Etude des patrons d'évolution asymétrique dans les séquences d'ADN

Anamaria Necsulea

N° 75-2008 Année 2007 - 2008

Thèse

Présentée

devant L'Université Claude Bernard - Lyon 1

pour l'obtention

du Diplôme de Doctorat

(arrêté du 7 août 2006)

soutenue le
20 Juin 2008

par

Anamaria Necşulea

Etude des patrons d'évolution asymétrique dans les séquences d'ADN

Directeur de thèse : Jean R. Lobry

Jury :	Jean R. Lobry	Directeur de thèse
	Eduardo P.C. Rocha	Rapporteur
	Claude Thermes	Rapporteur
	Laurence D. Hurst	Examinateur
	Dominique Mouchiroud	Examinatrice

UNIVERSITÉ CLAUDE BERNARD-LYON 1

Président de l'Université	**M. le Professeur L. COLLET**
Vice-Président du Conseil Scientifique	M. le Professeur J. F. MORNEX
Vice-Président du Conseil d'Administration	M. le Professeur J. LIETO
Vice-Président du Conseil des Etudes et de la Vie Universitaire	M. le Professeur D. SIMON
Secrétaire Général	**M. G. GAY**

SECTEUR SANTÉ

Composantes

UFR de Médecine Lyon R.T.H. Laënnec	Directeur : M. le Professeur P. COCHAT
UFR de Médecine Lyon Grange-Blanche	Directeur : M. le Professeur X. MARTIN
UFR de Médecine Lyon-Nord	Directeur : M. le Professeur J. ETIENNE
UFR de Médecine Lyon-Sud	Directeur : M. le Professeur F.N. GILLY
UFR d'Ontologie	Directeur : M. O. ROBIN
Institut des Sciences Pharmaceutiques et Biologiques	Directeur : M. le Professeur F. LOCHER
Institut Techniques de Réadaptation	Directeur : M. le Professeur MATILLON
Département de Formation et Centre de Recherche en Biologie Humaine	Directeur : M. le Professeur P. FARGE

SECTEUR SCIENCES

Composantes

UFR de Physique	Directeur : M. le Professeur S. FLECK
UFR de Biologie	Directeur : M. le Professeur H. PINON
UFR de Mécanique	Directeur : M. le Professeur H. BEN HADID
UFR de Génie Electrique et des Procédés	Directeur : M. le Professeur G. CLERC
UFR de Sciences de la Terre	Directeur : M. le Professeur P. HANTZPERGUE
UFR de Mathématiques	Directeur : M. le Professeur A. GOLDMAN
UFR d'Informatique	Directeur : M. le Professeur S. AKKOUCHE
UFR de Chimie Biochimie	Directeur : Mme. le Professeur H. PARROT
UFR STAPS	
Observatoire de Lyon	Directeur : M. le Professeur R. BACON
Institut des Sciences et des Techniques de l'Ingénieur de Lyon	Directeur : M. le Professeur J. LIETO
IUT A	Directeur : M. le Professeur M. C. COULET
IUT B	Directeur : M. le Professeur R. LAMARTINE
Institut de Science Financière et d'Assurances	Directeur : M. le Professeur J. C. AUGROS

Remerciements

Je remercie Jean Lobry, mon directeur de thèse, pour m'avoir soutenu et supporté pendant ces quatre dernières années. Je lui suis reconnaissante pour la confiance qu'il a su m'accorder, pour m'avoir donné l'opportunité de réaliser ce travail de thèse en parfaite autonomie, mais aussi pour ses conseils et pour son encadrement scientifique. Je dois aussi le remercier pour m'avoir donné le goût du logiciel libre (peut-on faire le café avec ?), et pour la qualité de ses enseignements, qui ont constitué ma principale source d'inspiration au cours de mon monitorat.

Je remercie Laurent Duret, pour les discussions scientifiques que nous avons eu vers la fin de ma thèse. Je suis toujours impressionnée par la façon dont Laurent trouve toujours les bonnes réponses, par sa formidable culture scientifique et par son inépuisable imagination. Sa gentillesse et son sens de l'humour font qu'on prend du plaisir dans toute discussion avec Laurent, que ce soit lors d'une réunion de travail ou simplement à la pause café. Je finis cette thèse avec un seul grand regret : le fait que ma timidité m'a empêché d'aller plus souvent l'embêter avec mes questions... J'espère que nous continuerons à collaborer à l'avenir, et je compte ne pas refaire la même erreur (ceci est un avertissement plus qu'une promesse !).

Je remercie Bastien Boussau, pour son amitié et pour toute son aide pendant ces quatre dernières années. Bien qu'un peu occupé par sa propre thèse, Bastien a toujours trouvé le temps de relire mes articles et même ce manuscrit, et ces travaux ont été considérablement améliorés grâce à ses critiques. Je le remercie également pour m'avoir proposé des collaborations, qui ont toujours été passionnantes et fructueuses ; travailler avec Bastien est un plaisir. Merci aussi pour sa compagnie et celle de Mathilde pendant les fins de semaines passées au laboratoire - la rédaction avance mieux quand on est plusieurs. J'espère qu'à l'avenir, comme jusqu'à présent, nos interactions auront lieu dans des conditions optimales, aussi bien de température, que de concentration en oxygène...

Je remercie Eduardo Rocha, Claude Thermes, Laurence Hurst et Dominique Mouchiroud, qui m'ont fait l'honneur de porter leur jugement sur ce travail de thèse. Je leur suis reconnaissante pour le temps qu'ils ont eu la gentillesse de m'accorder, et pour leurs critiques et encouragements. Leurs remarques m'ont permis non seulement d'enrichir les résultats présentés ici, mais aussi d'imaginer de futures directions de recherche.

Je remercie Guillaume Beslon, Hubert Charles, Frédéric Menu et Didier Piau,

membres du comité de pilotage de ma thèse. Nos réunions m'ont permis de faire régulièrement le point sur l'avancement (ou pas) de mes travaux, et je les remercie pour leurs conseils et pour leur soutien.

Je remercie Manolo Gouy, Dominique Mouchiroud, Christian Gautier, pour m'avoir accueillie au sein de l'équipe Bioinformatique et Génomique Évolutive, et au sein du Laboratoire de Biométrie et Biologie Évolutive. J'ai ainsi pu bénéficier tout le long de ma thèse d'excellentes conditions de travail et ressources matérielles. Je remercie également Marie-France Sagot, qui m'a généreusement reçue dans les locaux de l'équipe BAOBAB pendant mon année de Master.

Je remercie Nathalie Arbasetti, Misou Pieri, Agnès Python pour leur aide avec mes nombreux soucis administratifs. Je remercie Stéphane Delmotte, Lionel Humblot, Simon Penel, Bruno Spataro pour leur aide avec mes encore plus nombreux soucis informatiques ...

Je remercie tous les membres de l'équipe et du laboratoire, avec qui j'ai eu le plaisir d'interagir : mes collègues de bureau (Alexandra, Claire, Hugo, Jean-François, Leo, Yann, Yves), ou d'étage (Anne-Muriel, Anne-Sophie, Céline, Daniel, Eric, Gabriel, Guy, Jean T., Lauranne, Laurent G., Marie, Martin, Raquel, Sophie, Sylvain, Vincent D., Vincent L.), ou de l'étage en dessous (Claire, Élise, Émilie, Ludo, Manu, Marilia, Perrine, Sabine, Thibaut, Vincent).

Enfin, je remercie ma famille, pour tout leur soutien, et Florin, pour ses encouragements déguisés, façon psychologie inversée : "Je t'avais dit de pas faire une thèse..."

Table des matières

Préambule

Le 21^{e} siècle s'annonce propice pour la génomique. Nous disposons aujourd'hui de plusieurs centaines, bientôt plusieurs milliers de génomes complètement séquencés, dont notamment le génome humain (Lander *et al.*, 2001; Collins *et al.*, 2004). Le développement de nouvelles technologies, comme par exemple le pyroséquençage (Margulies *et al.*, 2005), permet actuellement d'obtenir des fragments d'ADN totalisant plusieurs millions de nucléotides en seulement quelques heures. L'analyse bioinformatique des génomes ne peut que profiter de cette avalanche des données. Cependant, un danger existe : immergé dans l'océan de A, C, G, T, le bioinformaticien risque d'oublier que derrière les suites de lettres qui représentent de manière simplifiée les séquences d'ADN se trouvent des structures biochimiques complexes.

La structure tridimensionnelle de l'acide désoxyribonucléique est une découverte fondamentale de la biologie moléculaire. La publication du modèle de la double hélice (Watson et Crick, 1953) a valu l'attribution du prix Nobel à ses auteurs, ainsi qu'à Maurice Wilkins. Malheureusement, toutes les données expérimentales à la base de ce modèle n'ont pas connu la même popularité. Les travaux de cristallographie de Rosalind Franklin représentent une des sources d'inspiration dans le développement du modèle de la double hélice, et les travaux de biochimie d'Erwin Chargaff en sont une autre. Dès 1949, Chargaff et ses collaborateurs s'étaient intéressés à la composition en nucléotides des molécules d'ADN (Vischer *et al.*, 1949; Chargaff *et al.*, 1949). Leur résultats suggèrent fortement (du moins *a posteriori*) une régularité de la composition, par les égalités entre le nombre d'adénines et le nombre de thymines, et entre le nombre de guanines et le nombre de cytosines, et cela pour différentes espèces (Vischer *et al.*, 1949). Cette régularité a été confirmée dans plusieurs études réalisées par les mêmes auteurs, ce qui les a conduit à proposer explicitement, en 1951, qu'il pourrait s'agir là d'une caractéristique structurelle des molécules d'ADN :

> *Not only the ratio of purines to pyrimidines but also that of adenine to thymine and of guanine to cytosine equals 1. As the number of examples of such regularity increases, the question will become pertinent whether it is merely accidental or whether it is an* expression

> of certain structural principles *that are shared by many desoxypentose nucleic acids, despite far reaching differences in their individual composition and the absence of a recognizable periodicity in their nucleotide sequence* (Chargaff *et al.*, 1951).

Cette propriété biochimique de l'ADN a mené Watson et Crick à proposer les règles d'appariement $A \cdot T$ et $G \cdot C$ pour la structure en double hélice.

E. Chargaff et ses collaborateurs ont ensuite mis en évidence une autre régularité compositionnelle des acides nucléiques : le nombre de bases 6-amino est égal à celui des bases 6-kéto ($[A + C] = [G + T]$), dans des séquences d'ADN à l'état simple brin (Karkas *et al.*, 1968; Rudner *et al.*, 1968, 1969). Les données biochimiques de Chargaff et ses collaborateurs suggèrent même une plus forte règle de symétrie : $[A] \approx [T]$ et $[C] \approx [G]$. Cette règle de parité des molécules d'ADN simple brin a été redécouverte et confirmée des années plus tard, lorsque les premières séquences génomiques complètes sont devenues disponibles (Prabhu, 1993; Lobry, 1995; Sueoka, 1995).

La justification biologique des équimolarités $[A] \approx [T]$ et $[C] \approx [G]$ pour l'ADN simple brin n'est pas triviale. Les fondements de l'explication ont été posés par Wu et Maeda (1987) et développés ensuite par Sueoka (1995); Lobry (1995). Lorsque les processus évolutifs (pressions de mutations et sélection) agissent de manière équivalente sur les deux brins d'ADN, les taux de substitution des nucléotides sont eux aussi symétriques. Il est montré que sous cette hypothèse de symétrie des processus évolutifs, à l'état d'équilibre les séquences d'ADN respectent la règle de parité $[A] = [T]$ et $[C] = [G]$ (Wu et Maeda, 1987; Lobry, 1995; Sueoka, 1995).

En biologie moléculaire, l'impact de cette deuxième découverte de Chargaff a certes été moins fort que celui des observations qui ont conduit à la proposition de la structure en double hélice. Mais d'un point de vue évolutif la règle de parité intra-brin est très importante, car elle a servi de base pour le développement d'un modèle falsifiable, d'une hypothèse nulle pour l'évolution des séquences d'ADN. Nous devons nous interroger : pourquoi les processus évolutifs seraient-ils symétriques sur les deux brins d'ADN ? Mais pourquoi ne le seraient-ils pas ? Rien dans la *structure* de la molécule d'ADN ne permet de distinguer *a priori* les deux brins. Si les deux brins d'ADN évoluent de façons différentes, la structure physique du génome n'est pas en cause, mais son *mode de fonctionnement*.

En effet, le rejet de l'hypothèse d'évolution symétrique a apporté une contribution significative à la compréhension actuelle de l'évolution des séquences d'ADN. La règle de parité intra-brin est respectée sur de longs fragments d'ADN, voire des chromosomes entiers, mais des déviations locales existent dans la quasi-totalité des espèces. L'asymétrie de composition des séquences a été mise en relation avec des processus cellulaires essentiels, comme la réplication (Lobry, 1996a; Touchon

et al., 2005) ou la transcription (Touchon *et al.*, 2003). Cette propriété des génomes a ouvert la voie pour le développement de certaines applications pratiques, notamment de nombreuses méthodes *in silico* pour la détection des origines de réplication (Lobry, 1996b; Grigoriev, 1998; Salzberg *et al.*, 1998; Frank et Lobry, 2000; Worning *et al.*, 2006). Par conséquent, dans les dernières années de nombreuses études se sont intéressées à l'asymétrie de composition, et le présent document ne fait pas exception.

L'objectif de ce travail de thèse est d'étudier les processus évolutifs qui sont à l'origine de l'asymétrie de composition. Dans un premier chapitre, nous passerons en revue les connaissances actuelles sur les mécanismes biologiques qui pourraient générer ce biais de composition : la réplication et la transcription. Le but de ce chapitre n'est pas de donner une description exhaustive de ces deux processus ; nous essayerons néanmoins d'illustrer les aspects de leur fonctionnement qui pourraient expliquer l'asymétrie d'évolution des deux brins d'ADN.

Si dans le premier chapitre la réplication et la transcription sont étudiées séparément, nous verrons par la suite qu'en réalité la distinction entre ces deux sources d'asymétrie n'est pas triviale. Dans le deuxième chapitre, nous discuterons de l'importance du mode de réplication de l'ADN sur la distribution des gènes sur les chromosomes procaryotes. Nous nous intéresserons en particulier au phénomène de co-orientation entre réplication et transcription. Nous présenterons notamment une approche *in silico* pour l'étude de l'asymétrie de composition qui permet de découpler les effets de ces deux mécanismes.

Nous nous intéresserons ensuite aux patrons de substitution ponctuelle qui sont à l'origine de l'asymétrie de composition. Dans le troisième chapitre, nous étudierons l'influence du contexte nucléotidique sur l'asymétrie d'évolution des séquences. Nous examinerons en particulier deux hypothèses qui sont couramment avancées pour expliquer le biais de composition des séquences transcrites : l'existence d'un biais de réparation de l'ADN au cours de la transcription et la mutagénicité de la transcription.

Dans le quatrième chapitre, nous continuerons notre analyse sur l'influence du contexte nucléotidique sur l'asymétrie des processus évolutifs, en étudiant l'évolution de séquences qui présentent une composition en nucléotide extrême : les microsatellites. Nous démontrerons que les processus évolutifs qui sont à l'origine de l'asymétrie de composition agissent de la même manière sur les microsatellites et sur les séquences à composition non-biaisée.

Nous devons conclure ce préambule en précisant que cette thèse ne prétend pas être une documentation exhaustive des mécanismes biologiques asymétriques. Certains aspects ont été abordés de manière très détaillée dans la littérature, et une des difficultés de ce travail de thèse a été d'éviter la redondance avec les résultats déjà existants. Les travaux décrits ici ont comme point de départ plusieurs études fondamentales, et il est important de les citer dès maintenant.

Pour une discussion de l'influence respective des processus évolutifs neutres ou sélectifs sur la mise en place de l'asymétrie de composition, nous considérons qu'une lecture de l'article de Frank et Lobry (1999) est nécessaire. L'impact du mécanisme de réplication sur l'organisation des génomes procaryotes a été décrit de manière très détaillée par Rocha (2004). Enfin, les problèmes liés à la détection des origines de réplication dans les génomes de mammifères, ainsi que l'impact de la réplication sur la structuration du génome humain sont discutés d'une manière exhaustive dans le manuscrit de thèse de Marie Touchon.

Chapitre 1

Introduction

Dans la plupart des organismes vivants, l'information génétique est emmagasinée sous la forme de molécules d'ADN double brin, dont la structure physique est intrinsèquement symétrique. Cette symétrie est importante d'un point de vue évolutif, car les mutations qui affectent les deux brins ne peuvent pas être discernées *a priori*. Une paire de nucléotides $A \cdot T$ peut se transformer dans une paire $G \cdot C$ soit par une mutation $A \to G$ dans le premier brin, soit par une mutation $T \to C$ dans le brin complémentaire.

Il est important de prendre en compte la complémentarité des molécules d'ADN lorsque l'on calcule des taux de mutation par rapport à un seul des deux brins. Les mutations d'une base X vers une base Y peuvent alors se décomposer en deux facteurs : $X \to Y = (X \to Y)_1 + (\bar{X} \to \bar{Y})_2$, où $\bar{X}$ et $\bar{Y}$ représentent les bases complémentaires de X et Y, et les indices 1 et 2 désignent les deux brins d'ADN. La mutation dans le sens complémentaire sera alors $\bar{X} \to \bar{Y} = (\bar{X} \to \bar{Y})_1 + (X \to Y)_2$. Si le mode d'évolution est identique sur les deux brins d'ADN, il s'ensuit que $(X \to Y)_1 = (X \to Y)_2$ et $(\bar{X} \to \bar{Y})_1 = (\bar{X} \to \bar{Y})_2$, ce qui implique que si l'on regarde les mutations du point de vue d'un seul brin, les changements complémentaires $X \to Y$ et $\bar{X} \to \bar{Y}$ doivent apparaître avec la même fréquence.

Si l'on modélise l'évolution des séquences d'ADN en imposant la contrainte d'égalité des taux de changements complémentaires, il a été démontré qu'à l'état d'équilibre les règles de parité $[A] = [T]$ et $[G] = [C]$ sont attendues sur chacun des brins (Wu et Maeda, 1987; Lobry, 1995). Cette prédiction théorique est confirmée lorsque l'on étudie des chromosomes entiers, ou simplement de très longues séquences (Lobry, 1995), mais localement les séquences d'ADN présentent de fortes asymétries de composition.

L'asymétrie de composition n'est pas localisée aléatoirement le long des chromosomes, au contraire, elle est associée à l'organisation fonctionnelle des génomes. Cette structuration est particulièrement visible sur les chromosomes procaryotes,

comme celui d'*Escherichia coli* (*cf.* figure 1.1), où l'asymétrie de composition est relativement constante sur chacune des deux moitiés, mais change de signe brutalement au niveau de l'unique origine de réplication de ce chromosome (Lobry, 1996a). L'association entre asymétrie de composition et structuration en réplichores est présente non seulement chez les espèces procaryotes, mais également dans le domaine eucaryote, notamment dans le génome humain (Touchon *et al.*, 2005). Plus important encore, l'asymétrie de composition associée à la réplication semble être quasi-universelle, car dans la grande majorité des espèces le sens des déviations par rapport aux règles de parités est le même ($[G] > [C]$ et $[T] > [A]$ sur le brin avancé pour la réplication ; seulement quelques exceptions à la règle $[G] > [C]$ sont connues à ce jour, bien que des biais $[T] < [A]$ soient rencontrés pour un plus grand nombre d'espèces) (Rocha *et al.*, 1999).

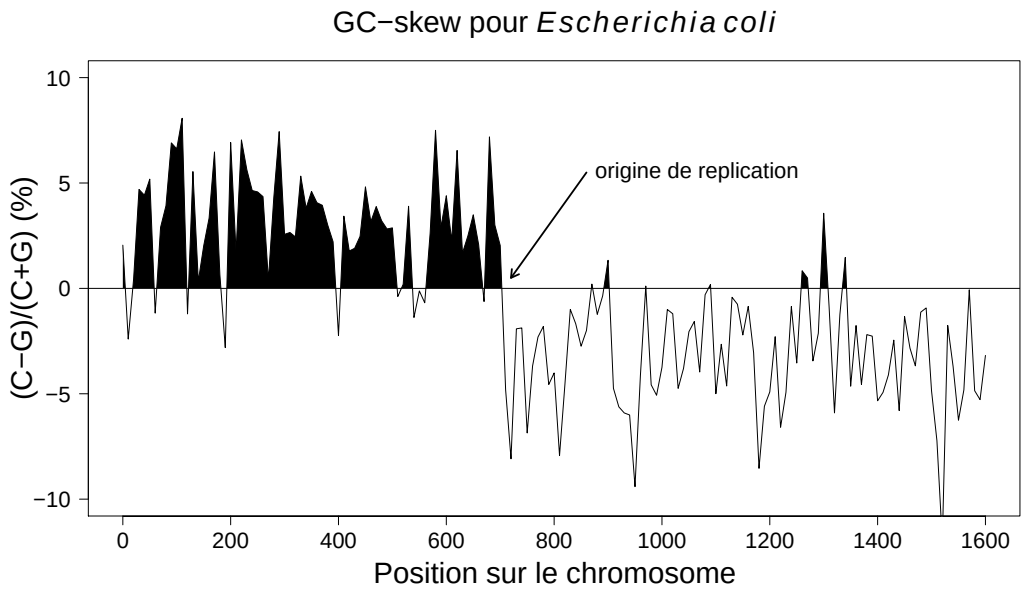

FIG. 1.1 – Déviation par rapport à la règle de parité $[G] = [C]$ (mesurée par le biais de GC, c'est à dire par le rapport $\frac{[C]-[G]}{[C]+[G]}$), le long du chromosome d'*Escherichia coli* (Lobry, 1996a).

L'asymétrie de composition est également associée avec un autre mécanisme cellulaire fondamental : la transcription. Ainsi, il a été démontré que dans le génome humain les régions transcrites présentent d'importantes déviations par rapport aux règles de parité $[A] = [T]$ et $[G] = [C]$ (*cf.* figure 1.2, Touchon *et al.* (2003)).

Ce biais de composition en bases associé à la réplication ou à la transcription peut en théorie être engendré aussi bien par des processus évolutifs neutres (mutation) que par la sélection naturelle. Une discussion détaillée de ces deux

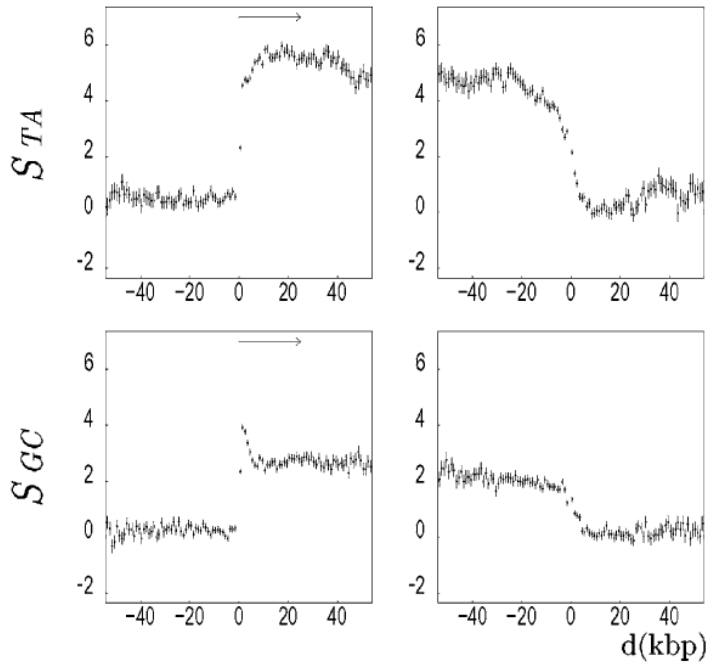

FIG. 1.2 – Déviation par rapport aux règles de parité $[A] = [T]$ et $[G] = [C]$ (mesurées par les ratios $S_{TA} = \frac{[T]-[A]}{[T]+[A]}$ et $S_{GC} = \frac{[G]-[C]}{[G]+[C]}$), dans des régions transcrites du génome humain (Touchon *et al.*, 2003). La flèche représente le sens de transcription du gène. Les graphiques sont centrées sur la position de début de la transcription. Les coordonnées négatives correspondent à des régions non-transcrites dans les graphiques à gauche, et à des régions transcrites dans les graphiques à droite.

possibilités, pour le cas de la réplication, a été donnée par Frank et Lobry (1999). L'hypothèse qui paraît aujourd'hui la plus vraisemblable est que dans les deux cas il s'agit en effet de processus évolutifs neutres, car l'asymétrie de composition est particulièrement forte dans les régions peu contraintes du génome (positions synonymes des régions codantes, régions intergéniques et introns). La nature de ces mécanismes n'est cependant toujours pas parfaitement élucidée.

Pour engendrer un tel biais de composition, les mutations doivent être distribuées de manière asymétrique sur les deux brins d'ADN. L'association avec les processus de réplication et transcription paraît d'autant plus logique, car ces deux mécanismes agissent de manière différente sur les deux brins d'ADN. Dans ce chapitre, nous ferons une analyse du fonctionnement de la réplication et de la transcription, en mettant l'accent sur les aspects qui permettraient d'expliquer

l'asymétrie des processus évolutifs.

1.1 Asymétrie du mécanisme de réplication

Au moment de la découverte de la structure en double hélice, le fait que l'ADN représente le support biochimique de l'hérédité était déjà connu (Avery *et al.*, 1944). Cependant, la manière dont le matériel génétique est multiplié pour être transmis à la génération suivante a été perçue pour la première fois grâce aux travaux de Watson et Crick, qui ont vu dans l'appariement des deux brins d'ADN un possible mécanisme de réplication :

> *It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.* (Watson et Crick, 1953)

Comme prédit par Watson et Crick, il a été démontré que la réplication de l'ADN se fait de manière semi-conservative, chacun des deux brins servant de matrice pour la synthèse du brin complémentaire (Meselson et Stahl, 1958).

La synthèse des molécules d'ADN est assurée par des complexes enzymatiques, appelés polymérases. Une caractéristique commune aux polymérases dans les trois domaines du vivant est le fait que l'élongation d'une chaîne d'ADN ne peut se faire que dans la direction $5' \rightarrow 3'$. Or, les deux chaînes d'ADN sont antiparallèles, ce qui implique que la synthèse des deux nouveaux brins à partir des brins parentaux se fait dans deux orientations divergentes. Pour que la synthèse des deux molécules d'ADN "filles" se fasse simultanément par la même machinerie de réplication, les cellules on trouvé une solution : l'un des deux brins est synthétisé de manière continue (il est appelé alors brin "leading" ou brin "avancé"), alors que le brin complémentaire (appelé "lagging" ou "tardif") est synthétisé de manière discontinue, sous la forme de fragments d'Okazaki (*cf.* figure 1.3, Kornberg (1988)). Il faut néanmoins noter que ce mode de réplication n'est pas universel ; par exemple, l'ADN mitochondrial de certaines familles d'eucaryotes est répliqué par le processus du "cercle roulant", au cours duquel la synthèse des deux brins est découplée (Nosek et Tomaska, 2003).

La réplication de l'ADN est le mécanisme fondamental de la cellule, et il n'est pas surprenant que ce processus assure la transmission du matériel génétique à la génération suivante avec une très grande fidélité. L'exactitude de la copie est garantie d'abord par la précision de l'insertion de nouveaux nucléotides par la polymérase, qui a lieu chez *E. coli* avec un taux d'erreur de seulement 10^{-4} par base répliquée (Kunkel et Alexander, 1986; Kornberg, 1988). Les erreurs éventuelles sont corrigées ensuite par une exonucléase, qui agit dans la direction $3' \rightarrow 5'$, et qui réduit le taux d'erreur jusqu'à 10^{-6} par base (Kornberg, 1988). Une troisième

étape de correction des mésappariements réduit encore davantage le taux d'erreur (Radman et Wagner, 1986). Le taux de mutation par génération est donc généralement très faible, mais il n'est certainement pas nul.

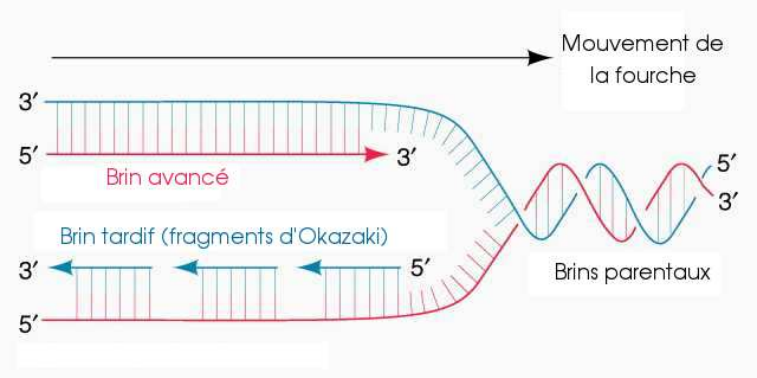

FIG. 1.3 – Schéma décrivant le mode de réplication semi-discontinue d'une molécule d'ADN. Le brin avancé est représenté en rouge et le brin tardif en bleu.

Vu que le mode usuel de réplication de l'ADN permet de distinguer le brin avancé du brin tardif, il est pertinent de demander si le taux de mutation sous-jacent au mécanisme de réplication peut être différent sur les deux brins d'ADN. Si une telle différence existe, elle peut venir de plusieurs sources. Une première possibilité serait que les taux d'erreur d'incorporation des nucléotides ne sont pas identiques pour la synthèse des deux brins. Cela pourrait s'expliquer si les polymérases qui répliquent les deux brins ont des niveaux de précision différents, ou si les mécanismes d'excision des nucléotides insérés de manière erronée ne sont pas aussi efficaces sur le brin avancé et sur le brin tardif. Une deuxième possibilité serait que les taux de mutation spontanée des nucléotides sont différents sur les deux brins, peut-être dû aux différences de structure biochimique au moment de la réplication. Nous discuterons ci-dessous de ces différentes possibilités.

1.1.1 Asymétrie de la machinerie de réplication

La machinerie de réplication de l'ADN n'est pas identique dans les trois domaines du vivant ; il existe des différences importantes entre les bactéries d'une part, et les archées et les eucaryotes d'autre part (Olsen et Woese, 1997). Les complexes enzymatiques qui assurent la réplication sont bien connus pour les espèces bactériennes *Escherichia coli* et (dans une moindre mesure) *Bacillus subtilis*. Chez les eucaryotes les connaissances disponibles viennent surtout des études effectuées chez la levure *Saccharomyces cerevisiae*, mais aussi chez l'humain.

Machinerie de réplication chez les Bactéries

Chez *Escherichia coli*, la synthèse de l'ADN est assurée par le complexe enzymatique Pol III (McHenry et Kornberg, 1977). Le complexe Pol III est constitué de dix sous-unités, codés par différents gènes (*cf.* figure 1.4). Les sous-unités α (codée par le gène *dnaE*), ϵ (gènes *dnaQ* et *mutD*) et θ (gène *holE*) constituent le "cœur" catalytique du complexe enzymatique ; ce cœur est présent en deux copies (Kelman et O'Donnell, 1995). La sous-unité α est l'unité responsable pour l'élongation de la chaîne d'ADN (Maki *et al.*, 1985), alors que la sous-unité ϵ effectue la réparation des erreurs d'incorporation des nucléotides, par une activité exonucléase $3' \rightarrow 5'$ (Scheuermann et Echols, 1984). Les sous-unités γ, δ, δ', χ et ψ forment le complexe γ, qui assure le chargement de Pol III sur la molécule d'ADN, conjointement avec deux sous-unités β (Kelman et O'Donnell, 1995). Les deux cœurs du complexe Pol III sont physiquement couplés par des interactions avec deux sous-unités τ (McHenry, 1982; Kim *et al.*, 1996). L'assemblage des sous-unités τ et du complexe γ est aussi appelé complexe DnaX (McHenry, 2003). Un schéma de l'organisation du complexe Pol III est donné dans la figure 1.4.

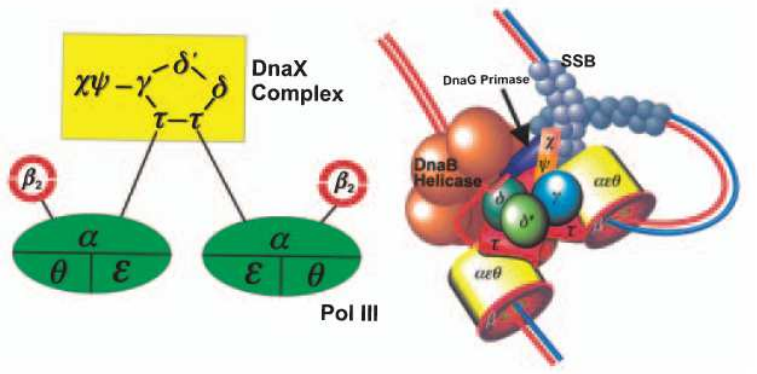

FIG. 1.4 – Structure de l'ADN polymérase III (Pol III) chez *Escherichia coli*. D'après McHenry (2003).

La réplication des brins avancé et tardif se fait au sein du même complexe Pol III, par les deux cœurs catalytiques (Johanson et McHenry, 1984; Maki *et al.*, 1988; Glover et McHenry, 2001). Chez *E. coli*, la réplication des deux brins est donc réalisée par deux polymérases identiques, mais le complexe Pol III est néanmoins asymétrique par rapport aux deux brins d'ADN (Yuzhakov *et al.*, 1996). L'asymétrie de Pol III est créée par l'orientation du complexe de chargement sur l'ADN ("clamp loader"), qui est dirigé vers le brin tardif, pour favoriser la formation de complexes d'initiation pendant la synthèse des fragments d'Okazaki (Glover et McHenry, 2001; McHenry, 2003).

La structure du complexe Pol III est moins bien déterminée chez les autres espèces bactériennes. Chez *Bacillus subtilis*, la sous-unité catalytique (α) est plus grande que l'enzyme correspondante chez *E. coli* et possède une activité $3' \rightarrow 5'$ exonucléase (Low *et al.*, 1976). Cet enzyme est codé par le gène *polC* (Sanjanwala et Ganesan, 1989). L'analyse des premières séquences génomiques complètes a démontré que les génomes des bactéries Gram positives à faible taux de GC (dont *B. subtilis*) contiennent deux gènes codant pour des polymérases réplicatives : le gène *polC* et le gène *dnaE*, qui est homologue au gène codant pour la sous-unité α chez *E. coli* (Koonin et Bork, 1996; Kunst *et al.*, 1997).

Une étude expérimentale a démontré que chez *Streptococcus pyogenes* ces deux enzymes fonctionnent au niveau de la fourche de réplication et il a été suggéré qu'elles pourraient avoir des rôles différents dans la synthèse des deux brins d'ADN (Bruck et O'Donnell, 2000). De même, ces deux enzymes sont essentielles pour la réplication chez *B. subtilis*, et il est probable que *polC* et *dnaE* sont responsables pour la réplication des brins avancé et tardif, respectivement (Dervyn *et al.*, 2001). Ces études ont été confirmées chez *Staphylococcus aureus* (Inoue *et al.*, 2001).

La polymérase DnaE de *B. subtilis* ne possède pas d'activité $3' \rightarrow 5'$ exonucléase, contrairement à PolC, et semble être également impliquée dans des processus de mutagénèse induite par les rayonnements UV (Le Chatelier *et al.*, 2004). Des études *in vitro* ont démontré que chez *S. pyogenes* DnaE a un taux élevé d'erreur d'incorporation des nucléotides (Bruck et O'Donnell, 2000; Bruck *et al.*, 2003). Si cet enzyme est effectivement resposable de la synthèse du brin tardif, il est donc possible que la synthèse du brin tardif se fasse avec plus d'erreurs que celle du brin avancé.

Machinerie de réplication chez les Eucaryotes

Chez les eucaryotes, la réplication de l'ADN est réalisée par trois complexes enzymatiques essentiels : Pol α (aussi appelée ADN polymérase I chez la levure), Pol δ (ou ADN polymérase III) et Pol ϵ (ou ADN polymérase II)(Campbell, 1993). La polymérase α intervient dans l'initiation de la réplication, pour les deux brins d'ADN (Nethanel et Kaufmann, 1990). Le rôle des polymérases δ et ϵ a été plus difficile a élucider. Les premiers modèles pour la réplication chez *S. cerevisiae* ont proposé que les polymérases δ et ϵ pourraient effectuer séparément la synthèse des brins avancé et tardif, ce qui serait cohérent avec le caractère essentiel de ces deux enzymes (Morrison *et al.*, 1990; Burgers, 1991). Cependant, une étude réalisée *in vitro* a montré que la polymérase δ peut synthétiser le brin avancé, ainsi que le brin tardif, une fois que les fragments d'Okazaki ont été initiés par la polymérase α (Waga et Stillman, 1994). Le rôle de la polymérase δ dans la maturation des fragments d'Okazaki a été confirmé plus tard (Jin *et al.*, 2003).

Des études de mutagénèse ont réussi à démontrer que les polymérases δ et

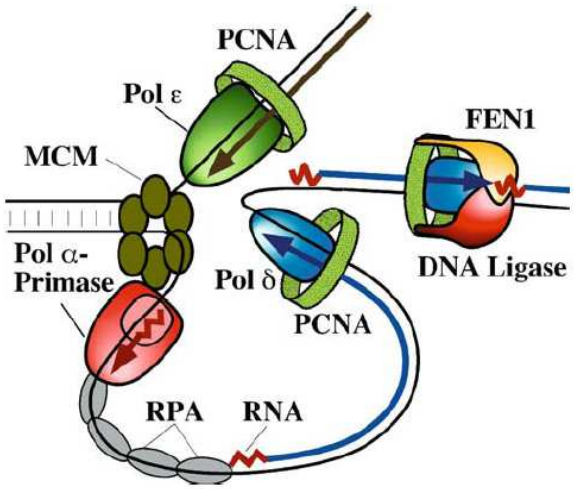

FIG. 1.5 – Complexes enzymatiques présents au niveau de la fourche de réplication chez les eucaryotes. D'après Garg et Burgers (2005).

ϵ sont chacune responsable de la réplication d'un brin d'ADN chez la levure. Les deux enzymes possèdent une activité $3' \rightarrow 5'$ exonucléase (Morrison *et al.*, 1991). En introduisant artificiellement des mutations sur les brins avancé et tardif, Shcherbakova et Pavlov ont prouvé que les domaines $3' \rightarrow 5'$ exonucléase des deux polymérases participent à la relecture (proofreading) sur des brins opposés d'ADN (Shcherbakova et Pavlov, 1996). Cette étude a été confirmée ultérieurement, toujours pour *S. cerevisiae* (Karthikeyan *et al.*, 2000). Avec le même type d'approche, il a été prouvé récemment que la polymérase ϵ participe à la réplication du brin avancé chez la levure (Pursell *et al.*, 2007). Du moins pour *S. cerevisiae*, le modèle selon lequel la synthèse des brins avancé et tardif serait effectuée par les enzymes ϵ et δ, respectivement, semble à ce jour consensuel.

Chez l'homme, la situation n'est pas encore claire. Il est maintenant bien établi que les deux polymérases interviennent dans la réplication (Fukui *et al.*, 2004). Une étude récente suggère que les polymérases δ et ϵ ont des activités au moins partiellement indépendantes, car elles ne sont que très peu co-localisées pendant la phase S de la division cellulaire (Rytkonen *et al.*, 2006). Cela représente un argument fort contre le modèle qui place les deux enzymes sur les deux brins opposés au moment de la réplication, si le brin avancé et le brin tardif sont synthétisés simultanément dans cette espèce.

Pour *S. cerevisiae*, des estimations de la fidélité de l'insertion de nucléotides ont été réalisées pour les deux types d'enzymes. Les résultats de l'équipe de A. Sugino (*cf.* table 1.1) indiquent que la fidélité de la polymérase ϵ, qui synthétise le brin avancé, est supérieure à celle de la polymérase δ (Shimizu *et al.*,

Référence	Taux d'erreur Pol δ	Taux d'erreur Pol ϵ
Shimizu *et al.* (2002)		$0.47 * 10^{-5}$ $(G \cdot G)$ $0.5 * 10^{-5}$ $(T \cdot G)$ $0.01 * 10^{-5}$ $(A \cdot G)$
Hashimoto *et al.* (2003)	$1.3 * 10^{-4}$ $(G \cdot G)$ $2.62 * 10^{-4}$ $(T \cdot G)$ $0.074 * 10^{-4}$ $(A \cdot G)$	
Shcherbakova *et al.* (2003)		$\leq 2 * 10^{-5}$ (substitution) $5 * 10^{-7}$ (délétion)
Fortune *et al.* (2005)	$\leq 1.3 * 10^{-5}$ (substitution) $3 * 10^{-4}$ (délétion)	

TAB. 1.1 – Estimations expérimentales de la fidélité des polymérases δ et ϵ chez *S. cerevisiae*. Les taux d'insertion sont différents selon le type de mésappariement introduit, précisé entre parenthèses. Pour les études de Shcherbakova *et al.* (2003) et Fortune *et al.* (2005), les estimations ne permettent pas de distinguer les types de mésappariements introduits, mais des estimations des taux de substitution et de délétion sont disponibles.

2002; Hashimoto *et al.*, 2003). Des études indépendantes ont été réalisées par T. Kunkel et ses collaborateurs. Leurs résultats ne permettent pas de tirer une conclusion sur la différence de fidélité entre les polymérases ϵ et δ pour ce qui est de l'insertion erronée de nucléotides (*cf.* table 1.1). Néanmoins, le taux de délétion dans les régions de faible complexité semble être nettement plus élevé pour Pol δ (Shcherbakova *et al.*, 2003; Fortune *et al.*, 2005).

Machinerie de réplication chez les Archées

La machinerie de réplication des Archées présente des homologies avec celle des Eucaryotes. Notamment, les protéines qui reconnaissent les origines de réplication, les complexes de chargement de la polymérase sur l'ADN ("clamp loader complex"), le "sliding clamp" et les ribonucléases qui dégradent les amorces d'ARN présentent de fortes similarités entre Archées et Eucaryotes (Edgell et Doolittle, 1997). Jusqu'à récemment, une seule famille de polymérases réplicatives était connue chez les Archées, et cette famille présente aussi des similarités avec les enzymes Eucaryotes (nommée famille B) (Edgell et Doolittle, 1997). Les travaux de Uemori *et al.* (1997) ont mis en évidence la présence d'une nouvelle famille de polymérases chez l'euryarchée *Pyrococcus furiosus* (nommée famille D). Les polymérases D sont constituées de deux sous-unités, dont l'une présente des simi-

larités de séquences avec une sous-unité de la polymérase δ des Eucaryotes (Cann *et al.*, 1998).

Une étude récente a suggéré que chez les Euryarchées la réplication des deux brins d'ADN pourrait se faire par deux types de polymérases : le brin avancé serait synthétisé par la polymérase de type B, et le brin tardif par la polymérase de type D (Henneke *et al.*, 2005). Ce modèle est encore à confirmer.

La fidélité des polymérases réplicatives chez les Archées semble être généralement élevée. Il existe des estimations expérimentales pour les polymérases de type B ; chez *P. furiosus*, le taux d'incorporation erronée est de seulement $1.6 * 10^{-6}$ (Lundberg *et al.*, 1991). Il en est de même chez une autre euryarchée, *Thermococcus litoralis* (Mattila *et al.*, 1991). Pour l'instant on ne dispose pas d'informations concernant une éventuelle différence de fidélité entre les polymérases de type B et D.

1.1.2 Conséquences de la synthèse discontinue de l'ADN

Toutes les ADN polymérases que l'on connaît jusqu'à présent possèdent une caractéristique commune : elles ne peuvent synthétiser une chaîne d'ADN que dans la direction $5' \rightarrow 3'$. Les deux brins d'ADN sont antiparallèles (l'un d'entre eux est orienté $5' \rightarrow 3'$ et l'autre $3' \rightarrow 5'$). Cependant, la réplication des deux brins a lieu de manière simultanée, par l'avancement d'une même fourche de réplication (Meselson et Stahl, 1958). Cet apparent paradoxe est résolu par le modèle de la réplication discontinue, proposé par Okazaki *et al.* (1967). Selon ce modèle, l'un des deux brins est synthétisé de façon discontinue, sous la forme de courts fragments d'ADN (appelés fragments d'Okazaki). L'élongation de la chaîne d'ADN peut se faire alors dans la direction $3' \rightarrow 5'$ (*cf.* figure 1.3).

Mutations spécifiques de l'ADN simple brin

Sous le modèle de réplication semi-discontinue de l'ADN, le brin avancé est présent plus longtemps à l'état simple-brin. En effet, l'hélicase doit dérouler la double hélice d'ADN d'une distance au moins égale à la taille d'un fragment d'Okazaki, pour permettre la synthèse du fragment suivant. Pendant ce temps, le brin avancé parental, qui sert de matrice pour la synthèse des fragments d'Okazaki, est à l'état non apparié, alors que le brin tardif parental est "couvert" par la machinerie de réplication. Or, on sait maintenant que certaines mutations (dont la désamination de la cytosine, (Frederico *et al.*, 1990)) sont beaucoup plus fréquentes sur l'ADN simple brin que sur l'ADN double brin. Il est donc possible que le mode de réplication semi-discontinue soit à l'origine de processus évolutifs asymétriques sur les deux brins d'ADN (Frank et Lobry, 1999).

Cependant, il ne faut pas perdre de vue que le temps que le brin avancé passerait à l'état non apparié est très court, de l'ordre de la seconde. La longueur moyenne des fragments d'Okazaki chez *E. coli* est d'environ 1 à 2 kilobases, et

la vitesse d'avancement de la polymérase est de 1000 nucléotides par seconde (Marians, 1992; Kelman et O'Donnell, 1995). Les fragments d'Okazaki des eucaryotes sont plus courts, de l'ordre de 200 nucléotides, mais la vitesse d'avancement des polymérases réplicatives est aussi considérablement plus faible que chez les procaryotes - environ 50 à 100 nucléotides par seconde (Johnson et O'Donnell, 2005).

Le temps passé par le brin avancé parental à l'état non-apparié serait donc de seulement quelques secondes. Il n'est pas clair pour l'instant si cette courte durée est suffisante pour que des mutations s'accumulent préférentiellement sur le brin avancé.

Mécanismes de réparation des mésappariements

La discontinuité de la réplication pourrait avoir une autre conséquence sur les taux de mutation des deux brins. Le processus de réparation des mésappariements nécessite la présence de discontinuités, ou "cassures" dans l'ADN (Modrich et Lahue, 1996). Si le brin tardif est le seul à être synthétisé de manière discontinue, il est possible que les mécanismes de réparation soient plus efficaces sur ce brin (Radman, 1998).

Cependant, il n'est sûr que le mode de réplication *in vivo* soit réellement semi-discontinu. Le modèle initial d'Okazaki et ses collaborateurs, soutenu par des données obtenues *in vivo*, proposait que la synthèse des *deux* brins d'ADN se fait de manière discontinue (Okazaki *et al.*, 1968). La plupart des expériences *in vivo* semblent soutenir le mode de réplication discontinu sur les deux brins, alors que les données *in vitro* sont plutôt en faveur du modèle semi-discontinu (Wang, 2005). Si la synthèse du brin avancé se fait aussi de manière discontinue, cela pourrait permettre une efficacité de réparation similaire à celle du brin tardif.

1.1.3 Arrêt de la réplication par des lésions dans l'ADN

Au cours de son cycle de vie, la cellule est soumise à une multitude de facteurs endogènes ou environnementaux qui peuvent provoquer des lésions dans l'ADN. Parmi les lésions à causes endogènes, on peut citer la désamination hydrolytique de la cytosine, l'oxydation de la guanine ou l'alkylation de l'adénine et de la guanine (Barnes et Lindahl, 2004). Les facteurs environnementaux comme les rayonnements UV peuvent également endommager la chaîne d'ADN, en provoquant par exemple la formation de dimères de pyrimidine (Setlow, 1966). Il existe plusieurs mécanismes qui permettent de réparer les lésions de l'ADN, notamment la voie de réparation par excision de bases (*cf.* Seeberg *et al.* (1995) pour revue) et la réparation par excision de nucléotides (*cf.* de Laat *et al.* (1999) pour revue).

La machinerie cellulaire de réplication, qui est très efficace et présente une grande fidélité dans des conditions normales, ne peut pas réaliser la synthèse de l'ADN lorsque les brins parentaux présentent des lésions. Lorsque la fourche de

réplication rencontre une lésion, elle est bloquée, et des polymérases spécifiques pour la réplication de l'ADN endommagé viennent continuer la synthèse (Rattray et Strathern, 2003). Ces polymérases présentent généralement de forts taux d'incorporation erronée des nucléotides, et elles sont donc impliquées dans la mutagénèse (Rattray et Strathern, 2003).

Vu que le mode de réplication de l'ADN est asymétrique, il est pertinent de se demander si les conséquences de la présence d'une lésion sur le brin parental avancé ou brin tardif sont équivalentes.

Plusieurs études récentes suggèrent que la progression de la fourche de réplication n'est pas affectée de la même manière selon que la lésion rencontrée se trouve sur le brin parental avancé ou sur le brin tardif. Une étude *in vitro* a montré que lorsque le complexe enzymatique Pol III d'*E. coli* effectue la synthèse de l'ADN, la fourche de réplication est complètement bloquée par des lésions présentes dans le brin parental tardif (qui sert de matrice pour le brin avancé)(Higuchi *et al.*, 2003). Si par contre la base endommagée est présente sur le brin parental avancé (qui sert de modèle pour la synthèse du brin tardif), la fourche de réplication peut continuer sa progression, non seulement pour la synthèse du brin avancé, mais aussi pour celle des fragments d'Okazaki qui se trouvent en aval de la lésion (Higuchi *et al.*, 2003).

La deuxième conclusion a été confirmée *in vivo* (Pages et Fuchs, 2003) et *in vitro* (McInerney et O'Donnell, 2004), mais pour la première les résultats semblent être contradictoires : selon Pages et Fuchs (2003), la synthèse du brin tardif peut continuer même si celle du brin avancé est arrêtée. La vérité est probablement quelque part au milieu : une étude récente a démontré que lorsque la synthèse du brin avancé est bloquée par une lésion dans l'ADN, la fourche de réplication continue sa progression sur une longueur de 0.5 à 3 milliers de paires de bases, et le brin tardif continue aussi à être synthétisé sur cette longueur (*cf.* figure 1.6, McInerney et O'Donnell (2007)). Une étude antérieure avait d'ailleurs aussi proposé que l'activité de l'hélicase n'est pas affectée par la présence d'une lésion sur le brin avancé (Veaute *et al.*, 2000).

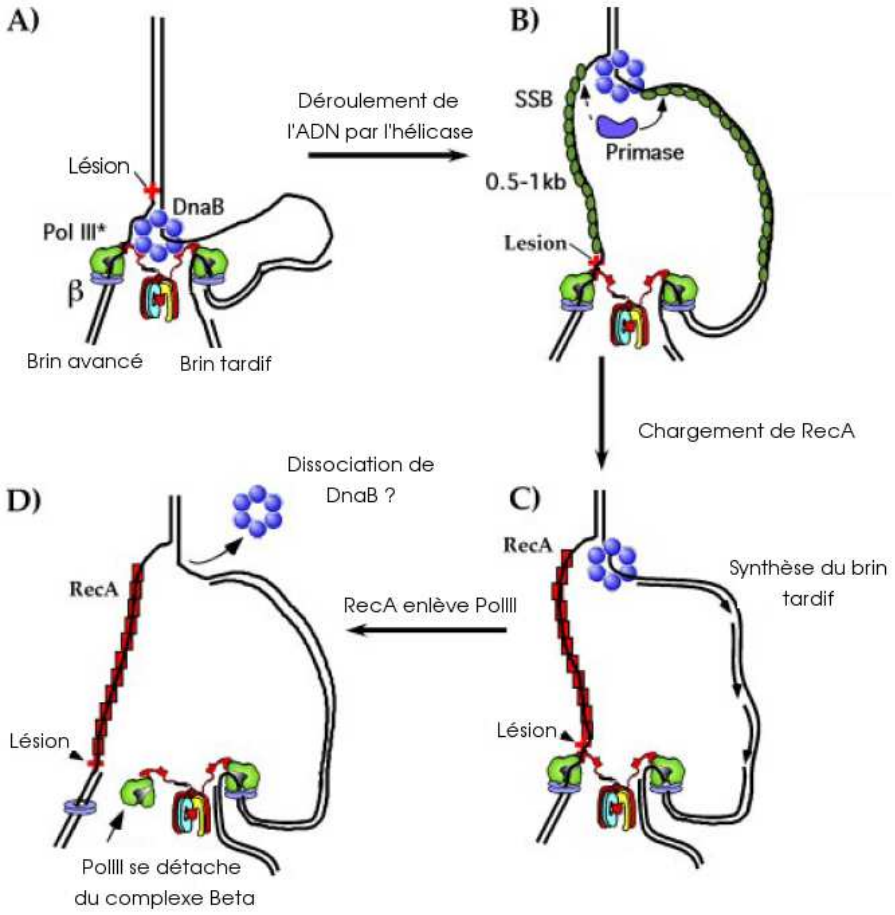

FIG. 1.6 – Schéma décrivant les mécanismes de reprise de la réplication lorsqu'une lésion est rencontrée sur le brin parental tardif. La synthèse du nouveau brin avancé est bloquée, alors que la synthèse du brin tardif peut continuer sur quelques milliers de paires de bases. La protéine RecA est impliquée dans la reprise de la progression de la fourche de réplication. D'après McInerney et O'Donnell (2007).

La reprise de la réplication lorsqu'une lésion de l'ADN est rencontrée sur le brin parental tardif est médiée entre autres par la protéine RecA (McInerney et O'Donnell, 2007). A ce sujet, il est significatif de noter que la perte du gène codant pour RecA dans certaines espèces bactériennes semble avoir comme conséquence une augmentation de l'asymétrie d'évolution des deux brins d'ADN (Klasson et Andersson, 2006).

Ces résultats suggèrent que le fonctionnement des machineries de réplication et de réparation est effectivement asymétrique selon que les lésions interviennent sur les brins avancé ou tardif. Notamment, il est possible que le brin parental tardif (qui sert de matrice pour le brin avancé) passe plus de temps à l'état non-apparié, lorsque des bases endommagées sont rencontrées. Cela pourrait avoir comme conséquence un plus fort taux de mutation sur ce brin (*cf.* discussion plus haut).

1.2 Mutations asymétriques associées à la réplication

Dans cette section, nous passerons en revue les études qui ont mis en évidence l'existence de biais mutationnels asymétriques sur les brins avancé et tardif, en lien direct avec la réplication de l'ADN.

1.2.1 Etudes expérimentales

La découverte du mode asymétrique de réplication de l'ADN a suscité beaucoup d'intérêt pour l'existence d'éventuelles différences dans les taux de mutation caractéristiques des brins avancé et tardif. Les premières analyses réalisées sur des cellules humaines ont indiqué que les taux d'incorporation erronée de nucléotides sont similaires sur le brin avancé et sur le brin tardif (quoique légèrement supérieures sur le brin tardif), dans différentes conditions (Roberts *et al.*, 1991; Basic-Zaninovic *et al.*, 1992).

Chez *E. coli*, les premières études *in vitro* ont indiqué que les taux d'insertion et de délétion sont sensiblement plus importants sur le brin tardif pour la réplication (Veaute et Fuchs, 1993; Rosche *et al.*, 1995; Iwaki *et al.*, 1996). Cependant, il faut noter que ces études ont été réalisées sur des plasmides qui ont un mode de réplication unidirectionnel, alors que le chromosome d'*E. coli* possède une origine de réplication bidirectionnelle. Une analyse *in vivo* a indiqué que les taux de délétion sont similaires pour les deux brins, contrairement à ce qui avait été proposé auparavant (Nagata *et al.*, 2005)

Pour ce qui concerne les changements de nucléotides chez *E. coli*, une étude réalisée par Fijalkowska *et al.* (1998) a montré que le brin avancé pour la réplication accumule plus de mutations que le brin tardif, indépendamment de la direction de la transcription du gène rapporteur utilisé. L'explication proposée par les auteurs est basée sur le fait que la polymérase qui synthétise le brin tardif peut

se dissocier plus facilement de la chaîne d'ADN lorsqu'elle rencontre des mésappariements (ce qui permettrait une réparation plus efficace), alors que l'affinité de la polymérase serait plus forte avec le brin avancé (Fijalkowska *et al.*, 1998). Les mêmes auteurs ont montré par la suite que la tendance est inversée lorsque le système de réparation SOS est actif, c'est à dire que le brin tardif est plus susceptible à la mutagénèse induite par le système SOS (Maliszewska-Tkaczyk *et al.*, 2000).

Wagner *et al.* (1997) ont montré que le taux de mutation (transversions $G \rightarrow T$) produites par l'oxydation de la guanine est similaire pour le brin avancé et pour le brin tardif, toujours chez *E. coli*.

Les études réalisées chez la levure indiquent que le taux de substitutions nucléotidiques est plus élevé sur le brin avancé pour la réplication (Pavlov *et al.*, 2002, 2003). Selon les auteurs, l'explication la plus vraisemblable pour cette asymétrie est une plus grande efficacité de la réparation des mésappariements sur le brin tardif, liée probablement à discontinuité de la synthèse de ce brin. Cette conclusion a été confirmée par une étude récente, qui montre que la réparation des mésappariements par MutSα est plus efficace sur le brin tardif (Kow *et al.*, 2007).

1.2.2 Etudes *in silico*

La première étude bioinformatique qui a proposé que la réplication est la cause de biais mutationnels asymétriques sur les deux brins d'ADN est celle de Wu et Maeda (1987). En analysant un alignement de séquences de primates, Wu et Maeda ont mis en évidence la présence d'une asymétrie significative dans les taux de substitution des nucléotides. Leurs résultats ont été contredits ultérieurement (Bulmer, 1991; Wu, 1991). L'article de Wu et Maeda (1987) reste toutefois fondamental pour l'étude de l'asymétrie de composition, car il décrit pour la première fois le modèle d'évolution symétrique des séquences d'ADN et il met en relation le rejet de ce modèle avec le mécanisme de réplication.

La disponibilité des séquences génomiques complètes a entraîné la publication de nombreuses études qui signalent l'existence de patrons de mutations asymétriques associés à la réplication. Cependant, la plupart de ces études se sont limitées à observer la présence de déviations par rapport aux règles de parité $[A] = [T]$ et $[G] = [C]$ dans les séquences génomiques, ce qui permet en effet de conclure sur l'existence de biais mutationnels asymétriques, mais qui ne permet pas de les identifier. Seulement quelques études ont fourni des estimations des patrons de mutations asymétriques (résumées dans le tableau 1.2).

L'asymétrie de composition associée à la réplication est quasi-universelle - presque toutes les espèces étudiées jusqu'à présent possèdent plus de G que de C sur le brin avancé. On pourrait donc s'attendre à ce que les causes de cette asymétrie soient aussi universelles. Vue la variabilité des machineries de réplication

dans les différents génomes, il est difficile d'imaginer que ce biais mutationnel universel soit provoqué par leurs caractéristiques. Jusqu'à récemment, l'explication la plus fréquemment proposée pour l'existence de l'asymétrie de composition était le mécanisme de la désamination de la cytosine sur l'ADN simple brin. Cette hypothèse a l'avantage d'être applicable à toutes les espèces qui ont un mode de réplication semi-discontinue, et elle devrait effectivement se traduire par une asymétrie $C \to T > G \to A$ sur le brin avancé. Les études réalisées jusqu'à 2005 semblaient confirmer cette hypothèse (*cf.* table 1.2). Cette théorie a été remise en question par Rocha *et al.* (2006), qui démontrent que les biais mutationnels responsables de l'asymétrie de composition sont beaucoup plus variables que ce qu'on avait imaginé auparavant (*cf.* table 1.2).

Il faut également mentionner l'étude de Klasson et Andersson (2006), qui a mis en évidence une association entre la perte de certains gènes impliqués dans la recombinaison et le re-démarrage des fourches de réplication bloquées, et l'augmentation de l'intensité de l'asymétrie de composition chez *Buchnera*.

Référence	Espèce étudiée	Mutations asymétriques
(Tanaka et Ozawa, 1994)	Mitochondries humaines	$C \to T > G \to A$ $A \to G > T \to C$
(Mitchell et Graur, 2005)	*Mycobacterium leprae*	$C \to T > G \to A$ $A \to G > T \to C$
(Klasson et Andersson, 2006)	*Buchnera aphidicola* (Bp)	$C \to T > G \to A$
(Rocha *et al.*, 2006)	*Bacillus anthracis*	$A \to C < T \to G$ $A \to G > T \to C$ $A \to T < T \to A$
	Bordetella	$C \to T > G \to A$
	Escherichia	$C \to T > G \to A$
	Neisseria	$C \to A > G \to T$ $C \to G > G \to C$ $C \to T > G \to A$
	Rickettsia	$C \to A < G \to T$
	Staphylococcus	$A \to G > T \to C$
	Streptococcus	$C \to T > G \to A$

TAB. 1.2 – Mutations asymétriques associées à la réplication. Pour les espèces bactériennes, le brin de référence est le brin avancé et pour la mitochondrie le brin H.

1.3 Asymétrie du mécanisme de transcription

Selon le "dogme central" de la biologie moléculaire, l'information génétique est portée par l'ADN, qui est ensuite transcrit pour donner un ARN messager, qui sera traduit pour donner une protéine. Dans la plupart des organismes, la transcription produit des molécules d'ARN messager à l'état simple brin, à partir des gènes codés par de l'ADN double brin. Par définition, ce mécanisme agit de manière asymétrique sur les deux brins d'ADN.

1.3.1 Fonctionnement de la transcription

Le fonctionnement de la transcription est similaire chez les procaryotes et chez les eucaryotes. La transcription est catalysée par un complexe enzymatique appelé ARN polymérase (*cf.* figure 1.7). Ce complexe se fixe à l'ADN au niveau du promoteur, localisé en $5'$ du gène. Pour que l'élongation de la transcription ait lieu, l'ARN polymérase déclenche l'ouverture de la double hélice d'ADN en aval du promoteur. L'ARN polymérase synthétise ensuite une molécule d'ARN simple brin en utilisant l'un des deux brins d'ADN comme matrice ; l'ARNm résultant sera identique (mis à part le remplacement de la thymine par l'uracile) au brin d'ADN complémentaire (que l'on appelle alors brin codant).

Les procaryotes possèdent une seule ARN polymérase, constituée de 5 sous-unités : α (présente en deux copies et responsable pour l'assemblage du complexe enzymatique), β (qui assure l'initiation et l'élongation de la chaîne d'ARNm), β' (qui se lie à la matrice d'ADN), σ (qui assure la fixation au promoteur) et ω (qui a un rôle dans la maintenance de la conformation de la sous-unité β') (Burgess, 1971; Yura et Ishihama, 1979; Mathew et Chatterji, 2006).

Chez les eucaryotes il existe trois classes d'ARN polymérases (Chambon, 1975), qui ont des rôles différents. L'ARN polymérase I synthétise les précurseurs des ARN ribosomiques, l'ARN polymérase II assure la transcription des gènes protéiques en ARNm, alors que l'ARN polymérase III synthétise les gènes d'ARN de transfert et les ARN ribosomiques 5S (Young, 1991). Chacun de ces trois complexes enzymatiques est constitué de 8 à 14 enzymes (Young, 1991).

1.3.2 Mutations spécifiques de l'ADN simple brin

Pour que la synthèse de la molécule d'ARNm puisse avoir lieu, la structure de la double hélice doit être temporairement dénaturée, par l'action de l'ARN polymérase. Il se forme ainsi une "bulle de transcription" (transcription bubble), qui permet l'avancement de la polymérase le long du brin matrice, laissant le brin codant non-apparié (Wang *et al.*, 1977). Comme mentionné précédemment, certaines mutations sont beaucoup plus fréquentes sur l'ADN simple brin par rapport à l'ADN double brin. Il est donc possible que le brin codant accumule

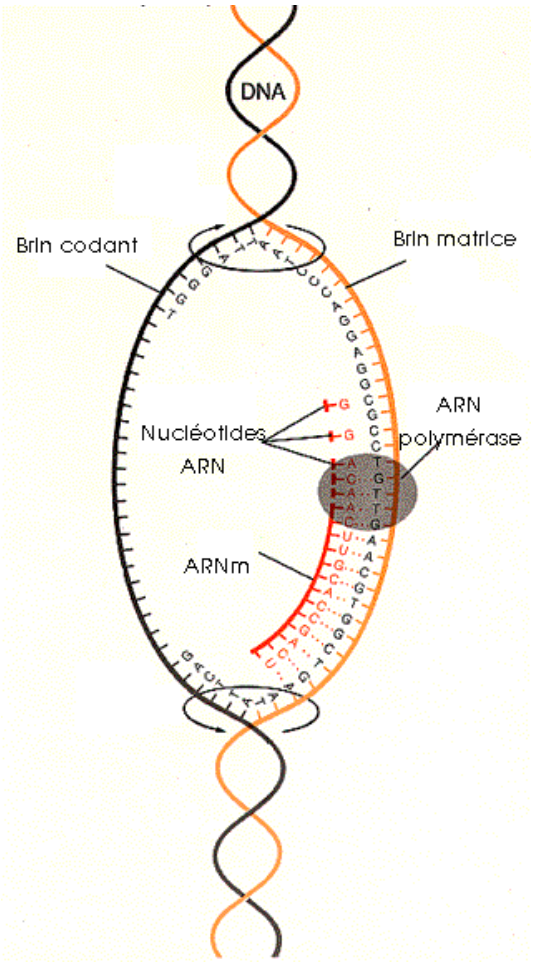

FIG. 1.7 – Schéma du fonctionnement de la transcription.

plus de mutations que le brin matrice, car pendant la transcription il se trouve essentiellement à l'état simple brin. L'intensité de cet éventuel biais mutationnel dépend de deux paramètres qui déterminent le temps passé par le brin codant à l'état non-apparié : la taille de la bulle de transcription et la vitesse d'avancement de la polymérase.

Les premières études sur l'initiation de la transcription chez *E. coli* ont montré que l'ARN polymérase perturbe la double hélice d'ADN sur un fragment d'environ 11 paires de bases, au niveau du promoteur (Siebenlist, 1979). La taille de la "bulle" varie entre 11 et 18 paires de bases au cours du déroulement de la transcription (Zaychikov *et al.*, 1995, 1997).

Les ARN polymérases synthétisent l'ARNm à une vitesse de plusieurs dizaines de nucléotides par seconde (30-50 nucléotides chez *E. coli* (Gotta *et al.*, 1991)). Nous pouvons en déduire que le temps passé par le brin codant à l'état non-apparié est relativement court, en moyenne inférieur à une seconde. Il n'est pas

clair pour l'instant si cette courte durée est suffisante pour que des mutations s'accumulent préférentiellement sur le brin codant.

1.3.3 Mécanismes de réparation associés à la transcription

L'existence de mécanismes de réparation préférentiellement associés à la transcription est maintenant bien établie. Des études réalisées sur des cellules eucaryotes avaient mis en évidence que les dimères de pyrimidine, qui sont connus pour bloquer la transcription, sont excisés plus rapidement dans une région codant pour un gène activement transcrit, par rapport aux régions non-codantes avoisinantes (Bohr *et al.*, 1985; Mellon *et al.*, 1986). Il a été montré par la suite que la réparation est beaucoup plus fréquente sur le brin matrice que sur le brin codant, ce qui est cohérent avec la nécessité de supprimer les lésions qui pourraient bloquer l'avancement de l'ARN polymérase (Mellon *et al.*, 1987). Le même mécanisme de couplage entre transcription et réparation existe chez *Escherichia coli*. Comme chez les eucaryotes, le brin matrice est préférentiellement réparé, et cela au moment même de la transcription du gène (Mellon et Hanawalt, 1989). L'association entre transcription et réparation est aussi valable pour la levure *S. cerevisiae* (Leadon et Lawrence, 1991).

Le mécanisme qui pourrait expliquer cette réparation préférentielle du brin matrice est la signalisation de la machinerie de réparation par la présence d'ARN polymérase bloquée par les lésions (Christians et Hanawalt, 1993). Cette proposition a été confirmée par l'étude de Selby et Sancar (1993), qui a démontré que le gène *mfd*, impliqué dans la réparation des lésions présentes sur le brin matrice chez *E. coli*, code pour une protéine similaire à une hélicase et peut intéragir directement avec l'ARN polymérase.

Le couplage entre transcription et réparation chez l'homme semble être valable seulement pour les gènes transcrits par l'ARN polymérase II. Ainsi, les gènes d'ARN ribosomiques, qui sont transcrits par l'ARN polymérase I, présentent un niveau de réparation des lésions qui est similaire aux régions non-actives (Christians et Hanawalt, 1993). La même absence de réparation préférentielle est valable pour les gènes d'ARNt de l'homme, qui sont transcrits par l'ARN polymérase I (Dammann et Pfeifer, 1997). Il semblerait par contre que chez la levure les gènes transcrits par l'ARN polymérase I sont sujets à réparation préférentielle (Conconi *et al.*, 2002).

1.4 Mutations asymétriques associées à la transcription

1.4.1 Etudes expérimentales

L'existence d'un mécanisme de réparation dirigé préférentiellement vers le brin matrice pour la transcription devrait avoir comme conséquence une asymé-

trie des taux de mutations par rapport aux deux brins d'ADN. Cette prédiction a été confirmée expérimentalement. Une étude *in vivo* réalisée chez *Escherichia coli* a démontré que le taux de mutation est en moyenne ≈ 14 fois plus élevé sur le brin codant par rapport au brin matrice (Oller *et al.*, 1992). Pour les changements $G \cdot C \rightarrow A \cdot T$, le taux de mutation est même jusqu'à 300 plus élevé sur le brin codant (Oller *et al.*, 1992). La relation cause-effet entre mécanisme de réparation et taux de mutation asymétrique a été validée, car lorsque le gène *mfd* (qui est responsable du couplage entre réparation et transcription) est inactivé, l'asymétrie disparaît (Oller *et al.*, 1992).

Le mécanisme de réparation couplée à la transcription n'est pas le seul en cause pour l'existence de mutations asymétriques. Chez *Escherichia coli*, la fréquence des mutations $C \rightarrow T$ est plus élevée sur le brin codant que sur le brin matrice, et cette fréquence augmente avec le niveau de transcription du gène (Beletskii et Bhagwat, 1996). Il ne peut pas s'agir là d'un mécanisme de réparation, car cela ne peut pas expliquer *l'augmentation* du taux de mutation avec le niveau d'expression. L'explication proposée par Beletskii et Bhagwat (1996) est en faveur d'une plus grande mutabilité de la cytosine sur le brin codant, due au fait que ce brin se trouve à l'état non-apparié pendant plus de temps que le brin matrice pendant la transcription. L'asymétrie des taux de mutations de $C \rightarrow T$ sur les brins codant et matrice a été confirmée par une étude indépendante (Bockrath et Li, 1998). Il a été également démontré que cette asymétrie des taux de mutation n'est pas restreinte à un seul gène, et qu'il s'agit au contraire d'un mécanisme valable pour tous les transcrits, du moins chez *E. coli* (Beletskii et Bhagwat, 2001).

Le spectre des mutations asymétriques par rapport aux deux brins n'est pas restreint aux transitions $C \rightarrow T$. Le taux de transversions de $G \rightarrow T$ est aussi plus élevé sur le brin codant chez *Escherichia coli* (Klapacz et Bhagwat, 2005). Une hypothèse qui pourrait expliquer cette asymétrie est une plus grande susceptibilité du brin codant aux lésions oxydatives de l'ADN, liée au fait que ce brin passe plus de temps à l'état non-apparié (Klapacz et Bhagwat, 2005). De manière générale, la présence de la transcription a comme effet une augmentation du taux de mutation (Klapacz et Bhagwat, 2002).

1.4.2 Etudes *in silico*

De nombreuses études bioinformatiques ont mis en évidence l'existence d'un patron d'évolution asymétrique par rapport aux deux brins d'ADN dans les séquences transcrites. La plupart de ces études se sont par contre limitées à démontrer que l'asymétrie de composition en bases observée sur les séquences d'ADN est associée à la transcription. Il existe seulement peu d'analyses qui ont effecti-

vement élucidé la nature du patron de substitution asymétrique qui engendre ce biais de composition.

La première étude *in silico* qui a amené des preuves en faveur de l'existence d'un patron de mutation asymétrique associé à la transcription est celle de Francino *et al.* (1996). En réalisant une étude comparative d'une dizaine de gènes séquencés dans plusieurs souches d'*Escherichia coli*, les auteurs ont démontré que le taux de substitution de $C \rightarrow T$ est plus élevé que celui de $G \rightarrow A$, lorsque les substitutions sont comptées sur le brin codant des gènes. Cette asymétrie est rencontrée aussi bien pour des gènes codés sur le brin avancé que pour des gènes codés sur le brin tardif pour la réplication. Il est donc possible d'exclure la possibilité que le mode de réplication des gènes soit un facteur confondant (Francino *et al.*, 1996).

Chez l'homme, les régions transcrites évoluent aussi de manière asymétrique. Par contre, il semblerait que la mutation qui présente le plus fort degré d'asymétrie n'est pas la transition de $C \rightarrow T$ comme pour *E. coli*, mais celle dans le sens contraire : il y a plus de mutations de $T \rightarrow C$ que de $A \rightarrow G$ sur le brin matrice (Green *et al.*, 2003). La transition $C \rightarrow T$ est elle aussi asymétrique, mais dans une moindre mesure.

Il est important de remarquer que chez l'homme l'intensité de l'asymétrie semble être dépendante du contexte : la nature des bases vosines en $5'$ et $3'$ semble influencer la différence entre les taux des mutations complémentaires (Hwang et Green, 2004).

1.5 Conclusion

Dans ce chapitre, nous avons essayé de résumer les connaissances actuelles en ce qui concerne l'asymétrie des processus de réplication et de transcription, et leur conséquence sur les processus évolutifs. Ces deux mécanismes ont été analysés séparément. Nous verrons dans le chapitre suivant que ces processus agissent en réalité de concert sur les séquences d'ADN, et nous présenterons par la suite quelques méthodes qui permettent de découpler les effets des deux types de mécanismes.

Chapitre 2

Séparation des bias de réplication et de transcription dans les génomes procaryotes

La réplication de l'ADN impose de fortes contraintes sur l'organisation des génomes procaryotes. D'une manière générale, les chromosomes procaryotes possèdent une unique origine de réplication bidirectionnelle et un terminus. La plupart des chromosomes procaryotes sont circulaires ; l'origine de réplication et le terminus sont en général localisés à des positions (approximativement) diamétralement opposées. A l'intérieur de chacun des segments délimités par l'origine et le terminus de réplication, les deux brins d'ADN sont asymétriques, non seulement pour ce qui concerne la composition en nucléotides, mais aussi pour le contenu en gènes.

L'asymétrie du contenu en gènes est très bien illustrée par l'organisation du chromosome de *Bacillus subtilis* (*cf.* figure 2.1). Comme la plupart des chromosomes procaryotes, celui de *B. subtilis* est constitué en proportion de plus de 80 % par des gènes. Sur le segment délimité en $5'$ par l'origine de réplication et en $3'$ par le terminus, la grande majorité des gènes sont transcrits dans le sens des aiguilles d'une montre. Sur le segment délimité en $5'$ par le terminus et en $3'$ par l'origine, les gènes sont majoritairement transcrits dans la direction opposée. La conséquence de ce biais de localisation des gènes est que la réplication et la transcription sont majoritairement co-orientées. Le brin codant des gènes correspond au brin parental avancé (qui sert de matrice pour le brin tardif lors de la réplication de l'ADN).

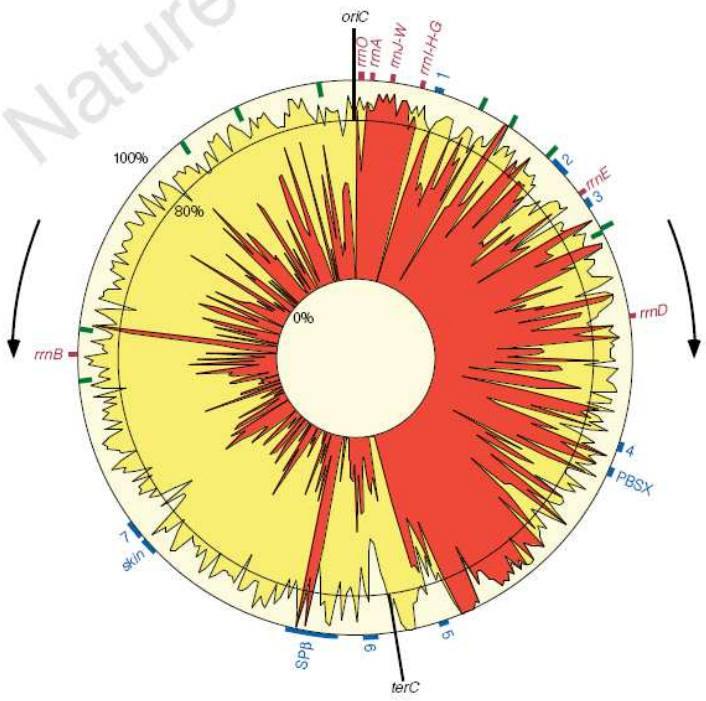

FIG. 2.1 – Schéma décrivant l'organisation du chromosome de *Bacillus subtilis*. Les positions de l'origine (*oriC*) et du terminus de réplication (*terC*) sont signalées. Les flèches noires représentent la direction d'avancement de la fourche de réplication. La surface jaune représente la densité de nucléotides codants pour des gènes (quelle que soit l'orientation du gène). La surface rouge représente la densité de nucléotides appartenant à des gènes transcrits dans le sens du mouvement des aiguilles d'une horloge. D'après Kunst *et al.* (1997).

L'influence du mécanisme de réplication sur l'organisation des chromosomes procaryotes a été mise en évidence bien avant que des séquences génomiques complètes soient disponibles. Ainsi, une analyse du chromosome d'*Escherichia coli* a montré que les gènes codant pour les protéines ribosomiques et pour les ARNr sont localisés sur le brin avancé pour la réplication (Nomura et Morgan, 1977). Cela revient à dire que la transcription des gènes a lieu dans la même direction que la réplication de l'ADN. Ce résultat a été confirmé et renforcé plus tard, par l'observation que la co-orientation entre réplication et transcription n'est pas restreinte aux gènes ribosomiques, mais s'étend aussi à d'autres classes de gènes protéiques chez *E. coli* (Brewer, 1988).

2.1 Variabilité taxonomique de la co-orientation entre réplication et transcription

L'analyse des premières séquences génomiques complètes a montré que l'intensité du biais d'orientation des gènes varie largement entre les espèces : par exemple, chez *Synechocystis sp.* et *E. coli*, la proportion des gènes codés sur le brin avancé atteint seulement 50 % et 54 %, respectivement, alors que chez *Bacillus subtilis* et *Mycoplasma genitalium* cette proportion atteint 74 % (Blattner *et al.*, 1997; McLean *et al.*, 1998).

Nous avons résumé dans la figure 2.2 la variabilité taxonomique de la co-orientation entre réplication et transcription. La famille des Firmicutes (des bactéries Gram positives à faible taux de GC) est celle qui présente les plus forts biais d'orientation des gènes. L'orientation des gènes n'est pas réellement biaisée chez la plupart des Archées et des Cyanobactéries. Le plus faible pourcentage de gènes codés sur le brin avancé (seulement 47 %) est rencontré chez *Lawsonia intracellularis*, une bactérie parasite intracellulaire appartenant à la famille des Protéobactéries.

Les bactéries de la famille des Firmicutes, pour lesquelles la tendance à la co-orientation entre réplication et transcription est la plus forte, ont une autre caractéristique particulière : l'ADN polymérase de ces organismes possède deux sous-unités α différentes, utilisées pour répliquer le brin avancé et le brin tardif. C'est du moins ce qui a été démontré expérimentalement pour *B. subtilis*, par Dervyn *et al.* (2001), ainsi que pour *Staphylococcus aureus* (Inoue *et al.*, 2001).

L'analyse des autres espèces de Firmicutes montre que les gènes codant pour les deux sous-unités α (*polC* et *dnaE*) sont présents dans leurs génomes, mais leur utilisation à la fourche de réplication reste pour l'instant hypothétique. Le lien entre l'asymétrie de la fourche de réplication et la distribution des gènes sur les deux brins semble être trop important pour être une coïncidence (Rocha, 2002).

La co-orientation entre réplication et transcription n'est pas spécifique des procaryotes. L'identification *in silico* des origines de réplication dans le génome humain a permis de mettre en évidence que les gènes protéiques sont aussi préférentiellement codés sur le brin avancé (Huvet *et al.*, 2007).

2.2 Hypothèse de la collision des polymérases

L'observation que les gènes ribosomiques se trouvent sur le brin avancé chez *E. coli* a conduit à la suggestion que cette organisation pourrait être favorable pour éviter les collisions frontales entre la machinerie de réplication et l'ARN polymérase (Nomura et Morgan, 1977). La co-orientation entre réplication et transcription n'est pas restreinte aux gènes ribosomiques, et il a été suggéré très tôt que l'évitement des collisions frontales entre les deux machineries cellulaires

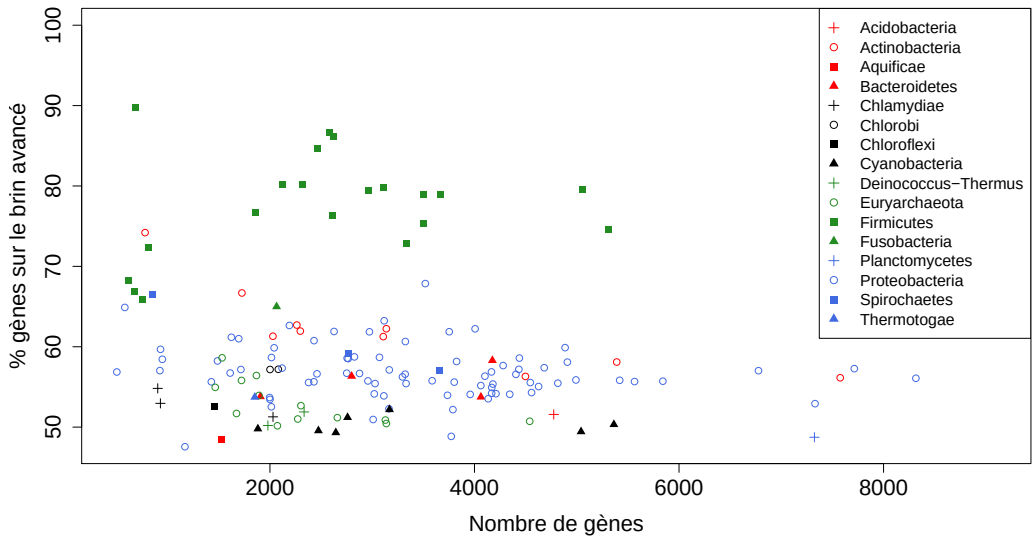

FIG. 2.2 – Co-orientation entre réplication et transcription pour 151 espèces dont le génome est complètement séquencé. Pour éviter la redondance, chaque genre taxonomique est représenté par un seul point. Nous avons représenté uniquement les espèces pour lesquelles la position de l'origine de réplication est connue ou peut être prédite *in silico*. Ce graphique représente une mise à jour de la figure présentée par Rocha (2002).

pourrait influencer de manière considérable l'organisation des génomes bactériens (Brewer, 1988).

Lorsque la cellule bactérienne est en phase de croissance exponentielle, plusieurs cycles de réplication démarrent en même temps, et la transcription de nombreux gènes se fait simultanément. Les complexes enzymatiques responsables de la réplication et de la transcription parcourent le chromosome en même temps (*cf.* figure 2.3), mais à des vitesses très différentes (environ 1000 nucléotides par seconde pour l'ADN polymérase et seulement 50-100 pour l'ARN polymérase). Les rencontres entre les deux complexes semblent donc inévitables. Pour les gènes codés sur le brin parental avancé, la collision peut réellement se faire entre les deux polymérases, car elles utilisent le même brin d'ADN comme matrice et elles avancent dans la même direction. Pour les gènes codés sur le brin parental tardif, le terme "collision des polymérases" doit être compris au sens large, car la polymérase qui synthétise les fragments d'Okazaki s'éloigne de l'ARN polymérase. Il peut par contre y avoir collision entre l'ARN polymérase et d'autres protéines présentes à la fourche de réplication, comme l'hélicase qui déroule le duplex d'ADN

(Brewer, 1988).

En théorie, les collisions entre polymérases pourraient avoir plusieurs conséquences délétères pour l'organisme. D'une part, la progression de la machinerie de réplication pourrait être ralentie ou même interrompue, ce qui empêcherait la division cellulaire, si la fourche de réplication n'est pas débloquée par des mécanismes de réparation spécifiques. D'autre part, la transcription des gènes impliqués dans la collision pourrait aussi être affectée, ce qui implique que la cellule sera incapable de synthétiser les protéines codées par ces gènes. De plus, vu que chez les procaryotes la traduction est couplée à la transcription, l'interruption de la transcription pourrait aussi donner naissance à des peptides non-fonctionnels, ce qui peut également avoir des effets délèteres sur la cellule (Rocha et Danchin, 2003a,b). Mais qu'en est-il en pratique ? Il existe relativement peu d'études expérimentales concernant les effets des collisions des machineries de réplication et de transcription, et les résultats sont souvent contradictoires. La plupart des études ont été réalisées *in vitro*, en utilisant la machinerie de réplication des bactériophages T4 ou Φ29, mais il existe également des expériences *in vivo*, pour *Escherichia coli*, *Bacillus subtilis* et *Saccharomyces cerevisiae*.

Effets des collisions sur le déroulement de la réplication

Une étude *in vitro* utilisant la machinerie de réplication du phage T4 a démontré qu'en absence de l'hélicase, la fourche de réplication est effectivement bloquée par la "bulle" de transcription (constituée par l'ARN polymérase d'*E. coli*) orientée dans la direction opposée, pendant plusieurs minutes. La présence de l'hélicase (ce qui devrait correspondre à la situation *in vivo*) réduit la durée de l'arrêt de la fourche de réplication à quelques secondes (Liu et Alberts, 1995). Cette pause de la fourche de réplication est néanmoins plus longue que celle observée lors des collisions co-orientées (Liu *et al.*, 1993; Liu et Alberts, 1995).

Pour *E. coli*, cette dernière conclusion a été confirmée récemment *in vivo*. L'élongation de la réplication n'est pas affectée par l'avancement de l'ARN polymérase dans la même direction, l'orientation opposée conduit à l'arrêt temporaire de la fourche de réplication (Mirkin et Mirkin, 2005). Nous disposons également d'informations qui soutiennent ces résultats chez *Bacillus subtilis*. Ainsi, une étude *in vivo* réalisée chez *Bacillus subtilis* a confirmé que la progression de la fourche de réplication est considérablement ralentie lorsque la transcription et la réplication avancent dans des directions opposées (Wang *et al.*, 2007).

Pour le bactériophage Φ29, la fourche de réplication s'arrête lorsqu'elle rencontre une ARN polymérase co-orientée bloquée sur son substrat, mais lorsque l'ARN polymérase reprend son mouvement, l'ADN polymérase peut elle aussi continuer sa progression. La vitesse de l'ADN polymérase est diminuée par des collisions co-orientées avec la machinerie de transcription (Elias-Arnanz et Sa-

las, 1997). Des résultats similaires ont été obtenus pour les collisions frontales : la machinerie de réplication ne peut pas avancer lorsqu'elle rencontre une ARN polymérase fixée à l'ADN, mais la réplication de l'ADN continue à une vitesse normale lorsque l'ARN polymérase est en mouvement (Elias-Arnanz et Salas, 1999). Il est important de remarquer que la réplication du phage Φ29 diffère de celle de *E. coli* par le fait qu'il n'y a pas de brin tardif : les deux brins d'ADN sont synthétisés de manière continue par deux polymérases avançant dans des directions convergentes (Elias-Arnanz et Salas, 1999).

Chez la levure *S. cerevisiae* des régions où la fourche de réplication s'arrête temporairement ont été observées *in vivo*. Un de ces sites de pause se trouve dans une région contenant des gènes d'ARNr, à l'extrémité 3′ d'un gène d'ARNr précurseur 35S. Cette observation a conduit à proposer que la transcription peut influencer le mouvement de la fourche de réplication, en créant une "barrière" en 3′ du transcrit, lorsque la réplication a lieu dans la direction opposée (Brewer et Fangman, 1988). Cependant, des études ultérieures ont démontré que l'arrêt de la fourche de réplication au niveau de l'extremité 3′ des gènes d'ARNr se fait même lorsque la transcription est inactivée, ce qui indique qu'il s'agit d'une barrière physique et non pas d'une conséquence de la collision des polymérases (Brewer *et al.*, 1992). Il semblerait qu'il existe un lien entre l'existence de cette barrière et le mécanisme de recombinaison, car au moins une protéine (Fob1) est impliquée aussi bien dans l'arrêt de la fourche de réplication et dans la mise en place de points chauds de recombinaison (Kobayashi et Horiuchi, 1996).

D'autres sites d'arrêt de la réplication pourraient par contre être la conséquence directe des collisions des polymérases. En effet, il a été démontré *in vivo* que la progression de la fourche de réplication est effectivement arrêtée, pendant quelques secondes, par la transcription de gènes d'ARN de transfert, uniquement quand la transcription se fait dans la direction opposée de la réplication (Deshpande et Newlon, 1996).

Effets des collisions sur le déroulement de la transcription

Une étude réalisée *in vivo* chez *Escherichia coli* a montré que la transcription des gènes d'ARNr est interrompue lors de la rencontre avec la fourche de réplication, quelle que soit son orientation (French, 1992). Du point de vue de la réplication, l'effet dépend de l'orientation. Lorsque les deux processus sont co-orientés, la fourche de réplication avance à une vitesse similaire à celle des régions non-transcrites, alors que dans la situation opposée son mouvement est considérablement ralenti (French, 1992).

Des études *in vitro* effectuées avec la machinerie de réplication du bactériophage T4 et l'ARN polymérase d'*E. coli* ont suggéré par contre que la transcription peut continuer son avancement lors des collisions entre polymérases, quelle

que soit l'orientation des deux machineries (Liu *et al.*, 1993; Liu et Alberts, 1995).

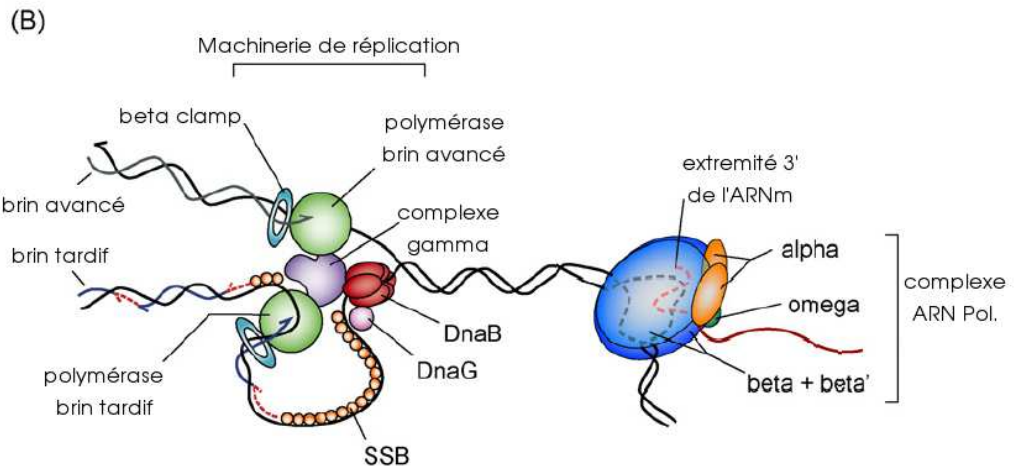

FIG. 2.3 – Représentation des machineries de réplication et transcription sur un chromosome bactérien. D'après Rudolph *et al.* (2007).

L'idée que l'orientation des collisions n'a d'importance que pour le déroulement de la réplication semble être contredite par l'organisation des génomes bactériens. De manière générale, si les collisions frontales sont délétères pour la cellule, on s'attend à ce que les gènes à haut niveau d'expression soient plus souvent codés sur le brin avancé, car un haut niveau de transcription implique aussi beaucoup plus de collisions entre polymérases. Si les deux orientations possibles des collisions ont le même effet négatif pour la transcription, les gènes qui sont essentiels pour l'organisme ne devraient pas présenter un biais d'orientation particulier. Si par contre les collisions frontales aboutissent plus fréquemment à l'interruption de la transcription que les collisions co-orientées, on s'attend à ce que les gènes essentiels soient plus souvent codés sur le brin avancé. C'est effectivement ce que l'on observe dans les génomes bactériens : les gènes essentiels sont préférentiellement codés sur le brin avancé, et le biais d'orientation est plus fort que pour les gènes non-essentiels à fort niveau d'expression (Rocha et Danchin, 2003a,b).

Effets des collisions sur la topologie des chromosomes

Les collisions frontales entre les machineries de réplication et transcription pourraient avoir des effets délétères non seulement à cause de l'interruption de ces deux processus cellulaires essentiels, mais aussi à cause des modifications qu'elles pourraient apporter à la topologie des chromosomes. Ainsi, il a été démontré que

l'orientation convergente de la réplication et de la transcription sur un plasmide d'*E. coli* a comme conséquence la formation d'un "nœud" dans la chaîne d'ADN (Olavarrieta *et al.*, 2002).

2.3 Co-localisation entre origines de réplication et promoteurs

Le besoin d'éviter les collisions frontales entre les machineries de réplication et transcription semble être l'explication biologique la plus probable pour l'orientation préférentielle des gènes sur le brin avancé. Cette explication pourrait cependant être complétée par une autre observation biologique : la co-localisation entre origines de réplication et promoteurs pour la transcription.

Les chromosomes procaryotes ne possèdent en général qu'une seule origine de réplication (seules quelques exceptions sont connues, dans des espèces d'Archées). Cette organisation ne pourrait pas convenir aux génomes eucaryotes, dont les tailles sont beaucoup plus importantes, étant donné de plus que la vitesse d'avancement de l'ADN polymérase eucaryote est beaucoup plus faible que celle des enzymes procaryotes. Il existe donc plusieurs origines de réplication. Les origines de réplication ne sont pas toutes actives en même temps, ni à la même fréquence (DePamphilis, 1999). Une manière optimale d'éviter les collisions entre les machineries de réplication et de transcription serait la régulation temporelle de l'activation des origines (ou bien la régulation temporelle de la transcription), pour que les deux processus n'aient pas lieu en même temps. C'est le contraire qui est observé en pratique : la réplication et la transcription de nombreux gènes semblent être coordonnées chez l'homme (Hassan *et al.*, 1994) et chez la drosophile (MacAlpine *et al.*, 2004). Cette corrélation temporelle entre les deux processus peut être expliquée par l'observation que la fixation des facteurs de transcription à l'ADN au niveau des origines peut déclencher la réplication (DePamphilis, 1993).

La co-localisation des origines de réplication et des régions promotrices pour transcription a été découverte pour la première fois dans les génomes de mitochondries et de virus (Baldacci et Bernardi, 1982; Clayton, 1991; Guo et DePamphilis, 1992). Chez l'homme, plusieurs études ont démontré que les origines de réplication coïncident souvent avec les promoteurs des gènes "de ménage" (Delgado *et al.*, 1998; Huvet *et al.*, 2007).

La coïncidence entre origines de réplication et promoteurs peut contribuer à expliquer la co-orientation entre les deux mécanismes. Les promoteurs se trouvent (généralement) dans la région en $5'$ des gènes. La fourche de réplication qui démarre à partir du promoteur sera donc *a fortiori* co-orientée avec la transcription

du gène en question. Imaginons maintenant qu'un promoteur bi-directionnel (situé entre deux gènes à orientations divergentes) soit recruté pour en faire une origine de réplication, elle aussi bi-directionnelle. Pour les deux gènes voisins de l'origine, la transcription sera nécessairement co-orientée avec la réplication. Au voisinage immédiat de l'origine de réplication, les gènes seront donc codés sur le brin parental avancé. Ce phénomène est évidemment très local. Cependant, il pourrait se propager aux gènes avoisinants, si leur orientation est contrainte par exemple pour éviter des collisions entre "bulles" de transcription convergentes (Prescott et Proudfoot, 2002).

Il faut toutefois remarquer que cette explication reste pour l'instant spéculative, et ne peut sans doute pas expliquer toute l'étendue des biais observés pour l'orientation des gènes.

2.4 Superposition des deux sources d'asymétrie de composition

Quelles qu'en soient les raisons biologiques, la co-orientation entre réplication et transcription a des conséquence importantes pour l'analyse bioinformatique des génomes. Nous avons vu précédemment (*cf.* chapitre 1) que ces deux mécanismes peuvent générer localement des compositions en nucléotides qui ne respectent pas les règles de parité $[A] = [T]$ et $[G] = [C]$. Pour les chromosomes procaryotes on observe généralement que le segment délimité en $5'$ par l'origine et en $3'$ par le terminus contient plus de guanine que de cytosine, et plus de thymine que d'adénine, alors que pour le segment délimité en $5'$ par le terminus et en $3'$ par l'origine l'asymétrie de composition est inversée.

Il faut remarquer que la première règle ($[G] > [C]$) semble réellement quasi-universelle, car seulement quelques espèces (*Streptomyces coelicolor*, *Halobacterium sp.*) présentent la tendance opposée. Le signe de l'asymétrie $[A] - [T]$ présente beaucoup plus de variabilité au sein des procaryotes. Cette propriété des chromosomes procaryotes a été utilisée avec succès pour déterminer la position de l'origine et du terminus (*cf.* figure 2.4, (Lobry, 1996b; Frank et Lobry, 2000)). L'asymétrie de composition a aussi servi de base pour le développement de méthodes plus complexes pour la recherche d'origines de réplication dans le génome humain (Touchon *et al.*, 2005).

Les méthodes *in silico* de recherche d'origines dans les génomes procaryotes ont toutes à la base la même idée : le calcul des rapports $\frac{([G]-[C])}{([C]+[G])}$ et $\frac{([A]-[T])}{([T]+[A])}$ (appelés aussi GC-skew et AT-skew) sur des fenêtres glissantes, et l'analyse des profils d'asymétrie qui en résultent. Les génomes procaryotes sont composés en très grande proportion de gènes (généralement plus de 80 %). Comme dit précédemment, l'orientation de ces gènes est généralement biaisée, de manière à ce que la réplication et la transcription se fasse dans la même direction. La composition

en bases des fenêtres glissantes est donc le résultat de plusieurs processus qui se superposent : des biais mutationnels associés à la réplication ou à la transcription, ainsi que des contraintes sélectives imposées par la structure des gènes codants. Une question se pose : **comment peut-on séparer les biais mutationnels dus à la réplication et les biais mutationnels ou sélectifs sous-jacents à la structure des gènes ?**

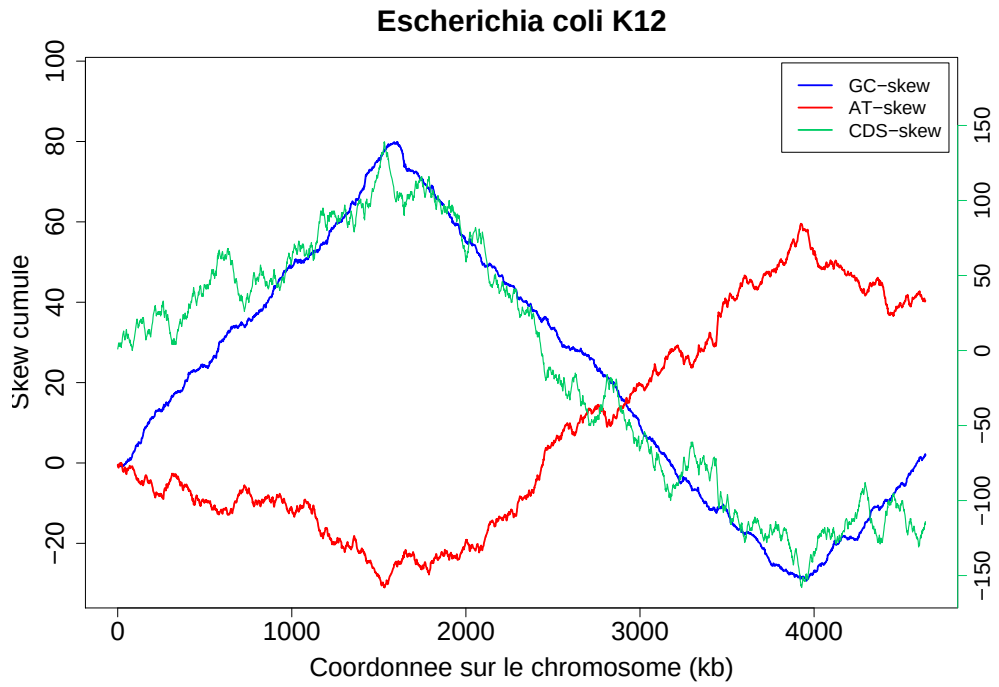

FIG. 2.4 – Asymétrie de composition et biais d'orientation des gènes le long du chromosome d'*Escherichia coli* K12. La courbe en bleu représente le rapport $\frac{([G]-[C])}{([C]+[G])}$ (aussi appelé GC-skew) calculé pour les troisièmes positions des codons des gènes protéiques, et cumulé le long du chromosome. La courbe en rouge représente le rapport $\frac{([A]-[T])}{([T]+[A])}$ (aussi appelé AT-skew), calculé de manière identique. La courbe verte représente le biais d'orientation des gènes : pour chaque gène on assigne la valeur 1 ou -1 selon qu'il est codé sur le brin direct (par rapport au brin d'ADN publié) ou sur le brin complémentaire, et on fait la somme cumulée de ces valeurs le long du chromosome. L'origine de réplication se trouve à environ 4Mb, et le terminus à environ 1.5 Mb. Ces deux positions correspondent aux changements de signe pour l'asymétrie de composition.

2.5 Méthodes de séparation des deux sources d'asymétrie de composition

Plusieurs méthodes statistiques ont été proposées pour répondre à cette question, dans la situation où les positions de l'origine et du terminus de réplication sont connues (Perriere *et al.*, 1996; Tillier et Collins, 2000; Lobry et Sueoka, 2002). Nous présenterons brièvement deux d'entre elles : celle proposée par Lobry et Sueoka (2002) et celle proposée par Tillier et Collins (2000).

2.5.1 Graphiques PR2 et comparaison des biais entre groupes de gènes

L'approche proposée par Lobry et Sueoka (2002) répose sur le calcul des déviations par rapport aux règles de parité dans des régions qui sont sous faible contrainte sélective : les troisièmes positions des codons. Il faut remarquer que ces positions n'évoluent pas de manière totalement neutre, car il peut y avoir une pression de sélection traductionnelle sur l'usage des codons synonyme. Dans les génomes procaryotes il est pourtant difficile de trouver des régions qui évoluent de manière totalement neutre, car les régions intergéniques sont courtes et contiennent souvent des signaux fonctionnels.

L'asymétrie de composition est mesurée par les rapports $\frac{A_3}{(A_3+T_3)}$ et $\frac{G_3}{(C_3+G_3)}$, sur le brin codant des gènes. Puisque l'origine et le terminus de réplication sont connus, il est possible de séparer les gènes codés sur le brin avancé des gènes codés sur le brin tardif. On peut donc faire une représentation bi-variée dans l'espace des deux mesures d'asymétrie, pour les deux groupes de gènes (*cf.* figure 2.5). Si les deux groupes de gènes sont centrés sur la même valeur, on peut en déduire que la réplication n'influence pas de manière significative l'asymétrie de composition - c'est le cas de *M. genitalium*. Si par contre les moyennes des deux groupes sont différentes, on peut conclure qu'il existe des biais mutationnels liées à la réplication qui influencent l'asymétrie de composition - c'est ce qui est observé pour *Borrelia burgdorferi* (*cf.* figure 2.5).

L'approche de Lobry et Sueoka (2002) ne se réduit pas à l'analyse des graphiques bi-variés. La significativité de la différence de l'asymétrie de composition des deux groupes est estimée par des tests de Student. La contribution relative des deux types de mécanismes (réplication et effets sous-jacents à la structure des gènes) peut aussi être estimée à partir de la différence des moyennes des deux groupes de gènes.

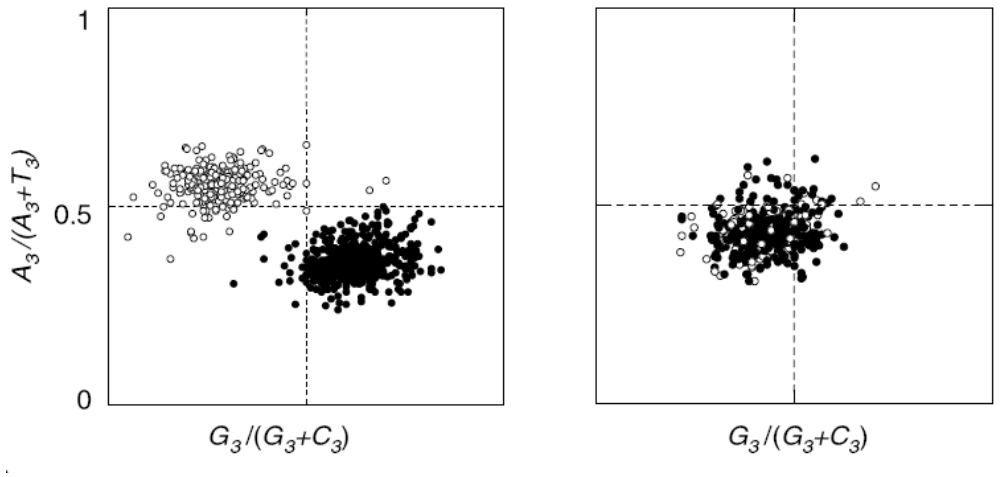

FIG. 2.5 – Application de la méthode de Lobry et Sueoka (2002) pour *Borrelia burgdorferi* et pour *Mycoplasma genitalium*.

2.5.2 Analyse de la variance

L'approche proposée par Tillier et Collins (2000) est très similaire à celle de Lobry et Sueoka (2002) par le fait qu'elle est basée sur la comparaison de moyennes de l'asymétrie de composition, pour les gènes codés sur le brin avancé et sur le brin tardif. L'asymétrie de composition est décomposée en deux fractions : le GC-skew ($\frac{([G]-[C])}{([C]+[G])}$) et le AT-skew ($\frac{([A]-[T])}{([T]+[A])}$). Ces deux valeurs sont calculées toujours par rapport au brin correspondant à la séquence publiée, séparément pour les trois positions des codons. Pour séparer les effets de la réplication et les effets sous-jacents à la structure des gènes, une analyse de variance est effectuée, avec comme variable expliquée les différentes mesures d'asymétrie, et comme variables explicatives l'orientation des gènes par rapport à la réplication (avancé ou tardif) et leur orientation par rapport au brin publié (sens direct ou complémentaire).

Ces méthodes permettent de distinguer de manière très efficace les effets des deux types de mécanisme sur la composition en bases. Elles ont cependant un grand désavantage : les positions de l'origine et du terminus de réplication doivent être connues *a priori*. Nous présenterons ci-dessous une méthode qui permet d'estimer la contribution du mécanisme de réplication à l'asymétrie de composition, et d'identifier simultanément les positions de l'origine et du terminus (Necsulea et Lobry, 2007).

2.5.3 Réarrangement artificiel des chromosomes

De nombreux travaux récents ont démontré que les biais mutationnels associés à la réplication ont un effet direct sur l'asymétrie de composition dans les génomes bactériens (Perriere *et al.*, 1996; McLean *et al.*, 1998; Tillier et Collins, 2000; Lobry et Sueoka, 2002). Certains auteurs ont aussi argumenté que les biais sous-jacents à la structure des gènes (biais mutationnels produits par la transcription, usage biaisé des codons) sont plus importants que les effets de la réplication.

Un argument intéressant a été apporté par Nikolaou et Almirantis (2005). A la base de cette étude se trouve l'observation que chez le procaryotes l'existence de fortes asymétries de composition en bases est associée à l'existence de fort biais d'orientation des gènes. Pour démontrer que les biais sous-jacents aux gènes sont suffisants pour générér de fortes asymétries de composition, les auteurs de l'article proposent un réarrangement artificiel des chromosomes bactériens, de manière à créer un biais parfait d'orientation des gènes. Pour cela, ils "déplacent" tous les gènes codés sur le brin direct (par rapport à la séquence publiée) sur la première moitié du chromosome et tous les gènes codés sur le brin complémentaire sur la deuxième moitié du chromosome (*cf.* figure 2.6). L'ordre relatif des gènes est conservé pour chacun de ces deux groupes.

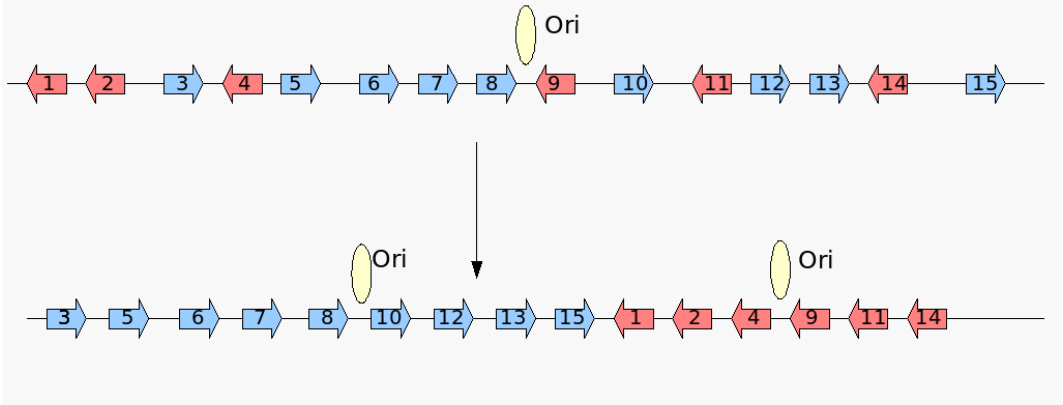

FIG. 2.6 – Schéma décrivant le réarrangement artificiel de l'ordre des gènes. Lors du réarrangement, l'ordre initial des gènes à l'intérieur des deux groupes est conservé. Nous pouvons donc retrouver la position de l'origine de réplication pour chacun des deux groupes de gènes. Dans cet exemple, nous savons que l'origine se place entre les gènes 8 et 10 (pour les gènes codés sur le brin direct) et entre les gènes 4 et 9 (pour les gènes codés sur le brin complémentaire).

Si l'on trace le diagramme cumulatif des mesures d'asymétrie (GC-skew et AT-

skew) le long du chromosome réarrangé, on peut remarquer que l'existence d'un biais parfait d'orientation des gènes entraîne une forte structuration de l'asymétrie de composition le long du chromosome. Pour le cas de *Gloeobacter violaceus*, cette structuration n'existait pas avant réarrangement (*cf.* figure 2.7). Ici, nous pouvons conclure que l'orientation des gènes est un prédicteur suffisant de leur asymétrie de composition. Pour cette espèce, le mécanisme de réplication n'a donc pas d'effet significatif sur la composition en bases. On peut donc en déduire que le biais d'orientation des gènes est un facteur *suffisant* pour la mise en place d'une structuration de la composition en nucléotides. Ce raisonnement ne nous permet pas d'affirmer que ce facteur serait aussi *nécessaire*. Nous considérons donc que la conclusion soutenue par Nikolaou et Almirantis (2005), qui est en faveur de l'absence d'un effet direct de la réplication sur l'asymétrie de composition, ne se justifie pas.

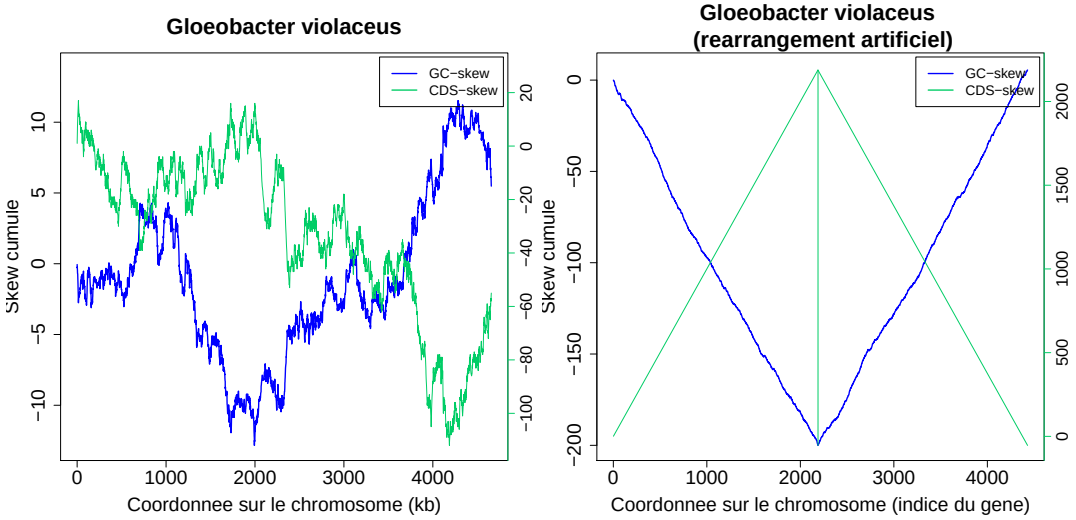

FIG. 2.7 – Asymétrie de composition le long du chromosome de *Gloeobacter violaceus*. A gauche, l'ordre initial des gènes est conservé. A droite, l'ordre des gènes est artificiellement réarrangé pour créer un biais d'orientation parfait.

Que se passe-t-il si l'on applique le même réarrangement artificiel de l'ordre des gènes à d'autres espèces ? L'exemple d'*Escherichia coli* est édifiant (*cf.* figure 2.8). Loin de suivre parfaitement l'orientation des gènes, le GC-skew présente deux points de cassure, sur chacune des moitiés du chromosome. Pour les gènes codés sur le brin direct (par rapport à la séquence publiée), les coordonnées des points de cassure sont approximativement 1.605 et 3.962 Mb, alors que pour les gènes codés sur le brin complémentaire, les coordonnées sont approximativement 1.603 et 3.858 Mb. Les positions de l'origine et du terminus de réplication pour ce génome ont été identifiés à environ 3.9 et 1.6 Mb (Blattner *et al.*, 1997). Nous

pouvons en déduire que les points de cassure observés sur le génome réarrangé correspondent à l'origine et au terminus de réplication.

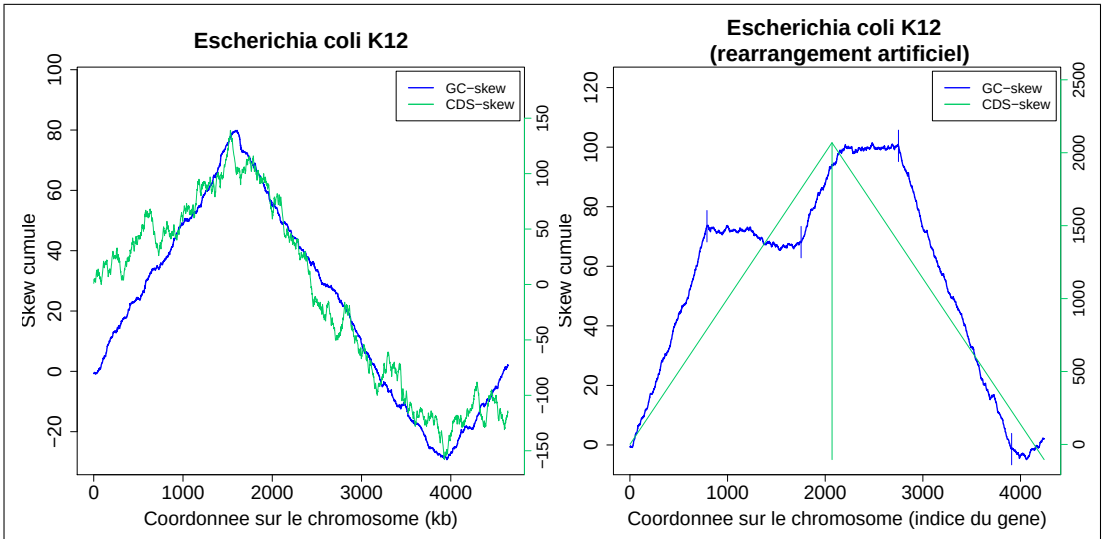

FIG. 2.8 – Asymétrie de composition le long du chromosome d'*Escherichia coli* K12. A gauche, l'ordre initial des gènes est conservé. A droite, l'ordre des gènes est artificiellement réarrangé pour créer un biais d'orientation parfait.

L'existence de points de cassure dans les biais de compositions calculés sur les génomes réarrangés est un argument fort en faveur d'un effet direct du mécanisme de réplication sur l'asymétrie de composition. Les segments délimités par ces points de cassure correspondent aux brins avancé et tardif pour la réplication. Dans le cas d'*E. coli*, le brin avancé est représenté par le premier et le troisième segment (*cf.* figure 2.8). La pente du GC-skew est donc significativement différente entre le brin avancé et le brin tardif; dans cet exemple, la pente est supérieure pour le brin avancé par rapport au brin tardif. Nous pouvons conclure donc que l'effet de la réplication est un enrichissement en guanine par rapport à la cytosine sur le brin avancé.

Résultats pour l'ensemble des génomes complètement séquencés

Les résultats précédents nous indiquent qu'il est possible d'appliquer le réarrangement artificiel de l'ordre des gènes pour évaluer l'effet de la réplication sur l'asymétrie de composition en bases. Nous avons détecté les points de cassure dans les mesures d'asymétrie de composition avec la méthode de Muggeo (2003) (*cf.* Matériel et méthodes), et nous avons estimé leur significativité grâce à une procédure de permutation (Necsulea et Lobry, 2007). Cette méthode a été

implémentée dans la bibliothèque `seqinr` (Charif et Lobry, 2006) dans R (R Development Core Team, 2005).

Nous avons appliqué cette méthode à l'ensemble des chromosomes procaryotes complètement séquencés - en juillet 2006 il y avait 389 chromosomes procaryotes dans la base de données des génomes complets du NCBI. La figure 2.9 présente les différentes situations qui peuvent être rencontrées dans l'analyse des skews réarrangés.

Dans le premier cas de figure, illustré par l'exemple d'*Anabaena variabilis*, notre méthode ne détecte pas de points de cassure significatifs, ni pour le GC-skew, ni pour le AT-skew, ou bien si des points de cassure sont détectés, leur position est distante de l'origine et du terminus de réplication. Notre interprétation dans cette situation est que le mécanisme de réplication n'a pas d'influence directe sur l'asymétrie de composition en bases. Cette situation semble être fréquente dans certaines familles de Bactéries, comme les Mollicutes ou les Cyanobactéries, ainsi que chez les Archées.

Dans la situation opposée, illustrée par l'exemple d'*Ehrlichia canis* Jake, les deux types de skews présentent des points de cassure significatifs qui coincident avec l'origine et/ou le terminus de réplication. Cette situation est de loin la plus fréquente (nous l'avons rencontrée dans 246 chromosomes sur 389). Nous pouvons confirmer la tendance générale pour le GC-skew : seulement pour quatre espèces peut-on observer un biais $[G] < [C]$ sur le brin avancé pour la réplication (*Halobacterium sp.* NRC-1, *Streptomyces coelicolor*, *Thermobifida fusca* et *Natronomonas pharaonis*). La direction du AT-skew présente par contre plus de variabilité. Dans la plupart des cas le brin avancé contient plus de thymine que d'adénine. La tendance opposée est rencontrée uniquement au sein des Firmicutes et pour la famille Fusobacteria.

Le cas d'*E. canis* Jake est aussi intéressant parce que le AT-skew présente deux points de cassure additionnels, qui ne semblent avoir aucun rapport avec l'origine ou le terminus de réplication. Dans certains cas, la présence de points de cassure supplémentaires peut être expliquée par l'existence d'inversions récentes ; d'ailleurs, les diagrammes d'asymétrie de composition ont été utilisées précédemment pour identifier certains réarrangements chromosomiques (Grigoriev, 1998, 2000). Il est également possible que ces points de cassure délimitent des suites de gènes avec un biais d'usage des codons inhabituel (peut-être acquis par transfert horizontal).

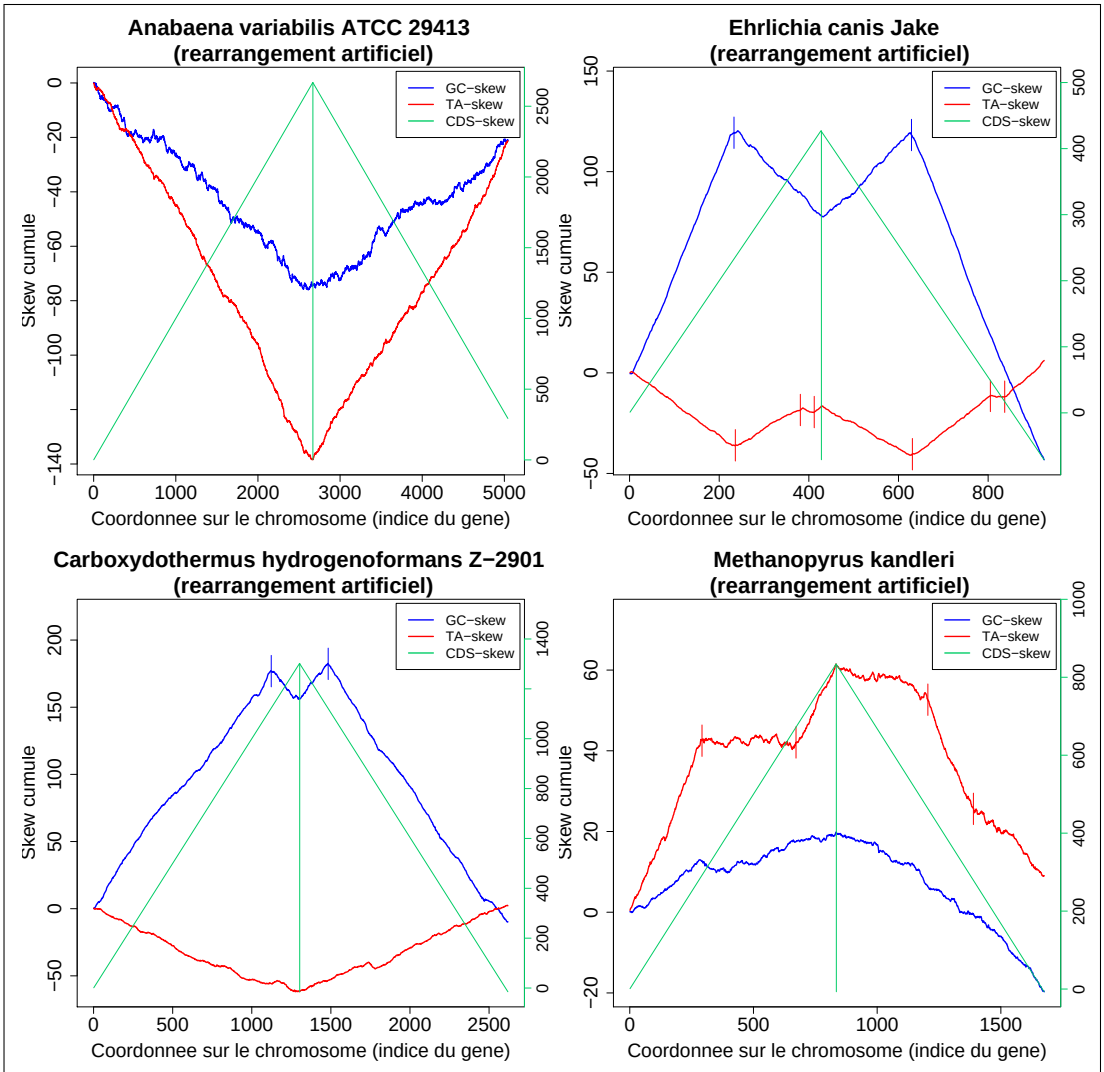

FIG. 2.9 – Asymétrie de composition dans les chromosomes réarrangés de *Anabaena variabilis* ATCC 29413, *Ehrlichia canis* Jake, *Carboxydothermus hydrogenoformans* Z-2901, *Methanopyrus kandleri*.

Dans de nombreuses espèces procaryotes le mécanisme de réplication a un effet direct seulement sur un type d'asymétrie de composition, soit GC soit AT. Par exemple, chez *Carboxydothermus hydrogenoformans* Z-2901 (figure 2.9) on trouve des points de cassure significatifs uniquement sur la courbe du GC-skew. Cette situation, où le mécanisme de réplication a un effet uniquement sur le GC-skew, est rencontrée pour 79 chromosomes sur 389 (74 Bactéries et 5 Archées). Le cas opposé, où seulement le AT-skew présente des points de cassure significatifs, est illustré ici par *Methanopyrus kandleri* (figure 2.9). Cette situation est rencontrée chez 14 chromosomes de Bactéries et 6 chromosomes d'Archées. Aucun patron taxonomique particulier n'est détecté.

Plusieurs origines de réplication dans les chromosomes d'Archées

Jusqu'à récemment, on pensait que les chromosomes d'Archées avaient une unique origine de réplication, comme les chromosomes des Bactéries. Des études expérimentales ont mis en évidence l'existence de trois origines de réplication chez *Sulfolobus acidocaldarius* et *Sulfolobus solfataricus* (Robinson *et al.*, 2004; Lundgren *et al.*, 2004). Des analyses *in silico* avaient suggéré que de multiples origines de réplication pourraient être aussi présentes chez *Halobacterium sp.* NRC-1 (Zhang et Zhang, 2005), malgré le fait qu'une seule origine ait été determinée expérimentalement (Berquist et DasSarma, 2003).

Nous avons voulu vérifier si notre approche pourrait détecter un effet significatif de la réplication sur la composition en bases lorsque les chromosomes possèdent plusieurs origines de réplication. Les résultats obtenus pour *S. acidocaldarius* et *Halobacterium sp.* sont présentés dans la figure 2.10. Pour ces deux espèces, nous avons fixé *a priori* le nombre de points de cassure à 5. Notre procédure de permutation indique que tous ces points de cassure ne sont pas significatifs ; cependant, ils correspondent aux positions des origines de réplication pour *S. acidocaldarius*. De même, pour *Halobacterium sp.* nous retrouvons les positions des origines prédites par Zhang et Zhang (2005).

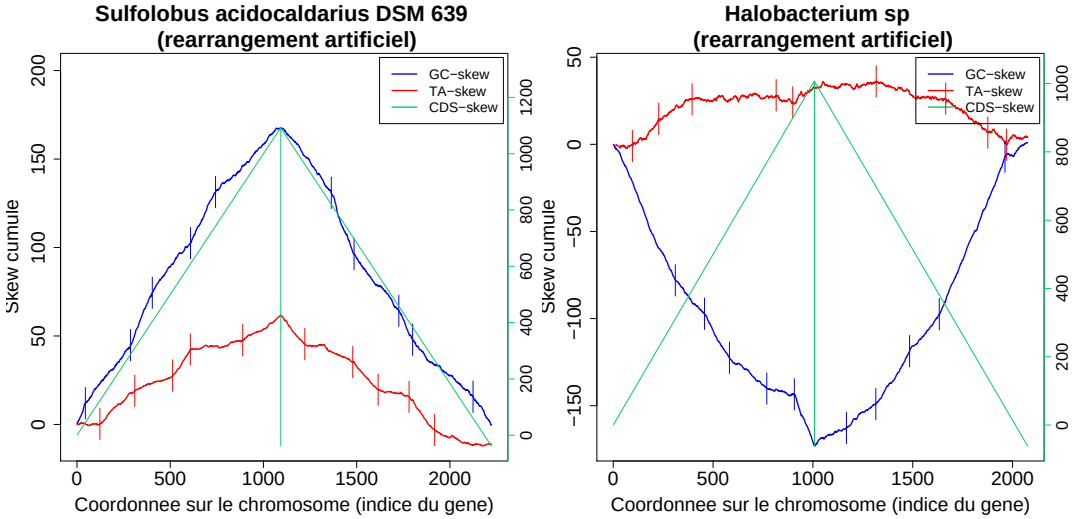

FIG. 2.10 – Asymétrie de composition dans les chromosomes réarrangés de *Sulfolobus acidocaldarius* et *Halobacterium sp.*

L'effet de la réplication sur le AT-skew chez les Firmicutes

Un article récent a suggéré que le signe du AT-skew sur le brin avancé pour la réplication pourrait être déterminé par l'identité de la sous-unité α de la polymérase qui réplique ce brin (Worning *et al.*, 2006). Chez *Bacillus subtilis*, les deux sous-unités α sont codés par deux gènes différents : *polC* et *dnaE*. La première sous-unité α synthétise le brin avancé, et la deuxième synthétise le brin tardif (Dervyn *et al.*, 2001). Les génomes des Firmicutes, Fusobacteria et Aquificales possèdent des homologues pour ces deux gènes, alors que pour les autres familles de bactéries les deux sous-unités α sont codés par le même gène. Lorsque l'on analyse l'asymétrie de composition en bases avec une méthode de type "fenêtre glissante", sans essayer de séparer les effets dus à la réplication des effets sous-jacents à la structure des gènes, il est possible de conclure que toutes ces espèces présentent aussi un enrichissement en A par rapport à T sur le brin avancé, alors que pour les espèces où les deux sous-unités α sont identiques, l'asymétrie va dans le sens opposé (Worning *et al.*, 2006). Mais que se passe-t-il lorsque les effets des deux types de mécanismes sont découplés ?

Nos résultats montrent que l'effet de la réplication sur le AT-skew est très variable même au sein des Firmicutes (*cf.* figure 2.11). Chez la majorité des espèces du genre *Bacillus*, *Lactococcus*, *Staphylococcus*, ainsi que chez *Clostridium* et *Lactobacillus salivarius* UCC 118, le brin avancé contient effectivement plus de A que de T. Mais la réplication n'a aucun effet sur le AT-skew chez *Lactobacillus acidophilus* NCFM. De plus, chez *Thermoanaerobacter tengcongensis* et *Geobacillus kaustophilus*, le brin avancé semble contenir plus de T que de A. Si l'enzyme PolC est effectivement utilisée par tous les Firmicutes pour répliquer le brin avancé, il est raisonnable de conclure que l'effet de la réplication sur le AT-skew n'est pas déterminé exclusivement par l'usage différentiel de DnaE ou de PolC.

Pourquoi cette inconsistence entre nos résultats et ceux de Worning *et al.* (2006) ? Les différences de méthodologie pourraient fournir des éléments de réponse. Dans notre approche, tout comme dans Oriloc (Frank et Lobry, 2000), l'asymétrie de composition (le GC-skew et le AT-skew) est calculée seulement pour les séquences codantes, dans les troisièmes positions des codons. La méthode développée par Worning *et al.* (2006) calcule des composition en oligomères, sur le chromosome entier. Plus important, notre approche est destinée spécifiquement à séparer les effets de la réplication et les biais sous-jacents aux séquences codantes, alors que la séparation entre les différentes sources de l'asymétrie n'est pas cherchée *a priori* par Worning *et al.* (2006). De plus, comme le biais d'orientation des gènes est plus fort chez les Firmicutes (Rocha, 2002), le fait de ne pas séparer les deux types de mécanismes peut être un facteur confondant important.

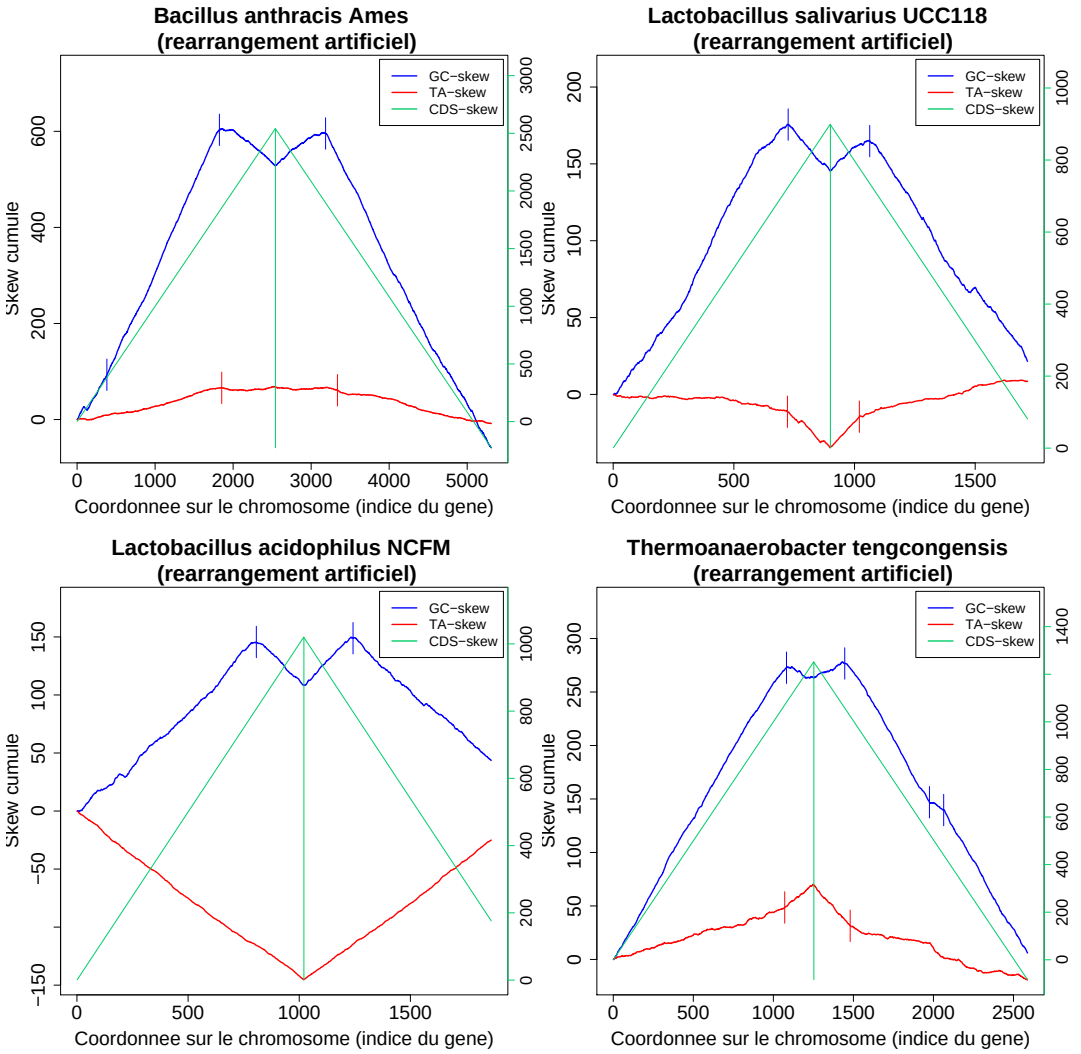

FIG. 2.11 – Asymétrie de composition dans les chromosomes réarrangés de quatre espèces de Firmicutes (*Bacillus anthracis* Ames, *Lactobacillus salivarius* UCC 118, *Lactobacillus acidophilus* NCFM, *Thermoanaerobacter tengcongensis*).

2.5.4 Comparaison des méthodes de séparation des deux sources de biais

Comme discuté précédemment, il existe à ce jour plusieurs méthodes qui permettent de découpler les deux sources d'asymétrie de composition (Perriere *et al.*, 1996; Tillier et Collins, 2000; Lobry et Sueoka, 2002). Quel est l'intérêt de proposer une nouvelle approche ? Il en existe au moins un : le fait que la méthode du

	AT + / GC +	AT + / GC -	AT - / GC +	AT - / GC -
AT + / GC +	115	0	0	0
AT + / GC -	4	3	1	0
AT - / GC +	13	0	8	1
AT - / GC -	4	1	3	5

TAB. 2.1 – Table de contingence permettant la comparaison des résultats obtenus avec la méthode du réarrangement artificiel (en ligne) et avec la méthode présentée par Lobry et Sueoka (2002) (en colonne). Le signe + signifie que pour ce type de skew nous trouvons que la réplication a un effet significatif, le signe - signifie qu'il n'y a pas d'effet de la réplication sur le skew. Par exemple, la première case du tableau (en haut à gauche) représente le nombre de chromosomes pour lesquels la méthode du réarrangement artificiel et la méthode de Lobry et Sueoka (2002) trouvent que la réplication a un effet significatif, aussi bien pour le AT-skew que pour le GC-skew.

réarrangement artificiel que nous avons présentée ici ne nécessite pas de connaissance *a priori* de la position de l'origine et du terminus de réplication.

En effet, si l'analyse des diagrammes d'asymétrie de composition sur les chromosomes réarrangés nous permettent de mettre en évidence des points de cassure significatifs, qui divisent le chromosome en deux parties de taille comparables, il est probable que ces points de cassure correspondent à l'origine et/ou au terminus de réplication. Notre méthode permet dans cette situation de conclure que la réplication a un effet significatif sur la composition en nucléotides, et d'identifier en même temps la position de l'origine et du terminus. Si par contre les mesures d'asymétrie sont parfaitement linéaires sur les chromosomes réarrangés, nous pouvons en déduire que la localisation des gènes sur les brins avancé ou tardif n'a aucun effet sur leur composition en nucléotides. Il n'est donc pas nécessaire nécessaire de connaître la position de l'origine ou du terminus de réplication pour tirer cette conclusion.

Nous pouvons toutefois nous demander si cette méthode permet d'aboutir à des conclusions comparables à celles obtenues avec des approches alternatives. Ici, nous avons appliqué les méthodes proposées par Lobry et Sueoka (2002) et Tillier et Collins (2000) pour estimer si la réplication a un effet significatif sur l'asymétrie de composition en bases. Cette étude a été réalisée sur 158 chromosomes procaryotes pour lesquels l'origine et le terminus peuevent être déterminés (*cf.* Matériel et méthodes). Pour éviter la redondance, nous avons inclus dans l'analyse seulement une espèce par genre taxonomique.

Nos résultats montrent qu'il existe un accord général entre les différentes approches (*cf.* tables 2.1 et 2.2). L'approche du "réarrangement artificiel" a ce-

	AT + / GC +	AT + / GC -	AT - / GC +	AT - / GC -
AT + / GC +	114	0	1	0
AT + / GC -	5	3	0	0
AT - / GC +	11	0	10	1
AT - / GC -	6	0	2	5

TAB. 2.2 – Table de contingence permettant la comparaison des résultats obtenus avec la méthode du réarrangement artificiel (en ligne) et avec la méthode présentée par Tillier et Collins (2000) (en colonne). Le signe + signifie que pour ce type de skew nous trouvons que la réplication a un effet significatif, le signe - signifie qu'il n'y a pas d'effet de la réplication sur le skew. Par exemple, la première case du tableau (en haut à gauche) représente le nombre de chromosomes pour lesquels la méthode du réarrangement artificiel et la méthode de Tillier et Collins (2000) trouvent que la réplication a un effet significatif, aussi bien pour le AT-skew que pour le GC-skew.

pendant tendance à être plus conservative. Par exemple, dans 18 chromosomes sur 158, notre méthode permet de conclure que la réplication n'a pas d'effet sur le AT-skew, alors que la méthode de Lobry et Sueoka (2002) conclut qu'il y a un effet significatif (*cf.* table 2.1). Les deux méthodes sont en désaccord pour 23 chromosomes sur 158.

De même, notre méthode est en désaccord avec la méthode de Tillier et Collins (2000) pour 26 chromosomes sur 158. La plupart des cas où les résultats des différentes approches sont contradictoires correspondent à des situations où le réarrangement artificiel indique qu'il n'y a pas d'effet de la réplication sur le GC-skew ou sur le AT-skew.

La méthode du "réarrangment artificiel" semble donc être assez conservative. Une explication possible pour cet aspect est que nous utilisons des tests non-paramétriques, de type permutation, pour détecter la significativité de l'effet de la réplication (*cf.* Matériel et méthodes). Les approches de Tillier et Collins (2000) et Lobry et Sueoka (2002) utilisent à ce but des tests statistiques paramétriques (tests de Student ou ANOVA). Ces tests sont connus pour être très puissants, notamment lorsque la quantité de données étudiées est importante.

2.6 Conclusion et perspectives

Nous avons présenté ici une nouvelle approche *in silico* pour l'analyse de l'asymétrie de composition en bases, avec l'objectif d'estimer l'effet direct des biais mutationnels associés à la réplication, et nous avons appliqué cette méthode à de

nombreux génomes complètement séquencés. Une conclusion que nous pouvons en déduire est que la réplication a des effets différents sur les deux mesures d'asymétrie (le GC-skew et le AT-skew). Les patrons d'asymétrie de composition, loin d'être universels, sont plus complexes que ce que l'on pensait auparavant. Cette conclusion est d'ailleurs soutenue par l'étude récente de Rocha *et al.* (2006), qui a montré que les processus mutationnels qui sont à l'origine de l'asymétrie de composition varient largement entre les espèces, même lorsque les biais de composition observés sur les séquences sont similaires. Pourquoi a-t-on cette variabilité des processus évolutifs asymétriques ? Cette question reste à présent ouverte.

2.7 Matériel et méthodes

2.7.1 Données génomiques

Le jeu de données que nous avons utilisé ici est constitué de 389 chromosomes complètement séquencés d'espèces procaryotes, extraits de la banque de données de génomes complets du NCBI en Juillet 2006. Les annotations des génomes ont été récupérées à partir de `ftp://ftp.ncbi.nih.gov/genomes/Bacteria/` et analysées en utilisant des fonctions de la bibliothèque `seqinr` (Charif et Lobry, 2006) de l'environnement statistique `R`. Les annotations ont été analysées pour extraire les coordonnées des gènes protéiques. Les gènes annotés comme partiels ou contenant des introns ont été ignorés.

2.7.2 Calcul des mesures d'asymétrie de composition

L'asymétrie de composition des chromosomes procaryotes est décrite par des diagrammes cumulatifs des mesures de GC-skew et de AT-skew, définies par : GC-skew$=\frac{([G]_3-[C]_3)}{([G]_3+[C]_3)}$ et GC-skew$=\frac{([A]_3-[T]_3)}{([A]_3+[T]_3)}$. Ces valeurs sont calculées uniquement pour les gènes codants, en troisièmes positions des codons.

Le biais d'orientation des gènes est mesuré par le CDS-skew. A chaque gène nous assignons la valeur 1 ou -1, selon l'orientation de la transcription (1 pour les gènes dont le brin codant est le brin d'ADN publié, et -1 pour les gènes dont le brin codant est le brin complémentaire). La somme cumulée de ces valeurs nous donne le GC-skew.

2.7.3 Réarrangements artificiels de l'ordre des gènes

Les chromosomes ont été réarrangés artificiellement pour obtenir un biais parfait d'orientation de gènes, comme décrit par Nikolaou et Almirantis (2005). Le brin d'ADN de référence correspond à la séquence publiée. Tous les gènes pour lesquels le brin codant correspond au brin d'ADN de référence ont été déplacés

vers la première moitié du chromosome, et tous les gènes pour lesquels le brin codant est complémentaire au brin de référence ont été déplacés vers la deuxième moitié. Ce réarrangement a été fait en conservant l'ordre des gènes dans chacun des deux groupes. Puisque nous avons calculé l'asymétrie de composition uniquement dans les séquences codantes, nous ignorons ici les régions intergéniques et les diagrammes d'asymétrie de composition le long des chromosomes réarrangés ont été tracés en fonction de l'indice des gènes (et non pas la coordonnée en kb).

2.7.4 Détection des points de cassure

Sous l'hypothèse que le biais de composition en bases est exclusivement provoqué par des mécanismes liés à la structure des gènes (par exemple un biais de mutation ou de réparation associé à la transcription, ou un processus sélectif tel que l'usage préférentiel des codons), lorsque l'on trace les diagrammes du GC-skew et du AT-skew le long des chromosomes réarrangés, nous devrions obtenir une relation linéaire. Si au contraire le mécanisme de réplication a un effet non-négligeable sur l'asymétrie de composition en bases, la différence entre le brin avancé et le brin tardif devrait être visible comme un changement de pente dans la régression linéaire des mesures d'asymétrie en fonction de la position sur le chromosome.

Nous souhaitons donc tester si la relation entre l'asymétrie de composition et la position sur les chromosomes réarrangés peut être décrite par un modèle linéaire classique, ou bien par un modèle linéaire avec points de cassure.

Nous avons utilisé la fonction `segmented.lm` dans la bibliothèque `segmented` (Muggeo, 2004) dans R pour chercher la position d'éventuels points de cassure dans la régression linéaire des GC- et AT-skew en fonction de la position sur le chromosome. Cette fonction implémente un algorithme itératif pour estimer la position des points de cassure, en essayant de maximiser une mesure de vraisemblance et de minimiser la somme des carrés des écarts résiduelle de la régression. Cette fonction nécessite des valeurs initiales pour les points de cassure (Muggeo, 2003).

Puisqu'il est impossible de savoir *a priori* combien de points de cassure peut présenter la régression linéaire, nous avons appliqué dans un premier temps la fonction `segmented.lm` pour chercher 5 points de cassure. Pour éviter la situation où les points de cassure estimés correspondent à des maxima locaux de la fonction de vraisemblance, nous avons répété la détection 150 fois, en donnant des valeurs différentes pour les points de cassure initiaux. Nous avons ensuite choisi le jeu de points de cassure qui maximise la fonction de vraisemblance. Ensuite, nous avons répété cette procédure en fixant le nombre de points de cassure à 4, 3, 2 et 1.

Nous avons également appliqué une procédure supplémentaire pour améliorer la précision de la détection, en analysant de la même manière les région restreintes autour des points de cassure initiaux.

2.7.5 Significativité statistique des points de cassure

Nous avons décidé de ne pas utiliser les tests statistiques proposés dans `segmented` pour déterminer la significativité des points de cassure, car nous craignons que le nombre important de points analysés (plusieurs milliers de gènes) ne puisse rendre les points de cassure artificiellement significatifs. Nous avons développé une méthode alternative de détection de la significativité. Pour chaque point de cassure, nous calculons la différence entre les pentes des deux segments qu'il délimite. Nous déterminons ensuite la distribution attendue (sous l'hypothèse nulle, selon laquelle la réplication n'a pas d'effet sur l'asymétrie de composition) des différences de pente avec une procédure de permutation. Pour chaque chromosome et pour chaque groupe de gènes (codés sur le brin de référence ou sur le brin complémentaire), nous permutons aléatoirement l'ordre des gènes, et ensuite nous détectons les points de cassure avec la procédure décrite plus haut. Les différences des pentes calculés pour ces permutations nous donnent la distribution attendue sous l'hypothèse nulle. Nous calculons ensuite pour chaque point de cassure "réel" une $p-value$ par rapport à cette distribution. Cette procédure de test nous permet de déterminer si les points de cassure que nous observons sont seulement une conséquence de la variation normale des mesures d'asymétrie entre les gènes, ou si au contraire ils sont dûs à la position des gènes sur le chromosome.

Nous avons choisi un seuil de significativité de 5 %. Pour déterminer le nombre optimal de points de cassure pour un chromosome, nous appliquons une approche "top-down". Nous analysons d'abord la situation où l'on a détecté 5 points de cassure. Si parmi eux il existe au moins un point de cassure dont la $p-value$ est supérieure à 5 %, nous descendons au niveau inférieur et nous analysons le jeu de données à 4 points de cassure ; et ainsi de suite jusqu'à ce que tous les points de cassure soient significatifs.

2.7.6 Origines de réplication

Pour vérifier si les points de cassure que nous avons détectés correspondent à l'origine et au terminus de réplication, nous avons besoin de connaître les positions de ces points sur les chromosomes.

Nous avons utilisé Oriloc (Frank et Lobry, 2000) pour prédire les positions de l'origine et du terminus, et nous avons utilisé comme complément les prédictions données par Worning *et al.* (2006) pour les chromosomes circulaires (`http://www.cbs.dtu.dk/services/GenomeAtlas/suppl/origin/`).

Chapitre 3

Effet du contexte nucléotidique sur les substitutions asymétriques

3.1 Introduction

Chez l'homme, comme chez les autres mammifères, la proportion du génome qui code pour des protéines est très faible, de seulement environ 1 % (Lander *et al.*, 2001). L'organisation du génome en région codantes et non-codantes est donc très déséquilibrée. Une caractéristique remarquable du génome humain est l'importante variabilité de la composition en nucléotides le long des chromosomes, qui établit un niveau de structuration plus frappant même que l'organisation fonctionnelle. Quels sont les processus évolutifs à l'origine de cette hétérogénéité compositionnelle ? Depuis la découverte de la variabilité de la composition en nucléotides dans les génomes de mammifères (Filipski *et al.*, 1973), cette question a suscité d'intenses débats.

L'aspect le plus étudié de la composition en nucléotides est le contenu en guanine et cytosine. Il est maintenant bien établi que la variation du contenu en guanine et cytosine le long des chromosomes des mammifères est engendrée par des processus évolutifs neutres, associés à un mécanisme fondamental de la cellule : la recombinaison (Meunier et Duret, 2004).

L'étude de l'asymétrie de composition a également permis de comprendre l'effet des mécanismes cellulaires essentiels sur l'évolution du génome humain. Ainsi, des déviations par rapport aux règles de parité $[A] = [T]$ et $[G] = [C]$ ont été mises en évidence dans les régions transcrites (Touchon *et al.*, 2003), et autour des origines de réplication (Touchon *et al.*, 2005). Cette propriété du génome humain a des applications importantes. Notamment, elle a permis d'inférer *in silico* les positions des origines de réplication, avant que des études expérimentales à grande échelle puissent répondre à cette question (Touchon *et al.*, 2005; Huvet

et al., 2007).

Dans ce chapitre, nous nous intéresserons aux processus évolutifs qui sont à l'origine de l'asymétrie de composition, en étudiant les patrons de substitution dans les régions transcrites et au voisinage des origines de réplication. Des études récentes sur ce sujet ont permis de décrire les patrons de substitutions asymétriques dans ces régions (Green *et al.*, 2003; Duquenne *et al.*, 2007). Nous nous intéresserons ici uniquement à un aspect particulier de l'évolution asymétrique des séquences, qui est la dépendance du contexte nucléotidique.

3.1.1 Asymétrie de composition dans le génome humain

Historiquement, la première étude qui a signalé la présence d'une asymétrie de composition dans le génome humain est celle de Wu et Maeda (1987). La séquence analysée était un court fragment (environ 3 kilobases) de la région de la β-globine, région qui contient une origine de réplication. Les résultats suggéraient l'existence de déviations par rapport aux règles de parité pour cette région ($[T] > [A]$ et $[C] > [G]$), ainsi que des patrons de substitution asymétriques dans six espèces de primates. La significativité de ces résultats a été mise en question plus tard, quand il s'est avéré que l'asymétrie était restreinte à une très courte séquence, et ne s'appliquait pas à l'ensemble de la région étudiée (Wu, 1991; Bulmer, 1991). Lorsque l'analyse à grande échelle est devenue possible grâce au projet de séquençage, de nombreuses études ont démontré que le génome humain présente des régions où la composition en nucléotides est significativement asymétrique.

Asymétrie de composition dans les régions transcrites

L'association entre l'asymétrie de composition et le mécanisme de transcription dans le génome humain a été confirmée par l'observation que les introns des gènes protéiques présentent un excès de T par rapport à A (Duret, 2002). Le biais est significatif aussi pour les nucléotides G et C, mais son amplitude est plus faible (Touchon *et al.*, 2003). L'intensité de l'asymétrie de composition n'est pas uniforme le long des introns ; les régions voisines des jonctions introns-exons présentent présentent des biais de composition extrêmes, qui diminuent vers les régions centrales des introns (Touchon *et al.*, 2004). Quel peut être le mécanisme biologique qui engendre cette asymétrie de composition ? Vraisemblablement, il s'agit d'une superposition de processus neutres et sélectifs. L'existence d'un biais de composition plus important dans les extrémités des introns pourrait être due à la présence de motifs fonctionnels impliqués dans l'épissage (Touchon *et al.*, 2004). En effet, il est maintenant reconnu que certains éléments fonctionnels impliqués dans la régulation de l'épissage des introns ont une composition asymétrique (Yeo

et al., 2007; Zhang *et al.*, 2008). Pour les régions centrales des introns, où la présence de signaux d'épissage devrait être moins fréquente, l'hypothèse qui paraît la plus vraisemblable pour expliquer le biais de composition est l'existence d'un mécanisme de mutation ou de réparation asymétrique par rapport aux deux brins d'ADN, dont la transcription serait la cause directe (Touchon *et al.*, 2004).

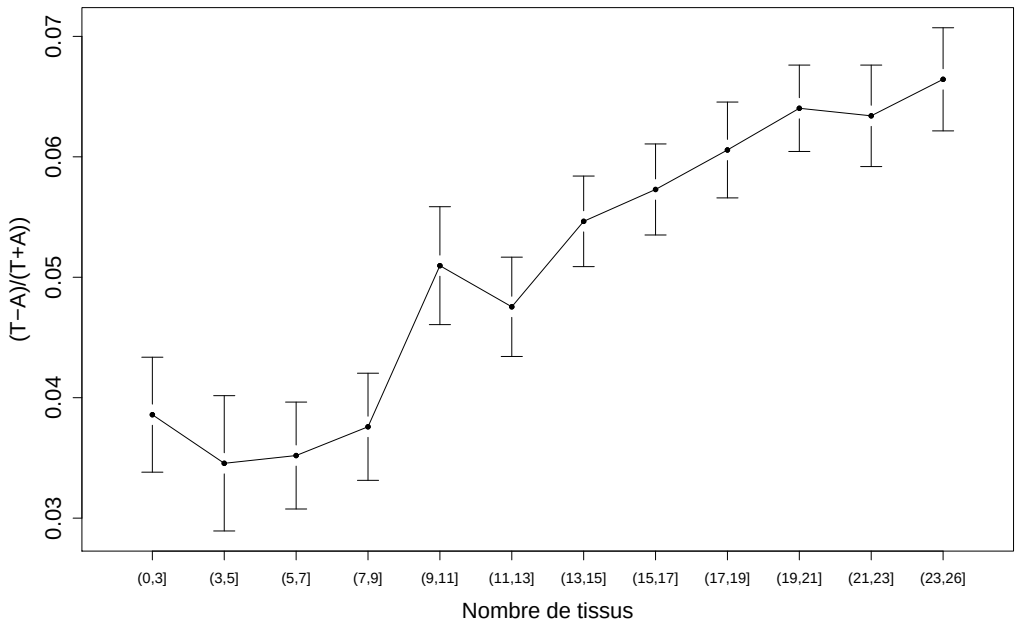

FIG. 3.1 – Asymétrie de composition (AT-skew=$\frac{[T]-[A]}{[T]+[A]}$) dans les introns des gènes, en fonction du nombre de tissus dans lesquels les gènes sont exprimés. Ce graphique représente une mise à jour de la figure présentée par Duret (2002)

Il est important de remarquer qu'il existe un lien entre l'asymétrie de composition des introns et le patron d'expression des gènes auxquels ils appartiennent. Ainsi, les gènes qui sont exprimés dans de nombreux tissus (et qui sont susceptibles d'être aussi fortement exprimés dans la lignée germinale) présentent un plus fort niveau d'asymétrie que les gènes tissu-spécifiques (*cf.* figure 3.1, Duret (2002)). De plus, parmi les gènes de ménage, ceux qui sont transcrits à fort niveau possèdent une composition plus biaisée que ceux qui sont faiblement exprimés (Majewski, 2003). Parmi les gènes tissu-spécifiques, ceux qui sont exprimés dans les testicules (et donc potentiellement dans la lignée germinale mâle) ont un biais de composition plus fort que les gènes exprimés dans les autres tissus (Comeron, 2004). Toutes ces informations renforcent l'idée que l'association entre transcription et asymétrie de composition est bien une relation cause à effet. Il faut toutefois remarquer que cette corrélation ne permet pas de privilégier

a priori l'hypothèse neutre ou celle sélective pour l'existence de la composition asymétrique. En effet, l'existence de biais de mutation ou de réparation associés à la transcription pourrait expliquer l'augmentation du biais de composition avec le niveau d'expression. Mais la même corrélation pourrait être observée si le biais de composition était généré exclusivement par la présence de signaux d'épissage, car il n'est déraisonnable d'imaginer que les gènes à plus fort niveau d'expression puissent être soumis à des contraintes plus importantes pour que l'épissage se réalise correctement.

Jusqu'à récemment, la plupart des études qui se sont intéressées à l'asymétrie de composition dans les régions transcrites ont dû se limiter à l'analyse des biais observés dans les séquences. Néanmoins, l'étude directe des patrons de substitution nucléotidique, qui est à présent réalisable, pourrait également aider à élucider les causes de ces biais de composition.

La différence d'amplitude entre les deux types de biais nucléotidiques (GC-skew et AT-skew) est informative pour les processus mutationnels sous-jacents, car elle suggère qu'il existe des transversions asymétriques par rapport aux deux brins. En effet, si les seules substitutions asymétriques sont les transitions ($A \leftrightarrow G$ et $T \leftrightarrow C$), comme ces changements affectent de la même manière les deux mesures de l'asymétrie, à l'équilibre on attend l'égalité des skews $[G] - [C]$ et $[A] - [T]$ (Touchon *et al.*, 2003).

L'étude directe des patrons de substitution dans les régions transcrites a démontré que les transitions montrent un très fort niveau d'asymétrie (Green *et al.*, 2003). Notamment, lorsque l'on comptabilise les changements par rapport au brin codant, le taux de substitution de $A \rightarrow G$ est significativement plus élevé que le taux de la substitution complémentaire. De même, on a $G \rightarrow A > C \rightarrow T$, mais la différence des deux taux est plus faible pour cette deuxième substitution (Green *et al.*, 2003). Une analyse plus détaillée a permis de confirmer que les transversions sont également asymétriques par rapport aux deux brins d'ADN (Hwang et Green, 2004), comme prédit par Touchon *et al.* (2003).

L'étude de Hwang et Green (2004) est innovante, car elle propose un modèle d'évolution des séquences d'ADN qui prend en compte la nature des nucléotides voisins en $5'$ et $3'$. Le contexte nucléotidique influence non seulement les taux de substitution, mais également leur niveau d'asymétrie par rapport aux deux brins. Ainsi, pour les substitutions de $A \rightarrow G$, l'asymétrie dans les régions transcrites est la plus forte lorsque les deux nucléotides voisins sur le brin codant sont des T (*cf.* figure 3.2, Hwang et Green (2004)).

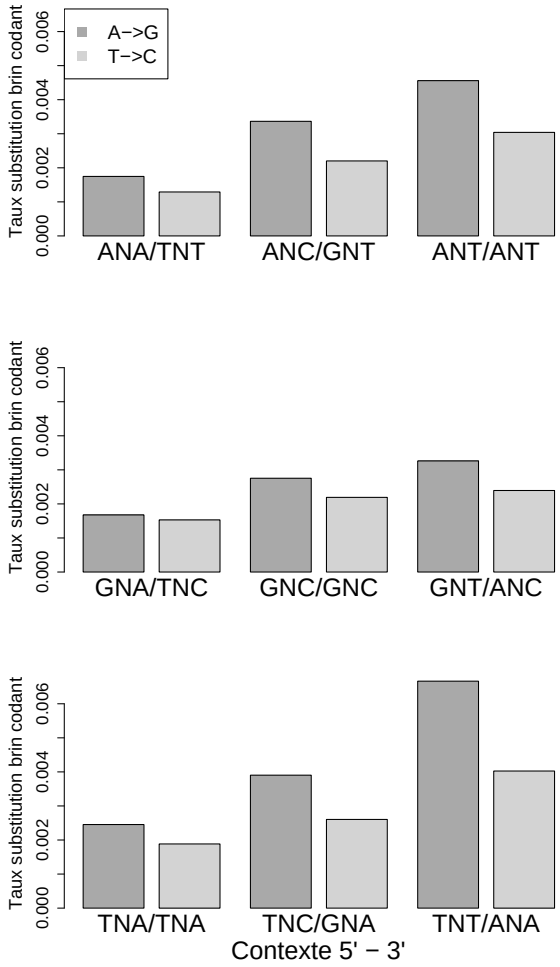

FIG. 3.2 – Patrons d'évolution asymétrique dans les séquences transcrites (introns) humains, en fonction du contexte. Les taux de substitution sont calculés toujours par rapport au brin codant. En gris foncé : taux de substitution de $A \rightarrow G$. En gris clair : taux de substitution de $T \rightarrow C$. La comparaison se fait entre substitutions complémentaires dans des contextes complémentaires ; par exemple $A \rightarrow G$ dans le contexte $5' - ANC - 3'$ est comparée avec $T \rightarrow C$ dans le contexte $5' - GNT - 3'$ (la base N représente la position où intervient la substitution). Les contextes qui pourraient se chevaucher avec des dinucléotides CpG ont été exclus de l'analyse.

Asymétrie de composition autour des origines de réplication

Les premières études qui ont discuté d'une possible association entre réplication et biais de composition en bases, notamment l'article de Wu et Maeda (1987), restent à ce jour controversées. Ces premières analyses se sont concentrées sur l'origine de réplication présente dans la région de la β-globine. La composition de cette région a permis de suggérer que dans le génome humain il n'existerait pas d'asymétrie de substitution associée à la réplication (Bulmer, 1991; Francino et Ochman, 2000). A ce jour, il est admis que l'origine voisine du gène de la β-globine représente une exception par rapport à la tendance générale observée dans le génome humain.

Plusieurs articles ont proposé que l'asymétrie de composition et sa variation régionale observée dans le génome humain pourrait être une conséquence la distribution des origines de réplication le long des chromosomes (Shioiri et Takahata, 2001; Niu *et al.*, 2003), mais la preuve indiscutable de l'association entre réplication et biais de composition a été donnée seulement en 2005, par Touchon *et al.* En analysant la composition en bases autour de neuf origines de réplication déterminées expérimentalement, les auteurs de cette étude ont démontré que les skews nucléotidiques présentent des changements abruptes de signe autour des origines (Touchon *et al.*, 2005). Cette situation peut paraître à première vue identique à celle rencontrée dans les génomes procaryotes, mais il existe une différence fondamentale. Dans les chromosomes bactériens, où le terminus de réplication se trouve dans une zone bien déterminée, l'asymétrie de composition est relativement homogène le long du brin avancé et le long du brin tardif. Pour le génome humain, le profil des biais de composition décroît linéairement entre deux origines de réplication consécutives. Ce profil est vraisemblablement le résultat de phénomènes de terminaison aléatoire de la réplication entre les deux origines, car chez les eucaryotes il n'existe pas une zone de terminaison spécifique (Touchon *et al.*, 2005).

L'analyse des biais de composition a permis d'inférer la position de plus d'un millier d'origines de réplication dans le génome humain, qui ont été confirmées expérimentalement par la suite (Touchon *et al.*, 2005; Huvet *et al.*, 2007). Ce chiffre paraît effectivement impressionnant, mais il semblerait qu'il ne s'agit que de la moitié du nombre total d'origines de réplication du génome humain (Touchon *et al.*, 2005).

3.1.2 Influence du contexte nucléotidique sur le patron de substitution

Les travaux de Hwang et Green (2004) ont permis de démontrer que l'asymétrie des substitutions est dépendante de la nature des nucléotides voisins en $5'$ et $3'$ (*cf.* figure 3.2). En ce qui concerne l'asymétrie du patron d'évolution, cette

démonstration de la dépendance du voisinage est effectivement la première. L'influence du contexte nucléotidique sur le patron de substitution a été par contre mise en évidence bien avant l'ère de la génomique.

L'étude pionnière dans ce domaine est vraisemblablement celle de Koch (1971). En utilisant un système de réversion des mutations non-sens chez le bactériophage T4, R. E. Koch a démontré que le taux de mutations du codon TAG vers le codon TGG était presque dix fois plus grand que le taux de changement $TAA \rightarrow TGA$, ce qui indique que la nature d'un nucléotide voisin (A ou G en $3'$ sur le brin de référence) a une forte influence sur le taux de mutation. Cette étude n'était pas en mesure de donner des indications sur les processus cellulaires responsables de ce biais de voisinage.

Influence du contexte sur la mutation spontanée

Dans les génomes de vertébrés, le taux de mutation spontanée peut être influencé par le contexte nucléotidique, mais jusqu'à présent le seul exemple documenté est représenté par l'hypermutabilité des dinucléotides CpG. Cette hypermutabilité peut s'expliquer par la méthylation de la cytosine dans le contexte CpG (Ehrlich et Yang, 1981), qui est suivie par la désamination en thymine (Duncan et Miller, 1980).

Influence du contexte sur la précision de la réplication

L'influence du contexte nucléotidique sur le taux de mutation semble être liée à la précision d'incorporation des nucléotides au moment de la réplication de l'ADN. Ainsi, il semblerait que lors de la synthèse d'un nouveau brin par l'ADN polymérase I d'*Escherichia coli*, le taux d'erreur est affecté par la stabilité de la double hélice d'ADN au niveau des paires de bases voisines (Patten *et al.*, 1984). Lorsque les paires de nucléotides voisines sont des paires $G \cdot C$, donc à forte stabilité, le taux d'erreur d'incorporation de nucléotides est plus fort que lorsque les paires voisines sont $A \cdot T$ (Patten *et al.*, 1984).

La stabilité des paires de nucléotides ne semble pas être le seul facteur qui joue dans le taux d'incorporation erronée lors de la réplication de l'ADN. Pour la polymérase α de *Drosophila melanogaster*, ainsi que pour la polymérase du bactériophage T4, le taux d'erreur est plus fort lorsque le nucléotide voisin est une pyrimidine sur le brin nouvellement synthétisé (Pless et Bessman, 1983; Mendelman *et al.*, 1989).

Influence du contexte sur l'efficacité de réparation des mésappariements

Le mécanisme de réparation des mésappariements est également influencé par le contexte nucléotidique. Ainsi, il semblerait que chez *E. coli* la réparation est plus efficace lorsque les nucléotides voisins sont des paires $G \cdot C$, ce qui suggère que la stabilité de la double hélice autour du mésappariement est importante pour la réparation (Radman et Wagner, 1986; Jones *et al.*, 1987). Chez l'homme, le mécanisme de réparation est également dépendent du contexte. Notamment, il a été démontré que la répartion des mésappariements $G \cdot T$ peut-être jusqu'à 12 fois plus efficace lorsque le nucléotide voisin en $5'$ est un C, dans le contexte CpG (Ullah *et al.*, 1996). Cela suggère que l'enzyme responsable de la réparation (la thymine ADN-glycosylase) s'est adaptée pour répondre à l'hypermutabilité des dinucléotides CpG chez les vertébrés (Ullah *et al.*, 1996).

Etudes in silico

Ces études expérimentales qui montrent une influence du contexte sur le taux de mutation ont été confirmées ultérieurement par des approches de génomique comparative *in silico*. Ainsi, l'analyse des patrons de substitution dans des séquences de pseudogènes de primates a mis en évidence l'existence d'un fort effet de voisinage, qui est plus fort pour, mais n'est pas restreint aux dinucléotides CpG (Blake *et al.*, 1992; Hess *et al.*, 1994). Les taux de substitution semblent être plus élevés dans les contextes qui présentent une alternance $5'$ purine - pyrimidine - purine $3'$, par rapport aux suites de pyrimidines (Blake *et al.*, 1992).

La dépendance du contexte semble être très variable selon l'espèce et le type de génome étudié. Par exemple, l'effet du voisinage dans les génomes des chloroplastes semble être particulièrement fort sur le rapport transitions/transversion. Notamment, lorsque les paires de bases voisines sont des $A \cdot T$, le taux de transversion est plus fort que le taux de transitions, alors que le rapport est inversé quand les nucléotides voisins sont des paires $C \cdot G$ (Morton et Clegg, 1995).

Ces études *in silico* ont été réalisées sur des séquences génomiques susceptibles d'évoluer de manière neutre (pseudogènes, régions intergéniques). Il est donc vraisemblable que le processus évolutif à l'origine des effets de voisinage soit neutre. Cependant, il n'est pas possible de discriminer entre biais de mutation ou de réparation en utilisant ce type d'approche.

3.2 Questions posées

Le point de départ de notre étude est l'article de Hwang et Green (2004), qui a démontré que les substitutions asymétriques associées au mécanisme de trans-

cription sont influencées par les nucléotides voisins en $5'$ et $3'$. Cette observation nous conduit à poser plusieurs questions.

Tout d'abord, quel est le mécanisme à l'origine de l'asymétrie de composition en bases dans les régions transcrites ? Nous nous intéresserons ici uniquement aux introns, et plus particulièrement aux parties centrales de ces derniers (cf. Matériel et méthodes). Comme discuté précédemment, pour ces régions la présence de motifs fonctionnels associés à l'épissage devrait être négligeable ; nous supposerons par la suite qu'à l'origine de l'asymétrie de composition observée dans ces régions se trouvent des processus évolutifs neutres, associés à la transcription. Mais s'agit-il d'un biais de mutation ou d'un biais de réparation ? Et comment pourrait-on interpréter l'influence du contexte nucléotidique sur l'asymétrie des substitutions à la lumière de ces deux hypothèses ? Un des objectifs de ce chapitre sera donc de déterminer lequel des deux mécanismes (mutation ou réparation) est le plus vraisemblable pour expliquer l'asymétrie de substitution observée, ainsi que sa dépendance du contexte.

Ensuite, existe-t-il une dépendance de contexte pour les substitutions asymétriques associées à la réplication ? Des travaux récents ont montré que les transitions $A \rightarrow G$ et $T \rightarrow C$, dont l'asymétrie est très forte dans les régions transcrites (Green *et al.*, 2003), sont aussi très asymétriques sur les brins avancé et tardif pour la réplication (Duquenne *et al.*, 2007). Il n'est cependant pas encore connu si le voisinage peut influencer l'intensité de l'asymétrie. L'existence (ou non) d'un effet de contexte pourrait fournir des informations importantes sur la nature des mécanismes qui sont à l'origine de ces patrons de substitution.

3.3 Effets de voisinage sur l'asymétrie de substitution associée à la transcription

3.3.1 La transcription est la cause de l'asymétrie du patron de substitution

Nous avons remarqué précédemment (*cf.* figure 3.1) que l'asymétrie de composition augmente avec le niveau d'expression, ce qui confirme que la transcription est la cause du biais de composition. Il est possible d'appliquer le même raisonnement pour l'asymétrie des substitutions. Par exemple, nous pouvons remarquer que la différence des taux de substitution $(A \rightarrow G) - (T \rightarrow C)$ augmente avec le nombre de tissus dans lesquels le gène est exprimé (*cf.* figure 3.3). Par conséquent, nous ne pouvons pas rejeter l'hypothèse que l'expression des gènes est la cause de l'asymétrie de cette substitution.

Par contre, pour la différence des taux de substitution $(C \rightarrow T) - (G \rightarrow A)$, la

corrélation avec le niveau de transcription n'est pas très forte. Le niveau d'expression n'explique que 1 % de la variabilité de cette mesure d'asymétrie. (*Nota bene* : Les taux de substitution présentés ici concernent uniquement les sites non-CpG, *cf.* Matériel et méthodes.) La transcription semble donc avoir très peu d'effet sur l'asymétrie de la transition $C \to T$. Pour la suite, nous discuterons uniquement des substitutions $A \to G$ et $T \to C$.

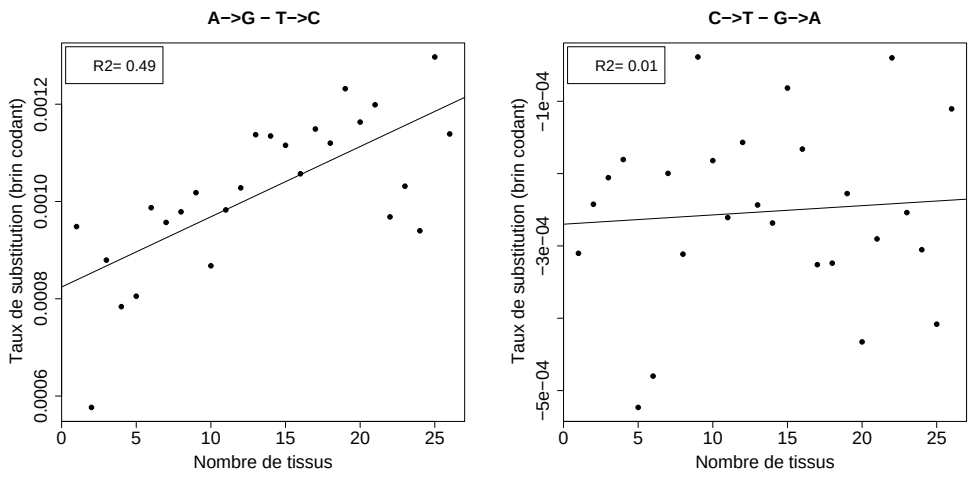

FIG. 3.3 – Différence des taux de substitution $(A \to G) - (T \to C)$ (à gauche) et $(C \to T) - (G \to A)$ (à droite) en fonction du nombre de tissus dans lesquels le gène est exprimé. Les taux de substitution sont calculés toujours par rapport au brin codant du gène.

3.3.2 Comment distinguer entre mutation et réparation ?

La variation des patrons de substitution asymétriques en fonction du niveau de transcription pourrait fournir des éléments de réponse à notre première question. Sous l'hypothèse que l'asymétrie de composition est générée par un taux plus fort de mutation sur le brin codant, nous devrions observer une augmentation du taux de substitution avec le niveau d'expression. Si par contre l'asymétrie de composition est créée par un biais de réparation, nous devrions voir une diminution du taux de substitution avec le niveau d'expression (*cf.* figure 3.4). Ce raisonnement a été utilisé par Beletskii et Bhagwat (1996) pour démontrer que la transcription augmente le taux de mutations de $C \to T$ chez *E. coli*.

Si l'on dispose d'estimations des taux de mutations qui ont effectivement lieu sur chaque brin, ce type de raisonnement permet d'identifier *avec précision* le

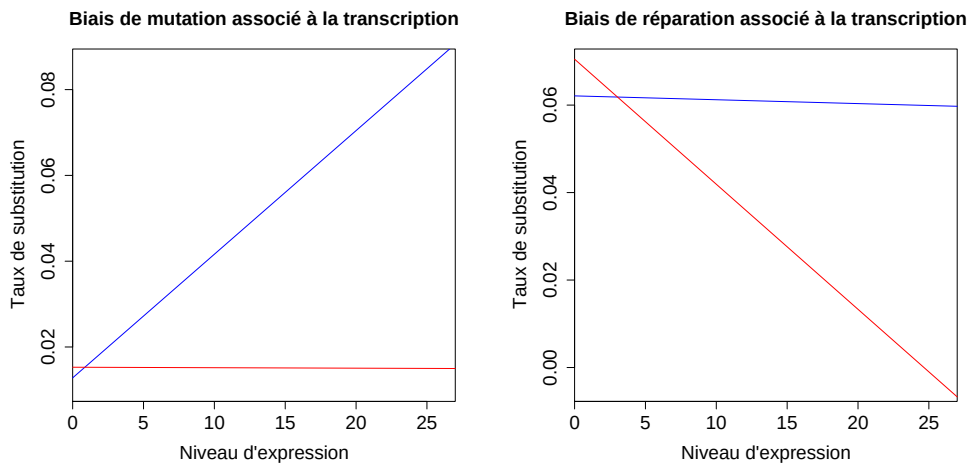

FIG. 3.4 – Exemple théorique de variation attendue des taux de substitution en fonction du niveau d'expression, dans le cas d'un biais de mutation associé à la transcription (à gauche) et dans le cas d'un biais de réparation associé à la transcription (à droite). Cet exemple est valable pour le cas où l'on connaît sur quel brin ont lieu les mutations. Le tracé bleu représente les mutations qui ont lieu sur le brin codant, le tracé rouge représente les mutations qui ont lieu sur le brin matrice.

mécanisme à l'origine de l'asymétrie. Par exemple, si l'on peut dire que le taux de mutation de $C \rightarrow T$ sur le brin codant augmente avec le niveau d'expression, alors que le même taux de mutation reste constant sur le brin matrice, nous pouvons en déduire qu'il s'agit en effet d'une plus grande mutabilité de la cytosine sur le brin codant. De même, si nous remarquons que le taux de mutation de $A \rightarrow G$ sur le brin codant reste constant, alors que sur le brin matrice il diminue avec le niveau d'expression, nous pouvons conclure que les mutations $A \rightarrow G$ sont sujettes à un mécanisme de réparation couplée à la transcription sur le brin matrice.

Malheureusement, les estimations des taux de mutation spécifiques de chaque brin peuvent seulement être obtenues expérimentalement. Lorsque le patron de substitution est inféré *in silico* avec une approche comparative, il n'est pas possible de déterminer sur quel brin ont eu lieu les mutations. Une substitution $A \rightarrow G$ *vue* du brin codant peut en effet être une substitution $T \rightarrow C$ qui a eu lieu sur le brin matrice. La variation des taux de substitution avec le patron d'expression pourra quand même nous donner des informations sur les mécanismes responsables de l'asymétrie, si l'on fait des hypothèses simplificatrices. Ainsi, pour la suite nous allons supposer que s'il y a un biais de réparation associé à la transcription, la réparation se fera préférentiellement sur le brin matrice. De

même, nous admettons que si la transcription induit une augmentation du taux de mutation, cette augmentation se fera sur le brin codant. Ces hypothèses simplificatrices sont réalistes, car en accord avec nos connaissances actuelles (Mellon *et al.*, 1987; Beletskii et Bhagwat, 1996).

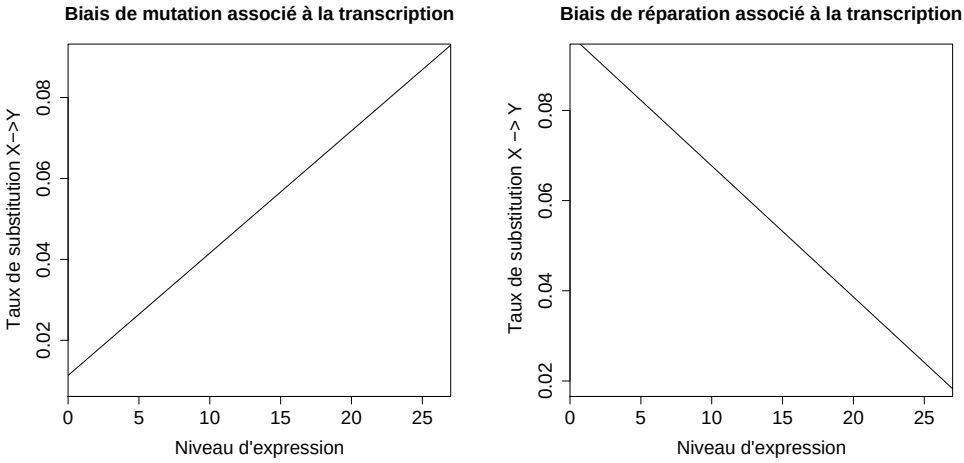

FIG. 3.5 – Exemple théorique de variation attendue des taux de substitution en fonction du niveau d'expression, dans le cas d'un biais de mutation associé à la transcription (à gauche) et dans le cas d'un biais de réparation associé à la transcription (à droite). Dans cette situation, nous ne savons pas sur quel brin ont réellement lieu les mutations. Les substitutions sont calculées par rapport au brin codant.

Avec ces hypothèses simplificatrices, nous pouvons interpréter la variation des taux de substitution en fonction du niveau d'expression en terme de biais de mutation ou de réparation. Prenons l'exemple hypothétique présenté dans la figure 3.5. Dans le premier graphique, le taux de substitution de $X \rightarrow Y$ calculé par rapport au brin codant augmente avec le niveau d'expression. Dans cette situation, nous concluons que la transcription entraîne une plus forte mutabilité des bases X sur le brin codant. Dans le deuxième graphique, le même taux de $X \rightarrow Y$ diminue avec le niveau d'expression. Nous en déduisons alors qu'il existe un biais de réparation associé à la transcription, et que ce biais de réparation affecte les bases $\bar{X}$ (complémentaires de X) sur le brin matrice.

Une question se pose : comment estimer le niveau d'expression des gènes ? D'un point de vue évolutif, les biais de mutation ou de réparation couplés à la transcription n'ont d'impact sur la composition des séquences que s'ils se produisent dans la lignée germinale. Idéalement, il faudrait donc mesurer les niveaux de transcription des gènes dans les cellules de la lignée germinale. A ce jour,

nous ne disposons pas d'informations quantitatives sur le niveau d'expression des gènes dans ces lignées. Nous avons décidé d'utiliser comme approximation l'étendue d'expression des gènes, c'est à dire le nombre de tissus dans lesquels ils sont transcrits. Cette approximation n'est pas déraisonnable. En effet, les gènes tissu-spécifiques ne devraient être exprimés dans la lignée germinale que de manière accidentelle, alors que les gènes de ménage devraient être transcrits à haut niveau. Une discussion plus détaillée des implications que ce choix peut avoir sur nos résultats sera donnée plus loin (*cf.* Discussion).

3.3.3 Variation de l'asymétrie de substitution en fonction du patron d'expression

La figure 3.6 montre la variation des taux de substitution $A \rightarrow G$ et $T \rightarrow C$, calculées par rapport au brin codant, avec le niveau d'expression des gènes. Ces substitutions sont calculées pour les sites non-CpG. Nous remarquons que les deux taux diminuent avec le niveau d'expression, et que cette diminution est plus forte pour le changement $T \rightarrow C$.

On peut expliciter $(A \rightarrow G)_{vuecodant} = (A \rightarrow G)_{codant} + (T \rightarrow C)_{matrice}$ et $(T \rightarrow C)_{vuecodant} = (T \rightarrow C)_{codant} + (A \rightarrow G)_{matrice}$. Les deux mutations sur le brin codant $(A \rightarrow G)_{codant}$ et $(T \rightarrow C)_{codant}$ soit augmentent avec le niveau de transcription, soit restent constantes, car nous savons que la réparation se fait préférentiellement sur le brin matrice (Mellon *et al.*, 1987). Puisque nous observons une diminution de $(A \rightarrow G)_{vuecodant}$ et $(T \rightarrow C)_{vuecodant}$, il s'ensuit que les deux taux de mutation $(T \rightarrow C)_{matrice}$ et $(A \rightarrow G)_{matrice}$ doivent diminuer avec le niveau de transcription, ce qui indique qu'il y a réparation de ces deux mutations. La diminution est plus forte pour $(T \rightarrow C)_{vuecodant}$. Il existe deux possibilités : soit la réparation de la mutation $A \rightarrow G$ est plus efficace que celle de $T \rightarrow C$ sur le brin matrice, soit cette dernière est compensée par une plus forte mutabilité de $A \rightarrow G$ sur le brin codant. L'explication la plus parcimonieuse est la première, si l'on admet que la mutabilité est la même pour $(A \rightarrow G)_{codant}$ et $(T \rightarrow C)_{codant}$.

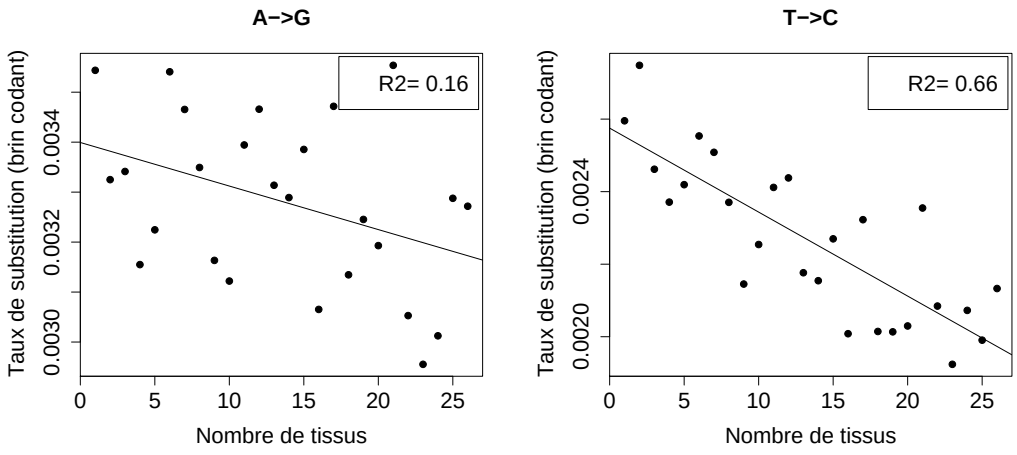

FIG. 3.6 – Variation du taux de substitution de $A \rightarrow G$ et de $T \rightarrow C$ en fonction du nombre de tissus dans lesquels le gène est exprimé. Les taux de substitution sont calculés toujours par rapport au brin codant du gène.

Que se passe-t-il lorsque le contexte nucléotidique est pris en compte ? La figure 3.7 montre la variation des taux de substitution $A \rightarrow G$ dans le contexte TNT et $T \rightarrow C$ dans le contexte complémentaire ANA, calculées par rapport au brin codant. Comme précédemment, nous avons :
$(A \rightarrow G)^{TNT}_{vuecodant} = (A \rightarrow G)^{TNT}_{codant} + (T \rightarrow C)^{ANA}_{matrice}$ et
$(T \rightarrow C)^{ANA}_{vuecodant} = (T \rightarrow C)^{ANA}_{codant} + (A \rightarrow G)^{TNT}_{matrice}$.

La première substitution ne varie pas selon le niveau de transcription, alors que la deuxième diminue très fortement. Il y a donc forcément réparation pour la mutation $(A \rightarrow G)_{matrice}$ dans le contexte TNT. Pour les mutations $(T \rightarrow C)_{matrice}$ dans le contexte ANA, il y a deux possibilités : soit il n'y a pas de réparation, soit cette réparation est contre-balancée par un plus fort taux de mutation de $A \rightarrow G$ sur le brin codant.

De même, lorsque l'on compare les substitutions $A \rightarrow G$ dans le contexte ANA et $T \rightarrow C$ dans le contexte complémentaire TNT (*cf.* figure 3.8), avec le même raisonnement nous pouvons déduire qu'il y a forcément réparation pour la mutation $(A \rightarrow G)_{matrice}$ dans le contexte TNT. La diminution $(T \rightarrow C)_{vuecodant}$ dans le contexte ANA est plus forte que celle de $(T \rightarrow C)_{vuecodant}$ dans le contexte TNT ($R^2 = 0.41$ et $R^2 = 0.17$, pentes $-0.39 * 10^{-4}$ et $-0.08 * 10^{-4}$ respectivement). Il est donc possible que la réparation de la mutation $A \rightarrow G$ sur le brin matrice soit plus efficace dans le contexte ANA que dans le contexte TNT.

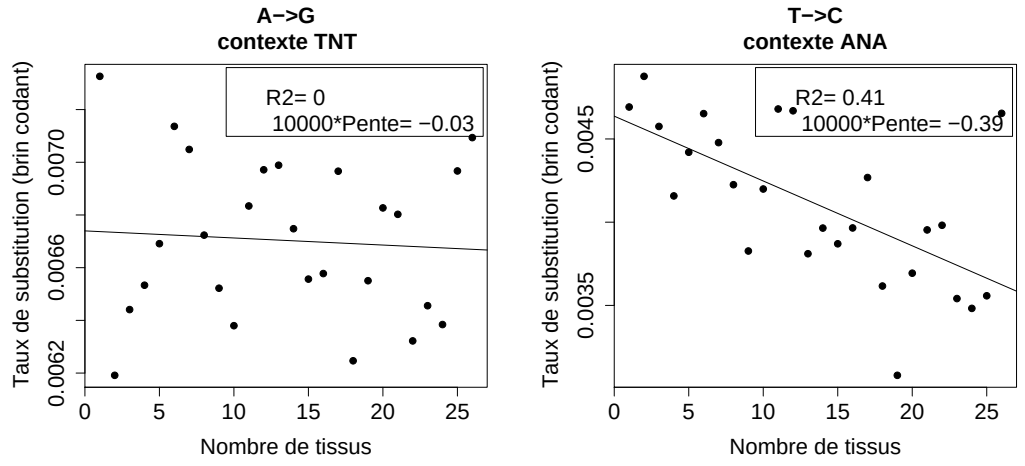

FIG. 3.7 – Variation du taux de substitution de $A \rightarrow G$ et de $T \rightarrow C$ en fonction du nombre de tissus dans lesquels le gène est exprimé, pour les contexte TNT et ANA. Les taux de substitution sont calculés toujours par rapport au brin codant du gène.

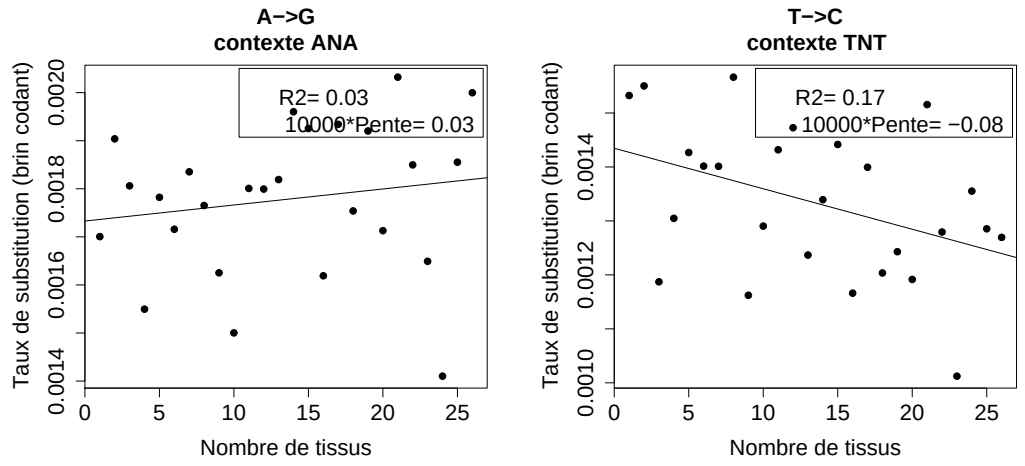

FIG. 3.8 – Variation du taux de substitution de $A \rightarrow G$ et de $T \rightarrow C$ en fonction du nombre de tissus dans lesquels le gène est exprimé, pour les contextes ANA et TNT. Les taux de substitution sont calculés toujours par rapport au brin codant du gène.

Les résultats que nous avons obtenus pour l'ensemble des contextes sont décrits dans le tableau 3.1.

Substitution	Pente x10000 (p)	Substitution	Pente x10000 (p)
$(A \to G)_{ANA}$	0.03 (0.418)	$(T \to C)_{TNT}$	-0.08 (0.037)
$(A \to G)_{TNT}$	-0.03 (0.74)	$(T \to C)_{ANA}$	-0.39 ($< 10^{-3}$)
$(A \to G)_{ANC}$	-0.05 (0.451)	$(T \to C)_{GNT}$	-0.25 ($< 10^{-3}$)
$(A \to G)_{GNT}$	-0.24 (0.012)	$(T \to C)_{ANC}$	-0.16 (0.063)
$(A \to G)_{ANG}$	-0.14 (0.003)	$(T \to C)_{CNT}$	-0.19 (0.001)
$(A \to G)_{CNT}$	-0.28 (0.024)	$(T \to C)_{ANG}$	-0.43 (0.004)
$(A \to G)_{ANT}$	-0.08 (0.28)	$(T \to C)_{ANT}$	-0.36 ($< 10^{-3}$)
$(A \to G)_{TNA}$	-0.12 (0.048)	$(T \to C)_{TNA}$	-0.19 ($< 10^{-3}$)
$(A \to G)_{CNA}$	-0.19 (0.047)	$(T \to C)_{TNG}$	-0.24 ($< 10^{-3}$)
$(A \to G)_{TNG}$	-0.08 (0.26)	$(T \to C)_{CNA}$	-0.36 ($< 10^{-3}$)
$(A \to G)_{CNC}$	-0.15 (0.153)	$(T \to C)_{GNG}$	-0.25 (0.008)
$(A \to G)_{GNG}$	-0.14 (0.001)	$(T \to C)_{CNC}$	-0.08 (0.122)
$(A \to G)_{CNG}$	-0.15 (0.178)	$(T \to C)_{CNG}$	-0.09 (0.275)
$(A \to G)_{GNC}$	-0.16 (0.119)	$(T \to C)_{GNC}$	-0.23 (0.004)
$(A \to G)_{GNA}$	-0.15 (0.002)	$(T \to C)_{TNC}$	-0.11 (0.008)
$(A \to G)_{TNC}$	-0.21 (0.009)	$(T \to C)_{GNA}$	-0.39 ($< 10^{-3}$)

TAB. 3.1 – Pente des régressions linéaires des taux de substitution en fonction du nombre de tissus dans lesquels les gènes sont exprimés. Les valeurs données en parenthèses représentent les p-values du test de Student pour l'hypothèse nulle pente = 0. Les taux de substitution sont donnés par rapport au brin codant. Nous remarquons que les pentes sont dans la grande majorité des cas négatives, ou très proches de 0. Nous n'observons jamais une augmentation significative du taux de substitution avec le niveau d'expression, donc nous ne pouvons pas conclure en faveur de la mutagénicité de la transcription, du moins pour les substitutions étudiées ici. La pente de la régression linéaire du taux de substitution en fonction du niveau d'expression varie en fonction du contexte. Cela nous suggère que la réparation associée à la transcription peut être plus ou moins efficace selon la nature des nucléotides voisins en $5'$ et en $3'$. Dans ce tableau, nous avons représenté tous les contextes possibles. Il faut noter cependant que certains d'entre eux chevauchent des dinucléotides CpG ; dans ces cas-là, il est possible que l'inférence par parcimonie soit erronée. Dans le texte, nous discutons uniquement des contextes non-CpG.

Il faut remarquer que nos résultats ne soutiennent jamais que la transcription est mutagène (pour les changements $A \rightarrow G$ et $T \rightarrow C$), car nous n'avons pas réussi à mettre en évidence une augmentation du taux de substitution avec le niveau d'expression. Cependant, dans ce cas-ci l'absence de preuve n'est pas synonyme avec la preuve de l'absence. Comme expliqué précédemment, dans chacune des situations il est possible d'imaginer un scénario moins parcimonieux qui fait apparaître en même temps un biais mutationnel et un biais de réparation associés à la transcription.

Sous l'hypothèse simplificatrice que la transcription n'augmente pas les taux de mutation de $A \rightarrow G$ et $T \rightarrow C$, nos résultats nous permettent de discuter de l'efficacité de la réparation sur le brin matrice dans les différents contextes nucléotidiques. De manière générale, nous pouvons remarquer que la variation du taux de substitution de $T \rightarrow C$ (calculé par rapport au brin codant) avec le niveau d'expression est plus forte que celle du taux de substitution de $A \rightarrow G$ (*cf.* figure 3.3, tableau 3.1). Cela suggère que la réparation des mutations $A \rightarrow G$ sur le brin matrice pourrait être plus efficace que celle des mutations $T \rightarrow C$.

Si l'on compare les pentes des régressions linéaires entre le taux de substitution de $A \rightarrow G$ et le patron d'expression dans les différents contextes (*cf.* tableau 3.1), nous pouvons conclure que l'efficacité de réparation des mutations $T \rightarrow C$ sur le brin matrice est la plus faible pour le contexte TNT. De même, en analysant les variations des taux de $T \rightarrow C$ avec le patron d'expression, nous pouvons dire que l'efficacité de réparation de la mutation $A \rightarrow G$ semble être la plus faible pour les contextes ANA et GNA, et la plus forte pour les contextes TNC et TNT.

Nous concluons ici que la dépendance de voisinage observée pour les substitutions asymétriques dans les régions transcrites pourrait être due à l'impact du contexte $5'$ - $3'$ sur l'efficacité de la réparation sur le brin matrice.

3.4 Effets de voisinage sur l'asymétrie de substitution associée à la réplication

L'asymétrie des substitutions $A \rightarrow G$ et $T \rightarrow C$ dans les régions transcrites du génome humain a été mise en évidence pour la première fois par Green *et al.* (2003). Les auteurs de cette étude avaient même suggéré que cette propriété du patron de substitution pourrait être utilisée pour détecter l'orientation et l'étendue des régions transcrites. Récemment, l'identification à grande échelle des origines de réplication dans le génome humain (Huvet *et al.*, 2007) a permis de démontrer que cette idée n'est pas applicable, car la transition $A \rightarrow G$ est également asymétrique entre le brin avancé et le brin tardif pour la réplication (Duquenne *et al.*, 2007).

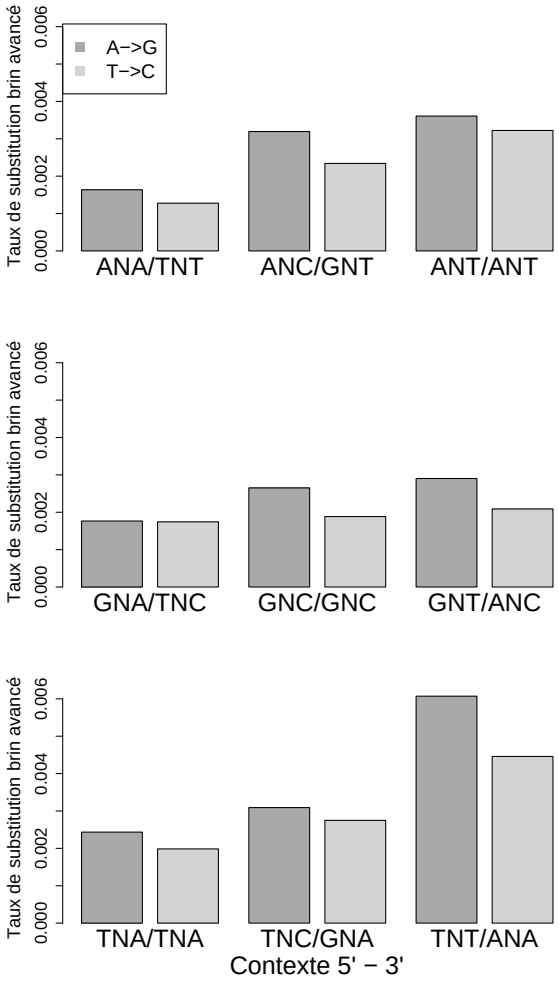

FIG. 3.9 – Asymétrie des substitutions $A \rightarrow G$ et $T \rightarrow C$ sur le brin avancé pour la réplication, en fonction du contexte $5'$ - $3'$. Les contextes qui pourraient se chevaucher avec des dinucléotides CpG ont été exclus de l'analyse.

Dans les régions transcrites, l'intensité de l'asymétrie des substitutions $A \rightarrow G$ et $T \rightarrow C$ est dépendante du contexte $5'$ - $3'$. Il est donc pertinent de se demander si les mêmes effets de voisinage sont observés pour l'asymétrie associée à la réplication. Pour répondre à cette question, nous avons analysé le patron de substitution dans les régions voisines des origines de réplication déterminées par Huvet *et al.* (2007).

Dans le génome humain, l'orientation des gènes protéiques n'est pas aléatoire ; il existe une forte tendance de co-orientation entre réplication et transcription (Huvet *et al.*, 2007). Cette co-orientation peut constituer un facteur confondant

pour notre analyse. Pour éviter cela, nous avons effectué notre analyse uniquement sur des régions intergéniques.

Les taux de substitution $A \rightarrow G$ et $T \rightarrow C$ sur le brin avancé pour la réplication sont présentés dans la figure 3.9. Nous remarquons que l'intensité de l'asymétrie de ces deux substitutions est affectée par le contexte nucléotidique. Dans le contexte TNT/ANA, les taux de substitution sont très élevés, et le niveau d'asymétrie est également très fort, alors que dans le contexte GNA/TNC la différence des deux taux de substitution est très faible. Ces deux observations sont en parfaite concordance avec les résultats obtenus pour les séquences transcrites (*cf.* figures 3.9 et 3.2). De manière générale, la variation des niveaux d'asymétrie en fonction du contexte est très similaire pour la réplication et pour la transcription (avec peut-être l'exception du contexte ANT, qui présente un faible niveau d'asymétrie pour la réplication, mais un fort niveau d'asymétrie pour la transcription).

3.5 Superposition de deux facteurs : réplication et transcription

Pourquoi cette similarité entre réplication et transcription ? La co-orientation des deux processus pourrait répondre au moins en partie à cette question. Pour l'analyse du patron de substitution sur le brin avancé nous avons étudié uniquement des régions intergéniques, pour éviter la superposition des deux sources d'asymétrie. Par contre, pour l'analyse des régions transcrites nous n'avons pas fait de distinction entre gènes codés sur le brin avancé et gènes codés sur le brin tardif. En effet, cette distinction est plus difficile à faire, d'une part parce que nous ne connaissons pas les positions de toutes les origines de réplication dans le génome humain, et d'autre part parce que la terminaison se fait de manière aléatoire.

Est-il possible que l'asymétrie de substitution que l'on observe dans les régions transcrites soit en effet provoquée par la réplication, du fait de la co-orientation des deux mécanismes ? Nos résultats indiquent que l'asymétrie des substitutions $A \rightarrow G$ et $T \rightarrow C$ augmente avec le niveau d'expression, ce qui suggère que la transcription pourrait être une cause directe de l'asymétrie. Mais la réplication peut là-aussi être un facteur confondant : en effet, dans le génome humain les gènes exprimés dans de nombreux tissus sont localisés préférentiellement au voisinage des origines de réplication, alors que les gènes tissu-spécifiques sont plus éloignés des origines (*cf.* figure 3.10, Huvet *et al.* (2007)).

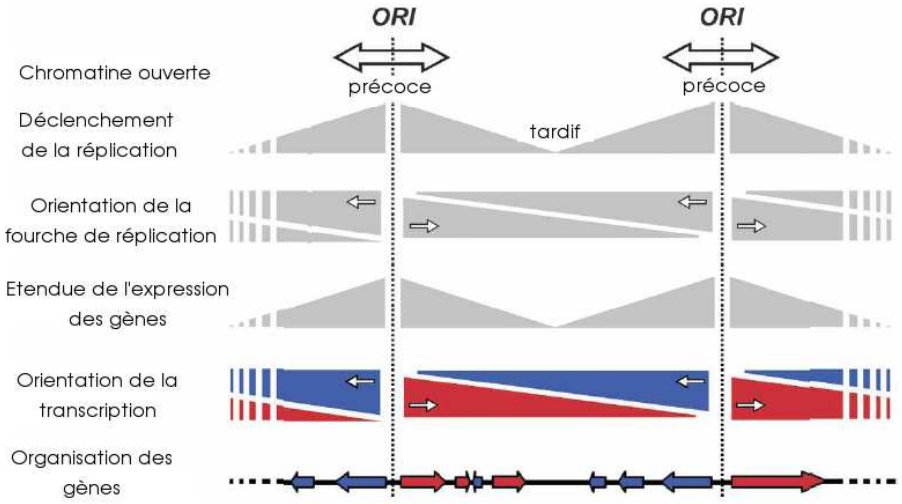

FIG. 3.10 – Organisation du génome humain autour des origines de réplication. Les gènes de ménage (qui sont exprimés dans de nombreux tissus) sont localisés préférentiellement près des origines. La plupart des gènes sont codés sur le brin avancé pour la réplication. D'après Huvet *et al.* (2007).

Nous avons également démontré que les taux de substitution $A \rightarrow G$ et $T \rightarrow C$ dans les régions transcrites ont tendance à diminuer avec le niveau d'expression des gènes, ce qui suggère que le mécanisme à l'origine de l'asymétrie pourrait être un biais de réparation couplée à la transcription.

Est-ce que là-aussi la position par rapport aux origines de réplication peut être un facteur confondant ? Cette possibilité ne peut pas être exclue. Cependant, il est difficile d'imaginer pourquoi le mécanisme de réplication de l'ADN aurait comme conséquence la diminution du taux de mutation au voisinage des origines. De même, il est peu probable que la diminution des taux de substitution soit due à des processus sélectifs, car les séquences que nous étudions sont sous faible contrainte (introns et régions intergéniques).

Les effets confondants dus à la co-orientation de la réplication et de la transcription mis à part, il est possible que la similarité entre les patrons de substitution asymétrique engendrés par les deux processus ait des causes biologiques réelles. L'hypothèse qui nous paraît la plus vraisemblable pour expliquer l'asymétrie associée à la transcription est l'existence d'un biais de réparation spécifique au brin matrice. Il est tout aussi possible que l'asymétrie de substitution associée à la réplication soit la conséquence d'un biais de réparation. Par exemple, Radman (1998) avait suggéré que la réparation du brin tardif pourrait se faire avec plus d'efficacité, du fait de la synthèse discontinue de ce brin - car la machinerie de

réparation nécessite la présence de cassures les deux brins d'ADN.

3.6 Discussion

Pour réaliser les analyses présentées dans ce chapitre, nous avons posé plusieurs hypothèses simplificatrices. Nous discuterons ci-dessous des arguments qui peuvent être amenés pour ou contre chacune de ces hypothèses, ainsi que de l'impact que leur inexactitude pourrait avoir sur nos résultats.

3.6.1 Distinction entre processus évolutifs neutres et sélectifs

Dans cette analyse, nous avons supposé que les substitutions asymétriques observées dans les régions centrales des introns sont dues à des processus évolutifs neutres, tels que des biais de mutation ou de réparation associés à la transcription. Avant de faire cette hypothèse simplificatrice, nous avons supprimé de l'analyse les séquences introniques qui sont le plus susceptibles de contenir des éléments fonctionnels pour l'épissage. Les extremités des introns sont connues pour contenir de telles séquences fonctionnelles (Black, 2003) ; par conséquent, nous avons décidé d'enlever de l'étude les fragments de 500 paires de bases situés à chaque extremité des introns. La taille des fragments à supprimer a été choisie sur la base des résultats présentés par Touchon *et al.* (2003). Il faut noter que les connaissances actuelles sur les éléments fonctionnels impliqués dans l'épissage rendent possible une analyse plus fine (Yeo *et al.* (2007), L. D. Hurst, communication personnelle).

Même en ayant pris ces précautions, il faut noter que nous n'avons pas la garantie absolue que l'asymétrie des substitutions est rééllement due à un processus neutre. Il est en effet possible d'imaginer un scénario purement sélectif, selon lequel l'asymétrie de composition des introns serait avantageuse car elle pourrait permettre la discrimination entre exons et introns au cours de l'épissage (hypothèse suggérée par Zhang *et al.* (2008), par exemple). L'analyse des patrons de substitutions n'est pas suffisante pour écarter ce type de scénario. La solution qui nous paraît la plus adéquate pour résoudre ce problème est d'analyser de manière simultanée le patron de polymorphisme et celui de substitution. En effet, si des processus neutres sont à l'origine de l'asymétrie de composition, cela devrait se voir dans le patron de polymorphisme : les mutations de $A \rightarrow G$ devraient être plus fréquentes que celles de $T \rightarrow C$, par exemple. Si par contre l'asymétrie est exclusivement générée par des processus sélectifs, cela devrait se réfléter uniquement dans la probabilité de fixation de mutations, et non pas dans leur fréquence d'apparition. Les données de polymorphisme actuellement disponibles pour le génome humain (The International HapMap Consortium, 2007) vont permettre de

réaliser cette analyse, que nous présentons comme une perspective de ce travail.

3.6.2 Niveaux d'expression des gènes dans la lignée germinale

Une autre hypothèse forte que nous avons posée ici est que l'étendue d'expression des gènes peut être utilisée comme approximation de leur taux d'expression dans la lignée germinale. Cette hypothèse simplificatrice a été souvent utilisée dans la littérature (voir par exemple Duret (2002); Lercher *et al.* (2004)), et elle peut être justifiée grâce à plusieurs arguments.

Tout d'abord, il paraît raisonnable de supposer que les gènes tissu-spécifiques ne sont pas exprimés dans la lignée germinale, alors que les gènes de ménage (qui sont exprimés dans tous les tissus) devraient l'être. Une confirmation de cette intuition vient de l'analyse des gènes qui produisent des retropseudogènes (et qui sont donc *a fortiori* exprimés dans la lignée germinale) dans le génome humain : ces gènes-là sont en général exprimés dans de nombreux tissus (Gonçalves *et al.*, 2000). Cependant, il faut noter qu'il existe des preuves en faveur d'un phénomène de "hypertranscription" (qui toucherait presque tous les gènes, y compris ceux tissu-spécifiques) dans la lignée germinale mâle (Schmidt, 1996; Kleene, 2001).

Il faut également remarquer que pour la question qui nous intéresse ici il est important d'analyser une mesure d'expression des gènes qui soit relativement stable dans le temps. En effet, pour qu'il existe une corrélation entre le niveau d'expression des gènes et les biais mutationnels provoqués par la transcription, ce niveau d'expression doit rester suffisamment stable dans le temps pour permettre l'accumulation graduelle des mutations. Or, des études récentes indiquent que le taux d'expression évolue à une vitesse importante : ainsi, il existe une forte divergence de niveau d'expression entre l'homme et le chimpanzé, bien que la séparation de ces deux espèces soit relativement récente (Khaitovich *et al.*, 2006). La vitesse d'évolution du niveau d'expression des gènes semble être particulièrement élevée pour la lignée germinale mâle, peut-être à cause de facteurs comme la compétition entre spermatozoïdes (Khaitovich *et al.*, 2006). Par opposition, l'étendue d'expression des gènes nous paraît intuitivement plus stable d'un point de vue évolutif. A ce jour, nous ne connaissons pas d'étude qui s'intéresse à la stabilité évolutive de l'étendue de l'expression ; cet argument intuitif est donc à prendre avec précaution.

3.6.3 La distinction entre réplication et transcription est-elle possible ?

Comme discuté précédemment, la co-orientation entre réplication et transcription et la distribution des gènes de ménage le long des domaines de réplication

sont d'importants facteurs confondants pour nos analyses. La séparation des effets de la réplication et de la transcription est difficile, sinon impossible à faire (du moins pour l'instant). A l'heure actuelle, nous ne disposons pas de données expérimentales à l'échelle du génome entier pour les positions des origines de réplication. L'inférence *in silico* des positions des origines de réplication représente certes une avancée importante dans ce sens, mais elle ne permet pas à ce jour d'identifier la totalité des origines de réplication du génome humain (Touchon *et al.*, 2005; Huvet *et al.*, 2007). Lorsque les origines de réplication seront déterminées expérimentalement, il sera important de réexaminer l'organisation des gènes protéiques par rapport à leurs positions. Si les phénomènes de co-orientation entre réplication et transcription et la distribution biaisée des gènes autour des origines sont valables à l'échelle du génome entier, le découplage entre les deux sources d'asymétrie sera impossible.

A la lumière de ces réflections, il est important de noter explicitement que l'hypothèse que nous proposons pour expliquer l'asymétrie des substitutions (et la dépendance du contexte), c'est à dire l'existence d'un biais de réparation associée à la transcription, n'est qu'une hypothèse parmi d'autres possibles.

3.6.4 D'autres facteurs confondants potentiels

Il peut paraître surprenant que cette analyse qui porte sur la variabilité des patrons de substitutions dans le génome humain ne fasse pas mention de l'impact de la recombinaison, ou de celui du taux de G et C local. En effet, ces deux facteurs sont connus pour avoir une forte influence sur les patrons (et sur les taux) de substitution (Duret et Arndt, 2008). La raison pour laquelle nous n'en avons pas discuté est simple : nous nous intéressons ici spécifiquement à l'asymétrie des substitutions. Pour cette question particulière, il est clair que le taux de G et C local ne peut pas influencer nos résultats : en effet, le taux de G et C est une mesure de composition intrinsèquement symétrique par rapport aux deux brins d'ADN, et donc il devrait avoir le même impact sur les deux changements complémentaires que nous étudions.

L'influence du taux de recombinaison sur l'asymétrie des substitutions est plus difficile à évaluer : certains auteurs ont évoqué la possibilité que la conversion génique biaisée soit asymétrique par rapport aux deux brins d'ADN dans les régions transcrites (Webster et Smith, 2004). Cependant, le mécanisme biologique à travers lequel pourrait avoir lieu cette conversion génique asymétrique n'est pour l'instant qu'une spéculation. Il faut également noter que cette étude reposait sur des données assez partielles, et ses conclusions devraient être réévaluées à la lumière des informations que nous avons à ce jour, notamment par l'analyse les données de polymorphisme à l'échelle du génome entier. Par souci de simplicité, ici nous avons décidé de ne pas prendre en compte ce potentiel effet de la recombinaison sur l'asymétrie de substitution. Nous considérons toutefois

qu'il sera nécessaire d'en discuter dans une future étude plus détaillée.

3.7 Conclusion et perspectives

L'analyse de la variation des taux de substitution en fonction du patron d'expression des gènes nous permet de suggérer que l'asymétrie des taux de transition $A \rightarrow G$ et $T \rightarrow C$ dans les régions transcrites pourrait être due à un mécanisme de réparation couplée à la transcription. Nous avons également démontré que l'asymétrie de substitution associée à la réplication est dépendante du contexte nucléotidique. De plus, cette dépendance de contexte est très similaire à celle que l'on observe dans les régions transcrites. Ces résultats nous conduisent à proposer que l'asymétrie des substitutions $A \rightarrow G$ et $T \rightarrow C$ pourrait être provoquée par un mécanisme de réparation brin - spécifique, aussi bien pour la transcription que pour la réplication. Nos résultats sont également en faveur d'une influence directe du contexte nucléotidique $5'$ - $3'$ sur l'efficacité de la réparation de l'ADN.

Pour réaliser cette étude, nous avons été amenés à poser plusieurs hypothèses simplificatrices. Notamment, nous avons admis que le nombre de tissus somatiques dans lesquels un gène est exprimé est une bonne approximation de son niveau d'expression dans la lignée germinale. Il serait sans doute intéressant de valider cette analyse lorsque des estimations expérimentales directes seront disponibles pour l'expression dans la lignée germinale.

Dans ce chapitre nous avons étudié l'influence du contexte immédiat (nature des nucléotides voisins en $5'$ et $3'$) sur les patrons de substitutions asymétriques. Cependant, l'influence du contexte sur les patrons de substitution n'est pas restreinte aux seuls nucléotides voisins, mais s'étend sur une région plus large. Dans le chapitre suivant, nous nous intéresserons à l'effet de la composition locale en nucléotides sur le patron de substitution.

3.8 Matériel et méthodes

3.8.1 Données génomiques et alignements de séquence

Les génomes étudiés ici correspondent aux assemblages *hg*18, *panTro*2 et *rheMac*2 de l'homme, du chimpanzé et du macaque, respectivement. Les annotations des gènes protéiques ont été extraites d'Ensembl, version 49. Les alignements entre les trois espèces correspondant aux introns ont été extraits en utilisant les interfaces de requête Galaxy (Giardine *et al.*, 2005) et Génomicro (V. Lombard et L. Duret, communication personnelle). Les calculs des taux de substitutions asymétriques ont été réalisés uniquement sur la région centrale des introns ; nous avons supprimé les 500 paires de bases à chaque extrémité des introns. Nous avons choisi la taille des fragments à éliminer en nous basant sur

les résultats présentés par Touchon *et al.* (2004), qui suggèrent que dans le génome humain l'asymétrie de composition directement déterminée par les signaux d'épissage s'étend sur environ 500 pb.

3.8.2 Données d'expression

Nous avons utilisé les données d'expression obtenues avec la méthode SAGE (Velculescu *et al.*, 1995), disponibles dans la banque de données "Human Cancer Anatomy Project" (`http://cgap.nci.nih.gov/SAGE`). Nous avons extrait ici uniquement les bibliothèques correspondant aux tissus normaux (par opposition aux tissus cancéreux), qui contiennent plus de 20000 étiquettes. Nous disposons ainsi de 100 bibliothèques, correspondant à 26 tissus. Un gène est considéré être exprimé dans un tissu, si l'étiquette correspondante est présente dans au moins une bibliothèque associée à ce tissu, quelle que soit sa fréquence observée.

Nombre de tissus	Gènes	Nombre de tissus	Gènes
0	492	14	468
1	597	15	504
2	588	16	524
3	629	17	494
4	585	18	450
5	597	19	470
6	572	20	453
7	520	21	489
8	554	22	472
9	539	23	381
10	490	24	415
11	488	25	377
12	466	26	521
13	519		

TAB. 3.2 – Distribution des patrons d'expression des gènes dans les bibliothèques SAGE.

Cette banque de données donne la correspondance entre chaque étiquette SAGE et un identifiant UniGene. Pour chaque gène protéique, nous avons extrait son identifiant UniGene en balayant les annotations fournies par Ensembl, version 49. Le jeu de données est constitué de 13654 gènes pour lesquels un identifiant UniGene et une étiquette SAGE est disponibles. Pour tracer les graphiques des

taux de substitution en fonction du patron d'expression, les substitutions ont été regroupées pour les gènes qui sont exprimés dans un même nombre de tissus.

3.8.3 Origines de réplication

Les positions des origines de réplication utilisées ici sont celles publiées par Huvet *et al.* (2007). Puisque dans le génome humain la terminaison de la réplication semble avoir lieu de manière aléatoire, la définition des fragments correspondant aux brins avancé et tardif n'est pas triviale. Nous avons décidé de restreindre notre étude aux régions qui se trouvent dans le voisinage immédiat des origines de réplication, pour avoir plus de confiance dans la définition des brins avancé et tardif. Ainsi, nous considérons que les fragments de 20 kilobases qui se trouvent en $5'$ des origines correspondent au brin tardif, et que les fragments de même taille en $3'$ des origines correspondent au brin avancé. Pour éviter que l'asymétrie de substitution associée à la transcription ne soit un facteur confondant, nous avons restreint notre jeu de données aux origines de réplication qui se trouvent dans des régions intergéniques, à une distance d'au moins 50 kilobases par rapport au plus proche gène. Avec cette restriction, le jeu de données analysé est restreint à 170 origines de réplication.

Ces coordonnées sont données par rapport à l'assemblage $hg17$ du génome humain. Les annotations des gènes protéiques pour cette version du génome ont été extraites de la banque Ensembl, version 37. Pour l'inférence des substitutions, nous avons utilisé les alignements entre les assemblages $hg17$, $panTro1$ et $rheMac1$, extraits de la base de données Génomicro, version 1 (V. Lombard et L. Duret, communication personnelle).

3.8.4 Inférence des patrons de substitution

L'inférence des patrons de substitution a été faite avec une méthode de parcimonie, comme décrit par Meunier et Duret (2004). Le contexte nucléotidique des substitutions est pris en compte. Pour compter une substitution dans un contexte XNY, il faut que les trois bases en $5'$ de l'homme, du chimpanzé et du macaque soient égales à X, et de même pour les trois bases voisines en $3'$. Les sites CpG sont définis par les voisinages CNY et XNG ; les sites non-CpG sont ceux pour lesquels $X \neq C$ et $Y \neq G$. Les taux de substitution présentés dans les graphiques sont ceux des sites non-CpG. Les taux de substitution dans les régions transcrites sont calculés toujours par rapport au brin codant du gène. Pour l'étude des origines de réplication, le calcul est fait par rapport au brin avancé. Les éléments répétés sont masqués lors de l'inférence des substitutions.

Chapitre 4

Patrons d'évolution asymétrique dans les séquences répétées

4.1 Introduction

La relation entre le patron de substitution et la composition en nucléotides d'une séquence d'ADN peut être facilement comprise comme une relation unidirectionnelle, de type cause-effet. La variation régionale dans le patron de substitution, par exemple comme une conséquence de la variation du niveau de recombinaison le long des chromosomes, peut produire une variation régionale similaire dans la composition en nucléotides (Meunier et Duret, 2004).

Il est certes moins intuitif, mais tout aussi correct, que la composition locale en nucléotides peut également avoir un effet direct sur le patron de substitution. A une échelle fine, la nature des nucléotides voisins en $5'$ et $3'$ a une influence importante sur le patron de substitution ponctuelle (*cf.* chapitre 3). A une plus grande échelle, nous savons à ce jour que le patron de substitution peut être influencé par le contenu en guanine et cytosine dans la région avoisinante. Ainsi, l'efficacité de la réparation des mésappariements semble être plus grande dans les régions riches en guanine et cytosine, peut-être à cause de la plus grande stabilité de la double hélice dans ces zones (Radman et Wagner, 1986; Jones *et al.*, 1987).

On a également proposé que le contenu local en G et C pourrait affecter le taux de mutation spontanée de la cytosine (Fryxell et Zuckerkandl, 2000), et l'explication invoquée est aussi en rapport avec la stabilité de la double hélice.

La molécule d'ADN est une structure très dynamique. Les liaisons hydrogène qui assurent l'appariement des deux brins sont dissociées localement, pour permettre le déroulement des processus tels que la réplication et la transcription. Cette dénaturation de l'ADN peut aussi se faire spontanément, par le mécanisme

qui est appelé "respiration" de l'ADN (Fryxell et Zuckerkandl, 2000). Les régions à fort taux de guanine et cytosine sont moins susceptibles à la dénaturation, car l'appariement de ces nucléotides se fait par trois liaisons hydrogène, au lieu de deux pour l'adénine et la thymine. Nous savons également que la plupart des mutations de $C \rightarrow T$ interviennent à cause du phénomène de désamination de la cytosine, qui arrive très fréquemment sur les molécules d'ADN à l'état simple brin (Frederico *et al.*, 1990). Il s'ensuit que les régions riches en G et C devraient présenter un plus faible taux de désamination de la cytosine (et donc de mutations de $C \rightarrow T$), car elles se trouvent moins souvent à l'état simple brin que les régions riches en A et T (Fryxell et Zuckerkandl, 2000). C'est effectivement ce que l'on observe dans le génome humain (*cf.* figure 4.1).

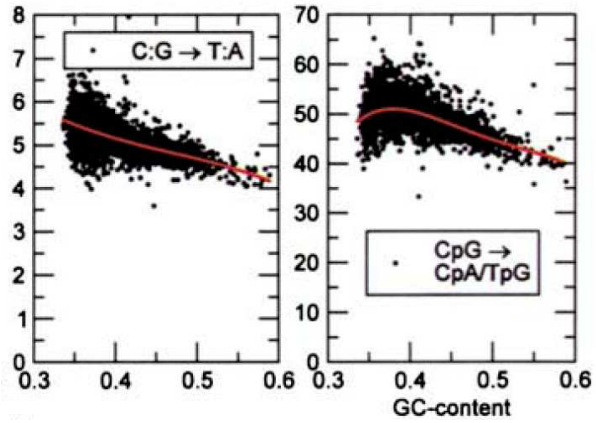

FIG. 4.1 – Variation du taux de substitution de $C \rightarrow T$ selon le contenu local en G et C. A gauche : le taux de substitution hors des sites CpG, à droite : le taux de substitution dans les dinucléotides CpG. D'après Arndt *et al.* (2005).

La variabilité régionale des taux de substitution et la corrélation avec la composition en nucléotides sont des sujets qui ont été étudiés de manière intensive au cours des dernières années (Smith *et al.*, 2002; Meunier et Duret, 2004; Arndt *et al.*, 2005; Duret et Arndt, 2008). Il faut cependant remarquer que ces études se sont intéressées uniquement à la relation entre le taux de G et C des séquences et le patron de substitution. Or, la conformation de la molécule d'ADN, qui peut influencer son mode d'évolution, n'est pas influencée uniquement par le contenu en guanine et cytosine. La *complexité* de la séquence peut également avoir un impact sur sa structure tridimensionnelle.

Les génomes eucaryotes sont parsemés de séquences répétées de très faible complexité, et qui possèdent une composition en nucléotides extrêmement biaisée : les microsatellites. Ces séquences sont connues pour générer des conformations inhabituelles de l'ADN, notamment des hélices de type Z (Hamada et Kakunaga, 1982). Pour mieux comprendre l'influence de la composition locale en

nucléotides sur le processus évolutifs, il serait intéressant de déterminer quel est le mode d'évolution des microsatelllites. C'est à cette question que nous essayerons de répondre dans le présent chapitre.

4.1.1 Les microsatellites : des séquences à composition extrême

Les microsatellites sont des répétitions en tandem d'un court motif (entre 1 et 6 nucléotides). Par exemple, les séquences $AAAAAAAAAA$ et $ACACACACAC$ peuvent être appelées des microsatellites, la première étant la répétition d'un motif A et la deuxième étant la répétition d'un motif AC. On dit que ces séquences sont à très faible complexité, car leur structure est en effet très régulière, et elles sont aussi caractérisées par une composition en nucléotides qui est très biaisée.

Ce type de séquence répétée a été découvert il y a presque 30 ans (Hamada et Kakunaga, 1982), et il est vite devenu évident que les microsatellites sont présents de manière ubiquitaire dans les génomes eucaryotes (Hamada *et al.*, 1982; Tautz et Renz, 1984). Les microsatellites sont aujourd'hui très bien connus, et cela pour une raison pratique : ces séquences sont très fréquemment utilisées en tant que marqueur moléculaire pour des études de génétique.

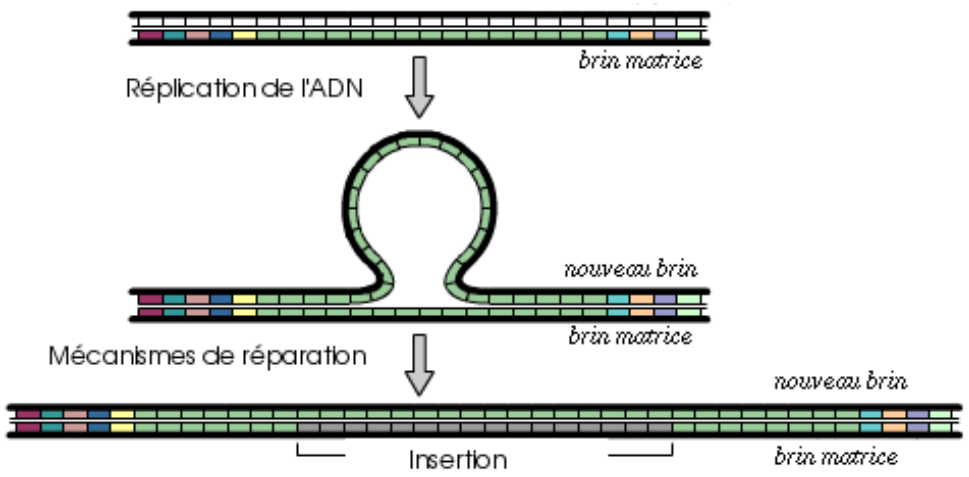

FIG. 4.2 – Schéma décrivant le glissement de l'ADN polymérase au moment de la réplication des séquences répétées.

La caractéristique des microsatellites qui les rend très utiles en tant que marqueur moléculaire est le niveau élevé de variabilité de longueur qu'ils présentent. Par conséquence, le mode d'évolution de la longueur des microsatellites a été très bien étudié, aussi bien de manière expérimentale que d'un point de vue théorique

(*cf.* Schlotterer (2000); Ellegren (2004); Buschiazzo et Gemmell (2006) pour revue). Il est maintenant admis que la principale source des mutations qui affectent la longueur des microsatellites est le mécanisme de glissement de la polymérase au moment de la réplication (*cf.* figure 4.2), qui a fréquemment comme conséquence l'apparition d'insertion ou de délétions dans les séquences répétées (Streisinger *et al.*, 1966; Levinson et Gutman, 1987).

4.1.2 Mutations ponctuelles dans les microsatellites

Les modèles couramment acceptés pour l'évolution des microsatellites sont fondés sur l'équilibre de deux mécanismes évolutifs : les événements de glissement de polymérase au moment de la réplication, ainsi que des mutations ponctuelles des nucléotides (Kruglyak *et al.*, 1998). Si le premier processus a été étudié de manière exhaustive au cours des dernières années, à ce jour la connaissance du patron de mutation ponctuelle dans les microsatellites reste très incomplète. Pourtant, des microsatellites qui présentent des changements ponctuels de nucléotides ont été découverts très tôt dans les génomes des eucaryotes, et il est maintenant admis qu'ils sont très fréquents (Estoup *et al.*, 1993). Un exemple de microsatellite qui présente un changement ponctuel de nucléotide est $ACACACATACACAC$; ce type de microsatellite est en général appelé "imparfait" ou "interrompu".

Les microsatellites imparfaits sont caractérisés par des réductions significatives de la variabilité en longueur, par rapport aux répétitions parfaites (Chung *et al.*, 1993; Blanquer-Maumont et Crouau-Roy, 1995; Angers et Bernatchez, 1997). Récemment, il a été démontré que la nature du nucléotide qui interrompt la séquence répétée peut moduler le gain en stabilité du microsatellite (Boyer *et al.*, 2008). Puisque les microsatellites imparfaits sont moins susceptibles au phénomène de glissement de l'ADN polymérase, l'accumulation des changements nucléotidiques peut mener à la dégénération et finalement à "la mort" du microsatellite (Taylor *et al.*, 1999). Néanmoins, les interruptions de la séquence répétée peuvent également être supprimées par le glissement de l'ADN polymérase, qui rétablit donc le microsatellite parfait (Harr *et al.*, 2002).

Les premières études qui ont discuté l'existence de substitutions ponctuelles dans les microsatellites se sont intéressées à leur distribution le long de la séquence, et n'ont pas fourni d'information sur la direction ou le taux des substitutions. Même pour ce qui concerne la distribution des substitutions sur la longueur de la séquence répétée, il n'existe pas à ce jour un consensus. Il a été d'abord suggéré que les substitutions ne sont pas distribuées de façon uniforme, mais se concentrent vers les extrémités des microsatellites (Brohede et Ellegren, 1999; Shepherd et Lambert, 2005). Par contre, il n'est pas clair si cette conclusion est valable pour tous les locus et pour toutes les espèces, car une étude des micro-

satellites situés sur les chromosomes sexuels humains a signalé une distribution homogène des substitutions (Balaresque *et al.*, 2003).

Les facteurs qui déterminent l'évolution de la longueur des microsatellites ont déjà été étudiés à l'échelle des génomes complets (Kelkar *et al.*, 2008). Par contre, les connaissances très partielles que nous possédons à ce jour sur le patron de mutation ponctuelle viennent d'études basées sur de faibles nombres de locus. L'absence d'un consensus clair n'est donc pas surprenante, et de nombreuses questions concernant cet aspect de l'évolution des microsatellites restent à présent ouvertes. L'objectif de ce chapitre est d'effectuer une étude exhaustive du patron de substitution nucléotidique dans les microsatellites du génome humain. Un de nos objectifs est évidemment de contribuer aux connaissances actuelles du mode d'évolution des microsatellites. Dans une perspective plus large, nous souhaitons également étudier l'influence de la composition locale en nucléotides sur le patron d'évolution des séquences. Un des aspects qui nous intéressent est évidemment l'asymétrie de patrons de substitution par rapport aux deux brins d'ADN.

4.2 Questions posées

La première question à laquelle nous souhaitons répondre est : comment peut-on estimer le patron de substitution dans les microsatellites ? La réponse n'est pas triviale, et c'est peut-être en partie pour cela que cet aspect du mode d'évolution des microsatellites a été ignoré jusqu'à présent. La manière habituellement employée pour inférer le patron de substitution dans des séquences d'ADN est la comparaison de séquences homologues : le même locus est séquencé dans plusieurs espèces, ensuite les séquences homologues sont alignées et l'inférence des substitutions se fait en fonction de la topologie de l'arbre phylogénétique des espèces prises en compte. Dans le cas des microsatellites, la composition en nucléotides est fortement biaisée, et la nature répétée de ces séquences rend l'alignement des sites homologue difficile, sinon impossible (*cf.* figure 4.3). L'approche comparative n'est donc pas directement applicable pour ce type de séquence.

Pour inférer le patron de substitution, nous baserons notre raisonnement sur l'observation suivante : du fait de la nature répétitive des séquences de microsatellites, la présence d'interruptions dans la séquence d'ADN est dans la plupart des cas synonyme avec l'apparition de mutations ponctuelles. Nous allons donc utiliser la distribution des imperfections dans les microsatellites pour estimer le patron de substitution des nucléotides.

Une deuxième question qui nous intéresse est de savoir si la composition extrême en nucléotides des microsatellites peut influencer le patron de substitution ponctuelle. Pour étudier cet aspect, nous comparerons les patrons de substitution entre microsatellites et régions non-répétées. L'estimation de ce deuxième patron

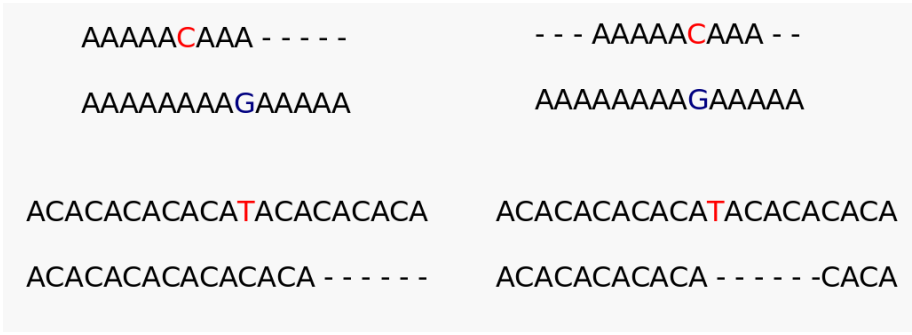

FIG. 4.3 – Schéma décrivant la difficulté de l'alignement des séquences de microsatellites. Dans le premier exemple, on veut aligner deux microsatellites homologues de type A_n, les deux étant imparfaits, mais présentant des substitutions nucléotidiques différentes. Les algorithmes d'alignement de séquences auront tendance à positionner face à face les deux changements nucléotidiques observés dans les microsatellites, pour maximiser le score de l'alignement. Dans le deuxième exemple, les deux microsatellites sont de type AC_n, et l'un d'entre eux présente une substitution. Les algorithmes d'alignement vont positionner préférentiellement cette substitution devant un "gap". Dans les deux cas, le scénario évolutif inféré peut être biaisé par la méthode d'alignement.

de substitution pourra se faire avec l'approche comparative classique, grâce à la disponibilité des séquences génomiques complètes pour l'homme, le chimpanzé et le macaque.

Enfin, nous essayerons de déterminer si le patron de substitution dans les microsatellites est influencé par des facteurs régionaux autres que la composition en nucléotides. Notamment, nous souhaitons savoir si l'évolution des microsatellites qui sont présents dans des régions transcrites est aussi caractérisée par une asymétrie des substitutions par rapport aux deux brins d'ADN.

4.3 Comptage des substitutions dans les microsatellites

L'idée d'estimer un patron de substitution en analysant une seule séquence génomique peut paraître extravagante, du moins à première vue. Néanmoins, la nature répétitive des microsatellites nous permet d'effectuer ce genre d'inférence. Ces séquences répétées peuvent présenter des interruptions, qui sont dans la plupart des cas équivalentes à des mutations ponctuelles. Par exemple, le scénario le plus parcimonieux pour expliquer l'existence du microsatellite imparfait

$ACACACATACACAC$ est que l'état ancestral était un microsatellite parfait d'unité AC, et qu'une seule mutation de $C \rightarrow T$ a eu lieu dans la quatrième unité de répétition. Avec ce type de raisonnement, l'analyse de la distribution des imperfections pour chaque type de microsatellite peut nous donner une estimation du patron de substitution nucléotidique dans ces séquences.

Il est important de noter qu'ici on utilise le terme "substitution" pour désigner tous les changements de nucléotides qui sont observés dans un microsatellite. Evidemment, il est impossible de distinguer entre mutations polymorphes et fixées en analysant un seul génome, donc le terme "substitution" ne doit pas être compris comme "mutation fixée dans la population".

4.3.1 Pièges du raisonnement par parcimonie

Comme tout raisonnement par parcimonie, le nôtre présente des points sensibles. Le mécanisme de glissement de la polymérase, qui détermine l'évolution de la longueur des microsatellites, peut dupliquer ou supprimer des unités de répétition imparfaites. Par exemple, pour expliquer l'existence du microsatellite $AAAAGAAGAAGAAAA$, on peut imaginer plusieurs scénarii évolutifs. D'une part, on peut supposer que l'état ancestral était un microsatellite parfait A_n, et que trois substitutions indépendantes de $A \rightarrow G$ ont eu lieu dans les unités centrales. D'autre part, on peut imaginer qu'une seule substitution de $A \rightarrow G$ a eu lieu initialement, et que le motif AAG au centre de la répétition a ensuite été multiplié par glissement de la polymérase. Le deuxième scénario semble plus raisonnable, et l'inférence du patron de substitution serait erronée si on appliquait le premier raisonnement. Pour éviter cette source de biais, nous avons enlevé de notre analyse tous les microsatellites "composés", c'est à dire ceux qui contiennent un mélange de deux unités de répétition (A et AAG dans l'exemple cité). Nous avons également effectué une analyse séparée en incluant uniquement les microsatellites qui ne présentent qu'un seul changement de nucléotide. Les résultats obtenus ne sont pas significativement différents avec les deux approches.

Cette correction permet de réduire le biais introduit par la duplication des unités imparfaites, mais elle ne peut pas résoudre le problème de l'élimination des unités imparfaites par le glissement de la polymérase. Ce problème semble être sans issue, mais il faut remarquer qu'il ne peut biaiser l'inférence du patron de substitution que si l'élimination des unités imparfaites ne se fait pas de manière aléatoire. Il n'est pas clair pour l'instant quel est l'impact de ce processus sur la distribution des imperfections dans les microsatellites.

Le fait de prendre en compte les substitutions qui ont lieu aux extrémités des microsatellites peut aussi poser des problèmes. Plusieurs études ont signalé que dans les microsatellites les substitutions ont lieu plus fréquemment aux extrémi-

tés (Brohede et Ellegren, 1999; Shepherd et Lambert, 2005). Cela pourrait en effet être une caractéristique réelle des séquences répétées, mais il est tout aussi probable que les frontières de la région répétée sont difficiles à définir. Pour éviter cette situation, nous avons pris en compte uniquement les changements qui ont eu lieu dans la région centrale des microsatellites, en ignorant les deux unités de répétition à chaque extrémité. De même, pour le comptage des substitutions nous avons imposé le contexte soit conservé : les deux unités de répétition voisines des deux côtés de la substitution doivent être parfaites.

Enfin, une autre source possible de biais est l'association entre microsatellites et éléments transposables. En effet, certains microsatellites imparfaits, notamment ceux de type A_n, sont souvent trouvés à l'intérieur des éléments *Alu* et *LINE*. Le fait de compter ces substitutions comme des événements indépendants pourrait mener à une estimation erronée du patron de substitution, car l'imperfection était vraisemblablement présente dans l'élément transposable ancestral, et a été ensuite multipliée par retrotransposition. Il est donc important d'exclure de l'analyse les microsatellites qui chevauchent des éléments transposables.

Il est important de noter que notre approche peut donner uniquement des estimations des taux *relatifs* de substitution. Pour calculer des taux absolus de substitution, l'échelle de temps doit être fixée de manière précise, et l'approche comparative peut fournir cette échelle de temps. Par contre, l'analyse des imperfections présentes dans les microsatellites ne permet pas d'estimer la période de temps au cours de laquelle ces substitutions ont eu lieu. Pour nous assurer néanmoins que nous analysons des événements évolutifs relativement récents, et pour minimiser la possibilité que des substitutions multiples aient lieu à un même site, nous avons pris en compte seulement des microsatellites qui présentent un haut niveau d'identité par rapport au microsatellite parfait correspondant (*cf.* Matériel et méthodes).

4.4 Distribution génomique des microsatellites

Nous avons identifié tous les microsatellites parfaits et imparfaits pour lesquels la longueur de l'unité de répétition est comprise entre 1 et 3. Les résultats que nous présentons ici concernent uniquement les microsatellites qui ne se superposent pas avec des éléments transposables, et qui se trouvent dans des régions intergéniques ou dans les introns.

Comme remarqué précédemment (Kelkar *et al.*, 2008), l'abondance des microsatellites dépend de la longueur de l'unité, c'est à dire que les microsatellites de type mononucléotide sont plus fréquents que les répétitions de di- et trinucléotides (*cf.* figure 4.4). La composition en bases de l'unité de répétition est aussi un facteur important dans leur fréquence d'apparition. Les microsatellites riches en guanines et cytosine sont relativement rares dans le génome humain, en particulier pour les motifs qui incluent le dinucléotide CpG.

Pour les microsatellites qui sont très peu fréquents dans le génome humain, le patron de substitution ne pourra pas être inféré avec précision. Par la suite, nous avons restreint notre étude aux répétitions d'unité A, AC, AG, AT, AAC et AAT, qui sont les plus abondants.

Les microsatellites sont des séquences qui ont une composition extrême en nucléotides. On peut se demander si les régions génomiques dans lesquelles ils apparaissent sont aussi caractérisées par des compositions biaisées. Pour répondre à cette question, nous avons calculé le taux de guanine et cytosine dans les régions flanquantes des microsatellites. La taille de la région flanquante a été fixée à 100 paires de bases. Les résultats sont présentés dans la figure 4.5. Nous remarquons que la composition des régions voisines des microsatellites n'est pas particulièrement biaisée. Les valeurs médianes des distributions se trouvent toujours au voisinage de 0.4, les valeurs les plus faibles étant rencontrées pour les microsatellites AT_n et AC_n.

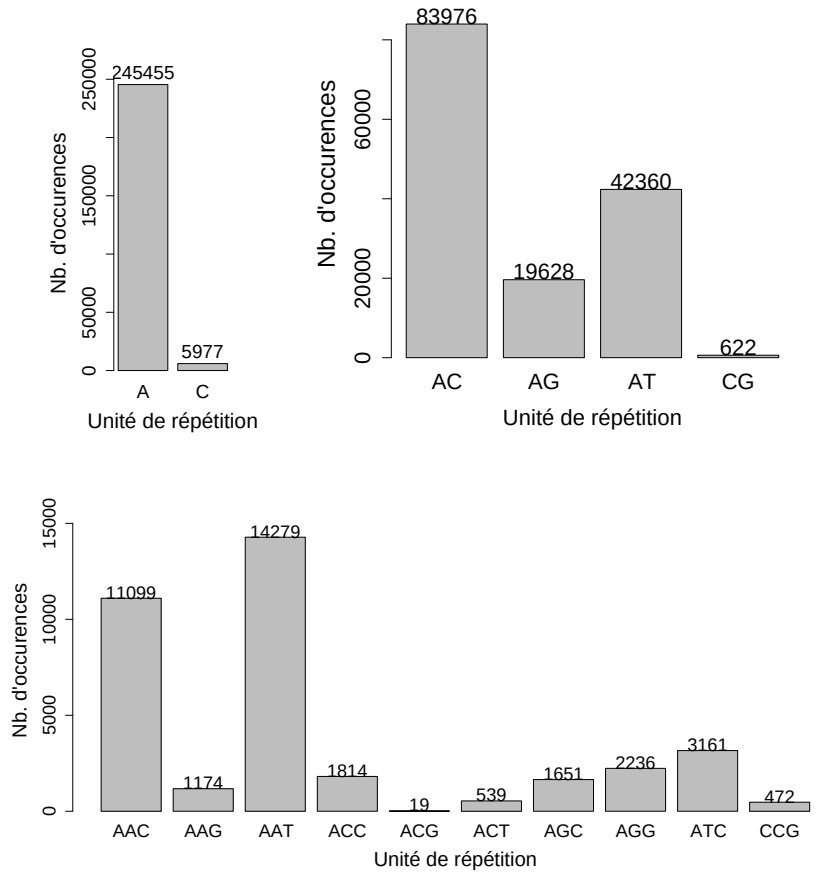

FIG. 4.4 – Distribution des microsatellites d'unité de longueur 1, 2 et 3 dans les régions non-codantes du génome humain. Il faut noter qu'ici les microsatellites complémentaires sont regroupés, nous ne faisons pas de distinction entre des répétitions de motifs AC et TG, par exemple. De même, aucune distinction n'est possible entre les microsatellites d'unités chevauchantes, par exemple AAC_n et ACA_n sont considérés être le même microsatellite.

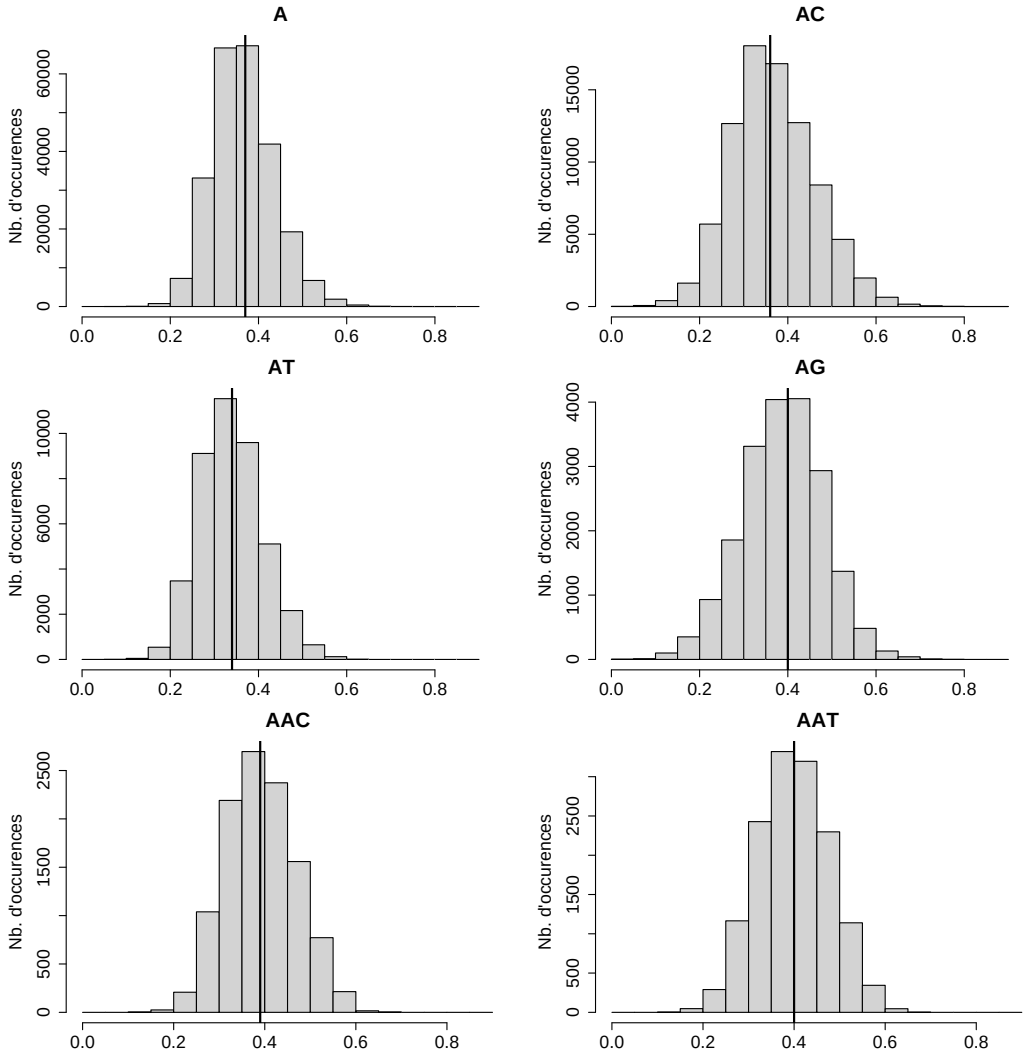

FIG. 4.5 – Contenu en G et C dans les régions flanquantes des microsatellites (50 paires de bases de chaque côté). Seulement les microsatellites les plus fréquents ont été représentés. La barre verticale représente la position de la médiane de la distribution.

4.5 Patrons de substitution dans les microsatellites

Les patrons de substitution nucléotidique dans les microsatellites humains sont présentés dans les figures 4.6 et 4.7. Pour cette partie de l'étude, nous avons calculé un patron de substitution symétrique, c'est à dire que les substitutions complémentaires (par exemple $A \rightarrow G$ et $T \rightarrow C$) sont regroupées. De même, aucune distinction n'est faite entre les microsatellites complémentaires (AC_n est

synonyme de TG_n). Le patron de substitution est dépendant du contexte, c'est à dire que chaque substitution est donnée en fonction des nucléotides voisins en $5'$ et $3'$. Dans le génome humain, les microsatellites les plus fréquents sont relativement riches en A et T, par conséquent nous avons pu estimer le patrons de substitution des nucléotides A/T avec plus de précision, et dans plus de contextes que celui des nucléotides G/C.

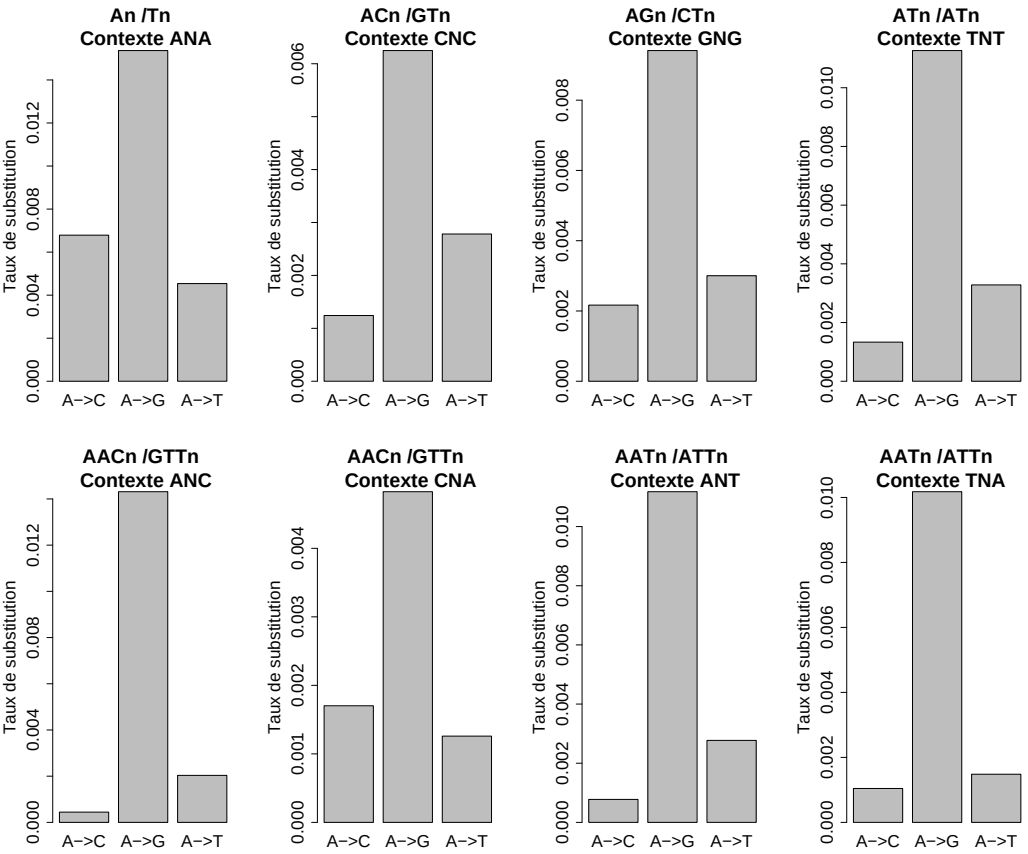

FIG. 4.6 – Patron de substitution des nucléotides A, dans les microsatellites A_n, AC_n, AG_n, AAC_n et AAT_n.

Nous pouvons remarquer que les patrons de substitutions varient selon le type de microsatellite. Par exemple, si on veut comparer les taux relatifs des transversions $A \rightarrow C$ et $A \rightarrow T$, le résultat ne sera pas le même selon l'unité de répétition du microsatellite (*cf.* figure 4.6 ; les comparaisons 2 à 2 par tests de χ^2 donnent des $p - values$ inférieures à $2 * 10^{-}10$). De même, le rapport entre les taux de transversions $C \rightarrow A$ et $C \rightarrow G$ n'est pas le même dans les microsatellites AC, AG et AAC. Les microsatellites semblent toutefois respecter une règle fondamentale du patron de substitution du génome humain, dans le

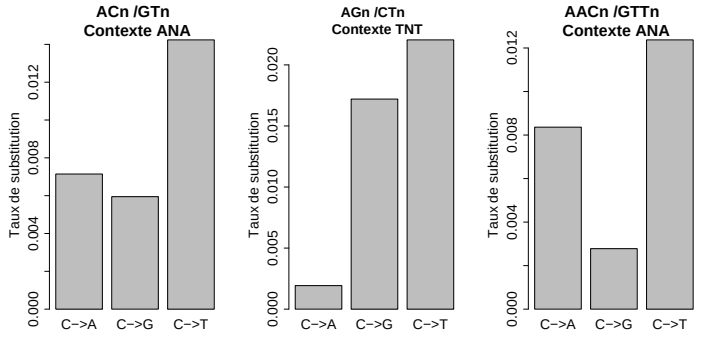

FIG. 4.7 – Patron de substitution des nucléotides C, dans les microsatellites AC_n, AG_n et AAC_n.

sens que les transitions sont beaucoup plus fréquentes que les transversions.

Pourquoi a-t-on cette variabilité du patron de substitution selon le type de microsatellite ? Une réponse possible est qu'il s'agit là uniquement de l'influence du contexte $5'$ - $3'$. Ce facteur ne semble pas être le seul en cause, car nous observons par exemple des patrons de substitution différents pour le nucléotide C dans le même contexte ANA, pour les microsatellites AC_n et AAC_n.

Nous pouvons aussi tester cette hypothèse en comparant les patrons de substitution entre les microsatellites et les régions non-répétées du génome humain, à contexte $5'$ - $3'$ égal. Nous présentons ces patrons de substitution dans les figures 4.8 et 4.9. Dans certains cas, les patrons de substitutions sont effectivement très similaires entre microsatellites et régions non-répétées. Par exemple, les substitutions qui ont lieu dans le contexte ANA sont similaires dans les microsatellites A_n et dans les régions non-répétées, dans le sens que l'on dans les deux cas $A \to T < A \to C < A \to G$. De même, les patrons de substitutions ont la même allure pour le contexte CNA, dans les microsatellites AAC_n et dans les régions non-répétées, car on a $A \to T < A \to C < A \to G$ pour les deux cas. (*Nota bene :* ici similaire ne veut pas dire statistiquement identique. Lorsque nous comparons les patrons de substitution avec des tests de type χ^2, nous rejetons l'hypothèse d'égalité des distributions avec des $p - values$ inférieures à 0.001.)

Dans la plupart des cas, les patrons de substitution sont différents entre microsatellites et régions non-répétées (la comparaison avec des tests de χ^2 donne des $p - values$ inférieures à 10^-10).

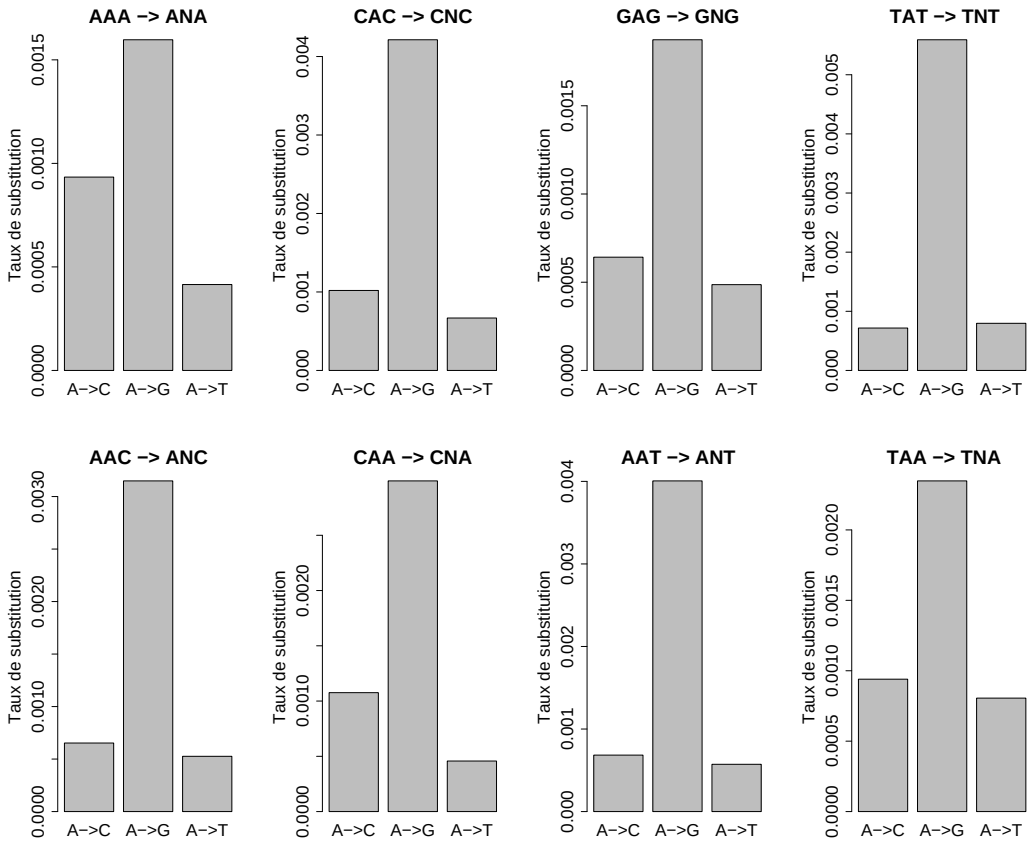

FIG. 4.8 – Patron de substitution des nucléotides A, dans les régions non-répétées, dans les contextes ANA, CNC, GNG, ANC, CNA, ANT, TNA. Ces contextes correspondent à ceux présents dans les microsatellites A_n, AC_n, AG_n, AAC_n et AAT_n. Cette figure est à comparer avec la figure 4.6.

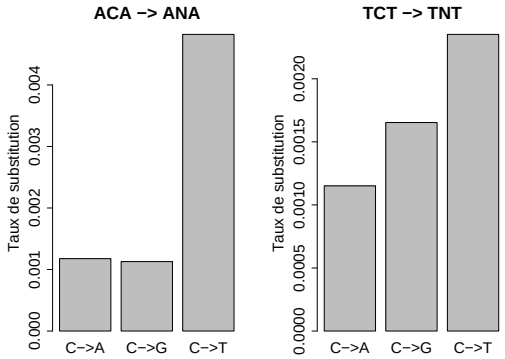

FIG. 4.9 – Patron de substitution des nucléotides C, dans les régions non-répétées, dans les contextes ANA et TNT. Ces contextes correspondent aux microsatellites AC_n, AG_n et AAC_n. Cette figure est à comparer avec la figure 4.7.

4.6 Variation du patron de substitution selon le contenu en G et C

Nous avons vu précédemment que la prise en compte du contexte immédiat (nucléotides voisins en $5'$ et en $3'$) ne suffit pas pour expliquer les différences du patron de substitution entre microsatellites et régions non-répétées. Nous avons voulu vérifier si ce désaccord peut être dû à des effets de voisinage au sens plus large, notamment à des différences de taux de G et C local.

Pour faire cela, nous devons d'abord définir la taille des fenêtres sur lesquelles on calculerait le taux de G et C local. A l'heure actuelle, nous ne connaissons pas à quelle échelle s'étend réellement l'influence du taux de G et C local sur le patron d'évolution. La motivation du choix sera ici uniquement technique. Si l'on choisit des tailles de fenêtre trop faibles, de l'ordre de 20-30 paires de bases, le voisinage des substitutions qui ont lieu dans les microsatellites se réduit en fait à la séquence de la répétition. La comparaison avec les régions non-répétées n'est alors pas toujours possible. Par exemple, pour avoir un terme de comparaison pour les substitutions qui ont lieu dans les microsatellites A_n ou AT_n, il faudrait trouver des régions à 0 % de guanine et cytosine, qui ne soient pas des microsatellites. Cela est évidemment impossible. Même pour les microsatellites dont le taux de G et C est moins biaisé (par exemple AC_n ou AG_n), la comparaison est difficile - car dans le génome humain il existe aussi peu de régions à 50 % de taux de G et C.

Nous avons donc choisi des tailles de fenêtres glissantes qui dépassent la longueur moyenne des microsatellites. Pour chaque imperfection observée dans un microsatellite, nous avons calculé la composition en G et C dans une fenêtre de 100 paires de bases centrée sur l'imperfection. La variation de la composition des régions flanquantes est présentée dans la figure 4.10. Puisque nos fenêtres glissantes incluent les microsatellites, il est normal d'observer que le taux de G et C autour des substitutions qui ont lieu dans les microsatellites AT_n, AAT_n et A_n est plus faible que celui calculé pour les microsatellites de type AG et AC.

Nous avons calculé le patron de substitution dans les régions non-répétées, en fonction du taux de G et C local, calculé comme pour les microsatellites dans des fenêtres de 100 pairs de bases. Les résultats sont présentés dans les figures 4.11, 4.12 et 4.13. Pour ces comparaisons nous tenons bien sûr compte aussi du contexte immédiat $5'$ - $3'$.

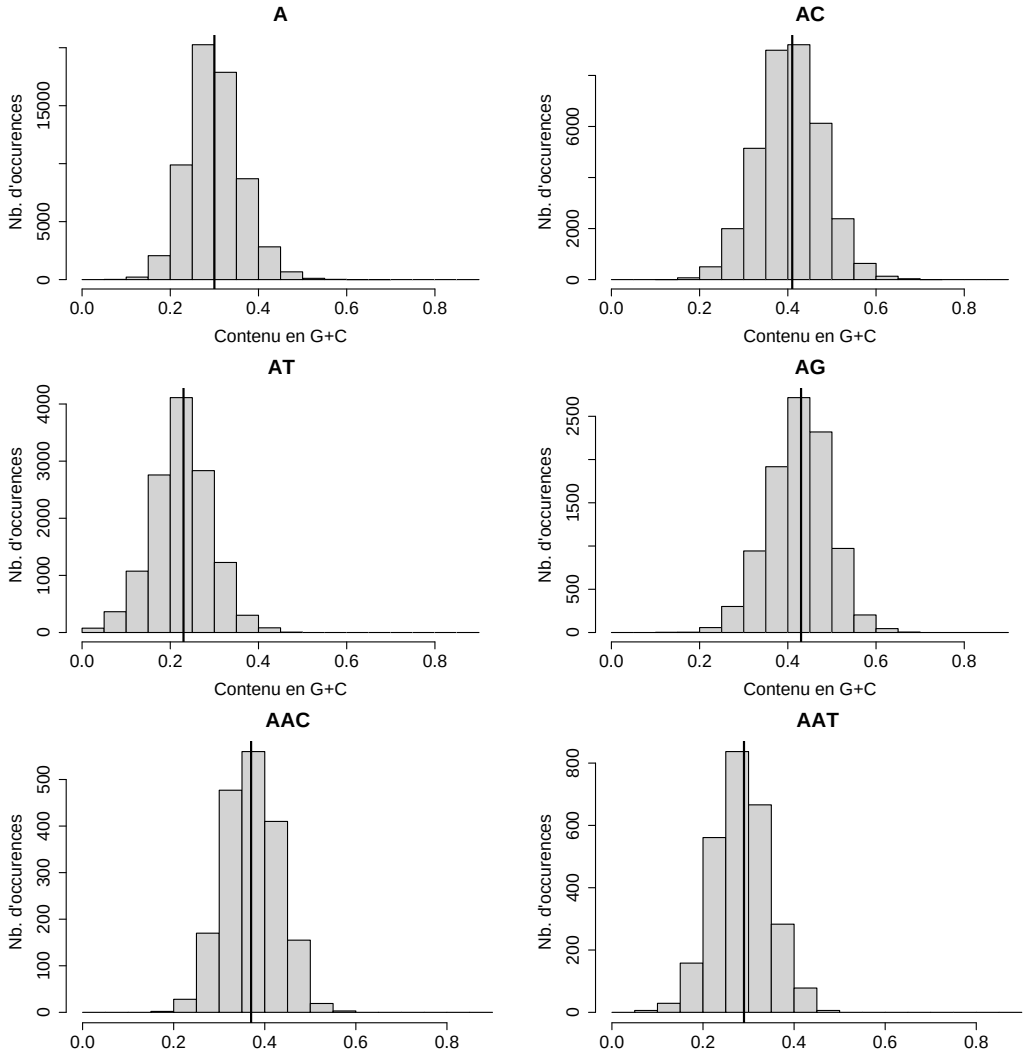

FIG. 4.10 – Contenu en G et C dans les régions flanquantes des microsatellites imparfaits. Les régions flanquantes sont centrées sur la substitution (50 paires de bases de chaque côté). Seulement les microsatellites les plus fréquents ont été représentés. La barre verticale représente la position de la médiane de la distribution.

Si le taux de G et C local suffit pour décrire le du mode d'évolution de séquence, on s'attend à ce que les patrons de substitutions soient similaires pour les régions répétées et non-répétées, lorsque ce taux de G et C est pris en compte.

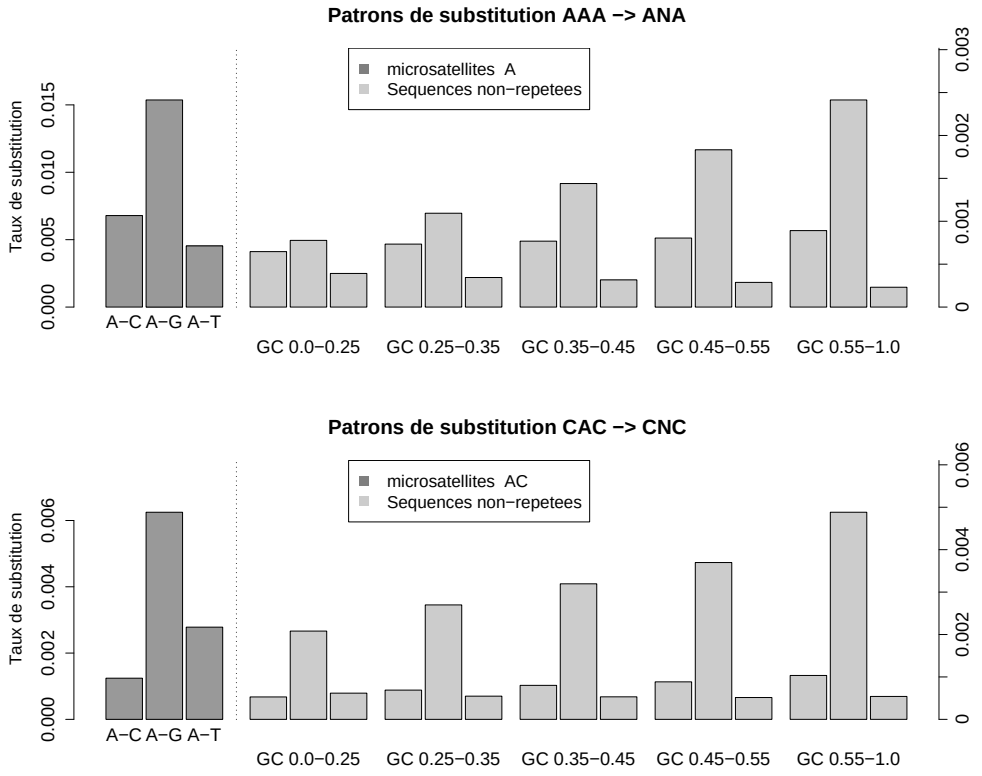

FIG. 4.11 – Patrons de substitution de A vers C, G et T dans les microsatellites A_n et AC_n et comparaison avec les régions non-répétées, en tenant compte du taux de G et C local et du contexte $5'$ - $3'$.

Nos résultats contredisent cette attendue, dans la plupart des cas. En effet, les patrons de substitutions sont souvent significativement différents (même si comparables) entre microsatellites et régions non-répétées, même si le contenu en G et C régional est pris en compte. Dans certains cas, la différence n'est pas très grande. Par exemple, le taux de G et C autour des substitutions qui ont lieu dans les microsatellites A_n varie entre 26 % et 34 %. Le patron de substitution calculé pour ce type de microsatellite est cependant plus similaire à celui observé pour des régions non-répétées dont le taux de G et C est compris entre 0.35 et 0.45 (*cf.* figures 4.11 et 4.14).

Dans d'autres situations, les différences peuvent être beaucoup plus flagrantes. Par exemple, le taux de G et C des régions autour des imperfections observées dans les microsatellites AC_n et AG_n est relativement élevé (au dessus de 0.35), mais le patron de substitution dans ces microsatellites est plus similaire à celui observé dans des régions non-répétées à très faible taux de G et C (*cf.* figures 4.11 et 4.14).

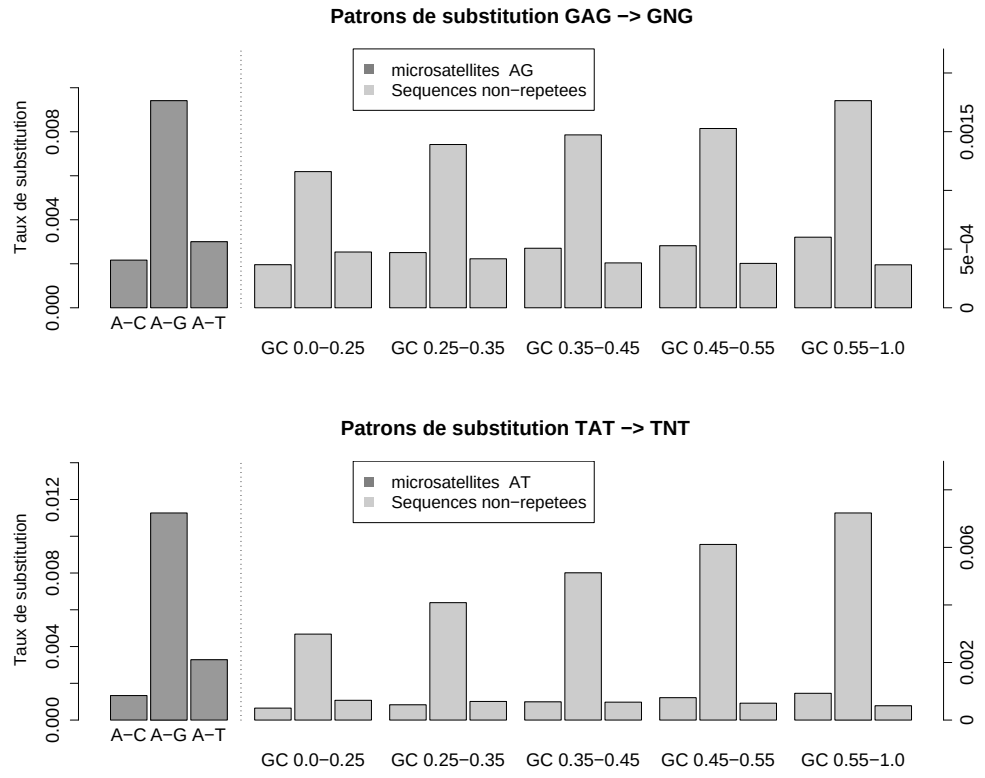

FIG. 4.12 – Patrons de substitution de A vers C, G et T dans les microsatellites AG_n et AT_n, et comparaison avec les régions non-répétées, en tenant compte du taux de G et C local et du contexte $5'$ - $3'$.

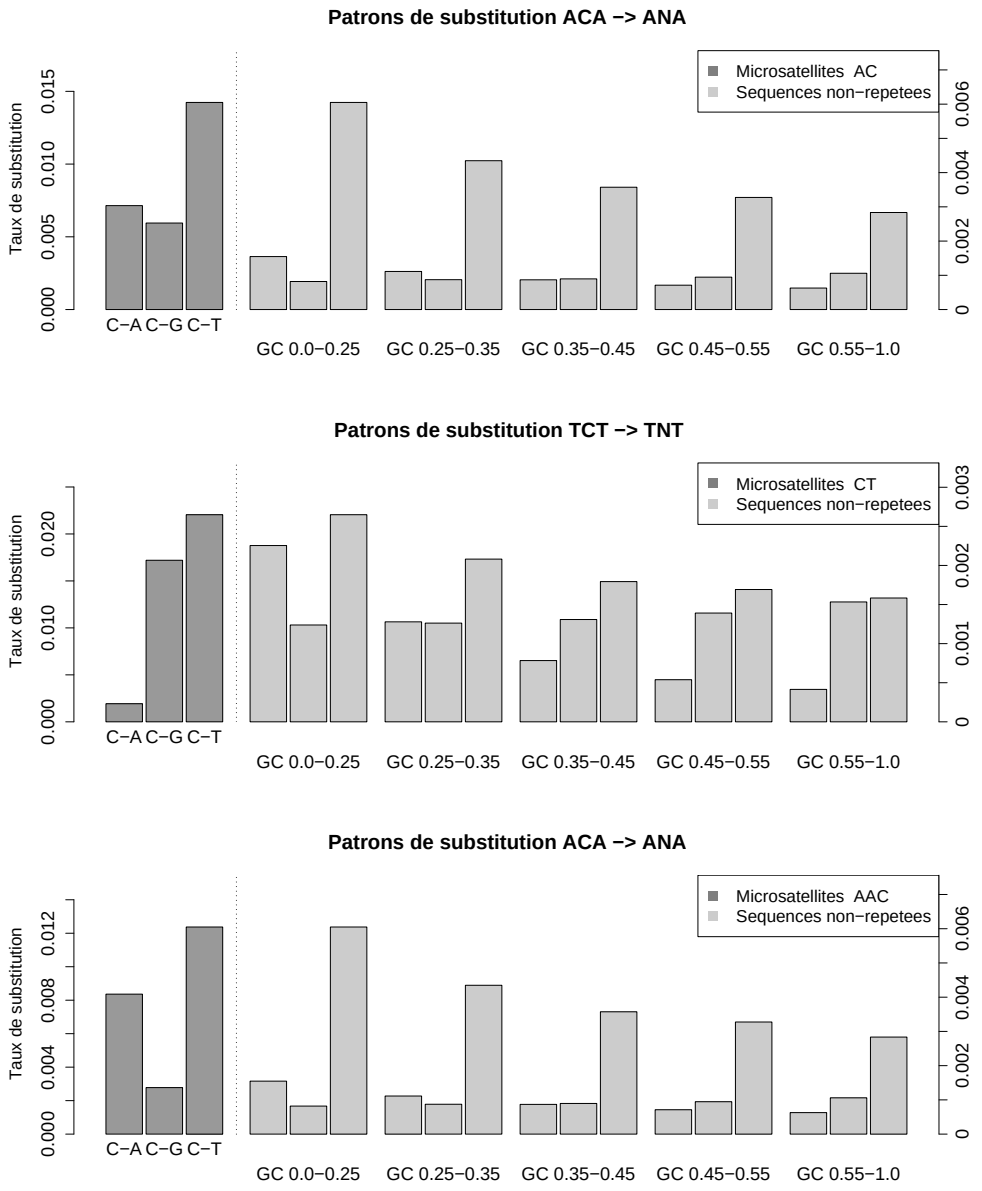

FIG. 4.13 – Patrons de substitution de C vers A, G et T dans les microsatellites AC_n, TG_n et AAC_n et comparaison avec les régions non-répétées, en tenant compte du taux de G et C local et du contexte $5'$ - $3'$.

Pour réaliser une comparaison globale entre microsatellites et régions non-répétées, nous avons effectué des analyses de correspondances sur les patrons de substitutions (*cf.* Matériel et Méthodes). Nous avons tout d'abord comparé les patrons de substitution du nucléotide A, pour les contextes $5' - 3'$ AAA, CAC, GAG et TAT (qui correspondent aux microsatellites A_n, AC_n, AG_n et AT_n). Pour les régions non-répétées, nous avons construit plusieurs classes selon le taux de G et C local (cf. figure 4.14). La carte factorielle présentée dans la figure 4.14 montre que la première source de variabilité dans le patron de substitution (qui correspond à l'axe horizontal) est représentée par le contexte $5' - 3'$. Le deuxième axe (sur la verticale) semble être lié à l'effet du taux de G et C local sur le patron de substitution. Pour les contextes CAC, GAG et TAT, nous remarquons que les patrons de substitutions des microsatellites présentent le plus de similarité avec les régions non-répétées à faible taux de G et C, alors que cela n'est pas le cas pour le contexte AAA.

Nous avons refait l'analyse pour les patrons de substitutions du nucléotide C, dans les contextes $5' - 3'$ ACA et TCT (microsatellites AC_n et AG_n). Le premier axe factoriel semble aussi correspondre au contexte $5' - 3'$, mais la séparation est moins nette que pour les substitutions du nucléotide A. Dans le contexte TCT, nous remarquons cette fois que le patron observé pour le microsatellite (AG_n) présente plus de similarité avec les régions non-répétées à fort taux de G et C.

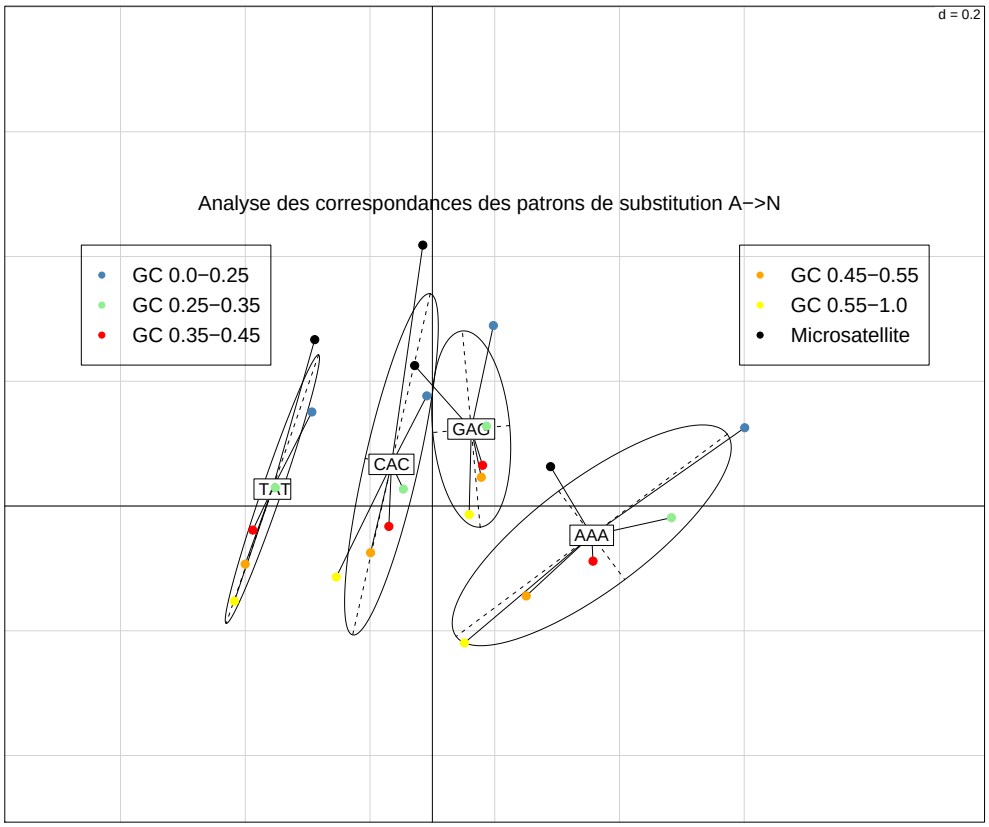

FIG. 4.14 – Carte factorielle pour l'analyse des correspondances des patrons de substitution de A vers C, G et T dans les microsatellites A_n, AC_n, AG_n et AT_n (contextes AAA, CAC, GAG, TAT) et dans les régions non-répétées, selon le taux de G et C local et du contexte $5'$ - $3'$. L'axe des X correspond à la coordonnée sur le premier axe factoriel (qui totalise 87 % de la variabilité totale des données), l'axe des Y représente la coordonnée sur le deuxième axe factoriel (12% de variabilité). Les classes de taux de G et C pour les régions non-répétées sont représentées en couleur ; les points noirs correspondent aux microsatellites.

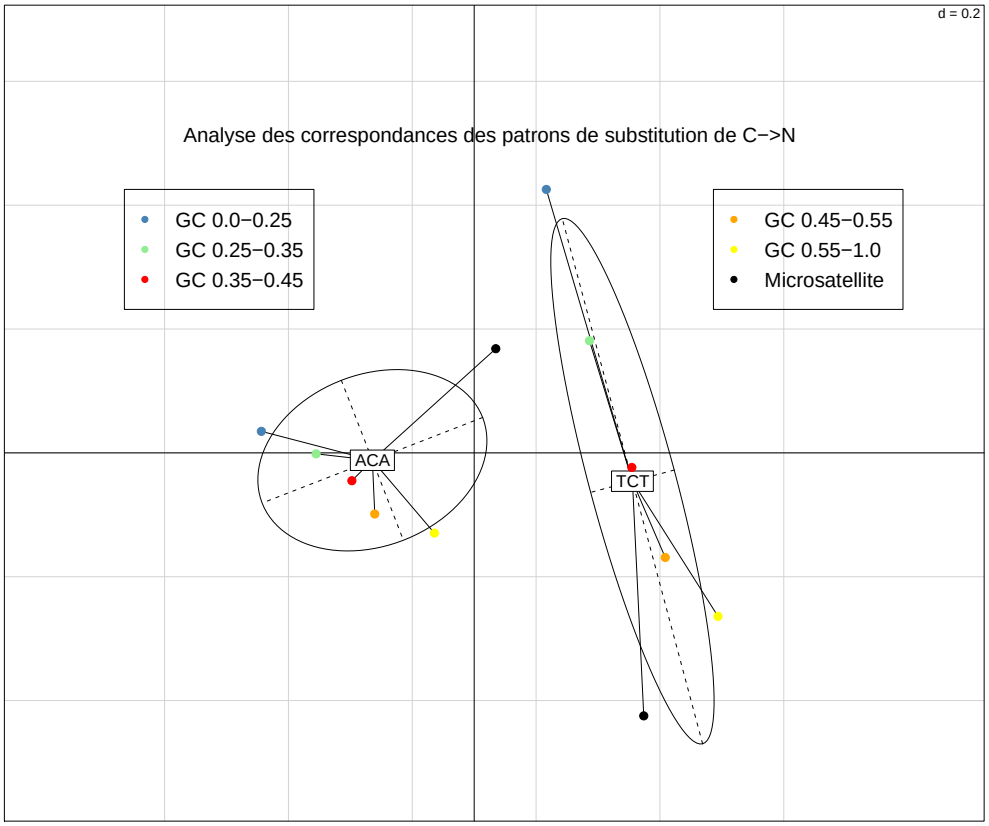

FIG. 4.15 – Carte factorielle pour l'analyse des correspondances des patrons de substitution de C vers A, G et T dans les microsatellites AC/TG_n et AG/TC_n (contextes ACA et TCT) et dans les régions non-répétées, selon le taux de G et C local et du contexte $5'$ - $3'$. L'axe des X correspond à la coordonnée sur le premier axe factoriel (qui totalise 73 % de la variabilité totale des données), l'axe des Y représente la coordonnée sur le deuxième axe factoriel (23% de variabilité). Les classes de taux de G et C pour les régions non-répétées sont représentées en couleur ; les points noirs correspondent aux microsatellites.

4.7 Substitutions asymétriques dans les microsatellites transcrits

Nous avons voulu vérifier ensuite si le patron de substitution dans les microsatellites est affecté par des facteurs régionaux, autres que la composition locale en nucléotides. Dans le génome humain, le mode d'évolution des séquences est très affecté par un processus cellulaire essentiel : la transcription. Comme discuté précédemment, les régions transcrites du génome humain sont caractérisées par un patron de substitution asymétrique par rapport aux deux brins. La substitution qui présente le plus fort degré d'asymétrie est la transition $A \rightarrow G$. Lorsque le brin codant est pris comme référence pour le calcul des substitutions, il y a un fort excès de transitions $A \rightarrow G$ par rapport à $T \rightarrow C$ (Green *et al.*, 2003). De plus, l'intensité de l'asymétrie est influencée par le contexte $5'$ - $3'$ (Hwang et Green, 2004).

Pour tester si la transcription affecte le patron de substitution dans les microsatellites, nous avons identifié les répétitions qui sont présentes dans les introns des gènes protéiques. Nous avons calculé les taux relatifs de substitution par rapport au brin codant, et nous avons comparé les substitutions complémentaires $A \rightarrow G$ et $T \rightarrow C$ dans des contextes complémentaires. Par exemple, le taux de $A \rightarrow G$ dans les microsatellites AG_n est comparé avec le taux de $T \rightarrow C$ dans les répétitions de type TC_n.

Il est important de prendre une précaution supplémentaire pour s'assurer que la différence des taux de substitution est effectivement due à l'asymétrie du mécanisme de transcription, et non pas à la différence d'âge dans les microsatellites : les échelles de temps auxquelles les deux patrons de substitution sont inférés doivent être identiques. Pour faire cela, nous avons pris en compte uniquement les microsatellites humains qui possèdent des orthologues chez le chimpanzé, et pour lesquels le microsatellite orthologue est parfait. Avec cette correction, nous avons plus de certitude que les substitutions observées ont eu lieu depuis le même moment dans le temps : la divergence entre l'homme et le chimpanzé.

Les résultats sont présentés dans la figure 4.16. Seulement les microsatellites mononucléotides et dinucléotides ont été pris en compte pour cette étude, car la quantité de données n'est pas suffisante pour des unités de répétition plus grandes. Le patron de substitution asymétrique observé pour les microsatellites est remarquablement similaire à celui des régions non-répétées. Comme attendu, il y a un excès de transitions $A \rightarrow G$ par rapport à $T \rightarrow C$, pour tous les types de microsatellites. L'asymétrie est plus forte lorsque les voisins en $5'$ et $3'$ sont des pyrimidines, aussi bien pour les microsatellites que pour les séquences uniques.

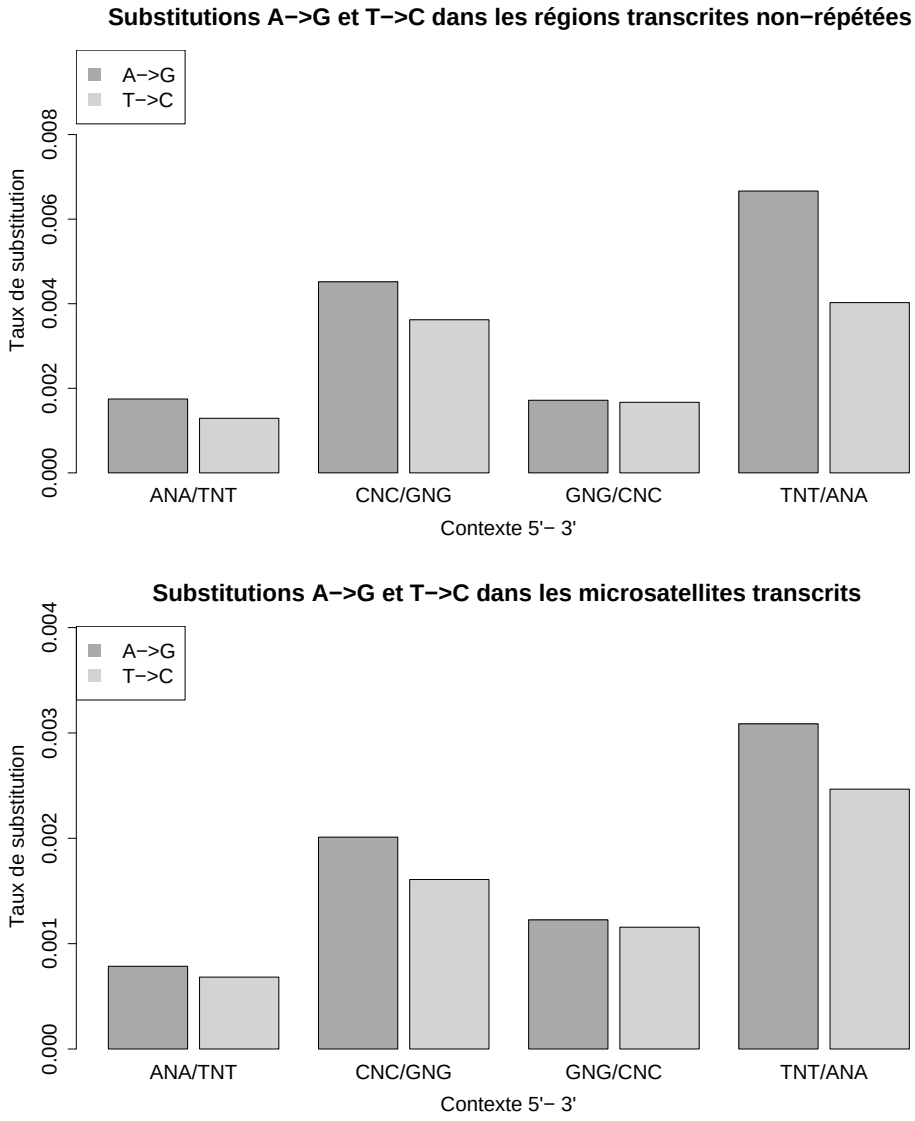

FIG. 4.16 – Patron de substitution asymétrique dans les régions non-répétées et dans les microsatellites.

4.8 Influence de la composition en nucléotides sur le patron de substitution

L'approche que nous avons présentée ici permet d'estimer le patron de substitution des nucléotides dans les répétitions de type microsatellite. Cette méthode est basée sur un raisonnement parcimonieux, et ce mode d'inférence est effectivement susceptible à être biaisé. Nous avons essayé d'éliminer, dans la mesure du possible, les sources d'erreur. Malheureusement, la précision de notre méthode d'inférence des substitutions ne peut pas être estimée par comparaison avec une méthode alternative, car pour les microsatellites les approches classiques d'inférence des substitutions ne sont pas applicables. La discussion qui suit admet comme hypothèse de travail que nos résultats ne sont pas affectés de manière significative par les biais sous-jacents à la méthode d'inférence.

Comme discuté précédemment, il est à ce jour admis que le taux local de guanine et cytosine peut avoir une influence significative sur le patron de substitution. Nous savons également que celui-ci est aussi dépendant de la nature des nucléotides immédiatement voisins en $5'$ et en $3'$. S'agit-il des seuls facteurs qui influencent le mode d'évolution des séquences ? Notre étude de l'évolution des microsatellites semble indiquer le contraire.

Nous avons comparé les patrons de substitution entre microsatellites et régions non-répétées, en tenant compte aussi bien du contexte immédiat que du taux de G et C dans les régions flanquantes. Malgré ces corrections, nous observons que les patrons de substitution sont sensiblement différents entre microsatellites et séquences non-répétées. Nous remarquons que dans la plupart des cas le patron de substitution dans les microsatellites est plutôt similaire au patron estimé pour des régions uniques à composition en bases très biaisée, c'est à dire avec des contenu en G et C soit très forts, soit très faibles. Cela nous semble cohérent, car les microsatellites ont aussi une composition très biaisée - seulement, il ne s'agit pas là seulement du contenu G et C, mais de la composition au sens plus large.

Il existe plusieurs raisons possibles pour les désaccords observés entre les microsatellites et les régions non répétées. Tout d'abord, les échelles de temps auxquelles les deux patrons de substitution sont inférés sont très différentes. Pour les régions non-répétées, les substitutions ont eu lieu depuis la divergence relativement récente (4 à 6 millions d'années) entre l'homme et le chimpanzé. Pour les microsatellites, nous n'avons aucun contrôle de l'échelle de temps avec notre approche. Néanmoins, étant donné le taux moyen de substitution dans les microsatellites et en admettant que la vitesse d'évolution est la même pour les deux types de séquences, nos résultats suggèrent que ces événements évolutifs sont considérablement plus anciens que la divergence homme-chimpanzé. Pour cette raison, et malgré l'élimination des microsatellites trop divergents, l'apparition de

substitution multiples peut biaiser notre estimation.

Une autre explication possible est que le patron de substitution n'est pas influencé uniquement par le taux de G et C, mais aussi par la composition locale en nucléotides, comprise dans un sens plus large. Les microsatellites sont par définition des régions de très faible complexité de séquence. L'explication biologique qui est le plus fréquemment invoquée pour la relation entre le taux de G et C et le patron de substitution est l'influence du contenu en guanine et cytosine sur la stabilité de la double hélice. La composition très biaisée des séquences de type microsatellite n'est pas sans conséquence sur la conformation de la molécule d'ADN. Au contraire, la caractéristique qui définissait les microsatellites, au moment de leur découverte, était le potentiel de ces séquences à former des doubles hélices de type Z (Hamada et Kakunaga, 1982). Au vu de nos résultats, nous pouvons proposer que le patron de mutation ponctuelle dans les microsatellites pourrait être influencé par cette conformation particulière de l'ADN.

Les différences observées pour le patron de substitution entre les microsatellites et les régions non-répétées ne sont certainement pas négligeables. Il existe cependant quelques caractéristiques communes qui peuvent être intéressantes. Notamment, nous avons démontré que dans le génome humain les microsatellites transcrits sont caractérisés par les mêmes patrons de substitution asymétrique que les régions non-répétées. Ce résultat ouvre une application possible pour notre approche. Dans de nombreuses espèces, l'approche comparative classique ne peut pas être appliquée pour inférer le patron de substitution, à cause de l'absence de séquences génomiques provenant d'espèces apparentées. Dans ces situations, l'inférence du patron de substitution dans les microsatellites, qui nécessite moins de données génomiques, peut se prouver utile pour étudier l'effet des mécanismes tels que la réplication ou la transcription sur le mode d'évolution des séquences.

4.9 Conclusion et perspectives

Dans ce chapitre, nous avons étudié le patron de substitution dans des régions caractérisées par une composition "extrême" en nucléotides, et nous avons comparé le mode d'évolution de ces séquences avec celui des régions non-répétées. Nos résultats suggèrent que la nature répétitive de ces séquences et leur contenu biaisé en nucléotides ont une forte influence sur leur patron de substitution. Cette influence de la composition en nucléotides sur le patron d'évolution pourrait s'expliquer par la présence d'une conformation particulière de l'ADN dans ces régions répétées.

Malgré cela, les microsatellites "obéissent" à une des règles fondamentales de l'évolution des séquences d'ADN : l'asymétrie des substitutions dans les régions transcrites. Nous pouvons conclure donc que le processus biologique qui est à l'origine du mode d'évolution asymétrique associé à la transcription n'est vrai-

semblablement pas influencé par la composition extrême en nucléotides, ni par l'éventuelle présence d'une conformation particulière de l'ADN.

4.10 Matériel et méthodes

4.10.1 Données génomiques et détection des microsatellites

Les données génomiques utilisées ici proviennent des assemblages $hg18$, $panTro2$ et $rheMac2$ des génomes de l'homme, du chimpanzé et du macaque, respectivement. Les données ont été extraites de la base de données de l'UCSC Genome Browser (Karolchik *et al.*, 2003). Les annotations des gènes protéiques ont été extraites de la base Ensembl, version 45 (Hubbard *et al.*, 2002), en utilisant le système de requête ACNUC (Gouy *et al.*, 1985).

Nous avons détecté les microsatellites avec le logiciel Tandem Repeats Finder (Benson, 1999), en utilisant comme poids pour l'alignement 2, 5 et 7 pour les appariements corrects, pour les mésappariements et pour les insertions-délétions, respectivement. Ce logiciel permet de sélectionner les microsatellites en fonction du score de l'alignement avec le microsatellite parfait correspondant. Nous avons fixé le seuil de score à 20 pour les microsatellites de type mononucléotide et à 30 pour les microsatellites d'unité di- ou trinucléotide. Avec ce seuil, les microsatellites mononucléotidiques parfaits sont inclus dans l'analyse si leur longueur dépasse 10 unités, et les microsatellites dinucléotides parfaits sont inclus si leur longueur dépasse 8 unités. Nous avons restreint le jeu de données aux microsatellites ayant au moins 80 % d'identité avec le microsatellite parfait correspondant, et nous avons supprimé les répétitions qui possèdent des insertions ou des délétions à l'intérieur des unités de répétition.

Nous avons également éliminé de l'analyse les microsatellites qui se chevauchent avec des éléments transposables, ainsi que les microsatellites composés (qui sont un mélange de deux unités de répétition, par exemple $ACACTCTCACAC$).

Les microsatellites analysés sont uniquement ceux présents dans des régions non-codantes, c'est à dire dans les régions intergéniques, à une distance d'au moins 2kb par rapport au gène voisin le plus proche, ainsi que dans les introns.

4.10.2 Patrons de substitution dans les microsatellites

Pour inférer les patrons de substitution des nucléotides dans les microsatellites, nous avons analysé les interruptions qui pourraient apparaître à travers des mutations ponctuelles. Par exemple, nous excluons les microsatellites imparfaits de type $AAAAACGTAAAAA$, où l'interruption ne peut pas s'expliquer par une simple substitution.

Pour éliminer le risque d'apparition de substitutions multiples au même site, nous avons éliminé les microsatellites qui étaient trop divergents par rapport aux

microsatellites parfaits correspondant (moins de 80 % d'identité). Nous prenons en compte uniquement les substitutions qui ont lieu dans un contexte conservé ; les deux unités voisines de chaque côté d'une imperfection doivent être parfaites pour que la substitution soit comptée.

Pour la première partie de l'analyse, nous avons inféré des patrons de substitution asymétriques, c'est à dire que nous avons regroupé les substitutions complémentaires. Par exemple, les substitutions $GAG \rightarrow GGG$ dans les microsatellites $(AG)_n$ et les substitutions $CTC \rightarrow CCC$ dans les microsatellites $(TC)_n$ ont été regroupées.

Pour la deuxième partie de l'analyse, les patrons de substitutions dans les microsatellites transcrits ont été calculés sur le brin codant, et les microsatellites complémentaires ont été étudiés séparément. Pour pouvoir comparer des taux absolus de substitution (et non pas uniquement des taux relatifs), nous avons fixé l'échelle de temps en analysant uniquement les microsatellites qui ont des orthologues chez le chimpanzé, et où cet orthologue est parfait.

4.10.3 Patrons de substitution dans les séquences non-répétées

Les alignements entre les trois espèces correspondant aux régions intergéniques et aux introns ont été extraits en utilisant les interfaces de requête Galaxy (Giardine *et al.*, 2005) et Génomicro (V. Lombard et L. Duret, communication personnelle). L'inférence des patrons de substitution a été faite avec une méthode de parcimonie, comme décrit par Meunier et Duret (2004). Le contexte nucléotidique des substitutions est pris en compte. Pour compter une substitution dans un contexte XNY, il faut que les trois bases en $5'$ de l'homme, du chimpanzé et du macaque soient égales à X, et de même pour les trois bases voisines en $3'$. Le taux de G et C local (dans une fenêtre de 100 paires de bases) est également pris en compte. Nous avons testé deux autres tailles de fenêtre (20 et 50 paires de bases) ; les résultats ne sont pas significativement affectés.

Comme pour les microsatellites, pour la première partie de l'analyse, nous avons inféré des patrons de substitution asymétriques, c'est à dire que nous avons regroupé les substitutions complémentaires (par exemple les $XAY \rightarrow XGY$ et $\bar{Y}T\bar{X} \rightarrow \bar{Y}C\bar{X}$ ont été regroupés, où $\bar{X}$ est le nucléotide complémentaire de X). Pour la deuxième partie de l'analyse, les patrons de substitutions dans les introns ont été calculés par rapport au brin codant.

4.10.4 Comparaison des patrons de substitutions

Pour comparer les patrons de substitution entre les microsatellites et les régions non-répétées, nous avons utilisé le test de χ^2 d'égalité des distributions, appliqué aux tables de contingence. Ce test permet de comparer des distributions 2 à 2. Nous avons également réalisé une comparaison plus global, grâce à

des analyses de correspondances (Greenacre, 1984). Ces analyses de correspondances ont été réalisées sur des tables de contingence qui croisent les différents types de substitutions (par exemple $A \to C$, $A \to G$ et $A \to T$ pour la première analyse) avec les différents types de régions (microsatellites, régions non-répétées avec un taux de G et C dans les classes $0-0.25$, $0.25-0.35$, $0.25-0.35$, $0.35-0.45$, $0.45-0.55$ et $0.55-1$).

Les calculs ont été réalisés en utilisant les fonctions proposées par la bibliothèque ADE4 (Chessel *et al.*, 2004) dans R (R Development Core Team, 2005).

Conclusion et perspectives

Dans les sciences du vivant, les avancées se font historiquement sur une échelle de complexité décroissante : l'étude des organismes a été révolutionnée par la découverte de la cellule, et la biologie cellulaire a été à son tour métamorphosée par découverte du matériel génétique. A présent, nous disposons de toutes les informations nécessaires pour étudier les plus petites "briques" du vivant : les nucléotides qui forment les séquences d'ADN. Mais pour comprendre l'évolution des séquences d'ADN, nous devons remonter l'échelle de la complexité. Si l'ADN fournit le plan de construction de la cellule vivante, il est tout aussi vrai que le fonctionnement de cette cellule est le principal moteur de l'évolution de l'ADN. Il est donc important de comprendre l'impact des processus cellulaires essentiels, tels que la réplication, la transcription ou la recombinaison sur le mode d'évolution des séquences d'ADN.

L'objectif de cette thèse a été d'étudier deux de ces processus cellulaires essentiels : la réplication et la transcription. Ces deux mécanismes ont un fonctionnement qui est asymétrique par rapport aux deux brins d'ADN, et ils ont tous les deux comme conséquence évolutive l'apparition d'une composition asymétrique dans les séquences génomiques. De plus, ces deux phénomènes agissent de manière coordonnée le long des chromosomes, et il est souvent très difficile de séparer leurs influences respectives sur l'évolution des séquences.

Dans la première partie de cette thèse, nous avons étudié un aspect particulier de l'organisation des chromosomes procaryotes : la co-orientation entre réplication et transcription. Nous avons proposé une nouvelle approche pour l'étude des biais de composition des séquences, qui permet de séparer ces deux sources d'asymétrie. Grâce à cette nouvelle méthode, qui a l'avantage qu'elle ne nécessite pas de connaissances *a priori* sur la position de l'origine et du terminus de réplication, nous avons réussi à évaluer l'impact de la réplication sur l'asymétrie de composition, pour l'ensemble des génomes procaryotes complètement séquencés. Nous avons démontré que les biais de composition associés à la réplication, loin d'être universels, présentent une très grande variabilité, même entre espèces proches de bactéries. Quels sont les fondements biologiques de cette variabilité ? Cette question reste à présent ouverte.

Par la suite, nous avons étudié le patron d'évolution des séquences d'ADN, dans les régions transcrites et autour des origines de réplication. Nous avons surtout porté notre attention sur un phénomène particulier : l'influence du contexte nucléotidique immédiat sur l'asymétrie d'évolution des séquences. Cette analyse a été réalisée sur le génome humain, car c'est pour cette espèce que l'on a démontré pour la première fois que la nature des nucléotides voisins en $5'$ et en $3'$ a un impact important sur le patron de substitution asymétrique dans le génome humain (Hwang et Green, 2004). Ici, nous avons réussi à démontrer que des biais de voisinages similaires sont en jeu pour l'évolution asymétrique des séquences autour des origines de réplication.

Nous avons également analysé la variation des taux de substitutions asymétriques en fonction du patron d'expression des gènes, et nos résultats nous ont permis de suggérer qu'un biais de réparation préférentiel sur le brin matrice de la transcription pourrait entraîner l'évolution asymétrique des séquences. L'influence du voisinage sur l'asymétrie de substitution pourrait s'expliquer alors par une variabilité de l'efficacité de réparation des mésappariements selon le contexte nucléotidique. Mais peut-on étendre la même conclusion pour l'asymétrie de composition engendrée par la réplication ? La similarité entre les patrons d'évolution asymétrique provoqués par la réplication et par la transcription suggère que le même type de mécanisme biologique pourrait en être la cause. Des analyses supplémentaires sont à présent nécessaires pour déterminer si l'évolution asymétrique associée à la réplication pourrait aussi être l'effet d'un mécanisme de réparation.

Nous nous sommes ensuite intéressés à l'influence du contexte génomique sur l'évolution des séquences d'ADN, cette fois dans un sens plus large, car nous avons étudié la relation entre la composition locale en nucléotides et le patron de substitution. En particulier, nous avons proposé une méthodologie d'étude du patron de substitution nucléotidique dans des régions répétées à composition en nucléotides très biaisée : les microsatellites. Nous avons appliqué cette méthode pour inférer le mode d'évolution des microsatellites présents dans le génome humain. Nous avons démontré que le patron de substitution des microsatellites est souvent différent de celui qui est caractéristique des régions non-répétées. Nous avons proposé que cette différence pourrait être due à l'existence d'une conformation particulière de la double hélice d'ADN dans ce type de séquence.

Malgré les différences observées entre microsatellites et les régions non-répétées, nos résultats démontrent que dans les régions transcrites les microsatellites sont sujets au mêmes processus évolutifs asymétriques que le reste du génome. Si la structure tridimensionnelle de l'ADN est effectivement en cause pour le mode d'évolution particulier des microsatellites, ce dernier résultat suggère que le mécanisme biologique qui engendre l'asymétrie de substitution liée à la transcription ne serait pas dépendant de la conformation de la double hélice. Cela soulève à nouveau des questions sur l'identité de ce mécanisme.

Une conclusion importante que nous pouvons tirer de ce travail est que lorsque l'on étudie l'évolution de l'ADN, il est important de comprendre que les nucléotides A, C, G, T ne sont pas entités indépendantes. Le mode d'évolution des séquences est fortement influencé par le contexte génomique, à échelles plus ou moins grandes.

A l'avenir, l'étude de l'influence du contexte génomique pourra être étendue à des niveaux supérieurs d'organisation. Notamment, les avancées récentes (aussi bien *in silico* qu'expérimentales) en ce qui concerne le positionnement des nucléosomes et la structure de la chromatine, rendent possible une analyse plus poussée de l'impact de l'environnement des séquences d'ADN sur leur mode d'évolution.

Bibliographie

Angers, B., et Bernatchez, L. 1997. Complex evolution of a salmonid microsatellite locus and its consequences in inferring allelic divergence from size information. *Mol. Biol. Evol.*, **14**, 230–238.

Arndt, P.F., Hwa, T., et Petrov, D.A. 2005. Substantial regional variation in substitution rates in the human genome : importance of GC content, gene density and telomere-specific effects. *J. Mol. Evol.*, **60**, 748–763.

Avery, O. T., MacLeod, C. M., et McCarty, M. 1944. Studies on the chemical nature of the substance inducing transformation of pneumococcal types. *J. Exp. Med.*, **79**, 137–158.

Balaresque, P., Toupance, B., Heyer, E., et Crouau-Roy, B. 2003. Evolutionary dynamics of duplicated microsatellites shared by sex chromosomes. *J. Mol. Evol.*, **57**, S128–S137.

Baldacci, G., et Bernardi, G. 1982. Replication origins are associated with transcription initiation sequences in the mitochondrial genome of yeast. *EMBO J.*, **1**, 987–994.

Barnes, D.E., et Lindahl, T. 2004. Repair and genetic consequences of endogenous DNA base damage in mammalian cells. *Annu. Rev. Genet.*, **38**, 445–476.

Basic-Zaninovic, T., Palombo, F., Bignami, M., et Dogliotti, E. 1992. Fidelity of replication of the leading and the lagging DNA strands opposite N-methyl-N-nitrosourea-induced DNA damage in human cells. *Nucleic Acids Res.*, **20**, 6543–6548.

Beletskii, A., et Bhagwat, A.S. 1996. Transcription-induced mutations : Increase in C to T mutations in the nontranscribed strand during transcription in *Escherichia coli*. *Proc. Natl. Acad. Sci. USA*, **93**, 13919–13924.

Beletskii, A., et Bhagwat, A.S. 2001. Transcription-induced Cytosine-to-Thymine mutations are not dependent on sequence context of the target Cytosine. *J. Bacteriol.*, **183**, 6491–6493.

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

Berquist, B.R., et DasSarma, S. 2003. An archaeal chromosomal autonomously replicating sequence element from an extreme halophile, *Halobacterium* sp. strain NRC-1. *J. Bacteriol.*, **185**, 5959–5966.

Black, D. L. 2003. Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem.*, **72**, 291–336.

Blake, R.D., Hess, S.T., et Nicholson-Tuell, J. 1992. The influence of nearest neighbors on the rate and pattern of spontaneous point mutations. *J. Mol. Evol.*, **34**, 189–200.

Blanquer-Maumont, A., et Crouau-Roy, B. 1995. Polymorphism, monomorphism, and sequences in conserved microsatellites in primate species. *J. Mol. Evol.*, **41**, 492–497.

Blattner, FR, Plunkett, G, Bloch, CA, Perna, NT, Burl and, V, Riley, M, ColladoVides, J, Glasner, JD, Rode, CK, Mayhew, GF, Gregor, J, Davis, NW, Kirkpatrick, HA, Goeden, MA, Rose, DJ, Mau, B, et Shao, Y. 1997. The complete genome sequence of Escherichia coli K-12. *Science*, **277**(5331), 1453–1462.

Bockrath, R., et Li, B-H. 1998. Transcriptional mutagenesis and DNA strand asymmetrical mutations expressed in *Escherichia coli* under restrictive metabolic conditions. *Mut. Res.*, **422**, 351–355.

Bohr, W., Smith, C. A., Okumoto, D.S., et Hanawalt, P.C. 1985. DNA repair in an active gene : Removal of pyrimidine dimers from the DHFR gene of CHO cells is much more efficient than in the genome overall. *Cell*, **40**, 359–369.

Boyer, J.C., Hawk, J.D., Stefanovic, L., et Farber, R.A. 2008. Sequence-dependent effect of interruptions on microsatellite mutation rate in mismatch repair-deficient human cells. *Mutat. Res.-Fund. Mol. M.*, **000**, 000–000.

Brewer, B. J. 1988. When polymerases collide : replication and the transcriptional organization of the *E. coli* chromosome. *Cell*, **53**, 679–686.

Brewer, B.J., et Fangman, W.L. 1988. A replication fork barrier at the 3′ end of yeast ribosomal RNA genes. *Cell*, **55**, 637–643.

Brewer, B.J., Lockshon, D., et Fangman, W.L. 1992. The arrest of replication forks in the rDNA of yeast occurs independently of transcription. *Cell*, **71**, 267–276.

Brohede, J., et Ellegren, H. 1999. Microsatellite evolution : polarity of substitutions within repeats and neutrality of flanking sequences. *Proc. R. Soc. Lond. B*, **266**, 825–833.

Bruck, I., et O'Donnell, M. 2000. The DNA replication machine of a Gram-positive organism. *J. Biol. Chem.*, **275**, 28971–28983.

Bruck, I., Goodman, M.F., et O'Donnell, M. 2003. The esssential C family DnaE polymerase is error-prone and efficient at lesion bypass. *J. Biol. Chem.*, **278**, 44361–44368.

Bulmer, M. 1991. Strand symmetry of mutation rates in the beta-globin region. *J. Mol. Evol.*, **33**, 305–310.

Burgers, P. M. J. 1991. *Saccharomyces cerevisiae* replication factor C. II. Formation and activity of complexes with the proliferating cell nuclear antigen and with DNA polymerases δ and ϵ. *J. Biol. Chem.*, **266**, 22698–22706.

Burgess, R. R. 1971. RNA polymerase. *Annu. Rev. Biochem.*, **40**, 711–740.

Buschiazzo, E., et Gemmell, N.J. 2006. The rise, fall and renaissance of microsatellites in eukaryotic genomes. *BioEssays*, **28**, 1040–1050.

Campbell, J.L. 1993. Yeast DNA replication. *J. Biol. Chem.*, **268**, 25261–25264.

Cann, I.K.O., Komori, K., Toh, H., Kanai, S., et Ishino, Y. 1998. A heterodimeric DNA polymerase : evidence that members of Euryarchaeota possess a distinct DNA polymerase. *Proc. Natl. Acad. Sci. USA*, **95**, 14250–14255.

Chambon, P. 1975. Eukaryotic nuclear RNA polymerases. *Annu. Rev. Biochem.*, **44**, 613–638.

Chargaff, E., Vischer, E., Doniger, R., Green, C., et Misani, F. 1949. The composition of the desoxypentose nucleic acids of thymus and spleen. *J. Biol. Chem.*, **177**, 405–416.

Chargaff, E., Lipshitz, R., Green, C., et Hodes, M.E. 1951. The composition of the desoxyribonucleic acid of salmon sperm. *J. Biol. Chem.*, **192**, 223–230.

Charif, D., et Lobry, J. R. 2006. SeqinR 1.0-2 : a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. *In :* Bastolla, U., Porto, M, Roman, H.E., et M., Vendruscolo (eds), *Structural approaches to sequence evolution : Molecules, networks, populations.* Biological and Medical Physics, Biomedical Engineering. New York : Springer Verlag. in press.

Chessel, D., Dufour, A.-.B., et Thioulouse, J. 2004. *The ade4 package-I- One-table methods*. R News.

Christians, F.C., et Hanawalt, P.C. 1993. Lack of transcription-coupled repair in mammalian ribosomal RNA genes. *Biochemistry*, **32**, 10512–10518.

Chung, M-Y., Ranum, L.P.W., Duvick, L.A., Servadio, A., Zoghbi, H.Y., et Orr, H.T. 1993. Evidence for a mechanism predisposing to intergenerational CAG repeat instability in spinocerebellar ataxia type I. *Nature Genet.*, **5**, 254–258.

Clayton, D.A. 1991. Replication and transcription of vertebrate mitochondrial DNA. *Annu. Rev. Cell. Biol.*, **7**, 453–478.

Collins, F.S., ander, E.S. L, Rogers, J., et Waterston, R.H. 2004. Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931–945.

Comeron, J. M. 2004. Selective and mutational patterns associated with gene expression in humans : influences on synonymous composition and intron presence. *Genetics*, **167**, 1293–1304.

Conconi, A., Bespalov, V. A., et Smerdon, M.J. 2002. Transcription coupled repair in RNA polymerase I-transcribed genes of yeast. *Proc. Natl. Acad. Sci. USA*, **99**, 649–654.

Dammann, R., et Pfeifer, G.P. 1997. Lack of gene- and strand-specific DNA repair in RNA polymerase III-transcribed human tRNA genes. *Mol. Cell. Biol.*, **17**, 219–229.

de Laat, W. L., Jaspers, N. G. J., et Hoeijmakers, J. H. J. 1999. Molecular mechanism of nucleotide excision repair. *Genes Dev.*, **13**, 768–785.

Delgado, S., Gomez, M., Bird, A., et Antequera, F. 1998. Initiation of DNA replication at CpG isl ands in mammalian chromosomes. *EMBO J.*, **17**, 2426–2435.

DePamphilis, M.L. 1993. Eukaryotic DNA replication : anatomy of an origin. *Annu. Rev. Biochem.*, **62**, 29–63.

DePamphilis, M.L. 1999. Replication origins in metazoan chromosomes : fact or fiction ? *BioEssays*, **21**, 5–16.

Dervyn, E., Suskin, C., Daniel, R., Bru, C., , Chapuis, J., Errington, J., Janniere, L., et Ehrlich, S.D. 2001. Two essential DNA polymerases at the bacterial replication fork. *Science*, **294**, 1716–1721.

Deshpande, A.M., et Newlon, C.S. 1996. DNA replication fork pause sites dependent on transcription. *Science*, **272**, 1030–1033.

Duncan, B.K., et Miller, J.H. 1980. Mutagenic deamination of cytosine residues in DNA. *Nature*, **287**, 560–561.

Duquenne, L., Huvet, M., d'Aubenton Carafa, Y., Thermes, C., Zaghloul, L., Nicolay, S., Arneodo, A., et Audit, B. 2007. Replication shapes the substitution landscape of the human genome. *In : Integrative Post-Genomics.*

Duret, L. 2002. Evolution of synonymous codon usage in metazoans. *Curr. Opin. Genet. Dev.*, **12**, 640–649.

Duret, L., et Arndt, P. F. 2008. The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet.*, **4**, 1–19.

Edgell, D. R., et Doolittle, W.F. 1997. Archaea and the origin of DNA replication proteins. *Cell*, **89**, 995–999.

Ehrlich, M., et Yang, R.Y.H. 1981. 5-methylcytosine in eukaryotic DNA. *Science*, **212**, 1350–1357.

Elias-Arnanz, M., et Salas, M. 1997. Bacteriophage ϕ29 DNA replication arrest caused by codirectional collisions with the transcription machinery. *EMBO J.*, **16**, 5775–5783.

Elias-Arnanz, M., et Salas, M. 1999. Resolution of head-on collisions between the transcription machinery and bacteriophage Φ29 DNA polymerase is dependent on RNA polymerase translocation. *EMBO J.*, **18**, 5675–5682.

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

Estoup, A., Solignac, M., Harry, M., et Cornuet, J-M. 1993. Characterization of (GT)n and (CT)n microsatellites in two insect species : *Apis mellifera* and *Bombus terrestris*. *Nucleic Acids Res.*, **21**, 1427–1431.

Fijalkowska, I.J., Jonczyk, P., Maliszewska-Tkaczyk, M., Bialoskorska, M., et Schaaper, R.M. 1998. Unequal fidelity of leading strand and lagging strand DNA replication on the *Escherichia coli* chromosome. *Proc. Natl. Acad. Sci. USA*, **95**, 10020–10025.

Filipski, J., Thiery, J.P., et Bernardi, G. 1973. An analysis of the bovine genome by Cs2SO4-Ag density gradient centrifugation. *J. Mol. Biol.*, **80**, 177–197.

Fortune, J.M., Pavlov, Y.I., Welch, C.M., Johansson, E., Burgers, P.M.J., et Kunkel, T.A. 2005. *Saccharomyces cerevisiae* DNA polymerase δ. High fidelity for base substitutions but lower fidelity for single- and multi-base deletions. *J. Biol. Chem.*, **280**, 29980–29987.

Francino, M.P., et Ochman, H. 2000. Str and symmetry around the β-globin origin of replication in primates. *Mol. Biol. Evol.*, **17**, 416–422.

Francino, M.P., Chao, L., Riley, M.A., et Ochman, H. 1996. Asymmetries generated by transcription-coupled repair in enterobacterial genes. *Science*, **272**, 107–109.

Frank, A.C., et Lobry, J.R. 1999. Asymmetric substitution patterns : a review of possible underlying mutational or selective mechanisms. *Gene*, **238**, 65–77.

Frank, A.C., et Lobry, J.R. 2000. Oriloc : prediction of replication boundaries in unannotated bacterial chromosomes. *Bioinformatics*, **16**, 560–561.

Frederico, L.A., Kunkel, T.A., et Shaw, B.R. 1990. A sensitive genetic assay for the detection of cytosine deamination : deamination of rate constants and the activation energy. *Biochemistry*, **29**, 2532–2537.

French, S. 1992. Consequences of replication fork movement through transcription units in vivo. *Science*, **258**, 1382–1385.

Fryxell, K.J., et Zuckerkandl, E. 2000. Cytosine deamination plays a primary role in the evolution of mammalian isochores. *Mol. Biol. Evol.*, **17**, 1371–1383.

Fukui, T., Yamauchi, K., Muroya, T., Akiyama, M., Maki, H., Sugino, A., et Waga, S. 2004. Distinct roles of DNA polymerases delta and epsilon and the replication fork in *Xenopus* egg extracts. *Genes to Cells*, **9**, 179–191.

Garg, P., et Burgers, P.M.J. 2005. DNA polymerases that propagate the eukaryotic DNA replication fork. *Crit. Rev. Biochem. Mol.*, **40**, 115–128.

Giardine, B., Riemer, C., Hardison, R.C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, Y., Taylor, J., Miller, W., Kent, W.J., et Nekrutenko, A. 2005. Galaxy : A platform for interactive large-scale genome analysis. *Genome Res.*, **15**, 1451–1455.

Glover, B.P., et McHenry, C.S. 2001. The DNA polymerase III holoenzyme : an asymmetric dimeric replicative complex with leading and lagging strand polymerases. *Cell*, **105**, 925–934.

Gonçalves, I., Duret, L., et Mouchiroud, D. 2000. Nature and structure of human genes that generate retropseudogenes. *Genome Res.*, **10**, 672–678.

Gotta, S. L., Miller, O. L., et French, S. L. 1991. rRNA transcription rate in *Escherichia coli*. *J. Bacteriol.*, **20**, 6647–6649.

Gouy, M., Gautier, C., Attimonelli, M., Lanave, C., et di Paola, G. 1985. ACNUC – a portable retrieval system for nucleic acid sequence databases : logical and physical designs and usage. *Bioinformatics*, **1**, 167–172.

Green, P., Ewing, B., Miller, W., Thomas, P.J., et Green, E.D. 2003. Transcription-associated mutational asymmetry in mammalian evolution. *Nature Genet.*, **33**, 514–517.

Greenacre, M. 1984. *Theory and applications of correspondence analysis*. Academic Press.

Grigoriev, A. 1998. Analyzing genomes with cumulative skew diagrams. *Nucleic Acids Res.*, **26**, 2286–2290.

Grigoriev, A. 2000. Graphical genome comparison : rearrangements and replication origin of *Helicobacter pylori*. *Trends in Genet.*, **16**, 376–378.

Guo, Z-S., et DePamphilis, M.L. 1992. Specific transcription factors stimulate Simian Virus 40 and polyomavirus origins of DNA replication. *Mol. Cell. Biol.*, **12**, 2514–2524.

Hamada, H., et Kakunaga, T. 1982. Potential Z-DNA forming sequences are highly dispersed in the human genome. *Nature*, **298**, 396–398.

Hamada, H., Petrino, M. G., et Kakunaga, T. 1982. A novel repeated element with Z-DNA-forming potential is widely found in evolutionarily diverse eukaryotic genomes. *Proc. Natl. Acad. Sci. USA*, **79**, 5901–5905.

Harr, B., Todorova, J., et Schlotterer, C. 2002. Mismatch repair-driven mutational bias in *Drosophila melanogaster*. *Mol. Cell*, **10**, 199–205.

Hashimoto, K., Shimizu, K., Nakashima, N., et Sugino, A. 2003. Fidelity of DNA polymerase δ holoenzyme from *Saccharomyces cerevisiae* : the sliding clamp proliferating cell nuclear antigen decreases its fidelity. *Biochemistry*, **42**, 14207–14213.

Hassan, A.B., Errington, R.J., White, N.S., Jackson, D.A., et Cook, P.R. 1994. Replication and transcription are colocalized in human cells. *J Cell Science*, **107**, 425–434.

Henneke, G., Flament, D., Hubscher, U., Querellou, J., et Raffin, J-P. 2005. The hyperthermophilic euryarchaeota *Pyrococcus abyssi* likely requires the two DNA polymerases D and B for DNA replication. *J. Mol. Biol.*, **350**, 53–64.

Hess, S.T., Blake, J.D., et Blake, R.D. 1994. Wide variations in neighbor-dependent substitution rates. *J. Mol. Biol.*, **236**, 1022–1033.

Higuchi, K., Katayama, T., Iwai, S., Hidaka, M., Horiuchi, T., et Maki, H. 2003. Fate of DNA replication fork encountering a single DNA lesion during *oriC* plasmid DNA replication *in vitro*. *Genes Cells*, **8**, 437–449.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., Durbin, R., Eyras, E., Gilbert, J., Hammond, M., Huminiecki, L., Kasprzyk, A., Lehvaslaiho, H., Lijnzaad, P., Melsopp, C., Mongin, E., Pettett, R., Pocock, M., Potter, S., Rust, A., Schmidt, E., Searle, S., Slater, G., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Stupka, E., Ureta-Vidal, A., Vastrik, I., et Clamp, M. 2002. The Ensembl genome database project. *Nucleic Acids Res.*, **30**, 38–41.

Huvet, M., Nicolay, S., Touchon, M., Audit, B., d'Aubenton Carafa, Y., Arneodo, A., et Thermes, C. 2007. Human gene organization driven by the coordination of replication and transcription. *Genome Res.*, **17**, 1278–1285.

Hwang, D.G., et Green, P. 2004. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc. Natl. Acad. Sci. USA*, **101**, 13994–14001.

Inoue, R., Kaito, C., Tanabe, M., Kamura, K., Akimitsu, N., et Sekimizu, K. 2001. Genetic identification of two distinct DNA polymerases, DnaE and PolC, that are essential for chromosomal DNA replication in *Staphylococcus aureus*. *Mol. Genet. Genomics*, **266**, 564–571.

Iwaki, T., Kawamura, A., Ishino, Y., Kohno, K., Kano, Y., Goshima, N., Yara, M., Furusawa, M., Doi, H., et Imamoto, F. 1996. Preferential replication dependent mutagenesis in the lagging DNA strand in Escherichia coli. *Mol. Gen. Genet.*, **251**, 657–664.

Jin, Y.H., Ayyagari, R., Resnick, M.A., Gordenin, D. A., et Burgers, P. M. J. 2003. Okazaki fragment maturation in yeast. II. Cooperation between the polymerase and $3' - 5'$ exonuclease activities of Pol δ in the creation of a ligatable nick. *J. Biol. Chem.*, **278**, 1626–1633.

Johanson, K.O., et McHenry, C.S. 1984. Adenosine 5'-*O*-(3-Thiotriphosphate) can support the formation of an initiation complex between the DNA polymerase III holoenzyme and primed DNA. *J. Biol. Chem.*, **259**, 4589–4595.

Johnson, A., et O'Donnell, M. 2005. Cellular DNA replicases : components and dynamics at the replication fork. *Annu. Rev. Biochem.*, **74**, 283–315.

Jones, M., Wagner, R., et Radman, M. 1987. Repair of a mismatch is influenced by the base composition of the surrounding nucleotide sequence. *Genetics*, **115**, 605–610.

Karkas, J.D., Rudner, R., et Chargaff, E. 1968. Separation of *B. subtilis* DNA into complementary strands, II. Template functions and composition as determined by transcription with RNA polymerase. *Proc. Natl. Acad. Sci. USA*, **60**, 915–920.

Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J., Weber, R.J., Haussler, D., et Kent, W.J. 2003. The UCSC Genome Browser Database. *Nucleic Acids Res.*, **31**, 51–54.

Karthikeyan, R., Vonarx, E.J., Straffon, A.F.L., Simon, M., Faye, G., et Kunz, B.A. 2000. Evidence from mutational specificity studies that yeast DNA polymerases δ andϵ replicate different DNA strands at an intracellular replication fork. *J. Mol. Biol.*, **299**, 405–419.

Kelkar, Y.D., Tyekucheva, S., Chiaromonte, F., et Makova, K.D. 2008. The genome-wide determinants of human and chimpanzee microsatellite evolution. *Genome Res.*, **18**, 30–38.

Kelman, Z., et O'Donnell, M. 1995. DNA polymerase III holoenzyme : structure and function of a chromosomal replicating machine. *Annu. Rev. Biochem.*, **64**, 171–200.

Khaitovich, P., Enard, W., Lachmann, M., et Paabo, S. 2006. Evolution of primate gene expression. *Nat. Rev. Genet.*, **7**, 693–672.

Kim, S., Dallmann, H.G., McHenry, C.S., et Marians, K.J. 1996. τ couples the leading- and lagging-strand polymerases at the *Escherichia coli* DNA replication fork. *J. Biol. Chem.*, **271**, 21406–21412.

Klapacz, J., et Bhagwat, A. S. 2005. Transcription promotes guanine to thymine mutations in the non-transcribed strand of an *Escherichia coli* gene. *DNA Repair*, **4**, 806–813.

Klapacz, J., et Bhagwat, A.S. 2002. Transcription-dependent increase in multiple classes of base substitution mutations in *Escherichia coli*. *J. Bacteriol.*, **184**, 6866–6872.

Klasson, L., et Andersson, S.G.E. 2006. Strong asymmetric mutation bias in endosymbiont genomes coincide with loss of genes for replication restart pathways. *Mol. Biol. Evol.*, **23**, 1031–1039.

Kleene, C. K. 2001. A possible meiotic function of the peculiar patterns of gene expression in mammalian spermatogenic cells. *Mech. Dev.*, **106**, 3–23.

Kobayashi, T., et Horiuchi, T. 1996. A yeast gene product, Fob1 protein, required for both replication fork blocking and recombinational hotspot activities. *Genes Cells*, **1**, 465–474.

Koch, R.E. 1971. The influence of neighboring base pairs upon base-pair substitution mutation rates. *Proc. Natl. Acad. Sci. USA*, **68**, 773–776.

Koonin, E.V., et Bork, P. 1996. Ancient duplication of DNA polymerase inferred from analysis of complete bacterial genomes. *Trends Biochem. Sci.*, **21**, 128–129.

Kornberg, A. 1988. DNA replication. *J. Biol. Chem.*, **263**, 1–4.

Kow, Y.W., Bao, G., Reeves, J.W., Jinks-Robertson, S., et Crouse, G.F. 2007. Oligonucleotide transformation of yeast reveals mismatch repair complexes to be differentially active on DNA replication strands. *Proc. Natl. Acad. Sci. USA*, **104**, 11352–11357.

Kruglyak, S., Durrett, R.T., Schug, M.D., et Aquadro, C.F. 1998. Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proc. Natl. Acad. Sci. USA*, **95**, 10744–10778.

Kunkel, T.A., et Alexander, P.S. 1986. The base substitution fidelity of eucaryotic DNA polymerases. *J. Biol. Chem.*, **261**, 160–166.

Kunst, F., Ogasawara, N., Moszer, I., Albertini, A. M., Alloni, G., Azevedo, V., Bertero, M. G., , Bessieres, P., Bolotin, A., Borchert, S., Borriss, R., Boursier, L., Brans, A., Braun, M., Brignell, S. C., Bron, S., Brouillet, S., , Bruschi, C. V., Caldwell, B., Capuano, V., Carter, N. M., Choi, S.-K., Codani, J.-J., Connerton, I. F., Cummings, N. J., Daniel, R. A., Denizot, F., Devine, K. M., Dusterhoft, A., Ehrlich, S. D., Emmerson, P. T., Entian, K. D., Errington, J., Fabret, C., Ferrari, E., Foulger, D., Fritz, C., Fujita, M., Fujita, Y., Fuma, S., Galizzi, A., Galleron, N., Ghim, S.-Y., Glaser, P., Goffeau, A., Golightly, E. J., andi, G. Gr, Guiseppi, G., Guy, B. J., Haga, K., Haiech, J., Harwood, C. R., Henaut, A., Hilbert, H., Holsappel, S., Hosono, S., Hullo, M.-F., Itaya, M., Jones, L., Joris, B., Karamata, D., Kasahara, Y., Klaerr-Blanchard, M., Klein, C., Kobayashi, Y., Koetter, P., Koningstein, G., Krogh, S., Kumano, M., Kurita, K., Lapidus, A., Lardinois, S., Lauber, J., Lazarevic, V., Lee, S.-M., Levine, A., Liu, H., Masuda, S., Mauel, C., Medigue, C., , Medina, N., Mellado, R. P., Mizuno, M., Moestl, D., Nakai, S., Noback, M., Noone, D., O'Reilly, M., Ogawa, K., Ogiwara, A., Oudega, B., Park, S.-H., Parro, V., Pohl, T. M., Portetelle, D., Porwollik, S., Prescott, A. M., Presecan, E., Pujic, P., Purnelle, B., Rapoport, G., Rey, M., Reynolds, S., Rieger, M., Rivolta, C., Rocha, E., , Roche, B., Rose, M., Sadaie, Y., Sato, T., Scanlan, E., Schleich, S., Schroeter, R., Scoffone, F., Sekiguchi, J., Sekowska, A., Seror, S. J., Serror, P., Shin, B.-S., Soldo, B., Sorokin, A., Tacconi, E., Takagi, T., Takahashi, H., Takemaru, K., Takeuchi, M., Tamakoshi, A., Tanaka, T., Terpstra, P., Tognoni, A., Tosato, V., Uchiyama, S., andenbol, M. V, Vannier, F., Vassarotti, A., Viari, A., Wambutt, R., Wedler, E., Wedler, H., Weitzenegger, T., Winters, P., Wipat, A., Yamamoto, H., Yamane, K., Yasumoto, K., Yata, K., Yoshida,

K., Yoshikawa, H.-F., Zumstein, E., Yoshikawa, H., et Danchin, A. 1997. The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. *Nature*, **390**, 249–266.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., *et al.*. 2001. Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Le Chatelier, E., Bécherel, O.J., d'Alençon, E., Canceill, D., Ehrlich, S. D., Fuchs, R.P.P., et Janniere, L. 2004. Involvement of DnaE, the second replicative DNA polymerase from *Bacillus subtilis*, in DBA mutagenesis. *J. Biol. Chem.*, **279**, 1757–1767.

Leadon, S. A., et Lawrence, D. A. 1991. Preferential repair of DNA damage on the transcribed strand of the human metallothionein genes requires RNA polymerase II. *Mutat. Res.*, **255**, 67–78.

Lercher, M. J., Chamary, J-V., et Hurst, L. D. 2004. Genomic regionality in rates of evolution is not explained by clustering of genes of comparable expression profile. *Genome Res.*, **14**, 1002–1013.

Levinson, G., et Gutman, G. A. 1987. Slipped-strand mispairing : a major mechanism for DNA sequence evolution. *Mol. Biol. Evol.*, **4**, 203–221.

Liu, B., et Alberts, B.M. 1995. Head-on collision between a DNA replication apparatus and RNA polymerase transcription complex. *Science*, **267**, 1131–1137.

Liu, B., Wong, M.L., Tinker, R.L., Geiduschek, E.P., et Alberts, B.M. 1993. The DNA-replication fork can pass RNA-polymerase without displacing the nascent transcript. *Nature*, **366**, 33–39.

Lobry, J. R., et Sueoka, N. 2002. Asymmetric directional mutation pressures in bacteria. *Genome Biol.*, **3**, research0058.1–research0058.14.

Lobry, J.R. 1995. Properties of a general model of DNA evolution under no-strand-bias conditions. *J. Mol. Evol.*, **40**, 326–330.

Lobry, J.R. 1996a. Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol. Biol. Evol.*, **13**, 660–665.

Lobry, J.R. 1996b. Origin of replication of *Mycoplasma genitalium*. *Science*, **272**, 745–746.

Low, R. L., Rashbaum, S.A., et Cozzarelli, N. R. 1976. Purification and characterization of DNA polymerase III from *Bacillus subtilis*. *J. Biol. Chem.*, **251**, 1311–1325.

Lundberg, K. S., Shoemaker, D. D., Adams, M. W. W., Short, J. M., Sorge, J. A., et Mathur, E. J. 1991. High-fidelity amplification using a thermostable DNA polymerase isolated from *Pyrococcus furiosus*. *Gene*, **108**, 1–6.

Lundgren, M., Andersson, A., Chen, L., Nilsson, P., et ander, R. Bern. 2004. Three replication origins in *Sulfolobus* species : Synchronous initiation of chromosome replication and asynchronous termination. *Proc. Natl. Acad. Sci. USA*, **101**, 7046–7051.

MacAlpine, D.M., Rodriguez, H.K., et Bell, S.P. 2004. Coordination of replication and transcription along a *Drosophila* chromosome. *Genes Dev.*, **18**, 3094–3105.

Majewski, J. 2003. Dependence of mutational asymmetry on gene expression levels in the human genome. *Am. J. Hum. Genet.*, **73**, 688–692.

Maki, H., Horiuchi, T., et Kornberg, A. 1985. The polymerase subunit of DNA polymerase III of *Escherichia coli*. I. Amplification of the *dnaE* gene product and polymerase activity of the α subunit. *J. Biol. Chem.*, **260**, 12982–12986.

Maki, H., Maki, S., et Kornberg, A. 1988. DNA polymerase III holoenzyme of *Escherichia coli*. *J. Biol. Chem.*, **263**, 6570–6578.

Maliszewska-Tkaczyk, M., Jonczyk, P., Bialoskorska, M., Schaaper, R.M., et Fijalkowska, I.J. 2000. SOS mutator activity : unequal mutagenesis on leading and lagging strands. *Proc. Natl. Acad. Sci. USA*, **97**, 12678–12683.

Margulies, M., Egholm, M., Altman, W.E., *et al.*. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.

Marians, K. J. 1992. Prokaryotic DNA replication. *Ann. Rev. Biochem.*, **61**, 673–719.

Mathew, R., et Chatterji, D. 2006. The evolving story of the omega subunit of bacterial RNA polymerase. *Trends Microbiol.*, **14**, 450–455.

Mattila, P., Korpela, J., Tenkanen, T., et Pitkanen, K. 1991. Fidelity of DNA synthesis by the *Thermococcus litoralis* DNA polymerase- an extremely heat stable enzyme with proofreading activity. *Nucleic Acids Res.*, **19**, 4967–4973.

McHenry, C., et Kornberg, A. 1977. DNA polymerase III holoenzyme of *Escherichia coli*. Purification and resolution into subunits. *J. Biol. Chem.*, **252**, 6478–6484.

McHenry, C.S. 1982. Purification and characterization of DNA polymerase III'. Identification of τ as a subunit of the DNA polymerase III holoenzyme. *J. Biol. Chem.*, **257**, 2657–2663.

McHenry, C.S. 2003. Chromosomal replicases as asymmetric dimers : studies of subunit arrangement and functional consequences. *Mol. Microbiol.*, **49**, 1157–1165.

McInerney, P., et O'Donnell, M. 2004. Functional uncoupling of twin polymerases. Mechanism of polymerase dissociation from a lagging-strand block. *J. Biol. Chem.*, **279**, 21543–21551.

McInerney, P., et O'Donnell, M. 2007. Replisome fate upon encountering a leading strand block and clearance from DNA by recombination proteins. *J. Biol. Chem.*, **282**, 25903–25916.

McLean, M. J., Wolfe, K. H., et Devine, K. M. 1998. Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. *J. Mol. Evol.*, **47**, 691–696.

Mellon, I., et Hanawalt, P. C. 1989. Induction of the *Escherichia coli* lactose operon selectively increases repair of its transcribed DNA strand. *Nature*, **342**, 95–98.

Mellon, I., Bohr, V. A., Smith, C.A., et Hanawalt, P. C. 1986. Preferential DNA repair of an active gene in human cells. *Proc. Natl. Acad. Sci. USA*, **83**, 8878–8882.

Mellon, I., Spivak, G., et Hanawalt, P.C. 1987. Selective removal of transcription-blocking DNA damage from the transcribed strand of the mammalian *dhfr* gene. *Cell*, **51**, 241–249.

Mendelman, L.V., Boosalis, M.S., Petruska, J., et Goodman, M.F. 1989. Nearest neighbor influences on DNA polymerase insertion fidelity. *J. Biol. Chem.*, **264**, 14415–14423.

Meselson, M., et Stahl, F.W. 1958. The replication of DNA in *Escherichia coli*. *Proc. Natl. Acad. Sci. USA*, **44**, 671–682.

Meunier, J., et Duret, L. 2004. Recombination drives the evolution of the GC-content in the human genome. *Mol. Biol. Evol.*, **21**, 984–990.

Mirkin, E. V., et Mirkin, S. M. 2005. Mechanisms of transcription-replication collisions in Bacteria. *Mol. Cell. Biol.*, **25**, 888–895.

Mitchell, A., et Graur, D. 2005. Inferring the pattern of spontaneous mutation from the pattern of substitution in unitary pseudogenes of *Mycobacterium leprae* and a comparison of mutation patterns among distantly related organisms. *J. Mol. Evol.*, **61**, 795–803.

Modrich, P., et Lahue, R. 1996. Mismatch repair in replication fidelity, genetic recombination and cancer biology. *Annu. Rev. Biochem.*, **65**, 101–133.

Morrison, A., Araki, H., Clark, A.B., Hamatake, R. K., et Sugino, A. 1990. A third essential DNA polymerase in *S. cerevisiae*. *Cell*, **62**, 1143–1151.

Morrison, A., Bell, J.B., Kunkel, T. A., et Sugino, A. 1991. Eukaryotic DNA polymerase amino acid sequence required for 3'->5' exonuclease activity. *Proc. Natl. Acad. Sci. USA*, **88**, 9473–9477.

Morton, B.R., et Clegg, M.T. 1995. Neighboring base composition is strongly correlated with base substitution bias in a region of the chloroplast genome. *J. Mol. Evol.*, **41**, 597–603.

Muggeo, V. M. R. 2003. Estimating regression models with unknown breakpoints. *Stat. Med.*, **22**, 3055–3071.

Muggeo, Vitto M. R. 2004. *segmented : Segmented relationships in regression models*. R package version 0.1-4.

Nagata, Y., Kawaguchi, G., Tago, Y., Imai, M., Watanabe, T., Sakurai, S., Ihara, M., Kawata, M., et Yamamoto, K. 2005. Absence of strand bias for deletion mutagenesis during chromosomal leading and lagging strand replication in *Escherichia coli*. *Genes Genet. Syst.*, **80**, 1–8.

Necsulea, A., et Lobry, J. R. 2007. A new method for assessing the effect of replication on DNA base composition asymmetry. *Mol. Biol. Evol.*, **24**, 2169–2179.

Nethanel, T., et Kaufmann, G. 1990. Two DNA polymerases may be required for synthesis of the lagging DNA strand of simian virus 40. *J. Virol.*, **64**, 5912–5918.

Nikolaou, C., et Almirantis, Y. 2005. A study on the correlation of nucleotide skews and the positioning of the origin of replication : different modes of replication in bacterial species. *Nucleic Acids Res.*, **33**, 6816–6822.

Niu, D.K., Lin, K., et Zhang, D-Y. 2003. Strand compositional asymmetries of nuclear DNA in eukaryotes. *J. Mol. Evol.*, **57**, 325–334.

Nomura, M., et Morgan, E. A. 1977. Genetics of bacterial ribosomes. *Ann. Rev. Genet.*, **11**, 297–347.

Nosek, J., et Tomaska, L. 2003. Mitochondrial genome diversity : evolution of the molecular architecture and replication strategy. *Curr. Genet.*, **44**, 73–84.

Okazaki, R., Okazaki, T., Sakabe, K., et Sugimoto, K. 1967. Mechanism of DNA replication - possible discontinuity of DNA chain growth. *Jpn. J. Med. Sci.*, **20**, 255–260.

Okazaki, R., Okazaki, T., Sakabe, K., Sugimoto, K., et Sugino, A. 1968. Mechanism of DNA chain growth, I. Possible discontinuity and unusual secondary structure of newly synthesized chains. *Proc. Natl. Acad. Sci. USA*, **59**, 598–605.

Olavarrieta, L., Hernandez, P., Krimer, D.B., et Schvartzman, J.B. 2002. DNA knotting caused by head-on collision of transcription and replication. *J. Mol. Biol.*, **322**, 1–6.

Oller, A.R., Fijalkowska, I.J., Dunn, R.L., et Schaaper, R.M. 1992. Transcription-repair coupling determines the strandedness of ultraviolet mutagenesis in *Escherichia coli*. *Proc. Natl. Acad. Sci. USA*, **89**, 11036–11040.

Olsen, G.J., et Woese, C. R. 1997. Archaeal genomics : an overview. *Cell*, **89**, 991–994.

Pages, V., et Fuchs, R.P. 2003. Uncoupling of leading- and lagging-strand DNA replication during lesion bypass *in vivo*. *Science*, **300**, 1300–1303.

Patten, J.E., So, A.G., et Downey, K.M. 1984. Effect of base-pair stability of nearest-neighbor nucleotides on the fidelity of deoxyribonucleic acid synthesis. *Biochemistry*, **23**, 1613–1618.

Pavlov, Y., Mian, I.M., et Kunkel, T.A. 2003. Evidence for preferential mismatch repair of lagging strand DNA replication errors in yeast. *Curr. Biol.*, **13**, 744–748.

Pavlov, Y.I., Newlon, C.S., et Kunkel, T.A. 2002. Yeast origins establish a strand bias for replicational mutagenesis. *Mol. Cell*, **10**, 207–213.

Perriere, G., Lobry, J.R., et Thioulouse, J. 1996. Correspondence discriminant analysis : a multivariate method for comparing classes of protein and nucleic acid sequences. *CABIOS*, **12**, 519–524.

Pless, R. C., et Bessman, M. J. 1983. Influence of local nucleotide sequence on substitution of 2-aminopurine for adenine during deoxyribonucleic acid synthesis in vitro. *Biochemistry*, **22**, 4905–4915.

Prabhu, V.V. 1993. Symmetry observations in long nucleotide sequences. *Nucleic Acids Res.*, **21**, 2797–2800.

Prescott, E.M., et Proudfoot, N.J. 2002. Transcriptional collision between convergent genes in budding yeast. *Proc. Natl. Acad. Sci. USA*, **99**, 8796–8801.

Pursell, Z.F., Isoz, I., Lundstrom, E-B., Johansson, E., et Kunkel, T.A. 2007. Yeast DNA polymerase ϵ participates in leading-strand DNA replication. *Science*, **317**, 127–130.

R Development Core Team. 2005. *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Radman, M. 1998. DNA replication : one strand may be more equal. *Proc. Natl. Acad. Sci. USA*, **95**, 9718–9719.

Radman, M., et Wagner, R. 1986. Mismatch repair in *Escherichia coli*. *Ann. Rev. Genet.*, **20**, 523–538.

Rattray, A. J., et Strathern, J. N. 2003. Error-prone DNA polymerases : when making a mistake is the only way to get ahead. *Annu. Rev. Genet.*, **37**, 31–66.

Roberts, J.D., Thomas, D.C., et Kunkel, T.A. 1991. Exonucleolytic proofreading of leading and lagging strand DNA replication errors. *Proc. Natl. Acad. Sci. USA*, **88**, 3465–3469.

Robinson, N. P., Dionne, I., Lundgren, M., Marsh, V. L., Bernander, R., et Bell, S. D. 2004. Identification of two origins of replication in the single chromosome of the archaeon *Sulfolobus solfataricus*. *Cell*, **116**, 25–38.

Rocha, E. P., Danchin, A., et Viari, A. 1999. Universal replication biases in bacteria. *Mol. Microbiol.*, **32**, 11–16.

Rocha, E. P. C. 2002. Is there a role for replication fork asymmetry in the distribution of genes in bacterial genomes ? *Trends Microbiol.*, **10**, 393–395.

Rocha, E. P. C. 2004. The replication-related organization of bacterial genomes. *Microbiol.*, **150**, 1609–1627.

Rocha, E. P. C., et Danchin, A. 2003a. Essentiality, not expressiveness, drives gene-strand bias in bacteria. *Nature Genet.*, **34**, 377–378.

Rocha, E. P. C., et Danchin, A. 2003b. Gene essentiality determines chromosome organisation in bacteria. *Nucleic Acids Res.*, **31**, 6570–6577.

Rocha, E.P.C., Touchon, M., et Feil, E.J. 2006. Similar compositional biases are caused by very different mutational effects. *Genome Res.*, **16**, 1537–1547.

Rosche, W.A., Trinh, T.Q., et Sinden, R.A. 1995. Differential DNA secondary structure-mediated deletion mutation in the leading and lagging strands. *J. Bacteriol.*, **177**, 4385–4391.

Rudner, R., Karkas, J.D., et Chargaff, E. 1968. Separation of *B. subtilis* DNA into complementary strands, III. Direct analysis. *Proc. Natl. Acad. Sci. USA*, **60**, 921–922.

Rudner, R., Karkas, J.D., et Chargaff, E. 1969. Separation of microbial deoxyribonucleic acids into complementary strands. *Proc. Natl. Acad. Sci. USA*, **63**, 152–159.

Rudolph, C.J., Dhillon, P., Moore, T., et Lloyd, R.G. 2007. Avoiding and resolving conflicts between DNA replication and transcription. *DNA Repair*, **6**, 981–993.

Rytkonen, A. K., Vaara, M., Nethanel, T., Kaufmann, G., Sormunen, R., Laara, E., Nasheuer, H-P., Rahmeh, A., Lee, M. Y. W. T., Syvaoja, J. E., et Pospiech, H. 2006. Distinctive activities of DNA polymerases during human DNA replication. *FEBS J.*, **273**, 2984–3001.

Salzberg, S.L., Salzberg, A.J., Kerlavage, A.R., et Tomb, J-F. 1998. Skewed oligomers and origins of replication. *Gene*, **217**, 57–67.

Sanjanwala, B., et Ganesan, A.T. 1989. DNA polymerase III gene of *Bacillus subtilis*. *Proc. Natl. Acad. Sci. USA*, **86**, 4421–4424.

Scheuermann, R. H., et Echols, H. 1984. A separate editing exonuclease for DNA replication : the ϵ subunit of *Escherichia coli* DNA polymerase III holoenzyme. *Proc. Natl. Acad. Sci. USA*, **81**, 7747–7751.

Schlotterer, C. 2000. Evolutionary dynamics of microsatellite DNA. *Chromosoma*, **109**, 365–371.

Schmidt, E. E. 1996. Transcriptional promiscuity in testes. *Curr. Biol.*, **6**, 768–771.

Seeberg, E., Eide, L., et Bjoras, M. 1995. The base excision repair pathway. *Trends Biochem. Sci.*, **20**, 391–397.

Selby, C.P., et Sancar, A. 1993. Molecular mechanism of transcription-repair coupling. *Science*, **260**, 53–58.

Setlow, R. B. 1966. Cyclobutane-type pyrimidine dimers in polynucleotides. *Science*, **153**, 379–386.

Shcherbakova, P.V., et Pavlov, Y.I. 1996. $3' \rightarrow 5'$ exonucleases of DNA polymerases ϵ and δ correct base analog induced DNA replication errors on opposite DNA strands in *Saccharomyces cerevisiae*. *Genetics*, **142**, 717–726.

Shcherbakova, P.V., Pavlov, Y.I., Chilkova, O., Rogozin, I.B., Johansson, E., et Kunkel, T.A. 2003. Unique error signature of the four-subunit yeast DNA polymerase ϵ. *J. Biol. Chem.*, **278**, 43770–43780.

Shepherd, L.D., et Lambert, D.M. 2005. Mutational bias in penguin microsatellite DNA. *J. Hered.*, **96**, 566–571.

Shimizu, K., Hashimoto, K., Kirchner, J.M., Nakai, W., Nishikawa, H., Resnick, M. A., et Sugino, A. 2002. Fidelity of DNA polymerase ϵ holoenzyme from budding yeast *Saccharomyces cerevisiae*. *J. Biol. Chem.*, **277**, 37422–37429.

Shioiri, C., et Takahata, N. 2001. Skew of mononucleotide frequencies, relative abundance of dinucleotides and DNA strand asymmetry. *J. Mol. Evol.*, **53**, 364–376.

Siebenlist, U. 1979. RNA polymerase unwinds an 11-base pair segment of a phage T7 promoter. *Nature*, **279**, 651–652.

Smith, N.G.C., Webster, M.T., et Ellegren, H. 2002. Deterministic mutation rate variation in the human genome. *Genome Res.*, **12**, 1350–1356.

Streisinger, G., Okada, Y., Emrich, J., Newton, J., Tsugita, A., Terzhaghi, E., et Inouye, M. 1966. Frameshift mutations and the genetic code. *Cold Spring Harbor Symp. Quant. Biol.*, **31**, 77–84.

Sueoka, N. 1995. Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. *J. Mol. Evol.*, **40**, 318–325.

Tanaka, M., et Ozawa, T. 1994. Strand asymmetry in human mitochondrial DNA mutations. *Genomics*, **22**, 327–335.

Tautz, D., et Renz, M. 1984. Simple sequences are ubiquitous repetitive components of eukaryotic genomes. *Nucleic Acids Res.*, **12**, 4127–4138.

Taylor, J.S., J.M.H., et Breden, F. 1999. The death of a microsatellite : a phylogenetic perspective on microsatellite interruptions. *Mol. Biol. Evol.*, **16**, 567–572.

The International HapMap Consortium. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851–861.

Tillier, E. R. M., et Collins, R. A. 2000. The contributions of replication orientation, gene direction and signal sequences to base-composition asymmetries in bacterial genomes. *J. Mol. Evol.*, **50**, 249–257.

Touchon, M., Nicolay, S., Arneodo, A., d'Aubenton Carafa, Y., et Thermes, C. 2003. Transcription-coupled TA and GC strand asymmetries in the human genome. *FEBS Lett.*, **555**, 579–582.

Touchon, M., Arneodo, A., d'Aubenton Carafa, Y., et Thermes, C. 2004. Transcription-coupled and splicing-coupled strand asymmetries in eukaryotic genomes. *Nucleic Acids Res.*, **32**, 4969–4978.

Touchon, M., Nicolay, S., Audit, B., of Brodie, E-B. Brodie, d'Aubenton Carafa, Y., Arneodo, A., et Thermes, C. 2005. Replication-associated strand asymmetries in mammalian genomes : toward detection of replication origins. *Proc. Natl. Acad. Sci. USA*, **102**, 9836–9841.

Uemori, T., Sato, Y., Kato, I., Doi, H., et Ishino, Y. 1997. A novel DNA polymerase in the hyperthermophilic archaeon, *Pyrococcus furiosus* : gene cloning, expression and characterization. *Genes Cells*, **2**, 499–512.

Ullah, S., Gallinari, P., Xu, Y-Z., Goodman, M.F., Bloom, L.B., Jiricny, J., et Day, R. S. 1996. Base analog and neighboring base effects on substrate specificity of recombinant human G :T mismatch-specific thymine DNA-glycosylase. *Biochemistry*, **35**, 12926–12932.

Veaute, X., et Fuchs, R.P.P. 1993. Greater susceptibility to mutations in lagging strand of DNA replication in *Escherichia coli* than in leading strand. *Science*, **261**, 598–599.

Veaute, X., Mari-Giglia, G., Lawrence, C.W., et Sarasin, A. 2000. UV lesions located on the leading strand inhibit DNA replication but do not inhibit SV40 T-antigen helicase activity. *Mutat. Res.*, **459**, 19–28.

Velculescu, V. E., Zhang, L., Vogelstein, B., et Kinzler, K. W. 1995. Serial analysis of gene expression. *Science*, **270**, 484–487.

Vischer, E., Zamenhof, S., et Chargaff, E. 1949. Microbial nucleic acids : the desoxypentose nucleic acids of avian tubercle bacilli and yeast. *J. Biol. Chem.*, **177**, 429–438.

Waga, S., et Stillman, B. 1994. Anatomy of a DNA replication fork revealed by reconstitution of SV40 DNA replication *in vitro*. *Nature*, **369**, 207–212.

Wagner, J., Kamiya, H., et Fuchs, R.P.P. 1997. Leading *versus* lagging strand mutagenesis induced by 7,8-dihydro-8-oxo-2'-deoxyguanosine in *Escherichia coli*. *J. Mol. Biol.*, **265**, 302–309.

Wang, J. C., Jacobsen, J. H., et Saucier, J-M. 1977. Physicochemical studies on interactions between DNA and RNA polymerase. Unwinding of the DNA helix by *Escherichia coli* RNA polymerase. *Nucleic Acids Res.*, **4**, 1225–1241.

Wang, J.D., Berkmen, M.B., et Grossman, A.D. 2007. Genome-wide coorientation of replication and transcription reduces adverse effects on replication in *Bacillus subtilis*. *Proc. Natl. Acad. Sci. USA*, **104**, 5608–5613.

Wang, T-C. V. 2005. Discontinuous or semi-discontinuous DNA replication in *Escherichia coli* ? *BioEssays*, **27**, 633–636.

Watson, J.D., et Crick, F.H.C. 1953. Molecular structure of nucleic acids : a structure for deoxyribose nucleic acid. *Nature*, **171**, 737–738.

Webster, M. T., et Smith, N. G. C. 2004. Fixation biases affecting human SNPs. *Trends Genet.*, **20**, 122–126.

Worning, P., Jensen, L.J., Hallin, P.F., Staerfeldt, H-H., et Ussery, D.W. 2006. Origin of replication in circular prokaryotic chromosomes. *Env. Microbiol.*, **8**, 353–361.

Wu, C. I. 1991. DNA strand asymmetry. *Nature*, **352**, 114.

Wu, C. I., et Maeda, N. 1987. Inequality in mutation rates of the two strands of DNA. *Nature*, **327**, 169–170.

Yeo, G. W., Nostrand, E. L. Van, et Liang, T. Y. 2007. Discovery and analysis of evolutionarily conserved intronic splicing regulatory elements. *PLoS Genet.*, **3**, e85.

Young, R. A. 1991. RNA polymerase II. *Annu. Rev. Biochem.*, **60**, 689–715.

Yura, T., et Ishihama, A. 1979. Genetics of bacterial RNA polymerases. *Annu. Rev. Genet*, **13**, 59–97.

Yuzhakov, A., Turner, J., et O'Donnell, M. 1996. Replisome assembly reveals the basis for asymmetric function in the leading and lagging strand replication. *Cell*, **86**, 877–886.

Zaychikov, E., Denissova, L., et Heumann, H. 1995. Translocation of the *Escherichia coli* transcription complex observed in the registers 11 to 20 : "Jumping" of RNA polymerase and asymmetric expansion and contraction of the "transcription bubble". *Proc. Natl. Acad. Sci. USA*, **92**, 1739–1743.

Zaychikov, E., Denissova, L., Meier, T., Gotte, M., et Heumann, H. 1997. Influence of Mg2+ and temperature on formation of the transcription bubble. *J. Biol. Chem.*, **272**, 2259–2267.

Zhang, C., Li, W-H., Krainer, A. R., et Zhang, M. Q. 2008. RNA landscape of evolution for optimal exon and intron discrimination. *Proc. Natl. Acad. Sci. USA*, **105**, 5797–5802.

Zhang, R., et Zhang, C. T. 2005. Identification of replication origins in archaeal genomes based on the Z-curve method. *Archaea*, **1**, 335–346.

TITRE en français : Étude des patrons d'évolution asymétrique dans les séquences d'ADN.

RÉSUMÉ en français : Cette thèse étudie l'effet de deux processus cellulaires essentiels, la réplication et la transcription, sur la composition en nucléotides des séquences d'ADN. Ces mécanismes ont un fonctionnement asymétrique par rapport aux deux brins d'ADN, et ils ont comme conséquence une composition asymétrique dans les séquences.Nous avons étudié la co-orientation entre réplication et transcription chez les procaryotes. Nous proposons une méthode pour l'étude des biais de composition qui découple ces deux sources d'asymétrie. Nous montrons que les biais associés à la réplication sont très variables, même entre espèces proches. Nous avons ensuite analysé le patron de substitution dans les régions transcrites et autour des origines de réplication du génome humain, et notamment l'effet du contexte 5'-3'. Les biais de voisinage sont similaires pour l'asymétrie associée à la réplication et à la transcription. La variation des taux de substitutions en fonction du patron d'expression des gènes suggère qu'un biais de réparation asymétrique et contexte-dépendant pourrait être en jeu. Enfin, nous avons proposé une méthode de calcul du patron de substitution dans des séquences à composition biaisée : les microsatellites. Nous avons démontré que les microsatellites transcrits sont sujets au mêmes processus asymétriques que les régions non-répétées.

MOTS-CLEFS en français : asymétrie, réplication, transcription.

TITRE en anglais : Patterns of asymmetric DNA sequence evolution

RÉSUMÉ en anglais : This thesis analyses the effect of two essential cellular mechanisms, replication and transcription, on the base composition of DNA sequences. These two processes function asymmetrically on the two DNA strands, and they have as a consequence an asymmetric nucleotide composition in genomic sequences.First, we studied the co-orientation between replication and transcription in prokaryotes. We proposed a method for the study of base composition biases which can separate these two sources of asymmetry. We show that the effect of replication on base composition can be highly variable even between closely related species. We then studied the substitution pattern in transcribed regions and around replication origins, in human, with special emphasis on the effect of the 5'-3' nucleotide context. Patterns of context-dependent asymmetric substitutions are similar for replication and transcription. The variation of substitution rate with the expression pattern suggests the presence of context-dependent asymmetric repair mechanisms. We proposed a computational approach for the study of the substitution pattern in microsatellites. We prove that transcribed microsatellites are subject to asymmetric evolution.

MOTS-CLEFS en anglais : asymmetry, DNA replication, transcription.

DISCIPLINE : Bioinformatique

INTITULE ET ADRESSE DE L'UFR OU DU LABORATOIRE

Laboratoire de Biométrie et Biologie Évolutive - UMR 5558 CNRS

Bâtiment Gregor Mendel - Université Claude Bernard Lyon 1

43, bd. du 11 Novembre 1918 - 69622 Villeurbanne CEDEX