\ S_{i-1,j} + w(a_i, 0) \\ S_{i,j-1} + w(0, b_j) \end{cases}$$

$w(a_i, b_j)$ est le score d'alignement entre le résidu a_i et le résidu b_j.

$w(a_i, 0)$ est le score d'alignement entre le résidu a_i et un « trou » (insertion d'un trou dans la séquence b).

w(0, b_j) est le score d'alignement entre le résidu b_j et un « trou » (insertion d'un trou dans la séquence a).

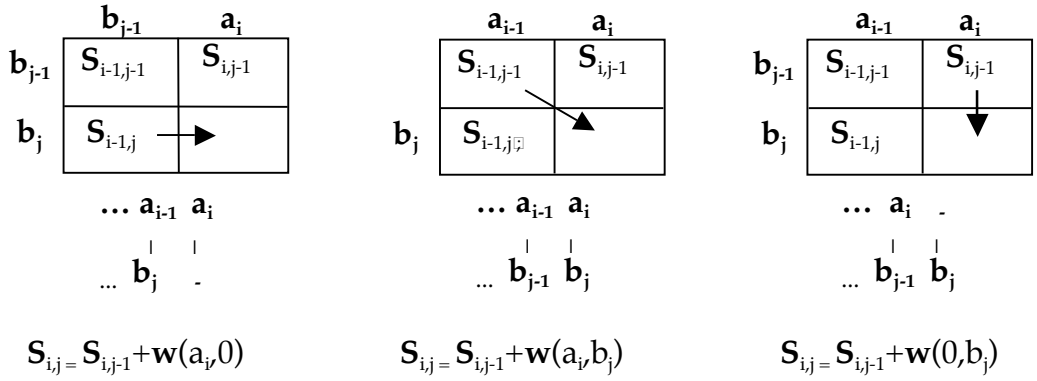

Figure 31 : Détermination du score maximum de la case (i,j).
Ce schéma illustre les trois possibilités pour chaque case i,j. Le score maximum $S_{i,j}$ est déterminé en comparant les trois scores possibles ($S_{i,j-1}+w(a_i,0)$; $S_{i-1,j-1}+w(a_i, b_j)$; $S_{i-1,j}+w(0,b_j)$) .

Soit deux séquences a et b de taille m et n. On note a_i, le $i^{ème}$ résidu de la séquence et b_j le $j^{ième}$ résidu de la séquence b. Le score maximum entre a_i et b_j est noté $S_{i,j}$. On calcule successivement les score $S_{i,j}$ pour des valeurs de i,j variant de 0,0 à m,n. $S_{m,n}$ sera le score maximum entre les deux séquence S. Les valeurs sont stockées dans un tableau à deux dimensions (chaque dimension est la longueur de la séquence + 1). La procédure débute avec $S_{0,0} = 0$. La valeur de chaque $S_{i,j}$ est calculée à partir des scores « précédents » ($S_{i-1,j}$ ou $S_{i,j-1}$ ou $S_{i-1,j-1}$). $S_{i,j}$ est calculé à partir de l'équation récursive suivante (illustrée dans la figure 31):

Le tableau à deux dimensions est rempli de proche en proche par cette méthode. Les scores $S_{i,j}$ (meilleur alignement jusqu'à la case i,j) sont notés dans ce tableau, et le chemin pour y arriver est également retenu. Un exemple de tableau est présenté dans la figure 32. Pour retrouver l'alignement optimum, il suffit de remonter le chemin depuis la case m,n. Il peut y avoir plusieurs alignements menant au même score maximum.

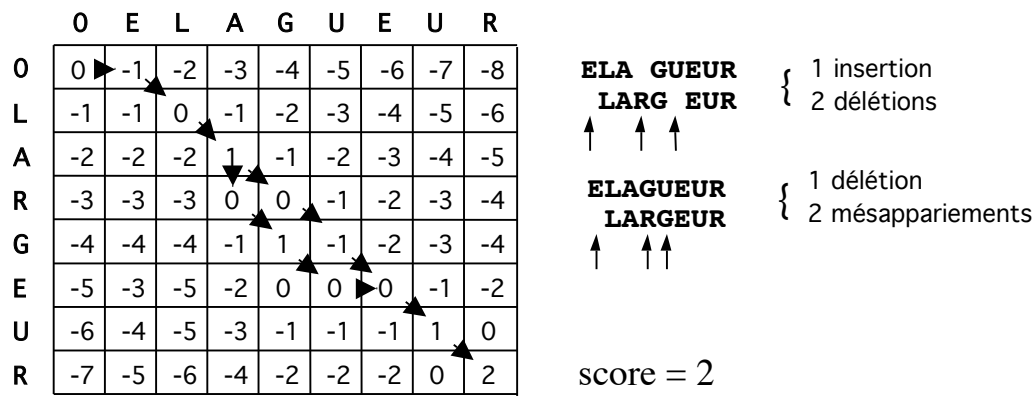

Figure 32 : Valeurs des scores maximaux pour un alignement global (d'après un cours de J. Pothier, 2002).
Dans cet alignement, l'appariement de deux résidus identiques vaut +1, les autres opérations (insertions, mésappariements) valent –1. Le score de l'alignement est 2. Les deux alignements présentés à droite possèdent le même score.

Pour effectuer un alignement local, le principe est identique. Cependant, deux différences majeures doivent être opérées lors de la construction du tableau :

- Dès qu'un score est négatif, il est remplacé par un score nul. Ainsi, on a des scores positifs dès que les deux régions sont similaires.
- On remonte le chemin à partir de la case contenant le score le plus élevé (la case ayant le meilleur sous-alignement). Le chemin est remonté jusqu'à ce qu'une valeur nulle soit rencontrée (fin de la similarité).

Grâce à ces deux modifications, l'alignement produit est le meilleur « sous-alignement » possible entre les deux séquences. Un exemple de tableau rempli pour un alignement local est montré sur la figure 33.

	0	E	L	A	G	A	G	E	S
0	0	0	0	0	0	0	0	0	0
L	0	0	1	0	0	0	0	0	0
A	0	0	0	2	1	2	0	0	0
G	0	0	0	0	3	2	3	2	1
A	0	0	0	1	2	4	3	2	1
F	0	0	0	0	1	3	3	2	1
F	0	0	0	0	0	2	2	2	1
E	0	1	0	0	0	0	1	3	2

LAGA
LAGA

score = 4

Figure 33 : Valeurs des scores maximaux pour un alignement local.
Dans cet alignement, l'appariement de deux résidus identiques vaut +1, les autres opérations (insertions, mésappariements) valent –1. Le score de l'alignement est 4.

Dans la version la plus simple, qui est présentée dans les figures 32 et 33, tous les appariements ont un score de +1, toutes les délétions et les mésappariements ont un score de –1. Pour les séquences biologiques, on utilise souvent des scores souvent plus nuancés. Par exemple, les insertions/délétions sont des événements plus rares que les simples substitutions. Cependant les insertions/délétions ont souvent une taille supérieure à 1 pb. On utilise donc des scores différents pour la « création d'une délétion » et pour « l'extension

d'une délétion». Nous avons, par exemple utilisé un score de –4 pour la création d'une délétion et de –1 pour son extension.

Il est également possible d'utiliser des matrices d'alignements permettant de nuancer les scores de ressemblance et dissemblance entre les symboles (Dayhoff *et al.* 1972). Par exemple, tous les petits acides aminés hydrophobes (comme leucine, valine, isoleucine, etc.) ont dans les matrices d'alignements de protéines des scores caractérisant leurs fortes ressemblances fonctionnelles. Dans les séquences d'acides nucléiques, on utilise souvent des matrices d'identité: +1 si le nucléotide est identique, -1 s'il est différent. L'utilisation d'une telle matrice fait l'hypothèse que tous les nucléotides ont la même chance de se retrouver aligné par hasard.

Les compostions globales en nucléotides sont très variables d'un génome à l'autre (Sueoka 1962). Par exemple, le génome de *Ureaplasma urealiticum* présente 26% de G+C, le génome de *E. coli* 50 % de G+C et le génome de *Halobacter sp* 68% de G+C. Il faut donc considérer que dans ces trois génomes, la chance d'apparier un G par hasard n'est pas la même. L'utilisation d'une matrice d'identité pour déterminer les scores d'alignement dans des génomes où la fréquence des nucléotides est très différente de 1/4 crée un biais.

Dans la détection des répétitions chez les bactéries (article 3), nous avons utilisé une matrice qui tient compte les fréquences de chaque nucléotide. Cette matrice de scores permet de donner un score plus important à un appariement entre des nucléotides rares et de pénaliser plus des mésappariements entre ces nucléotides rares. La matrice utilisée est présentée sur le tableau 5.

	A	**C**	**G**	**T**
A	$1-p_A \cdot p_A$	$-(1-p_C \cdot p_A)$	$-(1-p_G \cdot p_A)$	$-(1-p_T \cdot p_C)$
C	$-(1-p_A \cdot p_C)$	$1-p_C \cdot p_C$	$-(1-p_G \cdot p_C)$	$-(1-p_T \cdot p_C)$
G	$-(1-p_A \cdot p_G)$	$-(1-p_C \cdot p_G)$	$1-p_G \cdot p_G$	$-(1-p_T \cdot p_G)$
T	$-(1-p_A \cdot p_T)$	$-(1-p_C \cdot p_T)$	$-(1-p_A \cdot p_T)$	$1-p_T \cdot p_T$

Tableau 5: Matrice utilisée dans l'article 3 pour corriger les biais de composition.
pi *est la fréquence du nucléotide i. En pratique, comme l'implémentation que nous utilisons n'accepte pas de score décimal, la matrice est multipliée par 100 pour tenir compte de deux chiffres significatifs.*

Pour conclure, nous avons utilisé une version légèrement différente de l'algorithme dans la détection des répétitions (figure 34). La séquence s'étalant de 200 pb en amont jusqu'à 200 pb en aval de la graine est alignée localement. Si l'alignement local se termine tout près d'une des bordures, la répétition se prolonge probablement au delà des 200 pb prises en compte. 100 pb sont donc ajoutées à la bordure limitante et l'alignement est recommencé. Les ajouts et alignements se poursuivent ainsi jusqu'à ce qu'aucune bordure ne soit limitante. Or, l'alignement local ne peut pas être réalisé avec des séquences de grande taille (à cause du temps de calcul et de l'espace mémoire nécessaire). Nous avons donc choisi, si l'alignement dépasse 1kb, de conserver la région alignée et de ne poursuivre l'alignement que sur les séquences encore non alignées. Cette dernière opération nous permet d'aligner des répétitions aussi grandes que 600 kb (dans le chromosome 1 de *C. elegans*), mais ne garantissent plus un alignement optimal.

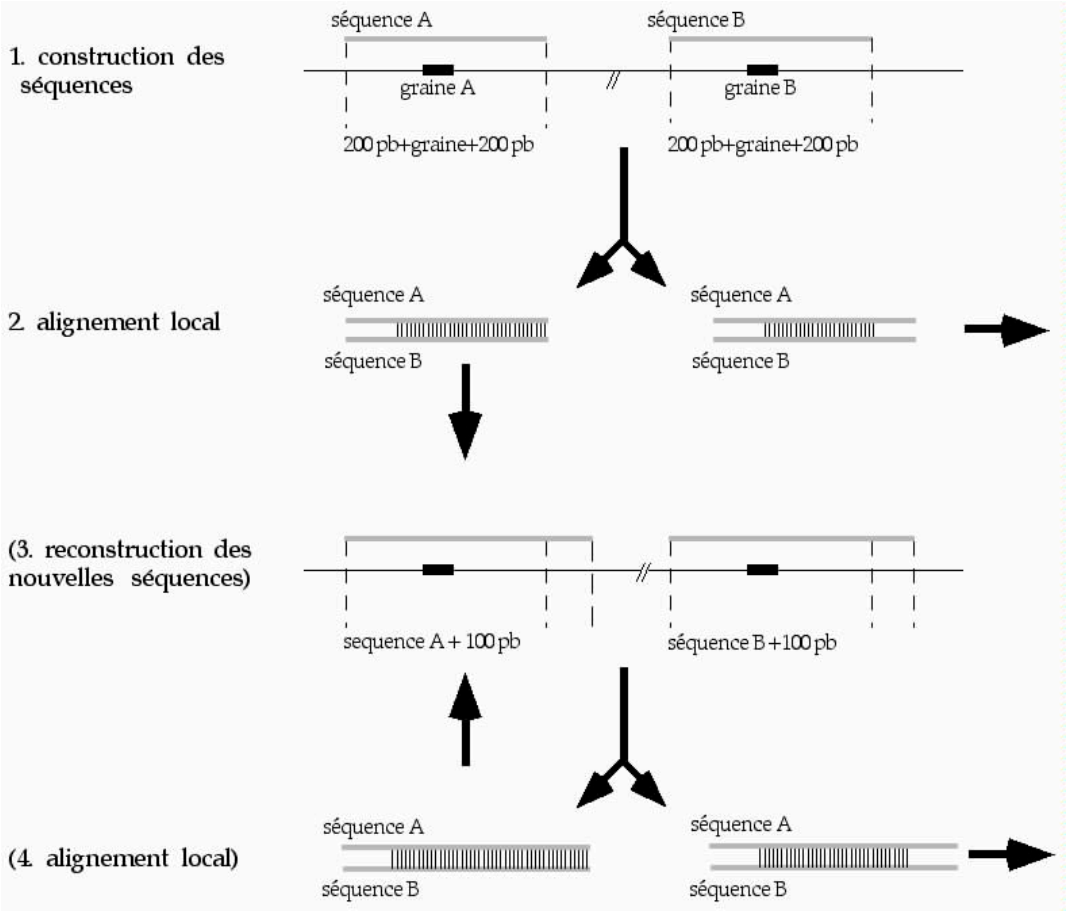

Figure 34 : algorithme utilisé pour détecter les répétitions.
Les séquences à aligner sont construites en considérant les 200 pb en aval et en amont de la graine(1). Les deux séquences sont alignées par alignement local (2). Si l'alignement se termine près d'une bordure, les séquences sont artificiellement limitantes. 100bp sont ajoutées aux séquences à aligner(3) et un nouvel alignement local est

opéré (4). L'ajout des 100 pb ne s'arrête quand l'alignement se termine loin d'une bordure (3,4). Les caractéristiques de la répétitions sont alors déterminées.

III. Résultats

Les résultats seront présentés en trois parties construites autour des trois étapes de notre exploration des chromosomes dont la séquence nucléique était achevée et publique en 2001. La première concerne, pour des raisons historiques, l'analyse du génome de *S. cerevisiae*, la seconde celle de 6 génomes eucaryotes et la troisième celle de 53 chromosomes bactériens (Bactéries et Archées).

A. Les duplications intrachromosomiques chez la levure *S. cerevisiae*.

La séquence complète du génome de la levure de boulanger *S. cerevisiae*, a été publiée en 1997 (Goffeau and authors) 1997). C'est l'aboutissement d'une collaboration internationale menée sur plusieurs années. L'équipe du Pr. Netter, où j'ai réalisé ma thèse, fut engagée dans ce projet et travailla à l'avènement de la séquence complète du chromosome VII de cet organisme (Tettelin *et al.* 1997). Les 39,4 kilobases séquencés par cette équipe (Coissac *et al.* 1996) contenaient 19 ORF, dont une petite ORF de quelques trois cents nucléotides, appartenant à une opulente famille multigénique de 23 membres, celle des séripaupérines (protéines pauvres en sérine) (Viswanathan *et al.* 1994).

Par la suite, l'équipe s'intéressa à l'ensemble des familles de gènes dans le génome de *S. cerevisiae* et plus particulièrement à leur localisation. Une analyse des gènes dupliqués de *S. cerevisiae*, fut menée conjointement avec l'analyse de quatre autres génomes : deux génomes bactériens complets, *H. influenzae* et *M. genitalium* et deux génomes partiels : *E. coli* et *C. elegans* (Coissac *et al.* 1997). Les duplications furent recherchées en déterminant les paralogues potentiels de tous les gènes. Les gènes étaient utilisés comme bornes physiques le long du chromosome sans que leur fonction soit prise en compte.

L'utilisation des séquences protéiques permet de détecter des répétitions plus anciennes. Trois effets concourent à cela :

- L'alphabet des protéines est constitué de 20 symboles et celui de l'ADN de 4 symboles. Il est donc plus rare d'observer des similarités dues au hasard entre deux séquences d'acides aminés qu'entre deux séquences d'acides nucléiques.

- Comme plusieurs combinaisons d'acides nucléiques codent pour le même acide aminé, certains changements au niveau de la séquence nucléique ne changent pas la séquence de la protéine.
- Les acides aminés peuvent être divisés en groupe : par exemple, leur hydrophobicité, leur fréquence de remplacement, etc.. Comme les protéines ont une fonction biologique, le remplacement d'un acide aminé d'un groupe biochimique par un autre de ce même groupe trouble moins (ou pas) la fonction de la protéine. On peut donc construire des matrices de ressemblance (ou dissemblance) entre acides aminés affinant la mesure de ressemblance globale entre deux protéines.

Cette étude des duplications, ainsi que celle de Wolfe et Shields (Wolfe and Shields 1997) mis en évidence de très larges segments dupliqués entre les différents chromosomes de la levure (une cinquantaine de segments couvrant 50% du génome). Ces segments n'étaient pas présents dans les autres génomes analysés. Si les observations concordent entre ces deux analyses, les causes proposées pour l' existence des segments dupliqués sont différentes. D'après KH Wolfe, ils sont la trace d'un événement de polyploïdisation ancestral suivi de nombreux réarrangements (Wolfe and Shields 1997). D'après E Coissac, elles sont la marque d'un processus continu de duplications segmentaires (Coissac 1996).

Quelle que soit la réponse à cet épineux problème, ces régions dupliquées sont coorientées et colocalisées vis-à-vis des télomères ; "coorientées" signifiant que l'orientation respective des deux régions par rapport aux télomères est identique et "colocalisées" signifie que la distance au télomère des deux blocs est similaire. Ces caractéristiques ne permettent pas de résoudre le problème sus-décrit, mais indiquent que les télomères jouent un rôle important dans les mécanismes qui ont conduit à ces régions segmentaires dupliquées.

Afin d'affiner l'analyse des répétitions dans le génome de *S. cerevisiae*, nous avons recherché des répétitions dans les séquences d'ADN (et non plus en utilisant les protéines). Si l'analyse des gènes paralogues permet de détecter des événements de duplications plus anciens, celle des répétitions d'ADN permet une analyse plus fine des événements de duplications récents. En effet, ce second type d'analyse permet de préciser les positions, la

longueur et la similarité des séquences répétées. Elle permet également de détecter des répétitions dans des régions non-codantes (ou non annotées). Cependant, la recherche exhaustive des séquences répétées dans les séquences d'ADN présente également des limites. Elle permet de détecter exclusivement des événements "récents". Les similarités dans les séquences d'ADN doivent être fortes pour être considérées comme significatives.

Par ailleurs, la méthode utilisée pour les détecter est, dans sa version la plus complète, très consommatrice de temps de calcul et de mémoire. Nous avons donc choisi de recourir à une heuristique (voir *Matériel et Méthodes*). Nous cherchons les répétitions totalement identiques d'une taille minimum et déterminons les similarités aux bornes de ces répétitions exactes. De plus, nous avons choisi, dans ce travail, de ne rechercher que les répétitions intrachromosomiques.

Les répétitions directes et les répétitions inversées (figure 16) ont été détectées dans les seize chromosomes de *S. cerevisiae*. Un certain nombre de répétitions étant déjà bien connues dans le génome de cette levure, elles ont été retirées de celles que nous avons détectées. Ont été retirées les répétitions de faible complexité (type microsatellites), celles dont l'une des deux copies étaient localisées dans les subtélomères ainsi que, sur la base des annotations, celles associées aux 275 ARN de transfert, aux 2 ARN ribosomiques, aux 50 Ty (rétroéléments à LTR) et aux 385 solos (LTR seuls issus d'une délétion interne des Ty). Nous avons choisi de ne garder que les répétitions les plus "génériques. Les ARNr ont, en réalité, plus que deux copies dans le génome de la levure et couvrent entre 1 et 2 Mb du chromosome XII. Ces répétitions en tandem, hautement variables (en taille), sont identiques en séquence. Les auteurs de la séquence publique ont décidé de ne laissé que deux des multiples copies dans la séquence disponible. Il en de même pour les répétitions de CUP1 (détoxication du cuivre) et de PMR2 (pompe à calcium) (Dujon 1996).

Chaque expérience ayant besoin d'un témoin, nous avons construit des génomes aléatoires de même longueur et de même composition que le vrai génome. Dans ces génomes, la même méthode de détection a été appliquée. Aux limites de détection que nous avons fixées, il faut environ 10 génomes aléatoires pour produire un nombre de répétitions

comparable à un génome réel. Tout au long de notre analyse, nous avons comparé ces répétitions « aléatoires » aux répétions « réelles » pour pouvoir séparer les effets biologiques des artefacts créés par la méthode.

A.1. Article 1.

Analysis of Intrachromosomal Duplications in Yeast *Saccharomyces cerevisiae*: A Possible Model for Their Origin

Guillaume Achaz, Eric Coissac,* Alain Viari,† and Pierre Netter**

*Structure et dynamique des génomes, Institut Jacques Monod, Paris, France; and †Atelier de Bioinformatique, Université Paris VI, Paris, France

The complete genome of the yeast *Saccharomyces cerevisiae* was investigated for intrachromosomal duplications at the level of nucleotide sequences. The analysis was performed by looking for long approximate repeats (from 30 to 3,885 bp) present on each of the chromosomes. We show that direct and inverted repeats exhibit very different characteristics: the two copies of direct repeats are more similar and longer than those of inverted repeats. Furthermore, contrary to the inverted repeats, a large majority of direct repeats appear to be closely spaced. The distance (delta) between the two copies is generally smaller than 1 kb. Further analysis of these ''close direct repeats'' shows a negative correlation between delta and the percentage of identity between the two copies, and a positive correlation between delta and repeat length. Moreover, contrary to the other categories of repeats, close direct repeats are mostly located within coding sequences (CDSs). We propose two hypotheses in order to interpret these observations: first, the deletion/conversion rate is negatively correlated with delta; second, there exists an active duplication mechanism which continuously creates close direct repeats, the other intrachromosomal repeats being the result, by chromosomal rearrangements of these ''primary repeats.''

Introduction

Since the first complete bacterial genome (Fleischmann et al. 1995), 22 new eubacteria, 6 archebacteria, and 3 eukaryote sequences have been completed, and several new genomics fields, such as ''functional genomics'' (deciphering the function of genes) and ''comparative genomics'' (comparison of entire genomes) (Chervitz et al. 1998), have emerged. Here, we focus on ''dynamical genomics,'' which can be seen as the study of chromosome history and dynamics through the analysis of the structure of current genomes. One way of studying these phenomena is through the analysis of chromosomal rearrangement remnants, such as duplications. Among eukaryotes, budding yeast *Saccharomyces cerevisiae,* which has been completely sequenced (Goffeau et al. 1996), is a good model because of its small size (12.1 Mb) and its comprehensive annotation.

The first evidence of sequence duplication in *S. cerevisiae* came from Lalo et al. (1993), who found a large duplication event between chromosomes II and XIV. More exhaustive studies, based on translated coding sequence (CDS) alignments, have brought prominence into large interchromosomal duplications (Coissac, Maillier, and Netter 1997; Wolfe and Shields 1997). Further analysis revealed that, for the two copies of a duplicated CDS, the distances to the closest telomere are similar (Coissac, Maillier, and Netter 1997). The importance of telomeres underlines the relation between nuclear organization and genome dynamics. Other studies were undertaken at the DNA level leading to the development of a ''duplication databank'' (Mewes et al. 1997) and to the definition of the X2 element in the subtelomeric region (Britten 1998).

However, to our knowledge, apart from the description of gene tandem duplication (CUP1, PMR2, rDNA, ASP3), no systematic study has yet been undertaken on intrachromosomal duplications. In the present work, we searched for intrachromosomal repeats at the level of nucleotide sequences. Through this analysis, we show that direct and inverted repeats exhibit very different characteristics. Moreover, we identify a special class of direct repeats (named close direct repeats) exhibiting several particular features. Finally, we propose a model based on the active flow of creation of these close direct repeats and their dispersion by chromosomal rearrangements.

Materials and Methods

Data

The *S. cerevisiae* complete sequences and annotations were extracted from the Saccharomyces Genome Database (SGD; http://genome-www.stanford.edu/Saccharomyces/). The total size of the 16 chromosomes is 12.1 Mb. We used the entire nuclear sequences as given in the database, including the three tandem clusters (CUP1, rDNA, and PMR2), which were reduced to a single repeat. We additionally built 10 ''random genomes'' by shuffling each chromosome independently with respect to its dinucleotide composition.

Construction of the Repeat Database

Our primary goal was to look for approximate repeats, i.e., repeats whose copies may not be strictly identical but may contain errors (mismatches and indels). The usual procedure for this purpose derives from dynamic programming (Smith and Waterman 1981) but is unfortunately not amenable to the study of very long sequences because of its quadratic time complexity. Although several heuristics have already been proposed to work around this problem (Leung et al. 1991; Vincens

Abbreviation: CDS, coding sequence.

Key words: genome dynamics, evolution, duplication, direct repeats, *Saccharomyces cerevisiae.*

Address for correspondence and reprints: Guillaume Achaz, Structure et dynamique des génomes, IJM, Tour 43–44, 1° étage, 4, place Jussieu, 75251 Paris CEDEX 05, France. E-mail: achaz@ijm.jussieu.fr.

Mol. Biol. Evol. 17(8):1268–1275. 2000

et al. 1998), we chose here to develop our own procedure in order to fit the biological problem more closely. Like most of the already-proposed heuristics, this procedure first looks for ''seeds'' of exact repeats and then extends the seeds by using dynamic programming techniques. This is done for each chromosome independently in four consecutive steps which are described as follows.

First Step: Searching for Seeds

Exact repeats were detected by using the Karp-Miller-Rosenberg (KMR) algorithm (Karp, Miller, and Rosenberg 1972), which finds the largest subword present at least r_{min} times (here $r_{min} = 2$) in a text (here, each chromosome). Since we were interested in ''unusually'' large repeats (i.e., repeats which did not appear by chance), we set a threshold (L_{min}) on the minimal length of repeats of interest. L_{min} was calculated using the statistics developed by Karlin and Ost (1985). For each chromosome, we chose L_{min} such that the probability of finding a three-copy repeat in a random sequence with the same length and base composition on the chromosome was less than 0.001. L_{min} typically ranges from 15 to 17 bp depending on the chromosome length.

In order to avoid the problem of any two subwords of a repeated word being themselves repeated, we devised the following heuristics (Rocha, Danchin, and Viari 1999): first, the longest repeat on the chromosome is sought, and its length is compared with the minimum preset value L_{min}. When a test is successful, both copies of the repeat are masked and excluded from further analysis. The process is iterated up to the point where the length of the largest repeats becomes smaller than L_{min}. It should be pointed out that this process is a heuristic. In particular, if there is a three-copy repeat where the third copy appears a little bit shorter, then this copy will be missed by the method. This explains the rationale behind the procedure to set up the threshold L_{min} (vide supra). We devised two versions of the program: one to detect direct repeats and the other to detect inverted repeats (repeats for which the second copy has the reverse orientation). The two orientation classes (direct and reverse) were further handled separately.

Second Step: Removing Low-Complexity, Overlapping, and Telomeric Seeds

In order to remove low-complexity repeats (like microsatellites), we used an entropy filter. The entropy is taken here in the sense of Shannon (Schneider et al. 1986) for dinucleotide distribution:

$$H = -\sum_{i=AA}^{TT} p_i \log_{16} p_i,$$

where p_i is the frequency of the ith dinucleotide. The entropy (H) is computed on the sequence of a repeat (H_{repeat}) and on the whole chromosome ($H_{chromosome}$). The values are then compared by computing the ratio $H_{repeat}/H_{chromosome}$. This ratio was calibrated by using artificial stretches of mono-, di-, tri-, and tetranucleotides to define a threshold: only repeats whose ratio was greater than 0.6 were kept.

Next, we discarded all repeats for which the two copies overlap. At this stage, these repeats generally correspond to multicopies of small words.

Finally, we removed all subtelomeric duplications. Several well-known elements are located in the subtelomeric regions (Y′ sequence [Louis and Haber 1990], X2 [Britten 1998], seripauperine [Viswanathan et al. 1994]). These elements have already been widely studied and are known to exhibit a highly special plasticity (for review, see Pryde, Gorham, and Louis 1997). We arbitrarily set a subtelomeric barrier at 30 kb and removed all repeats with at least one copy in a subtelomeric region.

Third Step: Extending the Seeds

Exact repeats (seeds) were extended into larger nonstrict repeats by using a local alignment program (Smith and Waterman 1981) developed by P. Hardy and M. Waterman (http://www.hto.usc.edu/software/seqaln/). The sequence of a seed was substituted with X's, and 100 bp were picked on both sides. For example, a seed of 30 bp will become $(A/C/G/T)_{100}$-$(X)_{30}$-$(A/C/G/T)_{100}$. The scoring matrix retained for the alignment was as follows: match(A/T/C/G) = +4; match(X) = +99; mismatch(A/T/G/C) = −4; mismatch(X) = −99; Gap_{open} = −16; $Gap_{extension}$ = −4. The value +99 will force the program to always align the two copies of the seed. When the best local alignment found by the program ended less than 10 bp from one of the sequences termini, the sequences were further extended 200 bp and a new run was performed. This operation was iterated until the alignment eventually ended more than 10 bp from both sides. It should be pointed out that after this step, several different initial seeds may give rise to the same (or a similar) extended repeat. Therefore, when two or more extended repeats occurred at the same location (with a tolerance of 20% of their length), we just kept the longest one.

Fourth Step: Removing Short or Biological Trivial Repeats

In order to remove repeats that were too short or too different, we decided to keep repeats with (1) a minimum percentage of identity and (2) a minimum number of matches between their two copies. These minima were arbitrarily set at 50% identity and 30 matches. Finally, we applied a last filter in order to remove all ''biologically trivial'' duplications, which have their own dynamics. Actually, many of the repeats were due to the 275 tRNAs, 2 rRNAs, 50 Ty's, or 385 solos widespread in the yeast genome and were therefore removed. The positions of these known repeated elements were extracted from the SGD annotations (http://genome-www.stanford.edu/Saccharomyces).

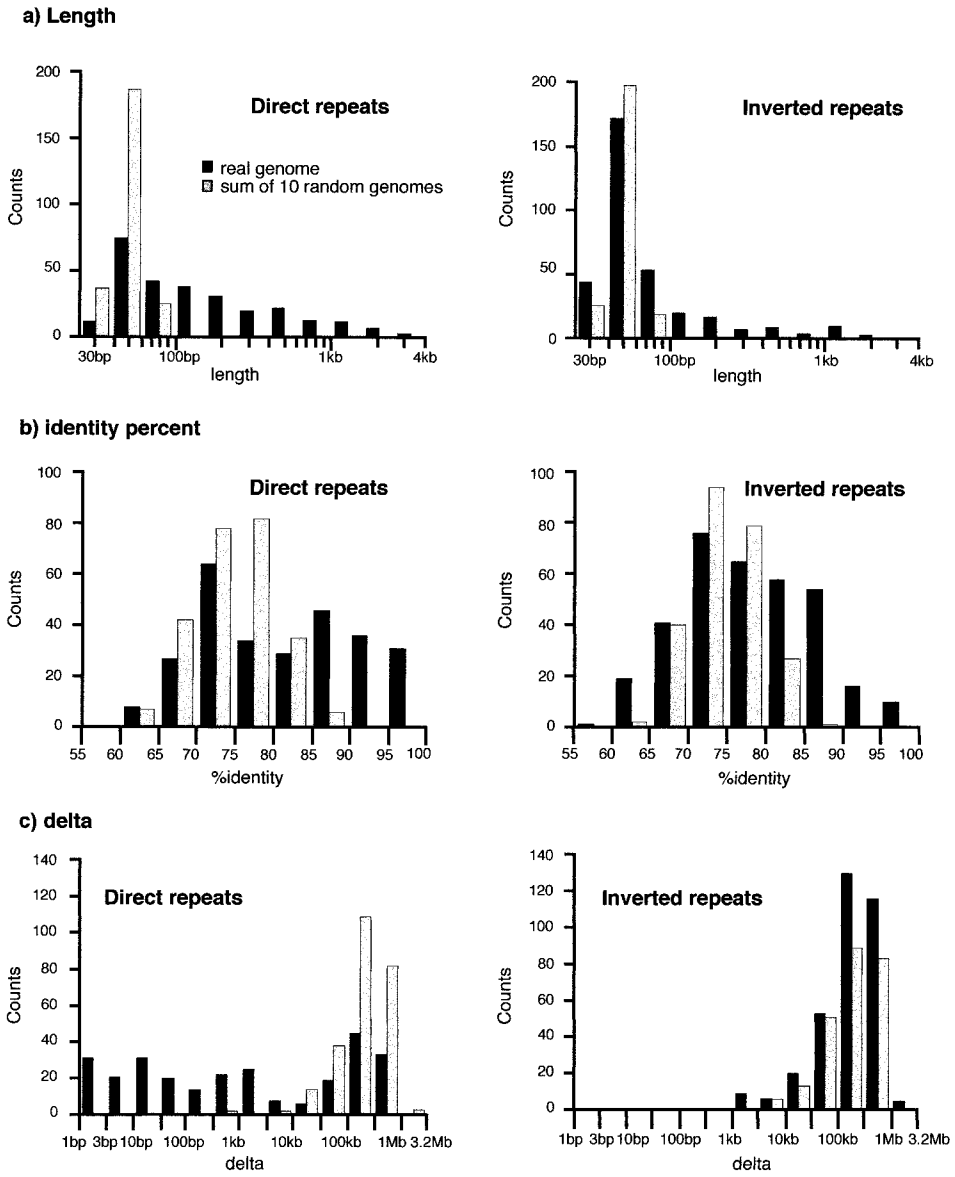

FIG. 1.—Distribution of the three parameters (length, identity, and delta) used in this study for each orientation (direct or inverted) of the repeats. Black boxes represent data observed for the real yeast genome, and gray boxes correspond to shuffled data. Since much fewer repeats are observed on random data (see text), repeats from 10 random genomes (each chromosome is shuffled with respect to the dinucleotide composition) have been pooled. *a,* Histogram of the length of the repeats. *b,* Histogram of the percentage of identity between the two copies of a repeat. *c,* Histogram of the distance (delta) between the two copies of a repeat.

Results

The application of the previously described method yields a total of 275 direct repeats and 340 inverted repeats on the yeast genome. In comparison, the random genomes (see *Materials and Methods*) produce an average of 25 direct repeats and 24 inverted repeats. The number and distribution of repeats differ from one chromosome to the other (data not shown). However, in the rest of this analysis, we pooled together all of the repeats in order to get sufficient statistics to study their global properties. In order to examine more closely the characteristics of the repeats, we focused on three parameters: ''length'' simply denotes the mean length of the two copies; ''identity'' is defined as the ratio of the number of matches between the two copies over the length of the largest copy; and ''delta,'' also called ''spacer'' in the literature (Klein 1995), is defined as the distance between the two copies. For both orientations, delta begins after the 3′ end of the first copy. It stops at the 5′ end of the second copy for direct repeats and at its 3′ end for inverted ones.

Differences Between Direct and Inverted Repeats

Figure 1 shows the distributions of the three parameters described above for the two orientation classes and for real and random genomes. The comparison of real direct repeats with random ones in figure 1*a* reveals important differences: random repeats are all shorter than

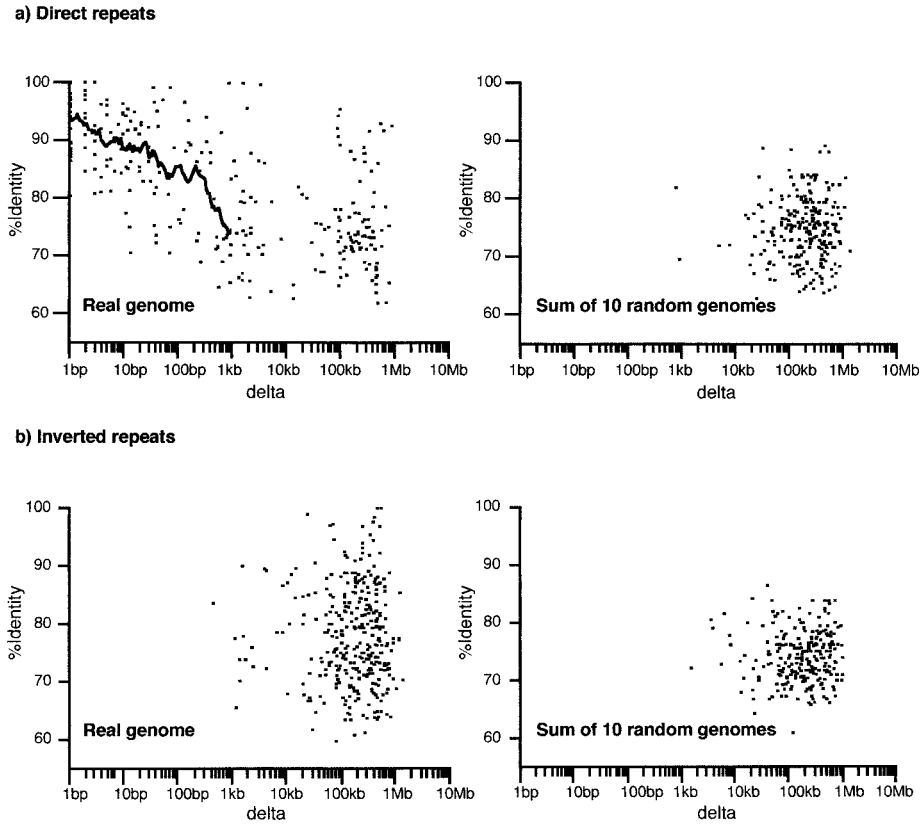

FIG. 2.—Negative correlation between the percentage of identity and the spacing (delta) between the two copies of a repeat. The percentage of identity (y-axis) is plotted as a function of delta (x-axis on a logarithmic scale) for both real (left side) and shuffled (right side) yeast genomes. Direct repeats are given in *a*, and inverted repeats are given in *b*. Since much fewer repeats are observed on random data (see text), and in order to give rise to a comparable total number of points, the plots on the right actually correspond to the sum of 10 random genomes. The black curve (for real data) represents the mean of the y values (identity) computed on a sliding window spanning 20 data points. This visual negative correlation is further confirmed by Kendall tau rank tests (see text).

100 bp, whereas a significant number (146/275) of real ones are much longer than 100 bp (up to 3,885 bp coming from the ENA family on chromosome IV). On the contrary, real inverted repeats behave much like random ones: only a few (71/340) real inverted repeats are significantly longer than 100 bp. Thus, on the sole basis of their length, it seems clear that real direct repeats are different from real inverted ones.

As shown in figure 1*b*, both real direct and inverted repeats show a higher percentage of identity than random ones. Moreover, by comparing the two orientation classes for real data, a major difference appears: direct repeats exhibit a higher degree of similarity than inverted ones (for instance, 103 direct repeats, against only 29 inverted repeats, are found above 90% identity).

Finally, the histograms of delta (fig. 1*c*) highlight another important structural difference between real direct and inverted repeats. Most (139/275) real direct repeats have deltas shorter than 1 kb, while random repeats exhibit almost exclusively deltas longer than 1 kb. In contrast, real inverted repeats display about the same distribution as random inverted ones.

In summary, these results show that both orientation classes are different from random distribution and that real direct and inverted repeats constitute two different populations with distinct properties. The main difference concerns the delta parameter, with the majority of direct repeats being closely spaced (delta smaller than 1 kb). Hereafter, we refer to them as ''close'' (as opposed to ''distant'') direct repeats.

Identity Is Negatively Correlated with Delta for Close Direct Repeats

In order to reveal possible correlations between the parameters, we plotted, for both orientation classes and for real and random genomes, the identity as a function of delta. Figure 2*a* suggests that close direct repeats are negatively correlated to delta. This visual observation is further confirmed by Kendall tau correlation measurement ($\tau = -0.36$; $P \approx 10^{-10}$). In contrast, no such correlation is found for larger deltas nor for random repeats.

Length Is Positively Correlated with Delta for Close Direct Repeats

Similarly, we searched for a correlation between length and delta for both orientation classes and for real and random genomes. Figure 3*a* reveals a peculiar variation of the length as a function of delta. More precisely, close direct repeats exhibit a positive correlation ($\tau = +0.26$; $P \approx 3 \times 10^{-6}$) between length and delta. It should be noted that no significant rank correlation

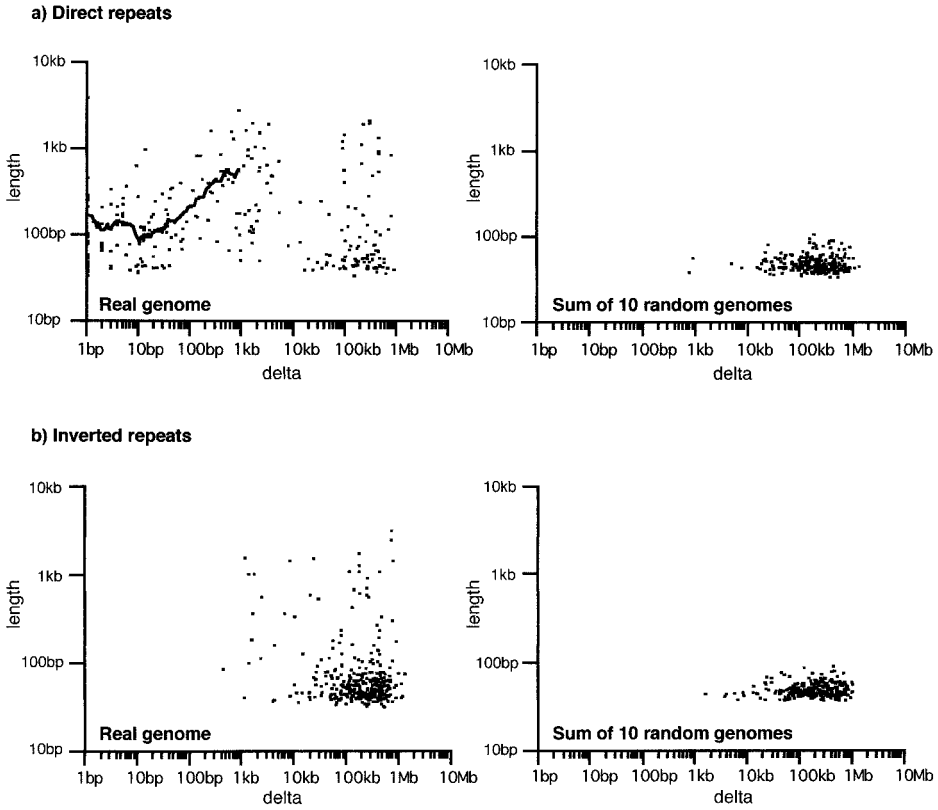

FIG. 3.—Positive correlation between the length and the spacing (delta) between the two copies of a repeat. The mean length of the two copies of a repeat (y-axis) is plotted as a function of delta (x-axis on a logarithmic scale) for both real (left side) and shuffled (right side) yeast genomes. Direct repeats are given in *a,* and inverted repeats are given in *b.* Since much fewer repeats are observed on random data (see text), and in order to give rise to a comparable total number of points, the plots on the right actually correspond to the sum of 10 random genomes. The black curve (for real data) represents the mean of the y values (identity) computed on a sliding window spanning 20 data points. This visual positive correlation is further confirmed by Kendall tau rank tests (see text).

between length and identity was observed for close direct repeats.

Close Direct Repeats Are Mostly ''Coding'' Sequences

In order to find out whether repeats are located inside CDSs, we examined positions of repeats in relation to the CDSs. This brought out two main results. Close direct repeats are mainly located within coding sequences: 85.6% (119/139) of them have their two copies completely included within CDSs and, with two exceptions, these repeats are always located within the same CDS. Moreover, it turns out that for 115 of these 117 repeats, the two copies are in the same coding frame, therefore giving rise to repeats at the protein level too.

In contrast, a much lower percentage of distant repeats (58%; 79/136) and inverted repeats (40.6%; 138/340) are completely included within CDSs. Moreover, only 50.6% (40/79) of distant direct DNA repeats and 34.1% (47/138) of inverted DNA repeats correspond to repeats at the protein level.

Discussion

This investigation on intrachromosomal duplications allows us to bring out several biological results and hypotheses about the dynamics of repeats. The first set of arguments comes from the analysis of the data presented in figures 2 and 3: for direct repeats, the main differences being observed between close repeats (delta $<$ 1 kb) and distant repeats (delta $>$ 1 kb) were as follows:

1. In figure 2, for close direct repeats, one can observe a negative correlation between the percentage of identity and delta: the shorter the delta, the higher the identity. Similar results have already been suggested for *Caenorhabditis elegans* (Semple and Wolfe 1999). This result could be understood if a high percentage of identity is (*i*) the mark of a recent adjacent duplication event and/or (*ii*) the result of an active conversion process (homogenization of the two copies). This latter process may depend upon the relative distance of the two copies (vide infra).
2. Figure 3 shows that, for close direct repeats, there is a positive correlation between length and delta: the shorter the delta, the shorter the length of the repeat. This correlation could be interpreted as the result of (*i*) a specific mechanism preferentially deleting large repeats (the loss of one copy leading to a single sequence) and/or (*ii*) genetic erosion due to the mutational events accumulated

from the initial duplication, therefore leading to a lower identity percent.

Some of the interpretations invoked above could be considered contradictory. In particular, the high percentage of identity of the close direct repeats is considered the mark of a recent duplication event (point 1, *i*), whereas their short length could be a consequence of the long time elapsed from the initial duplication event (point 2, *ii*). Therefore, as explained below, we shall consider the second explanation less probable.

Conversion Versus Deletion: A Plausible Explanation

For close direct repeats, it seems reasonable to think that the extent of the exchange process is negatively correlated with delta. The exchange process can be either deletion (loss of one copy by reciprocal recombination or replication slippage) or conversion (homogenization of the two copies by nonreciprocal recombination). In fact, experimental studies undertaken on *Bacillus subtilis* (Chedin et al. 1994) and *Escherichia coli* (Lovett et al. 1994) have highlighted similar results. Thus, the decrease in identity as a function of delta (fig. 2) could be explained by a decrease in the conversion rate.

To understand the correlation between length and delta (fig. 3), we must put the genetic exchange back in its dynamic context: actually, each close repeat can be submitted to a deletion or to a conversion event. If a deletion occurs, there is no way back. On the contrary, if a conversion occurs, the two copies are still present and a new round of exchange (i.e., conversion or deletion) is possible. So, during a long period, a bias in favor of deletion of one copy should be observed. Furthermore, several experiments have demonstrated a positive correlation between recombination rate and repeat length in yeast (Jinks, Michelitch, and Ramcharan 1993), bacteria (Peeters et al. 1988), and phages (Pierce, Kong, and Masker 1991). Briefly, for a short delta, a long repeat should be too unstable to persist, but by increasing delta, longer repeats could be maintained. This ''length tolerance'' effect could explain the observed positive correlation between length and delta.

Functional Pressures: A Protection from Deletion

Another important difference between close and distant repeats is related to their presence within CDS: close direct repeats are located mainly within CDSs and in the same frame, therefore giving rise to repeats at the protein level as well. On the contrary, distant direct repeats give rise to fewer protein repeats. These observations lead to two nonexclusive hypotheses:

1. Close direct DNA repeats are most probably submitted to an active recombination pressure leading to the deletion of one of the copies. However, the repeat can be fixed if it is submitted to functional pressures at the protein level. The consequence is that one very rarely observes close direct repeats which have not been protected from deletion by this selective advantage (i.e., located outside CDSs), because they have been massively lost by recombination. Finally, close direct repeats which have been fixed by functional pressures at the protein level are still submitted to an active conversion process (vide infra), preventing any further evolution.
2. On the other hand, distant repeats are submitted to less active recombination and conversion pressures. This allows the creation of different proteins from the same repeated DNA sequence located in different CDSs, and even sometimes translated in different reading frames.

It should be pointed out that Marcotte et al. (1998) recently reported that there is a high frequency of internal repeats within proteins sequences of eukaryotes (as compared with prokaryotes). This observation can be in line with the result presented in this study. To summarize, we probably observe a combination of DNA mechanisms which tend to (1) delete the close repeats and (2) keep them identical and which are constrained by functional pressures at the protein level.

Direct Versus Inverted Repeats

The last group of results we take into account concerns the important differences observed between direct and inverted repeats. These differences can be summarized as follows:

1. The direct repeats exhibit a higher similarity than the inverted ones (fig. 1*b*). If we assume that there is a higher conversion rate for the numerous close direct repeats (vide supra), this observation is not surprising.
2. The direct repeats are clearly longer that the inverted ones (fig. 1*a*). This result is surprising, since, as already mentioned, long direct repeats are experimentally known to be more easily deleted (Jinks, Michelitch, and Ramcharan 1993). Therefore, our interpretation is that the observed long direct repeats were produced more recently than inverted ones and have not yet been eliminated.
3. Finally, the repartition of deltas is the main difference between the two groups of repeats (fig. 1*c*). Inverted repeats do not seem to be constrained by delta (with the corresponding distributions being almost identical for real and random genomes). On the contrary, as shown above, close direct repeats are overrepresented and distant direct repeats are underrepresented as compared with inverted ones. We now discuss a possible model to account for these differences.

A Dynamic Model for Intrachromosomal Repeats

We propose a simple model, illustrated in figure 4, to explain all these observations and solve the apparent contradictions. In this model, the initial event is the continuous production of close direct repeats. Whatever the mechanism giving rise to it (unequal crossing over or replication slippage), when a close direct repeat is created, it can be either modified by mutation (although the

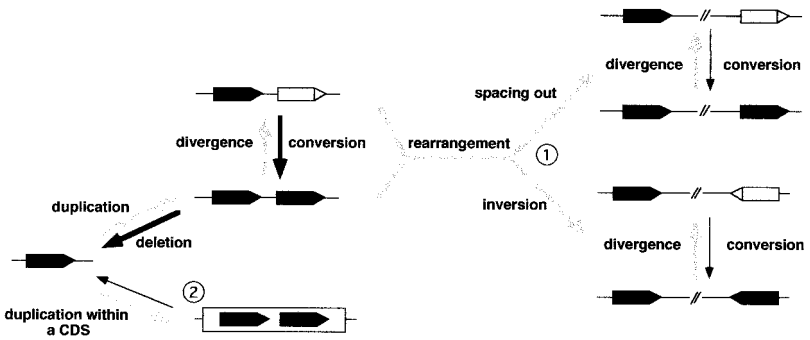

FIG. 4.—A model for the origin and dynamics of intrachromosomal repeats. The initial event is the close duplication of a sequence (oriented black boxes). The two copies can then diverge (oriented gray boxes) or be maintained identical through a conversion process. Alternatively, the repeat can also be deleted, leading back to a single copy. On a long timescale, this second situation prevails. Therefore the only two ways to maintain both copies are (case 1) to move them away through chromosomal rearrangements, since the relative conversion rate then decreases (thin arrows), and (case 2) to protect them from deletion by functional pressures; the two copies are located within CDSs.

high conversion rate will tend to maintain the two copies identical) or deleted (since the deletion rate is high). As long as the repeat remains, a new exchange (conversion/deletion) is possible. Therefore, the fate of a close direct repeat is to disappear sooner or later (depending on the conversion rate vs. the deletion rate). As a consequence, only two kinds of direct repeats can be conserved on a large time scale:

1. Coding repeats, which can be conserved by functional pressures. In this case, they must be short due to the length tolerance effect. One should note, however, that strong functional pressures and/or multiple-copy repeats can lead to maintenance of large tandem clusters (rDNAs, ENA family, ASP3 cluster, CUP1 cluster, etc).
2. Repeats in which one of the two copies is moved away by an interchromosomal (not represented in fig. 4) or intrachromosomal rearrangement: an inversion will lead to inverted repeats, and an insertion will lead to distant direct repeats. Under this model, we could explain the underrepresentation of distant direct repeats by a lower level of insertion as compared to the inversion one.

This model implies that most intrachromosomal repeats originate from close direct duplications but does not preclude any mechanism. Furthermore, it gives rise to several predictions that can be experimentally tested, like a negative correlation of the deletion/conversion rate with delta.

Acknowledgments

We thank E. Rocha, J. Pothier, E. Maillier, and D. Higuet for their scientific help and their friendly support. This work was supported by grants from Association pour la Recherche sur le Cancer. E.C. and P.N. are members of the Université Pierre et Marie Curie, Paris.

LITERATURE CITED

BRITTEN, R. J. 1998. Precise sequence complementarity between yeast chromosome ends and two classes of just-subtelomeric sequences. Proc. Natl. Acad. Sci. USA **95**:5906–5912.

CHEDIN, F., E. DERVYN, R. DERVYN, S. D. EHRLICH, and P. NOIROT. 1994. Frequency of deletion formation decreases exponentially with distance between short direct repeats. Mol. Microbiol. **12**:561–569.

CHERVITZ, S. A., L. ARAVIND, G. SHERLOCK et al. (13 co-authors). 1998. Comparison of the complete protein sets of worm and yeast: orthology and divergence. Science **282**: 2022–2028.

COISSAC, E., E. MAILLIER, and P. NETTER. 1997. A comparative study of duplications in bacteria and eukaryotes: the importance of telomeres. Mol. Biol. Evol. **14**:1062–1074.

FLEISCHMANN, R. D., M. D. ADAMS, O. WHITE et al. (11 co-authors). 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae Rd.* Science **269**:496–512.

GOFFEAU, A., B. G. BARELL, H. BUSSEY et al. (16 co-authors). 1996. Life with 6000 genes. Science **274**:546–567.

JINKS, R. S., M. MICHELITCH, and S. RAMCHARAN. 1993. Substrate length requirements for efficient mitotic recombination in *Saccharomyces cerevisiae.* Mol. Cell. Biol. **13**:3937–3950.

KARLIN, S., and F. OST. 1985. Maximal segmental match length among random sequences from a finite alphabet. Pp. 225–243 *in* L. M. L. CAM and R. A. OLSHEN, eds. Proceedings of the Berkeley Conference in honour of Jerzy Neyman and Jack Kiefer. Vol. 1. Association for Computing Machinery, New York.

KARP, R. M., R. E. MILLER, and A. L. ROSENBERG. 1972. Rapid identification of repeated patterns in strings, trees and arrays. Pp. 125–126 *in* Proceedings 4th Annual ACM Symposium Theory of Computing, New York.

KLEIN, H. L. 1995. Genetic control of intrachromosomal recombination. Bioessays **17**:147–159.

LALO, D., S. STETTLER, S. MARIOTTE, P. P. SLONIMSKI, and P. THURIAUX. 1993. Two yeast chromosomes are related by a fossil duplication of their centromeric regions. C. R. Acad. Sci. **316**:367–373.

LEUNG, M. Y., B. E. BLAISDELL, C. BURGE, and S. KARLIN. 1991. An efficient algorithm for identifying matches with errors in multiple long molecular sequences. J. Mol. Biol. **221**:1367–1378.

LOUIS, E. J., and J. E. HABER. 1990. The subtelomeric Y′ repeat family in *Saccharomyces cerevisiae*: an experimental system for repeated sequence evolution. Genetics **124**:533–545.

LOVETT, S. T., T. J. GLUCKMAN, P. J. SIMON, V. J. SUTERA, and P. T. DRAPKIN. 1994. Recombination between repeats in

Escherichia coli by a recA-independent, proximity-sensitive mechanism. Mol. Gen. Genet. **245**:294–300.

MARCOTTE, E. M., M. PELLEGRINI, T. O. YEATES, and D. EISENBERG. 1998. Census of protein repeats. J. Mol. Biol. **293**:151–160.

MEWES, H. W., K. ALBERMANN, M. BAHR et al. (12 co-authors). 1997. Overview of the yeast genome. Nature **387**: 7–65.

PEETERS, B. P., B. J. DE, S. BRON, and G. VENEMA. 1988. Structural plasmid instability in *Bacillus subtilis*: effect of direct and inverted repeats. Mol. Gen. Genet. **212**:450–458.

PIERCE, J. C., D. KONG, and W. MASKER. 1991. The effect of the length of direct repeats and the presence of palindromes on deletion between directly repeated DNA sequences in bacteriophage T7. Nucleic Acids Res. **19**:3901–3905.

PRYDE, F. E., H. C. GORHAM, and E. J. LOUIS. 1997. Chromosome ends: all the same under their caps. Curr. Opin. Genet. Dev. **7**:822–828.

ROCHA, E. P. C., A. DANCHIN, and A. VIARI. 1999. Analysis of long repeats in bacterial genomes reveals alternative evolutionary mechanisms in *Bacillus subtilis* and other competent prokaryotes. Mol. Biol. Evol. **16**:1219–1230.

SCHNEIDER, T. D., G. D. STORMO, L. GOLD, and A. EHRENFEUCHT. 1986. Information content of binding sites on nucleotide sequences. J. Mol. Biol. **188**:415–431.

SEMPLE, C., and K. H. WOLFE. 1999. Gene duplication and gene conversion in the *Caenorhabditis elegans* genome. J. Mol. Evol. **48**:555–564.

SMITH, T. F., and M. S. WATERMAN. 1981. Identification of common molecular subsequences. J. Mol. Biol. **147**:195–197.

VINCENS, P., L. BUFFAT, C. ANDRE, J. P. CHEVROLAT, J. F. BOISVIEUX, and S. HAZOUT. 1998. A strategy for finding regions of similarity in complete genome sequences. Bioinformatics **14**:715–725.

VISWANATHAN, M., G. MUTHUKUMAR, Y. S. CONG, and J. LENARD. 1994. Seripauperins of *Saccharomyces cerevisiae*: a new multigene family encoding serine-poor relatives of serine-rich proteins. Gene **148**:149–153.

WOLFE, K. H., and D. C. SHIELDS. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. Nature **387**:708–713.

MANOLO GOUY, reviewing editor

Accepted May 2, 2000

A.2. Résumé des résultats.

Au cours de cette étude, nous avons mis en évidence 275 répétitions directes et 340 répétitions inversées dans le génome de la levure. A titre de comparaison, nous avons détecté en moyenne 25 répétitions directes et 24 inversées dans les génomes aléatoires. La distribution des répétitions dans les chromosomes n'est pas uniquement fonction de la taille des chromosomes. Les variations de nombre de répétitions sont probablement la trace des histoires évolutives des chromosomes. Dans cette étude, nous nous sommes focalisés sur les caractéristiques globales de ces répétitions.

Nous nous sommes attachés à trois paramètres des répétitions : leur longueur (moyenne de la longueur des deux copies), l'identité et l'espace physique entre les deux copies ; cet espace est nommé *spacer*. Les distributions de ses trois paramètres sont présentés sur la figure 1 de l'article 1. La comparaison de ces distributions fait ressortir plusieurs observations importantes :

- Dans les génomes aléatoires, les distributions entre les divers paramètres sont similaires entre répétitions directes et répétitions inversées. Cela montre que la méthode n'induit pas de différence dans les caractéristiques des deux types de répétitions.
- Dans le génome réel, les grandes répétitions sont le plus souvent des répétitions directes. Il s'agit pour les plus grandes d'entre elles de gènes répétés comme la famille des gènes ENA (ou PMR2) répétés en tandem sur le chromosome IV. Les répétitions directes ont une identité plus forte que les répétitions inversées. Il existe beaucoup de répétitions directes proches (dont le *spacer* est inférieur à 1 kilobase), qui représente la moitié des répétitions directes. Ces répétitions proches sont absentes dans les répétitions inversées et dans les génomes aléatoires. Ces répétitions directes proches seront nommées, par la suite, CDR (*Close Direct Repeats*).

Nous avons recherché des relations entre ces trois paramètres. Pour les CDR uniquement, nous avons observé des corrélations entre identité, longueur et taille du *spacer*.

Plus précisément, l'identité est négativement corrélée avec la taille du *spacer* (article 1, figure 2) et la longueur est positivement corrélée à la taille du *spacer* (article 1, figure 3). Ces corrélations sont absentes des génomes aléatoires et des répétitions inversées.

Le dernier résultat concerne la localisation des répétitions par rapport à celle des gènes. La plupart des CDR (85 %) sont localisées dans des gènes. Plus précisément, pour 83% des CDR, les deux copies sont dans le même gène et codent pour des répétitions peptidiques dans la même protéine. A l'inverse, les répétitions directes éloignées et les répétitions inversées sont moins souvent localisées dans les gènes (respectivement 60 % et 40 %) et donnent peu lieu à des répétitions peptidiques (respectivement 29% et 14%).

A.3. Eléments de discussion.

L'analyse des répétitions intrachromosomiques de la levure *S. cerevisiae* a mis en évidence un certain nombre de caractéristiques biologiques que nous avons tenté d'interpréter.

En premier lieu, il est possible d'expliquer la présence des corrélations entre *spacer*, identité et longueur pour les CDR. Deux principaux mécanismes ciblent les répétitions : la recombinaison homologue et la recombinaison non homologue (*Introduction, chapitre A*). Ces mécanismes d'échange ont, pour les CDR, deux types de conséquence : la délétion mène à la perte d'une copie et du *spacer* (recombinaison réciproque) et la conversion (recombinaison non-réciproque) homogénéise les deux copies de la CDR. L'hypothèse que nous avançons pour expliquer ces corrélations est que la force de ces mécanismes (le taux d'échange) décroît quand la taille du *spacer* augmente. Des études, chez les bactéries *B. subtilis* (Chedin *et al.* 1994) et *E. coli* (Bi and Liu 1994; Lovett *et al.* 1994), montrent que le taux de délétion est négativement corrélé au *spacer*. Aucune de ces études ne s'est intéressée à la conversion pourtant étroitement liée aux mécanismes de crossing-over (la délétion peut se faire par crossing-over et/ou par un accident au cours de la réplication). Si notre hypothèse est valable, conversion et délétion ciblent potentiellement les CDR, avec d'autant plus d'efficacité que les deux copies des CDR sont proches.

Si un événement de conversion a lieu, les deux copies se retrouvent homogénéisées sur toute la longueur de l'échange. Les CDR, dont le *spacer* est petit, seraient plus souvent soumises à ces conversions, et tendraient à être maintenues plus identiques.

Si un événement de délétion se produit, la répétition est perdue. Or, il a été montré, chez *S. cerevisiae* (Jinks-Robertson *et al.* 1993) et chez *E. coli* (Peeters *et al.* 1988), que la fréquence de recombinaison entre deux copies est positivement corrélée à la longueur de ces copies. Le facteur longueur se combine alors à celui de la taille du *spacer*. On peut donc supposer que les grandes CDR dont le *spacer* est petit sont très instables et ne sont que très difficilement conservées dans les génomes. A l'inverse, quand le *spacer* est plus grand, les grandes répétitions sont mieux tolérées. Ceci pourrait expliquer la corrélation positive observée entre la longueur et la taille du *spacer*.

Nous proposons donc que les corrélations observées sont le reflet de contraintes structurales imposées aux CDR. Afin de mieux comprendre ces contraintes, il faut les replacer dans un contexte dynamique. A chaque instant, une CDR peut être soit ne pas être ciblée, soit la cible d'un événement de délétion, soit celle d'un événement de conversion. Dans le cas de la conversion, les deux copies sont homogénéisées et un nouvel échange est encore possible. En revanche, lorsqu'une délétion se produit, la répétition est perdue. Il existe donc un biais en faveur de la délétion ; les CDR devraient être perdues tôt ou tard. Alors comment expliquer cette abondance de CDR ?

Deux explications peuvent être envisagées : (1) il y a un fort taux de création de CDR qui contrebalance le fort taux de délétion et/ou (2) les CDR sont protégées de la délétion par des pressions fonctionnelles. La seconde explication semble confirmée par nos observations sur la localisation des CDR par rapport à celles des gènes, mais ne semble pas expliquer à elle seule l'abondance de ce type de répétition. En effet, il parait peu probable que seule les CDR soient protégées par les pressions fonctionnelles. Ces pressions peuvent donc expliquer comment les CDR sont maintenues, mais pourquoi elles sont si nombreuses. On peut donc envisager que la première explication cohabite avec la seconde.

La création massive de CDR expliquerait pourquoi l'on trouve les plus grandes répétitions (supposées plus récentes) dans les répétitions directes. Parmi les répétitions inversées, il n'y pas de répétitions proches. Cela suggère que les répétitions inversées ne sont pas créées proches. Les répétitions inversées et les répétitions directes éloignées sont globalement plus petites et moins identiques que les CDR, ce qui suggère qu'elles sont plus anciennes.

Nous proposons un modèle qui résume toutes les propriétés observées (article 1, figure 4). Dans ce modèle, il y a une création continue de CDR, qui sont ensuite soumises à d'importants taux de conversion et de délétion. Ces derniers tendent à les homogénéiser et à les faire disparaître. Cependant, deux voies peuvent sauvegarder une répétition :

- Elle peut être maintenue par des pressions fonctionnelles (contraintes sélectives),
- Elle peut se « transformer » en répétition éloignée (inversée ou non) par un ou des réarrangements chromosomiques. Ces derniers peuvent être des insertions, inversions ou des réarrangements interchromosomiques. Les deux copies ainsi écartées, sont alors peu soumises aux fort taux de conversion/délétion et peuvent diverger sans contrainte.

A.4. Les répétitions subtélomèriques

Dans notre étude, nous avons soustrait systématiquement toutes les répétitions dont l'une des deux copies était située dans un subtélomère. Ces subtélomères sont les régions situées juste en aval des télomères (sur quelques dizaines de kilobases). La structure intrinsèque de ces régions subtélomèriques a été soigneusement décrite (Louis 1995).

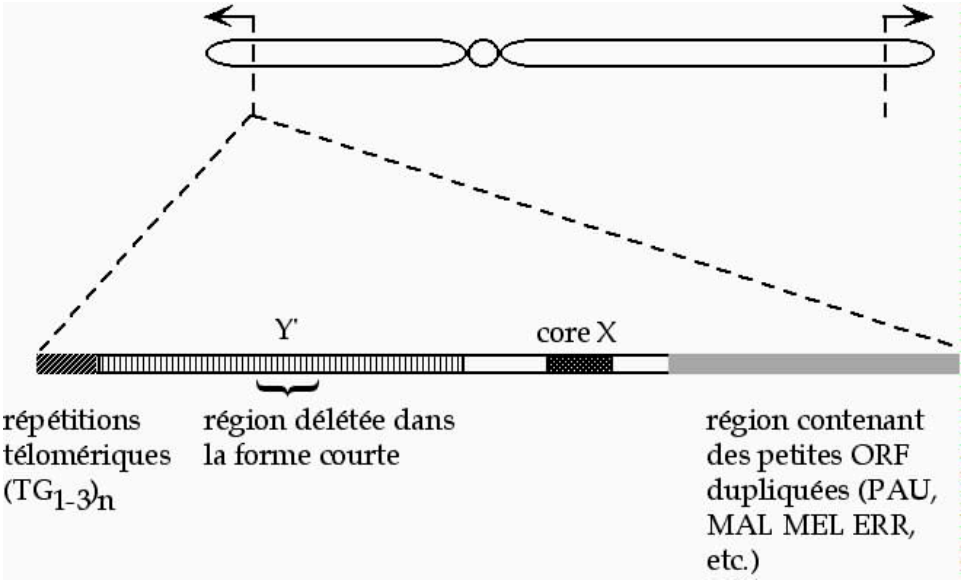

Figure 35 : Structure d'un subtélomère chez *S. cerevisiae*. *D'après (Louis 1995).*

Grossièrement, le télomère de *S. cerevisiae* est constitué depuis sa partie la plus distale (figure 35) :

a) de répétitions télomériques ajoutées par la télomèrase, une transcriptase inverse apparentée à celles des rétroéléments (Eickbush 1997).

b) d'une région susceptible de contenir une ORF nommée Y'. Cette ORF existe sous deux formes qui varient par une insertion/délétion : une forme courte et une forme longue. Ces éléments Y' sont présents en un nombre variable de copies allant de 0 à 4 par subtélomère selon les souches. Ils sont, de plus, le lieu de nombreuses conversions et de nombreux crossing-over (Louis and Haber 1990). Ces éléments contiennent un minisatellite interne de 36 paires de bases.

c) d'une région, nommée 'core X', contenant une ARS (*Autonomous Replication Sequence*). Cette région présente des divergences de 10 à 20 %. Aucune fonction ne lui a été encore attribuée, bien qu'elle semble contribuer à la stabilité de ségrégation (Enomoto *et al.* 1994). RJ Britten, à travers une étude des répétitions subtélomèriques, a suggéré de redéfinir une région plus large du 'core X' qu'il nomme X2 (Britten 1998).

d) d'une région présentant des ORF dupliquées. Ces ORF sont de plusieurs types et sont souvent présentes uniquement dans les subtélomères. Ces petites ORF dupliquées ont

été en partie décrites par RJ Britten (Britten 1998). Un exemple de ces petites ORF est la famille des séripaupérines (PAU).

En interphase, les 64 télomères de *S. cerevisiae* sont groupés, au niveau de la membrane nucléaire, en 6 à 8 foci distincts (Gotta *et al.* 1996). Au sein de ces groupes, ils sont organisés en bouquet et pourraient alors avoir des échanges préférentiels (Pryde *et al.* 1997). Ce type d'organisation est également retrouvé chez *H. sapiens* (Griffith *et al.* 1998), mais pas en périphérie du noyau. Cette organisation est également associée, chez *H. sapiens*, à un fort taux d'échange (Coleman *et al.* 1999).

A l'aide des répétitions, nous avons essayé de voir si les télomères se regroupent toujours dans les même foci. Si des télomères sont localisés toujours dans le même focus, les échanges génétiques devraient être favorisés entre certains télomères. Ces échanges créant parfois des répétitions, on devrait retrouver ces répétitions interchromosomiques en plus grand nombre entre les subtélomères d'un même focus. Nous avons donc étudié ces répétitions entre les différents subtélomères. Sur la base de la densité des répétitions interchromosomiques, nous n'avons pas pu mettre en évidence des groupes bien définis de télomères, seuls les subtélomères ayant des Y' formaient un groupe cohérent. Pour mieux comprendre la répartition des répétitions dans les subtélomères, nous avons cartographié les zones répétées entre les subtélomères. Sur la figure 36, est présenté un exemple de la répartition des répétitions subtélomèriques. D'autres subtélomères sont uniques, ou présentent un profil de répétitions plus complexe. Néanmoins, cela suggère que les subtélomères de *S. cerevisiae* sont construits par des additions successives d'autres régions subtélomèriques. Il semble ainsi raisonnable de penser que, bien que le taux de recombinaison entre certains subtélomères soit plus élevé (Eyre *et al.* 1999), les télomères ne sont pas toujours groupés dans les même foci.

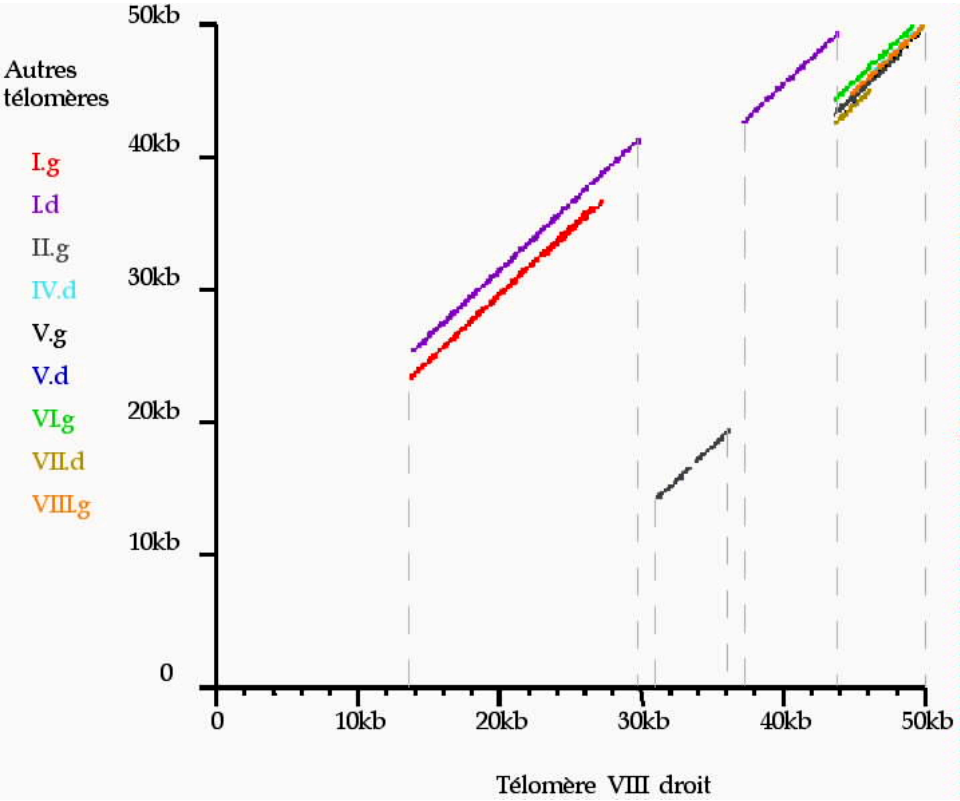

Figure 36 : Répétitions partagées par le subtélomère VIII droit avec les autres subtélomères.
Les 50 kilobases terminaux du bras droit du chromosome VIII ont été comparées avec les autres subtélomères et les répétitions ont été détectées. Ce subtélomère est répété avec de nombreux autres. 50 kb correspond à l'extrémité du télomère et 0 kb au début potentiel du subtélomère (50 kb vers le centromère)

B. Les duplications intrachromosomiques dans les génomes eucaryotes.

L'analyse des répétitions chez *S. cerevisiae* nous ayant permis de proposer un modèle d'évolution des séquences répétées intrachromosomiques, nous avons entrepris la recherche et l'analyse des répétitions dans les chromosomes eucaryotes dont la séquence était complète et publique. La question sous-jacente à cette seconde étude concerne la généralisation de notre modèle à d'autres chromosomes eucaryotes : les caractéristiques observées chez *S. cerevisiae* sont-elles une particularité de ce génome ou existent-elles aussi dans les autres chromosomes eucaryotes ? Les chromosomes que nous avons analysés se repartissent dans deux génomes complets (*S. cerevisiae* et *C. elegans*) et quatre génomes partiels (deux chromosomes de *Plasmodium falciparum*, deux chromosomes de *Arabidopsis thaliana*, six bras chromosomiques de *Drosophila melanogaster* et deux chromosomes de *H. sapiens*).

D'autres questions surgissent à l'esprit lorsque l'on entreprend la comparaison des répétitions de génomes de plusieurs espèces. L'une d'entre elles est la suivante : quelles sont les caractéristiques qui permettent de différencier chaque génome et quelles sont celles qui les rassemblent ? Dans cette idée nous avons cherché à travers cette seconde étude l'existence potentielle d'un « style d'amplification » propre à chaque génome.

La détection des répétitions dans des chromosomes aussi variés que les chromosomes 21 et 22 de *H. sapiens* (environ 35 Mb et très répétés) et les chromosomes de *S. cerevisiae* (individuellement souvent inférieure à 1Mb et peu répétés) a conduit à changer la méthode de détection. En effet, l'application de la méthode précédente donnant un nombre de répétitions exactes trop grand, il a fallu réduire le nombre de répétitions exactes pour la phase d'extension. Cette réduction s'est principalement opérée sur deux points :

- Une augmentation de la longueur minimum des répétitions exactes détectées.
- Le filtrage de toutes les répétitions multicopies. Nous n'avons conservé que les répétitions exactes présentes à deux copies dans le chromosome.

A ce niveau de stringence, aucune répétition n'est détectée dans les génomes aléatoires. Nous ne disposons donc plus de témoin aléatoire.

B.1. Article 2

Study of Intrachromosomal Duplications Among the Eukaryote Genomes

Guillaume Achaz, Pierre Netter, and Eric Coissac

Structure et Dynamique des Génomes, Institut Jacques Monod, Paris, France

Complete eukaryote chromosomes were investigated for intrachromosomal duplications of nucleotide sequences. The analysis was performed by looking for nonexact repeats on two complete genomes, *Saccharomyces cerevisiae* and *Caenorhabditis elegans*, and four partial ones, *Drosphila melanogaster*, *Plasmodium falciparum*, *Arabidopsis thaliana*, and *Homo sapiens*. Through this analysis, we show that all eukaryote chromosomes exhibit similar characteristics for their intrachromosomal repeats, suggesting similar dynamics: many direct repeats have their two copies physically close together, and these close direct repeats are more similar and shorter than the other repeats. On the contrary, there are almost no close inverted repeats. These results support a model for the dynamics of duplication. This model is based on a continuous genesis of tandem repeats and implies that most of the distant and inverted repeats originate from these tandem repeats by further chromosomal rearrangements (insertions, inversions, and deletions). Remnants of these predicted rearrangements have been brought out through fine analysis of the chromosome sequence. Despite these dynamics, shared by all eukaryotes, each genome exhibits its own style of intrachromosomal duplication: the density of repeated elements is similar in all chromosomes issued from the same genome, but is different between species. This density was further related to the relative rates of duplication, deletion, and mutation proper to each species. One should notice that the density of repeats in the X chromosome of *C. elegans* is much lower than in the autosomes of that organism, suggesting that the exchange between homologous chromosomes is important in the duplication process.

Introduction

All eukaryote genomes exhibit similar physical structures and constraints (i.e., linear chromosomes, scaffold attachment, nucleosome organization). However, many characteristics highlight important differences between them: (1) coding sequences represent 72% of the *Saccharomyces cerevisiae* genome (Goffeau et al. 1996) and only 2%–5% of the *Homo sapiens* genome (Dunham et al. 1999), (2) a centimorgan corresponds to kilobases in *S. cerevisiae* (Baudat and Nicolas 1997) and to megabases in humans (Dunham et al. 1999), (3) the number of introns per gene and the density of transposons increase with the genome size, and (4) the isochores organization has been ascribed mostly to vertebrate genomes (Bernardi 2000). Despite these differences, one would expect to find remnants of a similar nuclear organization in genome sequences. Events of DNA duplication were described in many eukaryote genomes, but are the duplication dynamics similar in all eukaryotes?

Within duplication events, four main subprocesses have been documented: abnormal segregation during cell division (leading to entire-chromosome[s] duplication, viz., hyperploidization, and sometimes to whole-genome doubling, viz. polyploidization), transposition (duplication of transposable elements), expansion of low-complexity sequences (microsatellites and minisatellites), and finally generic duplications of unspecific DNA regions within the same chromosome or between two chromosomes. We shall henceforth refer to this last subprocess as the iteration process. Polyploidization events were proposed to explain the large-scale duplications at the origin of vertebrates (Ohno 1970), in many angiosperms (Masterson 1994)—even in *Arabidopsis thaliana* (Blanc et al. 2000), in the fish lineage (Amores et al. 1998), and in the yeast *S. cerevisiae* (Wolfe and Shields 1997). However, it is not clear if these large-scale duplications are always the result of polyploidization, successive hyperploidizations, or bursts of large iterations (Holland 1999; Llorente et al. 2000; Vision, Brown, and Tanksley 2000; Hughes, Da Silva, and Friedman 2001; Robinson-Rechavi et al. 2001).

In order to investigate the iteration process, we focused our attention on intrachromosomal repeats in the chromosome sequences. Two complete genomes, *S. cerevisiae* and *Caernorhabditis elegans*, and four partial ones, *H. sapiens*, *Drosophila melanogaster*, *A. thaliana*, and *Plasmodium falciparum*, were analyzed. It should be noted that the genome of *S. cerevisiae* was already investigated for its repeats in a previous study (Achaz et al. 2000) in which we proposed a model for the dynamics of the iteration process based on a continuous genesis of close direct repeats (CDR). A CDR is defined here as a repeat with its copies in the same orientation and with a physical distance between them (the spacer) smaller than 1 kb. The model supposes that most of the intrachromosomal repeats originate from these CDRs, the others being the result of further chromosomal rearrangements. In the present study, the model established in yeast was tested for new eukaryote chromosomes. We focused on the differences between genomes and tried to connect them to the genome context. In our model, supposing that most of the intrachromosomal repeats originate from tandem repeats, the chromosome sequences had been investigated to find the remnants of the chromosomal rearrangements. Hence, we view repeats as the markers of genome dynamics.

Materials and Methods

Data

We analyzed the complete eukaryote genomes of *S. cerevisiae*—16 chromosomes—(Goffeau et al. 1996)

Key words: genome dynamics, evolution, duplication, eukaryotes.

Address for correspondence and reprints: Guillaume Achaz, Structure et Dynamique des Génomes, Institut Jacques Monod, Tour 43–44, 1° Étage, 4, Place Jussieu, 75251 Paris Cedex 05, France. E-mail: achaz@ijm.jussieu.fr.

Mol. Biol. Evol. 18(12):2280–2288. 2001

Table 1
Estimation of the Intrachromosomal Redundancy of Each Genome

Species	Chromosome Mean Size (Mb)	Two-copy Seeds[a] (% bp)	All Seeds[b] (% bp)
S. cerevisiae	0.75	1.77	3.83
P. falciparum	1.01	4.67	9.04
A. thaliana	18.60	4.83	10.13
C. elegans.	15.87	13.60	22.19
D. melanogaster.	19.07	1.10	3.22
H. sapiens.	33.65	3.90	18.68

[a] The proportion (as a percentage of the total base pairs) of the analyzed chromosomes included in the two-copy seeds (exact repeats). It should be noted that only two-copy seeds are kept for further analysis.
[b] The proportion of the analyzed chromosomes included in all seeds.

and *C. elegans*—six chromosomes—(Consortium 1998), and four partial genomes: *H. sapiens*—chromosomes 21 (Hattori et al. 2000) and 22 (Dunham et al. 1999), *P. falciparum*—chromosomes 2 (Gardner et al. 1998) and 3 (Bowman et al. 1999), *A. thaliana*—chromosomes 2 (Lin et al. 1999) and 4 (Mayer et al. 1999), and six chromosomal arms (X, 2L, 2R, 3L, 3R, 4) of *D. melanogaster* (Adams et al. 2000).

Sequences of *H. sapiens*, *C. elegans*, *P. falciparum*, and *A. thaliana* were extracted from GenBank (ftp://ncbi.nlm.nih.gov/genbank/genomes). The *S. cerevisiae* chromosomes were extracted from Saccharomyces Genome Database (http://genome-www.stanford.edu/Saccharomyces). Sequences of *D. melanogaster* were downloaded from Celera database (http://www.celera.com).

It should be pointed out that most sequences contain many gaps (stretches of N). For example, in chromosome 1 of *C. elegans*, 8.8% of its base pairs are N, and 29 gaps are longer than 10 kb. These stretches were not taken into account during the construction of the repeats' database.

Construction of the Repeats Database

General trends of repeats detection, like most of the heuristics already proposed (Leung et al. 1991; Vincens et al. 1998), are based on looking first for seeds (exact repeats) and then extending them with a local alignment program. The detailed methodology is described below through three main steps: searching, filtering, and extending.

First Step: Searching for Seeds

In this step, exact repeats (seeds) were detected by using the REPuter software (Kurtz and Schleiermacher 1999). This software detects all seeds (direct and inverted) in a given sequence that are any distance apart from the chromosome. As we are interested in unusually large seeds, the minimum length of seeds (L_{min}) was calculated using the statistics developed by Karlin and Ost (1985). For each chromosome, we chose L_{min} such that the probability of finding a two-copy word with at least this length in a same-size, same-nucleotide composition random sequence is 0.001. Typically, L_{min} ranges from 21 for the smallest chromosome (chromosome 1 of *S. cerevisiae*) to 28 for the largest ones (chromosomes 21 and 22 of *H. sapiens*).

Second Step: Filtering the Seeds

First, to remove all low-complexity seeds (i.e., microsatellites or poly-A stretches), we used an entropy filter based on dinucleotide composition (Achaz et al. 2000). Second, all multicopy seeds were removed. A chromosome map in which each position is linked to its n-plication degree (duplication, triplication, etc.) was established. To build this map, we counted for each chromosome position the number of times this position is found in seeds (direct and inverted seeds were pooled together). This map is used to estimate the degree of redundancy of chromosomes (i.e., the number of duplications, triplications, etc.). Table 1 presents, for each species, the mean size of the chromosomes, the percentage of chromosomes included in two-copy seeds, and the percentage of chromosomes represented by all the seeds. As we are only interested here in two-copy seeds, we used the map to remove all seeds in which one of the positions is included in a multicopy repeat.

Third Step: Extending the Seeds

Seeds were extended into larger nonstrict repeats by using a local alignment program (Smith and Waterman 1981) available at http://www-hto.usc.edu/software/seqaln. It should be pointed out that many seeds might give rise to the same extended repeat. Therefore, when two or more repeats occurred in the same location, we just kept the first one. Before the alignment is performed, 100 bp were picked up on both sides of the seeds. Thus, for a given seed of size N, the first alignment is computed with two sequences of $2 \times 100 + N$ bp. The following matrix, which was built empirically, was retained for local alignments: $match_{(A/T/C/G)} = 4$, mismatch $= -4$, $Gap_{open} = -16$, and $Gap_{extension} = -4$. When the best local alignment ends at less than 10 bp of a terminus, 200 bp were added at the termini, and a new run was done. As the alignment of a large sequence requires too many computer resources, we devised the following heuristic to compute the alignment of large sequences. If the alignment size was more than 1 kb, the partial alignment was memorized, and only the rest of the alignment was computed in a new

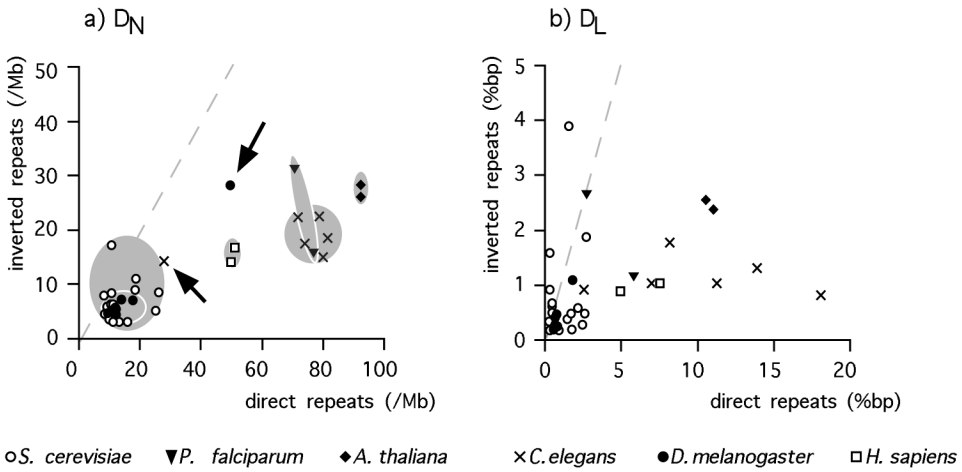

FIG. 1.—Occurrence and density of direct and inverted repeats in each chromosome. For each chromosome of each species, inverted repeats were compared to direct repeats. (*a*) Plot of D_N (density in number: number divided by chromosome length) of inverted repeats as a function of D_N of direct repeats. Chromosomes of the same species are grouped in gray areas, with the exception of the X chromosome of *C. elegans* and the fourth chromosomal arm of *D. melanogaster*, both indicated by black arrows. (*b*) Plot of D_L (density in length: sum of repeats length divided by chromosome length) of inverted repeats as a function of D_L of direct repeats. Each species is represented by a different symbol given just below the plot.

run. The process goes on until both sides of the complete alignment end at more than 10 bp of the termini. Thus, it provides a nonoptimal alignment but allows us to extend very large repeats. Then we removed all repeats in which the copies overlap because they generally correspond, at this stage, to three-copy repeats.

It should be mentioned that the methodology was similar to the one previously used in the *S. cerevisiae* analysis (Achaz et al. 2000), but was modified in order to analyze in the same way the chromosomes of yeast (<1.5 Mb) and man (35 Mb). The major modifications were applied to reduce the number of seeds and to keep only *sensu stricto* duplicated seeds (present only in two copies) for the alignment process.

Results and Discussion

The application of the methodology described above yields for direct and inverted repeats, respectively: 110 and 75 for *S. cerevisiae*, 136 and 48 for *P. falciparum*, 2,407 and 1,068 for *A. thaliana*, 6,885 and 1,845 for *C. elegans*, 1,479 and 691 for *D. melanogaster*, and 3,457 and 2,406 for *H. sapiens*.

Genome Style and History of Chromosomes

In order to analyze the relationship between chromosome size and redundancy level, we measured two parameters D_N and D_L, defined as follows:

$$D_N = \frac{\text{Number of repeats}}{\text{Size of chromosome}}$$

$$D_L = \frac{\sum \text{Length of repeats}}{\text{Size of chromosome}}$$

As predicted by the estimated redundancy of the genome in Table 1, if we exclude the Drosophila chromosomal arms (which are clearly underrepeated for their size), D_N and D_L are positively correlated with the chromosome size, using a Kendall tau-rank test ($\tau = 0.30$, $P < 0.05$ for D_N and $\tau = 0.40$, $P < 0.01$ for D_L). These observations are in agreement with an analysis of gene redundancy undertaken on partial genome sequences (Coissac, Maillier, and Netter 1997).

Two hypotheses can be proposed to explain the low densities of the chromosomal arms of *D. melanogaster*. The first one is a data bias: it should be noted that the analyzed sequences are constituted exclusively of euchromatine (only around two-thirds of the complete genome), and it is known that repeats are concentrated inside heterochromatine (Henikoff 2000). Moreover, assembly errors could lead to artificially deleted tandem repeats. The second hypothesis rests on biological grounds. One can imagine that Drosophila's genome has a special status in the duplication process (because there is no meiotic crossing-over in the male, the duplication process can be less active). The achievement of the complete sequence of *D. melanogaster* should solve this problem.

In order to investigate more precisely each chromosome, we analyzed D_N and D_L for direct and inverted repeats (fig. 1). It appears that D_N is similar for all chromosomes within the same species, whereas D_L is not. Thus, D_N could define the style of redundancy of the genome. We assume that D_N results from the iteration events combined with the loss of duplicated sequences, and then propose D_N to be connected to the biological machinery of each species. Because the machinery is clearly different for each species, but similar for all chromosomes within the same genome, D_N should be the consequence of each genome's dynamics. Furthermore, the differences between species come essentially from direct repeats, and less from inverted repeats. This suggests that the biological machinery is more connected to the creation and the loss of direct repeats than to the dynamics of inverted repeats.

The only two exceptions are the fourth chromosomal arm of *D. melanogaster* and the X chromosome of *C.*

elegans. The high density of the small fourth chromosomal arm of *D. melanogaster* could be the result of its particular structure (if there is no data bias): it is mostly constituted of heterochromatin, but, contrary to centromeric chromatin (or Y chromosome), it is partially visible in polytene chromosomes. On the contrary, the X chromosome of *C. elegans* exhibits a lower D_N than that of the other worm's chromosomes. This observation is in good agreement with the unequal distribution of repetitive elements, such as CeRep23 (Barnes et al. 1995), Cele1, Cele2, and Cele42 (Surzycki and Belknap 2000), between the autosomes and the X chromosome in *C. elegans*. It should be pointed out that exchanges between the homologous X chromosomes are only possible in hermaphrodite XX (males are X0), which could explain this lower D_N. If this is true for *C. elegans*, one may expect this to be true for all heterochromosomes. The X chromosomal arms of *D. melanogaster* seem similar to the other chromosomal arms; however, none of the Drosphila male chromosomes is submitted to meitoic crossing-over.

Contrary to D_N, D_L could reflect better the chromosome history than the effects of the cellular machinery: a unique event of iteration can lead to a high D_L for direct or inverted repeats. For example, direct repeats of the chromosome 1 of *C. elegans* exhibit a high D_L and a normal D_N (when compared with the other *C. elegans* chromosome values). This particularity is mainly caused by two large duplicated sequences, one 250-kb long (with an identity of 98.7%) and the other 600-kb long (fractionated into several segments of high identity, often more than 99%). Furthermore, the inverted repeats of the chromosome 1 of *S. cerevisiae* show a high D_L and a normal D_N, as a consequence of two internal regions inversely repeated in subtelomeres (Britten 1998).

A Model of Dynamics of Iteration

Our model of intrachromosomal iteration (Achaz et al. 2000) is based on a permanent genesis of CDR. The CDRs are then submitted to a high level of exchange (conversion and deletion). This high exchange rate tends to maintain the two copies identically (conversion) and also to eliminate them (deletion). At each round of exchange, both events are possible, but whereas conversion may still be followed by deletion, a deletion event cannot be followed by conversion.

Therefore, on a long timescale, a bias in favor of deletion should be observed. A CDR has to disappear sooner or later (depending on the relative rates of conversion and deletion). However, there are two situations where a repeat would be maintained: when it is protected from deletions by functional pressures (i.e., located inside a gene) or when the copies are spaced by further chromosomal rearrangements. This model was mainly based on three observations for CDR: (1) they are overrepresented, (2) they are mostly located inside the same gene, and (3) their length is positively correlated with the spacer (the physical distance between copies), and their identity is negatively correlated with it.

Through the present analysis, the model was tested with other eukaryotes. It should be mentioned that a model of tandem creation and further dispersion was already invoked for the families of two genes (*Hox* and *NBG*) in *C. elegans* (Ruvkun and Hobert 1998). The annotations of eukaryote chromosomes being partial, they were not taken into account. Thus, we did not analyze the relation of repeats position with genes location.

CDRs Are Overrepresented

The repartitions of spacer size for direct and inverted repeats (fig. 2) reveal that CDRs are overrepresented as compared with close inverted repeats. Moreover, in the previous study (Achaz et al. 2000), the repeats of *S. cerevisiae* were compared with the repeats that issued from random chromosomes. From this comparison, we showed that such close repeats (inverted or direct) are absent from random chromosomes. This strongly suggests that these CDRs are not the result of chance. The presence of many CDRs in all chromosomes is in good agreement with the model.

However, the repartition spacer's length indicates the existence of many direct repeats with a spacer between 1 and 10 kb in *A. thaliana* chromosomes (they represent more than one-third of all direct repeats). We looked for a plausible explanation for this overrepresentation in *A. thaliana* (as compared with other species), with particular attention to the sequence located between the two copies (the spacer). Several hypotheses can be envisaged and rejected: (1) repeats are not the edges of transposons because the spacers are not paralogous, (2) the hypothesis of campbell-like insertions of exogenous DNA, as it was proposed for *B. subtilis* (Rocha, Danchin, and Viari 1999*a*), can be eliminated because there is no difference in nucleotide composition between spacers and chromosomes, (3) there is no clear difference between these repeats and others—no high identity level, no special length, no special physical location. Similar observations can be established for the genome of *C. elegans* and *D. melanogaster*, where direct repeats with a spacer between 1 and 10 kb are also overrepresented.

In conclusion, we did not yet find any plausible hypothesis to understand why these repeats are overrepresented.

CDRs Are Identical and Short

We started by characterizing CDRs in terms of the distribution of their identity (fig. 3). Except for *S. cerevisiae*, CDRs have their two copies more identical than distant direct repeats ($P < 10^{-4}$, Mann-Whitney rank test).

In order to explain the *S. cerevisiae* exception, one should take into consideration that the distinction between close and distant repeats has been arbitrarily fixed at the same spacer size (1 kb) for each organism. The biological difference between close and distant repeats is connected to the recombination machinery. As this machinery varies from yeast to human, the limit between close and direct repeats should not be identical for all species. In that way, it can be shown that for *S. cerevisiae*, direct repeats with a spacer smaller than 500

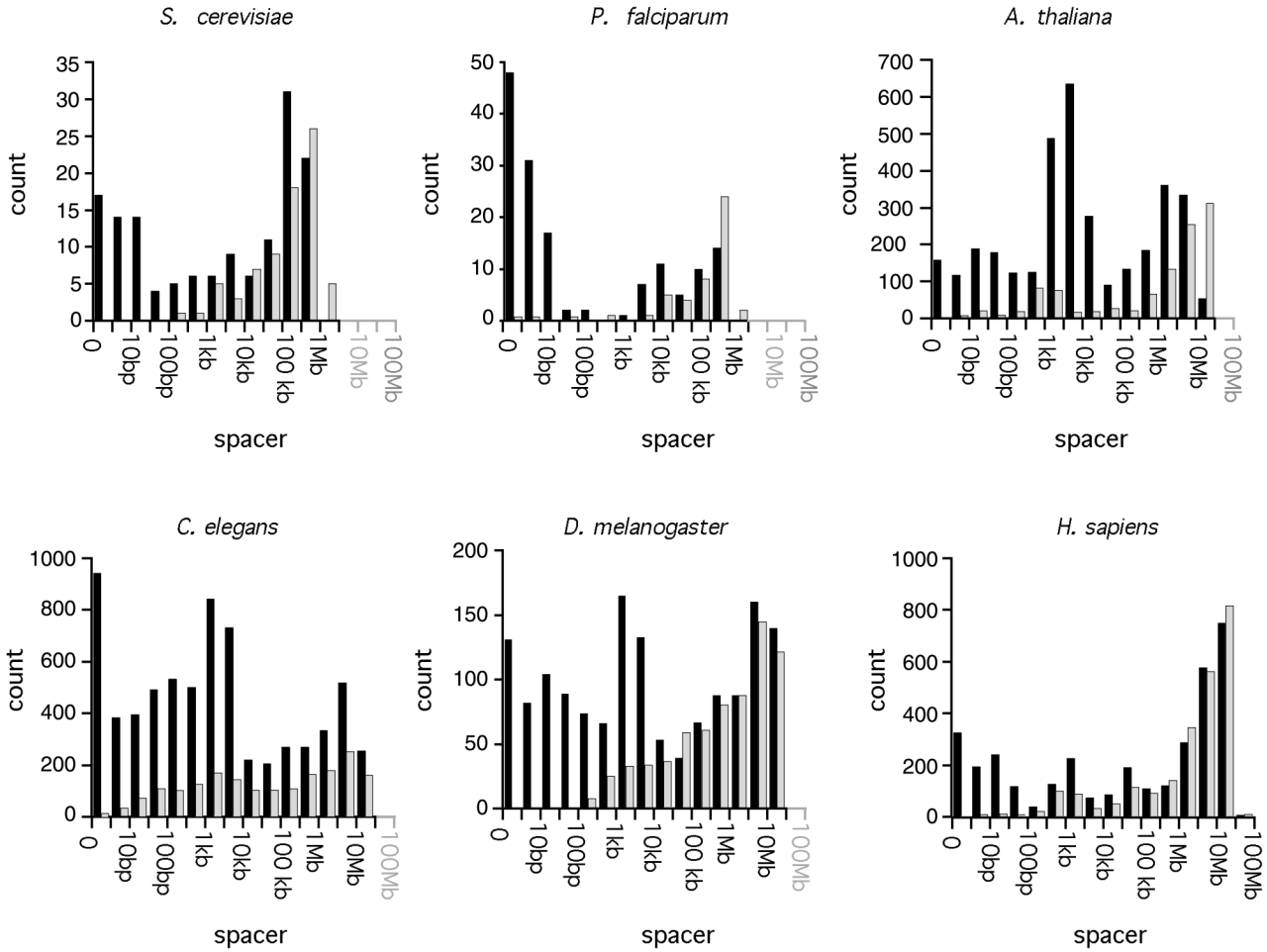

FIG. 2.—Distribution of spacers for each orientation (direct and inverted). Each histogram shows the distribution of spacers (the distance between the two copies) for direct and inverted repeats. Black boxes represent the direct repeats, and gray boxes represent the inverted ones. The distribution is established by the log 10 of the spacer size (in steps of 0.5). These histograms show clearly that CDRs (direct repeats with a spacer smaller than 1 kb) are overrepresented in all species.

bp are more identical than other direct repeats ($P < 0.05$, Mann-Whitney rank test).

This greater similarity could be explained, on the one hand, by the recent origin of these repeats and, on the other, by a high conversion rate between the two copies when they are close together. As previously discussed, CDR could also be submitted to a high deletion rate. It has been reported that recombination rate is positively correlated with repeat length in yeast (Jinks, Michelitch, and Ramcharan 1993) and in mammalian cells (Rubnitz and Subramani 1984). Thus, CDRs with long copies are too unstable to persist, and only small CDRs are conserved. In order to test this hypothesis, the length distributions of close and distant direct repeats were compared: it appeared that CDRs are smaller than the distant ones ($P < 10^{-4}$, Mann-Whitney rank test).

CDRs Exhibit an Exchange Rate Negatively Correlated with the Spacer Size

We previously observed a positive rank correlation between length and spacer and a negative rank correlation between identity and spacer for CDR in yeast: the closer the repeats, the more identical and shorter they are. Except for the *P. falciparum* chromosomes, correlations between identity, length and spacer were found in all eukaryotes (Table 2). This is in good agreement with an observation reported in *C. elegans* that the similarity between paralogous genes is negatively correlated with the physical distance between them (Semple and Wolfe 1999).

In order to understand such a result, we proposed that, as in bacteria (Peeters et al. 1988), the exchange rate between the two copies is negatively correlated with the spacer size. A higher conversion rate will increase the identity percentage, and a higher deletion rate will tend to remove large repeats.

In conclusion, the properties which supported the model of iteration dynamics established in *S. cerevisiae* are shared by other eukaryotes. This suggests that the model could be extended to all eukaryotes.

The Case of *P. falciparum*: How Parasitism Influences the Genome Style

P. falciparum chromosomes exhibit a high level of redundancy as compared with similar-sized chromo-

FIG. 3.—Distribution of the identity percentage for direct repeats. The histograms show the distribution of the identity percentage of a given species for distant direct repeats and CDRs. The hatched black boxes represent only the CDRs, and the plain black boxes represent the distant direct repeats. It can be shown using a Mann-Whitney rank test that, except for yeast, CDRs are more identical than distant direct repeats.

somes of *S. cerevisiae* (fig. 1), and their CDRs are extremely overrepresented: 74% have a spacer smaller than 1 kb (fig. 2). They are very identical (fig. 3) and very small (data not shown). However, no correlation between spacer, identity, and length can be highlighted (Table 2).

Two-thirds of the inverted repeats are located near the telomeres (one copy in each subtelomere), suggesting a peculiar history and a high exchange rate for these repeats. It was suggested that all subtelomeres exhibit a very plastic dynamics in *S. cerevisiae* (Pryde, Gorham, and Louis 1997) and in *H. sapiens* (Coleman, Baird, and Royle 1999). Their importance in the interchromosomal iteration process was demonstrated in *S. cerevisiae* (Coissac, Maillier, and Netter 1997).

All these observations are consistent with what was described previously: the highly repeated gene families and the special status of subtelomeres in *P. falciparum* (Gardner et al. 1998; Bowman et al. 1999).

Do These Observations Mean that This Ciliate Does Not Follow the Same Dynamics as the Other Eukaryotes?

P. falciparum is a human pathogenic parasite, the main agent of malaria. It has been reported that many bacterial pathogens exhibit a high redundancy level (Rocha, Danchin, and Viari 1999*b*) which has been related to high selective pressures for sequence variation. A significant number of repeats allows many recombination events, leading to a high plasticity of the genome, and then to a high evolution rate. As for these bacteria, the high redundancy level of *P. falciparum* could be a consequence of its parasitism.

Table 2
Computed Kendall Rank Correlations for CDRs of Each Species

	Close Direct Repeats[a]		Identity[b]		Length[b]	
Species	N	D_N	τ	P	τ	P
S. cerevisiae	60	5.0	−0.32	$<10^{-3}$	0.45	$<10^{-4}$
P. falciparum	100	49.8	−0.08	>0.05	0.06	>0.05
A. thaliana	889	23.9	−0.35	$<10^{-4}$	0.39	$<10^{-4}$
C. elegans	3,242	34.0	−0.31	$<10^{-4}$	0.24	$<10^{-4}$
D. melanogaster . . .	546	4.7	−0.36	$<10^{-4}$	0.41	$<10^{-4}$
H. sapiens	1,042	15.5	−0.30	$<10^{-4}$	0.33	$<10^{-4}$

[a] The number of CDRs (N) and the density in number (D_N) of these CDRs.

[b] The results of the tested correlations: the correlation coefficient (τ) and the probability value associated with (P) are given for the correlations between identity and spacer, and between length and spacer.

The quasi-absence of distant repeats and the absence of correlation indicate that there are almost only young repeats. The absence of correlation is, in this way, not caused by the absence of the mechanism leading to them but by too short a time of evolution. Population studies suggest that *P. falciparum* spread worldwide from a limited area (Rich and Ayala 2000). The absence of old repeats could be a consequence of the recent change in the ecological conditions of *P. falciparum*, associated with a burst of evolution. In conclusion, *P. falciparum* follows the same iteration dynamics as the other eukaryotes. However, because it is a recent parasite, its chromosomes are more repeated than those of the other eukaryotes (as a result of parasitism), and there are almost no ancient repeats (because of its recent emergence).

How Tandem Repeats Can Be Turned into Spaced Repeats

Intrachromosomal repeats, in our model, are mostly created in tandem (by recombination between sister chromatides or by replication slippage), and are turned into distant repeats by chromosomal rearrangements. Analyzing all the ending states after several rearrangements is difficult. However, it is interesting to examine all the theoretical resulting states obtained after only one rearrangement event. Three kinds of rearrangement have been taken into account (fig. 4): deletion of a part of the tandem, insertion of a sequence inside the tandem repeat, and inversion taking away a piece of the tandem. The insertion process can be the result of either the insertion of a transposable element or the reparation of a double-strand break by sequence conversion (Voelkel and Roeder 1990). Small inversions have been suggested to explain the evolution of the genes' order between *C. albicans* and *S. cerevisiae* (Seoighe et al. 2000), highlighting their role in genome dynamics.

If the model is valid, one should find the vestiges of tandem rearrangement in the chromosome sequences. Thus, we used the wublastn software (http://blast.wustl.edu) to look for paralogs of the spacers in the complete chromosomes. Only spacers with size between 50 bp and 10 kb and flanked by direct repeats were taken into account. It should be stressed that the queried databases were constructed for each species with complete chromosomes only (the same that we used for the detection of the repeats). A sequence was arbitrarily considered as a paralog of the spacer if the sequence length was at least 80% of the spacer length, and if the two sequences were identical by more than 80%. Using this approach, large insertions (fig. 4.2*a*) or some inversions (fig. 4.3*b*) can be undoubtedly identified, but small internal deletions and small internal insertions (fig. 4.1*b* and 4.2*b*) cannot be clearly differentiated. One should notice that deletion of an edge of a copy (fig. 4.1*a*) or a complete inversion of a copy (fig. 4.3*a*) cannot be detected by this method.

Results were sorted as a function of the number of paralogs detected in the chromosomes. For most spacers, no paralog was found. This has several possible reasons: (1) our criteria were very stringent, (2) the research was performed against the whole genome only for *S. cerevisiae* and *C. elegans*, and (3) we only detected paralogs for spacers issued from a recent unique event of rearrangement. Multiparalog families (when a spacer presented at least two paralogs) were separated because they give an idea of the relative transposition rate. All cases where the spacer had only one paralog have been analyzed more precisely as they appeared in figure 4.

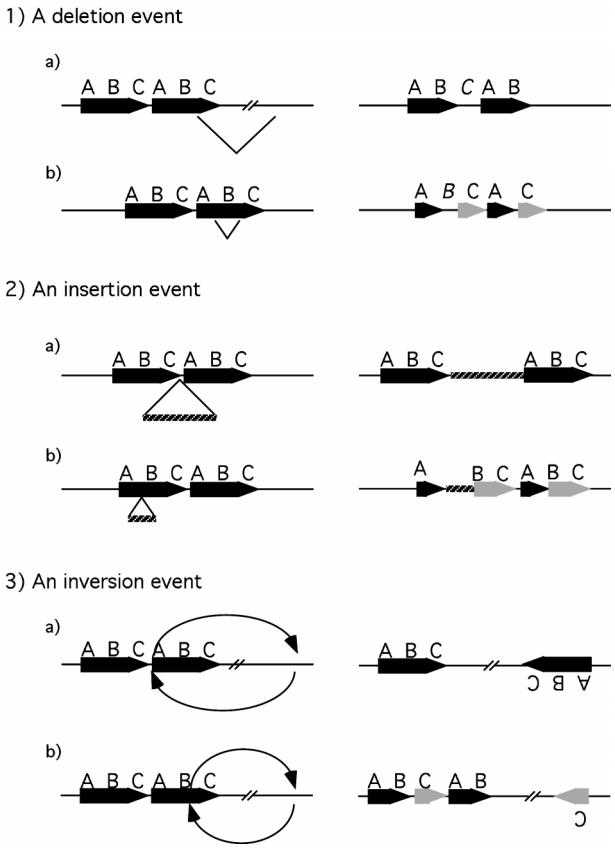

FIG. 4.—How tandem repeats can be turned into spaced repeats. This figure presents the three main ways to create a spacer between two tandem repeats: (1) a deletion event could occur inside the tandem repeats, leading to the creation of a spaced direct repeat (*a*) and two spaced direct repeats (*b*), (2) an insertion event could arise inside the tandem repeat, leading to one direct repeat with a spacer similar to another region in the genome (*a*) and two direct repeats (*b*), and (3) an inversion event could lead to one inverted repeat (*a*) or one direct repeat in which the spacer is an inverted repeat (*b*).

As shown in Table 3, all possible remnants of the tandem rearrangement were detected in the sequence of chromosomes. These observations indicate that the theoretical rearrangements arise in the genome history, reinforcing the model of the iteration dynamics.

A striking result was the overrepresentation of intrachromosomal direct paralogs in *C. elegans*. A detailed analysis of these paralogs revealed that they are mostly part of larger old tandem repeats. This observation has to be connected to the presence of large tandem repeats in the chromosomes of this species (i.e., a 600-kb repeat in the first chromosome), also recently described by Friedman and Hughes (2000). It seems probable that the

Table 3
Detected Paralogs for Spacers of Direct Repeats

Species	No Paralog	Only One Paralog[a]				At Least Two Paralogs	Total[c]
		Another Chromosome	Same Chromosome[b]				
			Close Direct	Direct	Inverted		
S. cerevisiae	20	1	1	—	—	6	28
P. falciparum	10	—	1	—	—	—	11
A. thaliana	1,318	11	46	17	6	77	1,475
C. elegans.	1,960	61	70	308	5	505	2,909
D. melanogaster[d] . . .	473	—	3	4	1	14	495
H. sapiens.	419	1	12	14	4	62	512

[a] The spacer and its paralog were either in two different chromosomes or in the same one.

[b] The spacer and its paralog were either in the same orientation and physically closer than 1 kb, in the same orientation but not close, or not in the same orientation.

[c] Blastn results obtained in paralogs research of the spacer sequences of direct repeats. Only spacers with a size between 50 bp and 10 kb were queried against the complete chromosomes of a given species. A sequence is considered as a paralog if its length is at least 80% of the spacer length and if the identity percentage is at least 80%.

[d] For *D. melanogaster*, the arms 2L and 2R (as well as 3L and 3R) were considered as the same chromosome and in the same orientation.

worm genome has exhibited an active process of intrachromosomal iteration.

All generic duplications of nonspecific DNA regions within the same chromosome or between two chromosomes were referred to in this study as iteration. However, this iteration process should be divided into at least two distinct mechanisms. The first is the creation of tandem repeats (by sister chromatide exchange or replication slippage), which creates (under our model) most of the intrachromosomal repeats. The second is the genesis of repeats (inter- or intrachromosomal) by a double-strand break repair. Actually, this repair can lead to duplication when the repair is associated with a conversion mechanism. This implies that the duplication process can at least be divided into four mechanisms: abnormal chromosome segregation (hyperploidization); transposition (transposable elements); sister chromatide exchange, replication slippage (tandem repeats and satellites), or both; and double-strand break repair (iteration by conversion).

Conclusions

Through this study of eukaryotes' intrachromosomal repeats, several biological results were highlighted. We extended our model, proposed for *S. cerevisiae*, to other eukaryote chromosomes (*S. cerevisiae*, *C. elegans*, *P. falciparum*, *A. thaliana*, *D. melanogaster*, and *H. sapiens*). This suggests that despite the differences in chromosomal properties, the iteration process follows globally the same dynamics in the eukaryote kingdom and thus has to be connected to structures and mechanisms shared by all eukaryote chromosomes.

The density of repeats number defines a genome style where the evolution rate results from iteration, deletion, rearrangement, and mutation. This rate is similar for all chromosomes within the same genome and is specific to each species. The main exception being the X chromosome of *C. elegans*, it suggests that exchanges between homologous chromosomes are important in the genesis of repeats. Thus, we propose that the genesis of tandem repeats is at least a consequence of exchange between homologous chromosomes.

Finally, we brought out the remnants of rearrangements of tandem repeats into spaced repeats. This suggests that tandem repeats, which can be easily created, are submitted to rounds of chromosomal rearrangements leading to the pattern of repeats observed today. Hence, repeats can be used to follow chromosome rearrangements and are markers of genome dynamics.

Acknowledgments

We thank I. Gonçalves, E. Rocha, D. Higuet, E. Maillier, J. Pothier, and A. Viari for their scientific help and their friendly support. This work was supported by grants from Association pour la Recherche sur le Cancer. E.C. and P.N. are members of Université Pierre et Marie Curie, Paris.

Literature Cited

Achaz, G., E. Coissac, A. Viari, and P. Netter. 2000. Analysis of intrachromosomal duplications in yeast *Saccharomyces cerevisiae*: a possible model for their origin. Mol. Biol. Evol. **17**:1268–1275.

Adams, M. D., S. E. Celniker, R. A. Holt et al. (195 co-authors). 2000. The genome of *Drosophila melanogaster*. Science **287**:2185–2195.

Amores, A., A. Force, Y. L. Yan et al. (13 co-authors). 1998. Zebrafish *hox* clusters and vertebrate genome evolution. Science. **282**:1711–1714.

Barnes, T. M., Y. Kohara, A. Coulson, and S. Hekimi. 1995. Meiotic recombination, noncoding DNA and genomic organization in *Caenorhabditis elegans*. Genetics **141**:159–179.

Baudat, F., and A. Nicolas. 1997. Clustering of meiotic double-strand breaks on yeast chromosome III. Proc. Natl. Acad. Sci. USA **94**:5213–5218.

Bernardi, G. 2000. Isochores and the evolutionary genomics of vertebrates. Gene **241**:3–17.

Blanc, G., A. Barakat, R. Guyot, R. Cooke, and M. Delseny. 2000. Extensive duplication and reshuffling in the Arabidopsis genome. Plant Cell **12**:1093–1101.

Bowman, S., D. Lawson, D. Basham et al. (36 co-authors). 1999. The complete nucleotide sequence of chromosome 3 of Plasmodium falciparum. Nature **400**:532–538.

Britten, R. J. 1998. Precise sequence complementarity between yeast chromosome ends and two classes of just-sub-

telomeric sequences. Proc. Natl. Acad. Sci. USA **95**:5906–5912.

COISSAC, E., E. MAILLIER, and P. NETTER. 1997. A comparative study of duplications in bacteria and eukaryotes: the importance of telomeres. Mol. Biol. Evol. **14**:1062–1074.

COLEMAN, J., D. M. BAIRD, and N. J. ROYLE. 1999. The plasticity of human telomeres demonstrated by hypervariable telomeres repeat array that is located on some copies of 16p and 16q. Hum. Mol. Genet. **8**:1637–1646.

CONSORTIUM. 1998. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science **282**: 2012–2018.

DUNHAM, I., N. SHIMIZU, B. A. ROE et al. (239 co-authors). 1999. The DNA sequence of human chromosome 22. Nature **402**:489–495.

FRIEDMAN, R., and A. L. HUGHES. 2000. Gene duplication and the structure of eukaryotic genomes. Genome Res. **11**:373–381.

GARDNER, M. J., H. TETTELIN, D. J. CARUCCI et al. (27 co-authors). 1998. Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. Science **282**:1126–1132.

GOFFEAU, A., B. G. BARRELL, H. BUSSEY et al. (16 co-authors). 1996. Life with 6000 genes. Science **274**:546.

HATTORI, M., A. FUJIYAMA, T. D. TAYLOR et al. (62 co-authors). 2000. The DNA sequence of human chromosome 21. The chromosome 21 mapping and sequencing consortium. Nature **405**:311–319.

HENIKOFF, S. 2000. Heterochromatin function in complex genomes. Biochem. Biophys. Acta **1470**:O1–O8.

HOLLAND, P. W. 1999. Gene duplication: past, present and future. Semin. Cell Dev. Biol. **10**:541–547.

HUGHES, A. L., J. DA SILVA, and R. FRIEDMAN. 2001. Ancient duplication did not structure the human *Hox*-bearing chromosomes. Genome Res. **11**:771–780.

JINKS, R. S., M. MICHELITCH, and S. RAMCHARAN. 1993. Substrate length requirements for efficient mitotic recombination in *Saccharomyces cerevisiae*. Mol. Cell. Biol. **13**:3937–3950.

KARLIN, S., and F. OST. 1985. Maximal segmental match length among random sequences from a finite alphabet. Pp. 225–243 *in* L. M. L. CAM and R. A. OLSHEN, eds. Proceedings of the Berkeley Conference in honor of Jerzy Neyman and Jack Kiefer, Vol. 1. Association for Computing Machinery, New York.

KURTZ, S., and C. SCHLEIERMACHER. 1999. REPuter: fast computation of maximal repeats in complete genomes. Bioinformatics **15**:426–427.

LEUNG, M. Y., B. E. BLAISDELL, C. BURGE, and S. KARLIN. 1991. An efficient algorithm for identifying matches with errors in multiple long molecular sequences. J. Mol. Biol. **221**:1367–1378.

LIN, X., S. KAUL, S. ROUNSLEY et al. (39 co-authors). 1999. Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. Nature **402**:761–768.

LLORENTE, B., A. MALPERTUY, C. NEUVEGLISE et al. (24 co-authors). 2000. Genomic exploration of the hemiascomycetous yeasts: 18. Comparative analysis of chromosome maps and synteny with *Saccharomyces cerevisiae*. FEBS Lett. **487**:101–112.

MASTERSON, J. 1994. Stomatal size in fossil plants: evidence for polyploidy in majority of angiosperms. Science **264**: 421–424.

MAYER, K., C. SCHULLER, R. WAMBUTT et al. (234 co-authors). 1999. Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*. Nature **402**:769–777.

OHNO, S. 1970. Evolution by gene duplication. Springer-Verlag, Heidelberg, Germany.

PEETERS, B. P. H., J. H. DE BOER, S. BRON, and G. VENEMA. 1988. Structural plasmid instability in *Bacillus subtilis*: effect of direct and inverted repeats. Mol. Gen. Genet. **212**: 450–458.

PRYDE, F. E., H. C. GORHAM, and E. J. LOUIS. 1997. Chromosome ends: all the same under their caps. Curr. Opin. Genet. Dev. **7**:822–828.

RICH, S. M., and F. J. AYALA. 2000. Population structure and recent evolution of *Plasmodium falciparum*. Proc. Natl. Acad. Sci. USA **97**:6994–7001.

ROBINSON-RECHAVI, M., O. MARCHAND, H. ESCRIVA, P. L. BARDET, D. ZELUS, S. HUGHES, and V. LAUDET. 2001. Euteleost fish genomes are characterized by expansion of gene families. Genome Res. **11**:781–788.

ROCHA, E. P., A. DANCHIN, and A. VIARI. 1999a. Analysis of long repeats in bacterial genomes reveals alternative evolutionary mechanisms in *Bacillus subtilis* and other competent prokaryotes. Mol. Biol. Evol. **16**:1219–1230.

ROCHA, E. P., A. DANCHIN, and A. VIARI. 1999b. Functional and evolutionary roles of long repeats in prokaryotes. Res. Microbiol. **150**:725–733.

RUBNITZ, J., and S. SUBRAMANI. 1984. The minimum amount of homology required for homologous recombination in mammalian cells. Mol. Cell Biol. **4**:2253–2258.

RUVKUN, G., and O. HOBERT. 1998. The taxonomy of developmental control in *Caenorhabditis elegans*. Science **282**: 2033–2041.

SEMPLE, C., and K. H. WOLFE. 1999. Gene duplication and gene conversion in the *Caenorhabditis elegans* genome. J. Mol. Evol. **48**:555–564.

SEOIGHE, C., N. FEDERSPIEL, and T. JONES et al. (20 co-authors). 2000. Prevalence of small inversions in yeast gene order evolution. Proc. Natl. Acad. Sci. USA **97**:14433–14437.

SMITH, T. F., and M. S. WATERMAN. 1981. Identification of common molecular subsequences. J. Mol. Biol. **147**:195–197.

SURZYCKI, S. A., and W. R. BELKNAP. 2000. Repetitive-DNA elements are similarly distributed on *Caenorhabditis elegans* autosomes. Proc. Natl. Sci. USA **97**:245–249.

VINCENS, P., L. BUFFAT, C. ANDRE, J. P. CHEVROLAT, J. F. BOISVIEUX, and S. HAZOUT. 1998. A strategy for finding regions of similarity in complete genome sequences. Bioinformatics **14**:715–725.

VISION, T. J., D. G. BROWN, and S. D. TANKSLEY. 2000. The origins of genomic duplications in *Arabidopsis*. Science **290**:2114–2117.

VOELKEL, K., and G. S. ROEDER. 1990. Gene conversion tracts stimulated by HOT1-promoted transcription are long and continuous. Genetics **126**:851–867.

WOLFE, K. H., and D. C. SHIELDS. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. Nature **387**:708–713.

MANOLO GOUY, reviewing editor

Accepted August 27, 2001

B.2. Résumé des résultats et de la discussion.

Cette étude nous a permis de détecter des répétitions à deux copies dans les génomes de 6 organismes eucaryotes. Les nombres de ces répétitions sont présentés dans le tableau 6. L'examen rapide de ce tableau montre que dans tous les génomes, nous détectons plus de répétitions directes que de répétitions inversées. Ce résultat peut sembler en contradiction avec celui obtenu dans l'étude précédente. Cependant, dans cette nouvelle étude, la stringence de détection est élevée. Dans l'étude précédente, les répétitions inversées étaient plus nombreuses mais de plus petite taille et moins similaires. Elles ne sont donc plus considérées comme significatives ici. Une seconde observation peut être faite : les chromosomes de *D. melanogaster* présentent très peu de répétitions étant donnée la taille totale analysée. La méthode de détermination de cette séquence est basée sur le séquençage aléatoire de fragments chromosomiques et de leurs assemblages en masse (technique de *Shotgun*). Il est possible que ce petit nombre de répétitions soit la conséquence d'un biais de la méthode d'assemblage des répétitions. Ce biais pourrait parfois considéré comme unique une séquence répétée en tandem. Mais cela peut évidemment aussi être la conséquence d'un processus biologique particulier à *D. melanogaster* : un taux de duplication plus faible ou un taux de délétion plus élevé.

Espèce	Taille totale analysée (Mb)	Nombre de répétitions directes	Nombre de répétitions inversées
S. cerevisiae	12,1	110	75
P. falciparum	2,0	136	48
A. thaliana	37,2	2 407	1 068
C. elegans	95,2	6 885	1 645
D. melanogaster	114,42	1 479	691
H. sapiens	67,3	3 457	2 406

Tableau 6 : Nombre de répétitions détectées dans les chromosomes eucaryotes.

Cette étude a également mis en évidence une propriété intéressante des répétitions à deux copies dans les chromosomes eucaryotes : la densité de répétitions (nombre de répétitions/Mb) est similaire dans tous les chromosomes d'un même génome (article 2,

figure 1). Cela suggère que le processus d'amplification (balance entre duplication et délétion) soit la conséquence d'un mécanisme global dont les caractéristiques sont spécifiques à chaque génome. Il existe pourtant deux chromosomes différents à cet égard : le quatrième chromosome de *D. melanogaster* et le chromosome X de *C. elegans*. Etant donné les résultats surprenants des répétitions chez *D. melanogaster*, le premier ne sera pas discuté. A l'inverse, il est intéressant de voir que le chromosome X de *C. elegans* est le seul chromosome hémiploïde : l'hermaphrodite étant X0 et la femelle XX. Cette particularité a pour conséquence que le chromosome X ne peut faire de recombinaison méiotique que chez les femelles. Si l'on imagine que la recombinaison méiotique est un des mécanismes de création des répétitions, il est logique que le chromosome X soit sous-répété. Si cette explication est correcte, alors tous les chromosomes X devraient être sous-répétés. Comme aucun chromosome n'effectue de recombinaison méiotique chez le mâle de *D. melanogaster*, le chromosome X de cet organisme ne peut pas être utilisé comme témoin.

Pour tester notre modèle de dynamique des répétitions intrachromosomiques, nous avons recherché si : (1) les CDR sont surreprésentées, (2) la longueur et l'identité des CDR sont corrélées à la taille du *spacer* et (3) l'on trouve des traces des remaniements des CDR. Comme les annotations des gènes dans les génomes eucaryotes sont encore imparfaites, nous n'avons pas recherché les positions relatives des répétitions par rapport aux gènes.

(1) L'examen de la distribution des tailles de *spacer* montre que, dans les six génomes étudiés, les CDR sont très abondantes (figure 2, article 2). Très peu de CDR étant attendues par hasard (article 1), les CDR sont surreprésentées dans tous les chromosomes étudiés. L'examen attentif de ces distributions met également en évidence une abondante population de répétitions directes espacées de 1 à 10 kilobases. Cette population est particulièrement importante chez *A. thaliana*, *C. elegans* et *D. melanogaster*. Nous n'avons pas encore d'hypothèses plausibles pour expliquer la présence de ces nombreuses répétitions. Il faut également mentionner que quasiment toutes les répétitions directes de *P. falciparum* sont des CDR.

(2) Pour les CDR de tous les organismes, à l'exception de *P. falciparum*, il y a une corrélation négative entre identité et taille de *spacer* et une corrélation positive entre longueur et taille du *spacer* (Table 2, article 2). Ceci tend donc à montrer que, comme chez *S. cerevisiae*, le taux de délétion et de conversion est négativement corrélé à la taille du *spacer*.

(3) Nous avons recherché des traces d'évènements de remaniements de répétitions en tandem. Bien qu'il soit difficile de caractériser de telles traces si de multiples réarrangements se sont produits après la duplication en tandem, il est possible de détecter les traces d'un seul événement de réarrangement. Trois types de réarrangements ont été pris en compte : la délétion, l'insertion et l'inversion. La plupart d'entre eux produit une répétition directe dont le *spacer* possède un paralogue dans le génome (figure 4, article 2). Si ce paralogue est une répétition directe proche, alors l'événement est soit une petite insertion, soit une petite délétion. Si le paralogue est une répétition inversée, alors on peut penser que le tandem a été remanié par une inversion. Si le paralogue est situé sur un autre chromosome, il s'agit vraisemblablement d'une insertion.

Nous avons donc entrepris la détection systématique des paralogues potentiels des *spacer* des répétitions directes. Seules les répétitions directes espacées de 50 paires de bases à 10 kilobases ont été examinées. Pour la plupart de ces répétitions, aucun paralogue n'a pu être clairement identifié. Cependant quelques *spacer* possèdent au moins un paralogue dans les chromosomes étudiés. Nous avons décidé de ne considérer que les cas où le paralogue est unique afin de ne pas prendre en compte les événements d'insertion de transposons. Dans les cas de paralogue unique, nous avons détecté tous les types de paralogues attendus avec un unique événement d'insertion, de délétion ou d'inversion (Table 3, article 2). Il semble donc que dans les six génomes étudiés, les répétitions en tandem sont remaniées.

Nous avons recherché une hypothèse pour expliquer l'absence de corrélation (entre identité, longueur et taille *spacer*) dans les chromosomes de *P. falciparum*. Ce dernier possède le génome qui présente la densité de CDR la plus importante (Table 2, article 2). De plus, la plupart des répétitions inversées sont localisées dans les subtélomères, qui ont leur propre dynamique (voir article 1). Il semble donc que les répétitions de ce génome soient pour la

plupart « récentes ». Cette hypothèse est étayée par des études récentes qui suggèrent une expansion géographique très récente (moins de 10 000 ans) de ce parasite. Si les répétitions sont très récentes, cela explique pourquoi elles ne sont pas encore dispersées dans le génome. Les corrélations sont le résultat d'un processus de délétion et de conversion sur un grand nombre de générations. Or, si ces répétitions sont pour la plupart récentes, le nombre de générations nécessaires à l'établissement de telles corrélations n'est peut-être pas encore révolu. Nous proposons donc que l'absence de corrélation chez *P. falciparum* ne soit pas la conséquence d'une absence des mécanismes de conversion et de délétion, mais plutôt d'un temps d'évolution trop court.

En conclusion, il semble, à la lumière de nos résultats, que les génomes eucaryotes partagent une dynamique commune concernant leurs répétitions intrachromosomiques. Notre modèle de création de répétitions en tandem puis espacement par réarrangement (Figure 4, article 1) rend compte relativement bien de cette dynamique. Cependant, comme l'indique les densités différentes de répétitions entre les génomes, il semble que chaque génome soit emprunt d'un « style d'amplification » propre. Ce dernier est probablement la conséquence de mécanismes globaux de gestion des chromosomes qui laissent des traces observables sur les caractéristiques des répétitions. Les répétitions sont donc un outil intéressant permettant l'étude précise des « forces » impliquées dans la dynamique des génomes.

B.3. Répétitions, teneur en GC et recombinaison.

Si le but premier de la détection des répétitions était de tester le modèle de dispersion des répétitions établi pour le génome de *S. cerevisiae*, d'autres analyses peuvent être entreprises. Une d'entre elles concerne l'étude précise des facteurs influençant la duplication en tandem et les mécanismes susceptibles de supprimer les répétitions en tandem ou de les homogénéiser. Les résultats présentés ci-dessous ne sont que préliminaires mais suggèrent que la teneur en GC de la région et le taux de recombinaison modulent ce processus d'amplification des répétitions en tandem.

B.3.1. La teneur en GC.

Les génomes des vertébrés, et en particulier celui des vertébrés à sang chaud, présentent des régions relativement homogènes du taux de GC le long de la séquence des chromosomes (Bernardi 2000). Ces régions sont nommées isochores. L'analyse minutieuse de ces isochores supposés homogènes en GC révèle des hétérogénéités locales (TIHGSC 2001). Les mécanismes pouvant expliquer ces isochores restent encore mal compris et ne seront pas abordés ici.

Nous avons recherché si la composition locale en GC avait une influence sur le processus d'amplification (balance entre duplication, délétion et conversion). Pour cela, nous avons comparé la distribution de la teneur en GC dans les chromosomes 21 et 22 de *H. sapiens* à celle des régions contenant les répétitions en tandem. Les estimations ont été réalisées comme suit :

- Pour estimer la distribution de GC dans les chromosomes, nous avons découpé les chromosomes en tranches de 150 kilobases chevauchantes sur 5 kilobases. Pour chacun d'entre eux, la teneur moyenne en GC est estimée.
- Pour estimer la distribution de GC dans les régions contenant des répétitions en tandem, nous avons pour chaque répétition en tandem extrait les séquences flanquantes sur 25 kilobases. Puis nous avons calculé le taux de GC moyen de cette séquence d'environ 50 kilobases (2 séquences de 25 kilobases + la répétition). Nous n'avons sélectionné, comme répétitions en tandem, que les répétitions directes espacées de moins de 10 paires de bases.

La distribution totale de teneur en GC des chromosomes et celle des régions contenant les répétitions en tandem sont présentées sur la figure 37. Pour les chromosomes 21 et 22, les distributions de GC total et de GC des régions environnant les répétitions en tandem ne sont pas équivalentes (Mann-Whitney, $p<10^{-4}$). Dans les deux chromosomes, les répétitions en tandem sont préférentiellement localisées dans les régions riches en GC.

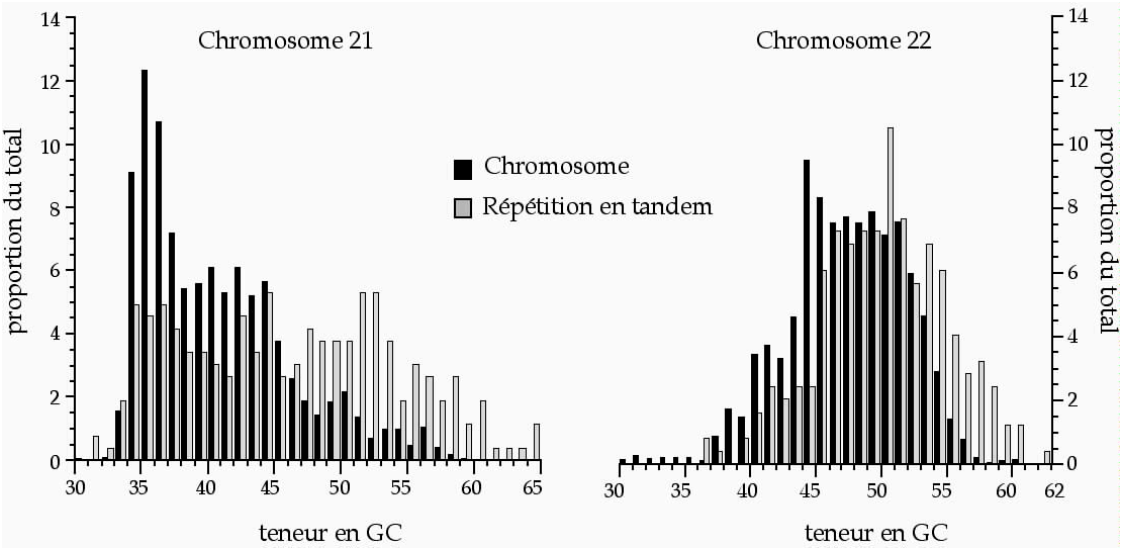

Figure 37 : Distributions de la teneur en GC du chromosome et des régions contenant des répétitions en tandem pour les chromosomes 21 et 22 de *H. sapiens*.
La distribution de la teneur en GC des répétitions en tandem est significativement différente de celle du chromosome. Les répétitions en tandem semblent localisées préférentiellement dans les zones riches en GC du chromosome.

Cette localisation préférentielle est également observable pour les 5 autosomes de *C. elegans* (Mann-Whitney, $p<10^{-4}$) , mais pas pour le chromosome X. Ce biais de localisation ne semble pas (ou très faiblement) s'appliquer aux répétitions en tandem des autres organismes (*S. cerevisiae*, *P. falciparum*, *A. thaliana* et *D. melanogaster*).

Nous n'avons encore aucune hypothèse satisfaisante pour expliquer ce biais de localisation des répétitions en tandem dans les régions les plus riches en GC. Par ailleurs, il est encore impossible de préciser si les répétitions en tandem sont préférentiellement créées dans les régions riches en GC ou sont créées partout et sont plutôt conservées dans les régions riches en GC. Enfin, ce biais existant également pour le génome de *C. elegans*, il faut envisager qu'il n'est pas une conséquence de la structuration des chromosomes en isochores (présente surtout chez des vertébrés).

B.3.2. La recombinaison.

La comparaison des cartes génétiques (en cM) aux cartes physiques (en paires de bases) des chromosomes a révélé, que pour la plupart des chromosomes, le taux de recombinaison n'est pas constant le long du chromosome (Barnes *et al.* 1995) (Nachman and

Churchill 1996) (Payseur and Nachman 2000). Chez *S. cerevisiae*, la carte des fréquences des cassures double brin (initiatrices de la recombinaison en méiose) le long du chromosome III confirme la très forte hétérogénéité des taux de recombinaison entre les régions des chromosomes (Baudat and Nicolas 1997).

Nous avons entrepris de rechercher un lien entre le taux de recombinaison le long du chromosome et la densité de répétitions en tandem. Pour cela, nous avons tiré profit des cartes génétiques et physiques établies pour les six chromosomes de *C. elegans* (Barnes *et al.* 1995). La figure 38 représente pour chacun de ces chromosomes, les variations de taux de recombinaison et les densités de répétitions en tandem.

- Les taux de recombinaison sont calculés en estimant le nombre de cM par Mb de séquence.
- La densité de répétitions en tandem est représentée pour chaque segment non chevauchant de 100 kilobases.

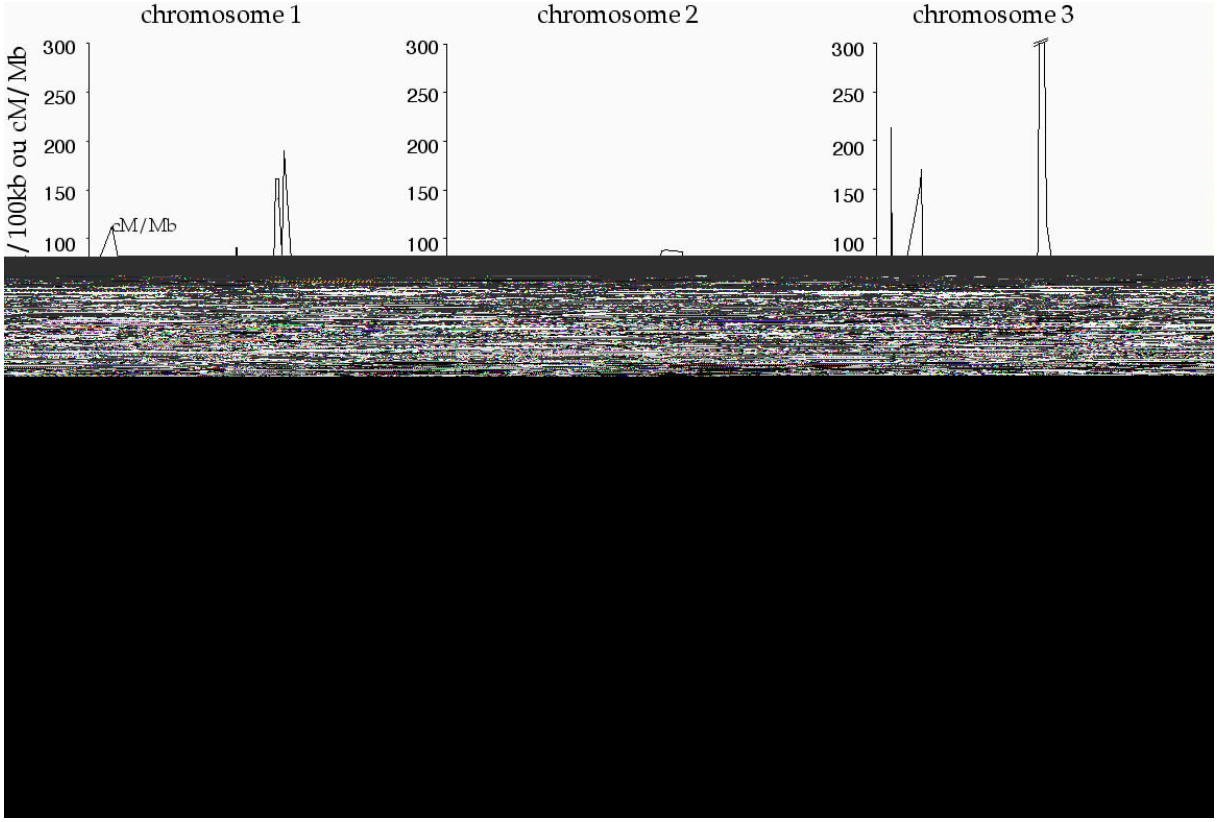

Figure 38 : Localisation des répétitions en tandem et variation du taux de recombinaison le long des chromosomes de *C. elegans*.
Le taux de recombinaison est estimé par les différences entre cartes génétiques (cM) et cartes physiques (Mb).

L'observation de la figure 38 suggère qu'il existe une corrélation entre taux de recombinaison et densité de répétitions en tandem. Les données utilisées pour les cartes sont anciennes, ce qui explique peut-être parfois le manque réel de recouvrement entre région à fort taux de recombinaison et région riche en répétition en tandem. Par ailleurs, comme exposé dans la discussion, les séquences des chromosomes que nous avons utilisé dans cette étude ne sont pas tout à fait de la bonne taille. On peut alors émettre l'hypothèse que la recombinaison méiotique est l'un des mécanismes impliqués dans la création de répétitions en tandem. Cette hypothèse, est renforcée par le fait que le chromosome X de *C. elegans* est sous-répété (figure 1, article 2).

C. Origine et destin des répétitions dans les génomes bactériens.

Les répétitions intrachromosomiques des génomes eucaryotes partagent une dynamique commune décrite par notre modèle de création en tandem puis dispersion. Nous avons entrepris de tester notre modèle dans les génomes bactériens. Pour cela, nous avons choisi 40 génomes de Bactéries et 11 génomes d'Archées dans lesquels nous avons recherché toutes les répétitions intrachromosomiques (Table 1, article 3). Seuls deux des 51 génomes, *Deinococcus radiodurans* et *Vibrio cholerae*, possèdent deux chromosomes ; aussi, pour la plupart des génomes bactériens, les répétitions intrachromosomiques constituent la totalité des répétitions. Nous nous sommes également intéressés aux facteurs pouvant influencer la création des répétitions en tandem. Les génomes bactériens pouvant présenter des compositions totales en G+C très différentes (Sueoka 1962), nous avons recherché un lien entre la composition nucléotidique globale des chromosomes et la densité de répétitions.

La taille des chromosomes bactériens est bien plus petite que celle de la plupart des chromosomes eucaryotes. Nous avons donc pu modifier la méthode de détection pour augmenter le nombre de répétitions pris en compte. Pour cela, nous avons opéré deux modifications majeures par rapport à la méthode utilisée dans la recherche des répétitions dans les chromosomes eucaryotes :

- Nous avons pris en compte toutes les répétitions en précisant pour chaque paire de copies, la taille de la famille répétée (duplications, triplications, etc…).
- Nous avons fixé une longueur minimum pour la détection des répétitions strictes très basse. Les répétitions non significatives ne sont enlevées qu'après la phase d'extension.

Après ces modifications, nous avons appliqué la méthode de détection aux 53 chromosomes bactériens, et entrepris de (1) caractériser les répétitions, (2) tester notre modèle et (3) regarder l'influence de la composition nucléotidique sur le processus d'amplification.

C.1. Article 3.

Nucleic Acids Research, 2002, Vol. 30, No. 13 2987–2994

Origin and fate of repeats in bacteria

G. Achaz*, E. P. C. Rocha[1,2], P. Netter and E. Coissac

Structure et Dynamique des Génomes, Institut Jacques Monod, Tour 43-44, 1° Étage, 4 Place Jussieu, F-75251 Paris Cedex 05, France, [1]Atelier de Bioinformatique, Université Pierre et Marie Curie, Paris, France and [2]URA2171, Unité GGB, Institut Pasteur, Paris, France

Received December 12, 2001; Revised April 12, 2002; Accepted May 8, 2002

ABSTRACT

We investigated 53 complete bacterial chromosomes for intrachromosomal repeats. In previous studies on eukaryote chromosomes, we proposed a model for the dynamics of repeats based on the continuous genesis of tandem repeats, followed by an active process of high deletion rate, counteracted by rearrangement events that may prevent the repeats from being deleted. The present study of long repeats in the genomes of Bacteria and Archaea suggests that our model of interspersed repeats dynamics may apply to them. Thus the duplication process might be a consequence of very ancient mechanisms shared by all three domains. Moreover, we show that there is a strong negative correlation between nucleotide composition bias and the repeat density of genomes. We hypothesise that in highly biased genomes, non-duplicated small repeats arise more frequently by random effects and are used as primers for duplication mechanisms, leading to a higher density of large repeats.

INTRODUCTION

DNA repeats can be defined as sequences sharing extensive similarity with other sequences of the same genome. It is usually supposed that repeats arise by successive duplications and several causal mechanisms, including hyperploidisation (even polyploidisation), tandem duplication, double-strand break repair by insertion or transposition, have been proposed to be involved. The underlying mechanisms are thought to act at different levels depending on the kingdom, or even on organism [i.e. polyploidisation has been proposed to explain the presence of large repeats in eukaryotes (1,2), but is probably absent in Archaea and Bacteria]. Once a repeat is created, it can be targeted by the recombination apparatus and be subject to deletion. Thus, genome size results from a balance between duplication and deletion events. The importance of deletion processes seems crucial in compact genomes, especially in those of intracellular endosymbionts or pathogens (3).

Usually, repeats in Bacteria are divided into two subclasses: low complexity repeats (sometimes mislabeled 'tandem repeats') and longer repeats (the centre of our interest). The first category is constituted of small oligonucleotides (typically ranging from mononucleotide to pentanucleotide in size) repeated many times in a head-to-tail configuration. These low complexity repeats, e.g. microsatellites, are very abundant in the genomes of eukaryotes, in which they have been widely studied (4). Although less abundant in bacterial and archaeal genomes (5), the mechanisms of their origin (6), their function (7), the consequences for genome dynamics (8) and the structural constraints imposed on the chromosome (9) have all been studied

Longer repeats include transposable elements, minisatellites (mostly in Eukarya), large tandem repeats and spaced repeats. DNA transposable elements (like IS) are widely distributed among the Archaea and Bacteria. As specific mechanisms for the duplication of mobile elements have been identified (10), such self-replicating elements have to be considered separately when the origin of repeats is analysed. However, they must be taken into account when the influence of repeats on genome stability is considered.

Several mechanisms have been proposed for the genesis of tandem repeats: slipped strand mispairing, unequal crossover (by homologous recombination), rolling circle and circle excision with reinsertion (11). Some of these mechanisms could also result in a tandem repeat deletion. These mechanisms render tandem repeats unstable, easy to create but also easy to delete. In contrast, distant repeats can almost only be deleted by homologous recombination and at the cost of large deletions of genetic material. As a consequence, they may persist more easily during genome evolution. Two mechanisms have been envisaged to create spaced repeats *ex nihilo*. The first, known as Campbell-like insertion, creates repeats by inserted exogenous sequences and has been proposed to explain the peculiar distribution of many repeats in *Bacillus subtilis* (12). The second, referred to as 'conversion' or 'insertion', repairs a double-strand break by copying a sequence sharing similarity with the edges of the broken sequence: this mechanism works either by break-induced replication or by gap repair (for reviews in yeast see 13,14).

The first question we tackled in this work concerns the origin of interspersed repeats (excluding transposable elements). Our previous studies (15,16) had led us to propose a model (Fig. 1) for the origin of eukaryote intrachromosomal repeats based on the permanent genesis of close direct repeats (CDR, repeats with copies separated by <1 kb). Since our model is compatible with all mechanisms, we do not assume any particular one for the creation of CDR. Newly created CDR are then subject to a strong rate of exchange (conversion and deletion). Experimental studies undertaken on *B.subtilis* (17) and *Escherichia coli* (18–20) have shown that the rate of illegitimate recombination is negatively correlated with the distance between the copies (spacer size) and positively correlated with repeat length. Recombination between close repeats tends to maintain

*To whom correspondence should be addressed. Tel: +33 1 44 27 76 94; Fax: +33 1 44 27 82 05; Email: achaz@ijm.jussieu.fr

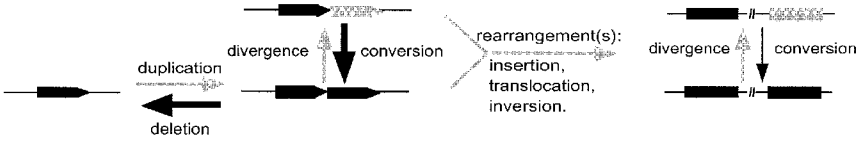

Figure 1. A model of interspersed repeats dynamics. In this model, interspersed repeats originate mainly from tandem repeats, which can be separated by further chromosomal rearrangements. In newly created repeats with a small spacer (i) the conversion rate is high, keeping the two copies identical and (ii) the deletion rate is also high, so that over a longer time scale only small repeats are retained. However, if one or more rearrangements (e.g. insertion, translocation and/or inversion) occur separating the two copies, both deletion and conversion rates decrease markedly. Both copies are then free to evolve.

neighbouring repeats identical (by conversion) but also to eliminate them (by deletion). At each round of exchange, both events are possible (although we ignore whether they are equally likely). If conversion can be followed by deletion, the opposite is not true: a deletion event cannot be followed by conversion. Over a long time, this will result in a bias in favour of deletions, with CDR disappearing sooner or later (depending on the relative rates of conversion and deletion). Thus, in the absence of strong selective pressure, long CDR are too unstable to persist, except if the copies are moved further apart by chromosomal rearrangements (i.e. insertion, translocation and inversion). In this case, the rate of illegitimate recombination will drop severely and the repeats may be maintained.

In this context, one expects CDR to be more similar than distant repeats, since either they are more recent or they are more subject to conversion. On the other hand, one expects that larger repeats will only survive fast deletion by frequent illegitimate recombination if they are placed distantly. Thus, under our model, CDR tend to have smaller and more identical repeats whereas distant repeats tend to be longer and less similar. This matches the observations we have made in eukaryote genomes, where repeats are both more identical and smaller when they are closer (15). The main goal of this work was to test if this model, first established in Eukarya, could be applied to Bacteria and Archaea.

The second focus of our attention concerns the factors influencing the dynamics of our model, i.e. rates of duplication, deletion and rearrangement. Here we analyse precisely the relation between the origin of tandem repeats and the genome composition biases. Duplication mechanisms typically require the pre-existence of a region of similarity. Levinson and Gutman (8) proposed that small non-duplicated repeats (afterwards referred to as repeats appearing by chance) are primers for mechanisms such as slipped strand mispairing, thus creating larger repeats. We have tried to analyse this proposition by deciphering the relations between repeat density and the relative frequencies of nucleotides in the chromosome.

MATERIALS AND METHODS

Data

We analysed the complete genomes of 40 Bacteria and 11 Archaea (Table 1). All sequences were extracted from GenBank (ftp://ftp.ncbi.nih.gov/genbank/Bacteria), except for those of *Pyrococcus furiosus*, downloaded from http://www.genome.utah.edu.

Construction of the repeats database

We followed the methodology previously developed to detect repeats in eukaryote genomes (15,16), but made an extra effort to detect smaller, but significant, repeats, since bacterial chromosomes are smaller. The methodology is described below and follows four main steps.

First step: detection of seeds. In this step, exact direct and inverse repeats (seeds) of 15 bp were detected using the REPuter software (21). Many seeds with lengths that are not statistically significant according to Karlin and Ost statistics were retained (22). The second step is intended to further extend these seeds into larger, non-strict repeats.

Second step: from seeds to repeats. Local alignment (23) is used to extend the edges of the seeds into larger repeats. Except for the construction of the score matrix, the extension process is the same we used to analyse eukaryote chromosomes (16). This method produces non-exact repeats by extending a seed on both sides when similarity is high. To do so, we used an algorithm based on a local alignment procedure (23).

Nucleotide frequencies differ widely between species genomes, from 25 to 75% (24). Therefore, if an identity matrix is used for the local alignment, seeds of the same size in chromosomes with a very unbalanced distribution of nucleotides (e.g. *Ureaplasma urealiticum* where A ≈ T ≈ 0.37 and C ≈ G ≈ 0.13) tend to produce larger repeats than in genomes with equal frequencies (e.g. *E.coli*). In order to avoid this effect, we used an empirical scoring matrix for each chromosome, which takes into account its specific composition. These matrices provide a better score for matches between rare nucleotides:

$\text{match}_{i/i} = 100 \times (1 - p_i^2)$; $\text{match}_{N/i} = 25$

$\text{mismatch}_{i/j} = -100 \times (1 - p_i \times p_j)$; $\text{gap}_{open} = -400$; $\text{gap}_{ext} = -100$

where p_i is the frequency of nucleotide i. By building these matrices for all species, we observed scores for matches ranging from 86 to 98 and scores for mismatches ranging from –98 to –86. Thus the score of gap_{open} is always less than 4× mismatch and the score of gap_{ext} always less than 1× mismatch. We also tried other matrices that gave similar results.

Third step: removing repeats that are not statistically significant. Since seeds are rather small, many repeats may not have statistically significant lengths. To remove these non-significant repeats, we built, for each chromosome, 10 additional random chromosomes by shuffling it with respect to its trinucleotide composition (Markov chains of order 2). In these random sequences, repeats were detected as in real sequences (steps 1 and 2). Afterwards, we built a distribution of observed alignment scores from the set of repeats detected in the 10 random chromosomes. We then defined a threshold of significance, corresponding to 0.001 of this distribution. Below this minimal score (S_{min}), repeats were regarded as non-significant and removed from further analysis. S_{min} depends essentially on the size and

Table 1. Organisms analysed

40 Bacteria	16 Proteobacteria	3 alpha subdivision	*Caulobacter crescentus*	*Cacr*
			Mesorhizobium loti	*Melo*
			Rickettsia prowazekii	*Ripr*
		2 beta subdivision	*Neisseria meningitidis MD58*	*NemeM*
			Neisseria meningitidis Z2491	*NemeZ*
		8 gamma subdivision	*Buchnera species*	*Busp*
			Escherichia coli K12	*EscoK*
			Escherichia coli O157 H7	*EscoO*
			Haemophilus influenzae	*Hain*
			Pasteurella multocida	*Pamu*
			Pseudomonas aeruginosa	*Psae*
			Vibrio cholerae	*Vich_1/Vich_2*
			Xylella fastidiosa	*Xyfa*
		3 delta and epsilon subdivision	*Campylobacter jejuni*	*Caje*
			Helicobacter pylori 26695	*Hepy*
			Helicobacter pylori J99	*HepyJ*
	10 Low G+C Gram-Positive Bacteria	2 Streptococcaceae	*Streptococcus pyogenes*	*Stpy*
			Lactococcus lactis	*Lala*
		2 Staphylococcus	*Staphylococcus aureus Mu50*	*StauM*
			Staphylococcus aureus N315	*StauN*
		2 Bacillus	*Bacillus halodurans*	*Baha*
			Bacillus subtilis	*Basu*
		4 Mycoplasmataceae	*Mycoplasma genitalium*	*Myge*
			Mycoplasma pneumoniae	*Mypn*
			Mycoplasma pulmonis	*Mypu*
			Ureaplasma urealyticum	*Urur*
	3 High G+C Gram-Positive Bacteria	3 Mycobacteriaceae	*Mycobacterium leprae*	*Myle*
			Mycobacterium tuberculosis CDC1551	*MytuC*
			Mycobacterium tuberculosis HR7Rv	*MytuH*
	5 Chlamydiales	5 Chlamydiaceae	*Chlamydia pneumoniae AR39*	*ChpnA*
			Chlamydia pneumoniae CWL029	*ChpnC*
			Chlamydia pneumoniae J138	*ChpnJ*
			Chlamydia muridarum	*Chmu*
			Chlamydia trachomatis	*Chtr*
	2 Spirochaetales	2 Spirochaetaceae	*Borrelia burgdorferi*	*Bobu*
			Treponema pallidum	*Trpa*
	1 Cyanobacteria	1 Chroococcales	*Synechocystis sp.*	*Sysp*
	1 Thermus/Deinococcus	1 Deinococcus	*Deinococcus radiodurans*	*Dera_1/Dera_2*
	1 Thermotogales	1 Thermotoga	*Thermotoga maritima*	*Thma*
	1Aquificales	1 Aquificaceae	*Aquifex aeolicus*	*Aqae*
11 Archea	2 Crenarchaeota	1 Aeropyrum	*Aeropyrum pernix*	*Aepe*
		1 Sulfolobaceae	*Sulfolobus solfataricus*	*Suso*
	9 Euryarchaeota	1 Archaeoglobaceae	*Archaeoglobus fulgidus*	*Arfu*
		1 Halobacteriaceae	*Halobacterium species*	*Hasp*
		1 Methanobacteriaceae	*Methanobacterium thermoautotrophicum*	*Meth*
		1 Methanococcaceae	*Methanococcus jannaschii*	*Meja*
		3 Thermococcaceae	*Pyrococcus abyssi*	*Pyab*
			Pyrococcus furiosus	*Pyfu*
			Pyrococcus horikoshii	*Pyho*
		2 Thermoplasmaceae	*Thermoplasma acidophilum*	*Thac*
			Thermoplasma volcanium	*Thvo*

composition of the genome (and naturally on our choice of scoring system) and ranges from 2052 (*Chlamydia pneumoniae*) to 2258 (*Mycoplasma pulmonis*). Using score (S), length (L) and identity (Id), characteristics of some pertinent repeats from these two organisms are given with more details: (i) for *C.pneumoniae*, the smallest score corresponds to $S = 2052, L = 36$ and $Id = 80.6\%$; the medians of the distributions being $S = 4505$, $L = 220$ and $Id = 63.1\%$; (ii) for *M.pulmonis*, the smallest score corresponds to $S = 2258, L = 82, Id = 71.7\%$; the medians being $S = 3005, L = 90$ and $Id = 68.9\%$.

Fourth step: determining family sizes. At this stage, all significant repeats are given as a series of pairs. However, many repeats are organised in multicopy families (i.e. IS and rRNA operons). Hence, we developed a procedure to detect such multicopy families in our data set.

To do so, we built, for each chromosome, a map in which each position is linked to its '*n*-plication' degree: unique, duplicated, triplicated, etc. These maps were built by counting, for each chromosome position, the number of times this position is found in repeats (direct and inverted ones were pooled

together). Each pair was then associated with the map and the family size of each repeat was determined.

Density of repeats

In order to characterise the repeats, we used two measures of density, the density in number and the density in length. They are defined as:

D_N = no. of copies/size of chromosome (Mb)

$D_L = 100 \times$ [size of repeat sequence (bp)/size of chromosome (bp)]

Nucleotide complexity

Complexity is frequently used as a compact measure of the difference of the nucleotide distribution to equal repartition. In this context, information entropy has been proposed to describe biases of mononucleotide distributions (25):

$$H = -\sum_{i=A}^{T} p_i \log_4 p_i$$

where p_i is the frequency of nucleotide i. If a sequence exhibits an equal repartition of its four nucleotides (maximum complexity), its entropy is 1. In bacterial chromosomes it ranges from 0.91 to 1.

Proportion of CDR

CDR were originally defined as repeats with a distance between their two copies of <1 kb. We estimated the proportion of CDR expected if repeats are spread randomly along a chromosome. The proportion of CDR is calculated as the ratio between the number of CDR and the total number of repeats. Two cases were taken into account. (i) If the chromosome is circular, the largest spacer size is $L/2$, where L is the chromosome length. The distribution of spacer size is constant from 0 to $L/2$. So, the proportion of CDR in a circular chromosome is $1000 \times 2/L$. (ii) If the chromosome is linear, the largest spacer size is L and the spacer distribution decreases linearly from 0 to L. Using the intercept theorem of Thales (or any analytical demonstration), it could easily be demonstrated that the proportion of CDR is $1000/L \times (2 - 1000/L)$.

RESULTS AND DISCUSSION

What repeats have we detected?

We have found a large number of repeats in most (but not all) bacterial genomes (Table 2). In order to characterise these repeats, we used two measures of repeat density, D_N and D_L (see Materials and Methods). As expected, both densities were positively correlated ($\tau = 0.63$, $P < 10^{-4}$, Kendall τ rank test): a chromosome with many repeats also exhibits a high proportion of duplications in its chromosome. However, the biological interpretation of these measures may be quite different: D_N can be assimilated to the rate of amplification (a balance between duplication and deletion processes) and D_L to the history of the chromosomes, a measure of the redundancy tolerated by a chromosome. Thus, D_N and D_L should be analysed in parallel as they give complementary information on chromosomal redundancy. The data in Table 2 brings to the fore two issues. (i) Chromosomes of related organisms often exhibit similar densities of repeats: both *Chlamydia trachomatis* strains, the three *C.pneumoniae* strains, the three *Pyrococcus* strains, both *Mycobacterium tuberculosis* strains, both *Staphylococcus aureus* strains, both *Neisseria meningitidis* strains and both *Helicobacter pylori* strains. However, exceptions do exist. *Escherichia coli* O157:H7 is more repeated than K12, in agreement with previous observations (26). Also, when we broaden the phylogenetic range, we observe that the four *Mycoplasma* spp. show very different densities (D_N and D_L), indicating fast divergence, possibly due to their rudimentary

Table 2. Densities of repeats

Species[b]	Size[c] (Mb)	D_N[d]			D_L[e]		
		D_N	D_{N2}	D_{N2}/D_N	D_L	D_{L2}	D_{L2}/D_L
Cacr[a]	4.0	1085.7	667.2	0.61	19.8	11.6	0.58
Melo[a]	7.0	987.1	543.8	0.55	20.8	12.4	0.60
Ripr	1.1	206.0	180	0.87	2.4	2.2	0.90
NemeM[a]	2.3	472.2	211.2	0.45	20.4	6.8	0.33
NemeZ[a]	2.2	461.0	218.8	0.47	19.1	6.0	0.32
Busp	0.6	561.9	178	0.32	11.3	5.3	0.47
EscoK[a]	4.6	261.3	151.8	0.58	11.4	5.9	0.52
EscoO[a]	5.5	322.3	153.6	0.48	17.9	6.9	0.38
Hain[a]	1.8	517.5	215.2	0.42	8.9	4.9	0.55
Pamu[a]	2.3	275.1	133	0.48	6.2	3.6	0.58
Psae[a]	6.3	1663.1	819	0.49	28.2	15.6	0.55
Vich_1[a]	3.0	150.3	96.6	0.64	5.5	2.7	0.48
Vich_2[a]	1.1	71.8	54	0.75	10.1	3.4	0.33
Xyfa[a]	2.7	229.9	119.4	0.52	13.7	8.5	0.62
Caje	1.6	503.8	359.4	0.71	9.1	5.8	0.64
Hepy[a]	1.7	463.5	259	0.56	11.5	7.1	0.62
HepyJ[a]	1.6	481.8	288.4	0.60	10.7	7.2	0.67
Stpy[a]	1.9	203.5	119.8	0.59	7.2	3.7	0.51
Lala[a]	2.4	528.8	257	0.49	14.1	7.4	0.52
StauM[a]	2.9	368.0	212	0.58	13.1	7.2	0.55
StauN[a]	2.8	354.9	199.6	0.56	13.4	6.3	0.47
Baha[a]	4.2	225.8	130.8	0.58	9.4	3.6	0.38
Basu	4.2	264.3	185	0.70	9.8	6.7	0.69
Myge	0.6	287.9	131	0.46	6.9	2.6	0.37
Mypn	0.8	400.5	95.6	0.24	24.0	5.5	0.23
MypuC[a]	1.0	632.9	292.6	0.46	18.3	10.3	0.56
Urur	0.8	530.8	367.2	0.69	14.0	11.1	0.8
Myle	3.3	232.9	186.6	0.80	6.1	3.4	0.56
MytuC[a]	4.4	496.2	273	0.55	21.0	8.7	0.41
MytuH[a]	4.4	501.2	265.6	0.53	21.3	8.7	0.41
ChpnA	1.2	132.5	84.6	0.64	4.9	3.1	0.63
ChpnC	1.2	144.7	96	0.66	4.9	3.0	0.62
ChpnJ	1.2	132.7	86.4	0.65	4.6	2.8	0.61
Chmu	1.1	86.0	56.2	0.65	3.4	2.0	0.58
Chtr	1.0	68.1	53.8	0.79	2.3	2.2	0.96
Bobu	0.9	295.4	226.2	0.77	4.6	3.6	0.77
Trpa	1.1	126.5	77.4	0.61	3.9	2.4	0.6
Sysp[a]	3.6	402.4	254.6	0.63	9.1	5.1	0.55
Dera_1[a]	2.7	871.4	618.4	0.71	14.0	8.2	0.59
Dera_2[a]	0.4	303.1	179.4	0.59	7.9	5.7	0.73
Thma	1.9	187.6	111.8	0.60	6.6	3.8	0.58
Aqae	1.6	226.9	176.6	0.78	5.3	4.3	0.81
Aepe	1.7	226.4	136.6	0.60	4.1	2.5	0.61
Suso[a]	3.0	577.8	157	0.27	25.0	9.1	0.36
Arfu	2.2	349.8	202	0.58	10.3	6.8	0.66
Hasp[a]	2.0	1366.8	712	0.52	19.8	12.7	0.64
Meth[a]	1.8	302.6	229.6	0.76	9.2	6.8	0.74
Meja[a]	1.7	773.0	408.4	0.53	13.5	7.9	0.58
Pyab[a]	1.8	186.4	132.6	0.71	4.7	3.4	0.72
Pyfu	1.9	181.3	141.6	0.78	7.5	4.4	0.59
Pyho	1.7	198.5	123.2	0.62	5.3	3.5	0.66
Thac[a]	1.6	118.2	74.2	0.63	2.9	2.1	0.71
Thvo	1.6	162.1	83.2	0.51	5.1	3.1	0.61

[a]Chromosomes containing transposable elements.
[b]Abbreviations and order are those used in Table 1.
[c]Size of the chromosome (in Mb)
[d]D_N, number of copies per Mb. D_{N2} is the density for two-copy CDR only.
[e]D_L is the proportion of the chromosome included in repeats. D_{L2} is the proportion of the chromosome included in two-copy CDR only.

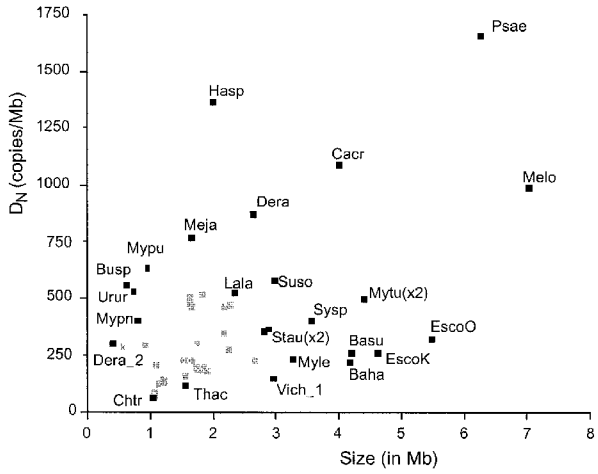

Figure 2. Repeat density as a function of chromosome size. Plot of D_N as a function of chromosome size. This figure illustrates the positive correlation between D_N and chromosome size.

repair mechanisms and to the selective pressure for variation in these pathogens (27). (ii) Both D_N and D_L exhibit a positive correlation with chromosome size ($\tau = 0.24$, $P < 10^{-3}$ for D_N and $\tau = 0.37$, $P < 10^{-4}$ for D_L),. These observations are in good agreement with previous observations on parts of both bacterial genomes and eukaryote genomes (16,28) (Fig. 2).

Since we were interested in the repeats' origins and in the supposition that it proceeds by duplication, we determined the proportions of two-copy repeats (and respective densities D_{N2} and D_{L2}) among all repeats (Table 2). As expected, D_{N2} is positively correlated with D_N ($\tau = 0.77$, $P < 10^{-4}$) and D_{L2} with D_L ($\tau = 0.73$, $P < 10^{-4}$). It could be noticed that, in contrast to eukaryote genomes in which D_{N2} is similar for chromosomes of the same species (16), densities varied between the two chromosomes of *Deinococcus radiodurans* and also between the two chromosomes of *Vibrio cholerae*.

Chromosomes containing transposable elements exhibit lower D_{N2}/D_N and D_{L2}/D_L ratios ($P < 0.01$, Mann–Whitney rank tests). Since transposable elements are mostly multicopy families, this can be easily understood. We observed few exceptions (low ratios in the absence of transposable elements), involving small genomes and, in particular, *Mycoplasma genitalium* and *Mycoplasma pneumoniae*. These repeats are associated with the immunodominant proteins of these genomes and are related to antigenic and tissue tropism variation (27).

Did interspersed repeats originate from tandems?

In order to test whether our model holds for Bacteria and Archaea we have tested its four major predictions. If interspersed repeats originate massively from tandem repeats, one might expect that (i) direct repeats are more numerous than inverted ones and that (ii) CDR are in large excess. Since the exchange rate between CDR is expected to be negatively correlated with spacer size and positively correlated with repeat length there should be (iii) a negative correlation between repeat similarity and spacer size and (iv) a positive correlation between repeat length and spacer size. Since we are interested in the origin of repeats, we decided to analyse only two-copy repeats further. This removed all low complexity repeats from our data set. Based on the annotations, we show

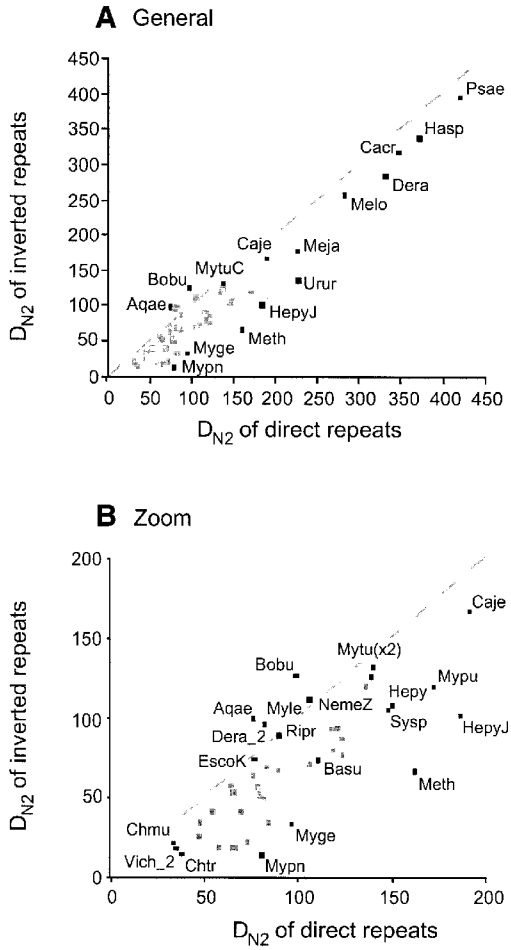

Figure 3. Densities of inverted repeats versus direct repeats. For each of the 53 chromosomes, we plotted the densities in number (D_{N2} = two-copy number/size in Mb) of inverted repeats as a function of D_{N2} of direct repeats. Because of the large difference in densities between genomes, two scales have been used. Abbreviations of species used in this figure correspond to those described in Table 1. The density of direct repeats is generally greater than the density of inverted repeats, but both are of the same order of magnitude.

that repeats located at least half in rRNA, tRNA or functional transposase represent ≤5% of our two-copy repeats, except for *C.trachomatis* (14%, four of 28) and the second chromosome of *V.cholerae* (7%, two of 29) (data not shown).

Direct repeats are more numerous than inverted ones. The large majority of the chromosomes (47 of 53) exhibit a higher density of two-copy direct repeats as compared with inverted ones ($P < 0.001$, binomial test), although sometimes the relative difference is not very high (Fig. 3). It is worth noticing that the two chromosomes that exhibit the largest excess of direct repeats are *M.genitalium* and *M.pneumoniae*. This is due to the previously described repeats located inside the adhesin genes.

CDR are over-represented. We estimated the numbers and densities of two-copy CDR, N_{2CDR} and D_{N2CDR}, respectively, and the theoretical number of CDR as a function of the number of direct repeats in linear and circular chromosomes (see

Table 3. Close direct repeats

Species[a]	N_{2CDR}[b]	p[c]	D_{N2cdr}[d]
Cacr	37	$<10^{-4}$	18.4
Melo	71	$<10^{-4}$	20.2
Ripr	4	$<10^{-4}$	7.2
NemeM	33	$<10^{-4}$	29.0
NemeZ	26	$<10^{-4}$	23.8
Busp	1	0.101	3.2
EscoK	23	$<10^{-4}$	10.0
EscoO	34	$<10^{-4}$	12.4
Hain	22	$<10^{-4}$	24.0
Pamu	23	$<10^{-4}$	20.4
Psae	51	$<10^{-4}$	18.2
Vich_1	12	$<10^{-4}$	8.2
Vich_2	4	$<10^{-4}$	7.4
Xyfa	46	$<10^{-4}$	34.4
Caje	16	$<10^{-4}$	19.6
Hepy	28	$<10^{-4}$	33.6
HepyJ	35	$<10^{-4}$	42.6
Stpy	14	$<10^{-4}$	15.2
Lala	26	$<10^{-4}$	22.0
StauM	30	$<10^{-4}$	20.8
StauN	29	$<10^{-4}$	20.6
Baha	26	$<10^{-4}$	12.4
Basu	31	$<10^{-4}$	14.8
Myge	7	$<10^{-4}$	14.2
Mypn	12	$<10^{-4}$	29.4
Mypu	17	$<10^{-4}$	35.2
Urur	12	$<10^{-4}$	32.0
Myle	19	$<10^{-4}$	11.6
MytuC	34	$<10^{-4}$	15.4
MytuH	33	$<10^{-4}$	15.0
ChpnA	12	$<10^{-4}$	19.6
ChpnC	12	$<10^{-4}$	19.6
ChpnJ	10	$<10^{-4}$	16.2
Chmu	6	$<10^{-4}$	11.2
Chtr	5	$<10^{-4}$	9.6
Bobu	9	$<10^{-4}$	19.8
Trpa	11	$<10^{-4}$	19.4
Sysp	29	$<10^{-4}$	16.2
Dera_1	37	$<10^{-4}$	28.0
Dera_2	5	$<10^{-4}$	24.2
Thma	5	$<10^{-4}$	5.4
Aqae	4	$<10^{-4}$	5.2
Aepe	10	$<10^{-4}$	12.0
Suso	18	$<10^{-4}$	12.0
Arfu	15	$<10^{-4}$	13.8
Hasp	22	$<10^{-4}$	21.8
Meth	71	$<10^{-4}$	91.0
Meja	24	$<10^{-4}$	28.8
Pyab	10	$<10^{-4}$	11.4
Pyfu	8	$<10^{-4}$	8.4
Pyho	11	$<10^{-4}$	12.6
Thac	4	$<10^{-4}$	5.2
Thvo	4	$<10^{-4}$	5.0

[a]Abbreviations and order are those used in Table 1.
[b]Observed number of two-copy CDR.
[c]Probability of finding N_{2CDR} or more under a random model. In the random model, one can estimate the probability of finding at least N_{2CDR} in N_2 two-copy direct repeats. This probability is $1 - B(0) + \ldots + B(N_{CDR} - 1)$, where $B(n)$ is the probability of finding n CDR in N direct repeats using a binomial law where the frequency of CDR is $2000/L$ for circular chromosomes and $2000/L - (1000/L)^2$ for linear ones (L = chromosome length). At an α risk of 10^{-4}, we assumed that 52/53 chromosomes are over-represented in CDR (with a risk of 0.005 of getting one or more false positives).
[d]Density in number (copies/Mb) for two-copy CDR.

Materials and Methods). As predicted by the model, CDR are over-represented in all chromosomes, taking into account the number of repeats (Table 3). The only exception is *Buchnera* sp., for which there are few CDR repeats, but it is unclear if this is a statistical artifact or has biological meaning. The *Buchnera* sp. genome is thought to be undergoing reductive evolution (3) and lacks an evident RecA homologue (29). Further, there is evidence that intracellular bacteria are subject to weaker selection (30). Thus, the absence of CDR could be the result of the reductive evolution process. Even if CDR are created, selection will not prevent them from being deleted. This deletion could arise easily since CDR deletion is mainly RecA independent.

Identity and length are constrained by spacer size. We looked for correlations between identity and spacer size within two-copy CDR for species in which there were at least 20 CDR (24 chromosomes). In 18 chromosomes identity was significantly negatively correlated with spacer size ($P < 0.01$, Table 4). In order to extend our analysis, we also took into account multi-copy repeats for chromosomes with less than 20 two-copy CDR or for those exhibiting a non-significant correlation for two-copy CDR (17 + 6 chromosomes). However, because the number of couples increases when families become very large [$c = n \times (n - 1)/2$, where c is the number of couples and n the number of copies], we retained only repeats with between two and five copies. This test identified significant positive correlations for 15 additional chromosomes ($P < 0.01$). Thus, out of the 41 chromosomes tested, 33 exhibited a significant negative correlation between identity and spacer size. Table 4 suggests that many others are weakly correlated.

Correlations between length and spacer size were tested under the same conditions as for identity (Table 5) and were also in agreement with the model. A negative correlation was found in 24 of the 41 chromosomes at $P < 0.01$ and in nine further chromosomes at a less significant α level ($P < 0.05$). Although very significant, these results are weaker than for the correlation between identity and spacer size and this deserves some comment. In the model, interspersed repeats are mostly created as identical tandem repeats, but their size can vary. Successive rounds of recombinational exchange constrain these repeats to be both highly identical and small due to the deletion bias mentioned above. Therefore, while the conversion process only maintains the pre-existing characteristics of the repeats (a high identity), the deletion process establishes an additional new constraint (small length). It is then conceivable that more rounds of exchange are required to establish the correlation between length and spacer size, thereby justifying weaker correlations.

Is tandem repeat creation modulated by chromosomal characteristics?

Since the previous results suggest the adequateness of our model, we proceeded to test the influence of chromosomal features on the duplication process, and in particular of nucleotide composition biases. Bacterial chromosomes exhibit large differences in their nucleotide composition, especially in terms of G + C composition, which can vary from 25 to 75% (24). We used the information entropy to measure the composition bias and found a significant negative correlation between entropy (and then composition bias) and the density of two-copy repeats, D_{N2} ($\tau = -0.34$, $P < 10^{-3}$, Fig. 4), as well as with total repeat densities, D_N ($\tau = -0.34$, $P < 10^{-3}$, Fig. 4). One would expect more biased random chromosomes to be more repetitive, since they use a subset of the possible symbols more frequently. However, our methodology to search for repeats already tackles this effect: we determined threshold scores

Table 4. Correlations between identity and spacer size for CDR

2-copy CDR[a]				2-, 3-, 4-, and 5-copy CDR[e]			
Species[b]	CDR	τ[c]	p[d]	Species[b]	CDR	τ[c]	p[f]
Baha	26	-0.31	0.013	*Aepe*	20	-0.06	0.371
Basu	31	-0.42	$<10^{-3}$	*Arfu*	38	-0.33	0.002
Cacr	37	-0.10	0.190	*Baha*	57	-0.38	$<10^{-3}$
Dera_1	37	-0.24	0.021	*Dera_1*	64	-0.29	$<10^{-3}$
EscoK	23	-0.54	$<10^{-3}$	*Bobu*	21	-0.57	$<10^{-3}$
EscoO	34	-0.50	$<10^{-3}$	*Cacr*	63	-0.27	$<10^{-3}$
Hain	22	-0.73	$<10^{-3}$	*Caje*	31	-0.39	0.001
Hasp	22	-0.17	0.141	*Chpn*	24	-0.34	0.011
Hepy	28	-0.21	0.066	*ChpnJ*	21	-0.31	0.024
HepyJ	35	-0.37	$<10^{-3}$	*Hasp*	44	-0.15	0.071
Lala	26	-0.51	$<10^{-3}$	*Hepy*	88	-0.15	0.022
Meja	24	-0.07	0.318	*Meja*	63	-0.19	0.015
Melo	71	-0.33	$<10^{-3}$	*Myge*	41	-0.15	0.090
Meth	71	-0.41	$<10^{-3}$	*Myle*	26	-0.72	$<10^{-3}$
Mytu	34	-0.52	$<10^{-3}$	*Mypn*	38	-0.26	0.010
MytuC	33	-0.48	$<10^{-3}$	*MypuC*	63	-0.40	$<10^{-3}$
NemeM	33	-0.43	$<10^{-3}$	*Pyab*	21	-0.39	0.008
NemeZ	26	-0.41	$<10^{-3}$	*Pyho*	21	-0.47	0.002
Pamu	23	-0.50	0.002	*Stpy*	42	-0.35	$<10^{-3}$
Psae	51	-0.31	$<10^{-3}$	*Suso*	40	-0.42	$<10^{-3}$
StauM	30	-0.34	0.004	*Sysp*	77	-0.32	$<10^{-3}$
StauN	29	-0.32	0.007	*Trpa*	22	-0.49	$<10^{-3}$
Sysp	29	-0.27	0.021	*Vich_1*	61	-0.10	0.130
Xyfa	46	-0.53	$<10^{-3}$				

[a]24 chromosomes with more than 20 two-copy CDR were used to test correlations.
[b]Abbreviations are those used in Table 1.
[c]Coefficients of Kendall τ rank tests between spacer size and identity for CDR.
[d]Probability associated with the Kendall τ rank tests. We assumed that, at an α risk of 0.01, 18/24 correlations are significant (with a risk of 0.21 of getting at least one false positive and of 0.02 of getting at least two false positives).
[e]We used two-copy, three-copy, four-copy and five-copy CDR to test 17 new chromosomes and re-test the six non-significant ones, where the two-copy CDR were less than 20 or were the tested correlation was $P < 0.01$.
[f]Probability associated with the Kendall τ rank tests. We assumed that, at an α risk of 0.01, 15/23 correlations are significant (with a risk of 0.21 of getting at least one false positive and of 0.02 of getting at least two false positives).

Table 5. Correlations between length and spacer size for CDR

2-copy CDR[a]				2-, 3-, 4-, and 5-copy CDR[e]			
Species[b]	CDR	τ[c]	p[d]	Species[b]	CDR	τ[c]	p[f]
Baha	26	0,46	$<10^{-3}$	*Aepe*	20	0,19	0.134
Basu	31	0,36	0,002	*Arfu*	38	0,39	$<10^{-3}$
Cacr	37	0.18	0.061	*Bobu*	21	0,50	$<10^{-3}$
Dera_1	37	0.16	0.089	*Cacr*	63	0,27	0.001
EscoK	23	0,50	$<10^{-3}$	*Caje*	31	0,30	0.009
EscoO	34	0,39	$<10^{-3}$	*Chpn*	24	-0,08	0.310
Hain	22	0,38	0,007	*ChpnJ*	21	0,01	0.488
Hasp	22	0.10	0.276	*Dera_1*	64	0,10	0.134
Hepy	22	0.28	0.018	*Hasp*	44	0,07	0.254
HepyJ	35	0,47	$<10^{-3}$	*Hepy*	88	0,25	$<10^{-3}$
Lala	26	0,26	0.033	*Lala*	66	0.18	0.019
Meja	24	0.12	0.212	*Meja*	63	0,19	0.015
Melo	71	0,30	$<10^{-3}$	*Myge*	41	-0,01	0.451
Meth	71	0,23	0.002	*Myle*	26	0,56	$<10^{-3}$
Mytu	34	0,41	$<10^{-3}$	*Mypn*	38	0,18	0.059
MytuC	33	0,38	$<10^{-3}$	*MypuC*	63	0,41	$<10^{-3}$
NemeM	33	0,14	0.127	*NemeM*	81	0.19	0.006
NemeZ	26	0,28	0.026	*NemeZ*	67	0.21	0.006
Pamu	23	0,32	0.018	*Pamu*	62	0.36	$<10^{-3}$
Psae	51	0,24	0.006	*Pyab*	21	0,20	0.113
StauM	30	0,35	0.003	*Pyho*	21	0,28	0.040
StauN	29	0,30	0.012	*StauN*	128	0.22	$<10^{-3}$
Sysp	29	0,19	0.079	*Stpy*	42	0,17	0.055
Xyfa	46	0,55	$<10^{-3}$	*Suso*	40	0,13	0.121
				Sysp	77	-0.05	0.264
				Trpa	22	0,35	0.011
				Vich_1	61	0,08	0.198

[a]24 chromosomes with more than 20 two-copy CDR were used to test correlations.
[b]Abbreviations are those used in Table 1.
[c]Coefficients of Kendall τ rank tests between spacer size and length for CDR.
[d]Probability associated with the Kendall τ rank tests. We assumed that, at an α risk of 0.01, 13/24 correlations are significant (with a risk of 0.21 of getting at least one false positive and of 0.02 of getting at least two false positives).
[e]We used two-copy, three-copy, four-copy and five-copy CDR to test 17 new chromosomes and re-test the 11 non-significant ones.
[f]Probability associated with the Kendall τ rank tests. We assumed that, at an α risk of 0.01, 11/27 correlations are significant (with a risk of 0.24 of getting at least one false positive and of 0.03 of getting at least two false positives).

based on empirical distributions for each genome and also defined specific scoring matrices, calculated taking into account the nucleotide compositions of the genomes (see Materials and Methods). This is why the minimal significant alignment score is larger for more biased genomes, such as some *Mycoplasma* spp. Since methodological biases were taken into account in the search for repeats, one is inclined to explain these results from a biological point of view.

Whatever the mechanism of tandem repeat genesis, it always requires pre-existing small repeats (11). Levinson and Gutman (8) have proposed that small repeats appear by chance and are at the origin of larger repeats that are created by slipped strand mispairing between these small repeats. It so happens that low complexity genomes, by chance alone, present a larger number of small repeats. If we accept the hypothesis that tandem genesis mechanisms are not down-regulated in low complexity genomes, then we are immediately led to the conclusion that tandem genesis must be more frequent in these genomes, simply due to their higher compositional bias. Thus, we propose that in such genomes a higher number of primers appear by chance and lead to more abundant repeats.

Small, non-duplicated repeats can be used as primers for initiation of tandem duplications. Thus, many types of repeats are related: small repeats are transformed into tandem repeats, which are then turned into interspersed repeats. As a consequence one gains by analysing these repeats together, instead of dividing them into different classes.

In this respect, it is interesting to note that chromosomes 2 and 3 of *Plasmodium falciparum* exhibit a very high density of repeats (as compared with eukaryote chromosomes of the same size) (16) which is associated with a very low G + C content (18%). It is therefore tempting to suggest that in eukaryote chromosomes complexity of the genome also plays an important role in the mechanisms of repeat generation. Naturally, the statistical testing of this generalisation will have to await the availability of a larger sample of complete eukaryote genomes.

CONCLUSION

We have shown that a model for the dynamics of repeats (previously established in Eukarya), based on tandem genesis with further dispersion, holds for most Bacteria and Archaea. As predicted by the model, we show that in most genomes (i) direct repeats are more numerous than inverted repeats, (ii) CDR are in large excess, (iii) there is a negative correlation between repeat identity and spacer size and (iv) there is a positive correlation between repeat length and spacer size. This strongly suggests that despite their diversity, intrachromosomal repeats of

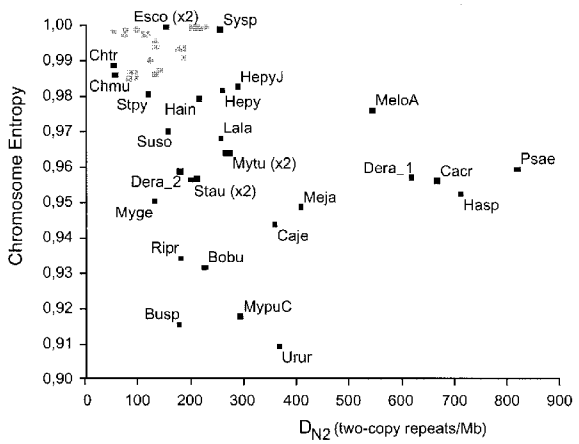

Figure 4. Complexity of chromosomes as a function of repeat density. Entropy (a measure of nucleotide complexity) of each of the 53 chromosomes as a function of their global repeat density. Entropy measures the nucleotide complexity of a sequence: if each nucleotide frequency is 0.25, then entropy is maximum (1), else it is lower. This figure illustrates that entropy is negatively correlated with repeat density.

all genomes share similar dynamics that are probably related to very ancient mechanisms shared by the three domains of life. Naturally, this model is not exclusive of other mechanisms of duplication (transposition, horizontal gene transfer, insertions, hyperploidisation, etc.).

We have also shown that nucleotide composition biases of the chromosome strongly influence the rate of tandem repeat creation and thus the rate of repeat amplification. Other effects are likely to shape the dynamics of bacterial repeats and the large availability of complete genomes will shed light on them. This will certainly provide new clues in deciphering the dynamics of repeats in bacterial genomes and shed additional light on genome evolution.

ACKNOWLEDGEMENTS

We would like to thank I. Gonçalves, D. Higuet, E. Maillier and J. Pothier for their scientific help and their friendly support. We would also like to thank P. Avner and E. Leguern for their helpful remarks on previous versions of this manuscript. This work was supported by grants from the Association pour la Recherche sur le Cancer. G.A. was funded by the Fondation pour la Recherche Médicale. E.C. and P.N. are members of Université Pierre et Marie Curie (Paris, France).

REFERENCES

1. Ohno,S. (1970) *Evolution by Gene Duplication*. Springer-Verlag, Heidelberg, Germany.
2. Wolfe,K.H. (2001) Yesterday's polyploids and the mystery of diploidization. *Nature Rev. Genet.*, **2**, 333–341.
3. Andersson,S.G. and Kurland,C.G. (1998) Reductive evolution of resident genomes. *Trends Microbiol.*, **6**, 263–268.
4. 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**, 1161–1167.
5. Le Fleche,P., Hauck,Y., Onteniente,L., Prieur,A., Denoeud,F., Ramisse,V., Sylvestre,P., Benson,G., Ramisse,F. and Vergnaud,G. (2001) A tandem repeats database for bacterial genomes: application to the genotyping of *Yersinia pestis* and *Bacillus anthracis*. *BMC Microbiol.*, **1**, 2.
6. Levinson,G. and Gutman,G.A. (1987) High frequencies of short frameshifts in poly-CA/TG tandem repeats borne by bacteriophage M13 in *Escherichia coli* K-12. *Nucleic Acids Res.*, **15**, 5323–5338.
7. van Belkum,A., van Leeuwen,W., Scherer,S. and Verbrugh,H. (1999) Occurrence and structure-function relationship of pentameric short sequence repeats in microbial genomes. *Res. Microbiol.*, **150**, 617–626.
8. Levinson,G. and Gutman,G.A. (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol. Biol. Evol.*, **4**, 203–221.
9. Yeramian,E. and Buc,H. (1999) Tandem repeats in complete bacterial genome sequences: sequence and structural analyses for comparative studies. *Res. Microbiol.*, **150**, 745–754.
10. Mahillon,J. and Chandler,M. (1998) Insertion sequences. *Microbiol. Mol. Biol. Rev.*, **62**, 725–774.
11. Romero,D. and Palacios,R. (1997) Gene amplification and genomic plasticity in prokaryotes. *Annu. Rev. Genet.*, **31**, 91–111.
12. Rocha,E.P.C., Danchin,A. and Viari,A. (1999) Analysis of long repeats in bacterial genomes reveals alternative evolutionary mechanisms in *Bacillus subtilis* and other competent prokaryotes. *Mol. Biol. Evol.*, **16**, 1219–1230.
13. Kraus,E., Leung,W.Y. and Haber,J.E. (2001) Break-induced replication: a review and an example in budding yeast. *Proc. Natl Acad. Sci. USA*, **98**, 8255–8262.
14. Paques,F. and Haber,J.E. (1999) Multiple pathways of recombination induced by double-strand breaks in *Saccharomyces cerevisiae*. *Microbiol. Mol. Biol. Rev.*, **63**, 349–404.
15. Achaz,G., Coissac,E., Viari,A. and Netter,P. (2000) Analysis of intrachromosomal duplications in yeast *Saccharomyces cerevisiae*: a possible model for their origin. *Mol. Biol. Evol.*, **17**, 1268–1275.
16. Achaz,G., Netter,P. and Coissac,E. (2001) Study of intrachromosomal duplications among the eukaryote genomes. *Mol. Biol. Evol.*, **18**, 2280–2288.
17. Chedin,F., Dervyn,E., Dervyn,R., Ehrlich,S.D. and Noirot,P. (1994) Frequency of deletion formation decreases exponentially with distance between short direct repeats. *Mol. Microbiol.*, **12**, 561–569.
18. Lovett,S.T., Gluckman,T.J., Simon,P.J., Sutera,V.J. and Drapkin,P.T. (1994) Recombination between repeats in *Escherichia coli* by a recA-independent, proximity-sensitive mechanism. *Mol. Gen. Genet.*, **245**, 294–300.
19. Bi,X. and Liu,L.F. (1996) recA-independent DNA recombination between repetitive sequences: mechanisms and implications. *Prog. Nucleic Acid Res. Mol. Biol.*, **54**, 253–292.
20. Peeters,B.P., de Boer,J.H., Bron,S. and Venema,G. (1988) Structural plasmid instability in *Bacillus subtilis*: effect of direct and inverted repeats. *Mol. Gen. Genet.*, **212**, 450–458.
21. Kurtz,S. and Schleiermacher,C. (1999) REPuter: fast computation of maximal repeats in complete genomes. *Bioinformatics*, **15**, 426–427.
22. Karlin,S. and Ost,F. (1985) Maximal segmental match length among random sequences from a finite alphabet. In Cam,L.M.L. and Olshen,R.A. (eds), *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*. Association for Computing Machinery, New York, NY, Vol. 1, pp. 225–243.
23. Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
24. Sueoka,N. (1962) On the genetic basis of variation and heterogeneity of DNA base composition. *Proc. Natl Acad. Sci. USA*, **48**, 582–592.
25. Schneider,T.D., Stormo,G.D., Gold,L. and Ehrenfeucht,A. (1986) Information content of binding sites on nucleotide sequences. *J. Mol. Biol.*, **188**, 415–431.
26. Hayashi,T., Makino,K., Ohnishi,M., Kurokawa,K., Ishii,K., Yokoyama,K., Han,C.G., Ohtsubo,E., Nakayama,K., Murata,T. *et al.* (2001) Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Res.*, **8**, 11–22.
27. Rocha,E.P.C. and Blanchard,A. (2002) Genomic repeats, genome plasticity and the dynamics of *Mycoplasma* evolution. *Nucleic Acids Res.*, **30**, 2031–2042.
28. Coissac,E., Maillier,E. and Netter,P. (1997) A comparative study of duplications in bacteria and eukaryotes: the importance of telomeres. *Mol. Biol. Evol.*, **14**, 1062–1074.
29. Shigenobu,S., Watanabe,H., Hattori,M., Sakaki,Y. and Ishikawa,H. (2000) Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS. *Nature*, **407**, 81–86.
30. Ochman,H. and Moran,N.A. (2001) Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis. *Science*, **292**, 1096–1099.

C.2. Résumé des résultats et de la discussion.

Cette étude nous a permis de détecter dans la plupart des génomes de nombreuses répétitions. L'examen des densités de répétitions (nombre de copies/Mb) dans tous les chromosomes bactériens (table 2, article 3) met en évidence que :

- La plupart des espèces proches présentent une densité similaire.
- La densité des répétitions est positivement corrélée à la taille des génomes.

L'ensemble des répétitions à deux copies formant un ensemble relativement homogène de répétitions (quasiment plus d'IS, d'ARNr, d'ARNt et autres familles de répétitions multicopies), nous avons décidé de tester notre modèle de la dynamique des répétitions intrachromosomiques sur ce sous-ensemble de nos données. La prise en compte des répétitions multicopies est bien plus complexe. Par exemple, les événements de conversion potentiels deviennent particulièrement nombreux lorsque le nombre de copies des répétitions augmente.

Nous avons donc, sur les répétitions à deux copies, testé si (1) les CDR sont plus abondantes que ce que le hasard prédit, (2) il existe, pour les CDR, des corrélations entre identité, longueur et taille du *spacer*.

(1) Dans un chromosome circulaire de taille L, la proportion de CDR (nombre de CDR/nombre de répétitions directes) est 2x1000/L si la dispersion des répétitions est aléatoire. Dans un chromosome linéaire de taille L, cette proportion est $2000/L-(1000/L)^2$. On peut donc calculer, à partir du nombre de répétitions directes, le nombre de CDR attendu dans un génome aléatoire. En comparant ce nombre au nombre réel de CDR, nous avons pu montrer que , à l'exception de *Buchnera sp.*, tous les chromosomes bactériens possèdent un nombre de CDR très supérieur à celui prédit par le hasard (Table 3, article 3).

(2) Pour beaucoup des CDR des chromosomes bactériens, l'identité des répétitions est négativement corrélée à la taille du *spacer* (Table 4, article 3) et la longueur est positivement corrélée à la taille du *spacer* (Table 5, article 4). Certains génomes ne possèdent pas suffisamment de CDR pour que l'on puisse tester les corrélations.

Nous avons regardé la localisation des répétitions proches par rapport à celle des gènes. Contrairement aux CDR du génome de *S. cerevisiae* (voir article 1), les CDR des génomes bactériens ne sont, en général, pas plus localisées dans les gènes que ce que le hasard prédit. Comme les génomes bactériens soient quasi-totalement composés de gènes (en incluant tous les signaux nécessaires à leur expression), cela rend peut-être la mesure caduque. En conclusion, les résultats suggèrent que notre modèle est valide pour la plupart des génomes bactériens. Cela suggère que les mécanismes qui sous-tendent notre modèle sont des mécanismes très anciens de gestion de l'ADN, apparus avant la divergence entre Bactéries, Archées et Eucaryotes.

En second lieu, nous avons cherché à mesurer l'influence de la composition nucléotidique des chromosomes bactériens sur le processus d'amplification. Nous avons utilisé l'entropie de Shannon (Schneider *et al.* 1986) pour mesurer le déséquilibre de composition. Cette entropie mesure l'écart à l'équirépartition des nucléotides. Ainsi, si la fréquence de chacun des nucléotides est de 0,25, l'entropie vaut 1. Si les fréquences ne sont pas équilibrées, l'entropie prend une valeur plus faible (jusqu'à 0,91 pour *Ureaplasma urealiticum*). Nous avons mis en évidence une corrélation négative entre l'entropie et la densité de répétitions (τ=-0,34 ; $p<10^{-3}$; Kendall-tau). Nous proposons pour expliquer ce résultat que dans les génomes les plus biaisés (entropie la plus faible), plus de répétitions fortuites apparaissent par simple juxtaposition de nucléotides. Les mécanismes proposés pour la création des répétitions nécessitent souvent une amorce, une petite zone de similarité préexistante (Introduction, chapitre A). Nous proposons donc que les chromosomes de faible entropie ont, par le fruit du hasard, plus d'amorces pour la duplication et ainsi créent plus souvent des séquences répétées. La composition en nucléotide serait donc un facteur clef déterminant la densité de répétitions d'un chromosome.

C.3. Répétition et biais de réplication.

On peut faire l'hypothèse que les génomes possédant une forte densité de répétitions sont plus souvent soumis à des réarrangements chromosomiques. Ce sont des chromosomes supposés instables. Une des mesures de la stabilité des chromosomes bactériens est la force

du biais de réplication. Le biais de réplication induit une différence de composition entre le brin tardif et le brin précoce au cours de la réplication. La conséquence de ce biais de réplication est donc une différence de composition de G et C (ou A et T) entre les deux demi-chromosomes (brins tardifs et précoces). Le biais de réplication est souvent mesuré par le 'GC skew' (Lobry 1996). Une étude plus récente a utilisé une analyse discriminante linéaire pour mesurer ce biais dans beaucoup de chromosomes bactériens (Rocha *et al.* 1999c). Ce biais est très variable d'une espèce à une autre : certains chromosomes, comme celui de *Borrelia burgdorferi*, présentent un biais très fort alors que ce biais est absent dans le génome de *Methanococus janaschii*. Une explication avancée pour expliquer l'absence de biais est que certains génomes qualifiés d'instables subiraient trop de réarrangements pour pouvoir présenter un biais fort (Rocha *et al.* 1999b). En estimant que la densité de répétitions est une mesure de l'instabilité des chromosomes, nous avons recherché une relation entre la force du biais et la densité des répétitions. Nous avons pu mettre en évidence une corrélation négative entre densité et biais de réplication (τ=-0.31, $p<0.01$, test de rang de Kendall).

Pour préciser les relations existant entre biais de réplication et répétitions, nous avons divisé les répétitions en quatre catégories. Ces catégories sont établies en fonction de l'orientation des copies (directes – D- ou inversées –I-) et de leur localisation par rapport aux deux demi-chromosomes (dans le même –1- ou une copie chacun –2-). Ces quatre catégories sont décrites dans la figure 39.

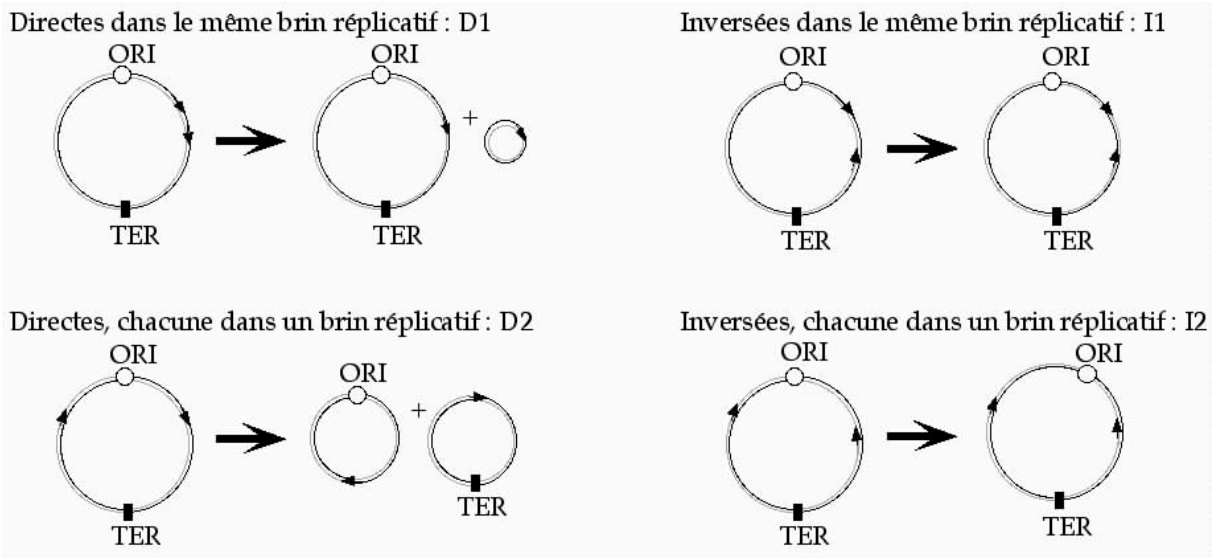

Figure 39 : Structure des quatre catégories de répétitions D1, D2, I1 et I2 et les conséquences d'un crossing-over entre leurs deux copies.

Afin d'étudier les différences entre les quatre catégories, nous avons mesuré la répartition des répétitions dans ces quatre catégories en fonction du biais de réplication.

Espèces[a]	Médiane du biais	Répétitions directes		Répétitions inversées	
		D1	D2	I1	I2
Meja Meth Urur Mypn MypuC HepyJ Hepy MytuC Cacr	0,63	1351	10706	10403	10089
MytuH EscoO Hain NemeZ EscoK NemeM Lala Psae Basu	0,71	88713	79596	83903	82271
Xyfa Ripr Baha Vich_1 StauM StauN Vich_2 Caje Busp	0,82	16808	2721	2754	4328
Stpy Myle Chpn ChpnA ChpnJ Trpa Chmu Chtr Bobu	0,90	1045	550	599	645

Tableau 7 : Répétitions en fonction du biais de réplication
a) le nom des espèces est celui utilisé dans la table 1 de l'article 3

Pour cela, nous avons utilisé 36 chromosomes dont l'ORI et TER sont bien définis et nous avons compté les répétitions de chaque catégorie. Les chromosomes ont ensuite été répartis en quatre groupes de même effectif : chromosomes « pas biaisés », « peu biaisés », « biaisés » et « très biaisés » (tableau 7). Un test de chi2 d'homogénéité montre que la répartition n'est pas la même en fonction du biais de réplication ($p<10^{-4}$).

Groupes[c]	Répétitions directes				Répétitions inversées			
	D1		D2		I1		I2	
	Ratio[a]	Chi2[b]	Ratio[a]	Chi2[b]	Ratio[a]	Chi2[b]	Ratio[a]	Chi2[b]
Pas biaisés	1,01	1,7	1,05	28,9	0,98	4,0	0,95	22,1
Peu biaisés	0,90	894,1	1,04	112,5	1,05	189,8	1,03	80,4
Biaisés	2,15	10392,1	0,45	1870,9	0,43	2048,8	0,68	640,6
Très biasés	1,26	54,3	0,84	15,6	0,88	9,4	0,95	1,5

Tableau 8 : Chi2 d'homogénéité des répétitions en fonction du biais de réplication.
a) Ratio entre valeurs observées et attendues (observé/théorique). b) Valeur du chi2. c) Les groupes sont ceux du tableau précédent.

L'examen du tableau 8 montre que :

- Dans les chromosomes pas et peu biaisés, les répétitions sont réparties dans les quatre catégories (avec un excès léger de D1).
- Dans les génomes biaisés, on observe un très grand excès de D1. Cela suggère que, dans ces génomes, les répétitions sont créées en tandem, mais ne subissent pas beaucoup de réarrangements pouvant e*spacer* les deux copies. Il faut également noter que les répétitions I2 sont moins sous-représentées que les autres (D2 et I1).
- Dans les génomes très biaisés, l'excès de D1 est moins important. Il faut noter que I2 est encore moins sous-représenté que D2 et I1.

Plusieurs explications doivent concourir à l'établissement d'une telle différence dans la répartition des répétitions dans les catégories en fonction du biais.

L'excès de répétitions en D1, pour les génomes peu ou pas biaisés, est attendu dans notre modèle, puisque les répétitions sont majoritairement créées en tandem.

Dans les génomes biaisés, il faut admettre que les réarrangements sont rares (sinon le biais de réplication serait perdu). Donc, si les répétitions nouvellement créées sont peu réarrangées, elles restent proches et directes, ce qui augmente la différence entre D1 et les autres catégories. Un crossing-over entre deux répétitions I2 situées à une distance similaire

de ORI (ou TER) ne change pas beaucoup le biais de réplication. De tels réarrangements peuvent donc avoir lieu dans les génomes biaisés.

Par ailleurs, dans les génomes biaisés, si deux copies sont situées sur le même type de brin (tardif ou précoce), elles subissent des mutations de même nature (par exemple, un excès de C). A l'inverse, si deux copies sont situées chacune sur un type de brin, les mutations subies par chacune des copies tendent à être très souvent différentes. Dans ce second cas, les copies divergent plus vite et sont donc plus vite considérées comme non significatives.

Dans les génomes très biaisés, il faut émettre l'hypothèse que le processus de duplication est ralenti (ou celui de délétion accéléré). Ainsi, le nombre total de répétitions est plus faible, et l'excès de D1 apparaît moins fort.

IV. Discussion

Dans l'avant-propos nous avons montré qu'il faut répondre (au moins partiellement) à trois interrogations pour étudier un niveau de la Biologie : Quels sont les contraintes structurales ? Quelles sont les contraintes sélectives ? Quelle est la « liberté » laissée par les contraintes structuralo-sélectives ? Je m'efforcerai donc d'extraire de mon travail des bribes de réponses à ces trois questions. Cependant, je m'attarderai auparavant à démasquer les biais qui se sont introduit dans nos analyses. Cette première étape est cruciale car elle permet de conserver un esprit critique vis à vis du travail effectué. Les conclusions et les modèles que nous proposons ne sont pas des « vérités », mais des interprétations vraisemblables de nos observations.

A. Quels sont les biais ?

Trois principaux biais existent dans les analyses de données biologiques (comme les séquences d'ADN). Le premier provient du matériel utilisé. En effet, il faut envisager la possibilité d'erreurs dans les séquences d'ADN que nous utilisons (erreurs de séquence et erreurs d'assemblage) . Ces erreurs conduisent à une analyse correcte des données mais à une interprétation biologique fausse. Pour illustrer mon propos, je développerai un exemple d'erreur dans les séquences du génome de *C. elegans*. Le second type d'erreur est d'ordre méthodologique. Comme les analyses de génomes sont, pour la plupart, encore en phase d'exploration, ce type d'erreur est fréquent. Il faut donc lors de la mise en place d'une méthode tester l'existence de ces biais en utilisant des témoins (comme les génomes aléatoires). J'illustrerai ce type de biais par un problème méthodologique pour lequel nous n'avons pas encore une solution totalement satisfaisante. Le troisième est le biais d'interprétation. Au cours de notre étude, nous avons dû préférer certaines hypothèses à d'autres. Ces choix ont souvent été résolus parcimonieusement, mais nous ne sommes pas à l'abri d'avoir fait des « mauvais » choix. J'illustrerai mon propos par une brève discussion autour de quelques-uns uns de ces choix.

A.1. Exemple de biais matériel, le cas du génome de *C. elegans*.

Comme indiqué dans la partie *Matériel et Méthodes*, nous avons choisi d'utiliser les séquences publiques du site ftp de Genbank. Ce choix fut motivé par des raisons de

simplicité. En effet, le format des génomes dans Genbank étant toujours le même, il est facile de construire des routines qui transforment ce format en d'autres formats utilisés pour les analyses (par exemple le format *Fasta*).

En ce qui concerne le génome de *C. elegans*, les séquences déposées sur le site ftp de Genbank datent d'avril - mai 1999. Or, le Consortium International de séquençage de ce génome met à jour relativement souvent les séquences sur le site ftp de wormbase, une base de données dédiée à *C. elegans* (ftp.wormbase.org/pub/wormbase/DNA_DUMPS). Des versions précédentes des chromosomes sont disponibles sur le site ftp du Sanger Center (ftp.sanger.ac.uk/pub/databases/C.elegans_sequences/CHROMOSOMES).

Banque	Date	Caractéristiques globales			Répétitions		
		Taille (Mb)	Taille (sans N)	% de N	N2[b]	D_{N2}[a,c]	D_{L2}[a,d]
Genbank	23/04/99	16,18	14,75	8.85	1518	102,9	20,77
Sanger	11/11/98	14,02	13,84	1.32	974	70,37	3,61
Sanger	27/07/99	13,47	12,73	5.49	874	68,66	3,69
Sanger	16/02/00	14,83	14,27	3.77	1042	73,01	3,73
Sanger	21/11/01	15,05	14,99	0.41	1133	75,57	3,75
Wormbase	29/03/02	15,08	15,08	0.00	1141	75,66	3,75

Tableau 9 : Evolution de la séquence du chromosome I de *C. elegans* dans les banques de données.
a) calculée sur les séquences sans les « trous » (sans N). b) Nombre de deux copies. c) N2/taille (Mb). d) proportion (%) du chromosome occupée par les répétitions à deux copies.

Nous sommes allés rechercher toutes les séquences disponibles pour le chromosome I de *C. elegans* (la plus récente provient de Wormbase et les autres du Sanger Center). Dans toutes ces séquences, nous avons détecté les répétitions par la méthode utilisée dans le second article (méthode « eucaryote »). Les caractéristiques des différents chromosomes I et leurs contenus en répétitions sont présentés dans le tableau 9.

L'examen de ce tableau laisse songeur quant aux biais qui existent dans les données de séquences. Il semble qu'au cours du temps, les versions des chromosomes sont de plus en plus homogènes : la taille ne varie presque plus, les « trous » de séquence (N) sont peu à peu comblés et le nombre de répétitions varie moins.

Comme il semble que trois ans et demi après la publication de la séquence de *C. elegans* (TCESC 1998), la séquence commence à devenir fiable, nous avons recommencé une analyse brève des données de *C. elegans*. Les résultats de la version actuelle de la séquence de Wormbase sont comparés avec ceux obtenus à partir de la séquence de Genbank (tableau 10).

Chr	Genbank[a]				Wormbase[b]			
	Taille (sans N)	N2	D_{N2}	D_{L2}	Taille (sans N)	N2	D_{N2}	D_{L2}
I	14,8	1518	102,9	20,8	15,1	1141	75,7	3,8
II	16,6	1659	99,9	15,6	15,1	1303	85,9	5,0
III	11,6	1092	94,1	8,3	13,9	1126	81,3	3,5
IV	14,4	1418	98,5	13,5	17,5	1443	82,5	5,2
V	20,5	2135	104,1	10,3	20,9	1720	82,2	6,1
X	17,3	723	41,8	3,5	17,7	602	34,1	1,8

Tableau 10 : Résultats obtenus avec les nouvelles séquences de *C. elegans*.
a) Résultats présentés dans l'article 2, établis sur la séquence de Genbank (avril 1999). b) Résultats établis sur la dernière version de Wormbase (mars 2002)

Un examen attentif de ce tableau permet de confirmer deux des résultats que nous avons discutés dans l'article 2 :

- Les chromosomes de *C. elegans* présentent une densité de répétition (D_{N2}) relativement homogène entre eux. Cela suggère que le processus d'amplification (balance entre duplication et délétion) est la conséquence d'un mécanisme global de la cellule.
- Le chromosome X est sous-répété par rapport aux autres chromosomes de *C. elegans*. Cette seconde observation suggère un rôle possible de la recombinaison méiotique dans les processus d'amplification.

Par contre, il existe une très importante différence entre les séquences de Genbank et les versions actuelles des séquences : la proportion du chromosome comprise dans les répétitions à deux copies (D_{L2}) est très supérieure dans les séquences de Genbank. Cela est du à la présence de très grandes répétitions en tandem (jusqu'à 600 kb) dans les séquences de Genbank qui ont été retirées dans les nouvelles séquences. L'explication la plus probable est que ces répétitions en tandem très ressemblantes (~ 99%) sont des erreurs d'assemblage.

L'analyse des versions plus récentes des chromosomes des autres organismes utilisés dans notre étude est en cours. Elle nous permettra de découvrir si les versions plus récentes des séquences publiques réservent d'autres surprises.

A.2. Exemple de biais méthodologique, les matrices d'alignements.

Comme expliqué dans la partie *Matériel et Méthodes*, notre méthode de détection des répétitions est fondée sur la détection de répétitions strictes (graines) et leur extension par alignement local par programmation dynamique. Pour tenter de prendre en compte le biais de composition globale en nucléotides des génomes, nous avons utilisé une matrice de scores spécifique à chaque génome, construite grâce aux fréquences des nucléotides.

Nous avons, depuis cette première matrice, entrepris une étude plus approfondie des matrices possibles permettant de corriger les biais de composition. Une mesure intéressante du biais de composition est le dGC, l'écart du pourcentage en GC à 50. Pour caractériser les matrices, nous avons également utilisé la valeur *E (Expect)*, définie par $E = \sum_{i=A}^{T} S_{ij} p_i p_j$, où S_{ij} est le score du nucléotide i avec le nucléotide j et où p_i est la fréquence du nucléotide i. La valeur *E* indique le score moyen distribué par la matrice. Nous avons comparé les résultats d'alignements réalisés avec quatre matrices, la matrice «□1□» (matrice d'identité), la matrice «□1-p□» (décrite sur le tableau 11), la matrice «□1/p□» (ou le score est ± $1/p_ip_j$) et la matrice «□log(p)□» (où le score est ± $\log(p_ip_j)$).

Pour tester ces matrices, nous avons choisi 8 génomes bactériens représentatifs des différents biais de composition (de dGC □ 0 à dGC > 20). Pour chacun de ces génomes, nous avons construit 10 génomes aléatoires de même taille et de même composition en trinucléotides. Dans ces derniers, nous avons détecté les répétitions en utilisant une longueur minimum de graines de 15 bases. Pour chacune des matrices testées, toutes les graines sont étendues par alignement local. La moyenne de la longueur des répétitions est calculée et reportée sur le tableau 11.

Espèces[a]	dGC	Matrices[b]

		1	1-p	1/p	log(p)
Meth	0,46	18,33	17,45	18,93	17,50
Xyfa	2,67	18,19	18,18	18,76	18,49
Thac	4,01	18,39	18,06	18,75	18,12
Pyho	8,12	18,65	18,24	18,94	18,68
Mypn	9,99	18,96	18,23	19,06	18,41
Hain	11,85	19,31	18,39	19,39	18,78
Meja	18,57	21,90	19,65	21,87	18,79
Busp	23,69	28,37	22,69	25,17	19,63
Kendall-tau[c]	-	0,93	0,85	0,79	0,79
E_{min} / E_{max}[d]	-	-14 / -14	-30,5 / -48,7	-8 / -8	-11,7 / -13,8

Tableau 11 : Tests des différentes matrices corrigeant les biais de composition.
Dans ce tableau sont présentées les longueurs moyennes des répétitions détectées sur 8x10 génomes aléatoires de composition en nucléotides variable. a) les noms des espèces sont ceux définis en table 1 de l'article 3. b) les différentes matrices testées. c) Coefficient de corrélation entre la longueur moyenne et le dGC. d). Variation de E entre les différents génomes. La matrice est d'autant meilleure que E varie peu.

Les comparaisons des effets des différentes matrices sur la longueur moyenne indiquent que le choix d'une matrice par rapport à une autre influe fortement sur les résultats. Pour les génomes dont la composition en nucléotides n'est pas ou peu biaisé (dGC 0), le choix de la matrice a peu d'influence sur la longueur moyenne des répétitions. Cela peut s'expliquer par le fait que si les fréquences sont égales toutes les matrices convergent vers une matrice d'identité. A l'inverse, pour les génomes fortement biaisés, le choix de la matrice fait considérablement varier la longueur moyenne. Si l'on fait l'hypothèse que la composition en nucléotides ne doit pas changer cette longueur moyenne, on peut considérer que la matrice log(p) est plus corrective. Si l'on examine les valeurs des coefficients de corrélation de rang (Kendall-tau) entre longueurs et dGC, il apparaît qu'aucune des matrices testées ne corrige totalement le biais de composition globale en nucléotides d'une séquence. L'observation des valeurs *E*, indique que la matrice 1-p présente une variation importante de cette valeur. Cela signifie que le score moyen des matrices « 1-p » dépend du biais de composition du génome.

Il semble donc que la matrice « log(p) » est plus « juste » que la matrice 1-p. Nous avons recommencé la détection des répétitions à l'aide de la matrice log(p). Les résultats obtenus sont très similaires à ceux obtenus avec la matrice 1-p. Seuls les génomes très biaisés

(dGC>20) montrent une diminution du nombre de répétitions. Ces dernières ont, de plus, une longueur moyenne plus petite et une identité moyenne plus forte que les répétitions détectées avec la matrice 1-p.

A.3. Biais d'interprétation, quelques choix importants.

Une des premières observations que nous avons faites sur les répétitions est que si les CDR sont très abondantes, il n'existe quasiment pas de répétitions inversées proches (*Close Inverted Repeats* – CIR). Deux interprétations concurrentes sont possibles pour cette observation. Soit les répétitions ne sont pas créées inversées et proches, et les quelques CIR observées ne sont que le produit de remaniements les ayant rapprochées, soit les CIR sont très fortement contre-sélectionnées. Cette seconde hypothèse pourrait s'expliquer si l'on considère que les CIR forment des palindromes imparfaits susceptibles de se replier en épingle à cheveux dans les chromosomes ou dans les ARN. Nous avons pourtant préféré considérer qu'il n'existe pas de création de CIR. Ce choix est motivé principalement par un argument de parcimonie. En effet, il est toujours possible d'expliquer l'absence d'un objet biologique par sa création accompagnée de sa sélection négative, mais il semble plus simple de proposer que cet objet n'existe tout simplement pas. Par le hasard, quasiment aucune CIR n'est attendue ; c'est ce qui est observé. Par ailleurs, il ne semble pas exister, à notre connaissance, de mécanismes pouvant créer des CIR. Nous préférons donc l'hypothèse selon laquelle les répétitions inversées sont le fruit d'une dispersion par le hasard.

La seconde interprétation que nous avons été amenés à faire concerne la surabondance de CDR. Plusieurs choix sont possibles pour expliquer cette observation. Le premier est que un grand nombre de répétitions en tandem sont créées et peuvent (à la suite de petites délétions, de petites insertions ou de mutations) devenir des CDR. La seconde hypothèse, non exclusive de la première, est que les CDR sont maintenues par des pressions sélectives. Cette hypothèse semble étayée par l'observation, chez *S. cerevisiae*, de la localisation des CDR. Ces dernières sont le plus souvent localisées dans le même gène et codent pour des répétitions peptidiques. Cependant si les pressions fonctionnelles expliquent comment les CDR sont « maintenues » au cours de l'évolution, elles n'expliquent

pas comment elles « naissent ». Comme quasiment aucune CDR n'est attendue par la dispersion au hasard des répétitions, on ne peut pas penser que les CDR sont seulement issues de cette dispersion. Il semble également peu vraisemblable que les pressions fonctionnelles s'exercent plus fortement sur les CDR que sur les autres répétitions. L'hypothèse la plus parcimonieuse permettant d'expliquer cette abondance de CDR est donc l'existence d'un mécanisme actif de création de répétitions en tandem.

Le troisième et dernier choix que je développerai est celui de l'hypothèse des réarrangements de CDR en répétitions dispersées. Plusieurs observations semblent étayer cette hypothèse : (1) les répétitions éloignées sont, en moyenne, plus divergentes que les CDR. Cette observation peut également s'expliquer par la forte conversion qui s'exerce entre les copies des CDR. (2) On observe des traces de réarrangements de répétitions en tandem dans la plupart des génomes eucaryotes. Ceci tend à montrer que les réarrangements de répétitions en tandem se produisent dans les génomes (3) Le génome de *P. falciparum*, pressenti comme un génome issu d'une histoire récente, ne contient quasiment que des CDR et des répétitions subtélomériques. Toutes ces observations indiquent que les répétitions en tandem peuvent être réarrangées en répétitions dispersées. Cependant, nos résultats ne permettent pas de donner une idée de la fréquence de ces évènements de réarrangements.

B. Quelles sont les conclusions biologiques ?

Bien qu'il existe quelques biais dans les données et dans les méthodes que nous avons utilisées, il semble, après une analyse rétrospective, que la plupart des résultats et des conclusions que nous avons présentés paraissent vraisemblables. Dans l'*avant-propos*, nous avons exposé les différents niveaux de la Biologie (du niveau de *l'écologie* à celui de la *molécule*). Nous avons proposé que chacun de ces niveaux subisse deux types de contraintes (voir *Avant-propos*) tout en conservant une certaine « liberté ». Pour comprendre un niveau, il faut donc caractériser les deux types de contraintes et la liberté qui en resulte.

B.1. Quelles sont les contraintes structurales subies par les répétitions ? (les apports sur les mécanismes de duplication).

Les répétitions peuvent être considérées comme des objets biologiques macromoléculaires. Les contraintes structurales de ces répétitions sont donc celles imposées par l'agencement chimique des molécules. Une forme d'expression de ces contraintes est constituée des mécanismes moléculaires impliqués dans la dynamique des chromosomes. La quantité et les caractéristiques des répétitions résultent d'un équilibre entre duplication et délétion. Nos résultats ne permettent pas de proposer un mécanisme unique pour ces processus de duplication et délétion. Cependant, ils apportent des informations qui peuvent être interprétées pour comprendre ces mécanismes.

Le modèle que nous proposons sous-tend l'hypothèse que la plupart des répétitions intrachromosomiques proviennent de répétitions en tandem. Nous favorisons donc l'idée que l'un des mécanismes majoritaires de duplication est la duplication en tandem. Plusieurs mécanismes peuvent être envisagés pour la duplication en tandem : le dérapage au cours de la réplication (surtout proposé pour les expansions de SSR), le crossing-over inégal au cours de la mitose ou la méiose, l'excision réinsertion, et la réplication circulaire. Tous ces mécanismes requièrent une répétition directe proche préexistante formant ainsi un amplicon (voir *Introduction*). Après un événement de duplication, le corps de l'amplicon (la région située entre les deux copies de la répétition préexistante) et la répétition préexistante sont dupliqués. Dans le cas des SSR, le *spacer* est inexistant et donc seule la répétition est dupliquée.

Nos résultats de l'analyse des génomes eucaryotes indiquent que les chromosomes d'un même organisme ont une même densité de répétition. Cela suggère que le principal mécanisme de duplication est un phénomène global à tous les chromosomes. Le chromosome X de *C. elegans*, seul hémiploïde des chromosomes analysés, est largement sous-répété (Article 2, figure 1 et tableau 10). Cette seconde observation suggère que la recombinaison méiotique serait un mécanisme important de duplication dans les chromosomes eucaryotes. Cette hypothèse est renforcée par nos observations concernant la répartition des répétitions dans les chromosomes de *C. elegans*. Nous avons pu montrer que,

chez *C. elegans*, la densité de répétitions en tandem est plus forte dans les régions où le taux de recombinaison est plus élevé (figure 38).

Dans les génomes de *H. sapiens* et *C. elegans*, nous avons observé que les répétitions en tandem sont préférentiellement localisées dans les régions plus riches en GC (figure 37). Une étude récente montre une faible corrélation entre le taux de recombinaison et la proportion en GC (Fullerton *et al.* 2001). On peut donc émettre l'hypothèse qu'un fort taux de recombinaison conduit, d'une part, à un fort taux de GC et, d'autre part, à une plus grande densité de répétitions en tandem. Pour valider de telles propositions, il semble nécessaire de poursuivre nos investigations et de formaliser les relations entre recombinaison, proportion en GC et densité de répétitions en tandem.

Cette proposition de mécanisme ne rejette pas la possibilité que les autres mécanismes coexistent avec la recombinaison méiotique. La recombinaison mitotique, par exemple, pourrait également être impliquée dans les duplications. Cependant, la recombinaison mitotique est bien plus faible que la recombinaison méiotique (d'un facteur au moins 1000 chez *S. cerevisiae* (Paques and Haber 1999)). Si l'on considère le nombre de mitoses effectuées entre un zygote et un gamète (environ 10 chez *C. elegans*, 25 pour la femelle et 64 pour le mâle chez *M. musculus*, et 400 chez un homme de 30 ans, pour revue, voir (Drake *et al.* 1998)), il semble que la recombinaison mitotique soit moins « importante » que la recombinaison méiotique. La recombinaison mitotique semble, au moins chez *S. cerevisiae*, plus uniforme le long du chromosome que la recombinaison méiotique. Elle ne rend donc pas compte de la distribution non uniforme des répétitions en tandem le long des chromosomes.

Dans les génomes bactériens, il paraît difficile d'imaginer que la recombinaison méiotique joue un rôle important dans la genèse des répétitions en tandem ! Il faut donc envisager que dans les génomes bactériens, les répétitions en tandem sont issues des autres mécanismes. Cette hypothèse est étayée par deux différences notables entre les répétitions des génomes eucaryotes et celles des génomes bactériens :

- Dans les génomes bactériens possédant deux chromosomes, les deux chromosomes ne présentent pas une densité de répétitions homogène.
- La proportion des répétitions directes très proches (*spacer* de taille inférieure à 10 paires de bases) est très inférieure dans les génomes bactériens.

Ces observations indiquent que les mécanismes « majoritaires » sont différents dans les génomes eucaryotes et bactériens. Par ailleurs, la seconde observation suggère que chez les Bactéries et les Archées, la duplication en tandem est moins fréquente (ou le processus de délétion est plus actif). Cela pourrait être la conséquence des pressions sélectives fortes qui tendent à maintenir compacts les génomes bactériens.

B.2. Quelles sont les contraintes sélectives subies par les répétitions ? (les apports sur l'étude des fonctions).

Les contraintes sélectives qui s'imposent aux répétitions ne sont pas simples à déterminer. Les pressions fonctionnelles subies par ces répétitions sont de telles contraintes. L'étude des fonctions des répétitions permettent de caractériser ces pressions fonctionnelles. Nos résultats ne sont pas principalement tournés vers les fonctions des répétitions, mais nous avons pu approcher deux types de fonctions des répétitions.

Dans le premier article, nous avons montré que la plupart des CDR de *S. cerevisiae* sont localisées dans le même gène et codent pour des répétitions de peptides. Il est donc vraisemblable que ces répétitions codent pour des répétitions de domaines protéiques. Il serait intéressant de poursuivre cette étude sur les fonctions des répétitions dans les protéines. Par exemple, il faudrait analyser la localisation des répétitions de peptides : existe-t-il une caractéristique fonctionnelle commune à toutes les protéines contenant ce type de répétition ? Où sont localisées les répétitions dans les séquences des protéines ? Comment évoluent les deux copies d'une répétition peptidique ?

Dans le second article, nous avons proposé que les caractéristiques des répétitions de *P. falciparum* soit liées au statut particulier de ce génome. Ce dernier est très riche en répétitions en tandem et en répétitions subtélomèriques, mais possède peu de répétitions dispersées. Ces deux caractéristiques pourrait être la marque de l'évolution récente d'un

pathogène. En effet, certains pathogènes de l'homme ont une grande abondance de répétitions (voir par exemple (Rocha and Blanchard 2002)). Les pathogènes de l'homme doivent échapper au système immunitaire pour « survivre ». Une des stratégies adoptée par les pathogènes est la variation importante des protéines exposées à la membrane. Les répétitions peuvent augmenter la variabilité de ces protéines et ceci par deux effets :

- Elles peuvent servir de « réservoir » d'information à partir duquel le génome peut, par conversion ou recombinaison, modifier la séquence d'un locus particulier.
- Elles peuvent également être à l'origine de changement de cadre de lecture, permettant l'arrêt (ou le démarrage) de la traduction d'un gène.

Ainsi, la présence de nombreuses répétitions proches pourrait chez *P. falciparum* être la marque d'un pathogène.

B.3. Contrainte structurale ou sélective ? (biais de réplication et répétition).

Nous avons mis en évidence que, dans les génomes bactériens, les répétitions sont moins denses dans les génomes ayant un fort biais de réplication. Cela s'explique par le fait que le biais est une mesure de la stabilité des chromosomes et que la densité de répétitions est une mesure de l'instabilité.

On peut envisager que dans les génomes bactériens fortement biaisés, il existe une sélection importante qui tend à conserver une certaine stabilité du génome ou à diminuer la taille du génome. Cette seconde proposition semble particulièrement bien adaptée aux génomes minimaux, dont certains auteurs pensent qu'il vont vers une réduction de génome. Dans ce cas, la corrélation négative entre densité de répétitions et biais de réplication est la conséquence de contraintes structurales.

On peut également proposer que dans ces génomes, certains mécanismes de réarrangements ou de duplications sont moins actifs. Si la fréquence des duplications (ou celle des réarrangements) est réduite, le nombre de répétitions diminue. Ceci entraîne une diminution du nombre de réarrangements et donc une augmentation du biais de réplication.

Dans ce second cas, la corrélation observée est plutôt la conséquence de contraintes structurales.

Cet exemple illustre bien la difficulté d'interprétation des observations et montre qu'il est parfois délicat de bien séparer les contraintes structurales des contraintes sélectives.

B.4. Quels sont les apports sur l'étude de la dynamique des répétitions ?

Les contraintes structutales et sélectives subies par les répétitions définissent un champ d'évolution pour les répétitions. La dynamique des génomes peut être vue comme l'étude de la « liberté » laissée aux génomes par lesdites contraintes. Nous avons proposé un modèle d'évolution des séquences répétées dans les génomes eucaryotes et bactériens. Ce modèle postule que la plupart des répétitions naissent en tandem et sont dispersées par des remaniements chromosomiques postérieurs. Si certaines répétitions dispersées sont d'anciennes répétitions en tandem, cela signifie que les catégories de répétitions que nous définissons couramment sont interconnectées. Un schéma des relations entre la plupart des répétitions est proposé dans la figure 40.

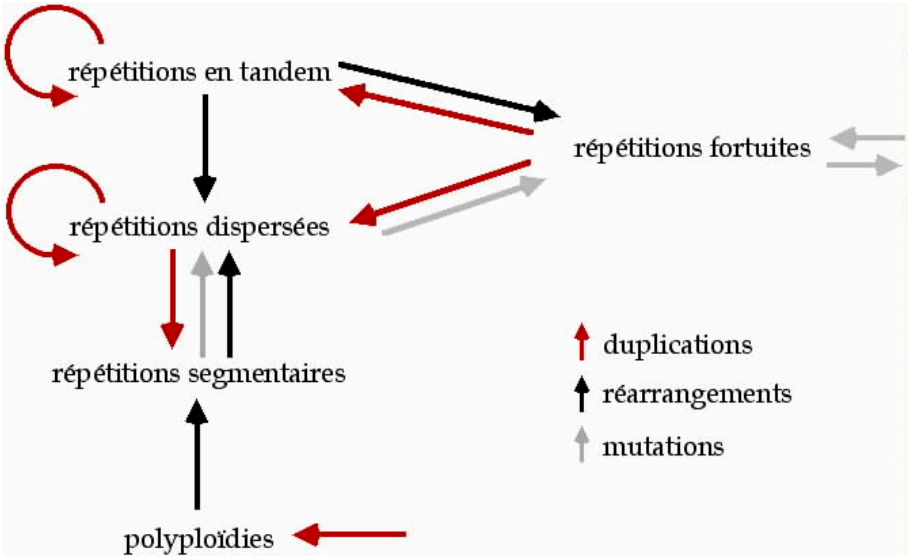

Figure 40 : Schéma illustrant la « transformation » des répétitions.
Dans ce schéma, les répétitions dérivent les unes des autres par des duplications, remaniements –délétions, inversions, translocations, etc.- et mutations. Dans ce schéma, toutes les répétitions proviennent soit de répétitions fortuites, soit de polyploïdies.

Dans ce schéma simpliste, toutes les répétitions dérivent de répétitions fortuites ou de polyploïdies. Si nous examinons les origines et les devenirs de chaque type de répétitions, nous pouvons proposer que :

- Les répétitions fortuites sont des répétitions de petite taille qui sont considérées comme non significatives. Elles peuvent apparaître *de novo* par des mutations ou être les reliquats de répétitions dispersées en train de disparaître. Elles servent d'amorces pour créer des répétitions en tandem ou dispersées. Elles peuvent disparaître par mutation.
- Les répétitions en tandem sont créées par des duplications amorcées par les répétitions fortuites. Elles peuvent être perdues par délétion. La taux de conversion que subissent ces répétitions en tandem est assez fort pour empêcher ces répétitions de disparaître ou de diverger par des mutations. Elles peuvent être remaniées en répétitions dispersées.
- Les répétitions dispersées forment une catégorie de répétitions très vaste regroupant différents types de répétitions. On peut mentionner, à titre d'exemple, les répétitions subtélomèriques, les répétitions interchromosomiques ou intrachromosomiques quelconques. Elles peuvent être créées par des mécanismes tels que la réparation des cassures double-brins qui nécessite une amorce de similarité. Cette amorce peut être une répétition fortuite ou une répétition dispersée. Par un mélange de mutations et réarrangements, elles peuvent provenir des répétitions segmentaires. Par ailleurs, elles peuvent être d'anciennes répétitions en tandem remaniées. Elles restent des répétitions dispersées et divergent par mutations avec peu de contraintes structurales. Elles disparaissent principalement par mutations.
- Les répétitions segmentaires proviennent le plus souvent de réarrangements impliquant des répétitions dispersées. En ce sens, elles dérivent des répétitions dispersées. Elles peuvent également être la marque de la diploïdisation d'une ancienne polyploïdie. Elles se transforment en répétitions dispersées par des remaniements et des mutations.

- Les polyploïdies forment le second type de répétitions qui peuvent potentiellement apparaître *de novo*. Elles sont remaniées par le processus de diploïdisation qui les transforment en répétitions segmentaires.

Le schéma présenté ci-dessus est sans aucun doute une vision très simplifiée d'une réalité complexe. Il fait pourtant ressortir plusieurs points intéressants.

Tout d'abord la plupart des répétitions sont issues d'autres types de répétitions. Cela constitue la « transformation » des répétitions. Ce processus est intimement lié aux réarrangements chromosomiques, causes majeures de cette transformation. La présence de répétitions dans un génome témoigne donc de l'activité et de l'instabilité de ce génome. Cette idée est très bien illustrée par l'existence d'une corrélation négative entre biais de réplication et densité de répétitions dans les génomes bactériens (voir *Résultats*). En effet, le biais de réplication témoigne de la stabilité d'un chromosome et la densité des répétitions de son instabilité.

La plupart des répétitions dérivent des répétitions fortuites. Or ces répétitions fortuites n'ont pas la même chance d'apparaître dans tous les génomes. Dans certains génomes, où le biais de composition est plus fort, les répétitions fortuites sont formées plus facilement. Si ces répétitions fortuites sont plus fréquentes, cela augmente la densité totale des répétitions. Ceci est illustré par la corrélation positive que nous observons entre densité de répétitions et biais de composition en nucléotides dans les génomes bactériens.

B.5. Les limites de l'analyse.

Le modèle d'évolution des répétitions décrit ci-dessus semble pouvoir rendre compte de la plupart des observations faites sur les séquences répétées. Cependant, la plupart des paramètres de ce modèles ne sont pas encore définis. Par exemple, nous n'avons pas d'informations pouvant nous renseigner sur la proportion de répétitions dispersées qui sont originaires (1) des répétitions fortuites (insertion), (2) des répétitions en tandem (réarrangement) et (3) des répétitions segmentaires (réarrangement ou mutations). Ces trois

types de genèse des répétitions dispersées sont très vraisemblablement toutes à l'œuvre dans les différents génomes, mais nous ne savons pas :

- Quels sont les taux relatifs de ces trois genèses de répétitions dispersées. Ces taux relatifs ne sont probablement pas les mêmes dans tous les génomes. Par exemple, les répétitions segmentaires n'existent pas dans les génomes bactériens, mais sont nombreuses dans les génomes de *S. cerevisiae* ou de *H. sapiens*.
- Quels est le temps nécessaire à la mise en place des répétitions dispersées dans les génomes. L'exemple du génome de *P. falciparum*, dont l'évolution semble récente (10 000 ans), nous suggère que la création de CDR est plus rapide que la création des répétitions dispersées. La comparaison de génomes bactériens phylogénétiquement proches nous renseignera sur les temps relatifs de réarrangement dans certains génomes bactériens.

Bien d'autres questions concernant notre modèle sont encore sans réponse. L'avancement des analyses des répétitions soulèvera peut-être d'autres limites et contradictions avec ce modèle. Ce dernier est donc bien une proposition vraisemblable de modèle pour la dynamique des répétitions qui permet d'interpréter parcimonieusement les observations. Il ne s'agit en aucun cas d'une « vérité » inamovible.

B.6. Quelques perspectives.

Ce travail a permis d'éclairer certaines facettes de la biologie moléculaire et de replacer les répétitions comme un outil de choix dans l'étude de la dynamique et l'évolution des génomes. Il ne constitue évidement pas une fin et de multiples voies sont envisageables pour le poursuivre.

Une première voie est la comparaison des répétitions dans des espèces proches. Ce type de travail est désormais possible en utilisant les génomes bactériens. En effet, dans ces derniers, il est possible de trouver un large éventail de distances phylogénétiques entre les organismes. La comparaison d'espèces plus ou moins proches, pourrait permettre de voir l'évolution des répétitions sur des temps (de divergence) plus ou moins longs.

Une seconde voie est la mise en place d'une méthode similaire permettant de détecter les répétitions interchromosomiques. Il sera ainsi possible de compléter nos données sur les répétitions et d'étudier plus finement les caractéristiques des répétitions dispersées. Sans ces répétitions interchromosomiques, nous perdons une information clef sur les répétitions. Par ailleurs, ceci pourra permettre d'étudier plus finement les traces de réarrangements des répétitions (comme les insertions, délétions et inversions des répétitions en tandem).

Bien d'autres voies sont ouvertes, comme par exemple la simulation du modèle, l'étude des relations entre répétitions et réarrangements, celle des relations entre répétitions, recombinaison et teneur en GC. Certaines d'entre elles sont déjà bien avancées et mériteraient d'être mieux formalisées.

B.7. Les répétitions de l'origine.

La figure 41 illustre les tailles des différents génomes des organismes les plus « simples » aux organismes les plus « complexes ». Sans aucun doute possible, la taille des génomes des organismes simples, comme Bactéries et Archées, est plus petite que celle des organismes « complexes », comme les Eucaryotes dits « supérieurs ». Comme la création *de novo* de matériel génétique ne semble pas très active, les duplications constituent la voie préférentielle pour expliquer ces grandes différences de tailles.

Cette figure illustre également la variabilité de taille qui existe entre des espèces relativement proches. Par exemple les génomes des amphibiens et des angiospermes ont des tailles de génomes pouvant varier respectivement d'un facteur 100 ou 1000. Il est intéressant de noter que ces phylums sont ceux pour lesquels de nombreuses polyploïdies ont été décrites (voir *Introduction*). On peut donc inférer que les changements massifs des tailles de génomes s'opèrent par polyploïdie. Curieusement dans les phylums de mammifères les tailles des génomes semblent plus proches les unes des autres. Les variations de taille entre ces génomes s'opèrent vraisemblablement par des duplications plus petites. Dans tous ces génomes de grande taille les répétitions segmentaires (comme celles décrites chez *H. sapiens*) peuvent presque être considérées comme des « petites » répétitions.

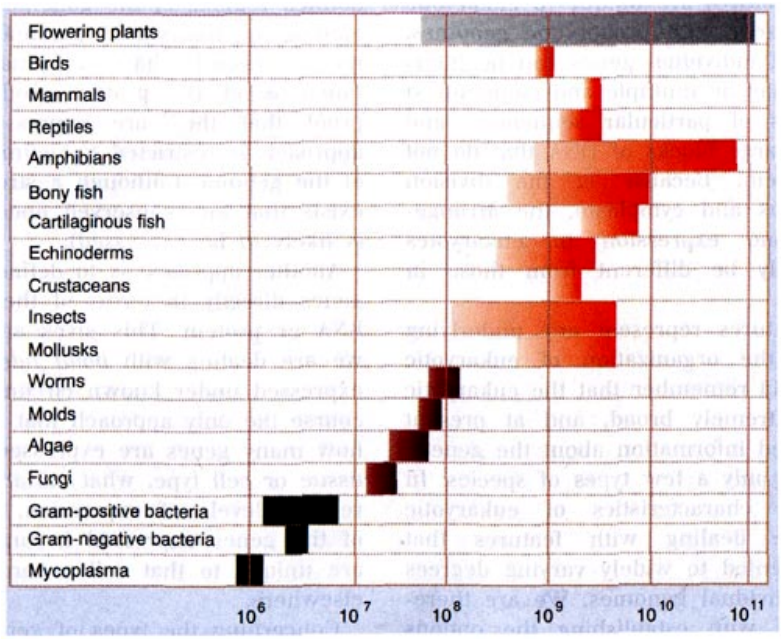

Figure 41 : Tailles des génomes des organismes « simples » et « complexes » (d'après (Lewin 1997)).

Chez les Bactéries et les Archées, les tailles des génomes varient au maximum d'un facteur 20. L'évolution de ces génomes se fait probablement en conservant un équilibre entre duplication et délétion. Cet équilibre tend à maintenir ces génomes à une taille petite et relativement constante. Il est avancé que les plus petits génomes, ceux des parasites de cellules eucaryotes, se sont engagés dans une voie de réduction de taille. Pour ces petits génomes, on peut penser que l'équilibre entre délétion et duplication est déplacé vers un excès de délétions.

L'étude des répétitions dans les génomes des Eucaryotes, des Archées et des Bactéries a montré que certaines caractéristiques de ces répétitions sont partagées par les trois règnes du vivant. Cela suggère que les répétitions sont, comme pressenti par S Ohno (Ohno 1970), une des clefs de l'évolution des génomes. Elles sont le reflet de contraintes structuralo-sélectives prédatant la divergence entre Eucaryotes, Bactéries et Archées (Table 1). Les répétitions et donc les duplications font donc partie de la dynamique des premiers génomes (les progénotes) et sont peut-être, comme l'a suggéré S Ohno, présentes depuis les premières étapes de l'élaboration de la vie en chimie prébiotique (Ohno 1987b).

Références bibliographiques

Agrawal, A., Q. M. Eastman and D. G. Schatz, 1998 Implication of transposition mediated by V(D)J-recombination proteins RAG1 and RAG2 for origins of antigen-specific immunity. Nature **394:** 744-751.

Ahn, B. Y., K. J. Dornfeld, T. J. Fagrelius and D. M. Livingston, 1988 Effect of limited homology on gene conversion in a Saccharomyces cerevisiae plasmid recombination system. Mol Cell Biol **8:** 2442-2448.

Alfenito, M. R., and J. A. Birchler, 1993 Molecular characterization of a maize B chromosome centric sequence. Genetics **135:** 589-597.

Allendorf, F. W., 1978 Protein polymorphism and the rate of loss of duplicate gene expression. Nature **272:** 76-78.

Amarger, V., D. Gauguier, M. Yerle, F. Apiou, P. Pinton *et al.*, 1998 Analysis of distribution in the human, pig, and rat genomes points toward a general subtelomeric origin of minisatellite structures. Genomics **52:** 62-71.

Amores, A., A. Force, Y. L. Yan, L. Joly, C. Amemiya *et al.*, 1998 Zebrafish hox clusters and vertebrate genome evolution. Science **282:** 1711-1714.

Anderson, P., and J. Roth, 1981 Spontaneous tandem genetic duplications in *Salmonella typhimurium* arise by unequal recombination between rRNA (rrn) cistrons. Proc Natl Acad Sci U S A **78:** 3113-3117.

Ayares, D., L. Chekuri, K. Y. Song and R. Kucherlapati, 1986 Sequence homology requirements for intermolecular recombination in mammalian cells. Proc Natl Acad Sci U S A **83:** 5199-5203.

Bach, M. L., F. Roelants, J. De Montigny, M. Huang, S. Potier *et al.*, 1995 Recovery of gene function by gene duplication in *Saccharomyces cerevisiae*. Yeast **11:** 169-177.

Bachtrog, D., S. Weiss, B. Zangerl, G. Brem and C. Schlotterer, 1999 Distribution of dinucleotide microsatellites in the *Drosophila melanogaster* genome. Mol Biol Evol **16:** 602-610.

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

Bailey, W. J., J. Kim, G. P. Wagner and F. H. Ruddle, 1997 Phylogenetic reconstruction of vertebrate Hox cluster duplications. Mol Biol Evol **14:** 843-853.

Baltimore, D., 1970 RNA-dependent DNA polymerase in virions of RNA tumour viruses. Nature **226:** 1209-1211.

Barnes, T. M., Y. Kohara, A. Coulson and S. Hekimi, 1995 Meiotic recombination, noncoding DNA and genomic organization in *Caenorhabditis elegans*. Genetics **141:** 159-179.

Baudat, F., and A. Nicolas, 1997 Clustering of meiotic double-strand breaks on yeast chromosome III. Proc Natl Acad Sci U S A **94:** 5213-5218.

Bernardi, G., 2000 Isochores and the evolutionary genomics of vertebrates. Gene **241:** 3-17.

Bi, X., and L. F. Liu, 1994 recA-independent and recA-dependent intramolecular plasmid recombination. Differential homology requirement and distance effect. J Mol Biol **235:** 414-423.

Bi, X., and L. F. Liu, 1996a DNA rearrangement mediated by inverted repeats. Proc Natl Acad Sci U S A **93:** 819-823.

Bi, X., and L. F. Liu, 1996b recA-independent DNA recombination between repetitive sequences: mechanisms and implications. Prog Nucleic Acid Res Mol Biol **54:** 253-292.

Bidenne, C., B. Blondin, S. Dequin and F. Vezinhet, 1992 Analysis of the chromosomal DNA polymorphism of wine strains of *Saccharomyces cerevisiae*. Curr Genet **22:** 1-7.

Bingham, P. M., M. G. Kidwell and G. M. Rubin, 1982 The molecular basis of P-M hybrid dysgenesis: the role of the P element, a P-strain-specific transposon family. Cell **29:** 995-1004.

Bingham, P. M., M. O. Scott, S. Wang, M. J. McPhaul, E. M. Wilson *et al.*, 1995 Stability of an expanded trinucleotide repeat in the androgen receptor gene in transgenic mice. Nat Genet **9:** 191-196.

Bisbee, C. A., M. A. Baker, A. C. Wilson, I. Haji-Azimi and M. Fischberg, 1977 Albumin phylogeny for clawed frogs (Xenopus). Science **195:** 785-787.

Biswas, I., V. Vagner and S. D. Ehrlich, 1992 Efficiency of homologous intermolecular recombination at different locations on the *Bacillus subtilis* chromosome. J Bacteriol **174:** 5593-5596.

Bjorkman, P. J., M. A. Saper, B. Samraoui, W. S. Bennett, J. L. Strominger *et al.*, 1987 The foreign antigen binding site and T cell recognition regions of class I histocompatibility antigens. Nature **329:** 512-518.

Blanc, G., A. Barakat, R. Guyot, R. Cooke and M. Delseny, 2000 Extensive duplication and reshuffling in the *Arabidopsis* genome. Plant Cell **12:** 1093-1101.

Blattner, F. R., G. Plunkett, 3rd, C. A. Bloch, N. T. Perna, V. Burland *et al.*, 1997 The complete genome sequence of *Escherichia coli* K-12. Science **277:** 1453-1474.

Bois, P., J. D. Stead, S. Bakshi, J. Williamson, R. Neumann *et al.*, 1998 Isolation and characterization of mouse minisatellites. Genomics **50:** 317-330.

Bosco, G., and J. E. Haber, 1998 Chromosome break-induced DNA replication leads to nonreciprocal translocations and telomere capture. Genetics **150:** 1037-1047.

Britten, R. J., 1998 Precise sequence complementarity between yeast chromosome ends and two classes of just-subtelomeric sequences. Proc Natl Acad Sci U S A **95:** 5906-5912.

Buard, J., A. Bourdet, J. Yardley, Y. Dubrova and A. J. Jeffreys, 1998 Influences of array size and homogeneity on minisatellite mutation. Embo J **17:** 3495-3502.

Casaregola, S., H. V. Nguyen, A. Lepingle, P. Brignon, F. Gendre *et al.*, 1998 A family of laboratory strains of *Saccharomyces cerevisiae* carry rearrangements involving chromosomes I and III. Yeast **14:** 551-564.

Charlesworth, B., P. Sniegowski and W. Stephan, 1994 The evolutionary dynamics of repetitive DNA in eukaryotes. Nature **371:** 215-220.

Chedin, F., E. Dervyn, R. Dervyn, S. D. Ehrlich and P. Noirot, 1994 Frequency of deletion formation decreases exponentially with distance between short direct repeats. Mol Microbiol **12:** 561-569.

Chien, Y., D. M. Becker, T. Lindsten, M. Okamura, D. I. Cohen *et al.*, 1984 A third type of murine T-cell receptor gene. Nature **312:** 31-35.

Chiu, C. H., H. Schneider, J. L. Slightom, D. L. Gumucio and M. Goodman, 1997 Dynamics of regulatory evolution in primate beta-globin gene clusters: cis-mediated acquisition of simian gamma fetal expression patterns. Gene **205:** 47-57.

Clare, J., and P. Farabaugh, 1985 Nucleotide sequence of a yeast Ty element: evidence for an unusual mechanism of gene expression. Proc Natl Acad Sci U S A **82:** 2829-2833.

Clarke, L., 1990 Centromeres of budding and fission yeasts. Trends Genet **6:** 150-154.

Coissac, E., 1996 Analyse structurale et fonctionnelle du génome de la levure *Sacchaomyces cerevisiae*., pp. 318. Université Pierre et Marie Curie, Paris.

Coissac, E., E. Maillier and P. Netter, 1997 A comparative study of duplications in bacteria and eukaryotes: the importance of telomeres. Mol Biol Evol **14:** 1062-1074.

Coissac, E., E. Maillier, S. Robineau and P. Netter, 1996 Sequence of a 39,411 bp DNA fragment covering the left end of chromosome VII of *Saccharomyces cerevisiae*. Yeast **12:** 1555-1562.

Coleman, J., D. M. Baird and N. J. Royle, 1999 The plasticity of human telomeres demonstrated by a hypervariable telomere repeat array that is located on some copies of 16p and 16q. Hum Mol Genet **8:** 1637-1646.

Colot, V., and J. L. Rossignol, 1999 Eukaryotic DNA methylation as an evolutionary device. Bioessays **21:** 402-411.

Consortium, The Dutch-Belgian Fragile X Consortium, 1994 Fmr1 knockout mice: a model to study fragile X mental retardation. Cell **78:** 23-33.

Copenhaver, G. P., K. Nickel, T. Kuromori, M. I. Benito, S. Kaul *et al.*, 1999 Genetic definition and sequence analysis of *Arabidopsis* centromeres. Science **286:** 2468-2474.

Critchlow, S. E., and S. P. Jackson, 1998 DNA end-joining: from yeast to man. Trends Biochem Sci **23:** 394-398.

Dawkins, R., 1976 *Le gène égoiste (traduction 1990)*, Paris.

Dayhoff, M. O., R. V. Eck and C. M. Park, 1972 A model of evolutionary change in peotein sequences, pp. 89-99 in *Atlas of Protein Sequence ande Structure*, edited by N. B. R. Found., Washington.

Dean, M., A. Rzhetsky and R. Allikmets, 2001 The human ATP-binding cassette (ABC) transporter superfamily. Genome Res **11:** 1156-1166.

Debrauwere, H., J. Buard, J. Tessier, D. Aubert, G. Vergnaud *et al.*, 1999 Meiotic instability of human minisatellite CEB1 in yeast requires DNA double-strand breaks. Nat Genet **23:** 367-371.

Delhomme, B., and P. Djian, 2000 Expansion of mouse involucrin by intra-allelic repeat addition. Gene **252:** 195-207.

Dianov, G. L., A. V. Kuzminov, A. V. Mazin and R. I. Salganik, 1991 Molecular mechanisms of deletion formation in *Escherichia coli* plasmids. I. Deletion formation mediated by long direct repeats. Mol Gen Genet **228:** 153-159.

Dietrich, F. S., S. Voegeli, T. Gaffney, C. Mohr, C. Rebischung *et al.*, 1999 Gene map of chromosome I of *Ashbya gossypii*, pp. 233 in *XIX International Conference on Yeast Genetics and Molecular Biology*, edited by F. Kaudewitz. springer-verlag, Rimini, Italy.

Dimpfl, J., and H. Echols, 1989 Duplication mutation as an SOS response in *Escherichia coli*: enhanced duplication formation by a constitutively activated RecA. Genetics **123:** 255-260.

Djian, P., K. Easley and H. Green, 2000 Targeted ablation of the murine involucrin gene. J Cell Biol **151:** 381-388.

Djian, P., and H. Green, 1989a The involucrin gene of the orangutan: generation of the late region as an evolutionary trend in the hominoids. Mol Biol Evol **6:** 469-477.

Djian, P., and H. Green, 1989b Vectorial expansion of the involucrin gene and the relatedness of the hominoids. Proc Natl Acad Sci U S A **86:** 8447-8451.

Djian, P., and H. Green, 1990 The involucrin gene of the gibbon: the middle region shared by the hominoids. Mol Biol Evol **7:** 220-227.

Djian, P., and H. Green, 1991 Involucrin gene of tarsioids and other primates: alternatives in evolution of the segment of repeats. Proc Natl Acad Sci U S A **88:** 5321-5325.

Djian, P., and H. Green, 1992 The involucrin gene of Old-World monkeys and other higher primates: synapomorphies and parallelisms resulting from the same gene-altering mechanism. Mol Biol Evol **9:** 417-432.

Djian, P., J. M. Hancock and H. S. Chana, 1996 Codon repeats in genes associated with human diseases: fewer repeats in the genes of nonhuman primates and nucleotide substitutions concentrated at the sites of reiteration. Proc Natl Acad Sci U S A **93:** 417-421.

Doolittle, R. F., D. F. Feng, M. S. Johnson and M. A. McClure, 1989 Origins and evolutionary relationships of retroviruses. Q Rev Biol **64:** 1-30.

Doolittle, R. F., D. F. Feng, M. A. McClure and M. S. Johnson, 1990 Retrovirus phylogeny and evolution. Curr Top Microbiol Immunol **157:** 1-18.

Drake, J. W., B. Charlesworth, D. Charlesworth and J. F. Crow, 1998 Rates of spontaneous mutation. Genetics **148:** 1667-1686.

Dubnau, D., 1993 Genetic exchange and homologous recombination. in *Bacillus subtilis and other Gram-positive bacteria*, edited by A. L. Sonenshien, J. A. Hoch and R. Losick. Ash Press, Washington.

Dubrova, Y. E., A. J. Jeffreys and A. M. Malashenko, 1993 Mouse minisatellite mutations induced by ionizing radiation. Nat Genet **5:** 92-94.

Dujon, B., 1996 The yeast genome project: what did we learn? Trends Genet **12:** 263-270.

Dunham, I., N. Shimizu, B. A. Roe, S. Chissoe, A. R. Hunt *et al.*, 1999 The DNA sequence of human chromosome 22. Nature **402:** 489-495.

Dutrillaux, B., 1997 Comment évoluent les chromosomes des mammifères. La Recherche **296:** 70-75.

Dutrillaux, B., and F. Richard, 1997 Notre nouvel arbre de famille. La Recherche **298:** 54-61.

Earnshaw, W. C., K. F. Sullivan, P. S. Machlin, C. A. Cooke, D. A. Kaiser *et al.*, 1987 Molecular cloning of cDNA for CENP-B, the major human centromere autoantigen. J Cell Biol **104:** 817-829.

Eckert, R. L., and H. Green, 1986 Structure and evolution of the human involucrin gene. Cell **46:** 583-589.

Eichler, E. E., N. Archidiacono and M. Rocchi, 1999 CAGGG repeats and the pericentromeric duplication of the hominoid genome. Genome Res **9:** 1048-1058.

Eichler, E. E., F. Lu, Y. Shen, R. Antonacci, V. Jurecic *et al.*, 1996 Duplication of a gene-rich cluster between 16p11.1 and Xq28: a novel pericentromeric-directed mechanism for paralogous genome evolution. Hum Mol Genet **5:** 899-912.

Eickbush, T. H., 1994 Origin and relationships of retroelements, pp. 121-157 in *The evolutionary biology of viruses*, edited by S. S. Morse. Raven Press, New-York.

Eickbush, T. H., 1997 Telomerase and retrotransposons: which came first? Science **277:** 911-912.

Elliott, J. F., E. P. Rock, P. A. Patten, M. M. Davis and Y. H. Chien, 1988 The adult T-cell receptor delta-chain is diverse and distinct from that of fetal thymocytes. Nature **331:** 627-631.

Emmons, S. W., L. Yesner, K. S. Ruan and D. Katzenberg, 1983 Evidence for a transposon in *Caenorhabditis elegans*. Cell **32:** 55-65.

Engels, W. R., 1996 P elements in Drosophila. Curr Top Microbiol Immunol **204:** 103-123.

Enomoto, S., M. S. Longtine and J. Berman, 1994 Enhancement of telomere-plasmid segregation by the X-telomere associated sequence in *Saccharomyces cerevisiae* involves SIR2, SIR3, SIR4 and ABF1. Genetics **136:** 757-767.

Eyre, D., H. C. Gorham and E. J. Louis, 1999 Telomere recombination: sequestring and non-random partners, pp. 350 in *XIX International Conference on Yeast Genetics and Molecular Biology*, edited by F. Kaudewitz. springer-verlag, Rimini, Italy.

Fawcett, D. H., C. K. Lister, E. Kellett and D. J. Finnegan, 1986 Transposable elements controlling I-R hybrid dysgenesis in *D. melanogaster* are similar to mammalian LINEs. Cell **47:** 1007-1015.

Ferris, S. D., and G. S. Whitt, 1977 Loss of duplicate gene expression after polyploidisation. Nature **265:** 258-260.

Feschotte, C., and C. Mouches, 2000 Evidence that a family of miniature inverted-repeat transposable elements (MITEs) from the *Arabidopsis thaliana* genome has arisen from a pogo-like DNA transposon. Mol Biol Evol **17:** 730-737.

Feuermann, M., M. J. de, S. Potier and J. L. Souciet, 1997 The characterization of two new clusters of duplicated genes suggests a 'Lego' organization of the yeast *Saccharomyces cerevisiae* chromosomes. Yeast **13:** 861-869.

Fischer, G., S. A. James, I. N. Roberts, S. G. Oliver and E. J. Louis, 2000 Chromosomal evolution in *Saccharomyces*. Nature **405:** 451-454.

Fischer, G., C. Neuveglise, P. Durrens, C. Gaillardin and B. Dujon, 2001 Evolution of gene order in the genomes of two related yeast species. Genome Res **11:** 2009-2019.

Flores, M., S. Brom, T. Stepkowski, M. L. Girard, G. Davila *et al.*, 1993 Gene amplification in Rhizobium: identification and in vivo cloning of discrete amplifiable DNA regions (amplicons) from Rhizobium leguminosarum biovar phaseoli. Proc Natl Acad Sci U S A **90:** 4932-4936.

Fogel, S., and J. W. Welch, 1982 Tandem gene amplification mediates copper resistance in yeast. Proc Natl Acad Sci U S A **79:** 5342-5346.

Force, A., M. Lynch, F. B. Pickett, A. Amores, Y. L. Yan *et al.*, 1999 Preservation of duplicate genes by complementary, degenerative mutations. Genetics **151:** 1531-1545.

Francis, J. C., and P. E. Hansche, 1972 Directed evolution of metabolic pathways in microbial populations. I. Modification of the acid phosphatase pH optimum in *S. cerevisiae*. Genetics **70:** 59-73.

Francis, J. C., and P. E. Hansche, 1973 Directed evolution of metabolic pathways in microbial populations. II.A repeatable adaptation in *Saccharomyces cerevisiae*. Genetics **74:** 259-265.

Friedman, R., and A. L. Hughes, 2001a Gene duplication and the structure of eukaryotic genomes. Genome Res **11:** 373-381.

Friedman, R., and A. L. Hughes, 2001b Pattern and timing of gene duplication in animal genomes. Genome Res **11:** 1842-1847.

Fullerton, S. M., A. Bernardo Carvalho and A. G. Clark, 2001 Local rates of recombination are positively correlated with GC content in the human genome. Mol Biol Evol **18:** 1139-1142.

Gaut, B. S., M. Le Thierry d'Ennequin, A. S. Peek and M. C. Sawkins, 2000 Maize as a model for the evolution of plant nuclear genomes. Proc Natl Acad Sci U S A **97:** 7008-7015.

Gellon, G., and W. McGinnis, 1998 Shaping animal body plans in development and evolution by modulation of Hox expression patterns. Bioessays **20:** 116-125.

Gilbert, N., and J. Allan, 2001 Distinctive higher-order chromatin structure at mammalian centromeres. Proc Natl Acad Sci U S A **98:** 11949-11954.

Gindullis, F., C. Desel, I. Galasso and T. Schmidt, 2001 The large-scale organization of the centromeric region in Beta species. Genome Res **11:** 253-265.

Goffeau, A., and *e. a.* authors), 1997 The yeast genome directory. Nature **387:** 5.

Goldberg, D. E., 1995 The enigmatic oxygen-avid hemoglobin of *Ascaris*. Bioessays **17:** 177-182.

Goldberg, I., and J. J. Mekalanos, 1986 Effect of a recA mutation on cholera toxin gene amplification and deletion events. J Bacteriol **165:** 723-731.

Goncalves, I., L. Duret and D. Mouchiroud, 2000 Nature and structure of human genes that generate retropseudogenes. Genome Res **10:** 672-678.

Goodman, M., G. W. Moore and G. Matsuda, 1975 Darwinian evolution in the genealogy of haemoglobin. Nature **253:** 603-608.

Goodman, M., J. Pedwaydon, J. Czelusniak, T. Suzuki, T. Gotoh *et al.*, 1988 An evolutionary tree for invertebrate globin sequences. J Mol Evol **27:** 236-249.

Gorbunova, V., and A. A. Levy, 1997 Non-homologous DNA end joining in plant cells is associated with deletions and filler DNA insertions. Nucleic Acids Res **25:** 4650-4657.

Gotta, M., T. Laroche, A. Formenton, L. Maillet, H. Scherthan *et al.*, 1996 The clustering of telomeres and colocalization with Rap1, Sir3, and Sir4 proteins in wild-type *Saccharomyces cerevisiae*. J Cell Biol **134:** 1349-1363.

Goyon, C., C. Barry, A. Gregoire, G. Faugeron and J. L. Rossignol, 1996a Methylation of DNA repeats of decreasing sizes in *Ascobolus immersus*. Mol Cell Biol **16:** 3054-3065.

Goyon, C., J. L. Rossignol and G. Faugeron, 1996b Native DNA repeats and methylation in *Ascobolus*. Nucleic Acids Res **24:** 3348-3356.

Green, H., and P. Djian, 1992 Consecutive actions of different gene-altering mechanisms in the evolution of involucrin. Mol Biol Evol **9:** 977-1017.

Greig, G. M., P. E. Warburton and H. F. Willard, 1993 Organization and evolution of an alpha satellite DNA subset shared by human chromosomes 13 and 21. J Mol Evol **37:** 464-475.

Griffith, J., A. Bianchi and L. T. de, 1998 TRF1 promotes parallel pairing of telomeric tracts in vitro. J Mol Biol **278:** 79-88.

Gumucio, D. L., D. A. Shelton, K. Blanchard-McQuate, T. Gray, S. Tarle *et al.*, 1994 Differential phylogenetic footprinting as a means to identify base changes responsible for recruitment of the anthropoid gamma gene to a fetal expression pattern. J Biol Chem **269:** 15371-15380.

Haaf, T., P. E. Warburton and H. F. Willard, 1992 Integration of human alpha-satellite DNA into simian chromosomes: centromere protein binding and disruption of normal chromosome segregation. Cell **70:** 681-696.

Haber, J. E., and W. Y. Leung, 1996 Lack of chromosome territoriality in yeast: promiscuous rejoining of broken chromosome ends. Proc Natl Acad Sci U S A **93:** 13949-13954.

Hagemann, S., and W. Pinsker, 2001 *Drosophila* P transposons in the human genome? Mol Biol Evol **18:** 1979-1982.

Hall, B. G., L. L. Parker, P. W. Betts, R. F. DuBose, S. A. Sawyer *et al.*, 1989 IS103, a new insertion element in *Escherichia coli*: characterization and distribution in natural populations. Genetics **121:** 423-431.

Hamada, K., Y. Nakatomi and S. Shimada, 1992 Direct induction of tetraploids or homozygous diploids in the industrial yeast *Saccharomyces cerevisiae* by hydrostatic pressure. Curr Genet **22:** 371-376.

Hansche, P. E., 1975 Gene duplication as a mechanism of genetic adaptation in *Saccharomyces cerevisiae*. Genetics **79:** 661-674.

Hansche, P. E., V. Beres and P. Lange, 1978 Gene duplication in *Saccharomyces cerevisiae*. Genetics **88:** 673-687.

Hansen, L. J., D. L. Chalker and S. B. Sandmeyer, 1988 Ty3, a yeast retrotransposon associated with tRNA genes, has homology to animal retroviruses. Mol Cell Biol **8:** 5245-5256.

Hardison, R., 1998 Hemoglobins from bacteria to man: evolution of different patterns of gene expression. J Exp Biol **201 (Pt 8):** 1099-1117.

Hardison, R., J. L. Slightom, D. L. Gumucio, M. Goodman, N. Stojanovic *et al.*, 1997 Locus control regions of mammalian beta-globin gene clusters: combining phylogenetic analyses and experimental results to gain functional insights. Gene **205:** 73-94.

Harris, S., K. S. Rudnicki and J. E. Haber, 1993 Gene conversions and crossing over during homologous and homeologous ectopic recombination in *Saccharomyces cerevisiae*. Genetics **135:** 5-16.

Hartley, S. E., and M. T. Horne, 1984 Chromosome relationships in the genus *Salmo*. Chromosoma **90:** 229-237.

Hedrick, S. M., D. I. Cohen, E. A. Nielsen and M. M. Davis, 1984 Isolation of cDNA clones encoding T cell-specific membrane-associated proteins. Nature **308:** 149-153.

Henikoff, S., K. Ahmad and H. S. Malik, 2001 The centromere paradox: stable inheritance with rapidly evolving DNA. Science **293:** 1098-1102.

Hood, L., M. Kronenberg and T. Hunkapiller, 1985 T cell antigen receptors and the immunoglobulin supergene family. Cell **40:** 225-229.

Horie, K., A. Kuroiwa, M. Ikawa, M. Okabe, G. Kondoh *et al.*, 2001 Efficient chromosomal transposition of a Tc1/mariner- like transposon Sleeping Beauty in mice. Proc Natl Acad Sci U S A **98:** 9191-9196.

Howman, E. V., K. J. Fowler, A. J. Newson, S. Redward, A. C. MacDonald *et al.*, 2000 Early disruption of centromeric chromatin organization in centromere protein A (Cenpa) null mice. Proc Natl Acad Sci U S A **97:** 1148-1153.

Hughes, A. L., 1994 The evolution of functionally novel proteins after gene duplication. Proc R Soc Lond B Biol Sci **256:** 119-124.

Hughes, A. L., 1998 Phylogenetic tests of the hypothesis of block duplication of homologous genes on human chromosomes 6, 9, and 1. Mol Biol Evol **15:** 854-870.

Hughes, A. L., J. da Silva and R. Friedman, 2001 Ancient genome duplications did not structure the human Hox-bearing chromosomes. Genome Res **11:** 771-780.

Hughes, A. L., and M. Nei, 1988 Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature **335:** 167-170.

Hughes, A. L., and M. Nei, 1989 Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection. Proc Natl Acad Sci U S A **86:** 958-962.

Hughes, A. L., and M. Yeager, 1997 Molecular evolution of the vertebrate immune system. Bioessays **19:** 777-786.

Hughes, D., 1999 Impact of homologous recombination on genome organization and stability, pp. 109-128 in *Organization of the prokaryotic genome*, edited by R. L. Charlebois. American society for Microbiology, Washington, D.C.

Hughes, T. R., C. J. Roberts, H. Dai, A. R. Jones, M. R. Meyer *et al.*, 2000 Widespread aneuploidy revealed by DNA microarray expression profiling. Nat Genet **25:** 333-337.

Ivics, Z., P. B. Hackett, R. H. Plasterk and Z. Izsvak, 1997 Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells. Cell **91:** 501-510.

Jackson, E. N., and C. Yanofsky, 1973 Duplication-translocations of tryptophan operon genes in *Escherichia coli*. J Bacteriol **116:** 33-40.

Jackson, M. S., M. Rocchi, G. Thompson, T. Hearn, M. Crosier *et al.*, 1999 Sequences flanking the centromere of human chromosome 10 are a complex patchwork of arm-specific sequences, stable duplications and unstable sequences with homologies to telomeric and other centromeric locations. Hum Mol Genet **8:** 205-215.

Jacobson, J. W., M. M. Medhora and D. L. Hartl, 1986 Molecular structure of a somatically unstable transposable element in Drosophila. Proc Natl Acad Sci U S A **83:** 8684-8688.

Jakubczak, J. L., Y. Xiong and T. H. Eickbush, 1990 Type I (R1) and type II (R2) ribosomal DNA insertions of *Drosophila melanogaster* are retrotransposable elements closely related to those of *Bombyx mori*. J Mol Biol **212:** 37-52.

Jeffreys, A. J., A. MacLeod, K. Tamaki, D. L. Neil and D. G. Monckton, 1991 Minisatellite repeat coding as a digital approach to DNA typing. Nature **354:** 204-209.

Jeffreys, A. J., D. L. Neil and R. Neumann, 1998 Repeat instability at human minisatellites arising from meiotic recombination. Embo J **17:** 4147-4157.

Jeffreys, A. J., and R. Neumann, 1997 Somatic mutation processes at a human minisatellite. Hum Mol Genet **6:** 129-132; 134-126.

Jeffreys, A. J., R. Neumann and V. Wilson, 1990 Repeat unit sequence variation in minisatellites: a novel source of DNA polymorphism for studying variation and mutation by single molecule analysis. Cell **60:** 473-485.

Jeffreys, A. J., N. J. Royle, V. Wilson and Z. Wong, 1988 Spontaneous mutation rates to new length alleles at tandem-repetitive hypervariable loci in human DNA. Nature **332:** 278-281.

Jeffreys, A. J., K. Tamaki, A. MacLeod, D. G. Monckton, D. L. Neil *et al.*, 1994 Complex gene conversion events in germline mutation at human minisatellites. Nat Genet **6:** 136-145.

Jeffreys, A. J., V. Wilson and S. L. Thein, 1985 Hypervariable 'minisatellite' regions in human DNA. Biotechnology **24:** 467-472.

Jeyaprakash, A., J. W. Welch and S. Fogel, 1991 Multicopy CUP1 plasmids enhance cadmium and copper resistance levels in yeast. Mol Gen Genet **225:** 363-368.

Ji, Y., E. E. Eichler, S. Schwartz and R. D. Nicholls, 2000 Structure of chromosomal duplicons and their role in mediating human genomic disorders. Genome Res **10:** 597-610.

Jinks-Robertson, S., M. Michelitch and S. Ramcharan, 1993 Substrate length requirements for efficient mitotic recombination in *Saccharomyces cerevisiae*. Mol Cell Biol **13:** 3937-3950.

Jinks-Robertson, S., and T. D. Petes, 1986 Chromosomal translocations generated by high-frequency meiotic recombination between repeated yeast genes. Genetics **114:** 731-752.

Johnston, J. R., C. Baccari and R. K. Mortimer, 2000 Genotypic characterization of strains of commercial wine yeasts by tetrad analysis. Res Microbiol **151:** 583-590.

Karlin, S., A. M. Campbell and J. Mrazek, 1998 Comparative DNA analysis across diverse genomes. Annu Rev Genet **32:** 185-225.

Karlin, S., and F. Ost, 1985 Maximal segmental match length among random sequences from a finite alphabet, pp. 225-243 in *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer*, edited by L. M. L. Cam and R. A. Olshen. Association for Computing Machinery, New-York.

Karp, R. M., R. E. Miller and A. L. Rosenberg, 1972 Rapid Identification of repeated patterns in strings, trees and array., pp. 125-136 in *Proceedings 4th annual ACM symposium theory of computing*. ACM.

Kasahara, M., M. Hayashi, K. Tanaka, H. Inoko, K. Sugaya *et al.*, 1996 Chromosomal localization of the proteasome Z subunit gene reveals an ancient chromosomal duplication involving the major histocompatibility complex. Proc Natl Acad Sci U S A **93:** 9096-9101.

Kasahara, M., J. Nakaya, Y. Satta and N. Takahata, 1997 Chromosomal duplication and the emergence of the adaptive immune system. Trends Genet **13:** 90-92.

Kato, M., A. Kato and N. Shimizu, 1999 A method for evaluating phylogenetic relationship of alpha-satellite DNA suprachromosomal family by nucleotide frequency calculation. Mol Phylogenet Evol **13:** 329-335.

Katti, M. V., P. K. Ranjekar and V. S. Gupta, 2001 Differential distribution of simple sequence repeats in eukaryotic genome sequences. Mol Biol Evol **18:** 1161-1167.

Kaufman, R. J., P. C. Brown and R. T. Schimke, 1979 Amplified dihydrofolate reductase genes in unstably methotrexate-resistant cells are associated with double minute chromosomes. Proc Natl Acad Sci U S A **76:** 5669-5673.

Kaufman, R. J., P. C. Brown and R. T. Schimke, 1981 Loss and stabilization of amplified dihydrofolate reductase genes in mouse sarcoma S-180 cell lines. Mol Cell Biol **1:** 1084-1093.

Kaufman, R. J., and R. T. Schimke, 1981 Amplification and loss of dihydrofolate reductase genes in a Chinese hamster ovary cell line. Mol Cell Biol **1:** 1069-1076.

Keogh, R. S., C. Seoighe and K. H. Wolfe, 1998 Evolution of gene order and chromosome number in *Saccharomyces*, *Kluyveromyces* and related fungi. Yeast **14:** 443-457.

Kimura, M., 1983 *La théorie neutraliste de l'évolution (traduction 1990)*, Paris.

Kipling, D., and P. E. Warburton, 1997 Centromeres, CENP-B and Tigger too. Trends Genet **13:** 141-145.

Klein, H. L., 1995 Genetic control of intrachromosomal recombination. Bioessays **17:** 147-159.

Koch, A. L., 1979 Selection and recombination in populations containing tandem multiplet genes. J Mol Evol **14:** 273-285.

Kraus, E., W. Y. Leung and J. E. Haber, 2001 Break-induced replication: a review and an example in budding yeast. Proc Natl Acad Sci U S A **98:** 8255-8262.

Ku, H. M., T. Vision, J. Liu and S. D. Tanksley, 2000 Comparing sequenced segments of the tomato and *Arabidopsis* genomes: large-scale duplication followed by selective gene loss creates a network of synteny. Proc Natl Acad Sci U S A **97:** 9121-9126.

Kurtz, S., and C. Schleiermacher, 1999 REPuter: fast computation of maximal repeats in complete genomes. Bioinformatics **15:** 426-427.

Lalioti, M. D., H. S. Scott and S. E. Antonarakis, 1999 Altered spacing of promoter elements due to the dodecamer repeat expansion contributes to reduced expression of the cystatin B gene in EPM1. Hum Mol Genet **8:** 1791-1798.

Lalo, D., S. Stettler, S. Mariotte, P. P. Slonimski and P. Thuriaux, 1993 Two yeast chromosomes are related by a fossil duplication of their centromeric regions. C R Acad Sci Iii **316:** 367-373.

Lampe, D. J., M. E. Churchill and H. M. Robertson, 1996 A purified mariner transposase is sufficient to mediate transposition in vitro. Embo J **15:** 5470-5479.

Lampson, B. C., S. Inouye and M. Inouye, 1991 msDNA of bacteria. Prog Nucleic Acid Res Mol Biol **40:** 1-24.

Laudet, V., C. Hanni, J. Coll, F. Catzeflis and D. Stehelin, 1992 Evolution of the nuclear receptor gene superfamily. Embo J **11:** 1003-1013.

Le Fleche, P., Y. Hauck, L. Onteniente, A. Prieur, F. Denoeud *et al.*, 2001 A tandem repeats database for bacterial genomes: application to the genotyping of *Yersinia pestis* and *Bacillus anthracis*. Bmc Microbiol **1:** 2.

Lebre, A. S., and A. Brice, 2001 Maladies par expansion de polyglutamine. Données moléculaires et physiopathologiques. Médecine/Sciences **17:** 1-9.

Lee, C., R. Critcher, J. G. Zhang, W. Mills and C. J. Farr, 2000 Distribution of gamma satellite DNA on the human X and Y chromosomes suggests that it is not required for mitotic centromere function. Chromosoma **109:** 381-389.

Leung, M. Y., B. E. Blaisdell, C. Burge and S. Karlin, 1991 An efficient algorithm for identifying matches with errors in multiple long molecular sequences. J Mol Biol **221:** 1367-1378.

Levinson, G., and G. A. Gutman, 1987 High frequencies of short frameshifts in poly-CA/TG tandem repeats borne by bacteriophage M13 in *Escherichia coli* K-12. Nucleic Acids Res **15:** 5323-5338.

Levy, D. D., and T. A. Cebula, 2001 Fidelity of replication of repetitive DNA in mutS and repair proficient *Escherichia coli*. Mutat Res **474:** 1-14.

Lewin, B., 1997 Genomes, pp. 645-662 in *Genes VI*. Oxford University Press, New-York.

Li, W. H., Z. Gu, H. Wang and A. Nekrutenko, 2001 Evolutionary analyses of the human genome. Nature **409:** 847-849.

Lichten, M., R. H. Borts and J. E. Haber, 1987 Meiotic gene conversion and crossing over between dispersed homologous sequences occurs frequently in *Saccharomyces cerevisiae*. Genetics **115:** 233-246.

Lichten, M., and J. E. Haber, 1989 Position effects in ectopic and allelic mitotic recombination in *Saccharomyces cerevisiae*. Genetics **123:** 261-268.

Lin, R. J., M. Capage and C. W. Hill, 1984 A repetitive DNA sequence, rhs, responsible for duplications within the *Escherichia coli* K-12 chromosome. J Mol Biol **177:** 1-18.

Liskay, R. M., A. Letsou and J. L. Stachelek, 1987 Homology requirement for efficient gene conversion between duplicated chromosomal sequences in mammalian cells. Genetics **115:** 161-167.

Llorente, B., C. Fairhead and B. Dujon, 1999 Genetic redundancy and gene fusion in the genome of the Baker's yeast *Saccharomyces cerevisiae*: functional characterization of a three-member gene family involved in the thiamine biosynthetic pathway. Mol Microbiol **32:** 1140-1152.

Llorente, B., A. Malpertuy, C. Neuveglise, J. de Montigny, M. Aigle *et al.*, 2000 Genomic Exploration of the Hemiascomycetous Yeasts: 18. Comparative analysis of chromosome maps and synteny with *Saccharomyces cerevisiae*. FEBS Lett **487:** 101-112.

Lo, A. W., J. M. Craig, R. Saffery, P. Kalitsis, D. V. Irvine *et al.*, 2001a A 330 kb CENP-A binding domain and altered replication timing at a human neocentromere. Embo J **20:** 2087-2096.

Lo, A. W., D. J. Magliano, M. C. Sibson, P. Kalitsis, J. M. Craig *et al.*, 2001b A novel chromatin immunoprecipitation and array (CIA) analysis identifies a 460-kb CENP-A-binding neocentromere DNA. Genome Res **11:** 448-457.

Lobry, J. R., 1996 Asymmetric substitution patterns in the two DNA strands of bacteria. Mol Biol Evol **13:** 660-665.

Loidl, J., 1995 Meiotic chromosome pairing in triploid and tetraploid *Saccharomyces cerevisiae*. Genetics **139:** 1511-1520.

Loidl, J., and K. Nairz, 1997 Karyotype variability in yeast caused by nonallelic recombination in haploid meiosis. Genetics **146:** 79-88.

Lopes, J., E. LeGuern, R. Gouider, S. Tardieu, N. Abbas *et al.*, 1996 Recombination hot spot in a 3.2-kb region of the Charcot-Marie-Tooth type 1A repeat sequences: new tools for molecular diagnosis of hereditary neuropathy with liability to pressure palsies and of Charcot- Marie-Tooth type 1A. French CMT Collaborative Research Group. Am J Hum Genet **58:** 1223-1230.

Lopes, J., S. Tardieu, K. Silander, I. Blair, A. Vandenberghe *et al.*, 1999 Homologous DNA exchanges in humans can be explained by the yeast double- strand break repair model: a study of 17p11.2 rearrangements associated with CMT1A and HNPP. Hum Mol Genet **8:** 2285-2292.

Lopes, J., A. Vandenberghe, S. Tardieu, V. Ionasescu, N. Levy *et al.*, 1997 Sex-dependent rearrangements resulting in CMT1A and HNPP. Nat Genet **17:** 136-137.

Lopreato, G. F., Y. Lu, A. Southwell, N. S. Atkinson, D. M. Hillis *et al.*, 2001 Evolution and divergence of sodium channel genes in vertebrates. Proc Natl Acad Sci U S A **98:** 7588-7592.

Lorenz, M. G., and W. Wackernagel, 1994 Bacterial gene transfer by natural genetic transformation in the environment. Microbiol Rev **58:** 563-602.

Louis, E. J., 1995 The chromosome ends of *Saccharomyces cerevisiae*. Yeast **11:** 1553-1573.

Louis, E. J., and J. E. Haber, 1990 Mitotic recombination among subtelomeric Y' repeats in *Saccharomyces cerevisiae*. Genetics **124:** 547-559.

Lovett, S. T., and V. V. Feschenko, 1996 Stabilization of diverged tandem repeats by mismatch repair: evidence for deletion formation via a misaligned replication intermediate. Proc Natl Acad Sci U S A **93:** 7120-7124.

Lovett, S. T., T. J. Gluckman, P. J. Simon, V. J. Sutera and P. T. Drapkin, 1994 Recombination between repeats in *Escherichia coli* by a recA- independent, proximity-sensitive mechanism. Mol Gen Genet **245:** 294-300.

Lupski, J. R., 1998 Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. Trends Genet **14:** 417-422.

Lupski, J. R., R. M. de Oca-Luna, S. Slaugenhaupt, L. Pentao, V. Guzzetta *et al.*, 1991 DNA duplication associated with Charcot-Marie-Tooth disease type 1A. Cell **66:** 219-232.

Lynch, M., and J. S. Conery, 2000 The evolutionary fate and consequences of duplicate genes. Science **290:** 1151-1155.

Lynch, M., and A. Force, 2000 The probability of duplicate gene preservation by subfunctionalization. Genetics **154:** 459-473.

Lyu, Y. L., C. T. Lin and L. F. Liu, 1999 Inversion/dimerization of plasmids mediated by inverted repeats. J Mol Biol **285:** 1485-1501.

Maestre, J., T. Tchenio, O. Dhellin and T. Heidmann, 1995 mRNA retroposition in human cells: processed pseudogene formation. Embo J **14:** 6333-6338.

Mahillon, J., and M. Chandler, 1998 Insertion sequences. Microbiol Mol Biol Rev **62:** 725-774.

Maloisel, L., and J. L. Rossignol, 1998 Suppression of crossing-over by DNA methylation in Ascobolus. Genes Dev **12:** 1381-1389.

Marcotte, E. M., M. Pellegrini, T. O. Yeates and D. Eisenberg, 1999 A census of protein repeats. J Mol Biol **293:** 151-160.

Marlor, R. L., S. M. Parkhurst and V. G. Corces, 1986 The *Drosophila melanogaster* gypsy transposable element encodes putative gene products homologous to retroviral proteins. Mol Cell Biol **6:** 1129-1134.

Mashkova, T. D., N. Y. Oparina, M. H. Lacroix, L. I. Fedorova, G. T. I *et al.*, 2001 Structural rearrangements and insertions of dispersed elements in pericentromeric alpha satellites occur preferably at kinkable DNA sites. J Mol Biol **305:** 33-48.

Masterson, J., 1994 Stomatal size in fossil plants : evidence for polyploidy in majority of angiosperms. science **264:** 421-424.

Masumoto, H., H. Masukata, Y. Muro, N. Nozaki and T. Okazaki, 1989 A human centromere antigen (CENP-B) interacts with a short specific sequence in alphoid DNA, a human centromeric satellite. J Cell Biol **109:** 1963-1973.

Mazzarella, R., and D. Schlessinger, 1997 Duplication and distribution of repetitive elements and non-unique regions in the human genome. Gene **205:** 29-38.

McClure, M. A., 1993 Evolutionary History of Reverse Transcriptase, pp. 425-441 in *Reverse Transcriptase*. Cold Sping Harbor Laboratory Press.

Mendiola, M. V., and F. de la Cruz, 1992 IS91 transposase is related to the rolling-circle-type replication proteins of the pUB110 family of plasmids. Nucleic Acids Res **20:** 3521.

Merlin, C., and A. Toussaint, 1999 Les éléments transposables. Médecine/Sciences **15:** I-XIII.

Mewes, H. W., K. Albermann, M. Bahr, D. Frishman, A. Gleissner *et al.*, 1997 Overview of the yeast genome. Nature **387:** 7-65.

Mitas, M., 1997 Trinucleotide repeats associated with human disease. Nucleic Acids Res **25:** 2245-2254.

Mizuuchi, K., 1992 Polynucleotidyl transfer reactions in transpositional DNA recombination. J Biol Chem **267:** 21273-21276.

Monckton, D. G., R. Neumann, T. Guram, N. Fretwell, K. Tamaki *et al.*, 1994 Minisatellite mutation rate variation associated with a flanking DNA sequence polymorphism. Nat Genet **8:** 162-170.

Moore, J. K., and J. E. Haber, 1996 Capture of retrotransposon DNA at the sites of chromosomal double- strand breaks. Nature **383:** 644-646.

Morange, M., 1998 *La part des gènes*, Paris.

Mortimer, R. K., 2000 Evolution and variation of the yeast (*Saccharomyces*) genome. Genome Res **10:** 403-409.

Mortimer, R. K., P. Romano, G. Suzzi and M. Polsinelli, 1994 Genome renewal: a new phenomenon revealed from a genetic study of 43 strains of *Saccharomyces cerevisiae* derived from natural fermentation of grape musts. Yeast **10:** 1543-1552.

Mount, S. M., and G. M. Rubin, 1985 Complete nucleotide sequence of the *Drosophila* transposable element copia: homology between copia and retroviral proteins. Mol Cell Biol **5:** 1630-1638.

Muro, Y., H. Masumoto, K. Yoda, N. Nozaki, M. Ohashi *et al.*, 1992 Centromere protein B assembles human centromeric alpha-satellite DNA at the 17-bp sequence, CENP-B box. J Cell Biol **116:** 585-596.

Murray, J., J. Buard, D. L. Neil, E. Yeramian, K. Tamaki *et al.*, 1999 Comparative sequence analysis of human minisatellites showing meiotic repeat instability. Genome Res **9:** 130-136.

Nachman, M. W., and G. A. Churchill, 1996 Heterogeneity in rates of recombination across the mouse genome. Genetics **142:** 537-548.

Nadeau, J. H., and D. Sankoff, 1997 Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution. Genetics **147:** 1259-1266.

Nargang, F. E., J. B. Bell, L. L. Stohl and A. M. Lambowitz, 1984 The DNA sequence and genetic organization of a *Neurospora* mitochondrial plasmid suggest a relationship to introns and mobile elements. Cell **38:** 441-453.

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

Neri, C., H. M. Cann and J. Dausset, 1996 Triplets répétés, maladies neurodégénératives et psychiatriques : mécanismes et gènes candidats. Médecine/Sciences **12:** 1361-1369.

Nunberg, J. H., R. J. Kaufman, R. T. Schimke, G. Urlaub and L. A. Chasin, 1978 Amplified dihydrofolate reductase genes are localized to a homogeneously staining region of a single chromosome in a methotrexate-resistant Chinese hamster ovary cell line. Proc Natl Acad Sci U S A **75:** 5553-5556.

Ohno, S., 1970 *Evolution by gene duplication*. Springer-Verlag, Heidelberg.

Ohno, S., 1984 Birth of a unique enzyme from an alternative reading frame of the preexisted, internally repetitious coding sequence. Proc Natl Acad Sci U S A **81:** 2421-2425.

Ohno, S., 1985 Immortal genes. Trends in Genetics **1:** 196-200.

Ohno, S., 1987a Early genes that were oligomeric repeats generated a number of divergent domains on their own. Proc Natl Acad Sci U S A **84:** 6486-6490.

Ohno, S., 1987b Evolution from primordial oligomeric repeats to modern coding sequences. J Mol Evol **25:** 325-329.

Ohta, T., 1987a A model of evolution for accumulating genetic information. J Theor Biol **124:** 199-211.

Ohta, T., 1987b Simulating evolution by gene duplication. Genetics **115:** 207-213.

Ohta, T., 1988 Time for acquiring a new gene by duplication. Proc Natl Acad Sci U S A **85:** 3509-3512.

Ohta, T., 1991 Multigene families and the evolution of complexity. J Mol Evol **33:** 34-41.

Paldi, A., G. Gyapay and J. Jami, 1995 Imprinted chromosomal regions of the human genome display sex-specific meiotic recombination frequencies. Curr Biol **5:** 1030-1035.

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

Paques, F., W. Y. Leung and J. E. Haber, 1998 Expansions and contractions in a tandem repeat induced by double-strand break repair. Mol Cell Biol **18:** 2045-2054.

Patten, P., T. Yokota, J. Rothbard, Y. Chien, K. Arai *et al.*, 1984 Structure, expression and divergence of T-cell receptor beta-chain variable regions. Nature **312:** 40-46.

Pavlicek, A., K. Jabbari, J. Paces, V. Paces, J. V. Hejnar *et al.*, 2001 Similar integration but different stability of Alus and LINEs in the human genome. Gene **276:** 39-45.

Payseur, B. A., and M. W. Nachman, 2000 Microsatellite variation and recombination rate in the human genome. Genetics **156:** 1285-1298.

Pebusque, M. J., F. Coulier, D. Birnbaum and P. Pontarotti, 1998 Ancient large-scale genome duplications: phylogenetic and linkage analyses shed light on chordate genome evolution. Mol Biol Evol **15:** 1145-1159.

Peeters, B. P., B. J. de, S. Bron and G. Venema, 1988 Structural plasmid instability in *Bacillus subtilis*: effect of direct and inverted repeats. Mol Gen Genet **212:** 450-458.

Pennisi, E., 2001 Genetics. Closing in on the centromere. Science **294:** 30-31.

Petit, M. A., J. M. Mesas, P. Noirot, F. Morel-Deville and S. D. Ehrlich, 1992 Induction of DNA amplification in the *Bacillus subtilis* chromosome. Embo J **11:** 1317-1326.

Pfeiffer, P., and T. Hohn, 1983 Involvement of reverse transcription in the replication of cauliflower mosaic virus: a detailed model and test of some aspects. Cell **33:** 781-789.

Phillips, M., P. Djian and H. Green, 1990 The involucrin gene of the galago. Existence of a correction process acting on its segment of repeats. J Biol Chem **265:** 7804-7807.

Plasterk, R. H., Z. Izsvak and Z. Ivics, 1999 Resident aliens: the Tc1/mariner superfamily of transposable elements. Trends Genet **15:** 326-332.

Pryde, F. E., H. C. Gorham and E. J. Louis, 1997 Chromosome ends: all the same under their caps. Curr Opin Genet Dev **7:** 822-828.

Pupko, T., and D. Graur, 1999 Evolution of microsatellites in the yeast *Saccharomyces cerevisiae*: role of length and number of repeated units. J Mol Evol **48:** 313-316.

Ricchetti, M., C. Fairhead and B. Dujon, 1999 Mitochondrial DNA repairs double-strand breaks in yeast chromosomes. Nature **402:** 96-100.

Rice, P., and K. Mizuuchi, 1995 Structure of the bacteriophage Mu transposase core: a common structural motif for DNA transposition and retroviral integration. Cell **82:** 209-220.

Richard, G. F., C. Hennequin, A. Thierry and B. Dujon, 1999 Trinucleotide repeats and other microsatellites in yeasts. Res Microbiol **150:** 589-602.

Richard, G. F., and F. Paques, 2000 Mini- and microsatellite expansions: the recombination connection. EMBO Rep **1:** 122-126.

Richardson, C., and M. Jasin, 2000 Coupled homologous and nonhomologous repair of a double-strand break preserves genomic integrity in mammalian cells. Mol Cell Biol **20:** 9068-9075.

Robinson-Rechavi, M., O. Marchand, H. Escriva, P. L. Bardet, D. Zelus *et al.*, 2001 Euteleost fish genomes are characterized by expansion of gene families. Genome Res **11:** 781-788.

Rocha, E. P. C., and A. Blanchard, 2002 Genomic repeats , genome plasticity and the dynamics of *Mycoplasma* evolution. Nucleic Acids Res **In press**.

Rocha, E. P. C., A. Danchin and A. Viari, 1999a Analysis of long repeats in bacterial genomes reveals alternative evolutionary mechanisms in *Bacillus subtilis* and other competent prokaryotes. Mol Biol Evol **16:** 1219-1230.

Rocha, E. P. C., A. Danchin and A. Viari, 1999b Functional and evolutionary roles of long repeats in prokaryotes. Res Microbiol **150:** 725-733.

Rocha, E. P. C., A. Danchin and A. Viari, 1999c Universal replication biases in bacteria. Mol Microbiol **32:** 11-16.

Roelants, F., S. Potier, J. L. Souciet and J. de Montigny, 1995 Reactivation of the ATCase domain of the URA2 gene complex: a positive selection method for Ty insertions and chromosomal rearrangements in *Saccharomyces cerevisiae*. Mol Gen Genet **246:** 767-773.

Romero, D., J. Martinez-Salazar, E. Ortiz, C. Rodriguez and E. Valencia-Morales, 1999 Repeated sequences in bacterial chromosomes and plasmids: a glimpse from sequenced genomes. Res Microbiol **150:** 735-743.

Romero, D., and R. Palacios, 1997 Gene amplification and genomic plasticity in prokaryotes. Annu Rev Genet **31:** 91-111.

Rosenberg, H., M. Singer and M. Rosenberg, 1978 Highly reiterated sequences of SIMIANSIMIANSIMIANSIMIANSIMIAN. Science **200:** 394-402.

Saffery, R., L. H. Wong, D. V. Irvine, M. A. Bateman, B. Griffiths *et al.*, 2001 Construction of neocentromere-based human minichromosomes by telomere-associated chromosomal truncation. Proc Natl Acad Sci U S A **98:** 5705-5710.

Sagot, M. F., A. Viari, J. Pothier and H. Soldano, 1995 finding flexible pattern in a text - an application to 3D molecaular matching, pp. 117-145 in *1st IEEE workshop on stage and patterns matching in computanionnal biology*. IEEE, Seattle.

Saigo, K., W. Kugimiya, Y. Matsuo, S. Inouye, K. Yoshioka *et al.*, 1984 Identification of the coding sequence for a reverse transcriptase-like enzyme in a transposable genetic element in *Drosophila melanogaster*. Nature **312:** 659-661.

Saito, H., D. M. Kranz, Y. Takagaki, A. C. Hayday, H. N. Eisen *et al.*, 1984 A third rearranged and expressed gene in a clone of cytotoxic T lymphocytes. Nature **312:** 36-40.

Sargent, R. G., M. A. Brenneman and J. H. Wilson, 1997 Repair of site-specific double-strand breaks in a mammalian chromosome by homologous and illegitimate recombination. Mol Cell Biol **17:** 267-277.

Sawyer, S. A., D. E. Dykhuizen, R. F. DuBose, L. Green, T. Mutangadura-Mhlanga *et al.*, 1987 Distribution and abundance of insertion sequences among natural isolates of *Escherichia coli*. Genetics **115:** 51-63.

Scheerer, J. B., and G. M. Adair, 1994 Homology dependence of targeted recombination at the Chinese hamster APRT locus. Mol Cell Biol **14:** 6663-6673.

Schlotterer, C., and D. Tautz, 1992 Slippage synthesis of simple sequence DNA. Nucleic Acids Res **20:** 211-215.

Schmidtke, J., and I. Kandt, 1981 Single-copy DNA relationships between diploid and tetraploid teleostean fish species. Chromosoma **83:** 191-197.

Schneider, T. D., G. D. Stormo, L. Gold and A. Ehrenfeucht, 1986 Information content of binding sites on nucleotide sequences. J Mol Biol **188:** 415-431.

Schueler, M. G., A. W. Higgins, M. K. Rudd, K. Gustashaw and H. F. Willard, 2001 Genomic and genetic definition of a functional human centromere. Science **294:** 109-115.

Schug, M. D., T. F. Mackay and C. F. Aquadro, 1997 Low mutation rates of microsatellite loci in *Drosophila melanogaster*. Nat Genet **15:** 99-102.

Seeger, C., D. Ganem and H. E. Varmus, 1986 Biochemical and genetic evidence for the hepatitis B virus replication strategy. Science **232:** 477-484.

Seoighe, C., and K. H. Wolfe, 1998 Extent of genomic rearrangement after genome duplication in yeast. Proc Natl Acad Sci U S A **95:** 4447-4452.

Seoighe, C., and K. H. Wolfe, 1999 Updated map of duplicated regions in the yeast genome. Gene **238:** 253-261.

Shen, P., and H. V. Huang, 1986 Homologous recombination in *Escherichia coli*: dependence on substrate length and homology. Genetics **112:** 441-457.

Shimeld, S. M., 1999 Gene function, gene networks and the fate of duplicated genes. Semin Cell Dev Biol **10:** 549-553.

Sia, E. A., R. J. Kokoska, M. Dominska, P. Greenwell and T. D. Petes, 1997 Microsatellite instability in yeast: dependence on repeat unit size and DNA mismatch repair genes. Mol Cell Biol **17:** 2851-2858.

Singer, M. F., 1982 SINEs and LINEs: highly repeated short and long interspersed sequences in mammalian genomes. Cell **28:** 433-434.

Singer, M. F., 1995 Unusual reverse transcriptases. J Biol Chem **270:** 24623-24626.

Singer, M. F., and J. Skowronski, 1985 Making sense out of LINES: long interspersed repeat sequences in mammalian genomes. Trends Bioch Science **82:** 119-122.

Skrabanek, L., and K. H. Wolfe, 1998 Eukaryote genome duplication - where's the evidence? Curr Opin Genet Dev **8:** 694-700.

Smith, G. P., 1976 Evolution of repeated DNA sequences by unequal crossover. Science **191:** 528-535.

Smith, T. F., and M. S. Waterman, 1981 Identification of common molecular subsequences. J Mol Biol **147:** 195-197.

Soldano, H., A. Viari and M. Champesme, 1995 Searching for a flexible repeated patterns in labelled objects using a non transitive similarity relation. Pattern Recognition Letters **16:** 233-246.

Souciet, J., M. Aigle, F. Artiguenave, G. Blandin, M. Bolotin-Fukuhara *et al.*, 2000 Genomic exploration of the hemiascomycetous yeasts: 1. A set of yeast species for molecular evolution studies. FEBS Lett **487:** 3-12.

Stallings, R. L., 1994 Distribution of trinucleotide microsatellites in different categories of mammalian genomic sequence: implications for human genetic diseases. Genomics **21:** 116-121.

Stephan, W., 1989 Tandem-repetitive noncoding DNA: forms and forces. Mol Biol Evol **6:** 198-212.

Stephan, W., and S. Cho, 1994 Possible role of natural selection in the formation of tandem-repetitive noncoding DNA. Genetics **136:** 333-341.

Stern, L. J., J. H. Brown, T. S. Jardetzky, J. C. Gorga, R. G. Urban *et al.*, 1994 Crystal structure of the human class II MHC protein HLA-DR1 complexed with an influenza virus peptide. Nature **368:** 215-221.

Sueoka, N., 1962 On the genetic basis of variation and heterogeneity of DNA base Composition. Proc Natl Acad Sci U S A **48:** 582-592.

Sullivan, K. F., and C. A. Glass, 1991 CENP-B is a highly conserved mammalian centromere protein with homology to the helix-loop-helix family of proteins. Chromosoma **100:** 360-370.

Sullivan, K. F., M. Hechenberger and K. Masri, 1994 Human CENP-A contains a histone H3 related histone fold domain that is required for targeting to the centromere. J Cell Biol **127:** 581-592.

Sun, X., J. Wahlstrom and G. Karpen, 1997 Molecular structure of a functional *Drosophila* centromere. Cell **91:** 1007-1019.

Surzycki, S. A., and W. R. Belknap, 1999 Characterization of repetitive DNA elements in *Arabidopsis*. J Mol Evol **48:** 684-691.

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

Sutherland, G. R., E. Baker and R. I. Richards, 1998 Fragile sites still breaking. Trends Genet **14:** 501-506.

Suzuki, D. T., A. J. F. Griffiths, J. H. Miller and R. C. Lewontin, 1989 Chromosome mutation II: changes in number, pp. 198-214 in *An introduction to genetic analysis (fourth edition)*, edited by W. H. a. C. Freeman, New-York.

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

TAGI, The *Arabidopsis* Genome Initiative, 2000 Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408:** 796-815.

Tanaka, Y., O. Nureki, H. Kurumizaka, S. Fukai, S. Kawaguchi *et al.*, 2001 Crystal structure of the CENP-B protein-DNA complex: the DNA-binding domains of CENP-B induce kinks in the CENP-B box DNA. Embo J **20:** 6612-6618.

TCESC, The *C. elegans* Sequencing Consortium, 1998 Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science **282:** 2012-2018.

Temin, H. M., 1989 Reverse transcriptases. Retrons in bacteria. Nature **339:** 254-255.

Temin, H. M., and S. Mizutani, 1970 RNA-dependent DNA polymerase in virions of Rous sarcoma virus. Nature **226:** 1211-1213.

Terryn, N., L. Heijnen, A. De Keyser, M. Van Asseldonck, R. De Clercq *et al.*, 1999 Evidence for an ancient chromosomal duplication in *Arabidopsis thaliana* by sequencing and analyzing a 400-kb contig at the APETALA2 locus on chromosome 4. FEBS Lett **445:** 237-245.

Tettelin, H., M. L. Agostoni Carbone, K. Albermann, M. Albers, J. Arroyo *et al.*, 1997 The nucleotide sequence of *Saccharomyces cerevisiae* chromosome VII. Nature **387:** 81-84.

Teumer, J., and H. Green, 1989 Divergent evolution of part of the involucrin gene in the hominoids: unique intragenic duplications in the gorilla and human. Proc Natl Acad Sci U S A **86:** 1283-1286.

Thiebaud, C. H., and M. Fischberg, 1977 DNA content in the genus Xenopus. Chromosoma **59:** 253-257.

Thomas, B. J., and R. Rothstein, 1991 Sex, maps, and imprinting. Cell **64:** 1-3.

TIHGSC, The International Human Genome Sequencing Consortium, 2001 Initial sequencing and analysis of the human genome. Nature **409:** 860-921.

Tomascik-Cheeseman, L., F. Marchetti, X. Lowe, F. L. Shamanski, J. Nath *et al.*, 2002 CENP-B is not critical for meiotic chromosome segregation in male mice. Mutat Res **513:** 197-203.

Troester, H., S. Bub, A. Hunziker and M. F. Trendelenburg, 2000 Stability of DNA repeats in *Escherichia coli* dam mutant strains indicates a Dam methylation-dependent DNA deletion process. Gene **258:** 95-108.

Trowsdale, J., 1993 Genomic structure and function in the MHC. Trends Genet **9:** 117-122.

Tseng, H., and H. Green, 1988 Remodeling of the involucrin gene during primate evolution. Cell **54:** 491-496.

Tseng, H., and H. Green, 1989 The involucrin gene of the owl monkey: origin of the early region. Mol Biol Evol **6:** 460-468.

Tudor, M., M. Lobocka, M. Goodell, J. Pettitt and K. O'Hare, 1992 The pogo transposable element family of *Drosophila melanogaster*. Mol Gen Genet **232:** 126-134.

Uyeno, T., and G. R. Smith, 1972 Tetraploid origin of the karyotype of catostomid fishes. Science **175:** 644-646.

Vasudevan, S. G., W. L. Armarego, D. C. Shaw, P. E. Lilley, N. E. Dixon *et al.*, 1991 Isolation and nucleotide sequence of the hmp gene that encodes a haemoglobin-like protein in *Escherichia coli* K-12. Mol Gen Genet **226:** 49-58.

Vergnaud, G., and F. Denoeud, 2000 Minisatellites: mutability and genome architecture. Genome Res **10:** 899-907.

Vergnaud, G., D. Mariat, F. Apiou, A. Aurias, M. Lathrop *et al.*, 1991 The use of synthetic tandem repeats to isolate new VNTR loci: cloning of a human hypermutable sequence. Genomics **11:** 135-144.

Vincens, P., L. Buffat, C. Andre, J. P. Chevrolat, J. F. Boisvieux *et al.*, 1998 A strategy for finding regions of similarity in complete genome sequences. Bioinformatics **14:** 715-725.

Vision, T. J., D. G. Brown and S. D. Tanksley, 2000 The origins of genomic duplications in *Arabidopsis*. Science **290:** 2114-2117.

Viswanathan, M., G. Muthukumar, Y. S. Cong and J. Lenard, 1994 Seripauperins of *Saccharomyces cerevisiae*: a new multigene family encoding serine-poor relatives of serine-rich proteins. Gene **148:** 149-153.

Voelkel-Meiman, K., and G. S. Roeder, 1990 Gene conversion tracts stimulated by HOT1-promoted transcription are long and continuous. Genetics **126:** 851-867.

Voytas, D. F., and F. M. Ausubel, 1988 A copia-like transposable element family in *Arabidopsis thaliana*. Nature **336:** 242-244.

Waldman, A. S., and R. M. Liskay, 1987 Differential effects of base-pair mismatch on intrachromosomal versus extrachromosomal recombination in mouse cells. Proc Natl Acad Sci U S A **84:** 5340-5344.

Walsh, J. B., 1987 Sequence-dependent gene conversion: can duplicated genes diverge fast enough to escape conversion? Genetics **117:** 543-557.

Walsh, J. B., 1995 How often do duplicated genes evolve new functions? Genetics **139:** 421-428.

Warburton, P. E., J. S. Waye and H. F. Willard, 1993 Nonrandom localization of recombination events in human alpha satellite repeat unit variants: implications for higher-order structural characteristics within centromeric heterochromatin. Mol Cell Biol **13:** 6520-6529.

Watt, V. M., C. J. Ingles, M. S. Urdea and W. J. Rutter, 1985 Homology requirements for recombination in *Escherichia coli*. Proc Natl Acad Sci U S A **82:** 4768-4772.

Waye, J. S., and H. F. Willard, 1986 Structure, organization, and sequence of alpha satellite DNA from human chromosome 17: evidence for evolution by unequal crossing-over and an ancestral pentamer repeat shared with the human X chromosome. Mol Cell Biol **6:** 3156-3165.

Waye, J. S., and H. F. Willard, 1989 Human beta satellite DNA: genomic organization and sequence definition of a class of highly repetitive tandem DNA. Proc Natl Acad Sci U S A **86:** 6250-6254.

Welch, J. W., S. Fogel, G. Cathala and M. Karin, 1983 Industrial yeasts display tandem gene iteration at the CUP1 region. Mol Cell Biol **3:** 1353-1361.

Welch, J. W., D. H. Maloney and S. Fogel, 1990 Unequal crossing-over and gene conversion at the amplified CUP1 locus of yeast. Mol Gen Genet **222:** 304-310.

Whitkus, R., J. Doebley and M. Lee, 1992 Comparative genome mapping of Sorghum and maize. Genetics **132:** 1119-1130.

Wierdl, M., M. Dominska and T. D. Petes, 1997 Microsatellite instability in yeast: dependence on the length of the microsatellite. Genetics **146:** 769-779.

Willard, H. F., 1990 Centromeres of mammalian chromosomes. Trends Genet **6:** 410-416.

Willard, H. F., and J. S. Waye, 1987 Hierarchical order in chromosome-specific human alpha satellite DNA. Trends Genet **3:** 192-198.

Wolfe, K. H., 2002 Gene order evolution and paleoploidy in hemiascomycete genomes, pp. in *Séminaire algorithmique et Biologie*, edited by M. F. Sagot, Lyon.

Wolfe, K. H., and D. C. Shields, 1997 Molecular evidence for an ancient duplication of the entire yeast genome. Nature **387:** 708-713.

Wyman, A. R., and R. White, 1980 A highly polymorphic locus in human DNA. Proc Natl Acad Sci U S A **77:** 6754-6758.

Yamada, Y., V. E. Avvedimento, M. Mudryj, H. Ohkubo, G. Vogeli *et al.*, 1980 The collagen gene: evidence for its evolutinary assembly by amplification of a DNA segment containing an exon of 54 bp. Cell **22:** 887-892.

Yoder, J. A., C. P. Walsh and T. H. Bestor, 1997 Cytosine methylation and the ecology of intragenomic parasites. Trends Genet **13:** 335-340.

Young, M., and J. Cullum, 1987 A plausible mechanism for large-scale chromosomal DNA amplification in streptomycetes. FEBS Lett **212:** 10-14.

Yu, X., and A. Gabriel, 1999 Patching broken chromosomes with extranuclear cellular DNA. Mol Cell **4:** 873-881.

Zawadzki, P., and F. M. Cohan, 1995 The size and continuity of DNA segments integrated in Bacillus transformation. Genetics **141:** 1231-1243.

Zhao, X. J., D. Raitt, V. B. P, A. S. Clewell, K. E. Kwast *et al.*, 1996 Function and expression of flavohemoglobin in *Saccharomyces cerevisiae*. Evidence for a role in the oxidative stress response. J Biol Chem **271:** 25131-25138.