Réarrangements chromosomiques dans les génomes de mammifères : caractérisation des points de cassure

Claire Lemaitre

N^o d'ordre : 218-2008 Année 2008

Thèse

Présentée

devant l'Université Claude Bernard - Lyon 1

pour l'obtention

du Diplôme de Doctorat
(arrêté du 7 août 2006)

et soutenue publiquement le
6 novembre 2008

par

Claire Lemaitre

Réarrangements chromosomiques dans les génomes de mammifères : caractérisation des points de cassure

---

Directrice de thèse : Marie-France Sagot
Co-directeur : Christian Gautier

Jury :	Pierre Capy,	Examinateur
	Nicolas Galtier,	Rapporteur
	Christian Gautier,	Directeur
	Roderic Guigó,	Rapporteur
	Hubert Pinon,	Président
	Eric Rivals,	Rapporteur
	Marie-France Sagot,	Directrice

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

A ma maman,

Remerciements

Je remercie tout d'abord mes directeurs de thèse, Marie-France et Christian qui m'ont encadrée et guidée dans ce travail durant ces trois années. Je les remercie sincèrement tous les deux pour la qualité de leur encadrement tant sur le plan scientifique que sur le plan humain. Merci en particulier à Marie-France qui a toujours été présente quand j'en avais besoin.

Je remercie mes rapporteurs Nicolas Galtier, Roderic Guigó et Eric Rivals, ainsi que mes examinateurs, Pierre Capy et Hubert Pinon, pour avoir accepté de lire ce manuscrit et d'évaluer ce travail.

Je remercie également les membres de mes comités de pilotage, en particulier Christophe Hitte, Stéphane Robin et Thomas Faraut qui ont fait tant de kilomètres à deux reprises pour m'apporter leur regard et critiques constructives aux étapes clés de ma thèse. Merci également à Laurent Duret pour les discussions toujours enrichissantes.

Ce travail est le fruit de diverses collaborations. Ainsi, Eric Tannier a été un collaborateur privilégié depuis le début de mon master et je le remercie notamment pour son aide en algorithmique. Merci également à Gabriel Marais pour m'avoir invitée sur ce sujet passionnant des chromosomes sexuels. Je remercie mes collaborateurs spécialistes des origines de réplication, Lamia, Benjamin Audit et Alain Arnéodo. Enfin, un merci particulier à Pierre pour les innombrables versions de classis et classus.

Je remercie également Misou, Nathalie, Agnès, Isabelle et Gaëlle, pour le soutien admnistratif, ainsi que Stéphane, Bruno, Simon et Lionel pour le soutien technique et informatique.

Durant cette thèse, j'ai eu l'opportunité de faire de l'enseignement, et je remercie à cette occasion mes collègues enseignants, en particulier Jean-François Pageaux et Hubert Charles pour m'avoir accordé leur confiance. Je tiens également à remercier Sandrine Charles, pour m'avoir sensibilisée à la pédagogie et à de nouvelles méthodes d'apprentissage.

Enfin, j'ai passé de très bons moments au sein de l'équipe Baobab et du laboratoire BBE et je tiens à remercier tous ceux qui ont rendu ce cadre de travail si agréable. Merci en particulier à Léo, les Vincents (L., N. et M.), Vicente, Paulo, Patricia, Janice, Thibaut, Perrine, Christelle et également à Marília pour ses récréations de portugais et de futebal. Mes voisins de bureau, Manu et Ludo, méritent également un grand merci pour m'avoir supportée pendant ces 3 ans, malgré mes "grattages" incessants et mes sautes d'humeur. Merci à Anouk, toujours disponible pour répondre à mes questions. Merci à Elise, pour tous ces moments de détente et également pour les enseignements en duo. Merci à Manue, l'altiligérienne, à Pantxo et Fanny pour le renfort berjallien.

Je remercie ma famille, Papa, Laure, et aussi Marcelle et Christian pour leur soutien. Et enfin un très grand merci à Fred pour sa présence et son soutien de tous les jours.

Table des matières

Introduction

Historiquement, le terme “génome” désignait l'ensemble des gènes d'un organisme. Cependant, il s'est avéré que le matériel génétique d'un organisme ne se limite pas aux gènes. Chez les eucaryotes, et notamment chez les vertébrés, la proportion de séquences codantes est très faible (moins de 2 % chez l'homme). Ainsi, on considère le génome comme l'ensemble de l'information génétique transmise à la descendance ; celle-ci est portée par un ensemble de molécules d'ADN, les chromosomes. Si le génome n'est pas seulement l'ensemble des gènes, il n'est pas non plus seulement un “sac” de séquences d'ADN. C'est un ensemble structuré et organisé, et ce à plusieurs niveaux : du niveau des chromosomes jusqu'à la structure complexe des gènes, en passant par divers paysages génomiques, comme les domaines de chromatine ou les isochores.

Cette structure n'est pas statique. Elle peut-être modifiée par des mutations appelées réarrangements chromosomiques. Ces derniers affectent des segments génomiques de quelques centaines de paires de base à des bras chromosomiques voire des chromosomes entiers, qui peuvent être déplacés, inversés, délétés ou dupliqués.

Ainsi, si on compare les génomes d'espèces différentes, on remarque une très grande diversité de caryotypes et de structure même entre espèces proches. La cartographie comparée, puis la génomique comparée avec le séquençage des génomes complets, consistent à comparer l'organisation et la structure de génomes d'espèces différentes. L'objectif est de comprendre comment les génomes évoluent, mais également comment ils fonctionnent. En effet, l'analyse des similitudes et des différences au niveau de l'organisation des génomes peut être interprétée en terme de contraintes fonctionnelles et nous informer sur l'origine et la fonction de ces structures.

Dans le but d'identifier des réarrangements chromosomiques, les méthodes de comparaison de génomes se sont considérablement sophistiquées et améliorées depuis la simple comparaison à l'oeil des caryotypes (du nombre et de la taille des chromosomes) jusqu'au séquençage de génomes complets et l'utilisation de méthodes bioinformatiques. La résolution augmentant, le nombre de réarrangements identifiés n'a fait que croître révélant leur importance dans les processus évolutifs et dans la diversité génétique. Jusque très récemment, on ne soupçonnait pas une telle quantité de réarrangements polymorphes dans les populations humaines. Ils sont également étudiés dans le domaine de la santé humaine, puisqu'ils peuvent être responsables de maladies génétiques et sont très nombreux dans les cellules cancéreuses. Enfin, certains réarrangements pourraient même engendrer ou faciliter les processus de spéciation.

Nous étudierons ici les réarrangements chromosomiques dans le cadre de l'évolution des génomes, et plus précisément l'évolution des génomes de mammifères. Si les mécanismes moléculaires à l'origine des réarrangements chromosomiques sont relativement bien compris à l'heure actuelle, il reste de nombreuses inconnues quant aux forces évolutives gouvernant ces mutations. Par exemple, on ne sait pas expliquer les différences de fréquences de réarrange-

ments observées entre lignées différentes, ni les différences de types de réarrangements. De même, la distribution des réarrangements le long des génomes est encore très mal comprise. Alors qu'on supposait que ces évènements se distribuaient de façon uniforme et indépendante le long du génome, des analyses à haute résolution, notamment grâce au séquençage de génomes complets, ont ouvert un débat à ce sujet. Ces analyses ont en effet révélé l'existence de régions sur le génome subissant plus fréquemment des réarrangements ; ces régions seraient plus fragiles que d'autres.

Afin d'étudier ces problématiques, nous nous sommes intéressés, au cours de cette thèse, spécifiquement aux régions du génome qui ont subit de tels évènements ; ces régions sont communément appelées des points de cassure (breakpoints en anglais). Nous nous sommes demandés si ces régions possèdent des caractéristiques différentes des autres régions génomiques. Si de telles caractéristiques existent, elles pourraient être reliées aux mécanismes moléculaires de réarrangements, soit en ayant favorisé ces évènements, soit en étant la conséquence de ces derniers. De plus, si les points de cassure ne sont pas distribués aléatoirement dans le génome, on peut espérer trouver dans ces régions des caractéristiques expliquant pourquoi une région a été cassée et pas sa voisine.

Nous avons choisi d'aborder ces questions de manière systématique à l'échelle d'un génome, en analysant simultanément l'ensemble des points de cassure d'un génome. Cette approche permet d'évaluer statistiquement les caractéristiques détectées et d'effectuer des analyses exploratoires sans *a priori*. Ce travail a nécessité l'utilisation d'outils mathématiques et informatiques pour manipuler et traiter des données volumineuses, telles que les séquences des génomes complets, mais également de développer de nouvelles méthodes bioinformatiques.

Une grande partie de ce travail a ainsi consisté à développer une méthode de détection précise des points de cassure sur un génome par comparaison avec le génome d'une autre espèce. Cette méthode, contrairement à celles existantes, sépare les deux tâches de détection des régions conservées et de délimitation des points de cassure. Elle permet d'obtenir des points de cassure à la fois fiables et précis. L'amélioration de la résolution des points de cassure a permis, dans un deuxième temps, d'analyser plus finement leurs séquences et leur distribution le long des génomes.

Le manuscrit est organisé en six chapitres. Les deux premiers présentent le contexte bibliographique au niveau biologique d'une part, et méthodologique d'autre part. Ainsi, nous commençons, dans le premier chapitre, par introduire les objets biologiques au coeur de cette thèse, les réarrangements chromosomiques. Nous décrivons les différents types de réarrangements ainsi que les mécanismes moléculaires pouvant les engendrer, puis leurs impacts au niveau cellulaire et populationnel, avant de nous intéresser plus en détail aux analyses des points de cassure de mammifères. Dans le deuxième chapitre, nous établissons un état de l'art sur les méthodes existantes permettant de détecter des réarrangements et leurs points de cassure par comparaison de génomes d'espèces différentes. Cette étude nous a amenés à proposer notre propre méthode de détection des points de cassure, qui est décrite dans les deux chapitres suivants. En effet, elle se décompose en deux étapes distinctes et indépendantes. La première propose un algorithme de comparaison de l'ordre des gènes entre deux génomes, permettant de définir des blocs de synténie et des points de cassure fiables. Dans la deuxième étape, décrite dans le Chapitre 4 les coordonnées des points de cassure sont affinées individuellement par alignement des séquences et en utilisant un algorithme de partitionnement. Enfin, les deux derniers chapitres présentent les résultats de caractérisation des points

de cassure obtenus. Dans le Chapitre 5, nous avons cherché à caractériser les séquences des points de cassure, notamment par rapport à des séquences non réarrangées et par rapport à leurs séquences adjacentes. Dans le Chapitre 6, nous étudions la distribution des points de cassure par rapport à différents niveaux d'organisation du génome.

Chapitre 1

Les réarrangements chromosomiques

Sommaire

Dans ce chapitre, nous définissons les objets biologiques qui sont au coeur de cette thèse : les réarrangements chromosomiques. Nous nous intéressons aux mécanismes moléculaires qui sont responsables de leur formation et aux impacts de tels évènements. Enfin, nous nous concentrerons sur les réarrangements impliqués dans l'évolution des génomes. Ces aspects seront étudiés principalement chez les eucaryotes et plus précisément chez les mammifères.

1.1 Qu'est-ce qu'un réarrangement ?

1.1.1 Le génome dynamique

Le génome subit constamment des mutations qui peuvent être d'origine exogène (agressions de l'environnement) ou bien endogène (dysfonctionnement de certains processus cellulaires). Ces mutations sont le plus souvent réparées par les mécanismes de réparation de la cellule. Dans certains cas, elles échappent à la réparation et, si elles se produisent dans les cellules germinales, peuvent être transmises à la descendance.

On distingue classiquement deux types de mutation en fonction de la taille de la région génomique affectée par celle-ci. Les mutations ponctuelles affectent un seul nucléotide à la fois. On regroupe sous ce terme, les substitutions (le remplacement d'un nucléotide par un autre), les délétions et les insertions d'un nucléotide. Par contre, les mutations chromosomiques affectent des segments d'ADN de plusieurs nucléotides, jusqu'à des chromosomes entiers. Une troisième classe de mutations est souvent distinguée des deux autres : ce sont les mutations spécifiques des répétitions. On y regroupe les phénomènes d'expansion et de contraction des répétitions en tandem (satellites, mini-satellites et micro-satellites) et les transpositions d'éléments transposables.

Les mutations chromosomiques regroupent les modifications en nombre des chromosomes d'une part et les modifications de leur structure d'autre part. Les modifications du nombre de chromosomes sont de deux types : la polyploïdie et la polysomie. La polyploïdie consiste à dupliquer tout le génome (obtenant un génome à 3n, ou 4n, etc. chromosomes au lieu de 2n pour un génome diploïde). La polysomie ou aneuploïdie concerne seulement un ou plusieurs chromosomes. Par exemple, la trisomie est le fait d'avoir trois copies d'un chromosome au lieu de deux dans un génome diploïde.

Les modifications de structure des chromosomes sont regroupées sous les termes **réarrangements** ou **remaniements chromosomiques**.

1.1.2 Les différents types de réarrangements

Il existe différents types de réarrangements. Nous les décrivons ici en fonction des modifications génomiques qui en résultent, les mécanismes moléculaires en jeu seront décrits dans la section suivante. On peut tout d'abord distinguer les réarrangements **équilibrés** de ceux **déséquilibrés**.

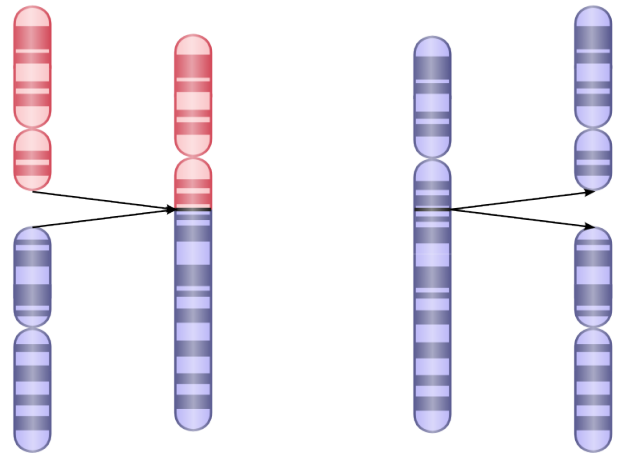
a. Réarrangements équilibrés

Les réarrangements équilibrés préservent la quantité d'ADN. Ils ne font que modifier la structure des chromosomes, sans perte ni ajout d'ADN.

La **fusion** et la **fission** de chromosomes sont deux réarrangements équilibrés qui affectent des chromosomes entiers. La fusion est la réunion de deux chromosomes en un seul, au niveau de leur télomère (voir Figure 1.1a). La fission est l'opération inverse qui consiste à "couper" un chromosome en deux (voir Figure 1.1b). Ces deux types de réarrangements modifient le nombre de chromosomes du génome. Pour être viables, ils doivent être associés à la perte ou l'inactivation d'un centromère dans le cas de la fusion, et à la création d'un centromère et de télomères dans le cas de la fission.

Un autre réarrangement impliquant un chromosome entier, mais sans modifier le nombre de chromosomes, est la formation d'un chromosome en anneau. Un chromosome en anneau est un chromosome dont les deux extrémités sont reliées. Ce type de réarrangement est observé dans des cas de maladie, mais il ne l'est jamais dans le caryotype normal d'une espèce.

Parmi les réarrangements équilibrés qui affectent des segments chromosomiques, on distingue principalement deux types : l'**inversion** et la **translocation**. Le premier est un réarrangement intra-chromosomique, c'est-à-dire qu'il affecte un seul chromosome. Comme son nom l'indique, l'inversion consiste à inverser un segment sur un chromosome (voir Figure 1.2a). Lorsque le segment inversé contient le centromère du chromosome, on parle d'inversion péricentrique, sinon d'inversion paracentrique. La translocation, ou translocation réciproque, est un réarrangement inter-chromosomique, impliquant deux chromosomes non homologues. C'est l'échange de deux extrémités entre ces chromosomes (voir Figure 1.2b). La translocation



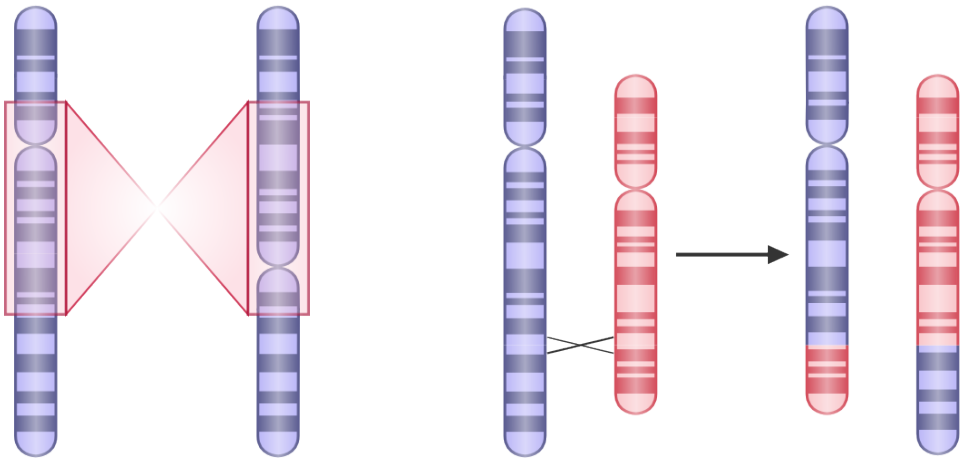
(a) Fusion de deux chromosomes.

(b) Fission d'un chromosome.

FIG. 1.1: Deux types de réarrangements équilibrés impliquant des chromosomes entiers.

robertsonienne est un cas particulier de translocation entre deux chromosomes acrocentriques au niveau de leur centromère. Il en résulte la fusion des deux grands bras des chromosomes acrocentriques et la perte des petits bras. Ce réarrangement s'apparente ainsi à la fusion de chromosomes, diminuant par un le nombre de chromosomes.

On peut trouver dans la littérature un troisième type de réarrangement de segment chromosomique appelé **transposition**. Il désigne le déplacement d'un segment chromosomique à une autre position du génome. Il peut être intra- ou inter-chromosomique en fonction de la position d'insertion. Même si on emploie le même terme lorsqu'il s'agit du déplacement d'éléments transposables, il ne faut pas les confondre, puisque les mécanismes moléculaires en jeu sont différents.



(a) Inversion péricentrique.

(b) Translocation.

FIG. 1.2: Deux types de réarrangements équilibrés.

b. Réarrangements déséquilibrés

Les réarrangements déséquilibrés modifient la quantité d'ADN. Il existe deux types de réarrangements déséquilibrés. La **délétion** diminue la quantité d'ADN, c'est la perte d'un segment chromosomique. Le réarrangement qui augmente la quantité d'ADN est la **duplication**. Un segment d'ADN se retrouve en deux copies dans le génome. Les duplications sont souvent en tandem, c'est-à-dire que les deux copies sont situées côte à côte sur le chromosome, dans la même orientation ou bien inversées. Lorsque les deux copies sont dispersées sur le chromosome, voire sur des chromosomes différents, on parle de **transposition duplicative**. Cependant, il n'est pas très clair dans quelle mesure les deux évènements (transposition et duplication) sont indépendants.

1.2 Mécanismes moléculaires des réarrangements

Dès les années 1930, bien avant qu'on ne connaisse la structure des chromosomes, plusieurs théories ont été proposées pour expliquer les aberrations chromosomiques visibles après des irradiations de cellules somatiques (lire la revue (Savage, 1998)). La théorie classique, appelée "cassure-et-réunion" ("breakage-and-reunion") et proposée en 1938, est encore largement acceptée à l'heure actuelle, même si les concepts de structure et de cassure des chromosomes sur lesquels la théorie était basée sont radicalement différents aujourd'hui.

Il est actuellement admis que les réarrangements se produisent à la suite d'une ou de plusieurs cassures double brin de l'ADN et que les mécanismes de réparation de ces lésions sont impliqués dans la formation des réarrangements (Pfeiffer *et al.*, 2000).

1.2.1 Les cassures double brin de l'ADN

Une cassure double brin est une cassure qui coupe les deux brins de la molécule d'ADN, à la même position, ou bien à des positions très proches. Ces lésions ne sont pas rares et peuvent être induites par de nombreux facteurs. D'une part, des agents exogènes, extérieurs à la cellule, comme des molécules réactives (ions à oxygènes, radicaux libres, péroxides) ou des radiations comme les rayons gamma et les rayons X, peuvent provoquer de telles lésions. Mais également, des éléments de la machinerie cellulaire peuvent induire des cassures double brin. Par exemple, certaines enzymes, les endonucléases, sont capables de couper l'ADN sur les deux brins. Des cassures double brin résultent souvent d'accidents durant la réplication. Si la fourche réplicative rencontre des obstacles, comme des structures secondaires formées par l'ADN simple brin, ou bien une cassure simple brin, cela peut alors engendrer une cassure double brin (Aguilera et Gómez-González, 2008).

Dans certains cas, les cassures double brin sont générées délibérément. Par exemple, une enzyme, la topoisomérase de type II, génère spontanément des cassures double brin pour modifier les tours d'hélice de la molécule d'ADN. Les cassures double brin font également partie de mécanismes moléculaires très complexes, comme la recombinaison V(D)J dans les lymphocytes ou les crossing-overs durant la méïose.

La recombinaison V(D)J est un processus qui participe à la protection immunitaire. Elle génère de la variabilité au sein des lymphocytes, les cellules chargées de reconnaître les agents pathogènes et étrangers au soi. En partant d'un ensemble de segments d'ADN, la recombinaison V(D)J génère un grand nombre de combinaisons entre eux, grâce à des cassures double brin à des sites spécifiques.

Lors de la formation des gamètes, durant la méïose, des cassures double brin sont également

requises pour effectuer la recombinaison génétique. Les crossing overs permettent d'échanger des fragments d'ADN à un même locus entre chromosomes homologues. Cela permet de mélanger sur un même chromosome des allèles d'origine maternelle avec des allèles d'origine paternelle. Ce mécanisme joue un rôle important pour créer et maintenir la diversité génétique.

1.2.2 Mécanismes de réparation des cassures double brin

Dans les cas précédents, les cassures double brin sont utiles et participent à des processus biologiques, mais elles sont toujours des dangers importants pour la cellule si elles ne sont pas réparées. L'intégrité du génome est en effet mise en danger. En conséquence, une seule cassure double brin est suffisante pour provoquer l'arrêt du cycle cellulaire, voire la mort de la cellule. Contrairement aux cassures simple brin, qui peuvent être facilement réparées en utilisant le brin sain, la réparation d'une cassure double brin nécessite des mécanismes moléculaires plus complexes. Parmi ces derniers, deux mécanismes principaux et très différents sont décrits dans la littérature : le Non-Homologous End Joining (NHEJ) et la recombinaison homologue (HR). Le premier est dit *biochimique* alors que le deuxième est dit *génétique*.

On dit que NHEJ est biochimique car il n'utilise pas (et ne dépend pas de) l'information génétique pour réparer la cassure. Il consiste simplement à "recoller" les extrémités séparées par la cassure. Ce mécanisme est très rapide, mais il en coûte une perte d'information : les extrémités de la cassure double brin sont souvent digérées avant d'être recollées. Ce mécanisme est aussi appelé recombinaison non homologue ou illégitime (Abeysinghe *et al.*, 2003; Pfeiffer *et al.*, 2000). Notons également que s'il n'a pas besoin d'homologie de séquence, il peut utiliser parfois des micro-homologies de quelques paires de bases (1 à 10 pb).

Par contre, le second mécanisme, HR, est plus conservatif et plus lent. Il restaure l'information endommagée en utilisant une information similaire, qui provient généralement du chromosome homologue à celui endommagé. Ce mécanisme de réparation est basé sur la recombinaison homologue et a été proposé par Szostak *et al.* (1983) : l'une des extrémités 3' de la cassure envahit la molécule double brin homologue, chaque extrémité 3' est "complétée" en utilisant le brin homologue, formant des jonctions de Holliday (c'est une structure réunissant deux molécules d'ADN double brin, dans laquelle chaque brin d'une molécule est hybridé avec celui de l'autre molécule double brin), les jonctions de Holliday sont résolues en deux molécules d'ADN sans cassure. La résolution peut se faire de deux manières : la conversion génique ou le crossing-over. Dans le premier cas, il y a très peu d'échange d'information génétique entre les deux molécules, seulement la région entre les jonctions de Holliday (hétéroduplex). Dans le cas du crossing-over, l'échange de matériel génétique entre les deux molécules est étendu jusqu'aux extrémités des molécules (voir Figure 1.3). Ce mécanisme nécessite que la séquence réparée et la séquence qui sert de matrice soient suffisamment similaires et ce sur plusieurs centaines de paires de base.

Ces deux mécanismes sont présents dans de nombreux organismes, mais en fonction des espèces l'un ou l'autre mécanisme est privilégié. Chez l'homme, c'est le NHEJ qui est prépondérant, notamment dans les phases G0, G1 et le début de la phase S du cycle cellulaire (Lieber *et al.*, 2003), HR serait utilisée après la réplication (Pfeiffer *et al.*, 2000). C'est le contraire chez la levure où c'est HR qui est le plus souvent utilisé. Une hypothèse pouvant expliquer cette différence est que le génome humain possède trop d'éléments répétés (plus de risques de réarrangements, voir section suivante) (Lieber *et al.*, 2003). Une seconde hypothèse est que chez les génomes à forte densité en gènes, le mécanisme NHEJ causerait trop de dégâts, car il est fortement mutagène (Pfeiffer *et al.*, 2000).

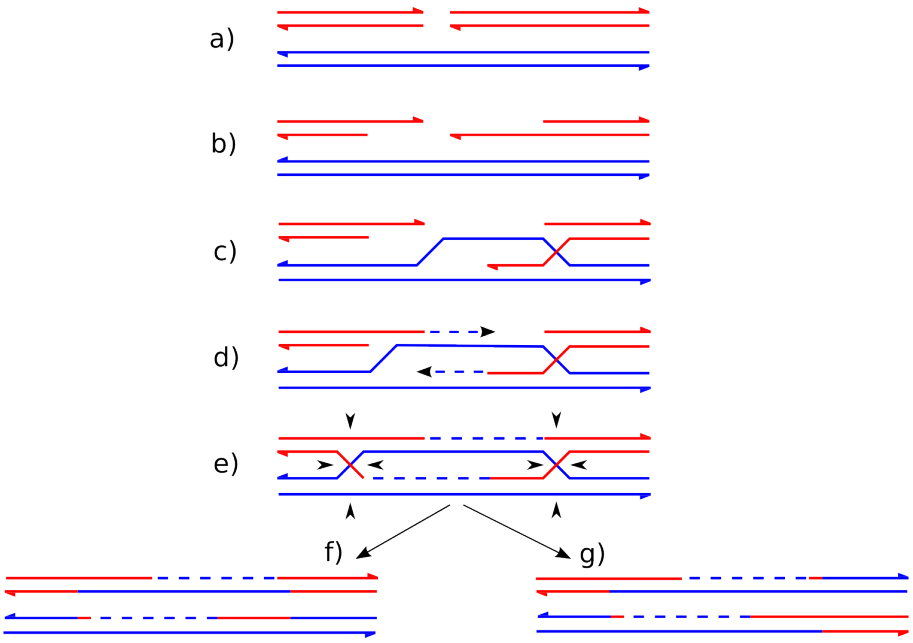


FIG. 1.3: La recombinaison homologue. Elle est initiée par une cassure double brin (a), dont les extrémités 5' sont ensuite digérées pour laisser libres les extrémités 3' (b). Une extrémité 3' migre et envahit une autre molécule d'ADN à un site homologue (c). Les extrémités 3' sont ensuite étendues par polymérisation (d), puis ligaturées avec les extrémités 5' pour former deux jonctions de Holliday (e). Celles-ci sont résolues par des cassures (flèches noires horizontales ou verticales). En fonction de celles-ci, on obtient deux produits différents : deux résolutions de même type (soit verticales, soit horizontales) produisent seulement des hétéroduplexes résolus par conversion génique (f), alors que deux résolutions de types différents produisent un échange réciproque (ou crossing over) en plus des hétéroduplexes (g).

Néanmoins, chez l'homme les deux mécanismes sont utilisés et sont contrôlés par deux voies de signalisation distinctes (revues dans (O'driscoll et Jeggo, 2006)). Le choix du mécanisme n'est pas aléatoire, il dépend de l'origine de la cassure et de la phase du cycle cellulaire dans laquelle se trouve la cellule.

On peut noter un troisième mécanisme de réparation des cassures double brin qui intervient dans les cellules eucaryotes : SSA (Single Strand Annealing). Ce mécanisme est classé avec HR parmi les mécanismes dépendant d'homologie, mais il est moins conservatif que HR (Pfeiffer *et al.*, 2000). Il intervient dans le cas particulier où la cassure se trouve entre deux répétitions directes (au moins 30 pb). Les deux extrémités 3' de la cassure s'hybrident au niveau de la répétition, et les gaps simple brin restants sont complétés. Le résultat de cette réparation est la délétion d'une des deux copies de la répétition et de la séquence comprise entre les deux.

1.2.3 Les réarrangements, des erreurs de la réparation

Les réarrangements sont, le plus souvent, causés par des erreurs des mécanismes de réparation. Des expériences chez la levure montrent que des mutations dans certaines protéines de la réparation impliquent une forte augmentation de la fréquence des réarrangements. Certaines maladies où on trouve de nombreux réarrangements sont également dues à des mutations

affectant des protéines impliquées dans ces systèmes de réparation (Pfeiffer *et al.*, 2000).

Même sans endommager les acteurs de ces mécanismes, la réparation peut commettre des erreurs. Ainsi, si plusieurs cassures double brin se produisent simultanément sur le génome et si elles sont proches dans le noyau, le système NHEJ peut se tromper et recoller les mauvais morceaux. Cela peut entraîner des inversions si les deux cassures sont sur le même chromosome, des translocations si elles sont sur des chromosomes différents et même des transpositions si plus de deux cassures sont en jeu.

Dans le cas de la réparation par recombinaison homologue, les erreurs sont possibles si une mauvaise matrice de réparation est utilisée. En effet, seule la similarité de séquence est nécessaire pour initier la recombinaison et il est possible que la matrice utilisée soit sur un locus différent. Si elle n'est pas au même locus sur le génome, l'échange d'ADN qui a lieu au cours de la recombinaison va entraîner un réarrangement (si les jonctions de Holliday sont résolues par un crossing over). Par exemple, si la recombinaison a lieu entre deux chromosomes non homologues, cela peut entraîner une translocation réciproque. Si la matrice est sur le même chromosome que la cassure mais dans l'orientation inverse, cela peut conduire à une inversion (voir des exemples dans la Figure 1.4).

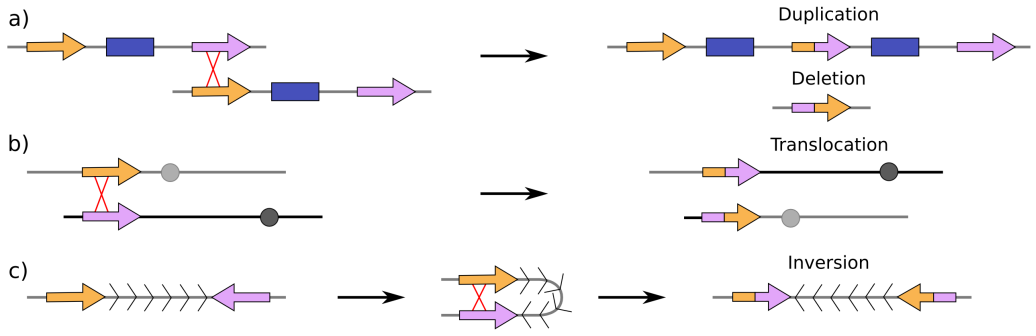


FIG. 1.4: Exemples de réarrangements obtenus par recombinaison homologue entre séquences à des loci différents (figure reproduite à partir de l'article de revue de Bailey et Eichler (2006) sur le rôle des duplications segmentaires dans l'évolution des génomes de primates). Les flèches violettes et saumons représentent des séquences dupliquées (séquences très similaires mais localisées à des loci différents). En a) il s'agit d'une duplication en tandem, le résultat de la recombinaison entre les deux copies est, soit la duplication de la séquence située entre les deux, soit sa délétion. En b), les deux copies sont sur des chromosomes différents, le résultat est une translocation entre ces deux chromosomes. Et en c), les deux copies sont sur un même chromosome mais sur des brins différents, ce qui entraîne l'inversion de la séquence située entre les deux.

Ce processus est appelé recombinaison homologue non allélique (Non Allelic Homologous Homologous Recombination (NAHR)) ou recombinaison ectopique. Il a été mis en cause dans de nombreuses maladies génétiques (appelées désordres génomiques), en particulier lorsqu'il conduit à des réarrangements non équilibrés (avec perte ou gain de matériel génétique) (lire les articles de revue (Ji *et al.*, 2000; Stankiewicz et Lupski, 2002)). En ce qui concerne SSA, des réarrangements peuvent se produire si plusieurs cassures double brin proches d'une séquence répétée sont réparées par SSA.

On note deux différences importantes entre ces mécanismes responsables de réarrange-

ments. La première porte sur le nombre de cassures double brin nécessaires. Une seule cassure double brin suffit pour générer un réarrangement par recombinaison ectopique, alors qu'au moins deux cassures sont nécessaires dans le cas de NHEJ et de SSA. La deuxième différence est l'implication d'homologie de séquence. Si NHEJ n'a pas besoin, en théorie, d'homologie de séquence pour joindre les extrémités de cassures double brin, des micro-homologies de quelques paires de bases sont souvent utilisées. Par contre, l'étendue d'homologie de séquence nécessaire pour SSA et surtout pour NAHR est plus importante : 30 pb semblent suffisantes pour initier la réparation par SSA, alors que NAHR nécessiterait au moins plusieurs centaines de paires de bases (Pfeiffer *et al.*, 2000). C'est grâce à ces longues séquences similaires aux points de cassure des réarrangements que l'on peut identifier le mécanisme NAHR comme responsable du réarrangement. Par exemple, on trouvera aux bornes d'un segment inversé par NAHR une même séquence répétée et inversée. En ce qui concerne NHEJ, il laisse moins de traces dans la séquence et est plus difficilement identifiable comme mécanisme responsable de réarrangements. Des petites délétions et insertions proches des points de cassure sont parfois des indices laissés par ce mécanisme.

1.2.4 Réarrangements et éléments transposables

Les transpositions d'éléments transposables ne sont pas considérées commes des réarrangements chromosomiques. Cependant, ces derniers peuvent jouer un rôle important dans la formation de réarrangements chromosomiques.

D'une part, du fait de leur présence en un grand nombre de copies dans les génomes, ils sont des candidats idéaux pour la recombinaison ectopique. Ainsi, si une cassure double brin se trouve dans (ou à proximité d') un élément transposable, la recombinaison homologue peut se produire entre des copies à des loci différents et générer un réarrangement.

D'autre part, les mécanismes de transposition des éléments transposables peuvent générer des réarrangements. Par exemple, l'insertion des éléments LINE peut entraîner des délétions de quelques paires de bases jusqu'à plusieurs dizaines de kilobases (Han *et al.*, 2005). Les transposons à ADN sont également mis en cause (Gray, 2000). Le mécanisme de transposition de ces éléments est de type "couper-et-coller", les deux extrémités de l'élément s'apparient et l'élément est excisé puis inseré à un autre locus. Des réarrangements peuvent se produire lorsque l'appariement se fait entre les extrémités de deux éléments différents. Divers types de réarrangements sont possibles, les plus observés sont les transpositions de séquences entre deux transposons successifs. Ces mécanismes ont été mis en évidence essentiellement chez les procaryotes, les plantes et les drosophiles (Gray, 2000).

1.2.5 Liens avec l'organisation spatiale des chromosomes dans le noyau

On peut se demander dans quelle mesure l'organisation du génome dans le noyau influence la formation de réarrangements. Les chromosomes ne sont pas positionnés au hasard les uns par rapport aux autres dans le noyau. Plusieurs modèles d'organisation nucléaire existent, allant de territoires chromosomiques excluant toute intéraction entre différents chromosomes, jusqu'à des niveaux de chevauchements des territoires (intermingling) de plus en plus importants (lire la revue (Branco et Pombo, 2006a)). Dans le modèle actuellement accepté de chevauchement des territoires, ces chevauchements ne sont pas aléatoires et des associations entre certains chromosomes semblent privilégiées (Branco et Pombo, 2006b). Ainsi, si la proximité spatiale des chromosomes dans le noyau joue un rôle dans la formation des réarrangements, on devrait observer certains réarrangements plus fréquemment que d'autres (les réarrangements associant des chromosomes ou des régions chromosomiques proches).

La question du rôle de la proximité spatiale des chromosomes dans le noyau sur les réarrangements n'est pas triviale et n'est pas tranchée à l'heure actuelle. Deux modèles existent. Le premier, "contact-first" suppose que les cassures double brin ne sont pas mobiles et il est donc nécessaire que les deux partenaires du réarrangement soient en contact ou proches lors de la (ou des) cassure(s). Le deuxième, "breakage-first" propose au contraire que les cassures sont mobiles et que le contact avec l'autre partenaire du réarrangement est postérieur à la cassure. Plusieurs résultats ont été obtenus dans ce domaine qui favorisent l'un ou l'autre modèle. En faveur du breakage-first, une mobilité importante des cassures a été observée, avec notamment des regroupements de cassures en clusters concentrant les protéines de réparation (Aten *et al.*, 2004). Cependant, l'analyse de la distribution des partenaires de réparation (par NHEJ ou HR) montre une préférence pour des loci situés sur le même bras chromosomique et à moins de 100 Kb, mais également certains loci inter-chromosomiques seraient privilégiés alors que d'autres évités (D'Anjou *et al.*, 2004). Enfin, Branco et al. ont observé une corrélation entre les patrons de chevauchement de territoires chromosomiques et les fréquences de translocation entre ces chromosomes (Branco et Pombo, 2006b). Ces deux derniers résultats sont plutôt en faveur du modèle contact-first et d'une influence de l'organisation spatiale des chromosomes dans le noyau sur la formation des réarrangements.

1.3 Impacts des réarrangements

Par rapport aux mutations ponctuelles, les modifications de structure du génome ont des impacts fonctionnels et évolutifs particuliers. Tout d'abord d'un point de vue mécanistique, il peuvent perturber des processus cellulaires qui dépendent de la structure des chromosomes. Ensuite, ils modifient l'organisation de l'information génétique et peuvent avoir un impact sur son expression dans la cellule et donc sur son phénotype. Enfin, au niveau populationnel, leur transmission et leur propagation peuvent parfois être complexes et engendrer des phénomènes d'isolement et de spéciation.

1.3.1 Le passage difficile de la méiose

Le premier impact des réarrangements est qu'ils peuvent perturber des processus cellulaires qui dépendent de la structure des chromosomes, comme la méiose. Nous nous intéressons plus particulièrement à la méiose car c'est une étape clé dans la gamétogénèse et donc dans la transmission des réarrangements à la descendance. L'impact peut être considérable pour la cellule, puisque certains réarrangements conduisent à l'arrêt de la méiose. Ainsi, tous les réarrangements ne sont pas "autorisés" et ne seront jamais transmis. Bien sûr, les réarrangements aboutissant à des chromosomes sans centromère ou télomère font partie des réarrangements non transmis. Mais la structure des chromosomes individuellement n'est pas le seul facteur important, il ne faut pas oublier que les chromosomes sont organisés par paires de chromosomes homologues dans les cellules diploïdes.

La méiose est l'évènement majeur de la gamétogénèse. Elle se déroule en deux étapes clés : la recombinaison entre les chromosomes homologues puis la réduction du nombre de chromosomes pour passer d'une cellule diploïde (2n) à une cellule haploïde (n). Ces deux étapes sont liées puisque la ségrégation des chromosomes dans chaque cellule fille n'est rendue possible que si la première étape de recombinaison s'est bien déroulée entre chaque paire de chromosomes homologues.

Un gros remaniement chromosomique à l'état hétérozygote peut perturber l'une et l'autre étape. En effet, la ségrégation des chromosomes dans les cellules filles nécessite que les chromo-

somes homologues soient appariés deux à deux. A cause de réarrangements chromosomiques, les appariements deux à deux sont plus complexes. Par exemple, une translocation sur un chromosome aboutit à la formation d'une structure complexe qui associe quatre chromosomes (les deux chromosomes ayant subi la translocation et leurs homologues), appelé un quadrivalent, lors de la recombinaison. Cela peut empêcher la bonne ségrégation des chromosomes et provoquer l'arrêt de la méiose. Des remaniements intra-chromosomiques peuvent également rendre difficile l'appariemment de chromosomes homologues (hétérozygotes) et empêcher la recombinaison homologue de se dérouler correctement, provoquant l'arrêt de la méiose.

1.3.2 Impacts fonctionnels

Les réarrangements chromosomiques, en modifiant l'organisation de l'information génétique, ont des effets sur son expression dans la cellule et donc sur le phénotype de cette dernière. On leur a souvent attribué des effets négatifs, notamment parce qu'ils ont beaucoup été étudiés lorsqu'ils sont associés à une pathologie. De nombreuses maladies génétiques sont causées par des réarrangements chromosomiques ; on les appelle des désordres génomiques. Cependant l'importance du polymorphisme de réarrangement suggère qu'un certain nombre de réarrangements sont neutres et ont très peu d'effets phénotypiques (Tuzun *et al.*, 2005; Feuk *et al.*, 2006; Korbel *et al.*, 2007).

De nombreuses associations entre réarrangement et phénotype ont été observées, dans le cas des désordres génomiques mais également dans des cas d'adaptations locales. Par exemple, dans des populations de drosophiles et de moustiques, on observe des corrélations entre des inversions polymorphes et la température, les changements climatiques, la latitude ou le type d'habitat. Un exemple frappant est celui du criquet d'Australie qui montre des changements de taille et de forme importants en fonction de la présence ou absence de deux inversions polymorphes (Coghlan *et al.*, 2005).

a. Ajout et perte de matériel génétique

L'impact le plus important est bien sûr la perte ou l'ajout de matériel génétique. C'est le cas des réarrangements déséquilibrés. A grande échelle, l'ajout de chromosomes entiers (polysomie) est très rarement viable. Chez l'homme, seules les trisomies des plus petits chromosomes sont observées (par exemple, la trisomie 21), mais elles sont associées à des pathologies. A plus petite échelle, si le matériel en cause (perdu ou dupliqué) contient un ou plusieurs gènes, cela peut avoir des conséquences sur les réseaux d'interaction des gènes et sur de nombreuses fonctions cellulaires. Ainsi, la plupart des désordres génomiques sont dus à des délétions ou duplications de gènes.

Récemment, le polymorphisme de délétion et de duplication (ou polymorphisme de nombre de copies) a été extrêmement étudié chez l'homme, notamment grâce à de nouvelles technologies (voir Section 2.1.3) (par exemple, lire (Feuk *et al.*, 2006; Freeman *et al.*, 2006; Redon *et al.*, 2006; Kidd *et al.*, 2008)). Ces études ont révélé un très grand nombre de variants du nombre de copies au sein des populations humaines, dont beaucoup impliquent des gènes. Les phénotypes des individus testés n'étaient *a priori* associés à aucune maladie, indiquant que les effets de ces variants sont, soit très peu délétères, soit neutres, voire positifs.

Enfin, la duplication de gène peut avoir des effets positifs à plus long terme, puisque c'est un des mécanismes permettant de créer de nouvelles fonctions (subfonctionalisation ou néofonctionalisation).

b. Modification de l'information génétique

Les réarrangements chromosomiques peuvent également endommager l'information génétique portée par les éléments fonctionnels tels que les gènes. Si les cassures double brin à l'origine des réarrangements se produisent dans des zones fonctionnelles, tels que les gènes ou les régions régulatrices, ces dernières peuvent être endommagées. On peut ainsi perdre un gène ou modifier son niveau d'expression. De nouveaux gènes peuvent également être créés par fusion de deux gènes au niveau des cassures double brin. Les cas de fusion de gènes par translocation ou inversion sont très courants dans les cellules cancéreuses (Bashir *et al.*, 2008).

c. Effets de position

Enfin, les éléments fonctionnels du génome ne sont pas placés aléatoirement dans le génome et leur contexte génomique peut influencer leur activité et leur fonction. Dans les génomes bactériens, les gènes sont organisés en opérons et sont co-transcrits et co-régulés. On comprend dans ce cas que la position d'un gène a une grande influence sur son activité. Dans les génomes eucaryotes et plus précisément des vertébrés, ces effets de position sur l'expression des gènes sont moins clairs. Plusieurs cas d'effets de position ont été mis en évidence dans le cadre de l'étude de maladies génétiques chez l'homme (lire la revue (Kleinjan et van Heyningen, 1998)) : le point de cassure du réarrangement a été identifié proche du gène responsable de la maladie mais n'a pas altéré la séquence ni du gène, ni de sa région promotrice. La modification de son expression s'explique souvent alors par son changement d'environnement, soit par la perte, soit par l'ajout, d'un élément régulateur distant.

Même si cela n'est pas systématique comme dans les génomes procaryotes, on a pu identifier, dans les génomes eucaryotes, certains groupes de gènes voisins qui sont co-exprimés ou co-régulés, comme par exemple le complexe CMH ou la région des β-globines. Ces exemples sont très anecdotiques et représentent souvent des groupes de gènes paralogues (lire la revue (Hurst *et al.*, 2004)). Sémon et Duret (2006) ont montré que seulement 3 à 5 % des gènes de l'homme seraient maintenus en clusters de gènes co-exprimés par la sélection naturelle et ne seraient donc pas dus au hasard.

Cependant, les génomes eucaryotes sont également organisés à plus grande échelle, par exemple, en domaines de chromatine, en domaines de réplication, ou en isochores. On suppose que le changement d'environnement pour un gène par rapport à ces structures aura une influence sur son activité. Par exemple, le déplacement d'un gène d'une région de chromatine ouverte à une région hétérochromatique aura pour effet d'éteindre le gène du point de vue de son expression (Sproul *et al.*, 2005; Hurst *et al.*, 2004). A une échelle encore plus grande, l'organisation 3D des chromosomes dans le noyau ne semble pas non plus aléatoire et paraît liée à la transcription. Il existerait des interactions physiques entre gènes distants (par exemple à travers les usines de transcription) (Martin et Pombo, 2003; Verschure, 2006). Les réarrangements chromosomiques pourraient avoir des effets sur ces interactions.

1.3.3 La spéciation chromosomique

On observe que la plupart des espèces de plantes et d'animaux ont des caryotypes différents. Parfois, même entre espèces très proches, le nombre et la forme des chromosomes sont différents. De plus, ces caractéristiques sont importantes pour le bon déroulement de la méiose et de la gamétogénèse, et donc pour la reproduction. Ces observations ont conduit à impliquer les réarrangements chromosomiques dans les processus de spéciation.

Si un réarrangement se produit dans la lignée germinale d'un individu, même s'il a des effets bénéfiques pour la cellule et l'individu, il doit encore passer de nombreuses étapes difficiles avant d'être fixé dans la population. En effet, contrairement aux mutations ponctuelles, la transmission parentale des gros réarrangements de génération en génération est souvent problématique. Ces réarrangements conduisent très souvent à une hypofertilité de l'individu qui le porte à l'état hétérozygote.

a. Hypofertilité des hétérozygotes

Chez un individu porteur d'un réarrangement à l'état hétérozygote, même si la méiose n'est pas stoppée, elle peut conduire dans de nombreux cas à des gamètes déséquilibrés. Dans le cas d'un réarrangement inter-chromosomique, la répartition des chromosomes dans les deux cellules filles peut conduire à deux cellules n'ayant pas le même contenu en ADN. Par exemple, dans le cas d'une translocation réciproque, les gamètes seront équilibrés seulement si les deux chromosomes non homologues ayant été transloqués se retrouvent dans la même cellule fille. Sinon les parties échangées seront, soit présentes en deux copies, soit délétées dans les cellules filles. Il en est de même pour les fusions et fissions de chromosomes, mais dans ces cas ce sont des chromosomes entiers qui sont perdus ou ajoutés. Par exemple, dans 5 % des cas de trisomie 21 chez l'homme, un des parents est porteur d'une translocation robertsonienne avec le chromosome 21.

Les réarrangements intra-chromosomiques peuvent également conduire à des duplications et délétions par l'intermédiaire de la recombinaison homologue. Si la recombinaison s'effectue au niveau d'un segment inversé, les segments chromosomiques échangés par le crossing over ne seront pas homologues et il en résulte des chromosomes déséquilibrés avec le segment échangé dupliqué ou délété (voir Figure 1.5).

Les déséquilibres quantitatifs (duplication ou délétion d'une partie ou d'un chromosome entier) sont souvent associés à des pathologies ou ne sont pas viables. Ainsi cela réduit fortement la fertilité des individus porteurs de remaniements. De plus, même si les remaniements sont transmis avec succès à la descendance, les descendants encourent ensuite les mêmes risques de fertilité de génération en génération, jusqu'à ce qu'on atteigne un état homozygote pour le réarrangement.

Plusieurs modèles de spéciation chromosomique (ou spéciation stasipatrique) sont basés sur cette hypothèse de barrière gamétique provoquée par les réarrangements chromosomiques (on retiendra notamment celui de White (1978)). Si un réarrangement est installé dans une sous-population, les individus de cette sous-population se reproduiront plus facilement entre eux qu'avec le reste de la population. Ainsi, il y a une réduction du flux de gènes entre la population originale et la sous-population possédant le réarrangement. Ces deux populations vont acquérir de nouvelles mutations indépendamment, les rendant de moins en moins interfertiles, jusqu'à l'isolement reproductif total et la spéciation.

b. Autres modèles de spéciation

Cependant, ces modèles sont basés sur un paradoxe : si les réarrangements chromosomiques réduisent beaucoup la fitness des hétérozygotes, comment de tels réarrangements peuvent-ils se fixer dans une sous-population ? De plus, on s'est aperçu qu'un grand nombre de réarrangements n'avaient que très peu d'effets sur la fertilité des individus hérétozygotes, et qu'il existe au sein des espèces un polymorphisme important de réarrangements. Les modèles de spécia-

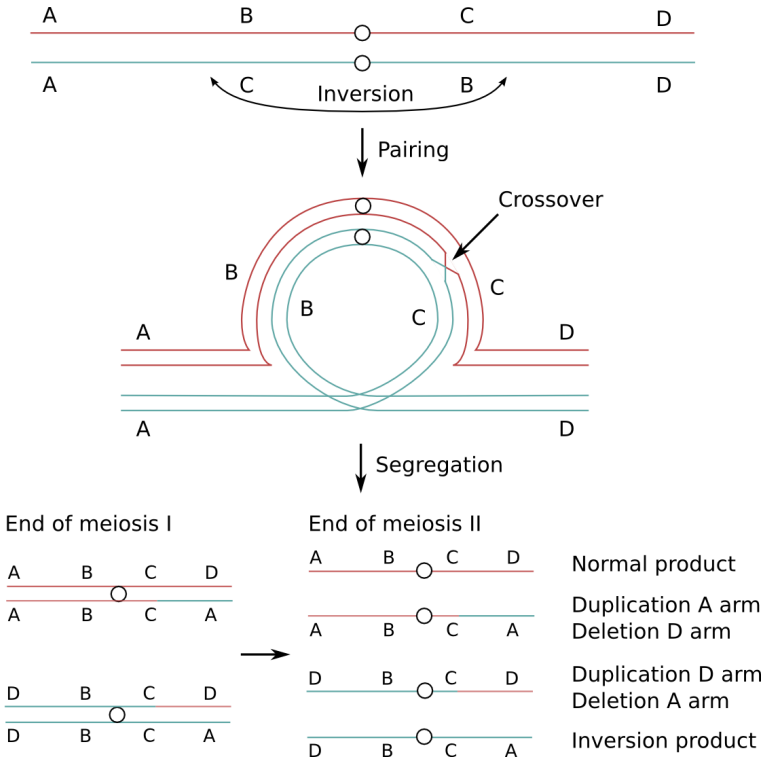


FIG. 1.5: Exemple de la formation de gamètes déséquilibrés par recombinaison dans un segment inversé (figure reproduite à partir de Rieseberg (2001)).

tion chromosomique se sont alors complexifiés, faisant intervenir de l'isolement géographique, de la dérive, la valeur adaptative du réarrangement, etc. (lire la revue (Rieseberg, 2001)).

On peut citer, par exemple, le modèle en cascade ou en chaîne qui propose que les réarrangements ne provoquent pas une barrière gamétique individuellement, mais c'est la combinaison de plusieurs réarrangements différents qui engendre la barrière gamétique (Dutrillaux, 1997; Britton-Davidian *et al.*, 2000; Piálek *et al.*, 2001).

Un autre modèle suggère que les réarrangements réduisent le flux de gènes entre populations, non pas par leur effet sur la fitness mais par leur effet sur la recombinaison (Noor *et al.*, 2001; Rieseberg, 2001; Navarro et Barton, 2003). Les réarrangements peuvent réduire la recombinaison localement, soit par sélection contre les gamètes déséquilibrés, ou bien directement sans réduction de la fertilité. C'est le cas notamment des inversions pour lesquelles on observe très peu de recombinaison dans le segment inversé à l'état hétérozygote. Si le segment inversé contient un ou plusieurs gènes d'isolement (allèle incompatible entre les deux populations), alors la spéciation est facilitée par l'inversion. La réduction du flux de gènes dans la zone inversée permet l'accumulation d'allèles incompatibles entre les deux populations, jusqu'à la spéciation. Plusieurs observations soutiennent ce modèle. Notamment, chez les espèces de drosophiles, les gènes responsables de l'isolement (par exemple, responsables de la stérilité des hybrides ou controlant les préférences des femelles) de plusieurs espèces proches se trouvent très souvent groupés dans des inversions différenciant ces espèces (Noor *et al.*, 2001; Machado *et al.*, 2002). De plus, on trouve plus d'inversions dans les espèces qui occupent des régions géographiques communes (en sympatrie) que dans les espèces séparées géographiquement (en parapatrie) (Noor *et al.*, 2001).

1.4 Les réarrangements et l'évolution des génomes

On appelle **réarrangements évolutifs**, les réarrangements qui différencient les génomes d'espèces différentes. On les distingue ainsi des réarrangements que l'on peut trouver dans des cellules somatiques, comme par exemple dans les cellules cancéreuses. En effet, les réarrangements évolutifs se sont produits dans des cellules de la lignée germinale et ont été transmis à la descendance. On les distingue également des réarrangements polymorphes car ces derniers ne sont pas (encore ?) fixés dans la population. Par contre, les réarrangements évolutifs ont été fixés dans la population, soit par dérive, soit par la sélection naturelle. Ainsi, on ne peut vraisemblablement pas observer de réarrangements évolutifs qui soient fortement défavorables pour l'individu et l'espèce. Cela les distingue, par exemple, des réarrangements responsables de désordres génomiques.

Les réarrangements évolutifs sont très nombreux. Depuis la découverte en 1921 d'une inversion sur un chromosome de drosophile (Sturtevant, 1921), le nombre de réarrangements identifiés dans les génomes n'a cessé de croître. En effet, avec l'amélioration des techniques d'identification, on a pu détecter des réarrangements à des échelles de plus en fines (voir Chapitre 2). On a ainsi mis en évidence un très grand nombre de réarrangements chromosomiques différenciant les génomes d'espèces différentes.

A grande échelle, on observe une grande diversité de caryotypes entre espèces. Au sein des mammifères, par exemple, le nombre de chromosomes varie de 2n=6 pour le muntjac indien à 2n=102 pour le rat-viscache roux d'Argentine. Même entre espèces proches, les différences peuvent être importantes. L'exemple le plus frappant est celui du muntjack : le muntjack indien et le muntjack de Chine sont très semblables au niveau phénotypique, cependant ils possèdent respectivement 2n=6 et 2n=42 chromosomes (Ferguson-Smith et Trifonov, 2007).

Lorsque la résolution augmente, le nombre de réarrangements évolutifs identifiés devient très important. Alors que dans les années 80, on pensait que l'homme et le chimpanzé ne différaient que par 9 inversions et une fusion (Yunis *et al.*, 1980), Newman *et al.* (2005) ont détecté aujourd'hui 93 inversions supplémentaires entre 12 kb et 1 Mb. Ces réarrangements ont une plus grande couverture du génome que l'ensemble des mutations ponctuelles réunies (divergence de séquence de 1.2 % (Chimpanzee Sequencing and Analysis Consortium, 2005)).

Ainsi, les réarrangements chromosomiques jouent un rôle important dans l'évolution des génomes. Cependant, si les mécanismes moléculaires responsables de leur formation sont bien compris à l'heure actuelle, il reste de nombreuses inconnues quant à ce qui détermine leur fréquence et leurs localisations dans les génomes. On ne sait expliquer, par exemple, les différences en fréquence et en type de réarrangements observées dans les différents groupes du vivant.

Certains réarrangements semblent spécifiques à certaines lignées évolutives. Par exemple, les primates évoluent essentiellement par des inversions, chez les cercopithèques ce sont les fissions de chromosomes qui dominent, chez les lémuriens les translocations robertsoniennes (Dutrillaux, 1997). Les taux de réarrangements sont également très différents. Les poissons et les oiseaux ont des taux de réarrangements nettement plus faibles que les mammifères, eux mêmes plus faibles que chez les insectes (Coghlan *et al.*, 2005). Même au sein des mammifères, les taux de réarrangements peuvent varier de plus d'un ordre de grandeur sur une période de 500 millions d'années (Burt *et al.*, 1999). Burt et collègues ont identifié, chez les mammifères, trois périodes depuis leur divergence avec le poulet, ayant des taux de réarrangements très différents. En effet, le taux varie de moins de 0.2 réarrangements par million d'années au

moment de la divergence avec le poulet, jusqu'à 2.3 actuellement chez certaines espèces de singes.

Quant à la distribution des réarrangements le long des génomes, elle est considérée aléatoire jusque dans les années 2000. On privilégie alors le modèle de cassures aléatoires, qui sera remis en question ensuite par les récentes données issues du séquençage des génomes complets.

1.4.1 Modèle de cassures aléatoires

En 1984, Nadeau et Taylor proposent que les réarrangements ayant eu lieu entre l'homme et la souris suivent le modèle de cassures aléatoires (Random Breakage Model) proposé par Ohno en 1973. Ce modèle suppose que les réarrangements (fixés) sont distribués uniformément et indépendamment le long du génome.

Nadeau et Taylor ont analysé la distribution de 83 marqueurs homologues sur les chromosomes de l'homme et de la souris ; ils ont obtenu 46 segments conservés, c'est-à-dire des segments qui contiennent au moins deux marqueurs qui se trouvent dans le même ordre dans les deux génomes. Sous le modèle de cassures aléatoires, la distribution des longueurs des segments conservés suit une loi exponentielle. Ils ont montré que la distribution de 13 segments identifiés avec les marqueurs suit effectivement une loi exponentielle, confirmant ainsi le modèle aléatoire. Ils ont également estimé le nombre de réarrangements ayant eu lieu entre les deux espèces : 178 plus ou moins 38 réarrangements.

Dans les années qui suivent, de plus en plus de marqueurs sont identifiés et cartographiés dans les génomes. Le nombre de segments conservés détectés entre l'homme et la souris augmente mais leur taille reste compatible avec les prédictions du modèle de cassures aléatoires. Par exemple, en 1996, 1416 marqueurs définissent 181 segments conservés (DeBry et Seldin, 1996) et on estime que la majorité des segments conservés ont été identifiés. Le fait que le modèle reste valide avec l'augmentation de la densité de marqueurs renforce l'hypothèse d'un seul processus aléatoire de distribution des réarrangements (Nadeau et Sankoff, 1998). Cependant, les derniers segments manquants sont vraisemblablement petits et des méthodes à plus fine résolution seront nécessaires pour valider cette hypothèse. C'est le nombre de ces petits segments conservés qui déterminera si un autre processus est en jeu.

1.4.2 Modèle des régions fragiles

En 2003, les génomes de l'homme et de la souris sont entièrement séquencés et une première version de leur assemblage est disponible. Des méthodes d'alignement de génomes complets sont alors développées qui permettent d'identifier les segments conservés à une résolution encore jamais atteinte (Pevzner et Tesler, 2003a; Kent *et al.*, 2003).

A grande échelle, les résultats sont compatibles avec ceux obtenus avec des cartes génétiques et avec le modèle de cassures aléatoires, avec 281 segments conservés de plus de 1 Mb (issus de "macro-réarrangements" pour Pevzner et Tesler (2003a)) et 344 de plus de 100 Kb avec la méthode de Kent *et al.* (2003).

Par contre, à plus petite échelle, le nombre de petits segments conservés est plus important qu'attendu et ne peut être expliqué par le modèle de cassures aléatoires (Kent *et al.*, 2003). Pevzner et Tesler (2003a) ont détecté plus de 3000 micro-réarrangements à l'intérieur des 281 segments conservés issus de macro-réarrangements (> 1 Mb). De plus, la répartition de ces petits segments ne semble pas uniforme le long du génome.

Pevzner et Tesler (2003a) poursuivent leur étude en calculant le nombre minimal de macro-réarrangements (inversion, translocation, fusion et fission) nécessaires pour passer d'un génome à l'autre. Ils obtiennent le chiffre de 245 réarrangements (borne inférieure) entre l'homme

et la souris à partir des 281 segments conservés détectés. Le nombre de réarrangements obtenu est grand par rapport au nombre de points de cassure (régions séparant deux segments conservés), 258. En effet, un réarrangement tel qu'une inversion donne généralement deux points de cassure. Or, pour obtenir seulement 258 points de cassure avec au moins 245 réarrangements, il est nécessaire que certains points de cassure aient été utilisés pour plusieurs réarrangements. En moyenne, chaque point de cassure a été utilisé 1.9 fois. Les auteurs appellent cela la ré-utilisation de points de cassure. Ils proposent alors un modèle différent du modèle de cassures aléatoires : le modèle de régions fragiles (fragile breakage model) dans lequel certaines régions du génome sont plus fragiles que d'autres et subissent plus souvent des réarrangements (Pevzner et Tesler, 2003b).

Ce nouveau modèle donnera lieu à une importante polémique entre partisans du modèle aléatoire et ceux du modèle fragile (Pevzner et Tesler, 2003b; Trinh *et al.*, 2004; Sankoff et Trinh, 2005; Peng *et al.*, 2006; Sankoff, 2006; Alekseyev et Pevzner, 2007). Cependant, la polémique porte principalement sur l'utilisation de l'indice du taux de ré-utilisation. Ainsi, Sankoff et collègues soutiennent que ce taux de ré-utilisation ne reflète pas une propriété biologique de l'évolution des génomes, mais est dû à des artefacts de la méthode. D'une part, le fait d'éliminer les petits blocs (inférieurs à 1 Mb) peut produire un fort taux de ré-utilisation (Trinh *et al.*, 2004; Sankoff et Trinh, 2005). D'autre part, la perte de signal dans les génomes produisant des arrangements de segments conservés proches de l'aléatoire serait également responsable de ce fort taux de ré-utilisation (Sankoff, 2006).

Cependant, le modèle fragile est soutenu par d'autres arguments. Notamment, plusieurs analyses montrent une ré-utilisation des points de cassure au cours de l'évolution. Il s'agit ici, contrairement au taux de Pevzner et collègues, de cas de ré-utilisation observés et non théoriques. En comparant plusieurs espèces de mammifères, de mêmes régions sur le génome de l'homme sont identifiées comme ayant "cassé" plusieurs fois dans des lignées différentes (Murphy *et al.*, 2005; Yue *et al.*, 2005; Hinsch et Hannenhalli, 2006; Gordon *et al.*, 2007). De même, certaines régions impliquées dans des réarrangements de désordres génomiques et même de cancers sont co-localisées avec des réarrangements évolutifs (Murphy *et al.*, 2005; Darai-Ramqvist *et al.*, 2008).

Cependant, ces résultats dépendent fortement de la résolution des points de cassure comparés. Dans une autre analyse des réarrangements évolutifs chez les mammifères, seulement 8 % des points de cassure sont trouvés ré-utilisés au cours de l'évolution (Ma *et al.*, 2006), alors que l'étude de Murphy et collègues reportait un taux d'environ 20 % mais avec des régions de cassure de l'ordre de 1 Mb (Murphy *et al.*, 2005). Ces différences soulignent également le fait qu'on ne sait pas à l'heure actuelle quelle est la taille de ces régions fragiles (Kehrer-Sawatzki et Cooper, 2008).

Ces récentes observations soutiennent l'existence de régions dans le génome qui seraient plus fragiles et subiraient fréquemment des réarrangements au cours de l'évolution, et d'autre part des régions "solides" qui ne casseraient que très rarement. Notons cependant que ce que nous observons est le résultat de la sélection. Ainsi, la fragilité et la solidité apparentes de certaines régions ne sont pas forcément le reflet de susceptibilités différentes de la molécule d'ADN aux cassures et aux réarrangements, mais peut-être de pressions de sélection différentes.

1.4.3 Analyse des points de cassure

S'il existe des régions plus fragiles que d'autres dans le génome, on peut alors se demander quelles sont leurs caractéristiques et qu'est-ce qui détermine la fragilité d'une région.

De nombreuses analyses de régions de cassure ont été effectuées. Avant le séquençage de génomes complets, des analyses ponctuelles de réarrangements évolutifs ou bien impliqués dans des désordres génomiques étaient effectuées. Ces analyses portent sur un point de cassure ou un réarrangement à la fois avec l'objectif, le plus souvent, d'identifier les mécanismes moléculaires responsables de leur formation. Elles ont l'avantage d'être très détaillées. Par contre, on ne peut tirer de conclusion générale et évaluer statistiquement les associations trouvées. Avec l'augmentation des données de réarrangements, des analyses ont été menées de manière systématique, prenant en compte tous les points de cassure détectés sur un génome. On a alors cherché des caractéristiques de séquences communes à plusieurs points de cassure et à corréler les positions des points de cassure avec d'autres informations génomiques. Nous présentons ici les principaux résultats de la littérature, ils sont également revus de manière plus détaillée dans l'article de revue que nous avons publié (Lemaitre et Sagot, 2008).

a. Duplications segmentaires

Les duplications de séquence ont depuis longtemps été recherchées dans les séquences de points de cassure pour leur rôle potentiel dans la formation des réarrangements. En effet, comme nous l'avons vu dans la Section 1.2, la recombinaison homologue entre des séquences dupliquées (ou NAHR) peut engendrer toutes sortes de réarrangements (exemple de l'inversion dans la Figure 1.6).

Les duplications segmentaires (SD, pour Segmental Duplications en anglais) sont des grandes duplications ($>$ 1 à 15 kb selon les études) très similaires ($>$ 95 % d'identité de séquence) et qui sont présentes en un faible nombre de copies (par opposition aux éléments transposables). Elles sont également appelées LCRs pour Low Copy Repeats en anglais. Elles ont été beaucoup étudiées dans le génome humain depuis son séquençage, notamment car leur fréquence était plutôt inattendue. Les duplications segmentaires couvrent en effet plus de 5.2 % du génome humain. Leur distribution dans le génome n'est pas uniforme, 35 % sont localisées dans les régions subtélomériques ou péricentromériques et elles apparaissent le plus souvent en clusters (lire la revue très complète de Bailey et Eichler (2006)).

Avant le séquençage du génome humain, elles avaient déjà été mises en cause dans des désordres génomiques (lire les revues (Ji *et al.*, 2000; Stankiewicz et Lupski, 2002)) et dans certains points de cassure chez les primates (Dennehey *et al.*, 2004; Goidts *et al.*, 2004, 2005; Stankiewicz *et al.*, 2001; Szamalek *et al.*, 2006; Cáceres *et al.*, 2007). Dans ces cas, les auteurs ont montré que les duplications étaient la cause des réarrangements, par le mécanisme de NAHR. Notamment pour les désordres génomiques, il est facile de montrer que les SDs étaient présentes avant le réarrangement. Kehrer-Sawatzki et Cooper (2008) proposent également que les SDs ne seraient pas forcément impliqués dans le réarrangement par le mécanisme de NAHR mais permettraient d'apparier (de mettre en contact) les régions réarrangées ; cela expliquerait pourquoi les duplications sont parfois éloignées du point de cassure.

En ce qui concerne les analyses systématiques des réarrangements évolutifs, plusieurs groupes ont trouvé une corrélation significative entre les SDs et les points de cassure entre l'homme et la souris (Bailey *et al.*, 2004; Armengol *et al.*, 2003). Armengol et collègues ont trouvé que 53 % des points de cassure entre l'homme et la souris contiennent au moins une SD dans une fenêtre de 25 Kb autour du point de cassure. Le chiffre de Bailey et collègues est plus faible, 26 % (dans des fenêtres de 10 Kb autour du point de cassure), mais l'association reste

très significative. Alors que les premiers proposent un effet causal des duplications, Bailey et collègues sont plus prudents et observent notamment que des associations existent entre des duplications spécifiques à l'homme et des points de cassure qui se sont produits dans la lignée de la souris. Ils expliquent alors cette co-localisation par le modèle des régions fragiles. En effet, on peut considérer ces duplications comme un type de réarrangement particulier. Ainsi, comme les réarrangements équilibrés, leurs localisations suivraient le même modèle des régions fragiles.

Cette association a été confirmée par de nombreuses études ultérieures chez l'homme (Murphy *et al.*, 2005; Hinsch et Hannenhalli, 2006; Newman *et al.*, 2005; Carbone *et al.*, 2006), ainsi que chez la souris (Armengol *et al.*, 2005) et également avec des réarrangements non évolutifs, polymorphes chez l'homme (Sharp *et al.*, 2005; Tuzun *et al.*, 2005; Feuk *et al.*, 2006; Korbel *et al.*, 2007) et dans les cellules cancéreuses (Darai-Ramqvist *et al.*, 2008).

Une autre relation a été mise en évidence lors de l'analyse des points de cassure d'une inversion séparant l'homme et le chimpanzé (Kehrer-Sawatzki *et al.*, 2005b). Les deux points de cassure chez l'homme correspondent à deux duplications chez le chimpanzé de 83 et 23 Kb respectivement. L'arrangement des deux copies de chacune des deux duplications, et le fait que le réarrangement s'est produit dans la lignée du chimpanzé sont cohérents avec un modèle où les duplications ne sont pas la cause du réarrangement mais la conséquence de celui-ci. Le mécanisme invoqué met en jeu des cassures double brin à bouts collants dont les extrémités simple brin seraient complétées après l'inversion du segment engendrant des duplications inversées de part et d'autre du segment inversé (voir Figure 1.6 à droite). Les duplications sont très grandes dans ce cas, et on peut se demander si une cassure double brin à bout collant longue de 80 Kb est réaliste. Cependant, récemment ce modèle a été invoqué pour un grand nombre de réarrangements chez la drosophile, cette fois pour des duplications de quelques centaines de paires de bases jusqu'à 2 Kb (Ranz *et al.*, 2007). Sur 55 points de cassure analysés, 34 présentent des duplications inversées qui ne sont présentes que dans le génome ayant subit le réarrangement.

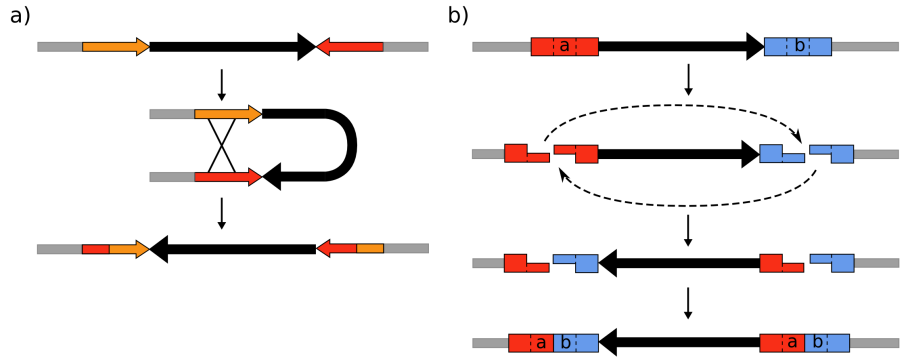


Fig. 1.6: Image reproduite à partir de Casals et Navarro (2007) montrant les deux mécanismes de réarrangement associant une inversion et des duplications inversées à chaque point de cassure. En a) la recombinaison homologue entre les deux copies de la duplication inversée est la cause de l'inversion (NAHR). En b) c'est l'inversion qui est la cause des duplications, à cause de cassures double brin à bouts collants.

Ainsi, si la corrélation entre les points de cassure de réarrangements (équilibrés) et les duplications de séquence est très forte, les liens de cause à effet entre les deux types d'évènements peuvent être dans les deux sens ou bien inexistants (s'il s'agit d'évènements indépendants). En conséquence, dans la majorité des cas, cette corrélation ne permet pas d'expliquer la localisation des points de cassure et la fragilité de ces régions, si elle existe.

b. Eléments transposables

D'autres types d'éléments répétés ont été recherchés dans les régions de cassure. Notamment les éléments transposables sont souvent recherchés car ils peuvent être impliqués dans le mécanisme de réarrangement, soit par recombinaison ectopique (NAHR) entre copies, soit par des mécanismes spécifiques liés à leur transposition (voir Section 1.2.4). Ainsi, de nombreuses analyses individuelles identifient des éléments transposables dans les régions de cassure ou à proximité. Des éléments des différentes classes sont mis en cause : des éléments de type LINE (Goidts *et al.*, 2005), de type SINE, dont notamment des éléments Alu, (Dennehey *et al.*, 2004; Kehrer-Sawatzki *et al.*, 2002; Goidts *et al.*, 2005), et LTR (Kehrer-Sawatzki *et al.*, 2002, 2005a). Cependant, ces résultats sont très anecdotiques.

Plusieurs analyses portant sur un plus grand nombre de points de cassure ont également révélé un enrichissement en éléments transposables. Par exemple, Dehal *et al.* (2001) ont trouvé un enrichissement en éléments L1 (de type LINE) et LTR dans 15 points de cassure du chromosome 19 humain comparé au génome de la souris. Les éléments L1 sont également surreprésentés dans les points de cassure de réarrangements polymorphes chez l'homme (Korbel *et al.*, 2007). Cependant, Dobigny *et al.* (2004) n'ont pas trouvé d'association significative entre les éléments LINE et les points de cassure chez les mammifères. Dans une autre analyse, ce sont les éléments de type SINE qui sont trouvés majoritairement dans les points de cassure (Ma *et al.*, 2006). Dans une analyse d'environ 200 points de cassure entre le génome humain et celui de la vache, les auteurs n'ont pas trouvé de classe d'ET sur-représentée par rapport aux autres, mais ils mettent en évidence un lien entre l'origine du réarrangement et celle des éléments transposables présents dans les points de cassure (Schibler *et al.*, 2006). Par exemple, les points de cassure issus de réarrangements de la lignée humaine seraient enrichis en éléments Alu spécifiques des primates, contrairement aux autres points de cassure ; de même, les points de cassure sur le génome de la souris, issus de réarrangements survenus dans le lignée de la souris, seraient enrichis en éléments SINE spécifiques de cette lignée.

On observe que toutes ces analyses ne s'accordent pas sur la classe d'éléments transposables mise en cause et certaines ne reportent de corrélation significative avec aucune classe d'éléments transposables (Hinsch et Hannenhalli, 2006). De plus, la distribution de ces différentes classes le long des chromosomes est corrélée à d'autres structures du génome à plus grande échelle, telles que la densité en gènes, le contenu en GC et les isochores. On peut donc se demander dans quelle mesure la présence de ces éléments est une caractéristique locale liée aux points de cassure, ou bien reflète une corrélation de ces derniers avec l'organisation à plus grande échelle des génomes.

Ainsi, la présence d'éléments transposables dans les points de cassure de réarrangement n'est pas systématique et les liens entre ces deux phénomènes sont mal établis à l'heure actuelle.

c. Sites fragiles

Une autre corrélation intéressante a été mise en évidence avec les points de cassure de réarrangements évolutifs ; elle concerne les sites fragiles.

Les sites fragiles sont des sites sur le génome qui ont tendance à "casser" dans certaines conditions de culture des cellules. Ils sont détectés par des méthodes cytogénétiques avec une résolution assez faible de l'ordre de la bande cytogénétique ($\sim$ 4 Mb). On distingue deux types de sites fragiles en fonction de leur fréquence dans les populations : les sites fragiles communs (environ une centaine annotés) et les sites fragiles rares (présents dans moins de 5 % de la population humaine, environ une trentaine annotés) (lire les revues (Handt *et al.*, 2000; Schwartz *et al.*, 2006)). Très peu de choses sont connues sur les bases moléculaires de ces fragilités. Aucun motif ou répétition particuliers n'ont été trouvés dans la vingtaine de sites fragiles communs caractérisés au niveau moléculaire, alors que certains sites fragiles rares présentent des régions AT-riches, des motifs CCG répétés et des séquences capables de former des structures secondaires (Ruiz-Herrera *et al.*, 2006). Les sites communs se répliqueraient tardivement et auraient des caractéristiques de bandes G à l'intérieur de bandes R. Ils seraient conservés au moins chez les primates. Enfin, ils seraient impliqués dans certains cancers (Arlt *et al.*, 2006) et on pense naturellement que ces fragilités peuvent être liées aux réarrangements évolutifs.

En effet, ils ont été trouvés associés à des points de cassure de réarrangements évolutifs à plusieurs reprises (Ruiz-Herrera *et al.*, 2002, 2005a,b, 2006; Ruiz-Herrera et Robinson, 2007) ; ils montrent également, dans certains cas, une association avec des séquences télomériques intra-chromosomales (ITs). Mais la corrélation est parfois faible, la résolution des sites fragiles et des points cassure est grossière (la bande chromosomique >1 Mb) et aucun modèle n'est proposé pour expliquer cette corrélation.

En ce qui concerne les ITs, Nergadze *et al.* (2004) proposent qu'ils ne représentent pas des sites fragiles pour les réarrangements mais que, au contraire, ils seraient causés par les cassures double brin et seraient insérés par la machinerie NHEJ lors de la réparation de ces dernières.

d. Motifs et séquences particulières

Certains motifs ou séquences particulières ont également été recherchés dans les points de cassure. Notamment, les analyses se sont concentrées sur trois types de séquences : (1) des motifs associés à la recombinaison, tels que les motifs spécifiques de la recombinaison V(D)J, le site χ, les sites de fixation de la topo-isomérase, les motifs associés aux points chauds de recombinaison (Abeysinghe *et al.*, 2003), (2) des séquences capables de former des structures secondaires, comme des répétitions simples directes ou inversées de quelques paires de bases, des séquences palindromiques, des mini- et micro-satellites, et (3) des séquences pouvant engendrer des conformations particulières de la molécule d'ADN, comme des séquences poly-purines $(R)_n$, poly-pyrimidines $(Y)_n$ ou alternant purines et pyrymidines $(RY)_n$ (Bacolla *et al.*, 2004). Ces trois types de séquences pourraient engendrer des cassures double brin.

Ponctuellement, toutes ces séquences ont été trouvées dans certains points de cassure évolutifs. On peut noter par exemple la présence de courtes répétitions simples et de satellites, de séquences $(R)_n$, $(Y)_n$ et $(RY)_n$, de palindromes riches en AT, de sites de fixation de topo-isomérases dans des points de cassure entre l'homme et le chimpanzé (Kehrer-Sawatzki *et al.*, 2002, 2005b; Szamalek *et al.*, 2005; Goidts *et al.*, 2005). Récemment, l'étude de points chauds de réarrangements (sites subissant des réarrangements de manière récurrente, observés dans la population humaine) a mis en évidence un rôle important des palindromes riches en AT (ou PATRRs) dans la formation des cassures double brin et leur réparation par NHEJ (Kurahashi *et al.*, 2003, 2007; Gotter *et al.*, 2007; Babcock *et al.*, 2007; Kato *et al.*, 2008). Notons que dans ces cas, le point de cassure est identifié au centre du palindrome.

De manière systématique, ces associations ont essentiellement été étudiées pour les points de cassure de cancers et de désordres génomiques chez l'homme. Des associations assez fortes ont été mises en évidence, parfois spécifiques de certains types de réarrangements. Par exemple, Abeysinghe *et al.* (2003) rapportent que les séquences $(RY)_n$ sont sur-représentées dans les points de cassure de délétions, alors que pour les translocations on trouverait plutôt des séquences $(Y)_n$. Chuzhanova *et al.* (2003) ont trouvé plus de 83 % des points de cassure étudiés (39/47) contenant des petites répétitions de quelques paires de bases. Enfin, dans une étude systématique de 222 points de cassure de cancers et de maladies, la distance de ces derniers à une séquence $(R)_n$, $(Y)_n$ ou $(RY)_n$ est significativement plus faible qu'attendu par hasard (Bacolla *et al.*, 2004).

Si ces associations sont relativement fortes pour les points de cassure de cancers et de maladies, elles n'ont pas été reportées de manière systématique dans les points de cassure évolutifs. Comme ces différents types de points de cassure sont souvent co-localisés sur le génome (Eichler et Sankoff, 2003; Murphy *et al.*, 2005), on pourrait espérer retrouver ces corrélations. La principale difficulté est sûrement d'obtenir une aussi bonne précision des points de cassure pour conduire de telles analyses. Une autre raison qui pourrait expliquer que ces caractéristiques ne sont pas présentes dans les points de cassure évolutifs serait le temps. Si les points de cassure sont très anciens (comparés aux réarrangements polymorphes et de cancers), on peut s'attendre à ce que les caractéristiques de séquence ayant causé ou résultant des réarrangements se sont effacées avec l'évolution des séquences.

1.4.4 Travail de thèse

Le travail de thèse se place dans ce contexte et ces problématiques. Si le modèle de cassures aléatoires a été rejeté, on ne sait toujours pas ce qui détermine la distribution non uniforme des points de cassure et quels facteurs génomiques influencent leurs localisations dans les génomes. Les analyses des points de cassure évolutifs effectuées jusqu'à présent sont en fait limitées principalement par la faible précision de leurs coordonnées génomiques. Nous proposons dans cette thèse d'améliorer la définition des points de cassure sur les génomes de mammifères afin d'analyser de manière systématique et à grande résolution leurs caractéristiques génomiques.

Chapitre 2

Détection des réarrangements par comparaison de génomes

Sommaire

L'étude du contenu des régions de cassure nécessite au préalable de les identifier sur les génomes. Cette étape est très importante et influence la suite des analyses. Bien sûr, nous voulons que les données soient fiables, non biaisées et les plus précises possible. Dans ce chapitre, nous établissons un état de l'art des méthodes permettant de détecter ces régions sur les génomes.

Pour identifier les réarrangements évolutifs et leurs points de cassure, les seules données disponibles sont les arrangements actuels des génomes. Les génomes actuels sont issus d'un génome ancêtre commun et ont divergé en accumulant des mutations ponctuelles mais aussi des réarrangements chromosomiques. L'approche privilégiée sera donc une approche comparative. En comparant les génomes actuels, on cherche tout d'abord à identifier ce qui a été conservé, c'est-à-dire ce qui n'a pas été réarrangé. On appelle ces régions, des **régions conservées**. Il s'agit de régions contiguës, une dans chaque génome comparé qui sont issues d'une même région chez leur ancêtre commun (on dit qu'elles sont **homologues**) et qui n'ont subit aucun réarrangement depuis la divergence des espèces.

Dans un deuxième temps, on pourra comparer les arrangements respectifs de ces régions dans les différents génomes comparés, pour inférer les réarrangements qui se sont produits au

cours de l'évolution et identifier les régions de cassure. Nous nous attachons, dans ce chapitre, à la première partie, c'est-à-dire l'identification des régions conservées.

Les méthodes pour identifier des régions conservées entre plusieurs génomes sont variées. Elles sont toutes basées sur la recherche de similarité, mais à des niveaux différents. Elles se distinguent principalement par le type de données qu'elles comparent. Ainsi, historiquement, lorsque les seules données disponibles sur les génomes étaient les caryotypes, on a recherché des similarités au niveau des caryotypes. Puis, avec les développements de la cytogénétique et de l'hybridation *in situ*, on a identifié des similarités de séquences grâce au phénomène d'hybridation des molécules d'ADN. Avec l'accumulation de plus en plus d'informations sur les génomes, des méthodes bioinformatiques ont été développées pour comparer l'ordre de marqueurs homologues dans différents génomes. Pour finir, nous disposons actuellement de la séquence complète (ou presque) de certains génomes, permettant ainsi de les comparer au niveau nucléotidique.

2.1 Méthodes expérimentales

Avant le séquençage de génomes complets, les méthodes d'identification des réarrangements étaient principalement expérimentales.

2.1.1 Comparaison de caryotypes

La première méthode consistait à comparer les caryotypes de différents organismes. Le caryotype est une image du noyau en métaphase, où les chromosomes sont dans une configuration très compacte. On peut distinguer les différents chromosomes et les identifier par leur taille. Ainsi on peut comparer le nombre et la taille approximative des chromosomes entre plusieurs génomes. Puis, dans les années 1970, une technique appelée "chromosome banding" a permis d'identifier un peu mieux la structure des chromosomes. Différentes colorations permettent de distinguer différentes parties du chromosome appellées bandes (par exemple, les bandes G et les bandes R), caractérisées par des niveaux de compaction et des compositions nucléotidiques différentes. Ainsi, on peut caractériser un chromosome par son patron en bandes (succession et taille des différents types de bandes). Cela a permis la comparaison des caryotypes à une échelle plus fine : celle de la bande chromosomique (Figure 2.1). La taille d'une bande étant approximativement de 4 Mb, seuls les grands réarrangements, tels que les inversions, les translocations de grands segments chromosomiques, ainsi que les fusions et fissions de chromosomes, peuvent être identifiés. De plus, si les organismes sont trop éloignés évolutivement, le patron en bande peut prêter à confusion ; en effet, deux patrons similaires ne sont pas synonymes d'homologie de séquence, mais indiquent seulement des propriétés structurales similaires.

2.1.2 Hybridation in situ : FISH

Dans les années 1990, ce problème a pu être résolu, puisqu'on a pu identifier des similarités de séquence grâce au développement d'une technique majeure dans le domaine de la cytogénétique : la technique FISH ("Fluorescent In Situ Hybridization"). Comme son nom l'indique, cette technique est basée sur le principe d'hybridation. L'hybridation est un processus moléculaire qui joint deux molécules d'ADN simple brin complémentaires pour former une molécule d'ADN double brin. FISH utilise ce processus pour localiser des sondes sur des chromosomes

FIG. 2.1: Caryotypes de deux mammifères, (a) *Dasypus novemcinctus* (tatou à neuf bandes) et (b) *Lepus europaeus* (lièvre), les chromosomes sont colorés de façon à faire ressortir les bandes chromosomiques de type R. Figure extraite de (Richard *et al.*, 2003), étude effectuée dans le but de reconstruire le caryotype ancestral des mammifères euthériens.

cibles. Les sondes sont des petites molécules d'ADN simple brin, marquées par une substance fluorescente. Elles s'hybrident sur les chromosomes cibles sous forme simple brin lorsque leurs séquences sont complémentaires. La fluorescence est ensuite capturée sur une image permettant de localiser les sondes sur les chromosomes cibles. Par exemple, par cette technique, si on crée des sondes avec l'ADN du chromosome 4 du chimpanzé et qu'on les hybride sur une préparation avec l'ensemble des chromosomes de l'homme, on pourra identifier tous les chromosomes de l'homme qui sont homologues en partie au chromosome 4 du chimpanzé. La résolution reste faible, puisque la fluorescence ne peut être détectée que si elle couvre une distance physique suffisammment importante. En fonction de la condensation des chromosomes cibles, la résolution varie de 50 Kb pour les chromosomes en interphase à 3 Mb pour les chromosomes en métaphase. Cette méthode, appelée peinture chromosomique, a été étendue (zoo-FISH) pour être appliquée sur des espèces plus éloignées (Figure 2.2). Mais son principal inconvénient est qu'elle ne permet de détecter que les réarrangements inter-chromosomiques.

2.1.3 Hybridation comparée : CGH

Inspirée de la technique FISH, l'hybridation génomique comparée (CGH en anglais) permet d'identifier les réarrangements quantitatifs, tels que les duplications et délétions, mais n'est pas capable d'identifier les réarrangements équilibrés tels que les inversions et translocations. Cette technique consiste à hybrider deux solutions d'ADN, une de référence et l'autre d'intérêt, marquées avec des fluorochromes différents, sur une préparation normale de chromosomes en métaphase. En comparant les niveaux relatifs de fluorescence, on peut identifier les gains et les pertes le long des chromosomes.

Limitée en résolution comme la technique FISH, cette méthodologie s'est développée grâce à l'essort d'un autre support : les puces ou micro-puces à ADN. Avec les puces CGH, au lieu de chromosomes en métaphase, l'hybridation est réalisée sur une puce microscopique contenant

FIG. 2.2: Exemple d'analyse par zoo-FISH ; les photographies à gauche montrent l'hybridation des chromosomes de chat (FCA) et de gibbon (HCO) sur des chromosomes de baleine, les positions des signaux sont schématisées à droite sur le dessin. Cette figure est tirée de (Froenicke, 2005).

des milliers de clones ou sondes d'ADN couvrant tout le génome de référence. La résolution de cette méthode dépend du type de sondes présentes sur la puce, allant de quelques centaines de kilobases avec des BACs au nucléotide près avec des oligonucléotides chevauchants ("tiling arrays"). L'analyse de ces puces n'est pas triviale et nécessite des méthodes de segmentation du signal. Elles sont très utilisées actuellement, notamment pour l'identification de variants structurels (polymorphisme de nombre de copies) et l'analyse de cellules cancéreuses.

Cependant, ces deux techniques ne permettent pas de localiser l'emplacement des régions variantes : on connaît la localisation sur le génome normal d'une région amplifiée, mais on ne connaît pas les localisations des copies supplémentaires dans le génome anormal.

2.1.4 Séquençage d'extrémités appariées

Le PEM ("Paired-End Mapping") est une technique développée très récemment qui permet d'identifier presque tous les types de réarrangements (équilibrés ou non) à grande échelle et à haute résolution. Le principe est de découper de façon aléatoire le génome d'intérêt en clones de taille constante, puis de séquencer systématiquement les deux extrémités de chaque clone. Ces séquences sont ensuite cartographiées sur le génome de référence par alignement de séquences, et les réarrangements sont identifiés lorsque la distance ou l'orientation des deux extrémités d'un clone ne sont pas concordantes (Figure 2.3).

Cette technique présente quelques limitations. Bien sûr, elle nécessite que le génome de référence soit entièrement séquencé. De plus, les extrémités séquencées sont en général courtes (les plus longues sont de l'ordre de 500 paires de base), il est donc nécessaire que les génomes comparés ne soient pas trop distants en terme de séquence pour pouvoir cartographier les extrémités sans ambiguïté.

Cette méthodologie a été utilisée d'abord avec des fosmides (clones de 40 Kb environ), permettant d'identifier des réarrangements de plus de 12 Kb entre l'homme et le chimpanzé (Newman *et al.*, 2005), et des réarrangements polymorphes chez l'homme de plus de 8 Kb (Tuzun *et al.*, 2005). Plus récemment, la combinaison de l'utilisation de clones plus petits

FIG. 2.3: Cartographie des extrémités de fosmides et exemples de positions "invalides" indiquant un réarrangement. Figure extraite de (Tuzun *et al.*, 2005)

(environ 3 Kb) avec les nouvelles technologies de séquençage à haut débit a permis d'augmenter considérablement la résolution tout en diminuant le coût (identification des réarrangements polymorphes chez l'homme $>$ 3 Kb avec une résolution moyenne de quelques centaines de paires de bases sur la localisation des points de cassure (Korbel *et al.*, 2007)). Cette technologie est également très utilisée pour détecter des aberrations chromosomiques dans les cellules cancéreuses (Bashir *et al.*, 2008).

Cette méthode est d'autant plus prometteuse qu'avec l'essort des nouvelles techniques de séquençage à haut débit, elle sera de moins en moins chère et de plus en plus précise.

Ainsi, les méthodes expérimentales sont encore largement utilisées et évoluent constament. Elles sont combinées de plus en plus avec les données de séquence et les méthodes bioinformatiques. Par exemple, les points de cassure des réarrangements identifiés par FISH, sont ensuite caractérisés au niveau moléculaire par hybridation de BACs et cartographie sur la séquence génomique (voir, par exemple, les caractérisations des points de cassure des 9 grandes inversions entre l'homme et le chimpanzé, revues dans (Szamalek *et al.*, 2005)). Elles sont complémentaires des méthodes bioinformatiques également parce qu'elles peuvent s'appliquer à un plus grand nombre d'espèces, à des organismes d'une même espèce et même à différents types de cellules (dans le cas des cancers).

2.2 Comparaison de l'ordre des gènes

Lorsqu'on a commencé à accumuler des informations de cartographie sur les génomes, des méthodes algorithmiques et bioinformatiques ont été développées afin de détecter des régions conservées dans plusieurs génomes. L'objectif de ces méthodes est d'identifier des régions conservées en comparant les positions relatives des gènes dans les différents génomes. Lorsque l'ordre et l'orientation des gènes orthologues sont conservés dans les espèces comparées, on

fait généralement l'hypothèse la plus parcimonieuse qu'elles sont issues d'une même région chez leur ancêtre commun et qu'elles n'ont subit aucun réarrangement depuis la divergence des espèces.

Notons que cette méthode ne s'applique pas seulement aux gènes, elle peut s'appliquer à n'importe quel marqueur orthologue dont on connaît la position (relative ou physique) sur les génomes comparés. On parle le plus souvent de gènes car ce sont les marqueurs les plus conservés, les plus cartographiés et les plus séquencés.

Les méthodes de comparaison de l'ordre des gènes utilisent deux types d'information :
- la localisation des gènes sur chacun des génomes comparés,
- les relations d'orthologie entre les gènes de génomes différents.

2.2.1 Localisation des gènes sur les génomes

L'information de la localisation des gènes sur le génome n'est pas triviale à obtenir. On peut distinguer deux types de méthodes : les méthodes expérimentales qui s'appliquent généralement lorsque le génome en question n'est pas entièrement séquencé et les méthodes *in silico* d'annotation des séquences complètes des génomes.

a. Cartographie expérimentale

Les techniques de cartographie expérimentales produisent principalement deux types de cartes : les cartes génétiques et les cartes d'irradiation. Les cartes génétiques sont produites par les méthodes d'analyse de liaison. Le principe est le suivant : si deux loci sont "liés", ils seront hérités ensemble. Ainsi la distance relative entre deux loci peut être approchée en estimant la fréquence avec laquelle on les observe hérités ensemble, si on suppose que la distribution des crossing-overs est uniforme le long du génome. Les cartes d'irradiation sont obtenues selon un principe similaire. Si on segmente aléatoirement l'ADN en fragments, deux loci proches ont plus de chances de rester sur le même fragment que deux loci éloignés. Concrètement, l'ADN est fragmenté par irradiation et les fragments obtenus sont indépendamment gardés ou perdus au cours des différentes générations cellulaires. On peut ainsi estimer la distance relative entre deux loci par la fréquence avec laquelle ils sont retrouvés dans la même cellule.

La résolution de ces cartes reste faible (1cR=50 à 100 kb pour les cartes d'irradiation, 1cM=1000 kb pour les cartes génétiques) et elles ne fournissent pas l'information d'orientation des gènes. Cependant, elles permettent d'étudier l'ordre des gènes dans les organismes non séquencés (Murphy *et al.*, 2005; van der Wind *et al.*, 2004).

b. Annotation des séquences de génomes

L'annotation d'une séquence génomique consiste à associer l'information biologique à la séquence. La première étape est d'identifier les éléments fonctionnels sur le génome et la deuxième est d'associer une fonction biologique à ces éléments. Dans la première étape, on cherche notamment à identifier les gènes protéiques, avec des méthodes de **prédiction de gènes**. Il existe trois types de méthodes de prédiction de gènes :
- les méthodes *ab initio* qui n'utilisent que l'information de la séquence pour prédire les gènes. Ces méthodes se basent sur les connaissances *a priori* de la structure des gènes : code génétique, longueur des gènes, des introns et des exons, signaux d'épissage, promoteurs, etc.

- les méthodes qui utilisent les séquences d'ARN messagers ou de protéines. Lorsque les séquences d'ARNm ou de protéines sont séquencées, on essaie de les positionner sur la séquence génomique, grâce à des méthodes d'alignement de séquences.
- les méthodes comparatives qui utilisent les annotations et informations acquises sur d'autres génomes d'espèces proches.

Ces méthodes sont complémentaires les unes des autres et sont souvent combinées entre elles. Cependant, les résultats de ces méthodes sont seulement des prédictions et elles ont souvent tendance à sur-estimer le nombre de gènes en annotant, par exemple, des pseudo-gènes comme des gènes. Ces prédictions ont besoin d'être vérifiées expérimentalement.

2.2.2 Assignation d'orthologie

L'assignation d'orthologie est une étape clé pour l'identification des segments conservés. Nous disposons, ici, pour chaque génome à comparer d'un ensemble de gènes (ou marqueurs) avec leur séquence. L'objectif de cette étape est d'identifier quels gènes sont homologues, et plus précisément on recherche le plus souvent les gènes orthologues.

a. Homologie, orthologie, paralogie

Deux séquences sont homologues si elles dérivent d'une même séquence. On distingue deux types d'homologie : l'**orthologie** et la **paralogie**. Deux séquences sont orthologues si elles ont un ancêtre commun (donc si elles sont homologues) et si elles ont divergé à la suite d'un évènement de spéciation, alors que deux séquences paralogues ont divergé à la suite d'un évènement de duplication. Deux séquences orthologues appartiennent à des espèces différentes, on dit aussi que deux gènes sont orthologues s'ils ont dérivés d'un unique gène ancestral dans le dernier ancêtre commun des espèces.

L'exemple couramment utilisé pour illustrer ces concepts est celui de l'insuline (Figure 2.4). Chez l'ancêtre de l'homme, de la souris et du rat, nous avons un gène INS. Dans la branche des rongeurs, avant la spéciation du rat et de la souris, le gène INS est dupliqué ; on appelle INS1 et INS2 les deux copies résultant de cette duplication. Ainsi, actuellement, l'homme possède une copie INS, et le rat et la souris possèdent chacun deux copies INS1 et INS2. L'arbre de la Figure 2.4 représente l'histoire évolutive de ce gène depuis l'ancêtre commun à l'homme, la souris et le rat. On y distingue deux types de noeuds, ceux qui représentent des évènements de spéciation, entre l'homme et les rongeurs, puis entre le rat et la souris, et un noeud représentant la duplication du gène. En utilisant l'arbre, il est facile d'identifier les relations d'orthologie et de paralogie. Si on prend deux gènes actuels, il suffit de remonter l'arbre jusqu'à leur premier noeud commun ; s'il s'agit d'un noeud de spéciation, ces deux gènes sont orthologues, s'il s'agit d'un noeud de duplication, ce sont des paralogues. Ainsi le gène INS de l'homme est orthologue aux deux copies INS1 et INS2 de la souris (comme de celles du rat), le gène INS1 de la souris est orthologue au gène INS1 du rat, mais paralogue du gène INS2 du rat.

On remarque alors que les relations d'orthologie peuvent être *multiples*, c'est-à-dire que n gènes d'une espèce peuvent être orthologues à m gènes d'une autre espèce. Il suffit que les duplications à l'origine des n copies d'une part, et des m copies d'autre part, soient postérieures à la spéciation. On appelle les n copies de l'espèce 1 des co-orthologues ou in-paralogues.

Ainsi, on aimerait aller plus loin et définir l'**orthologie de position** qui serait cette fois 1-1. Parmi les orthologues n-m, on est alors intéressés par la paire de gènes dont la position est orthologue.

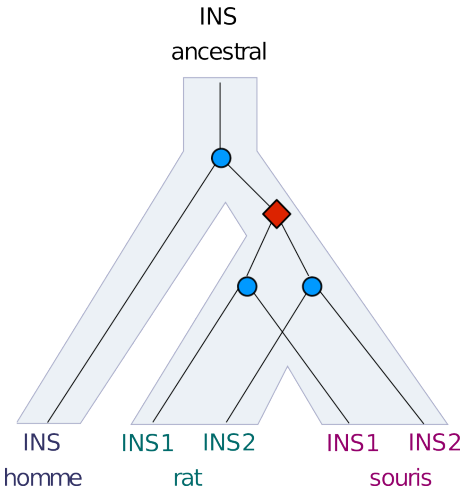


FIG. 2.4: Exemple du gène de l'insuline pour illustrer les concepts d'orthologie et de paralogie. Arbre du gène INS depuis la divergence des trois espèces homme, souris, rat. Les ronds bleus représentent des évènements de spéciation, alors que le losange rouge représente une duplication.

Reprenons l'exemple de l'insuline, le gène INS a été dupliqué dans la lignée des rongeurs. Plus précisément, supposons qu'on puisse identifier une source et une cible parmi les deux copies : c'est le gène INS1 de la souris qui a été copié en gène INS2 (Figure 2.5). Par définition, le gène de l'insuline humain INS est orthologue à INS1 et INS2 chez la souris. Cependant INS1 a gardé la position ancestrale, alors que INS2 est une copie de INS1 insérée à un autre locus. INS et INS1 sont orthologues de position. L'orthologie de position est définie dans (Burgetz *et al.*, 2006; Bourque *et al.*, 2005; Swidan *et al.*, 2006), par le fait que les gènes gardent le même contexte génomique. Deux types d'orthologie de position sont distingués dans (Dewey et Pachter, 2006) en fonction du type de duplication. Les auteurs définissent une duplication comme étant *dirigée* si on peut identifier une source et une cible. Par exemple, dans le cas d'une duplication en tandem, il paraît difficile d'identifier quelle copie a été la source et laquelle a été la cible puisqu'elles ont toutes deux le même contexte génomique. Les auteurs définissent alors les topo-orthologues qui sont des orthologues qui depuis leur divergence n'ont été ni l'un ni l'autre la cible d'une duplication dirigée. Parmi ces topo-orthologues, ils définissent une sous-division : les monotopo-orthologues, des topo-orthologues qui depuis la divergence n'ont pas subit de duplication non dirigée. Cette dernière relation est la seule à être 1-1.

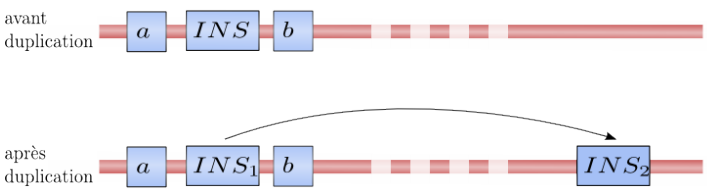


FIG. 2.5: Schématisation de la duplication dirigée du gène INS. INS1 est la source, son contexte génomique est inchangé (schématisé par les carrés a et b), et INS2 est la cible.

L'importance de l'orthologie dans l'étude des réarrangements Le plus souvent, les duplications ne sont pas prises en compte dans l'étude des réarrangements. La principale justification est que ce sont des évènements qui touchent des zones limitées sur le génome (souvent limitées à un seul gène). Il y a, bien sûr, des exceptions, par exemple, concernant les duplications segmentaires chez les primates, ou bien les duplications complètes de génomes. Mais le plus souvent les duplications sont étudiées séparément des autres réarrangements. Ainsi, lorsqu'on recherche les modifications de l'ordre des gènes dues uniquement aux réarrangements équilibrés tels que les inversions, translocations, etc., il est important d'utiliser des marqueurs orthologues de position. Si ce n'est pas le cas, les deux marqueurs auront des positions différentes qu'on imputera à tort à des réarrangements équilibrés.

b. Méthodes d'assignation d'orthologie et alignement de séquences

On détecte l'homologie de séquence avec un critère de similarité de séquence. En effet, deux séquences homologues sont issues d'un ancêtre commun, elles proviennent donc d'une même séquence qui a évolué indépendamment dans les deux lignées. Si le temps évolutif n'est pas trop long, on espère que ces deux séquences sont similaires, c'st-à-dire qu'elles possèdent encore un certain nombre de nucléotides inchangés.

L'alignement de séquences Pour identifier la similarité de séquences, on utilise le plus souvent l'alignement de séquences. Un alignement de deux séquences est une façon de placer ces séquences l'une en face de l'autre, en autorisant des décalages ponctuels, appelés indels. A chaque position de l'alignement, on distingue trois types d'états : l'appariement lorsque les deux lettres sont identiques, le mésappariement lorsque les deux lettres sont différentes, et l'indel lorsqu'une lettre est en face d'un décalage de l'autre séquence. On peut donc définir un score de similarité associé à cet alignement, simplement en comptant positivement les appariements et négativement les deux autres états. Le problème de l'alignement de deux séquences devient alors : comment placer les indels pour maximiser le score de similarité. Nous ne décrivons pas ici les différentes approches et algorithmes d'alignement (voir Section 4.2.1), nous utiliserons seulement le fait qu'à partir d'un alignement, on dispose d'une mesure de la similarité entre deux séquences.

Il suffit alors de fixer un seuil de similarité entre deux séquences à partir duquel on les considère comme étant homologues. Même si le choix de ce seuil n'est pas trivial, la distinction entre orthologues et paralogues est encore plus complexe.

Plusieurs méthodes d'assignation d'orthologie La plupart des méthodes d'assignation d'orthologie partent du principe que les séquences de gènes orthologues sont plus similaires entre elles qu'elles ne le sont avec les séquences des autres gènes.

Meilleur hit réciproque La méthode d'assignation la plus simple est celle du meilleur hit réciproque, ou meilleur hit bidirectionel (en anglais : "Reciprocal Best Hit", RBH, ou "Bidirectional Best Hit", BBH). Elle n'assigne que des relations d'orthologie 1-1. La méthode est assez intuitive, le gène x_1 du génome G_1 et le gène x_2 du génome G_2 sont meilleurs hits réciproques si x_2 est le gène le plus similaire à x_1 parmi tous les gènes de G_2, et x_1 est le gène le plus similaire à x_2 parmi tous les gènes de G_1. En pratique, on effectue tous les alignements deux à deux des gènes de G_1 et de G_2 et on ne retient pour chaque gène que son meilleur alignement. On identifie ensuite les relations réciproques.

Cette méthode a l'avantage d'être simple et rapide. Cependant, elle possède plusieurs défauts. Le premier est qu'elle n'est fiable que si on possède l'ensemble complet des gènes de chaque génome. Le second est qu'elle ne s'applique qu'à deux génomes simultanément. Le troisième est qu'elle ne peut identifier que des relations d'orthologie 1-1, ainsi elle n'est pas très fiable lorsque les relations d'orthologie sont n-m et elle est très sensible aux grandes familles multigéniques. Reprenons l'exemple de l'insuline entre la souris et le rat, si les copies INS1 et INS2 sont peu différentes, il se peut que le meilleur hit de INS1 souris soit INS1 rat mais que le meilleur hit de ce dernier soit INS2 souris. Dans ce cas, la méthode ne détectera pas d'orthologue du gène INS1 souris chez le rat. Une autre source d'erreurs est la perte différentielle de copies dans les deux espèces. Par exemple, si la souris a perdu sa copie INS1 et le rat sa copie INS2, alors les copies restantes seront assignées orthologues alors qu'elles sont paralogues.

Assignation d'orthologie n-m A partir des orthologues 1-1 identifiés par RBH, on peut également essayer de compléter ces paires de gènes pour former des relations n-m lorsque c'est possible. C'est ce que fait la méthode Inparanoid (O'Brien *et al.*, 2005), en cherchant des gènes au sein d'une espèce qui sont plus similaires entre eux qu'avec n'importe quel gène de l'autre espèce. Alors que cette méthode est limitée à des comparaisons deux à deux de génomes, la méthode COG (Tatusov *et al.*, 1997) ("Clusters of Orthologous Groups") s'applique à plus de deux génomes. Elle regroupe d'abord les in-paralogues potentiels (gènes plus similaires au sein d'une espèce qu'avec les gènes des autres espèces), puis applique un critère similaire au RBH entre groupes d'in-paralogues de différentes espèces.

Approche phylogénétique L'approche phylogénétique semble la plus fiable puisque son objectif est d'inférer l'histoire évolutive des familles de gènes homologues. Elle permet d'assigner des relations d'orthologie n-m et de paralogie. Elle est basée sur la réconciliation de l'arbre phylogénétique des gènes avec celui des espèces en inférant sur l'arbre des gènes les évènements de spéciation, de duplication et de perte de gènes. Cette méthode s'applique sur un grand nombre d'espèces simultanément. En comparant tous les gènes des différents génomes deux à deux, on les regroupe en familles de gènes en fonction de leur score de similarité. On effectue ensuite un alignement multiple des gènes d'une même famille et on calcule un arbre phylogénétique. Parmi tous les noeuds de l'arbre, on veut distinguer les noeuds correspondant à un évènement de spéciation de ceux correspondant à un évènement de duplication. On connaît la topologie de l'arbre des espèces, cela restreint donc la façon d'assigner les noeuds de spéciation. On infère alors les évènements de duplication en utilisant le principe de parcimonie : on veut minimiser le nombre d'évènements de duplication et de pertes de gènes (Page et Charleston, 1997; Dufayard *et al.*, 2005).

L'inconvénient principal de cette méthode est qu'elle est très lourde à mettre en place et son exécution peut être assez longue en fonction du nombre d'espèces comparées. Elle est également très sensible aux artéfacts de reconstruction phylogénétique telles que le phénomène d'attraction des longues branches. Enfin, elle suppose que les gènes sont uniquement transmis verticalement, elle est donc peu fiable dans le cas de transferts horizontaux.

Ces méthodes ne permettent pas d'identifier les orthologues de position parmi les orthologues n-m. Cependant, elles minimisent le risque de mauvaises assignations d'orthologues 1-1, dues à des orthologies n-m. Au contraire, la méthode RBH identifiera souvent une paire de gènes parmi les orthologues n-m, les plus similaires, or il ne s'agit pas forcément de gènes orthologues de position. Ainsi, une étude menée sur des génomes bactériens a montré que,

dans le cas d'orthologues n-m, seulement 60 % des paires de gènes les plus similaires sont également des orthologues de position (Notebaart *et al.*, 2005).

Dans les méthodes de comparaison de l'ordre des gènes, les orthologues n-m seront le plus souvent éliminés, ou bien on cherchera les orthologues de position grâce à l'information contextuelle des autres gènes.

2.2.3 Identification des segments conservés

Nous disposons des paires de gènes orthologues, avec leur chromosome, leur position sur le chromosome et leur orientation, et ce sur les deux génomes. Et nous recherchons des paires de régions, une sur chaque génome, qui contiennent les mêmes ensembles de gènes dans le même ordre et la même orientation. Cela paraît trivial, et en effet, si on ne s'autorise aucune flexibilité, chaque fois que deux gènes sont consécutifs sur un génome et consécutifs avec des orientations relatives concordantes sur l'autre, on considère qu'ils appartiennent à une même région conservée. Et on procède ainsi en parcourant un génome.

Mais le problème se complique si on veut introduire un peu de flexibilité, par exemple autoriser quelques gènes à ne pas respecter l'ordre et l'orientation de la région conservée dans laquelle ils se trouvent. Cette flexibilité est parfois nécessaire lorsque les données en entrée sont bruitées et contiennent des erreurs, telles que des erreurs d'assignation d'orthologie ou des erreurs de positionnement (par exemple dues à des erreurs d'assemblage). Dans d'autres cas, cette flexibilité est voulue lorsque l'objectif est d'identifier les segments d'une certaine taille et de négliger des petits réarrangements, par exemple lorsque les espèces comparées sont très éloignées et ont subi beaucoup de réarrangements.

Il faut noter qu'il existe des méthodes dont le but est d'identifier des régions dont le contenu en gènes est similaire mais l'ordre des gènes au sein des régions n'est pas forcément conservé entre les deux génomes. Ces méthodes, appelées "Gene teams" ou "max-gap clusters", sont moins pertinentes du point de vue d'une étude des réarrangements puisque de grosses différences d'ordre peuvent exister au sein des groupes identifiés. Elles sont surtout utilisées dans le but d'identifier des gènes reliés fonctionnellement. Elles sont bien décrites et revues dans les références (Bergeron *et al.*, 2002; Hoberman *et al.*, 2005).

Nous étudierons ici les méthodes qui recherchent des régions où non seulement le contenu en gènes est conservé mais également l'ordre et l'orientation des gènes. Nous appellerons ces régions des blocs de synténie.

Nous ouvrons ici une parenthèse sur l'usage du terme "synténie". L'expression "blocs de synténie" est utilisée actuellement pour désigner les régions conservées (notamment en ordre et en orientation). Mais le terme "syntenie" a été détourné de son sens premier (lire (Passarge *et al.*, 1999)). Il a été introduit en 1971 par John H. Renwick pour désigner le fait que deux loci sont sur un même chromosome. En effet, les méthodes de cartographie de l'époque pouvaient permettre de confirmer la localisation de deux loci sur un même chromosome sans connaître leur position ni la distance qui les séparait. Ainsi, le terme "syntenie" ne s'appliquait, à l'origine, qu'à un seul génome. On a alors utilisé le terme "synténie conservée", pour désigner le fait pour deux marqueurs ou plus d'être sur un même chromosome dans plusieurs génomes comparés, indépendamment de leur ordre et orientation sur ces chromosomes (Andersson *et al.*, 1996; Ehrlich *et al.*, 1997). D'autres expressions telles que "segment conservé", et plus précisément "segment d'ordre conservé", ont alors été utilisées pour désigner la contrainte supplémentaire d'ordre. Cependant, il semble qu'aujourd'hui, le terme le plus utilisé soit "blocs de synténie" (par exemple par (Pevzner et Tesler, 2003a; Kent *et al.*, 2003; Choi *et al.*, 2007;

Sinha et Meller, 2007)).

a. Un grand nombre de méthodes

Les méthodes de construction de blocs de synténie basées sur l'ordre des gènes orthologues sont très nombreuses. Ainsi, il n'existe pas UNE méthode de détection des blocs de synténie générique qui serait utilisée par la majorité de la communauté comme il existe un algorithme d'alignement tel que Blast. Au contraire, il semble que chacun développe sa méthode selon ses besoins. Cela provient peut-être du fait que ce problème semble facile de prime abord. Ainsi, un grand nombre d'entres elles sont des méthodes "ad hoc", décrites comme des suites d'opérations, elles ne définissent pas formellement les objets utilisés et recherchés et ne sont pas implémentées dans un programme disponible. La deuxième raison est que les données en entrée peuvent être de nature différente nécessitant des méthodes différentes. Par exemple, les génomes bactériens et eucaryotes ont des structures très différentes (nombre de chromosomes, organisation des gènes, taux de duplications et de transferts horizontaux). Selon la nature des données, si elles proviennent de cartes génétiques ou bien d'annotation de génomes complets, certaines informations sont différentes et avec des niveaux de bruit différents.

Sans décrire chacune d'entre elles en détail, nous décrivons les grandes lignes des approches, en nous intéressant plus particulièrement aux propriétés biologiques des objets produits.

b. Une approche cummune, avec des variantes...

Avant de décrire les méthodes, il est nécessaire de décrire plus précisément les objets traités. Nous nous placerons dans le cas général de la comparaison de deux génomes, G_1 et G_2, avec des paires de gènes orthologues sans duplication ni chevauchements.

Ancre Nous appellerons une paire de gènes orthologues, une **ancre**. Une ancre x est définie par sa localisation sur chacun des deux génomes et une orientation. Plus précisément $x = (c_1, x_1, c_2, x_2, \sigma_x)$, avec x_1 (resp. x_2) la position du gène x sur le chromosome c_1 (resp. c_2) de G_1 (resp. G_2). L'orientation de l'ancre σ_x prend la valeur $+1$ si les deux gènes ont la même orientation sur leur chromosome respectif, -1 sinon.

On peut représenter les ancres dans l'espace à deux dimensions définis par les deux génomes (c'est la représentation en dotplot), ou bien de façon linéaire, en plaçant un génome en dessous de l'autre (voir Figure 2.6).

Colinéarité On dit que deux ancres x et y, appartenant à une même paire de chromosomes et telles que $x_1 < y_1$, sont **colinéaires**, si $x_2 < y_2$ et $\sigma_x = \sigma_y = +1$ ou bien $x_2 > y_2$ et $\sigma_x = \sigma_y = -1$. Dans l'exemple de la Figure 2.6, seules les paires d'ancres (a,b), (a,c), et (d,e) sont colinéaires.

Schéma de base La plupart des méthodes sont basées sur le schéma suivant : deux ancres appartiennent à un même groupe si elles sont colinéaires et distantes (en pb ou en nombre de gènes) de moins d'un certain paramètre de gap, qu'on appellera G. On ne retient ensuite que les groupes dont la taille est supérieure à un certain paramètre de taille S (en pb ou en nombre de gènes).

On peut citer les méthodes suivantes qui suivent ce schéma, avec quelques variantes : SynBrowse (Pan *et al.*, 2005), Cinteny (Sinha et Meller, 2007), SyntQL (Zdobnov *et al.*, 2002; Bourque *et al.*, 2005), GeneSyn (Pavesi *et al.*, 2004), FISH (Calabrese *et al.*, 2003),

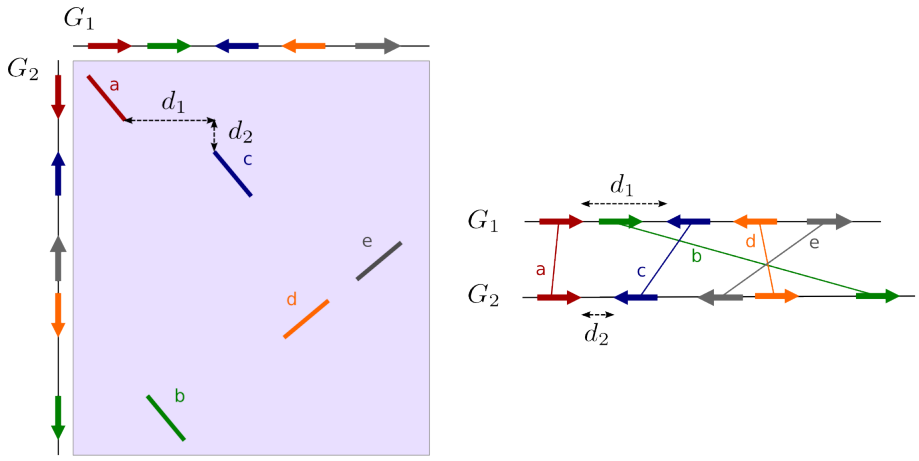


FIG. 2.6: Exemple de deux génomes G_1 et G_2, ayant chacun un seul chromosome, avec 5 ancres a, b, c, d et e. A gauche, les ancres sont représentées sous la forme d'un dotplot, à droite sous la forme linéaire. Les distances d_1 et d_2, respectivement dans les génomes G_1 et G_2, entre les ancres a et c, sont montrées sur les deux types de représentation.

DAGchainer (Haas *et al.*, 2004), DiagHunter (Cannon *et al.*, 2003), ADHoRE (Vandepoele *et al.*, 2002), LineUp (Hampson *et al.*, 2003), AutoGraph (Derrien *et al.*, 2007).

Les différences entre les méthodes résident principalement dans les conditions sur les ancres en entrée (chevauchantes ou non, avec duplications ou non), la définition de la distance entre deux ancres, les paramètres utilisés (en paires de bases ou en nombre de gènes), les critères pour retenir un bloc (taille, ou critère statistique), le relâchement de la condition de colinéarité, etc. En fonction de ces différents choix, les propriétés des blocs de synténie obtenus pourront être assez différentes.

Propriétés des ancres en entrée La majorité des algorithmes requièrent que les relations d'orthologies soient 1-1 et que les gènes ne se chevauchent sur aucun des deux génomes (Hampson *et al.*, 2003; Pan *et al.*, 2005; Derrien *et al.*, 2007). Lorsque les orthologies n-m sont autorisées, il existe souvent une limite sur les nombres n et m, ou bien les données subissent une étape de prétraitement pour éliminer les dupliqués (Calabrese *et al.*, 2003; Sinha et Meller, 2007).

LineUp et AutoGraph n'utilisent pas l'information d'orientation des ancres, ils sont en fait spécialement adaptés aux données de cartes génétiques qui produisent des localisations de gènes non orientées (Hampson *et al.*, 2003; Derrien *et al.*, 2007). LineUp gère également l'incertitude d'ordre sur les marqueurs proches qui existe avec ce type de données.

GeneSyn (Pavesi *et al.*, 2004) s'applique à plus de deux génomes simultanément.

Distance entre deux ancres Le plus souvent, la distance rend compte de l'espacement des ancres en 2D, c'est-à-dire qu'elle prend en compte la distance sur le premier génome, qu'on appelle d_1, et sur le second, d_2. La distance $d_1(x, y)$ (resp. $d_2(x, y)$) entre deux ancres x et y sur le génome G_1 (resp. G_2) est définie par $d_1(x, y) = |x_1 - y_1|$ (resp. $d_2(x, y) = |x_2 - y_2|$) (voir Figure 2.6). Cette distance peut être exprimée en paires de bases (x_1 est la position physique du gène sur le chromosome), ou bien en nombre de gèness (x_1 est le rang du gène sur le chromosome).

La manière dont ces deux distances sont combinées pour donner une distance globale est variable. La distance la plus souvent utilisée est la distance de Manhattan qui est la somme de d_1 et d_2 ($d = d_1 + d_2$). L'algorithme FISH et plusieurs algorithmes d'alignement de génomes complets l'utilisent (Calabrese *et al.*, 2003; Pevzner et Tesler, 2003a; Hubbard *et al.*, 2007). On peut également ne s'intéresser qu'à la distance maximale $d = max(d_1, d_2)$ (Zdobnov *et al.*, 2002; Pan *et al.*, 2005).

Certaines distances prennent en compte la différence de distance sur les deux génomes, c'est-à-dire la différence entre d_1 et d_2. Ainsi, la distance entre deux ancres sera d'autant plus petite que deux ancres sont situées sur la même diagonale ($d_1 = d_2$). Par exemple, l'algorithme ADHoRE (Vandepoele *et al.*, 2002) utilise la distance suivante : $d = 2 \times max(d_1, d_2) - min(d_1, d_2)$.

Critère d'évaluation d'un bloc Le critère d'évaluation d'un groupe d'ancres est assez différent entre les méthodes. Le critère le plus souvent utilisé est celui de la taille du groupe en nombre d'ancres qu'il contient (Calabrese *et al.*, 2003; Hampson *et al.*, 2003; Vandepoele *et al.*, 2002; Zdobnov *et al.*, 2002). Soit l'objectif est d'éliminer les petits réarrangements, soit, lorsqu'on veut éliminer le bruit et les erreurs contenues dans les données, on part du principe que les erreurs, si elles existent, sont isolées. Ainsi, plus un groupe contient d'ancres, plus il est fiable. Il reste alors à définir le nombre minimal d'ancres qu'un groupe doit contenir pour être considéré comme fiable, soit le paramètre S. C'est souvent un paramètre que peut faire varier l'utilisateur, ou bien il peut être estimé statistiquement. Par exemple, plusieurs méthodes effectuent des simulations où les génomes sont randomisés (les gèness sont redistribués aléatoirement sur le génome) afin d'estimer la taille des groupes attendue par hasard (Hampson *et al.*, 2003; Vandepoele *et al.*, 2002). FISH estime la probabilité d'un groupe à k ancres de manière analytique, sous un modèle nul où les ancres sont réparties aléatoirement dans une matrice $n \times n$ où n est le nombre d'ancres (Calabrese *et al.*, 2003).

Ce critère de taille peut également être exprimé en paires de bases, il représente alors la distance couverte par le groupe d'ancres sur les génomes (Sinha et Meller, 2007).

DAGchainer définit un critère plus complexe, qui prend en compte le degré de similarité des ancres, la distance entre les ancres et la différence de distance dans les deux génomes (Haas *et al.*, 2004).

Le degré de désordre au sein du bloc peut également être un critère. Par exemple, ADHoRE (Vandepoele *et al.*, 2002) teste si les ancres au sein d'un groupe forment une droite dans le dotplot. Il effectue une régression linéaire dans chaque groupe avec les coordonnées des ancres dans le dotplot et évalue les groupes en fonction du degré d'ajustement des points à la droite de régression.

Algorithmes Du point de vue algorithmique, il existe également des différences entre les méthodes. Cependant, les algorithmes ne sont pas toujours décrits précisément. De nombreuses méthodes utilisent des graphes pour représenter les relations de distance entre les ancres sur les différents génomes (Pavesi *et al.*, 2004; Haas *et al.*, 2004; Calabrese *et al.*, 2003). La programmation dynamique et des fonctions récursives sont utilisées pour chercher les meilleures chaînes d'ancres dans une matrice ou dotplot (Calabrese *et al.*, 2003; Haas *et al.*, 2004; Cannon *et al.*, 2003). Enfin, d'autres procèdent de façon itérative ou récursive, en modifiant les valeurs des paramètres à chaque étape, augmentant ainsi la taille des blocs petit à petit (Vandepoele *et al.*, 2002).

c. Les conflits et l'approche de Sankoff

La flexibilité de ces méthodes a une conséquence sur les blocs de synténie obtenus, qui est peu prise en compte ou commentée. Autoriser de chaîner des ancres qui ne sont pas consécutives entraîne qu'une ancre peut être chaînée à plus d'une ancre : une ancre peut avoir plusieurs successeurs qui satisfont les critères de chaînage (distance). Ainsi un bloc de synténie ne peut plus être représenté par une chaîne unique d'ancres toutes colinéaires les unes avec les autres (voir Figure 2.7). Il peut également arriver que deux blocs se chevauchent même si les ancres sont non-chevauchantes et de type 1-1, par exemple si une ancre "sautée" dans un bloc appartient à un autre bloc. Les chevauchements et les conflits de chaînage ne sont pas toujours aisés à manipuler et à interpréter dans la suite des analyses. Or, les méthodes présentées ici ne mentionnent pas ces possibilités et ne semblent pas les gérer, ou bien, dans certains cas, des choix sont effectués pour éviter conflits et chevauchements, mais les critères utilisés ne sont pas justifiés, ni commentés.

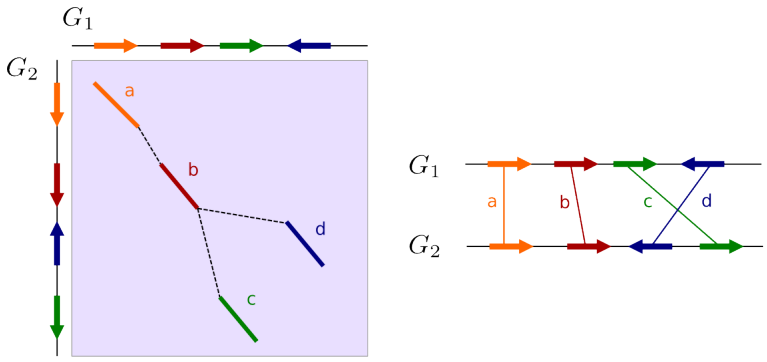


FIG. 2.7: Exemple schématique de conflit de chaînage au sein d'un bloc de synténie. Le bloc de synténie est construit de la façon suivante : deux ancres colinéaires appartiennent à un même bloc si elles sont distantes de moins de G ancres. Dans cet exemple, les ancres a, b, c et d appartiennent au même bloc de synténie : a et b sont colinéaires et $d(a,b) < G$, b et c sont colinéaires et $d(b,c) < G$ et b et d sont colinéaires et $d(b,d) < G$. On observe que b possède deux successeurs possibles c et d. On peut donc avoir deux chaînages des ancres possibles au sein de ce bloc. On observe également que dans ce cas, toutes les ancres ne sont pas colinéaires deux à deux : c et d ne sont pas colinéaires.

Sankoff et collègues se sont attachés à proposer une définition formelle de blocs de synténie sans conflit (Zheng et Sankoff, 2006; Choi *et al.*, 2007). Leur approche est assez différente des méthodes décrites précédemment. Le principe est d'éliminer des ancres du jeu initial afin d'obtenir des blocs de synténie sans aucune flexibilité, qu'ils appellent *chaînes pures* (*"pure strip"*), c'est-à-dire où les ancres sont colinéaires et consécutives dans les deux génomes. Le critère pour éliminer les ancres est global : il s'agit de minimiser le nombre d'ancres à éliminer pour n'obtenir que des chaînes pures. L'intérêt de cette approche est que les objets recherchés ainsi que le problème sont formellement définis. Cependant ce problème est NP-difficile, et sa résolution sur des jeux de données réels n'est pas réalisable. Les auteurs ont alors proposé plusieurs heuristiques qui permettent de limiter l'espace de recherche. Elles sont basées essentiellement sur l'ajout de contraintes, comme par exemple une contrainte de distance entre les ancres controlée par un paramètre de gap, similaire au paramètre G utilisé par les méthodes décrites précédemment.

d. Tableau récapitulatif

Méthode (ref)	Type d'ancres	critère(s) pour former les blocs (paramètre G)	critère(s) pour évaluer les blocs (paramètre S)	conflits	
				dans un bloc	entre blocs
Cinteny Sinha (2007)	n-m (élimination de copies)	colinéarité, distance en pb	S_1 en pb et S_2 en nb d'ancres	oui	non (comment ?)
DAGchainer Haas (2004)	n-m associées à une E-value	colinéarité, G en pb	S en nb d'ancres et score calculé en fonction de 5 paramètres (E-value, distance, diagonale)	non	oui
SynBrowse Pan (2005)	1-1	colinéarité, G en pb	S en nb d'ancres	oui	oui
DiagHunter Cannon (2003)	n-m	colinéarité, G en ? (et distance à la diagonale)	S en nombre d'ancre et score du bloc (def. ? ?)	non	oui
LineUp Hampson (2003)	1-1 non orientées	G en nb d'ancres, nb d'ancres non colinéaires limité par un paramètre	S_1 en pb (ou cM) et S_2 en nb d'ancres, évalués par permutations	oui	oui
AutoGraph Derrien (2007)	1-1 non orientées, non chevauchantes	colinéarité, G en nb d'ancres		non *	non *
Sankoff Choi (2007)	1-1	colinéarité, G en nb d'ancres	$S = 2$ ancres, minimiser le nb total d'ancres éliminées	non	non
GenSyn Pavesi (2004)	n-m (limités)	colinéarité, G en nb d'ancres	S en nb d'ancres, q quorum	non	oui
SyntQL Zdobnov (2002)	n-m non orientées, avec E-value	colinéarité, $G = 5$ ancres	$S = 2$ ancres	? ?	oui
FISH Calabrese (2003)	n-m (limités et dup. en tandem regroupées), non chevauchantes	G en nb d'ancres	S en nb. d'ancres, fixé statistiquement (p-value)	non	oui
ADHoRE Vandepoele (2002)	n-m (dup. en tandem regroupées)	colinéarité, G en nb. d'ancres (et distance à la diagonale)	régression linéaire, S en nb d'ancres évalué par permutations	oui	oui
GRIMM Pevzner (2003a)	1-1 non chevauchantes	G en pb	S en pb	oui	oui
CHAINNET Kent (2003)	? ?	colinéarité	score (def. ? ?)	non	non (génome de réf.)
CP Couronne (2003)	? ?	G en pb	S (en ? ?) et score d'alignement global	non	oui ?
MAUVE Darling (2004)	1-1	colinéarité, G=0 ancres (ancres successives)	S en pb	non	non

TAB. 2.1: Tableau récapitulatif des principaux aspects des méthodes de construction de blocs de synténie. En bas du tableau, sont ajoutées quatres algorithmes similaires utilisés dans le cadre de l'alignement de génomes complets (voir Section 2.3). Nous appelons G le paramètre qui désigne la distance maximale permise entre deux ancres pour être groupées, et S la taille minimale d'un bloc pour être retenu. Les points d'interrogations (? ?) représentent des informations manquantes. Le symbole * signifie que la résolution des conflits est arbitraire et asymétrique.

2.3 Alignement de génomes complets

Lorsqu'on compare des cartes de marqueurs, la résolution des segments conservés dépend fortement de la densité en marqueurs. Or, les gènes orthologues, entre l'homme et la souris par exemple, couvrent moins de 30 % du génome humain. Avec le séquençage des génomes complets on dispose de toute l'information et on espère atteindre la résolution maximale : identifier les réarrangements au nucléotide près. Dans cette partie, nous analysons et comparons différentes méthodes d'alignement de génomes complets. Elle est en grande partie reprise de l'article de revue que nous avons publié (Lemaitre et Sagot, 2008).

L'alignement de génomes complets, et en particulier pour les génomes de vertébrés, pose plusieurs problèmes qui nous empêchent d'utiliser directement les algorithmes classiques d'alignement de séquences (alignement local ou global).

Tout d'abord, la taille des séquences concernées rend impossible l'utilisation d'algorithmes exacts. En effet, le génome de l'homme compte presque trois milliards de nucléotides. Ainsi, souvent ce sont des heuristiques qui sont utilisées pour traiter de telles séquences.

Deuxièmement, les génomes sont des séquences très hétérogènes en terme de conservation de séquence. Par exemple, l'homme et la souris ont divergé il y a 75 millions d'années, et seulement 40 % de leurs génomes sont alignables. Seulement 5 % du génome humain est sous pression de sélection (estimation d'après des alignements), les régions codantes représentent seulement 1.5 % du génome (Waterston *et al.*, 2002). Alors que beaucoup d'algorithmes d'alignement ont été conçus pour aligner des séquences codantes, aligner des séquences intergéniques est plus difficile, car les séquences sont moins conservées, elles peuvent contenir de gros indels et des segments qui ne présentent aucune similarité avec leurs orthologues. Les algorithmes utilisés doivent donc être capables de "sauter" ces régions sans trop de coût.

Enfin, les génomes ont peut-être subit des réarrangements et des duplications. Or les algorithmes d'alignement global *classiques* requièrent que l'ordre et l'orientation des nucléotides homologues soient les mêmes dans les deux séquences et un nucléotide ne peut être aligné qu'une seule fois. Il existe des algorithmes d'alignement global qui permettent de détecter des réarrangements et/ou des duplications dans l'alignement (Brudno *et al.*, 2003b; Alves *et al.*, 2005). Cependant, ils ne peuvent être appliqués sur des génomes entiers de vertébrés car, en général, ils ne permettent que des petits réarrangements locaux et des duplications en tandem. De plus, leur complexité algorithmique peut être très importante et les rend inutilisables sur de très longues séquences telles que les génomes entiers.

L'alignement global de génomes ayant subit des réarrangements n'est donc pas adapté à ce problème. Mais on peut espérer que, à l'intérieur des régions conservées, un alignement global est possible. La stratégie est donc de détecter les régions conservées, non réarrangées, puis d'aligner globalement chacune d'entre elles indépendamment. La détection des régions conservées se fait également par alignement, mais cette fois avec un algorithme d'alignement local qui identifie des paires de petites séquences très similaires, appelées ancres. La difficulté ensuite est de distinguer parmi tous les résultats, les paires constituant des séquences réellement orthologues, des autres non orthologues qui sont similaires par hasard, ou par duplication (appelés faux positifs). Une étape intermédiaire de filtrage est donc nécessaire. Pour filtrer les faux positifs, on utilise l'information contextuelle, et on suppose que les faux positifs sont représentées par des ancres isolées. Cette étape est très similaire aux algorithmes de construction de blocs de synténie décrits à la section précédente. L'objectif est le même, mais les données en entrée sont différentes, il s'agit ici d'alignements locaux et non de gènes orthologues. Par conséquent, les ancres sont plus nombreuses et beaucoup plus bruitées, ce qui impose des méthodes différentes (Sankoff et Nadeau, 2003).

Ainsi, l'alignement de génomes complets suit une stratégie en 3 étapes : 1. détection des ancres, 2. filtrage des ancres, 3. alignement des régions homologues détectées. La première étape est la plus similaire entre les méthodes existantes, alors que la troisième dépend de la finalité de chaque méthode : soit obtenir des alignements de séquences, soit étudier les réarrangements.

Nous avons étudié principalement 4 méthodes d'alignement de génomes complets : GRIMM (Pevzner et Tesler, 2003a), CHAINNET (Kent *et al.*, 2003), la méthode décrite par Couronne et collaborateurs (Couronne *et al.*, 2003), que nous appellerons CP, et enfin MAUVE (Darling *et al.*, 2004). Ces méthodes ont été publiées quasiment en même temps, lorsque le génome de la souris a été rendu public. Il était désormais possible de comparer à l'echelle du nucléotide les génomes de l'homme et de la souris.

2.3.1 Détection des ancres

L'objectif de cette étape est d'identifier des similarités locales entre deux génomes, en un temps raisonnable étant donné la taille des génomes comparés. Cette étape est commune à toutes les méthodes. Cependant, elles utilisent toutes un algorithme différent. Les ancres peuvent représenter des mots exacts en commun, ou presque exacts, ou même des alignements locaux avec ou sans indels. Elles sont de taille variable et doivent parfois respecter certaines conditions d'unicité et/ou de non chevauchement.

CHAINNET utilise un algorithme d'alignement local, Blastz (Schwartz *et al.*, 2003) mais ne garde comme ancres que les sous-parties sans indels des alignements obtenus. GRIMM utilise les alignements locaux avec indels produits par l'algorithme Pattern Hunter (Ma *et al.*, 2002), mais il ne retient que celles qui ne se chevauchent pas (sur aucun des deux génomes) et qui sont uniques (pas de duplications possibles).

Ces deux algorithmes d'alignement local sont très similaires, ils ont été développés à la même époque, dans le but d'aligner des séquences peu conservées telles que les séquences non codantes. Ils sont basés sur le même principe que l'algorithme Blast (Altschul *et al.*, 1990, 1997) : de type "ancrer-et-étendre" ("seed-and-extend" en anglais). Cette stratégie est utilisée pour accélérer la recherche de similarités locales. On part du principe que deux séquences similaires ont de fortes chances de contenir des mots exacts en commun. Ainsi pour limiter l'espace de recherche, on commence par ne chercher que des mots exacts en commun, appelées graines (étape d'ancrage). Les graines servent ensuite comme points d'ancrage des alignements (étape d'extension). On étend d'abord la graîne de part et d'autre sans autoriser d'indels jusqu'à ce que le score chute d'un certain seuil. Puis on étend encore mais en autorisant les indels avec un algorithme de type programmation dynamique, et on ne garde que les alignements ayant obtenu un score suffisant.

Blastz et Pattern Hunter diffèrent de Blast principalement par le type de graines utilisées. Au lieu de chercher des mots exacts, ils cherchent des mots de taille l dans lesquels certaines positions (fixées) doivent être exactes : ce sont des graines espacées. De plus, Blastz autorise une transition[1] à la place d'un appariement à une position de la graine. Ce type de graine a été montré très sensible (Sun et Buhler, 2006), et la même étude a montré que Blastz était l'un des algorithmes d'alignement les plus sensibles pour aligner des séquences non codantes.

[1]Les mutations de nucléotides sont classées en deux catégories : les transitions et les transversions. Une transition est une mutation d'une purine à une purine ou d'une pyrimidine en une pyrimidine (A-G ou C-T), alors qu'une transversion est une mutation qui entraîne un changement du type de base : purine en pyrimidine ou vice versa. Par abus de langage, un mésappariement dans un alignement de type A-G ou C-T est appelé transition. Les substitutions de types transition sont plus fréquentes que les transversions. C'est pourquoi l'algorithme permet une transition à la place d'un appariement.

CP utilise l'algorithme d'alignement local Blat (Kent, 2002) pour détecter les ancres. Cet algorithme fait partie des algorithmes "seed-and-extend". Les graines utilisées sont des mots exacts mais il y a une étape de sélection des graînes avant celle d'extension : seuls les groupes de graines proches et sur une même diagonale seront étendues. Blat n'a pas été conçu pour l'alignement inter-espèces mais il présente l'avantage d'être très rapide.

Ces trois méthodes d'ancrage possèdent plusieurs paramètres, comme la taille des graines, les matrices de substitution, la pénalisation des indels, et les seuils d'extension ou finaux. Le choix des valeurs à donner aux paramètres a une influence sur la sensibilité et la spécificité de l'algorithme. Ce choix dépend également des espèces considérées. Mais il est très peu discuté et est souvent arbitraire.

L'algorithme Mauve est différent des trois autres méthodes pour cette partie d'ancrage. Alors que l'objectif des méthodes précédentes est d'être très sensible, Mauve est très stringent et spécifique dans sa méthode d'ancrage. Mauve cherche des mots exacts et uniques (qui n'apparaissent qu'une seule fois dans chaque génome) : des MUMs (Maximum Unique Match).

La taille des ancres trouvées par ces diverses méthodes varie de quelques paires de base (Mauve), à de longs segments de 500 paires de bases (Grimm), avec une taille moyenne intermédiaire de 30 paires de bases (Chainnet).

2.3.2 Filtrage

Le but de cette étape est d'éliminer les faux positifs de l'ensemble d'ancres détectées et/ou les petits réarrangements locaux. L'idée principale est d'utiliser l'information contextuelle et de supposer que les erreurs, s'il y en a, seront isolées. Ainsi deux ancres qui se trouvent proches l'une de l'autre et à une même distance dans les deux génomes ont moins de chance d'être des erreurs qu'une ancre isolée.

On retrouve dans cette étape des idées et des approches similaires aux méthodes de construction de blocs de synténie basées sur la comparaison de l'ordre des gènes. On peut distinguer deux types d'approches : la première consiste à former des chaînes d'ancres, la deuxième à les regrouper ("clustering"). Le seul critère utilisé par cette dernière est la distance entre deux ancres, alors que la première approche requiert que l'ordre et l'orientation des ancres soit conservé au sein d'une chaîne. Grimm et CP ont adopté l'approche de regroupement, alors que Mauve et Chainnet celle du chaînage.

La distance utilisée par Grimm est la distance de Manhattan exprimée en paires de bases (voir Section 2.2.3). Deux ancres appartiennent au même cluster si leur distance est inférieure à un certain seuil G (pour gap), même si elles ne sont pas colinéaires. L'algorithme pour identifier les clusters construit un graphe. Chaque noeud correspond à une ancre, deux noeuds sont reliés par une arête si leur distance est inférieure au seuil G. Les clusters correspondent aux composantes connexes du graphe. Seuls les clusters qui couvrent une distance suffisante sur le génome (paramètre S) sont retenus. Ainsi, au sein d'un cluster, les ancres ne forment pas forcément une chaîne et peuvent avoir des orientations différentes.

La stratégie de CP semble similaire (le papier n'est pas très clair sur ce point) puisque les ancres sont regroupées en fonction de leur distance (mais celle-ci n'est pas définie). Cependant, même si l'ordre et l'orientation des ancres ne sont pas pris en compte à cette étape, ces critères sont indirectement pris en compte à l'étape suivante qui consiste à aligner globalement les séquences d'un cluster. Ainsi des clusters qui contiennent trop de désordre obtiendront un score d'alignement global trop faible pour être retenus.

Les deux autres méthodes, Chainnet et Mauve, prennent en compte directement l'ordre et l'orientation des ancres, puisqu'ils cherchent à faire des chaînes où toutes les ancres sont

colinéaires les unes avec les autres. MAUVE n'utilise pas l'information de distance physique entre les ancres, mais les ancres chaînées doivent être colinéaires et consécutives dans les deux génomes, n'autorisant aucun désordre. Seules les chaînes dont les ancres couvrent une longueur suffisante sont retenues. On obtient ici des chaînes sans conflit : ordre strict et pas de chevauchement. Cette méthode est très stringente, puisqu'elle ne présente aucune flexibilité. Elle requiert donc qu'il y ait très peu d'erreurs et de bruit dans l'ensemble d'ancres. On comprend mieux l'utilisation des MUMs dans la première étape.

CHAINNET chaîne les ancres en prenant en compte l'ordre, l'orientation et la distance entre les ancres. L'algorithme utilise une structure algorithmique complexe, une structure d'arbre qui permet de partitionner les données dans un espace à k dimensions. Ici, elle est utilisée pour trouver tous les points qui sont dans un hyper-rectangle donné, afin de chercher les ancres colinéaires qui peuvent être chaînées (Zhang *et al.*, 1994). Contrairement à MAUVE, CHAINNET produit dans cette étape des chaînes qui peuvent se chevaucher ou qui peuvent "sauter" certaines ancres qui ne sont pas dans le bon ordre ou la bonne orientation. En fait, CHAINNET produit toutes les chaînes possibles et assigne un score à chacune en fonction du nombre d'ancres, de la distance couverte, des distances inter-ancres, etc (la fonction de score n'est pas définie précisément, ni comment fixer les poids relatifs de ces différents critères). Nous verrons dans la troisième étape comment CHAINNET traite ces chaînes.

2.3.3 Alignement final, extension ou récursivité

La dernière étape est la plus différente entre les méthodes comparées. La raison principale est que ces méthodes ont en fait des objectifs différents. GRIMM, par exemple, a été conçu pour l'analyse des réarrangements qui séparent deux génomes, c'est-à-dire la reconstruction de l'histoire des réarrangements en utilisant des algorithmes de tris de permutations. Ainsi, l'arrangement des différentes régions conservées dans les deux génomes est une sortie suffisante de cet algorithme. Comment les séquences s'alignent à l'intérieur des régions conservées n'est pas utile pour l'analyse des scénarios de réarrangements. De plus, les régions conservées produites doivent être suffisamment longues et peuvent contenir (masquer) des petits réarrangements locaux, appelés micro-réarrangements. C'est pourquoi l'arrangement des ancres dans les clusters n'est pas contraint. Dans cette troisième étape, GRIMM chaîne les clusters qui se trouvent colinéaires et consécutifs dans les deux génomes, quelle que soit la distance physique qui les sépare. Il forme ainsi de long blocs de synténie. Cependant, il n'est pas très clair comment il attribue une orientation aux clusters, étant donné que les ancres que ces derniers contiennent peuvent avoir des orientations différentes.

Par contre, l'objectif de CHAINNET et de CP est de produire des alignements entre les deux génomes et que tous les nucléotides de régions potentiellement homologues soient alignés. Ces alignements sont alors utilisés pour analyser les processus évolutifs à grande comme à petite échelle. Cependant, ces deux méthodes ont leur troisième étape très différente. CP ne cherche pas à étendre les clusters obtenus à la deuxième étape, mais il aligne simplement chaque cluster avec un algorithme d'alignement global, AVID (Bray *et al.*, 2003). Seuls les alignements obtenant un score suffisant (au dessus d'un certain seuil) sont retenus. Cette stratégie aboutit à de nombreux petits alignements.

L'idée derrière CHAINNET est plutôt originale. L'algorithme traite les chaînes, obtenues à la deuxième étape, les unes après les autres dans l'ordre décroissant de leurs scores ; il les place sur le génome de référence et marque les positions du génome qui sont couvertes par les ancres d'une chaîne. Une position du génome de référence ne peut être couverte que par une ancre d'une seule chaîne. Ainsi, si la chaîne considérée à un instant donné couvre des positions

déjà marquées, l'algorithme ne garde et ne positionne que les sous-parties de cette chaîne qui couvrent des positions non marquées. De plus, si une chaîne peut s'insérer dans un gap d'une chaîne déjà positionnée sur le génome de référence, celle-ci est ajoutée mais à un niveau hiérarchique inférieur. Ainsi l'algorithme forme un réseau de chaînes, avec des chaînes enfants de chaînes parentes. Cette idée est originale, l'une des motivations pour former un tel réseau est d'identifier des réarrangements (comme des inversions) qui sont insérés ("embedded") dans des plus grands blocs de synténie (voir un exemple dans la Figure 2.8).

FIG. 2.8: Figure tirée de (Kent *et al.*, 2003), représentant le réseau de chaînes obtenu sur une partie du chromosome 7 humain (positions sur le chromosome indiquées en haut de la figure) aligné avec le génome de la souris. Au niveau 1, la partie représentée du chromosome 7 humain est couverte par une seule chaîne orange (chaque couleur correspond à un chromosome de la souris). Les rectangles ponctuant la chaîne représentent les ancres, c'est-à-dire les parties effectivement alignées, alors qu'entre les rectangles on a des gaps. Ainsi au deuxième niveau, on a pu placer d'autres chaînes sur cette partie du chromosome 7 dans les gaps de la première chaîne. On observe ici une inversion de 15 Kb et d'autres petits réarrangements.

Enfin, MAUVE a lui aussi une troisième étape très différente des autres méthodes. Elle consiste à appliquer les deux premières étapes de manière récursive, avec des paramètres moins stringents (taille minimale des MUMs) et dans des espaces plus confinés. Entre chaque paires de MUMs consécutifs d'une chaîne, on applique les deux premières étapes : on recherche des MUMs, et on les chaîne. On cherche également de nouvelles chaînes entre celles existantes. Ainsi, les chaînes formées à une étape d'itération ne sont jamais remises en question. Lorsqu'on n'arrive plus à trouver de nouvelles chaînes, les séquences couvertes par des chaînes sont alignées globalement.

2.3.4 Discussion

Ces méthodes d'alignement de génomes complets rencontrent souvent les mêmes problèmes que ceux mentionnés dans la partie sur les blocs de synténie : les chevauchements de blocs, le traitement des duplications et le choix des seuils qui détermine quels réarrangements sont détectés et lesquels sont masqués. Ce dernier point est accentué, avec ces méthodes, car les données sont plus nombreuses et plus bruitées.

Les micro-réarrangements GRIMM appelle micro-réarrangements, les petits réarrangements qui sont éliminés ou masqués à l'intérieur des blocs de synténie. C'est le choix des paramètres de gap (G) et de taille (S) qui détermine quels réarrangements sont gardés et lesquels sont éliminés. Or, il semble que ces paramètres sont souvent fixés de manière arbitraire, sans justification biologique. Ainsi, lors de la publication de la méthode, les paramètres étaient tous deux fixés à 1 Mb, ce qui implique qu'aucun réarrangement de moins d'1 Mb ne peut être détecté. Cela diminue considérablement la résolution de la méthode et son intérêt par

rapport aux "anciennes" méthodes. En diminuant ces valeurs, on s'expose alors à augmenter le nombre de faux positifs et d'erreurs.

La méthode CP étant similaire pourrait également masquer des réarrangements, mais l'alignement global des clusters proscrit l'existence de petits réarrangements à l'intérieur des clusters. De plus, elle est surtout paramétrée pour produire des petits blocs et la méthode ne cherche pas à les regrouper lorsqu'ils sont colinéaires et consécutifs dans les deux génomes (voir le nombre et la taille des blocs dans le Tableau 2.2). Actuellement, afin d'augmenter la fiabilité des blocs obtenus, les auteurs utilisent désormais les parties codantes annotées (exons) comme ancres potentielles, plutôt que les alignements génomiques (méthode Mercator [2]). C'est également la stratégie souvent utilisée pour comparer des génomes distants (Bourque *et al.*, 2005).

Chainnet ne possède pas de paramètre de taille et ne semble pas éliminer de réarrangements sur ce critère. Cependant, les données produites ne peuvent être utilisées directement pour analyser les réarrangements. Ainsi, lors de la publication de la méthode, les auteurs ont dû appliquer finalement un seuil sur la taille (de 100 Kb) pour effectuer une analyse des réarrangements. De plus, seuls les inversions (visibles au niveau 2) et les réarrangements non chevauchants (niveau 1) ont été analysés. On peut noter, ici, que la méthode développée postérieurement par la plateforme Ensembl (Hubbard *et al.*, 2007) est une combinaison de Grimm et de Chainnet : les alignements issus du réseau de chaînes de Chainnet sont utilisés comme ancres dans un algorithme de regroupement en blocs similaire à celui de Grimm. La structure hiérarchisée de chaînes produite par Chainnet s'avère donc difficile à utiliser. En effet, à un même niveau, on peut trouver aussi bien de très grandes chaînes que de toutes petites chaînes, voire des ancres isolées, dont on peut douter de la fiabilité. Ainsi, cette hiérarchie est difficile à interpréter d'un point de vue biologique.

Enfin, Mauve n'est pas adapté aux génomes de vertébrés. Son étape de chaînage n'est pas assez flexible, et son étape de recherche d'ancres est trop stringente pour des génomes peu codants, avec beaucoup de répétitions. Cet algorithme est surtout utilisé pour comparer des génomes bactériens proches.

Le Tableau 2.2 présente quelques statistiques sur les alignements et les blocs obtenus par les méthodes Grimm, Chainnet et CP, sur un même jeu de données : les génomes de l'homme et de la souris. Bien sûr, les chiffres ne sont pas directement comparables, puisque les paramètres utilisés sont différents. Cependant, on observe de très grandes différences en terme de nombre de blocs et de leurs tailles. Notons que Grimm et Chainnet ont été comparés dans la publication de la séquence du génome du rat (Gibbs *et al.*, 2004), conduisant globalement aux mêmes segments conservés. Cependant, cette comparaison est très peu décrite dans l'article (aucun chiffre ni aucun détail quant aux critères de comparaison utilisés) et seuls les segments conservés de plus de 1 Mb ont été comparés.

Les chevauchements et les duplications En ce qui concerne le chevauchement des blocs, Chainnet et Mauve produisent des alignements qui ne se chevauchent pas. Cependant, si un chevauchement était possible, ils effectuent un choix entre les différentes possibilités. Par contre, Grimm et CP ne contrôlent pas si les blocs se chevauchent ou pas. La question des chevauchements rejoint celle des duplications, car d'un point de vue biologique, deux blocs qui se chevauchent sur un génome constituent, soit des réarrangements supplémentaires, soit une duplication sur l'autre génome. Or, aucune des méthodes présentées ici ne tente d'identifier des duplications. Grimm et Mauve les éliminent dès la première étape en restreignant

[2] `http://www.biostat.wisc.edu/~cdewey/mercator/`

Méthode	Longueur des blocs retenus	Couverture du génome	% du génome aligné	Nombre de blocs de synténie	Longueur moyenne des blocs	Longueur moyenne inter-bloc
Chainnet	>100 Kb	90.9 %	32.9%	579	983 Kb	450 Kb*
Grimm	>1 Mb	93%	?	281	9.6 Mb	668 Kb
CP [1]	>100 Kb	76%	<35.2 %	8080	270 Kb*	86 Kb*

Tab. 2.2: Comparaison des blocs de synténie entre les génomes de l'homme et de la souris, obtenus avec les différents algorithmes (sauf pour CP). Un point d'interrogation signifie que l'information n'a pas été trouvée dans la publication. Une étoile indique que l'information n'a pas été trouvée dans la publication, mais calculée avec les formules suivantes. La longueur moyenne inter-bloc est donnée par : $\simeq$ taille du génome $\times (1 -$ couverture$))$/nombre de blocs. La taille moyenne des blocs est donnée par : $\simeq$ taille du génome$\times$couverture/nombre de blocs. [1] données obtenues dans (Brudno *et al.*, 2004).

le jeux d'ancres à celles qui n'apparaissent qu'une seule fois dans chaque génome. Grimm élimine également les ancres qui se chevauchent. De tels critères éliminent un grand nombre d'ancres, alors que l'étape suivante de filtrage aurait pu permettre de sélectionner les bonnes et éliminer les mauvaises. De plus, pour Grimm, étant donné que l'algorithme permet des chevauchements, l'identification de duplications pourrait être conduite en même temps que l'identification des blocs de synténie. La description de la méthode CP n'indique pas clairement comment elle traite les duplications, il semble donc qu'elle les garde si elles sont dans un cluster significatif, puis dans un alignement significatif. Quant à Chainnet, les duplications sont gérées de manière asymétrique : elles sont autorisées dans l'un des génomes (le génome de référence) et bannies dans l'autre. Dans le génome de référence, si une position appartient à plusieurs chaînes, donc si une position est potentiellement dupliquée dans l'autre génome, une des copies est en fait choisie : c'est celle qui appartient à la chaîne de meilleur score. Cette asymétrie implique qu'on n'obtient pas les mêmes alignements en fonction du génome choisi comme référence.

Chapitre 3

Construction de blocs de synténie

Sommaire

Pour analyser les points de cassure, nous voulons des blocs de synténie et surtout des points de cassure très résolus. Les méthodes d'alignement de génomes complets semblent alors les plus appropriées puisqu'elles ont la plus forte densité en marqueurs. Cependant, ces méthodes présentent certains inconvénients (voir Section 2.3.4) et n'aboutissent pas en général à des résolutions satisfaisantes. En effet, les moyennes des tailles des points de cassure sont de l'ordre de 500 Kb (voir Tableau 2.2).

Le problème de résolution provient du compromis nécessaire entre sensibilité et spécificité. Les alignements de séquences, et notamment de séquences non codantes, peuvent comporter un grand nombre de faux positifs (séquences alignées non orthologues). Ces faux positifs sont le plus souvent dus à des séquences répétées, des duplications ou parfois des erreurs d'assemblage. De plus, lorsque les génomes comparés sont éloignés phylogénétiquement, peu de séquences sont conservées et alignables. Par exemple, 40% du génome de l'homme est alignable avec celui de la souris. Ainsi, à l'étape d'ancrage il faut utiliser des paramètres assez sensibles, mais on augmente le risque d'obtenir des alignements de séquences non orthologues. L'étape de filtrage permet d'éliminer un certain nombre de ces faux positifs et plus elle est stringente, le plus d'erreurs sont éliminées. Bien sûr, en contre-partie, on perd en résolution. Ce problème vient du fait qu'on détecte l'emplacement des blocs de synténie et leurs limites simultanément, alors que ces deux tâches mériteraient des stratégies et des paramètres différents.

L'idée que nous avons adoptée est donc de procéder en deux étapes :

1. identifier des blocs de synténie fiables mais pas forcément très résolus,
2. affiner chaque point de cassure indépendamment, en essayant d'étendre les limites des blocs.

Ainsi, dans la première étape, on peut être assez stringent pour éviter les erreurs, et dans la deuxième, on a réduit l'espace de recherche et on peut être plus sensible. Nous avons décidé d'utiliser, pour la première étape, l'information la plus fiable, les gènes, et pour la seconde, les séquences génomiques afin d'obtenir une meilleure précision. L'ensemble de la méthode a été publié (Lemaitre *et al.*, 2008b).

Nous décrivons dans ce chapitre la première étape, le chapitre suivant étant consacré à l'étape d'affinement des points de cassure. Nous présentons tout d'abord, dans la Section 3.1, les données d'orthologie et les problèmes rencontrés qui nous ont amenés à proposer une méthode de construction de blocs de synténie, détaillée dans la Section 3.2. Enfin, dans la dernière section, nous présentons une application de cette méthode à la comparaison homme-souris.

3.1 Analyse des données d'orthologie

Dans un premier temps, nous n'avons autorisé aucune flexibilité pour construire les blocs de synténie à partir de l'ordre des gènes orthologues. Ainsi, lorsque deux gènes sont colinéaires et consécutifs dans les deux génomes, ils appartiennent au même bloc de synténie. Lorsque ce n'est pas le cas, on définit un point de cassure et on crée un nouveau bloc. Cela implique qu'une seule paire de gènes orthologues peut constituer un bloc à part entière.

La motivation était de ne pas éliminer un certain type de réarrangements, notamment les réarrangements de petite taille. En effet, les méthodes d'alignement de génomes complets ont un paramètre de taille minimale des blocs de synténie et éliminent ainsi systématiquement les petits blocs. Il est vrai que, par cette méthode, nous ne détectons pas tous les blocs de synténie, nous ne pouvons détecter que les blocs contenant au moins un gène orthologue, mais la sélection ici n'est pas faite sur la taille du bloc.

Cependant, l'analyse des gènes orthologues de l'homme et de la souris nous a amené à mettre en doute la fiabilité des blocs et des points de cassure obtenus avec cette stratégie.

3.1.1 Comparaison de plusieurs jeux de données

Nous avons utilisé deux jeux de données de gènes orthologues entre l'homme et la souris, accessibles depuis la base de données de cartographie comparée GeMCore (Navratil, 2005)[1]. Les données de séquences et leurs annotations sont issues de la base de données Ensembl (version 24) (Hubbard *et al.*, 2007). Le premier jeu d'orthologues, que nous appelons RBH, est construit avec la méthode de meilleur hit réciproque, alors que le deuxième, TREEPATTERN, est obtenu avec une approche phylogénétique de réconciliation d'arbres (Dufayard *et al.*, 2005) (voir Section 2.2.2b.). Les deux jeux de données sont constitués de relations d'orthologie 1-1.

On définit un point de cassure entre deux gènes orthologues a_r et b_r consécutifs dans le génome de référence G_r, mais dont les orthologues a_o et b_o ne sont pas consécutifs sur l'autre génome G_o, ou bien sont non colinéaires (voir Figure 3.1).

Le jeu RBH contient 15268 paires de gènes orthologues 1-1, résultant en 1480 points de cassures sur le génome de l'homme. A partir des 12474 paires d'orthologues de TREEPATTERN,

[1] `http://pbil.univ-lyon1.fr/gem/gem_home.php`

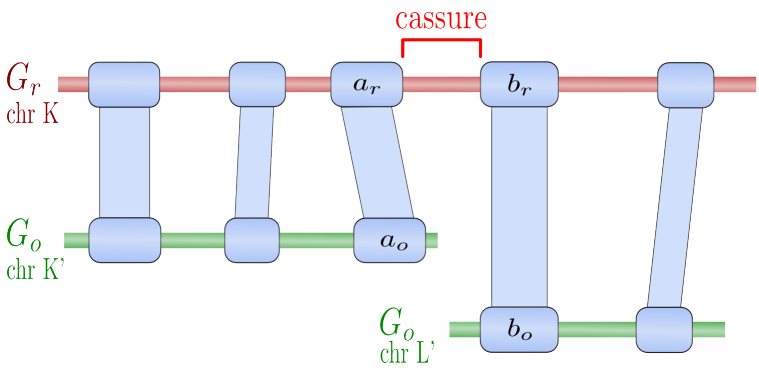


FIG. 3.1: Représentation schématique d'un point de cassure. Les gènes sont représentés par des carrés bleus et sont reliés entre les deux génomes par des ombres bleues s'ils sont orthologues. On définit un point de cassure entre les deux gènes successifs a_r et b_r, sur le génome de référence G_r, car leurs orthologues, respectivement a_o et b_o, sont sur deux chromosomes différents dans le génome G_o. Les autres gènes représentés ne sont pas impliqués dans des cassures, puisqu'ils sont colinéaires et consécutifs dans les deux génomes.

nous obtenons 986 points de cassure (voir Tableau 3.1). La différence en nombre de points de cassure entre les deux jeux de données est très importante et n'est pas proportionnelle à la différence de taille des jeux de données. De plus, seulement 308 points de cassure sont identiques entre les deux jeux de données (définis par les mêmes gènes a_r et b_r). Ainsi les données de points de cassure sont très sensibles à la méthode d'assignation d'orthologie.

Jeu de données	Nombre de gènes orthologues	Nombre de cassures	Nombre (et %) de cassures de type I ou II
RBH	15268	1480	510 (34 %)
TREEPATTERN	12474	986	352 (36 %)
INTERSECTION	12027	657	134 (20 %)

TAB. 3.1: Statistiques des trois jeux de données d'orthologie.

On construit alors un troisième jeu de données, en ne sélectionnant que les paires de gènes prédites par les deux méthodes. Ce jeu, appelé INTERSECTION, contient 12027 gènes, résultant à 657 points de cassure. Le nombre de points de cassure est là encore fortement diminué par rapport à la réduction en nombre de gènes. Ainsi, par rapport à TREEPATTERN, on a enlevé presque autant de gènes que de points de cassure.

Il faut noter que la majorité des différences entre les deux jeux de données RBH et TREEPATTERN provient de gènes qui ont un orthologue dans un jeu mais pas dans l'autre (seuls 74 gènes de l'homme ont un orthologue chez la souris différent entre les deux jeux de données ; dans ce cas au moins une des assignations (orthologie de position) est erronée). Ainsi, on peut se demander si ces gènes "en plus" dans l'un ou l'autre jeu de données ont plus de chances d'être impliqués dans des réarrangements, ou bien si ce sont des assignations erronées qui engendrent de "faux" points de cassure. Plusieurs raisons pourraient être invoquées pour expliquer la première hypothèse : par exemple les gènes impliqués dans des réarrangements pourraient avoir une histoire évolutive particulière due au réarrangement et qui rendrait les

assignations d'orthologie plus difficiles. Cependant, cela n'a pas été démontré, alors que certains biais des méthodes d'assignation sont connus. De plus, le grand nombre de points de cassure obtenus nous oriente plutôt vers l'hypothèse d'erreurs d'assignation d'orthologie. En effet, plusieurs études comparatives des génomes de l'homme et de la souris rapportent un nombre de points de cassure de l'ordre de 200 à 400 (Pevzner et Tesler, 2003a; Waterston *et al.*, 2002). Cependant, les méthodes utilisées sont différentes et nous avons vu dans la Section 2.3.4 qu'elles ont tendance à masquer les petits réarrangements.

3.1.2 Typologie des points de cassure

Nous effectuons une typologie des points de cassure en fonction de l'arrangement des gènes adjacents à la cassure dans les deux génomes (a_r, a_o, b_r et b_o dans la Figure 3.1). Tout d'abord, nous pouvons distinger les cassures intra-chromosomiques, lorsque a_o et b_o sont sur un même chromosome dans le génome G_o, de celles inter-chromosomiques, lorsque a_o et b_o sont sur deux chromosomes différents. Pour les cassures intra-chromosomiques, nous pouvons les classifier en fonction du nombre de gènes séparant a_o et b_o sur leur chromosome. Nous pouvons aller plus loin en prenant en compte les orientations relatives de ces deux paires de gènes.

La typologie des points de cassure homme-souris présente des particularités. Notamment, nous avons identifié deux classes de cassures qui regroupent de 20 % à 35 % des points de cassures en fonction des jeux de données (voir Tableau 3.1).

Elles sont représentées schématiquement dans la Figure 3.2. Dans le type I, la cassure est intra-chromosomique entre les gènes a_r et b_r et elle est due à la seule présence du gène x_o entre a_o et b_o. Dans le type II, il s'agit de deux cassures inter-chromosomiques, entre c_r et d_r d'une part, et entre d_r et e_r d'autre part. Encore une fois, les cassures sont dues à la seule présence du gène d_r entre c_r et e_r.

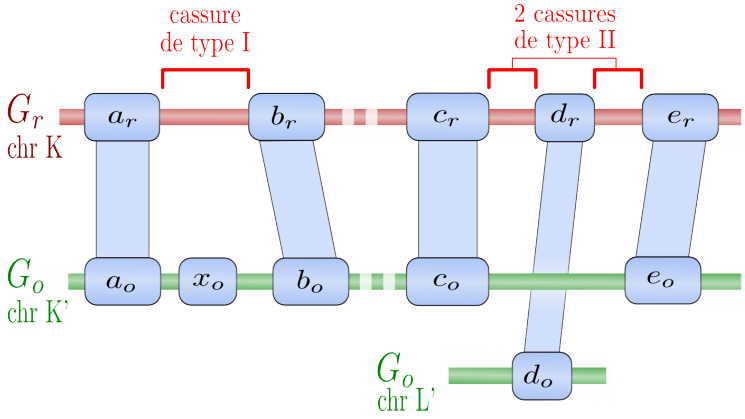


FIG. 3.2: Représentation schématique de deux types de cassures particulières. La cassure de type I est intra-chromosomique entre les gènes a_r et b_r sur le génome G_r. Le deuxième type est composé de deux cassures inter-chromosomiques consécutives entre c_r et d_r d'une part, et entre d_r et e_r d'autre part. Dans les deux cas, les cassures sont provoquées par la présence d'un seul gène, x_o pour le type I et d_r pour le type II.

Ces deux types de points de cassure ont les configurations attendues dans le cas d'erreurs d'assignation d'orthologie au sein de blocs de synténie. On peut alors douter de la fiabilité des assignations d'orthologie des gènes x_o du type I et (d_r, d_o) du type II.

3.1.3 Analyse des points de cassure de type I et II

Les points de cassure de type I et II se distinguent des autres en fréquence, mais également lors de l'étape d'affinement des points de cassure.

Nous avons appliqué l'étape d'affinement des points de cassures sur les séquences bordées par les deux gènes orthologues a_r et b_r. Cette étape d'affinement est décrite très précisément dans le chapitre suivant, nous indiquons seulement ici qu'elle est basée sur l'alignement des séquences adjacentes aux gènes bordant la cassure. L'affinement n'est efficace que si ces séquences sont suffisamment similaires.

Notamment, dans le cas des cassures de type II, un seul côté de la région de cassure semble pouvoir être affiné. Plus précisément, les séquences voisines du gène d_r ne sont pas similaires aux séquences voisines du gène d_o dans l'exemple de la Figure 3.2. Cela laisse supposer que ces deux gènes ne sont pas orthologues de position. Dans ce cas, on peut éliminer ces points de cassure qui ne sont pas générés par des réarrangements équilibrés. De même pour le type I, si l'orthologie du gène x_o est mal assignée, la cassure entre les gènes a_r et b_r peut être éliminée.

A partir des 134 cassures de type I ou II (du jeu INTERSECTION), nous avons identifié 78 paires de gènes "suspectes". Une des hypothèses pour expliquer des erreurs d'assignation de gènes isolés est le phénomène de rétrotransposition. Il arrive qu'un ARN messager d'un gène soit réinséré dans le génome à une position aléatoire. Souvent ce gène n'est plus actif car il lui manque sa région promotrice, il devient alors un pseudogène et on l'appelle rétropseudogène. Cependant, il peut être annoté comme un gène. Une de ses caractéristiques est qu'il ne contient pas d'intron. Ainsi, nous avons examiné les nombres d'introns des paires de gènes suspectes. Pour 29 d'entres elles (37%), l'un des deux gènes orthologues ne possède aucun intron alors que l'autre en possède au moins un. Avant d'affirmer que l'un des deux gènes est issu d'une rétrotransposition, il faudrait les analyser plus en détail. Cependant, cela indique qu'il peut y avoir des erreurs dans les données d'orthologie, notamment dues à la rétrotransposition.

Finalement, nous avons également confirmé que les résultats d'affinement sont meilleurs lorsque les gènes adjacents à la cassure forment des blocs de synténie de plusieurs gènes de part et d'autre de la cassure (comme représenté dans la Figure 3.1).

3.1.4 Conclusion

Le nombre et les types de points de cassure obtenus dépendent fortement du jeu de gènes orthologues utilisés. Cela vient du fait que les méthodes d'assignation d'orthologie ne sont pas fiables à 100 % et ont un taux de faux positifs non négligeable, surtout en ce qui concerne les orthologues de position. Cependant, même avec une méthode d'assignation très stringente (intersection de deux méthodes), nous avons pu mettre en évidence encore certaines de ces erreurs d'assignation. Or cela représente une part non négligeable de l'ensemble des points de cassure obtenus. Il semble ainsi nécessaire de construire des blocs de synténie plus flexibles qui permettent de s'affranchir de ces erreurs potentielles, quitte à perdre certains petits réarrangements.

3.2 Une méthode de construction de blocs de synténie

Nous avons identifié des erreurs potentielles d'assignation d'orthologie grâce à l'arrangement particulier des gènes bordant les points de cassure. Afin d'assainir les données, il suffirait d'éliminer ces gènes facilement identifiables. Pour les deux types de cassures douteuses (types

I et II), il s'agit de gènes orthologues isolés qui cassent un bloc de synténie à eux seuls. Cependant, ce ne sont pas les seuls cas où un seul gène casse un bloc. De plus, l'élimination de certains gènes douteux peut en faire apparaître ou disparaître d'autres. Ainsi, en fonction de l'ordre dans lequel on élimine ces gènes, les résultats seront différents.

Il est donc nécessaire de définir formellement les objets que l'on recherche et quels sont les critères pour éliminer ou garder des gènes orthologues. Nous recherchons en fait une méthode de construction de blocs de synténie flexible où les gènes n'appartenant à aucun bloc de synténie sont éliminés.

Un autre point important est que les blocs doivent respecter certaines conditions. Il est nécessaire pour l'étape d'affinement des points de cassure que les blocs ne se chevauchent sur aucun génome et que leurs extrémités soient définies par le même marqueur orthologue dans les deux génomes (voir Section 4.4.1). Nous appelons ce type de blocs, les blocs de synténie *sans conflit*. Nous avons vu dans la Section 2.2.3 que ces conditions sont rarement respectées et prises en compte par les méthodes existantes. C'est la raison pour laquelle nous avons développé notre propre méthode de construction de blocs de synténie. Notre approche est très proche de certaines méthodes existantes, mais elle a deux avantages : elle fait appel à une définition formelle des objets que nous recherchons, et elle produit des blocs sans conflit. Elle possède un paramètre k qui contrôle le degré de flexibilité autorisé au sein des blocs de synténie.

3.2.1 Définition formelle des blocs de synténie

Notations La méthode prend en entrée un nombre entier k et un ensemble d'ancres entre deux génomes G_r et G_o. On appelle une ancre une paire de marqueurs, un dans chaque génome, qui sont orthologues. Seuls les marqueurs qui ne se chevauchent pas sur les deux génomes et qui font partie d'une seule ancre sont pris en compte. Les marqueurs sans orthologue ne sont pas considérés et les relations d'orthologie sont seulement 1-1. Si on oriente arbitrairement les chromosomes avec une position de début et un brin direct, on peut identifier un marqueur par son chromosome, sa position sur le chromosome (par rapport à la position de début) et son orientation.

Ici, nous ne sommes intéressés que par l'ordre relatif des marqueurs les uns par rapport aux autres, nous abandonnons donc les positions physiques des marqueurs sur le chromosome au profit de leur rang sur ce chromosome par rapport aux autres marqueurs.

Une ancre est définie par une paire de chromosomes, une paire de rangs (rang de chaque marqueur sur son chromosome respectif) et une orientation relative. Par exemple, soit a une ancre, elle est identifiée par l'objet $(c_r, c_o, a_r, a_o, \sigma_a)$, avec c_r et a_r le chromosome et le rang respectivement du marqueur sur G_r, c_o et a_o le chromosome et le rang de son marqueur orthologue sur G_o, et enfin σ_a est égal à $+1$ si les deux marqueurs ont la même orientation, -1 sinon.

Définition d'une distance Si deux ancres distinctes a et b sont sur les mêmes chromosomes dans les deux génomes, la distance entre a et b, denotée $d(a, b)$, est le maximum des différences de rang entre a et b sur chaque génome : si a_r, a_o (resp. b_r, b_o) sont les rangs des ancres a (resp. b) sur les génomes G_r et G_o, alors $d(a, b) = \max(|b_r - a_r|, |b_o - a_o|)$. Ainsi, si a et b sont consécutifs sur chaque génome, la distance entre les deux est égale à 1. Si deux ancres contiennent des marqueurs qui ne sont pas sur le même chromosome dans au moins un des deux génomes, alors la distance est infinie.

Définition d'un graphe d'ancres On appelle $\mathcal{G}_k$ le graphe dirigé, tel que les noeuds sont les ancres et deux ancres sont reliées par un arc si elles sont colinéaires et à une distance inférieure à k. Plus formellement, on a un arc, dénoté par ab, de l'ancre a à l'ancre b si :

- $a_r < b_r$,
- $d(a,b) \leq k$,
- et ($a_o < b_o$ et $\sigma_a = \sigma_b = 1$) ou ($a_o > b_o$ et $\sigma_a = \sigma_b = -1$).

Un chemin dans ce graphe représente un ensemble d'ancres colinéaires proches dans les deux génomes. On appelle cela une chaîne. Une ancre peut appartenir à plusieurs chaînes et nous allons donc chercher des composantes connexes de ce graphe, c'est-à-dire des ensembles de noeuds connectés. Plus précisément, une composante connexe d'un graphe non orienté est un sous-graphe connecté maximal : il existe un chemin entre n'importe quels noeuds d'une composante connexe. Une composante connexe d'un graphe orienté, est une composante connexe du graphe non orienté dérivé en remplaçant les arcs par des arêtes.

Définition des conflits A ce stade, si k est plus grand que 1, une ancre peut être chaînée à plusieurs ancres distinctes. Ainsi, on peut obtenir des composantes connexes ayant des ancres qui ne sont pas colinéaires deux à deux. On peut également obtenir des composantes connexes dont les coordonnées génomiques se chevauchent. On parle dans ce cas de conflits entre chaînes.

Nous définissons alors deux types de conflits qui s'appliquent sur les arcs du graphe :

- le conflit de type I : deux arcs ab et cd qui appartiennent à la même composante connexe de $\mathcal{G}_k$ sont dits conflictuels si les ancres a, b, c, et d ne sont pas toutes colinéaires deux à deux (voir un exemple où $a = c$ dans la Figure 3.3).
- le conflit de type II : un arc ab dans une composante connexe C est conflictuel s'il existe une ancre x, dont au moins un des marqueurs est localisé entre les marqueurs de a et b dans un des deux génomes, et x appartient à une autre composante connexe qui possède au moins k ancres (voir un exemple dans la Figure 3.4).

Les arcs conflictuels représentent plusieurs possibilités de chaînage au sein d'une composante connexe (conflit de type I), ou bien des chaînes qui se chevauchent (conflit de type II).

Définition des *k-blocs* On appelle, maintenant, $\mathcal{H}_k$ le sous-graphe de $\mathcal{G}_k$ qui ne contient que les arcs non conflictuels de $\mathcal{G}_k$.

Un *k-bloc* est une composante connexe de $\mathcal{H}_k$ contenant au moins k noeuds. Les coordonnées des ancres aux extrémités des k-blocs délimitent les blocs de synténie sur le génome.

L'absence d'arcs conflictuels de type II garantit que les blocs de synténie ne se chevauchent pas sur l'un ou l'autre génome. Et l'absence d'arcs conflictuels de type I garantit que les ancres à l'intérieur d'un bloc sont totalement ordonnées, c'est-à-dire toutes colinéaires deux à deux. Ainsi on est sûr que les extrémités des blocs sont bien définies par une même ancre dans les deux génomes.

3.2.2 Complexité algorithmique et implémentation

En utilisant directement la définition, le temps de calcul des k-blocs est polynomial : si n est le nombre d'ancres, le calcul des rangs nécessite une procédure de tri sur toutes les ancres sur les deux génomes (temps en $O(n \log n)$). Ensuite, la construction du graphe prend un temps proportionnel à $n \times k$ et produit au plus $n \times k$ arcs. Le temps de calcul des composantes connexes est proportionnel au nombre d'arcs (temps en $O(n \times k)$). Pour chaque

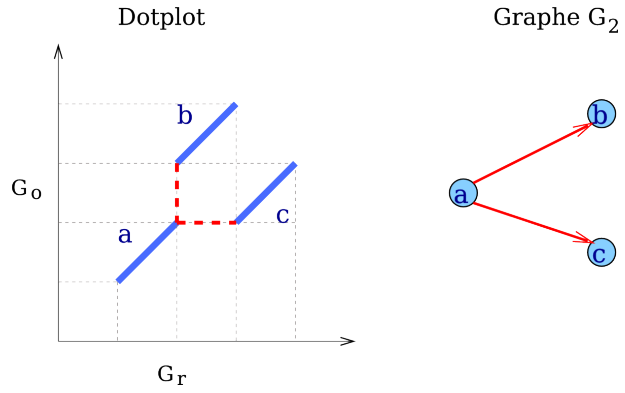


FIG. 3.3: Exemple de deux arcs conflictuels de type I. A gauche, la représentation en dotplot montre les positions des ancres a, b et c sur les deux génomes G_r et G_o. Elles sont sur le même chromosome sur les deux génomes et $d(a,b) = d(a,c) = 2$. Le graphe correspondant $\mathcal{G}_2$ ($k = 2$) est représenté à droite. Les arcs ab et ac sont conflictuels car l'ordre des marqueurs de a, b et c dans G_r est a, b, c, alors que c'est a, c, b dans G_o : les ancres b et c ne sont pas colinéaires.

arc, l'identification de conflits nécessite la comparaison avec au plus $2k$ autres arcs ou noeuds. La complexité algorithmique en temps de l'ensemble est donc $O(n \times k^2 + n \log n)$, avec k un paramètre fixé généralement faible (inférieur à 10).

Le pseudo-code de l'algorithme de construction des k-blocs est présenté dans l'algorithme 1. Il a été implémenté, pour les valeurs 1, 2 et 3 du paramètre k, dans une fonction R.

3.2.3 Discussion sur la méthode

La méthode décrite est flexible : les blocs sont construits en chaînant les marqueurs orthologues qui sont colinéaires et en autorisant un certain nombre de marqueurs "désordonnés". Le degré maximum de flexibilité autorisée est controlé par le paramètre k. Si $k = 1$, aucune flexibilité n'est autorisée et seules les ancres colinéaires et consécutives dans les deux génomes sont chaînées.

Le fait de n'avoir qu'un seul paramètre diffère avec les autres méthodes existantes et précédemment décrites (Section 2.2.3), qui utilisent en général au moins deux paramètres (Pevzner et Tesler, 2003a; Pan *et al.*, 2005; Hubbard *et al.*, 2007; Sinha et Meller, 2007), le premier, G, pour la distance maximale autorisée entre deux ancres chaînées et le deuxième, S, pour la taille minimale d'un bloc retenu. Notre choix d'un seul paramètre vient de l'observation suivante : ces deux paramètres sont liés. Si on veut empêcher un bloc d'être contenu dans un autre, il est nécessaire que $G \leq S$. Mais d'autre part, on élimine les blocs de taille inférieure à S car on suppose que ces blocs sont des erreurs. Or de telles erreurs ne devraient pas nous empêcher de chaîner des ancres de part et d'autre. On doit alors fixer $G \geq S$. Nous avons donc choisi de fixer $G = S$, paramètre que nous avons appelé k. En fait, ces deux paramètres ont souvent la même valeur dans les applications des méthodes (par exemple dans (Pevzner et Tesler, 2003a; Bourque *et al.*, 2005; Hubbard *et al.*, 2007; Sinha et Meller, 2007)).

Une conséquence de permettre de la flexibilité est la possibilité d'obtenir des conflits : soit des blocs peuvent se chevaucher (bien que les ancres ne se chevauchent pas), soit il peut exister plusieurs possibilités de chaînage des ancres au sein d'un bloc. Au lieu d'introduire

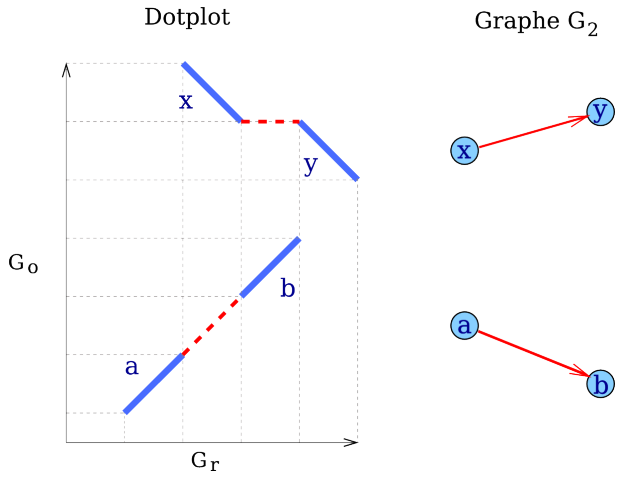


FIG. 3.4: Exemple d'un arc conflictuel de type II. A gauche, la représentation en dotplot montre les positions des ancres a, b, x et y sur les deux génomes G_r et G_o. Les ancres a et b sont sur le même chromosome dans les deux génomes et $d(a,b) = 2$, de même pour les ancres x et y. Le graphe correspondant $\mathcal{G}_2$ ($k = 2$) est représenté à droite. Il possède deux composantes connexes : $\{a, b\}$ et $\{x, y\}$. L'arc ab est conflictuel car $a_r < x_r < b_r$ et x appartient à une autre composante connexe avec au moins deux ancres.

des contraintes dont le sens biologique n'est pas toujours évident, nous avons choisi de ne pas résoudre les conflits mais d'éliminer les régions conflictuelles des blocs. C'est une solution qui peut paraître radicale car cela réduit la taille des blocs. Mais dans notre stratégie, la construction des blocs de synténie ne constitue que la première étape avant d'affiner les points de cassures. En effet, dans la majorité des cas, l'information perdue en supprimant les conflits pourra être récupérée à l'étape d'affinement des points de cassure. Si les arcs conflictuels supprimés ne font que raccourcir un bloc à une de ses extrémités, les ancres supprimées feront partie des séquences alignées dans l'étape d'affinement des points de cassure. Par contre, si la suppression des arcs conflictuels implique la suppression d'un bloc en entier, alors ce bloc ne sera probablement pas retrouvé à l'étape d'affinement.

La flexibilité est nécessaire pour éliminer les erreurs d'assignation d'orthologie qui engendrent de "faux" points de cassure. Il reste cependant à fixer le degré de flexibilité voulu, c'est-à-dire donner une valeur au paramètre k. Plus k est grand, plus les blocs obtenus sont fiables, mais moins on a de blocs. Le problème lorsqu'on rate des blocs n'est pas seulement de manquer des points de cassure, mais également de réduire la résolution de certains points de cassure restants. Si un bloc est manqué entre deux blocs réarrangés, c'est-à-dire dans un point de cassure, cela peut nous empêcher d'affiner ce dernier efficacement. Il s'agit donc de faire un compromis entre le taux de faux positifs (fiabilité) d'un côté, et de l'autre, le taux de faux négatifs (points de cassures manqués) et la précision des points de cassure restants.

3.2.4 Perspectives

Cette définition formelle des blocs de synténie pourrait être étendue pour l'utilisation d'ancres chevauchantes et avec des relations d'orthologie n-m.

De nombreux gènes se chevauchent sur le génome, qu'ils soient sur le même brin ou non, avec parfois des arrangements complexes. Ils peuvent se chevaucher sur une petite partie,

Algorithme 1 : Construction des k-blocs

Entrées : ensemble d'ancres avec leurs coordonnées sur les deux génomes G_r et G_o, k
Sorties : ensemble des k-blocs

trier les ancres selon G_r, assigner le rang a_r à chaque ancre a
trier les ancres selon G_o, assigner le rang a_o à chaque ancre a
// Construction du graphe
pour *chaque ancre* $a(a_r, a_o)$ **faire**
    pour *chaque ancre* $b(b_r, b_o)$ *telle que* $b_r = a_r + i$, $0 < i \leq k$ **faire**
        si *a et b colinéaires et* $d(a, b) \leq k$ **alors**
            créer un arc $a \rightarrow b$
calculer les composantes connexes du graphe
// Identification des conflits de type I
pour *chaque ancre* $a(a_r, a_o)$ *appartenant à une composante connexe de taille* $\geq k$ **faire**
    pour *chaque ancre* $b(b_r, b_o)$ *telle que* $(b_r = a_r + i,\ 0 < i \leq k)$ *ou* $(b_o = a_o + i,\ -k \leq i \leq k)$ **faire**
        si *b appartient à la même composante connexe que a, et a et b ne sont pas colinéaires* **alors**
            marquer tous les arcs entrant et sortant de a et b comme conflictuels
// Identification des conflits de type II
pour *chaque arc* $a(a_r, a_o) \rightarrow b(b_r, b_o)$ *appartenant à une composante connexe de taille* $\geq k$ **faire**
    pour *chaque ancre* $c(c_r, c_o)$ *telle que* $(a_r < c_r < b_r)$ *ou* $(a_o < c_o < b_o)$ **faire**
        si *c appartient à une composante connexe de taille* $\geq k$ *différente de celle de a et b* **alors**
            marquer l'arc $a(a_r, a_o) \rightarrow b(b_r, b_o)$ comme conflictuel
// Calcul des k-blocs
calculer les composantes connexes du graphe privé des arcs marqués comme conflictuels, en notant pour chaque composante les ancres de rangs min et max dans les deux génomes
pour *chaque composante connexe de taille* $\geq k$ **faire**
    créer un k-bloc dont les coordonnées sont celles des ancres de la composante de rangs min et max dans les deux génomes
retourner les k-blocs

parfois non codante (extrémités UTRs), ou bien être entièrement imbriqués l'un dans l'autre, sur le même brin ou non. Pour l'instant, ces gènes sont traités avant l'étape de construction des blocs : si un ensemble de gènes qui se chevauchent sur l'un des deux génomes sont dans le même ordre et la même orientation dans les deux génomes, alors ils sont groupés en une seule ancre, sinon ils sont éliminés. Ce traitement n'est pas très fin et il suffit qu'il y ait un seul gène mal ordonné pour éliminer un grand nombre de gènes. Ce traitement pourrait se faire simultanément à la construction des blocs de synténie. Il faudrait sûrement, dans ce cas, modifier l'attribution des rangs des ancres et la distance entre deux ancres.

La deuxième extension possible serait de généraliser les ancres utilisées aux relations d'or-

thologie n-m. La synténie peut en effet permettre de détecter parmi les n-m copies lesquelles sont orthologues de position. Jusqu'à présent, seuls les orthologues 1-1 sont utilisés, ce qui laisse de côté un nombre non négligeable de gènes. Il faudrait alors prendre en compte dans les définitions le fait que plusieurs ancres peuvent avoir le même rang sur un chromosome. Par exemple, pour identifier des conflits de type I, il faudrait comparer l'ordre partiel des ancres dans chaque génome afin de ne pas considérer une duplication en tandem comme un conflit.

On pourrait aller plus loin, en utilisant les familles entières de gènes homologues, c'est-à-dire considérer également les paralogues. Dans ce cas, on pourrait envisager d'identifier également des blocs de gènes dupliqués.

Ces perspectives mériteraient d'être approfondies. Il semble que ces problèmes peuvent être énoncés à partir de la formalisation des k-blocs. La complexité et l'implémentation seront sûrement plus problématiques, notamment dans le cas de relations n-m.

3.3 Application à la comparaison homme-souris

Nous avons appliqué cette méthode afin de détecter les blocs de synténie entre l'homme et la souris en utilisant les gènes orthologues comme marqueurs.

Nous avons utilisé les gènes orthologues 1-1 prédits par Ensembl (méthode de réconciliation d'arbres), sur l'assemblage NCBI35 (Mai 2004) de l'homme, et l'assemblage NCBI m35 (Décembre 2005) de la souris (Hubbard *et al.*, 2007). Les chevauchements de gènes ont été éliminés de la façon suivante : si plusieurs gènes se chevauchent, soit dans le génome de l'homme, soit dans celui de la souris, et si, de plus, ils sont tous colinéaires, alors ils sont groupés en un seul gène, sinon ils sont tous éliminés.

De 13557 orthologues 1-1, nous obtenons 12223 paires de gènes (ou groupes de gènes) non chevauchants.

3.3.1 Blocs de synténie pour plusieurs valeurs de k

La méthode de construction de blocs de synténie a été appliquée sur ces 12223 ancres avec trois valeurs pour le paramètre k : 1, 2 et 3. Le nombre de blocs et le nombre de gènes retenus dans les blocs sont présentés dans le Tableau 3.2 pour les trois valeurs de k. Avec $k = 1$, tous les gènes sont retenus puisque cela correspond à aucune flexibilité. Entre $k = 1$ et $k = 2$, les différences en nombre de gènes et en nombre de cassures sont importantes : le nombre de cassures est diminué de moitié en éliminant seulement 205 gènes, soit 1.7 % des gènes orthologues (de 763 à 366 points de cassure). De $k = 2$ à $k = 3$, les différences sont plus faibles, 67 gènes et 45 points de cassures éliminés. Ainsi, il semble que le "grand nettoyage" se fait essentiellement lors du passage de $k = 1$ à $k = 2$.

k	1	2	3
Nombre de gènes orthologues	12223	12018	11953
Nombre de blocs de synténie	786	389	344
Nombre de points de cassure	763	366	321

TAB. 3.2: Résumé des résultats obtenus pour k=1, 2 et 3 avec les données d'orthologie homme-souris. Pour k=2 et 3, les $k-$blocs colinéaires et consécutifs sur les deux génomes ont été groupés pour ne former qu'un seul bloc.

Nous détaillons, dans le Tableau 3.3, les caractéristiques des blocs obtenus avec $k = 2$.

Longueur	min	max	médiane	moyenne
Taille des blocs (en nb. de gènes)	2	473	13	31
Taille des blocs (en pb)	36 647	79 896 236	2 446 592	6 720 033
Taille des points de cassure (en pb)	1 057	5 311 140	267 891	515 890

TAB. 3.3: Description des 2−blocs obtenus entre l'homme et la souris. Les blocs colinéaires et consécutifs sur les deux génomes ont été groupés pour ne former qu'un seul bloc. Les points de cassure situés dans un centromère humain ont été enlevés de cette analyse. Les statistiques portent sur 389 blocs et 355 points de cassure.

3.3.2 Choix de la valeur de k

La question que l'on se pose est : est-ce qu'avec k=2, il subsiste des erreurs, c'est-à-dire des points de cassure qui n'en sont pas ? Sachant que les ancres considérées ici sont des gènes, il est en fait peu probable que deux erreurs d'assignation d'orthologie se retrouvent dans un même bloc : il faut pour cela que deux gènes mal assignés chez l'homme soient très proches, et que leurs "mauvais" orthologues le soient aussi sur le génome de la souris, et cela de façon colinéaire.

Avec $k = 2$, on obtient 33 blocs n'ayant que deux gènes ; ces blocs ont une taille moyenne de 205 Kb. Comme dans le cas de gènes isolés, il est plus probable que les erreurs se trouvent à l'intérieur de blocs plus grands. Parmi les 33 blocs de taille deux, 14 "cassent" un plus grand bloc de synténie, c'est-à-dire qu'ils sont masqués dans un bloc obtenu avec k=3. De plus, 9 correspondent en fait à une inversion de deux gènes. Or, les inversions sont des réarrangements plutôt courants. Ces 9 blocs nous semblent donc fiables.

Par contre, 17 blocs de taille 2 se trouvent dans un point de cassure identifié avec k=3. Si ces blocs de taille 2 sont réellement orthologues, il est alors très probable que nous ne serons pas capable d'affiner efficacement ces points de cassure définis avec k=3.

Pour ces raisons, nous estimons que les blocs définis avec k=2 sont fiables et nous utiliserons cette valeur par la suite.

Chapitre 4

Détection fine des points de cassure

Sommaire

Comme nous l'avons expliqué au début du chapitre 3, la méthode de détection des points de cassure que nous avons développée est composée de deux étapes : la première consiste à détecter des blocs de synténie (nous l'avons décrite dans le Chapitre 3, Section 3.2), et la deuxième a pour objectif d'améliorer la précision des points de cassure définis à la première étape. Un **point de cassure** est défini entre deux blocs de synténie consécutifs sur un génome s'ils ne sont pas consécutifs sur l'autre génome, ou bien s'ils ne sont pas colinéaires (voir Figure 4.1). Le point de cassure est ainsi délimité par les extrémités des deux blocs de synténie qui le bordent.

Contrairement aux méthodes existantes, qui s'arrêtent à cette étape, nous avons décidé d'aller plus loin et d'étudier ces régions plus en détail. En se concentrant sur chaque région indépendamment nous espérons l'affiner, c'est-à-dire étendre les extrémités des blocs voisins. La motivation principale est de gagner en précision afin d'effectuer des analyses des points de cassure plus précises et plus puissantes.

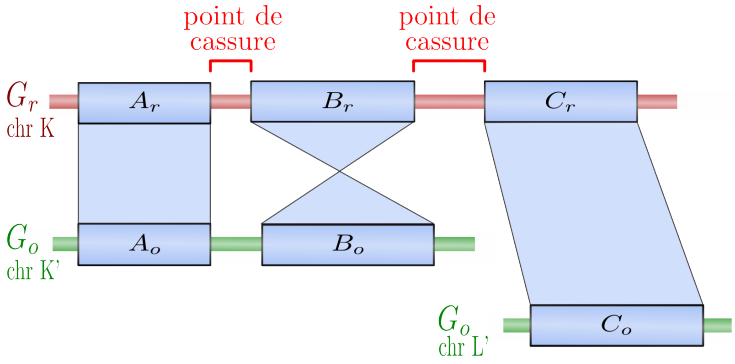


FIG. 4.1: Deux exemples de points de cassure sur le génome G_r, comparé au génome G_o, définis par les blocs de synténie (A_r,A_o), (B_r,B_o) et (C_r,C_o). Les deux blocs (B_r,B_o) et (C_r,C_o) sont consécutifs sur le génome G_r mais ne le sont pas sur le génome G_o (ils sont sur des chromosomes différents). On définit donc une région de cassure sur G_r entre ces deux blocs. On définit également une région de cassure entre les blocs (A_r,A_o) et (B_r,B_o). Ces derniers sont consécutifs dans les deux génomes mais le bloc (B_r,B_o) est inversé sur G_o.

Il est nécessaire ici de commenter cette expression, "point de cassure", car elle peut porter à confusion, principalement pour deux raisons.

La première vient de l'emploi du terme "cassure". Cela suggère une cassure physique de l'ADN, telle qu'une cassure double brin, et lui donne un sens biologique incorrect. Un point de cassure, tel que nous l'avons défini, est en fait une séquence qui n'a pas forcément été "cassée" depuis la divergence des génomes comparés. Cette séquence sur un génome est définie par comparaison avec un autre génome. Nous avons identifié un réarrangement mais nous ne pouvons pas, en général, savoir dans laquelle des deux lignées a eu lieu le réarrangement. Supposons, par exemple, que nous comparons les génomes de l'homme et de la souris, et que l'arrangement ancestral d'un des chromosomes est composé des blocs successifs A, B et C (ABC). Supposons maintenant que l'arrangement humain est le même que l'arrangement ancestral, ABC, alors que celui de la souris est ACB. En comparant l'homme et la souris, la région entre les blocs A et B sur le génome humain est un point de cassure, comme l'est la région entre A et C sur le génome de la souris. Cependant, aucune de ces deux régions ne contient le "vrai" point de cassure, c'est-à-dire la séquence qui a subit la cassure double brin (c'est la séquence entre A et B chez un ancêtre de la souris). Ces deux régions sont néanmoins toutes deux homologues à la région "cassée".

Le terme "point" dans l'expression "point de cassure" porte également à confusion. Bien entendu, le plus souvent, la localisation d'un point de cassure n'est pas aussi précise qu'un point, c'est-à-dire la position entre deux nucléotides successifs d'une séquence génomique. Un point de cassure consiste plutôt, généralement, à une région assez grande. Cette dernière est définie par la séquence bordée par deux blocs de synténie successifs, ce qui implique qu'aucune homologie n'a été détectée dans cette séquence (c'est-à-dire ayant un niveau de similarité significatif) permettant de l'ajouter aux blocs voisins ou de créer un nouveau bloc. Cependant, on ne sait pas *a priori* si cette imprécision est due à des propriétés biologiques de cette région ou au fait qu'on n'a pas été capable de détecter l'orthologie au delà des blocs. L'objectif de cette étape, que nous avons appelée *affinement des points de cassure*, est d'éliminer cette dernière hypothèse.

Nous présentons, dans ce chapitre, la méthode que nous avons développée pour affiner les points de cassure et les résultats obtenus sur les points de cassure de mammifères. Nous commençons par une vue d'ensemble de la méthode, puis chaque étape sera détaillée dans les sections suivantes. Enfin, nous montrons l'efficacité et les avantages de cette méthode par rapport aux méthodes existantes à travers une application aux génomes de mammifères.

4.1 Vue d'ensemble de la méthode

On dispose de deux génomes entièrement séquencés et de leurs blocs de synténie. L'objectif de la méthode est d'affiner les points de cassure sur un génome, qu'on appellera le génome de référence ou G_r, en utilisant les données de l'autre génome, G_o. On rappelle qu'un bloc de synténie est défini par une paire de coordonnées (A_r, A_o) et une orientation σ_A ; A_r (A_o) sont les coordonnées (chromosome, début et fin) du bloc A sur le génome G_r (G_o respectivement).

4.1.1 Définition du point de cassure et des séquences d'intérêt

On définit un **point de cassure** entre les blocs (A_r, A_o) et (B_r, B_o), consécutifs sur le génome G_r, s'ils ne sont pas consécutifs sur le génome G_o, ou bien s'ils ne sont pas colinéaires (voir Figure 4.2).

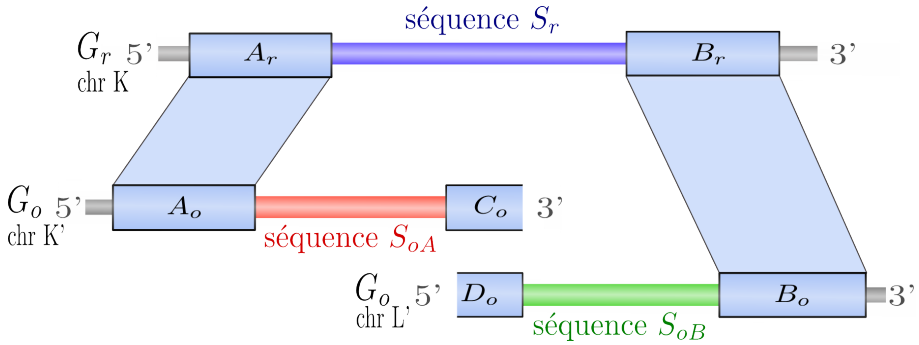


FIG. 4.2: Un exemple de point de cassure et ses séquences d'intérêt S_r, S_{oA} et S_{oB}. Les blocs de synténie (A_r,A_o) et (B_r,B_o) sont consécutifs sur le génome G_r mais ne le sont pas sur le génome G_o. Dans cet exemple, les orientations des deux blocs sont positives. On définit donc un point de cassure sur G_r, dont la séquence S_r est localisée entre les deux blocs (A_r,A_o) et (B_r,B_o). On définit également la séquence S_{oA}, adjacente au bloc (A_r,A_o) sur le génome G_o et bordée par le prochain bloc de synténie (C_r,C_o) sur G_o. La séquence S_{oB} est définie de façon similaire, adjacente au bloc (B_r,B_o) sur G_o et bordée par le bloc précédent (D_r,D_o) sur G_o.

Nous définissons également trois séquences d'intérêt pour affiner le point de cassure : la séquence du point de cassure sur G_r et deux séquences sur G_o qui doivent être en partie orthologues à cette dernière. Afin de faciliter la compréhension, nous avons indiqué l'orientation des brins d'ADN sur le schéma de la Figure 4.2, et nous nous référons à ce schéma pour définir les séquences d'intérêt :

- la séquence S_r est la séquence localisée entre les blocs (A_r, A_o) et (B_r, B_o) sur le génome G_r,

- la séquence S_{oA} est la séquence sur G_o adjacente en 5' au bloc (A_r, A_o) et bordée en 3' par le prochain bloc sur G_o,
- la séquence S_{oB} est la séquence sur G_o adjacente en 3' au bloc (B_r, B_o) et bordée en 5' par le bloc précédent sur G_o.

Cette définition dépend bien entendu de l'orientation des blocs de synténie (A_r, A_o) et (B_r, B_o). Dans l'exemple présenté, ces orientations sont toutes les deux positives. Ainsi, si les blocs sont corrects, l'extrémité 3' du bloc (A_r, A_o) sur G_r doit être orthologue à l'extrémité 3' de ce bloc sur G_o. Si l'orthologie ne s'arrête pas à l'extrémité du bloc, la partie 5' de la séquence S_r devrait être orthologue à la partie 5' de la séquence S_{oA}. Si l'orientation du bloc était négative, ce serait l'extrémité 5' du bloc sur G_o qui serait orthologue à son extrémité 3' sur G_r, et il faudrait choisir la séquence adjacente en 3' au bloc sur G_o comme séquence S_{oA}.

4.1.2 Principe

Ainsi, on s'attend à ce que la partie 5' de la séquence S_r soit orthologue à la partie 5' de la séquence S_{oA}, et que sa partie 3' soit orthologue à la partie 3' de la séquence S_{oB} (voir Figure 4.3). Entre les deux, on aura précisé le point de cassure. L'idée est donc de quantifier la similarité le long de la séquence S_r avec la séquence S_{oA} d'une part, et la séquence S_{oB} d'autre part.

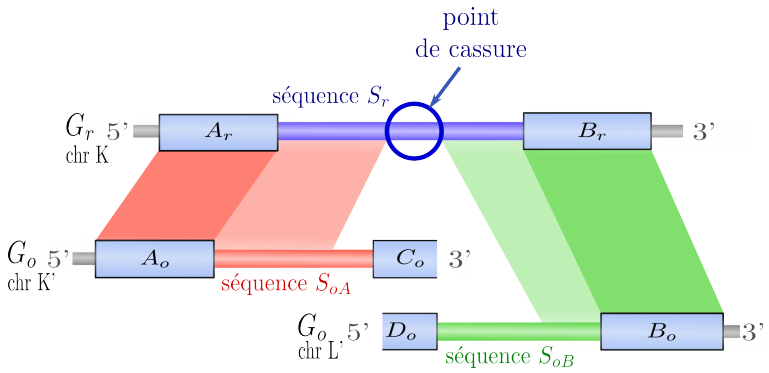


FIG. 4.3: Nous reprenons l'exemple de la Figure 4.2. Nous représentons ici l'orthologie qu'il reste à détecter entre les séquences S_r et S_{oA} d'une part (en rouge) et entre les séquences S_r et S_{oB} d'autre part (en vert). Entre les deux, on aura affiné le point de cassure sur la séquence S_r.

On peut donc distinguer deux étapes à notre méthode :

- la détection de similarité entre S_r et S_{oA}, puis entre S_r et S_{oB}.
- l'identification du point de cassure, grâce à l'information de similarité.

Nous verrons qu'il est nécessaire d'ajouter une troisième étape d'évaluation et de validation du point de cassure obtenu.

La première étape est décrite dans la Section 4.2, la deuxième et la troisième dans la Section 4.3. Cependant, il faut noter que ces deux étapes ont été développées, non pas l'une après l'autre, mais en parallèle. En fait, cette méthode a été développée et améliorée tout au long de la thèse. Nous essaierons donc de montrer les différentes stratégies testées et les motivations des choix qui ont abouti à la méthode actuelle.

4.1.3 Représentation graphique du point de cassure

Avant d'entrer dans les détails de la méthode, nous commençons par présenter un outil important de ce développement méthodologique. Il s'agit d'une représentation graphique du point de cassure et de ses similarités avec les deux séquences orthologues.

Nous disposons ici d'une mesure de similarité entre S_r et S_{oA}, puis entre S_r et S_{oB}. Supposons qu'on puisse colorier chaque position sur la séquence S_r en fonction des similarités mesurées avec les deux séquences. Par la suite, nous adopterons le code couleur suivant : rouge pour ce qui est similaire entre S_r et S_{oA}, et vert pour les similarités entre S_r et S_{oB}. Par exemple, dans le cas de l'utilisation de l'alignement de séquences pour détecter la similarité, nous disposons de la répartition des hits d'alignement (zones de similarités) le long de la séquence S_R avec les séquences S_{oA} et S_{oB} (Figure 4.4).

On s'attend à ce que le début de la séquence contienne beaucoup de positions de couleur rouge (similaires à S_{oA}) et peu de vertes (similaires avec S_{oB}) et inversement pour la fin de la séquence. Ainsi, on devrait pouvoir distinguer deux régions dans la séquence S_R : la partie homologue à S_{oA} et la partie homologue à S_{oB}. Un "bon" point de cassure sera une position sur S_R qui sépare ces deux régions.

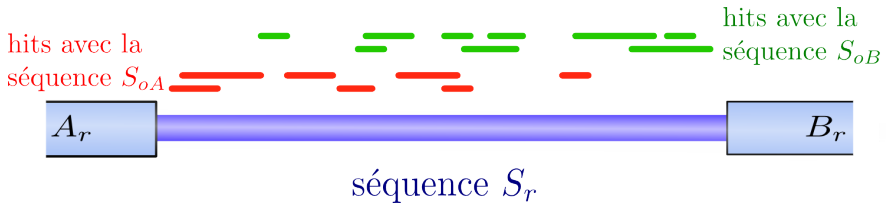


FIG. 4.4: Positionnement des hits d'alignement (ou autre mesure de similarité) le long de la séquence S_r.

On définit une fonction $sc(k)$ qui attribue un score à chaque position k sur I, considérée comme point de cassure potentiel. Le score d'une position k est d'autant plus élevé que les similarités avant et après celle-ci avec, respectivement, S_{oA} et S_{oB} sont fortes et avec, resp., S_{oB} et S_{oA} sont faibles. La fonction $sc(k)$ est définie par :

$$sc(k) = (g(S_{oA}, k) + d(S_{oB}, k)) - (d(S_{oA}, k) + g(S_{oB}, k))$$

avec :

- $g(X, i)$ = nombre de paires de bases de la séquence S_R similaires à la séquence X dont la position sur S_R est inférieure à k (g pour à *gauche* de k).
- $d(X, i)$ = nombre de paires de bases de la séquence S_R similaires à la séquence X dont la position sur S_R est supérieure à k (d pour à *droite* de k).

On compte positivement ce qui correspond à ce qu'on attend d'un point de cassure et on pénalise négativement ce qui y est contraire.

Si les deux régions mentionnées précédemment existent, cette fonction devrait croître dans la première région (homologue à S_{oA}), atteindre un maximum au point de cassure, puis décroître dans la deuxième région (homologue à S_{oB}). Ainsi, la position k pour laquelle $sc(k)$ est maximale représente la rupture entre les deux régions (voir l'exemple de la Figure 4.5). Intuitivement, on a confiance en ce point et on distingue les deux régions d'autant mieux que la croissance à gauche et la décroissance à droite de ce point sont fortes.

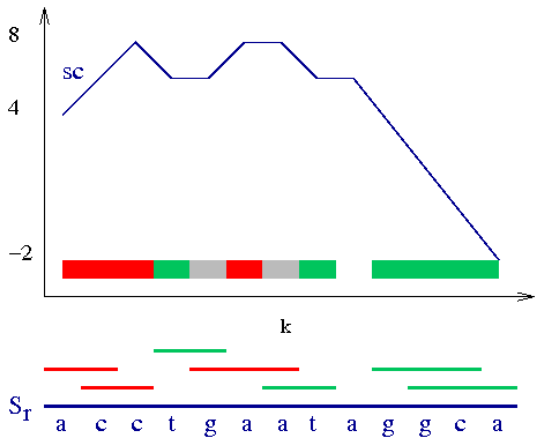


FIG. 4.5: Exemple de représentation graphique de la fonction de score *sc*. Sur le graphique, en dessous de la courbe *sc*, nous avons représenté les positions de la séquence S_r par des barres verticales colorées en fonction des hits qui les couvrent : en rouge (resp. vert) si la position n'est couverte que par des hits d'alignement (ou autre mesure de similarité) avec S_{oA} (resp. S_{oB}), en gris si la position est couverte par des hits des deux séquences et en blanc si la position n'est couverte par aucun hit.

Bien entendu, il sera nécessaire d'évaluer quantitativement ces caractéristiques et de définir une méthode rigoureuse pour définir le point de cassure. Pour l'instant, l'analyse qualitative des points de casure par l'intermédiaire de cette représentation permet d'évaluer la qualité des données de points de cassure et d'identifier les propriétés qui semblent caractériser un "bon" point de cassure. Une autre utilité de cette représentation est qu'elle permet de comparer pour un même point de cassure les différentes stratégies de détection de similarité. Dans ce cas, elle a également été utilisée de manière quantitative, grâce aux valeurs des pentes de part et d'autre du maximum utilisées comme critère de comparaison.

4.2 Detection de similarité

Pour quantifier le long de la séquence S_r la similarité avec les deux séquences S_{oA} et S_{oB}, nous avons testé plusieurs stratégies. Bien sûr, la façon classique de mesurer la similarité entre deux séquences est l'alignement de séquences. Nous avons alors comparé différents programmes d'alignement. Mais nous avons également envisagé une stratégie sans alignement, basée principalement sur le comptage de mots. Enfin, nous nous sommes intéressés à l'influence des éléments répétés présents dans les séquences sur les résultats de l'affinement.

4.2.1 Alignement des séquences

Il existe de nombreux algorithmes d'alignement de séquences nucléiques. On distingue tout d'abord deux types d'alignement : l'alignement global et l'alignement local. L'alignement global de deux séquences a pour objectif d'aligner les deux séquences dans leur intégralité, c'est-à-dire du début à la fin des deux séquences. On utilise ce type d'alignement lorsqu'on sait que les deux séquences à aligner sont homologues dans leur intégralité. Par contre, l'alignement local de deux séquences recherche les sous-séquences les plus similaires entre les deux séquences.

On peut distinguer ensuite les algorithmes exacts des heuristiques. Les premiers renvoient le ou les meilleurs alignements (en fonction d'un système de score). Ces algorithmes sont basés sur la programmation dynamique (Needleman et Wunsch (1970) pour l'alignement global et Smith et Waterman (1981) pour l'alignement local) et leurs complexités en temps et en mémoire sont proportionnelles au produit des tailles des séquences. Des heuristiques ont ensuite été développées pour accélérer les temps de calcul et limiter l'espace mémoire requis, cependant le résultat n'est pas forcément optimal. L'heuristique la plus utilisée pour détecter des similarités locales est Blast (Altschul *et al.*, 1990, 1997) (voir description Section 2.3.1).

Pour détecter des similarités entre les séquences S_r, S_{oA} et S_{oB}, nous avons utilisé un algorithme d'alignement local de type Blast. Plusieurs raisons motivent ce choix.

Tout d'abord, les séquences à aligner peuvent être très longues, certaines de plusieurs millions de nucléotides. Nous ne pouvons donc pas utiliser d'algorithmes exacts à cause de leur complexité algorithmique en temps et en mémoire.

Ensuite, les séquences ne s'alignent *a priori* pas sur toute leur longueur : seulement le début de S_r est censé s'aligner avec le début de S_{oA} (dans l'exemple de la Figure 4.2). Ce n'est donc pas pertinent de chercher un alignement global des séquences. Un autre argument pour utiliser un algorithme d'alignement local, est que les séquences ont peut-être subit des petits réarrangements, ou des insertions et délétions de segments d'ADN. Il existe des algorithmes d'alignement global qui permettent certains réarrangements et duplications (Alves *et al.*, 2005; Brudno *et al.*, 2003b), mais tous les types de réarrangements ne sont pas permis, les duplications, par exemple, ne sont pas autorisées sur les deux séquences dans le cas de sLAGAN (Brudno *et al.*, 2003b), et le temps de calcul peut être long, notamment dans (Alves *et al.*, 2005) qui est un algorithme exact.

Enfin, les espèces comparées peuvent être éloignées phylogénétiquement, et les séquences étant majoritairement intergéniques sont souvent peu similaires. La sensibilité de l'algorithme est donc un facteur important à prendre en compte.

Ainsi, la stratégie que nous avons adoptée est de chercher de petits alignements locaux entre les séquences S_r et S_{oA} d'une part, puis S_r et S_{oB} d'autre part. Nous obtenons deux listes d'alignements locaux (appelés *hits*) avec leurs positions sur les deux séquences alignées. Notons que nous nous intéressons uniquement aux positions des hits sur la séquence S_r, ainsi l'ordre et l'orientation des hits sur la séquence S_{oA} (ou S_{oB}) n'ont pas d'importance.

Nous avons testé plusieurs heuristiques d'alignement local, telles que Blast (Altschul *et al.*, 1997), Blastz (Schwartz *et al.*, 2003), CHAOS (Brudno *et al.*, 2003a), Repseek (Achaz *et al.*, 2007). Ces quatre algorithmes sont basés sur le même principe d'"ancrage-et-extension" (lire la Section 2.3.1 pour une description plus complète). C'est Blastz qui s'est avéré le plus sensible sur les séquences de points de cassure. Ce programme a été conçu pour aligner des génomes entiers et notamment des séquences non codantes peu conservées ; ces résultats confirment une étude montrant que c'est l'algorithme d'alignement le plus sensible à l'heure actuelle sur des séquences non codantes (Sun et Buhler, 2006).

Cependant, il faut noter que ces programmes possèdent un grand nombre de paramètres qui influent sur la sensibilité et la spécificité de l'algorithme. Un exemple de point de cassure dans la Figure 4.6 montre des différences importantes en fonction du paramétrage de l'algorithme Blast (Figures 4.6a et 4.6b). Ainsi, le paramétrage par défaut de Blast n'est pas adapté aux similarités recherchées. La paramétrisation de ces algorithmes n'est pas aisée : elle nécessite une bonne compréhension des différents paramètres, et elle peut être très coûteuse en temps

étant donné le nombre de paramètres à ajuster. Enfin, elle doit être renouvelée pour chaque couple d'espèces comparées, ayant des divergences de séquence différentes.

Or, Blastz présente un avantage sur ce point. Etant très utilisé pour l'alignement de génomes complets, des jeux de paramètres pour différents couples d'espèces sont disponibles (par exemple, sur le site du navigateur de génomes UCSC[1]).

4.2.2 Une alternative à l'alignement : Classus

La façon classique de mesurer la similarité est l'alignement de séquences. Mais l'alignement de séquences fait bien plus que mesurer la similarité : il effectue une correspondance entre les deux séquences alignées en identifiant les nucléotides homologues et en plaçant des indels. Or, dans ce cas précis, nous n'avons pas besoin de connaître précisément la correspondance des nucléotides homologues. Par exemple, on veut quantifier la similarité de la séquence S_r à la position i avec la séquence S_{oA}, et s'il y a similarité avec une région de cette séquence ; on n'a pas besoin de connaître la localisation de cette région sur S_{oA}.

En collaboration avec Pierre Peterlongo de l'université de Marne-la-Vallée, nous avons développé un algorithme, Classus, basé sur le comptage des mots en commun dans les deux séquences comparées. Durant son doctorat, Pierre Peterlongo a développé plusieurs algorithmes de filtrage de séquences nucléiques, Nimbus, ed'Nimbus et Tuiuiu (Peterlongo, 2006; Peterlongo *et al.*, 2008). Le filtrage des séquences est une étape préalable à l'alignement multiple des séquences. L'objectif de cette étape est de sélectionner les régions des séquences les plus similaires. On effectue alors l'alignement multiple seulement sur ces régions, réduisant ainsi considérablement le temps de calcul. Le filtrage des séquences est basé sur le comptage de mots de taille k, appelés k-facteurs, qui sont communs aux séquences à aligner. Grâce à une structure d'indexation appelée *tableau des k-facteurs*, le filtrage des séquences est très rapide.

L'algorithme Classus s'inspire de ces filtres. L'idée est d'utiliser le nombre de k-facteurs communs (appelés k-hits) à deux séquences comme mesure de similarité. Le comptage des k-hits se fait sur des fenêtres glissantes de taille L. Ainsi, on fait glisser une fenêtre le long de la séquence S_r, pour chaque fenêtre on compte le nombre maximum de k-hits avec une fenêtre de l'autre séquence (S_{oA}, par exemple).

On définit F_X l'ensemble des fenêtres de taille $2 \times L$ glissantes avec un pas de L sur la séquence X. Pour une position j sur S_r, on considère la fenêtre $f_j = [j - \frac{L}{2}, j + \frac{L}{2}]$ et pour chaque fenêtre f_k de F_X on compte le nombre d_{jk} de *k-facteurs* présents dans f_j et dans f_k. On définit $n_j(X)$ par :

$$n_j(X) = \max_{f_k \in F_X} (d_{jk})$$

Finalement, on se donne un seuil S, et on définit le score de la position j par rapport à la séquence X, par :

$$s_j(X) = \begin{cases} 1 & \text{si} \quad \max(n_j(X^+), n_j(X^-)) > S \\ 0 & \text{sinon} \end{cases}$$

(X^+ est la séquence X, X^- est son reverse complément).

On obtient pour S_{oA} une séquence de 0 et de 1, de même pour S_{oB}. Ces données sont du même type que celles obtenues avec les hits d'alignement et pourront être utilisées de la même manière (voir sections suivantes) : $s_j(S_{oA}) = 1$ peut être interprété comme le fait que

[1] `http://genome.ucsc.edu/`

la position j est couverte par un hit de S_{oA}.

Le premier avantage de cette méthode est qu'elle est très rapide. Elle est plus rapide que les algorithmes d'alignement Blast et Blastz. Le deuxième est qu'elle ne dispose que de trois paramètres, k, L et S, dont la signification est claire.

Cependant, son inconvénient principal est qu'elle est très sensible à la longueur des séquences comparées. Plus les séquences sont longues, plus il y a de fenêtres testées sur la séquence X, et donc plus le nombre $n_j(X)$ peut être grand par hasard. Or, nous comparons ici les scores obtenus avec les deux séquences S_{oA} et S_{oB}. Une différence importante de longueur de ces deux séquences peut donc biaiser les scores de similarité en faveur de la séquence la plus longue. Il faudrait alors fixer le seuil S en fonction de la longueur des séquences. On pourrait, par exemple, calculer le nombre $n_j(X)$ attendu sous un modèle nul de séquences aléatoires.

Finalement, cette piste n'a pas été approfondie, devant les performances bien meilleures de Blastz (voir Figure 4.6e et 4.6c). Dans cet exemple, Classus donne des résultats plus bruités et présente un maximum différent de ceux obtenus avec les autres algorithmes (Figure 4.6e).

4.2.3 Les éléments répétés

Un facteur important à prendre en compte lorsqu'on compare des séquences au niveau nucléotidique est la présence d'éléments répétés dans les séquences. Plusieurs types d'éléments répétés sont à distinguer. Il y a tout d'abord les *répétitions simples* et les *régions de faible complexité* d'une part, et puis les *éléments transposables* d'autre part.

Les *répétitions simples* sont des répétitions en tandem d'un motif très court. Les *régions de faible complexité* sont des régions où la composition en nucléotides est très biaisée vers un ou deux nucléotides (par exemple, les régions composées à plus de 90 % de nucléotides A). Ces types de séquences peuvent biaiser les méthodes de recherche d'homologie. En effet, lorsque la complexité des séquences est réduite (par exemple, si l'alphabet est réduit à deux lettres), il est plus probable d'observer des séquences très similaires par hasard. Ainsi, dans ce cas, la similarité n'est pas un indice d'homologie de séquence. Une stratégie consiste donc à éliminer ces répétitions et les régions de faible complexité des séquences avant de les comparer ; on appelle cela *masquer* les séquences.

Les *éléments transposables* peuvent également être masqués des séquences. Ce sont des séquences très répétées, c'est-à-dire qu'elles sont présentes en un très grand nombre de copies dans les génomes. Elles sont également *mobiles* : elles ont la capacité de se "déplacer" dans le génome, par des mécanismes de type "couper-et-coller" ou bien "copier-et-coller". On comprend que ces éléments sont problématiques lorsqu'on veut aligner des séquences ou identifier des orthologies de séquences. En effet, lorsqu'on compare deux espèces, les positions de leurs éléments transposables respectifs ne sont pas forcément les mêmes et ne sont pas comparables. Ainsi, des similarités de séquence au niveau d'éléments transposables ne sont pas synonymes d'orthologie de position. Ces éléments représentent environ 45 % du génome humain, ce n'est donc pas négligeable.

Le programme RepeatMasker[2] permet de masquer tous ces types de répétitions. Notamment, pour les éléments transposables, il utilise une base de données des séquences d'éléments transposables connus.

[2] A.F. Smit, R. Hubley et P. Green. `http://repeatmasker.org`

Revenons à notre problème et à nos séquences S_r, S_{oA} et S_{oB}. S'il semble pertinent de masquer les répétitions simples et les régions de faible complexité, la question des éléments transposables est plus difficile. En effet, représentant presque la moitié des séquences (45 % du génome humain), les éliminer serait une perte importante de résolution. De plus, en fonction de la divergence des espèces comparées, certains éléments transposables peuvent être informatifs. Si au moment de la divergence des espèces, ils n'étaient plus actifs (ne se déplaçaient plus), ils ont donc gardé des positions ancestrales dans les séquences des espèces actuelles et peuvent permettre de détecter l'orthologie de position.

Ainsi, en fonction de la divergence des espèces comparées, nous avons masqué ou pas les éléments transposables des séquences avant de les aligner. Notamment, pour la comparaison homme-souris, il est préférable de masquer les séquences. Dans l'exemple de la Figure 4.6, nous obtenons une plus faible couverture de la séquence par les hits d'alignement lorsque celle-ci est masquée. Cependant, les courbes de score sont plus "nettes", avec tous les hits rouges à gauche du maximum et tous les hits verts à droite. On y distingue également plus précisément le point de cassure (voir Figures 4.6c et 4.6d avec l'algorithme Blastz).

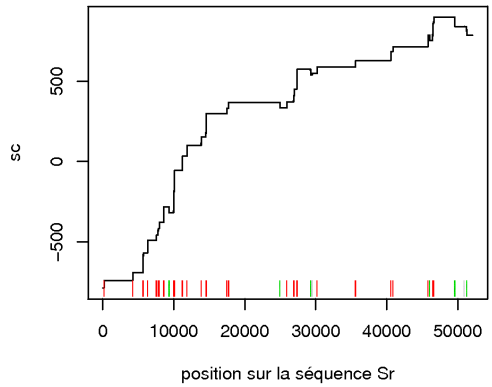
4.2.4 Bilan : choix de Blastz

Nous montrons dans la Figure 4.6, les représentations graphiques de la fonction de score *sc* obtenues pour un même point de cassure entre l'homme et la souris, avec différentes stratégies de détection de similarité. Notons que ce point de cassure est jugé représentatif des points de cassure étudiés entre l'homme et la souris. On peut comparer les formes des courbes et également l'amplitude des variations (axe des ordonnées). On observe notamment sur cette figure que l'algorithme Blastz (Figure 4.6c) donne les meilleurs résultats en terme d'amplitude et de pentes de part et d'autre du maximum. Lorsque les séquences sont masquées, les données sont moins bruitées, avec tous les hits rouges concentrés à gauche du maximum et les verts à droite.

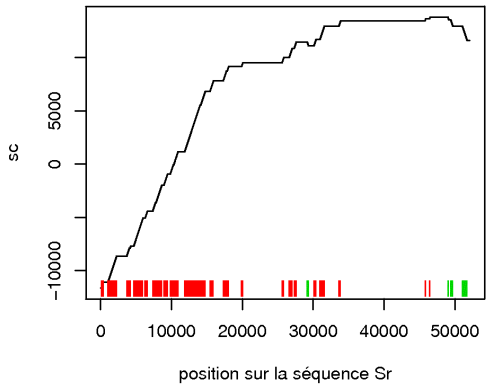
De manière plus générale, nous avons quantifié ces différences en calculant la pente à gauche du maximum, pour un ensemble de 126 points de cassure homme-souris, pour les différentes stratégies comparées. Ces résultats, présentés dans le Tableau 4.1, confirment le choix de Blastz.

	Valeur de la pente à gauche du maximum	
	médiane	moyenne
Blast - paramètres par défaut	0.02	0.045
Classus - k=8, L=200, S=13	0.04	0.143
Blast - paramétré	0.188	0.296
Blastz	0.298	0.473
Blastz - séquences masquées	0.483	0.572

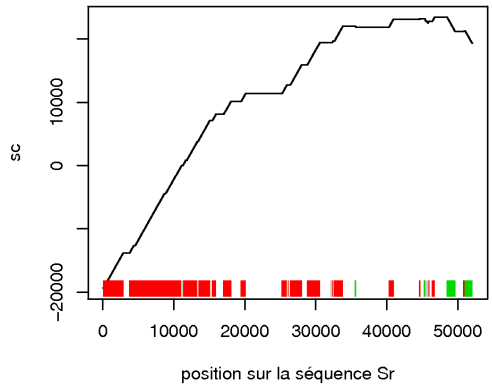
TAB. 4.1: Comparaison des différentes stratégies de détection de similarité pour 126 points de cassure entre l'homme et la souris. Les résultats sont rangés par ordre croissant des valeurs des pentes. Lorsque les séquences sont masquées, la valeur de la pente est calculée sans compter les positions masquées.



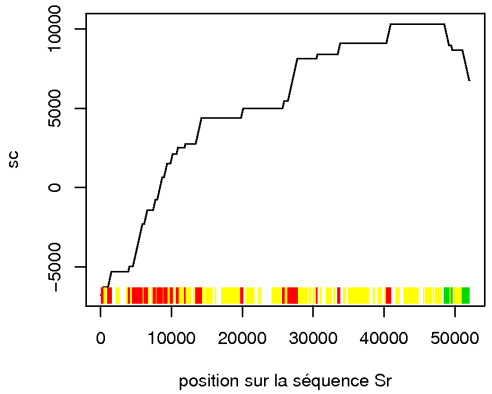
(a) Algorithme d'alignement Blast, paramètres par défaut.



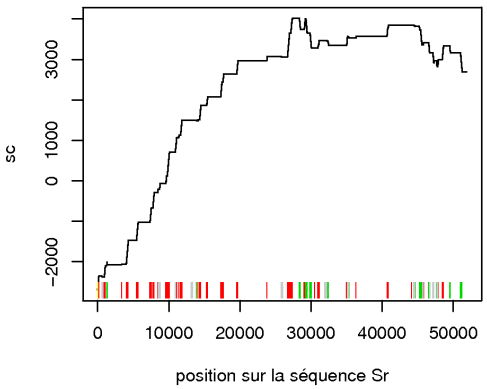
(b) Algorithme d'alignement Blast, paramétré (pénalité d'un mismatch plus faible : par défaut, un match compte +1 et un mismatch -3, ici le mismatch compte pour -0.5).



(c) Algorithme d'alignement Blastz (paramètres par défaut, excepté K=2000).



(d) Algorithme d'alignement Blastz, avec les séquences masquées de tous leurs éléments répétés. Les positions de la séquence S_r qui sont masquées sont représentées en jaune en dessous de la courbe.



(e) Algorithme classus (k=8, L=200 pb et S=13).

FIG. 4.6: Exemples de graphiques obtenus pour un point de cassure sur le génome humain comparé au génome de la souris, avec différentes stratégies de détection de similarité.

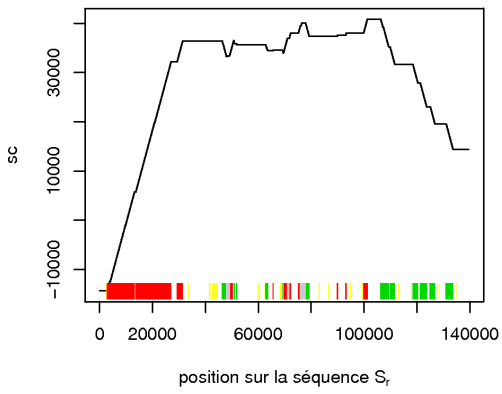
4.3 Méthode de détection quantitative du point de cassure

Nous avons développé une méthode pour déterminer les coordonnées du point de cassure dans la séquence S_r de manière quantitative et automatique, et qui permet également d'évaluer statistiquement la région détectée. Cette méthode est inspirée d'un problème classique en statistiques : le problème de la détection de ruptures dans un signal.

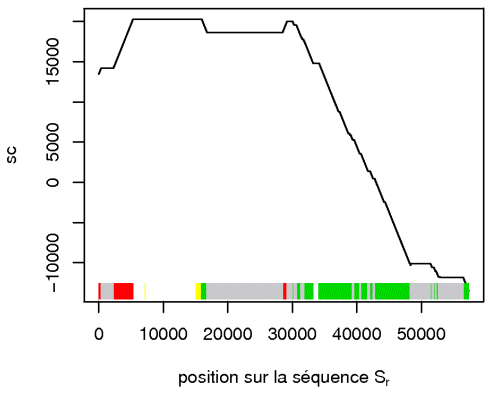
4.3.1 Point ou région de cassure : le plateau

En représentant la fonction *sc* pour les vraies données, on s'aperçoit que très rarement cette fonction a une forme de "pointe". Le plus souvent on observe un "plateau", avec de nombreuses valeurs égales ou proches du maximum.

Il est vrai que les hits d'alignements ne sont pas collés les uns aux autres le long de la séquence S_r ; en moyenne la distance entre deux hits est de quelques kilobases environ. Il ne faut donc pas s'attendre à obtenir une seule position pour laquelle la fonction *sc* est maximale, c'est-à-dire une position telle que la position voisine à gauche est couverte par un hit avec S_{oA} et la position voisine à droite est couverte par un hit avec la séquence S_{oB}. On s'attend plutôt à obtenir un ensemble de positions contiguës pour lesquelles la fonction *sc* est maximale et qui ne sont couvertes par aucun hit. Cependant, les plateaux que nous observons dans les courbes *sc* sont plus complexes : cela peut être des alternances de positions couvertes par des hits rouges puis des verts et ainsi de suite, ou bien des positions couvertes par les deux types de hits. On observe également que, dans le plateau, la densité en hits est plus faible que de part et d'autre (voir deux exemples dans la Figure 4.7).



(a) Exemple avec alternance de hits rouges et verts dans le plateau.



(b) Exemple avec une duplication dans le plateau.

FIG. 4.7: Exemples de graphiques obtenus pour deux points de cassure sur le génome humain comparé au génome du chimpanzé. Deux types de plateaux.

Cela signifie qu'entre le début de S_R qui s'aligne avec S_{oA} et la fin qui s'aligne avec S_{oB}, il existe une région non négligeable qui s'aligne, soit avec aucune des deux séquences, soit avec les deux séquences simultanément.

Cela pose un problème méthodologique : il ne s'agit plus de détecter un point, le maximum de la fonction *sc* par exemple, mais une région où les valeurs de la fonction sont proches du maximum, ou bien où la pente change de signe plusieurs fois. Nous recherchons donc désormais trois régions distinctes dans la séquence S_r.

4.3.2 Méthode de détection de ruptures dans un signal

Les problèmes de détection de ruptures dans un signal consistent à chercher la meilleure partition du signal en segments dans lesquels les caractéristiques du signal sont homogènes dans chaque segment, et différentes d'un segment à l'autre.

Nous présentons ici le cas où le signal est indépendant et gaussien ; la caractéristique étudiée est la moyenne du signal et le nombre de segments est fixé.

a. Modélisation

Soit y le vecteur de longueur n des valeurs observées. On note y_t le t-ème élément de y. On se donne $N < n$ le nombre de segments. On suppose que les données suivent le modèle suivant :

$$y_t = s(t) + \epsilon_t \ , \ t = 1..n$$

On note $\epsilon = (\epsilon_t)$, une suite de variables aléatoires indépendantes et identiquement distribuées selon une loi gaussienne centrée de variance σ^2. La fonction s est supposée constante par morceaux. On note $u_1, u_2, .., u_{N-1}$ les $N-1$ points de rupture de la fonction s (tels que $0 < u_1 < .. < u_j < .. < u_{N-1} < n$ et on pose $u_0 = 0$ et $u_N = n$), et $(s_1, s_2, .., s_N)$ les valeurs prises par s, telle que $s(t) = s_j$ pour $u_{j-1} < t \leq u_j$ (voir Figure 4.8).

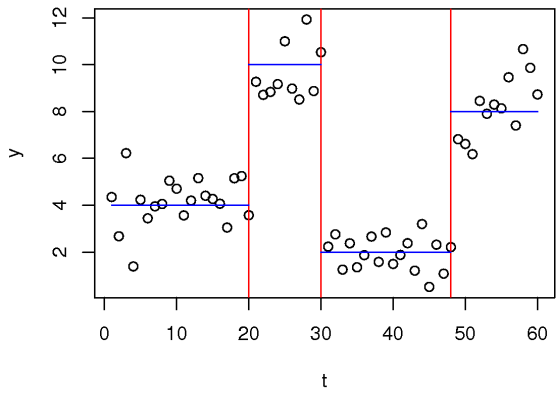


FIG. 4.8: Exemple d'un signal y modélisé par une fonction constante par morceau, s. Le vecteur y est représenté en fonction de t par les ronds noirs. La fonction s est représentée en bleu, avec les points de rupture représentés par des droites rouges verticales.

Le problème est d'estimer la fonction inconnue s, c'est-à-dire estimer les $N-1$ points de rupture et les N valeurs prises entre ces derniers. C'est un problème de partitionnement (un ensemble de $N-1$ points de rupture est une partition en N segments).

Pour une partition donnée, l'estimation des valeurs s_j dans chaque segment n'est pas un problème. Dans le segment j, la valeur s_j est estimée classiquement par minimisation des moindres carrés (ou maximum de vraisemblance) : c'est la moyenne des valeurs observées sur le segment, soit :

$$s_j = \frac{1}{u_j - u_{j-1}} \sum_{t=u_{j-1}+1}^{u_j} y_t$$

On peut alors évaluer une partition par une fonction, appelée fonction de contraste, qui rend compte de la qualité d'ajustement des données au modèle. On utilise en général la somme des carrés des écarts des données au modèle, appelée aussi **erreur quadratique**. On cherchera alors la partition qui minimise l'erreur quadratique.

L'erreur quadratique s'écrit dans ce cas :

$$RSS(u_1, u_2, .., u_{N-1}) = \sum_{j=1}^{N} \sum_{t=u_{j-1}+1}^{u_j} (y_t - s_j)^2$$

Il est donc possible de calculer l'erreur quadratique de toutes les partitions possibles de y en N segments. Cependant le nombre de partitions possibles est C_{n-1}^{N-1} et l'exploration de cet espace serait alors de l'ordre de $O(n^N)$.

b. Programmation dynamique

Un algorithme de programmation dynamique permet de calculer progressivement la meilleure partition, sans évaluer toutes les partitions possibles (Auger et Lawrence, 1989). Il est nécessaire pour cela que la fonction de contraste soit additive par rapport aux contributions de chaque position. La fonction de contraste d'une partition est alors la somme des contributions de chaque segment. C'est le cas pour l'erreur quadratique qui est la somme des carrés des écarts à chaque position.

Cette méthode est basée sur le principe de Bellman : si $\{u_1, u_2, .., u_{N-2}, u_{N-1}\}$ est la meilleure partition en N segments de $y_{1..n}$, alors $\{u_1, u_2, .., u_{N-2}\}$ est la meilleure partition de $y_{1..u_{N-1}}$ en $N-1$ segments (Bellman et Dreyfus, 1962). On va donc procéder progressivement en cherchant les partitions en 1 segment, puis en 2, jusqu'à N segments.

On pose $\Delta(u_{j-1}+1 : u_j)$ la contribution à la fonction de contraste du segment j, et pour $M < v \leq n$:

$$I_M(v) = \min_{u_0=0<u_1<...<u_{M-1}<u_M=v} \sum_{j=1}^{M} \Delta(u_{j-1}+1 : u_j).$$

$I_M(v)$ est la fonction de contraste minimale pour partitionner $y_{1..v}$ en M segments. D'après le principe précédent, on peut décomposer $I_M(v)$:

$$I_M(v) = \min_{u_{M-1}} \{I_{M-1}(u_{M-1}) + \Delta(u_{M-1}+1 : v)\}$$

La valeur qu'on recherche est en fait $I_N(n)$. Pour la calculer, il suffit de calculer les contributions de tous les segments possibles : $\Delta(i : j)$ pour $1 \leq i < j \leq n$ (on a une matrice triangulaire). Ainsi, on dispose des valeurs $I_1(v)$ pour $v = 2, .., n$, puis on obtient $I_2(v) = \min_r\{I_1(r) + \Delta(r+1 : v)\}$ pour $v = 3, .., n$. Et ainsi de suite jusqu'à $I_N(n)$.

La programmation dynamique ne nécessite donc pas l'évaluation de toutes les partitions. Sa complexité est quadratique avec n ($O(n^2)$).

c. Evaluation de la segmentation

Il existe des méthodes pour évaluer statistiquement la partition obtenue, mais elles sont basées principalement sur les différences de moyenne entre les segments. Par contre, il n'existe pas de méthode statistique pour évaluer les positions des points de rupture, ou qui donnerait, par exemple, un intervalle de confiance pour chaque point de rupture.

d. Estimer le nombre de segments

Un problème non trivial est celui de l'estimation du nombre de segments N, qui est souvent inconnu. On pourrait effectuer plusieurs segmentations avec différentes valeurs de N et choisir la meilleure des partitions. Cependant, la comparaison de partitions ayant un nombre de segments différents n'est pas triviale. Si on utilise un critère d'ajustement comme celui adopté pour la segmentation (les moindres carrés ou le maximum de vraisemblance), plus le nombre de segments est grand, meilleures seront les partitions (les modèles sont plus complexes, les données sont donc mieux ajustées). La technique classique est de combiner ce critère d'ajustement avec une pénalité liée au nombre de segments ou au nombre de paramètres ajustés. Cependant la fonction de pénalité utilisée n'est pas aisée à définir et dépend des modèles et des données utilisées.

4.3.3 Adaptation au point de cassure

Le problème qui nous intéresse pour les points de cassure est d'identifier différents segments dans la séquence S_r. Le critère pour différencier ces segments est leur contenu en les différents types de hits (les rouges et les verts). Pour utiliser la méthode de détection de ruptures, nous avons tout d'abord codé ces informations de hits dans une séquence numérique.

a. Codage numérique

L'information de similarité le long de la séquence S_r est codée dans une séquence numérique I. Sa longueur est n, la taille de la séquence S_r. A chaque position k de la séquence S_r correspond une valeur I_k.

Le codage adopté est le suivant :

- I_k vaut 1 si la position k est couverte par au moins un hit avec S_{oA} et aucun hit avec S_{oB},
- I_k vaut -1 si la position k est couverte par au moins un hit avec S_{oB} et aucun hit avec S_{oA},
- I_k vaut 0 dans les autres cas (c'est-à-dire si la position k est couverte par les deux types de hit ou par aucun hit).

Ce codage est similaire à celui utilisé dans la fonction sc : un hit à une position donnée a un poids de 1 ou -1 en fonction de son origine (avec S_{oA} ou S_{oB}). En fait, à une constante près, sc est (le double de) la fonction cumulée de I : $sc(x) = 2\sum_{k=1}^{x} I_k + C$ (avec $C = -\sum_{k=1}^{n} I_k$).

b. Modèle utilisé

On peut appliquer la méthode de détection de ruptures décrite précédemment à la séquence I et identifier des segments de moyennes différentes.

La première question qui se pose est combien de segments cherche-t-on ? On pourrait appliquer l'algorithme de segmentation avec différentes valeurs pour le nombre de segments et sélectionner la meilleure segmentation. On obtiendrait ainsi X segments avec les valeurs des moyennes dans ces segments. Il faudrait alors avoir une méthode qui permette de décider lequel de ces segments est la région de cassure, en fonction de la longueur et des moyennes des segments. Cela peut être vite complexe d'envisager tous les cas de figures dès que le nombre de segments est important. De plus, il faudrait fixer des seuils sur les moyennes pour différencier les différents états qu'on souhaite obtenir : orthologie avec S_{oA}, la région de cassure et l'orthologie avec S_{oB}. Nous n'avons fait que déplacer le problème ; il est simplifié certes, mais l'identification de la région de cassure n'est pas immédiate.

Pour éviter ce problème, nous allons utiliser l'information connue *a priori* pour segmenter la séquence I. On s'attend à ce que la séquence I soit divisée en trois segments (ou moins, si l'un des trois est nul). Le premier doit avoir une moyenne positive, puisqu'il représente l'orthologie avec S_{oA}, et doit donc contenir plus de hits rouges que de hits verts. De même, le troisième doit avoir une moyenne négative. En ce qui concerne le segment du milieu, on ne sait pas très bien quelle propriété il doit avoir, mais pour le différencier des deux autres, on le modélise par la valeur nulle. Ainsi, on fixe le modèle dans le segment du milieu et on impose des contraintes de signe aux deux autres segments.

Plus formellement, le modèle s'écrit (en conservant les notations précédentes) :

$$I_t = s(t) + \epsilon_t \ , \ t = 1..n$$

La fonction s est une fonction constante par morceaux définie par deux points de rupture (u_1, u_2) (tels que $0 = u_0 < u_1 < u_2 < u_3 = n$), telle que $s(t) = s_j$ pour $u_{j-1} < t \leq u_j$. Les valeurs (s_1, s_2, s_3) sont définies par :

- $s_1 = \begin{cases} \frac{\sum_{t=1}^{u_1} I_t}{u_1} & \text{si} \quad \sum_{t=1}^{u_1} I_t > 0 \\ \infty & \text{sinon} \end{cases}$
- $s_2 = 0$
- $s_3 = \begin{cases} \frac{\sum_{t=u_2+1}^{n} I_t}{n-u_2} & \text{si} \quad \sum_{t=u_2+1}^{n} I_t < 0 \\ \infty & \text{sinon} \end{cases}$

On cherche alors le couple de positions (u_1, u_2) qui minimisent la fonction de contraste (erreur quadratique) :

$$RSS(u_1, u_2) = \sum_{j=1}^{3} \sum_{t=u_{j-1}+1}^{u_j} (I_t - s_j)^2 = \sum_{t=1}^{u_1} (I_t - s_1)^2 + \sum_{t=u_1+1}^{u_2} I_t^2 + \sum_{t=u_2+1}^{n} (I_t - s_3)^2 \quad (4.1)$$

c. Algorithme

Autoriser moins de trois segments Pour autoriser moins de 3 segments, il suffit d'autoriser 0 pour la valeur u_1, et la valeur n pour u_2, et également u_1=u_2. Dans le premier cas, le premier segment est nul ; dans le deuxième, le segment 3 est nul et dans le troisième cas, le segment du milieu est nul.

Réduction de la complexité Le modèle étant plus contraint que dans le cas général, la recherche des points de rupture est simplifiée. Il s'agit en fait de maximiser deux sommes indépendamment, ce qui réduit la complexité algorithmique à $O(n)$, au lieu de $O(n^2)$ avec la programmation dynamique (cf. Lemme 1).

Lemma 1. *Etant donné la séquence I de taille n, telle que pour tout $k \in \{1, n\}, I_k \in \{-1, 0, 1\}$, les positions u_1 et u_2, avec $u_1 \leq u_2$, qui minimisent la fonction de contraste $RSS(u_1, u_2)$ (voir la formule (4.1)) sont tels que :*

- $\frac{1}{\sqrt{u_1}} \sum_{t=1}^{u_1} I_t$ *est maximal,*
- $\frac{1}{\sqrt{n-u_2}} \sum_{t=u_2+1}^{n} I_t$ *est minimal.*

Preuve :

Si on développe les termes carrés de la formule 4.1, on obtient la fonction suivante à minimiser :

$$RSS(u_1, u_2) = \sum_{t=1}^{n} I_t^2 - \frac{1}{u_1}(\sum_{t=1}^{u_1} I_t)^2 - \frac{1}{n-u_2}(\sum_{t=u_2+1}^{n} I_t)^2$$

On voit que le premier terme est indépendant de u_1 et de u_2, et les deux autres termes sont indépendants l'un de l'autre. Ainsi l'erreur quadratique est minimale lorsque ces deux termes sont maximaux. On peut donc trouver u_1 qui maximise $rss_1(u_1) = \frac{1}{u_1}(\sum_{t=1}^{u_1} I_t)^2$ et trouver u_2 indépendamment en maximisant $rss_2(u_2) = \frac{1}{n-u_2}(\sum_{t=u_2+1}^{n} I_t)^2$. Cependant, la solution doit respecter la condition $u_1 \leq u_2$. Nous montrons ici que cette condition est toujours respectée.

On note $A(u_1) = \frac{1}{\sqrt{u_1}} \sum_{t=1}^{u_1} I_t$ et $B(u_2) = \frac{1}{\sqrt{n-u_2}} \sum_{t=u_2+1}^{n} I_t$. Maximiser rss_1 revient à maximiser A (puisque $A > 0$) et maximiser rss_2 revient à minimiser B (puisque $B < 0$).

Soit x_1 (resp. x_2) la position sur I qui maximise $A(u_1)$ (resp. minimise $B(u_2)$). Supposons que $x_1 > x_2$. Alors :

$A(x_1) = \frac{1}{\sqrt{x_1}} \sum_{t=1}^{x_1} I_t = \frac{1}{\sqrt{x_1}} \sum_{t=1}^{x_2} I_t + \frac{1}{\sqrt{x_1}} \sum_{t=x_2+1}^{x_1} I_t$

$B(x_2) = \frac{1}{\sqrt{n-x_2}} \sum_{t=x_2+1}^{n} I_t = \frac{1}{\sqrt{n-x_2}} \sum_{t=x_2+1}^{x_1} I_t + \frac{1}{\sqrt{n-x_2}} \sum_{t=x_1+1}^{n} I_t$

- si $\sum_{t=x_2+1}^{x_1} I_t \geq 0$, alors $B(x_2) \geq \frac{1}{\sqrt{n-x_2}} \sum_{t=x_1+1}^{n} I_t > \frac{1}{\sqrt{n-x_1}} \sum_{t=x_1+1}^{n} I_t = B(x_1)$, donc $B(x_2)$ n'est pas minimal.
- sinon ($\sum_{t=x_2+1}^{x_1} I_t < 0$), $A(x_1) < \frac{1}{\sqrt{x_1}} \sum_{t=1}^{x_2} I_t < \frac{1}{\sqrt{x_2}} \sum_{t=1}^{x_2} I_t = A(x_2)$, donc $A(x_1)$ n'est pas maximal.

L'hypothèse de départ $x_1 > x_2$ n'est donc pas possible, et on a $x_1 \leq x_2$. □

Implémentation L'algorithme de détection du point de cassure à partir d'une séquence I de $\{0, 1, -1\}$ a été implémenté dans une fonction R.

d. Test de significativité

Quelle que soit la séquence I et sa structure, la méthode va produire la meilleure segmentation en trois segments (ou moins). Il est alors important de tester si les données sont effectivement structurées en trois segments, respectant les contraintes mentionnées ci-dessus, ou bien s'il n'existe pas de telle structure dans ces données. Dans ce dernier cas, les points de rupture obtenus n'ont pas de sens et nous devons conclure que nous ne sommes pas capables d'affiner le point de cassure sur la base des alignements.

Plus la séquence I est structurée en trois segments respectant les contraintes mentionnées ci-dessus, plus la valeur de la fonction de contraste minimisée sera faible : l'ajustement sera meilleur. Nous devons donc tester si cet ajustement est significativement meilleur que celui que nous obtiendrions avec une séquence non structurée.

Nous créons des séquences non structurées à partir de la séquence I en permutant les valeurs de cette dernière, et pour chaque séquence permutée nous calculons son ajustement au modèle, c'est-à-dire la valeur de la fonction de contraste 4.1 minimisée. Puisque la séquence I représente des hits d'alignement, les positions ne sont pas indépendantes les unes des autres et les valeurs de 1 et de -1 apparaissent regroupées. Il est important de prendre en compte cette structuration dans la procédure de permutation. Au lieu de permuter les positions de I individuellement, nous permutons les blocs de valeurs identiques consécutives données par les extrémités des hits.

Nous acceptons l'hypothèse nulle que I n'est pas structurée en trois segments respectant les contraintes du modèle si plus de 5 % des séquences permutées ont une valeur d'ajustement plus faible que celle obtenue pour I.

4.3.4 Autres modèles envisagés et perspectives

a. Le problème des "gaps"

L'efficacité de la méthode de segmentation dépend des caractéristiques du signal I, c'est-à-dire de la densité en hits et du niveau de bruit (répartition des hits rouges et verts). La stratégie de détection de similarité adoptée est de masquer les séquences de tous leurs éléments répétés et d'utiliser des paramètres d'alignement qui limitent le bruit. On a ainsi une bonne ségrégation des deux types de hits (rouges et verts) le long de la séquence S_r. En contre partie, les hits d'alignement sont moins nombreux et plus espacés. Notons ici que lorsque les séquences sont masquées, les positions masquées ne sont bien évidemment pas prises en compte dans la segmentation. Ces espaces sans hit, que nous appelons "gaps", peuvent poser problème lors de la segmentation. Lorsqu'ils sont à proximité du point de cassure, ils ont tendance à être ajoutés au segment du milieu, ce qui agrandit le point de cassure. En effet, ils sont souvent beaucoup plus grands que les hits, et ont donc plus de poids dans l'erreur quadratique à minimiser (voir un exemple dans la Figure 4.9). Ce problème est assez fréquent, car on observe une diminution de la densité en hits plus on se rapproche du point de cassure (voir Chapitre suivant).

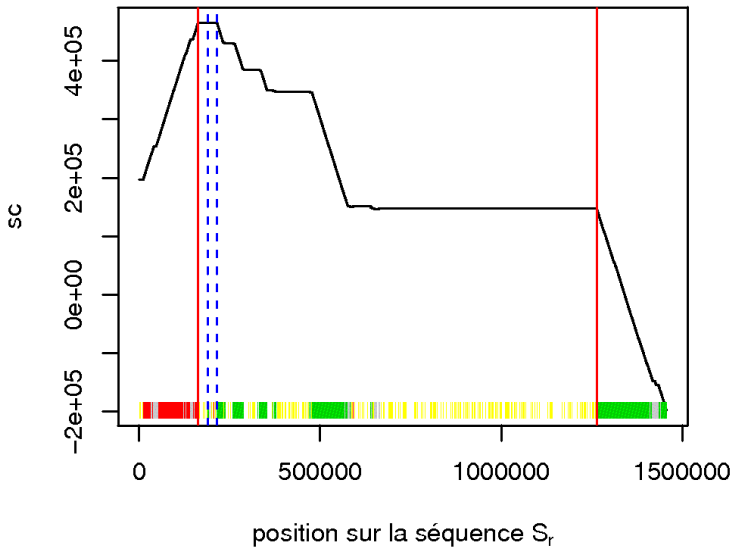


FIG. 4.9: Exemple de point de cassure homme-chimpanzé avec un problème de gap. Une grande région sans hit, aux environs de 1 Mb sur la séquence S_r, entraîne une segmentation peu satisfaisante (les points de rupture sont désignés par les droites verticales en rouge) lorsque les positions couvertes par aucun hit sont comptées. Les droites en pointillés verticales en bleu indiquent les résultats de la segmentation lorsque les positions couvertes par aucun hit ne sont pas prises en compte.

La solution adoptée pour éviter ce problème est de ne pas prendre en compte les positions sans hit dans la segmentation. On construit la séquence I' correspondant à la séquence I excisée des positions sans hit. La longueur de I' est donc le nombre de positions de S_r couvertes par au moins un hit. Cette solution est radicale, mais elle donne de meilleurs résultats dans la majorité des cas. Cependant, dans d'autres cas, elle peut conduire à affiner plus qu'on ne

devrait (par rapport à la représentation de la fonction de score). En fait, il suffit d'un petit hit dans une grande région sans hit pour déplacer les bornes du point de cassure sur de longues distances. Ainsi, cette solution est "risquée" s'il existe des faux positifs au sein des hits d'alignements utilisés. C'est pourquoi nous masquons les séquences de leurs éléments répétés et nous paramétrons l'algorithme d'alignement de façon à minimiser le nombre de faux positifs (augmenter la spécificité) sans perdre trop en sensibilité.

Une meilleure prise en compte des gaps est l'une des perspectives d'amélioration de cette méthode. On pourrait envisager, par exemple, une pondération des gaps dépendant de leur longueur, par analogie avec les pondérations affines des gaps dans les algorithmes d'alignements de séquences.

Ce problème des "gaps" illustre bien la dépendance entre les deux étapes de la méthode : la détection de similarité et la segmentation. En effet, la stratégie de détection de similarité influe sur les caractéristiques des hits obtenus et sur leur distribution le long de la séquence à segmenter. Or, les méthodes de segmentation sont plus ou moins efficaces en fonction de ces caractéristiques. Ces deux étapes ne peuvent donc pas être développées indépendamment et il est nécessaire de faire des aller-retours entre les deux. Par exemple, le problème des "gaps" ne se pose pas lorsque les séquences ne sont pas masquées, puisque la densité en hits est plus importante ; par contre, l'affinement est moins efficace car le signal est plus bruité. Ainsi, la modélisation choisie est adaptée si le signal est peu bruité et les hits d'alignements sont très spécifiques.

b. Les autres modèles testés

Nous venons de présenter la méthode dans son état final et abouti (telle qu'elle est utilisée et publiée (Lemaitre *et al.*, 2008b)), mais avant d'en arriver là, nous avons envisagé et testé d'autres méthodes et modèles pour segmenter la séquence I. Plusieurs aspects de la méthode ont évolué, comme par exemple le codage numérique de I, le modèle de segmentation, ou encore l'évaluation de la segmentation. Par exemple, nous avons d'abord envisagé de modéliser les ruptures de pentes de la fonction sc avant de s'intéresser aux données brutes de contenus en hits. On peut également mentionner la prise en compte de la qualité des hits (niveau de similarité) dans le codage numérique de I, et l'évaluation des segmentations par les différences de moyennes des segments.

Nous nous attardons plus longuement sur les aspects de modélisation. Notons que, ici, les "gaps" ou positions sans hit sont pris en compte.

Le modèle utilisé est assez contraint puisque nous imposons des moyennes positives et négatives dans les segments 1 et 3 respectivement et nous fixons la valeur du segment central à 0. Nous avons envisagé de contraindre encore plus le modèle, en fixant les valeurs de la fonction constante par morceaux dans chaque segment. Nous avons tout d'abord fixé les valeurs suivantes : $s_1 = 1$, $s_2 = 0$ et $s_3 = -1$. Ce sont les valeurs qu'on attend dans le meilleur des cas où le segment 1 est entièrement couvert par des hits rouge et aucun hit vert et inversement pour le segment 3. Cette modélisation s'est avérée trop stringente. Les régions de cassure obtenues sont très souvent très grandes avec les segments 1 et 3 réduits au minimum. En effet, il y a une asymétrie au niveau des contributions de chaque position à l'erreur quadratique en fonction du segment dans lequel elle se trouve. Dans le segment central, la contribution d'une position est soit 1 ou 0, alors que dans les segments extrêmes

elle peut être 0, 1 ou 4 ($4 = (-1-1)^2 = (1-(-1))^2$, lorsque $I_t = -1$ dans le segment 1, ou $I_t = 1$ dans le segment 3). Ainsi, la pénalité d'un "mauvais hit" (par exemple un hit vert dans le segment 1 ou 2) est 4 fois plus grande dans les segments 1 et 3 que dans le segment du milieu. De plus, cela impose au segment 1 (resp. 3) d'avoir une moyenne supérieure à 0.5 (resp. inférieure à -0.5). Ce seuil de 0.5 n'est pas justifié et n'est pas adapté à toutes les comparaisons. Par exemple, lorsque les espèces comparées sont éloignées, la densité en hits peut être faible, souvent inférieure à 50 %.

On pourrait alors envisager de fixer les valeurs de s_1 et s_3 en fonction de la divergence des séquences. Pour estimer la moyenne des segments attendue dans le cas de séquences homologues, il faudrait aligner des séquences du même couple d'espèces dont on est sûr qu'elles sont homologues. Cette estimation est coûteuse en temps et elle doit être répétée pour chaque couple d'espèces comparées. De plus, le niveau de similarité des séquences varie le long du génome, notamment en fonction de la densité en gènes. Enfin, nous montrons dans le chapitre suivant, que dans les régions de cassure et ses régions voisines, le niveau de similarité n'est pas représentatif des autres régions du génome, il est plus faible. Ainsi, il est peu probable de disposer d'une estimation de la moyenne attendue des segments qui soit adaptée à chaque point de cassure.

Une autre stratégie testée est de réduire la région de cassure d'un côté, puis de l'autre de façon indépendante. On effectue alors deux segmentations indépendantes, une pour la similarité avec S_{oA} et l'autre pour S_{oB}. Dans ce cas, les deux segmentations sont plus simples, il s'agit de chercher deux segments, l'un avec beaucoup de hits, l'autre avec très peu. On dispose donc de deux séquences I_A et I_B contenant chacune des 0 et des 1 (0 pour les positions sans hit, 1 pour celles avec hit). On recherche un point de rupture pour chacune d'elles. Par exemple, I_A est modélisée par une fonction constante par morceaux avec deux segments (on note son point de rupture x_1), et on impose la contrainte que le premier segment doit avoir une moyenne plus grande que le deuxième (c'est l'inverse pour I_B, on note son point de rupture x_2). On considère que le segment de la position 1 à x_1 est homologue à S_{oA} et que celui de x_2 à n est homologue à S_{oB}. La région de cassure affinée est alors délimitée de chaque côté par le point de rupture de chaque segmentation. L'avantage de cette méthode est qu'elle ne modélise pas la région de cassure en tant que telle. C'est un avantage car on ne sait pas très bien *a priori* si cette région existe dans tous les cas et quelles sont ses caractéristiques. Cependant, il reste à distinguer deux cas : le cas où $x_1 < x_2$, les deux régions homologues à S_{oA} et S_{oB} ne se chevauchent pas, et le cas où $x_1 > x_2$, la région de cassure est homologue aux deux séquences S_{oA} et S_{oB}.

Cette méthode n'a pas donné de bons résultats. Dans le cas où les données sont assez bruitées, les ruptures détectées sont souvent non significatives (test de significativité similaire à celui décrit dans la section précédente), alors que la fonction de score montre clairement que la région de cassure peut être affinée. Même lorsque les données ne sont pas bruitées avec les hits rouges et verts bien séparés, les ruptures obtenues ne correspondent pas systématiquement à la région de cassure observée avec la fonction de score. Par exemple, si la séquence contient des régions plus conservées que d'autres (par exemple des gènes), cela peut générer des ruptures qui ne correspondent pas au point de cassure. Il semble donc préférable de prendre en compte simultanément les deux types d'information de similarité (avec S_{oA} et avec S_{oB}).

c. Modèle multinomial

Un autre type de modélisation qui n'a pas été testé est l'utilisation de modèles multinomiaux. Dans la modélisation présentée ici, la position t de la séquence I est supposée être une valeur numérique émise par une loi normale. On ne prend pas en compte le fait que la séquence I est composée de valeurs discrètes : les seules valeurs possibles pour I_t sont 0, 1 et -1. En fait, on peut modéliser I_t par un modèle mutlinomial défini par les probabilités de chaque état 0, 1 et -1. On peut donc imaginer que les positions au sein d'un même segment suivent un même modèle, mais les modèles diffèrent entre segments (avec des paramètres différents) :

$$I_t \sim \mathcal{M}(1, \theta_j) \text{ pour } u_{j-1} < t \leq u_j$$

avec $\mathcal{M}(1, \theta_j)$ le tirage dans une loi multinomiale de paramètres $\theta_j = (p_j^-, p_j^0, p_j^+)$ telle que :

$$\begin{cases} \mathbb{P}(I_t = -1) = p_j^- \\ \mathbb{P}(I_t = 0) = p_j^0 \\ \mathbb{P}(I_t = 1) = p_j^+ \end{cases}$$

On peut calculer pour un modèle donné (c'est-à-dire une partition et les paramètres des modèles dans chaque segment), la vraisemblance de I, c'est-à-dire la probabilité d'obtenir la séquence I sachant le modèle θ :

$$\mathcal{V}(I, \theta) = \prod_{j=1}^{N} \prod_{t=u_{j-1}+1}^{u_j} \prod_{\alpha \in \{-1,0,1\}} (\mathbb{P}(I_t = \alpha))^{\mathbb{I}(I_t = \alpha)}$$

avec $\mathbb{I}$ la fonction indicatrice :

$$\mathbb{I}(I_t = \alpha) = \begin{cases} 1 & \text{si} \quad I_t = \alpha \\ 0 & \text{sinon} \end{cases}$$

De la même manière que pour la détection de ruptures dans la moyenne, on estime les paramètres du modèle dans chaque segment par les fréquences de 0, 1 et -1 et on cherche la partition qui maximise la vraisemblance de I. Avec la transformation logarithmique, les contributions de chaque position sont additives et on peut utiliser l'algorithme de programmation dynamique décrit précédemment pour trouver la meilleure partition. On a :

$$\log(\mathcal{V}(I, \theta)) = \sum_{j=1}^{N} \sum_{t=u_{j-1}+1}^{u_j} \sum_{\alpha \in \{-1,0,1\}} \mathbb{I}(I_t = \alpha) \log(\mathbb{P}(I_t = \alpha))$$

$$\log(\mathcal{V}(I, \theta)) = \sum_{j=1}^{N} \sum_{\alpha \in \{-1,0,1\}} N_j(\alpha) \log(p_j^\alpha)) \qquad \text{avec} \quad N_j(\alpha) = \sum_{t=u_{j-1}+1}^{u_j} \mathbb{I}(I_t = \alpha)$$

Cette modélisation a l'avantage de prendre en compte le caractère discret des données. De plus, on peut envisager dans ce cas d'ajouter un quatrième état. Dans le cas précédent, la valeur 0 peut signifier deux choses : soit la position est couverte par aucun hit, soit elle est couverte par des hits de "couleurs" différentes. Ici, on peut différencier ces deux types de 0, en créant deux états ayant des probabilités différentes.

Cette modélisation n'a pas été implémentée, ni testée. Il reste encore à déterminer comment contraindre les paramètres des modèles dans chaque segment afin d'obtenir les 3 segments voulus (homologie avec S_{oA}, point de cassure et homologie avec S_{oB}). Par exemple, dans le segment 1, on pourrait imposer que $p_1^+ > p_1^-$.

4.4 Affinement des points de cassure de mammifères

Nous avons appliqué la méthode sur les points de cassure de mammifères. Avant de présenter les résultats obtenus et les comparaisons effectuées avec d'autres méthodes, nous discutons des données qu'il faut fournir en entrée à cette méthode.

4.4.1 Prérequis sur les blocs de synténie et délimitations des séquences

Cette méthode d'affinement des points de cassure s'applique à des génomes entièrement séquencés, pour lesquels on a identifié les blocs de synténie. Cependant, plusieurs propriétés des blocs de synténie sont requises pour la réussite de cette étape.

Tout d'abord, les blocs ne doivent pas se chevaucher sur l'un ou l'autre génome, puisque les séquences S_r, S_{oA} et S_{oB} sont définies par les extrémités de blocs consécutifs (séquences inter-blocs).

Ensuite, les extrémités des blocs doivent être des séquences orthologues. Si l'extrémité du bloc (A_r, A_o) sur G_r n'est pas orthologue à l'extrémité du bloc (A_r, A_o) sur G_o, alors il est peu probable que les parties 5' de S_r et de S_{oA} soient orthologues. Cela peut arriver avec certaines méthodes de construction de blocs de synténie qui n'interdisent pas l'existence de conflits au sein des blocs (voir Section 2.2.3). C'est la raison pour laquelle nous avons développé notre propre méthode pour construire des blocs de synténie non chevauchants et sans conflit (décrite dans le Chapitre 3).

Un exemple de conflit à l'extrémité du bloc (A_r, A_o) est représenté dans la Figure 4.10. Dans ce cas, l'extrémité 5' de la séquence S_r est orthologue à une séquence comprise à l'intérieur du bloc de synténie (A_r, A_o) sur G_o, qui n'appartient donc pas à la séquence S_{oA}. En conséquence, on ne pourra détecter l'orthologie sur cette partie de la séquence S_r, ce qui peut empêcher d'affiner efficacement le point de cassure.

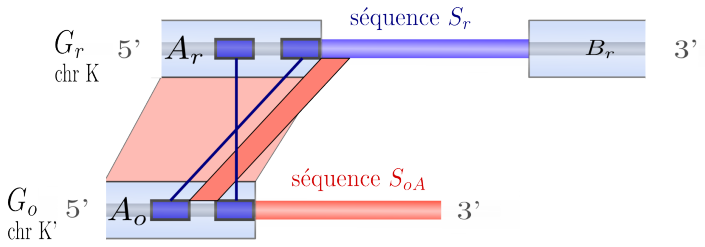


FIG. 4.10: Exemple de conflit à l'extrémité d'un bloc de synténie et ses conséquences sur l'étape d'affinement. Le parallélogramme rouge représente l'orthologie de séquence entre l'extrémité 5' de la séquence S_r et une séquence interne au bloc A_o.

Avec des blocs sans conflit, l'extrémité d'un bloc est représentée par un seul marqueur orthologue. Il reste alors à vérifier que l'extrémité de ce marqueur correspond à des nucléotides orthologues dans les deux génomes. Dans le cas de gènes orthologues, pris comme marqueurs, cette propriété n'est pas toujours respectée. En effet, l'orthologie des gènes est assignée par

alignement des séquences protéiques, or les séquences protéiques ne s'alignent peut-être pas sur toute leur longueur et les coordonnées des gènes sur le génome ne correspondent pas forcément aux coordonnées des séquences protéiques alignées. Ainsi, on ne peut pas exclure l'hypothèse que le point de cassure se trouve à l'intérieur du gène, à l'extrémité du bloc.

C'est pour cette raison que lorsque les blocs de synténie sont définis par des gènes orthologues, nous préférons inclure les gènes aux extrémités de chaque bloc dans les séquences S_r, S_{oA} et S_{oB} (voir Figure 4.11).

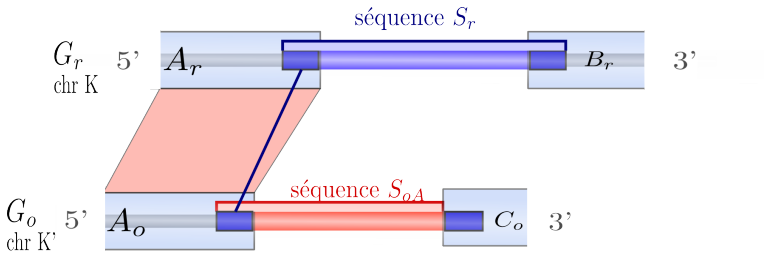


FIG. 4.11: Exemple de comment les gènes orthologues aux extrémités des blocs de synténie sont ajoutées dans les séquences S_r et S_{oA}.

A titre d'exemple, 40 points de cassure homme-souris, sur un total de 354 (soit 11 %) chevauchent l'un des deux gènes orthologues aux extrémités de S_r après affinement. Le chevauchement est le plus souvent faible (moins de 10 % de la longueur totale du gène dans la majorité des cas).

4.4.2 Application à plusieurs couples d'espèces

Nous avons appliqué cette méthode pour détecter les points de cassure sur le génome de l'homme, comparé avec les autres génomes de mammifères séquencés : souris, rat, chien, macaque et chimpanzé.

Les blocs de synténie utilisés ont été obtenus avec la méthode de construction de blocs de synténie que nous avons développée (voir Section 3.2). Nous avons utilisé les gènes orthologues 1-1 prédits par Ensembl pour construire les blocs de synténie avec le paramètre k fixé à deux (Hubbard *et al.*, 2007). Tous les points de cassure détectés ont été affinés, excepté ceux qui contenaient un centromère humain.

En fonction des espèces comparées, nous avons masqué plus ou moins les séquences et utilisé des paramètres d'alignement différents. Ces choix sont basés sur les paramètres utilisés dans la littérature (par exemple, la paramétrisation de Blastz pour l'alignement des génomes complets [3]), et sur l'analyse qualitative des points de cassure avec la fonction de score.

Pour les cinq comparaisons, la très grande majorité des points de cassure ont été affinés efficacement (voir Tableau 4.2). La taille des points de cassure a été réduite en moyenne de plusieurs centaines de Kilobases. La Figure 4.12 représente l'histogramme des tailles de points de cassure avant et après affinement pour la comparaison homme-souris. Nous obtenons, pour les cinq comparaisons, des points de cassure très résolus, dont plus de la moitié ont une taille finale inférieure à 50 Kb.

[3] `http://genome.ucsc.edu`

Seuls quelques points de cassure (5 en tout) n'ont pas pu être affinés, le test de permutation étant non significatif. Ces cas correspondent à des points de cassure pour lesquels l'intégralité de la séquence S_r s'aligne avec les deux séquences S_{oA} et S_{oB}.

Points de cassure	Effectif (non significatif)	Taille médiane avant affinement	Taille médiane après affinement	Réduction moyenne
homme - chimpanzé	36 (1)	228 Kb	22 Kb	493 Kb
homme - macaque	96 (0)	292 Kb	33 Kb	360 Kb
homme - chien	240 (0)	242 Kb	33 Kb	349 Kb
homme - souris	355 (1)	268 Kb	51 Kb	371 Kb
homme - rat	421 (3)	275 Kb	50 Kb	438 Kb

Tab. 4.2: Résumé des résultats d'affinement des points de cassure pour les cinq comparaisons. La deuxième colonne indique le nombre total de points de cassure qui ont été affinés avec, entre parenthèses, le nombre de points de cassure dont le test de permutation est non significatif.

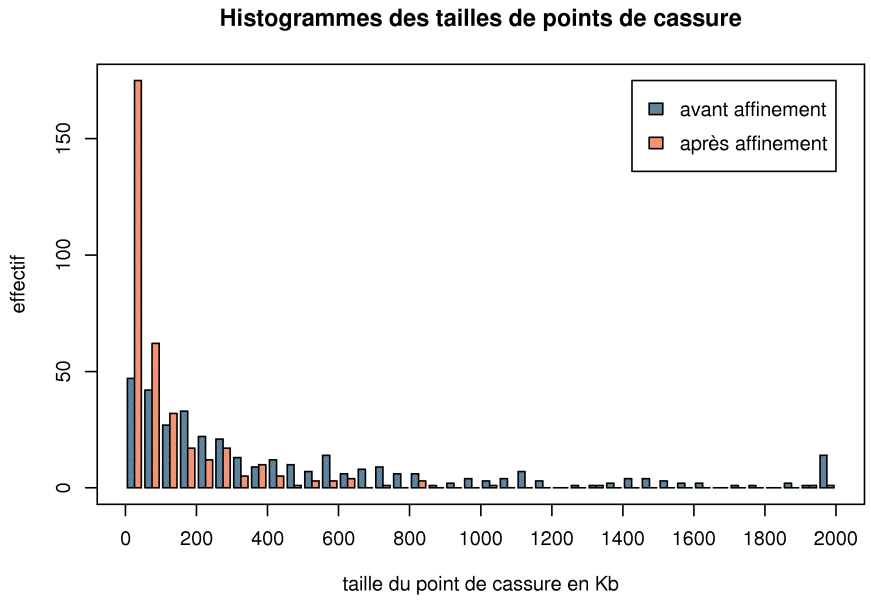


Fig. 4.12: Comparaison des tailles de points de cassure homme-souris, avant (en bleu) et après (en orange) l'étape d'affinement. L'intervalle à 2000 Kb regroupe toutes les valeurs supérieures à 2000 Kb.

4.4.3 Comparaison avec d'autres méthodes

Il est normal d'arriver à étendre les blocs de synténie au delà des gènes : les espèces comparées ne sont pas si éloignées que seuls les gènes seraient alignables et il n'y a pas de raison pour que les blocs de synténie s'arrêtent aux bordures des gènes. Pour juger de l'efficacité de la méthode, nous avons comparé les résultats avec ceux de méthodes ayant une résolution équivalente, en l'occurrence les méthodes d'alignement de génomes complets (voir

Section 2.3).

Données de points de cassure disponibles La majorité des données de points de cassure de mammifères disponibles publiquement portent sur la comparaison homme-souris. Ainsi, lors de la publication de la méthode GRIMM, les blocs de synténie obtenus entre l'homme et la souris ont été diffusés (Pevzner et Tesler, 2003a). Lorsque la méthode a été étendue à la comparaison de trois génomes simultanément, les données obtenues pour homme-souris-rat ont été également publiées (Bourque *et al.*, 2004). Nous utiliserons ces deux jeux de données, le premier que l'on appelle GRIMM2 et le deuxième GRIMM3. Les blocs de synténie de GRIMM2 sont localisés sur l'assemblage NCBI 30 du génome humain, ceux de GRIMM3 sur la version NCBI 33. Des paramètres très stringents ont été utilisés pour le premier : seuls les blocs de plus de 1 Mb ont été retenus ($G = 1$ Mb et $C = 1$ Mb), alors que dans le deuxième, les blocs de synténie doivent couvrir chacun au moins 300 Kb du génome humain ($G = 300$ Kb et $C = 450$ Kb).

La plateforme Ensembl (Hubbard *et al.*, 2007) met à disposition des blocs de synténie entre chaque couple d'espèces et pour chaque version de la base de données. Ces blocs sont calculés avec une méthode de clustering semblable à GRIMM, appliquée aux données d'alignement issues de CHAINNET (description très succinte de la méthode sur le site internet[4], les valeurs des paramètres utilisés ne sont pas indiquées). Nous avons utilisé les blocs de synténie homme-souris de la version actuelle d'Ensembl (version 49), qui sont localisés sur l'assemblage NCBI 36 du génome humain ; nous appelons ce jeu de données ENSEMBL. Nous n'avons pas pu utiliser les versions précédentes, et donc les blocs calculés sur la même version d'assemblage que nos données de points de cassure, puisque la méthode d'Ensembl était erronée dans les versions précédentes.

Les autres méthodes d'alignement de génomes complets, décrites dans la Section 2.3, soit n'ont pas de données publiques sur des comparaisons de mammifères (par exemple MAUVE), soit ne donnent pas directement des blocs de synténie et des points de cassure (c'est le cas de CHAINNET).

Pour les trois jeux de données, nous avons extrait les points de cassure, c'est-à-dire les régions entre deux blocs consécutifs sur le génome de l'homme dont les blocs orthologues ne sont pas consécutifs sur le génome de la souris, ou bien sont non colinéaires. Nous avons éliminé les points de cassure contenant un centromère humain, et lorsque les blocs se chevauchent sur le génome humain, nous considérons l'intersection comme le point de cassure.

a. Comparaisons globales

Tout d'abord, nous avons comparé globalement la distribution des tailles de points de cassure entre les différentes méthodes. Les points de cassure affinés par notre méthode sont globalement plus petits que ceux des trois autres jeux de données, avec une taille moyenne de 129 Kb contre 364, 454 et 1513 Kb pour, respectivement, GRIMM2, GRIMM3 et ENSEMBL (voir Tableau 4.3 et Figure 4.13). Les différences sont significatives au test de Wilcoxon (test non paramétrique de la somme des rangs), avec des p-values de $2.085e-14$, $< 2.2e-16$ et $4.977e-05$, lorsque les points de cassure affinés sont comparées à GRIMM2, GRIMM3 et ENSEMBL respectivement.

[4] `http://www.ensembl.org`

Taille des points de cassure (en pb)	minimum	maximum	médiane	moyenne
AFFINÉS	21	2 185 434	51 136	128 644
GRIMM2	313	5 418 383	155 816	364 199
GRIMM3	2 490	4 953 520	267 609	454 490
ENSEMBL	2	2 804 561	95 456	223 421

TAB. 4.3: Comparaison des distributions des tailles de points de cassure pour les quatre jeux de données.

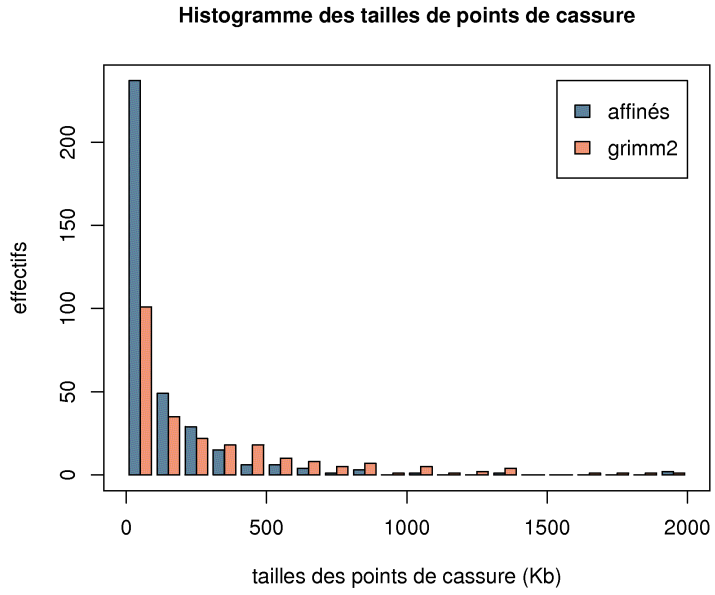


FIG. 4.13: Histogrammes des tailles des points de cassure affinés (en bleu) et des points de cassure du jeu de données GRIMM2 (en orange). Le dernier intervalle à 2000 Kb regroupe les points de cassures dont la taille est supérieure à 2000 Kb.

b. Comparaisons deux à deux

Tous les jeux de données ne contiennent pas le même nombre de points de cassure : nous avons 354 points de cassure affinés, alors que les jeux de données GRIMM2, GRIMM3 et ENSEMBL contiennent respectivement 246, 306 et 270 points de cassure. On pourrait alors se demander si on compare les mêmes points de cassure et si les différences de taille observées ne sont pas seulement dues à des points de cassure spécifiques de l'un ou l'autre des jeux de données comparés.

Pour tester cela, nous avons effectué des comparaisons deux à deux des points de cassure en commun aux jeux de données comparés. Pour identifier des points de cassure en commun, il faut comparer les coordonnées des points de cassure sur le génome humain. Or, les jeux de données ne sont pas basés sur la même version d'assemblage du génome humain, les coordonnées ne sont donc pas comparables. Nous avons utilisé l'outil LIFTOVER du navigateur de génome UCSC (Karolchik *et al.*, 2008) pour passer d'une version d'assemblage à une autre.

Les coordonnées sont comparées sur la version NCBI 35 de l'assemblage humain.

Deux points de cassure sont considérés communs à deux jeux de données si leurs intervalles se chevauchent sur le génome humain de manière 1-1 (c'est-à-dire un point de cassure du jeu X chevauche un seul point de cassure du jeu Y et vice versa) et si les blocs de synténie adjacents ont la même orientation et sont localisés sur le même chromosome chez la souris. Pour chaque point de cassure en commun, nous calculons la différence de taille entre notre jeu de données affinées et l'un des autres jeux de données.

Nous avons effectué ces comparaisons entre le jeu de points de cassure affinés et GRIMM3 d'une part, puis ENSEMBL d'autre part. Nous avons éliminé le jeu de données GRIMM2 car l'assemblage du génome humain sur lequel il repose est trop ancien.

Les coordonnées de 270 points de cassure du jeu de données GRIMM3 ont été converties sur l'assemblage utilisé pour les points de cassure affinés. Nous avons alors identifié 213 points de cassure en commun. Dans la Figure 4.14 est représenté l'histogramme des différences de taille. Les points de cassure de GRIMM3 sont plus grands de 270 Kb en moyenne et seulement 6 points de cassure de GRIMM3 ont une taille inférieure à celle affinée. La différence de taille est significative au test de Wilcoxon pour des données appariées (p-value inférieure à $2.2e-16$).

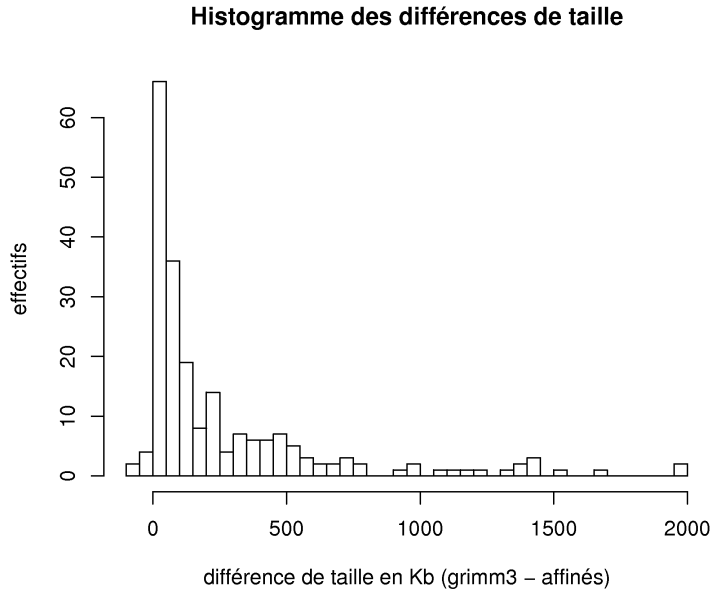


FIG. 4.14: Comparaison deux à deux des tailles de points de cassure entre GRIMM3 et les points de cassure affinés. Une valeur positive signifie que le point de cassure est plus grand dans GRIMM3 que dans notre jeu de cassures affinées. Le dernier intervalle à 2000 Kb regroupe les valeurs supérieures à 2000 Kb.

Il faut noter que les points de cassure de GRIMM3 ont été calculés sur un assemblage plus ancien du génome humain. Or, des erreurs et des trous (appelés gaps) dans cet assemblage ont pu compromettre certains alignements et nous ne pouvons exclure l'hypothèse que la grande taille de certains points de cassure soit due à des problèmes d'assemblage.

Par contre, le jeu de données ENSEMBL repose, lui, sur un assemblage plus récent. Les coordonnées de 270 points de cassure de ce jeu de données ont été converties sur l'assemblage

NCBI 35 (utilisé pour les points de cassure affinés). Nous avons alors identifié 185 points de cassure en commun. Dans la Figure 4.15 est représenté l'histogramme des différences de taille. Les points de cassure de ENSEMBL sont plus grandes de 102 Kb en moyenne et 38 points de cassure de ENSEMBL ont une taille inférieure à celle affinée. On remarque que pour ces 38 points, la différence de taille est souvent faible (voir Figure 4.15). La différence de taille est significative au test de Wilcoxon pour des données appariées (p-value inférieure à $2.2e-16$).

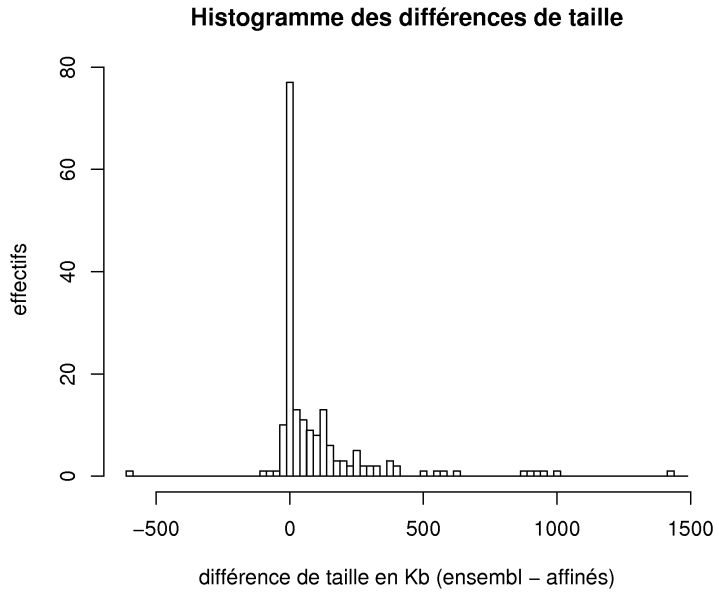


FIG. 4.15: Comparaison deux à deux des tailles de points de cassure entre ENSEMBL et les points de cassure affinés. Une valeur positive signifie que le point de cassure est plus grand dans ENSEMBL que dans notre jeu de cassure affinées. Le dernier intervalle à 2000 Kb regroupe les valeurs supérieures à 2000 Kb.

c. Conclusion

Ces comparaisons montrent que nous obtenons une meilleure résolution des points de cassure que les méthodes d'alignement de génomes complets. Ce résultat vient du fait que nous procédons en deux étapes : dans la première, nous identifions les blocs de synténie et dans la deuxième nous nous concentrons sur chaque point de cassure indépendamment. Ayant réduit l'espace de recherche d'homologie à la première étape, nous pouvons être plus sensibles dans la détection de similarité dans la deuxième étape. Les méthodes d'alignement de génomes complets opèrent en une seule étape. Cela nécessite d'utiliser des paramètres assez stringents dans la détection des blocs de synténie pour éviter les faux orthologues. En contre-partie, de faibles similarités dans les points de cassure sont ratées.

Nous avons montré ici que cette similarité existe dans les points de cassure et que l'on peut s'en servir pour affiner les points de cassure. La résolution obtenue avec notre méthode pourra permettre d'effectuer des analyses des points de cassure plus précises (voir les chapitres suivants).

4.4.4 Comparaison deux à deux ou multiple ?

Un autre argument est à avancer pour rendre compte du gain en précision de notre méthode. Il est basé sur le nombre de génomes comparés. Avec le nombre croissant de génomes entièrement séquencés, la tentation est grande de comparer simultanément plus de deux génomes. Plusieurs méthodes ont été développées avec cet objectif (Bourque *et al.*, 2004, 2005; Murphy *et al.*, 2005; Karolchik *et al.*, 2008). Cependant, nous avons décidé de développer une méthode deux à deux et non multiple. La motivation est de gagner encore en précision.

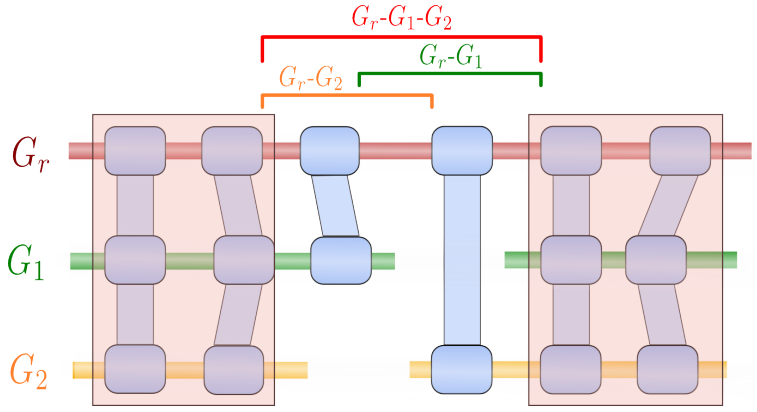
Si on compare les tailles des points de cassure entre les jeux de données GRIMM2 et GRIMM3, on s'aperçoit que le second issu d'une comparaison multiple à trois génomes donne des points de cassure moins résolus que le premier, qui est deux à deux. De plus, les paramètres utilisés dans le deuxième jeu de données sont beaucoup moins stringents, on s'attendrait alors à obtenir des points de cassure plus précis. Ce n'est pas le cas et cela provient du fait que la comparaison est multiple. Les ancres utilisées dans GRIMM3 sont "3-way", c'est-à-dire qu'une ancre représente un marqueur orthologue dans chacun des trois génomes. L'avantage est que les ancres sont plus fiables, cependant cela réduit le jeu d'ancres et donc la taille des blocs de synténie (voir un exemple dans la Figure 4.16a). Par exemple, un couple de marqueurs orthologues entre l'homme et la souris ne sera peut-être pas présent dans le jeu d'ancres 3-way si le marqueur orthologue chez le rat n'est plus présent, n'est plus assez similaire ou bien est dupliqué.

Les comparaisons multiples peuvent être utiles pour construire les blocs de synténie de notre première étape, puisqu'elles améliorent la fiabilité des blocs. Cependant, une méthode deux à deux est préférable dans la deuxième étape, celle d'affinement des points de cassure. En effet, cela permet d'être plus sensible dans la détection de similarités.

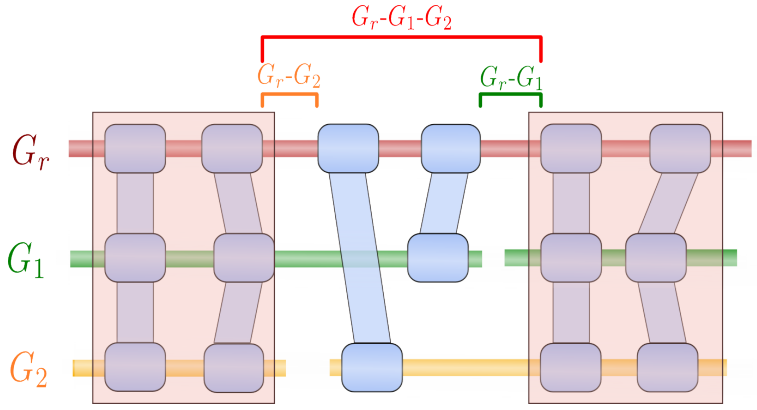
De plus, cela permet également de distinguer s'il y a un ou plusieurs évènements de réarrangement dans une même région. Supposons, par exemple, que deux points de cassure sont très proches l'un de l'autre sur le génome humain. Le premier provient d'un réarrangement ayant eu lieu dans le génome de la souris, le second d'un autre réarrangement ayant eu lieu dans le génome du chien. On appelle ce phénomène la ré-utilisation de point de cassure au cours de l'évolution. L'identification des deux points de cassure indépendamment par des comparaisons deux-à-deux, permet de déterminer si les deux points de cassure se chevauchent sur le génome humain. Par contre, la comparaison multiple des trois génomes peut conduire à n'identifier qu'un seul point de cassure, dans le cas par exemple où la distance entre les deux points de cassure est inférieure à la taille minimale d'un bloc de synténie (voir Figure 4.16b).

A titre d'exemple, nous avons identifié 5 cas où un point de cassure homme-souris est situé à moins de 50 kb d'un point de cassure homme-chien sur le génome de l'homme, sans qu'ils ne se chevauchent (10 cas lorsqu'on fixe la distance maximale à 100 Kb, et 39 cas pour 300 Kb). Cette stratégie peut donc être efficace pour évaluer le nombre de ré-utilisation de points de cassure dans différentes lignées.

Il semble ainsi préférable d'identifier précisément les points de cassure entre deux génomes, dans un premier temps, puis de comparer les coordonnées obtenues avec différentes comparaisons deux à deux. On pourra alors essayer d'inférer les relations évolutives entre les différents points de cassure qui se chevauchent sur le génome de référence par une analyse phylogénétique. Nous avons développé une méthode basée sur le principe de parcimonie pour inférer la position des évènements de réarrangement dans l'arbre des espèces à partir des points de cassure identifiés deux à deux entre l'homme et les 5 autres espèces de mammifères (cette



(a) Chevauchement des deux points de cassure.



(b) Deux points de cassure distincts.

FIG. 4.16: Exemple schématique présentant les avantages des comparaisons deux à deux plutôt que multiples. Nous disposons de trois génomes G_r, G_1 et G_2. Les carrés bleus représentent des zones de similarités entre les différents génomes qui servent à identifier les blocs de synténie. Les blocs roses sont des blocs de synténie identifiés par une comparaison multiple des trois génomes. On identifie donc le point de cassure correspondant en rouge. Or, des comparaisons deux à deux, d'une part entre G_r et G_1 (en vert), d'autre part entre G_r et G_2 (en orange) permettent d'identifier plus finement les points de cassure. Deux cas de figure sont représentés ici : en a) les points de cassure identifiés deux à deux se chevauchent sur G_r, en b) ils ne se chevauchent pas. Dans ce dernier cas, les comparaisons deux à deux mettent en évidence deux points de cassures distincts.

méthode est décrite dans le chapitre 6, Section 6.1.1a.).

Cela permet également d'affiner encore plus les points de cassure : dans l'exemple de la Figure 4.16a, si l'analyse phylogénétique confirme l'existence d'un seul évènement de réarrangement, le point de cassure est réduit à l'intersection des deux points de cassure identifiés deux à deux ($G_r - G_1 \bigcap G_r - G_2$).

Chapitre 5

Caractéristiques des séquences de points de cassure

Sommaire

Grâce à l'amélioration de la résolution des points de cassure, nous pouvons analyser plus en détail le contenu de ces séquences. De nombreuses analyses de séquences de points de cassure ont déjà été effectuées. Cependant, il s'agit principalement d'analyses ponctuelles de points de cassure. La majorité de ces analyses portent sur un réarrangement impliqué dans un désordre génomique et l'objectif est d'identifier la (ou les) cause(s) du réarrangement. Ces analyses ont permis d'identifier notamment des duplications segmentaires (lire les revues (Ji *et al.*, 2000; Stankiewicz et Lupski, 2002)), des éléments transposables (Dehal *et al.*, 2001), des séquences particulières telles que de courtes répétitions ou des palindromes (Kato *et al.*, 2008) qui sont impliqués dans la formation du réarrangement (voir Section 1.4.3). Elles sont donc utiles pour comprendre les mécanismes de certains réarrangements, mais elles ne permettent pas d'évaluer si ces mécanismes concernent tous les réarrangments ou bien seulement certains types. Ainsi, des analyses systématiques pourraient permettre de répondre à ces questions et d'identifier statistiquement des caractéristiques de séquences spécifiques des réarrangements sans *a priori*.

Les analyses systématiques des séquences de points de cassures ne sont pas nombreuses, notamment car la résolution des points de cassure est limitée. Les principaux résultats issus de ces analyses sont une perte de similarité au niveau des points de cassure (Kent *et al.*, 2003; Trinh *et al.*, 2004) et la présence de duplications segmentaires aux environs de ces points (Armengol *et al.*, 2003; Bailey *et al.*, 2004).

Nous nous proposons dans ce chapitre d'étudier si les séquences de points de cassure présentent des caractéristiques particulières. Notamment nous revenons sur les résultats de similarité et de duplications segmentaires déjà mis en évidence, et nous recherchons également d'autres éléments qui ont été identifiés dans des analyses ponctuelles mais pas dans le cadre d'une analyse à l'échelle du génome. Enfin, nous étudierons un cas particulier de points de cassure dans lequel des duplications de séquences peuvent être directement reliées au mécanisme de réarrangement.

5.1 Evolution des séquences dans et autour des points de cassure

Les méthodes d'alignement de génomes complets appliquées aux génomes de l'homme et de la souris ont mis en évidence de grandes régions entre les blocs de synténie présentant très peu de similarité (Pevzner et Tesler, 2003a; Kent *et al.*, 2003). Cette absence de similarité et l'existence d'un grand nombre de tous petits blocs dans les régions de cassure a été interprétée comme une indication de fragilité de ces régions et la trace de micro-réarrangements (Pevzner et Tesler, 2003b). Ces micro-réarrangements auraient tellement réarrangé ces régions qu'elles ne seraient plus alignables avec le génome de la souris. Trinh et collègues s'opposent à ces arguments et ont proposé plusieurs alternatives (Trinh *et al.*, 2004). D'une part, certaines parties des points de cassures pourraient être rattachées aux blocs de synténie voisins ; c'est ce que nous avons effectivement confirmé avec la méthode d'affinement. D'autre part, la perte de similarité pourrait être due à une accélération de la divergence des séquences par des processus mutationnels causés par le réarrangement et non par d'autres réarrangements.

Grâce à la méthode d'affinement nous avons confirmé la première hypothèse de Trinh et collègues et nous avons donc mieux délimité les points de cassure. Dans ces conditions, on peut analyser plus en détail la divergence des séquences de points de cassure et de celles avoisinantes. La méthode d'affinement des points de cassure permet d'étudier les similarités de séquence le long de la séquence S_r avec ses deux séquences orthologues S_{oA} et S_{oB}. En l'appliquant sur des séquences non cassées, ou colinéaires, on peut donc évaluer si les séquences de points de cassure présentent une évolution particulière. Nous avons notamment comparé la taille des régions affinées et la divergence des séquences du point de cassure et de ses séquences adjacentes.

5.1.1 Jeu de données de fausses cassures

Pour chaque jeu de données de points de cassure entre deux génomes G_r et G_o, nous construisons un jeu de données de régions colinéaires, qu'on appellera des fausses cassures, à partir des mêmes données en entrée, c'est-à-dire les blocs de synténie entre G_r et G_o.

a. Définition des fausses cassures et de leurs séquences d'intérêt

Les régions colinéaires se trouvent par définition à l'intérieur de blocs de synténie. Les blocs de synténie ont été construits avec la méthode présentée dans le Chapitre 3, en utilisant

les gènes orthologues comme marqueurs. Un vrai point de cassure est donc situé entre deux gènes orthologues consécutifs sur G_r, appartenant à deux blocs de synténie différents. Nous définissons un faux point de cassure, ou **fausse cassure**, par la région entre deux gènes orthologues consécutifs sur G_r appartenant à un même bloc de synténie ; ces deux paires de gènes orthologues sont donc colinéaires (voir la Figure 5.1).

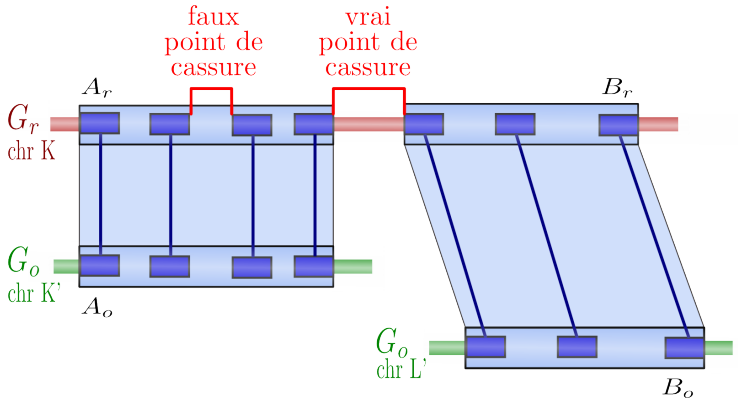
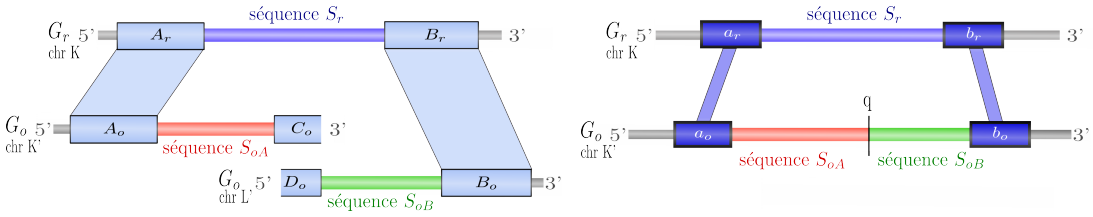


FIG. 5.1: Schéma représentant les vraies et fausses cassures entre deux génomes G_r et G_o. Les gènes orthologues sont représentés par des rectangles bleus foncés à l'intérieur des deux blocs de synténie (A_r, A_o) et (B_r, B_o).

Nous définissons une fausse cassure entre les gènes a_r et b_r, consécutifs sur le génome G_r, appartenant à un même bloc de synténie (leurs orthologues respectifs a_o et b_o sont donc également consécutifs sur le génome G_o et colinéaires) (voir Figure 5.2b).

Afin d'appliquer la méthode d'affinement sur ces régions de la même manière que pour les vrais points de cassure, nous devons définir les trois séquences d'intérêt à aligner pour chacune d'elles : les séquences S_r, S_{oA} et S_{oB} (voir Chapitre 4, Section 4.1.1, voir aussi la Figure 5.2a).

La séquence S_r est la séquence localisée entre les gènes a_r et b_r sur le génome G_r. Si les blocs de synténie sont corrects, S_r est orthologue à la séquence entre les gènes a_o et b_o sur le génome G_o. On tire alors dans cette séquence une position q qui la coupe en deux séquences. Ces deux séquences constitueront (en partie) les séquences S_{oA} et S_{oB} respectivement (voir le schéma de la Figure 5.2b).



(a) Vraies cassures.

(b) Fausses cassures.

FIG. 5.2: Définition des séquences d'intérêt pour les vraies et les fausses cassures.

Dans le cas des vrais points de cassure, les séquences S_{oA} et S_{oB} ne sont pas entièrement orthologues à S_r et possèdent *a priori* une partie non orthologue à S_r. Pour modéliser cela dans les fausses cassures, on ajoute aux séquences S_{oA} et S_{oB} des séquences du génome G_o

qui ne sont pas orthologues à S_r. Ces séquences ajoutées sont prises parmi des séquences S_{oA} ou S_{oB} d'autres régions colinéaires.

b. Echantillonnage

Bien sûr, il existe beaucoup plus de régions colinéaires que de points de cassure. Pour créer le jeu de fausses cassures, nous échantillonnons le même nombre de régions colinéaires que de points de cassure.

L'objectif de cet échantillonnage est d'obtenir un jeu de fausses cassures ayant des caractéristiques similaires aux vrais points de cassure. Les caractéristiques que nous voulons contrôler sont la longueur de la séquence S_r et la différence de longueur entre les séquences S_{oA} et S_{oB}, car elles peuvent influencer la méthode d'affinement. Plus les séquences sont grandes, plus on peut obtenir de hits d'alignements simplement dus au hasard. Ainsi, la longueur des séquences individuellement et les différences de longueur peuvent influencer la quantité de faux positifs et donc la qualité de l'affinement.

On échantillonne donc les fausses cassures parmi l'ensemble des régions colinéaires du génome G_r en veillant à obtenir une distribution des longueurs S_r similaire à celle des vrais points de cassure. Pour contrôler la différence de longueur des séquences S_{oA} et S_{oB}, on joue sur la position de la limite q définissant ces deux séquences. On place q de façon à respecter la distribution du rapport des longueurs de S_{oA} et S_{oB} des vrais points de cassure.

c. Espèces comparées

Nous effectuons les analyses sur le génome de l'homme (G_r) comparé à trois autres espèces de mammifères : la souris, le chien et le chimpanzé. Ces choix sont motivés par plusieurs raisons.

Le choix du génome humain comme référence est naturel, puisque c'est le génome de mammifères le plus étudié. On possède ainsi énormément d'informations génomiques localisées sur les chromosomes humains, que l'on pourra étudier avec les points de cassure.

En ce qui concerne les génomes de comparaison, tout d'abord ces trois génomes ont un assemblage d'assez bonne qualité ; le génome de la souris est séquencé depuis 2003 et c'est le seul génome des mammifères avec le génome humain dont le statut de l'assemblage est "complet". Les assemblages du chien et du chimpanzé sont encore à l'état de "brouillon" mais en sont à leur deuxième version, contrairement aux autres génomes de mammifères (euthériens) séquencés (excepté le rat) comme le macaque, la vache ou le cheval. Ensuite, ils présentent des niveaux de divergence avec l'homme assez différents. La souris et le chien sont suffisamment éloignés pour obtenir un grand nombre de points de cassure. Par contre, le chimpanzé est très proche de l'homme (les deux espèces ont divergé il y a environ 6 millions d'années) ; le nombre de points de cassure est donc faible, mais la faible divergence des séquences est intéressante pour les analyses. D'une part, cela permet d'avoir des points de cassure plus précis et d'autre part, on peut éliminer ce facteur comme cause potentielle de séquences non alignées. En effet, les deux génomes sont tellement proches que si on ne peut aligner certaines séquences, on ne peut en général pas invoquer l'impuissance de la méthode d'alignement.

Au niveau des réarrangements, il existe également des différences : les réarrangements détectés entre l'homme et le chimpanzé sont essentiellement des inversions, alors que presque la moitié des points de cassure entre l'homme et la souris (de même pour le chien) sont dus à des réarrangements inter-chromosomiques. Cependant, le génome de la souris est un peu

particulier du point de vue des réarrangements ; les rongeurs en général ont un taux de réarrangements bien supérieur aux autres mammifères (Burt *et al.*, 1999). Ainsi, l'utilisation du génome du chien permet de vérifier que les caractéristiques observées ne sont pas spécifiques de l'évolution des rongeurs.

Nous disposons de 36 points de cassure entre l'homme et le chimpanzé, 354 entre l'homme et la souris et 240 entre l'homme et le chien. Nous avons généré trois jeux de fausses cassures pour ces trois couples d'espèces avec les mêmes effectifs que les jeux de points de cassure.

5.1.2 Taille des régions affinées

Si la méthode d'affinement des points de cassure permet d'obtenir des points de cassure plus précis que les autres méthodes existantes, il en reste un grand nombre dont la précision est décevante. Un tiers des points de cassure homme-souris ont une taille supérieure à 100 Kb après affinement (117/354), un tiers de ceux homme-chimpanzé font plus de 50 Kb (12 sur 36). Ces régions ne présentent pour la plupart aucune similarité avec leurs séquences supposées orthologues (S_{oA} et S_{oB}). Or, il ne faut pas oublier qu'entre l'homme et la souris par exemple, seulement 40 % des séquences génomiques sont alignables. Ainsi, on peut se demander si la taille des points de cassure ne reflète pas simplement la distance moyenne entre deux hits d'alignement successifs entre les deux génomes comparés, ou bien si cela peut être un artefact de la méthode d'affinement.

a. Comparaison avec les fausses cassures

La taille de la région affinée est significativement plus grande pour les vraies cassures que pour les fausses. Dans les régions colinéaires, elle est en moyenne de 7 Kb pour la comparaison homme-souris et de 2 Kb pour la comparaison homme-chimpanzé (voir Table 5.1).

		Taille de la région affinée (pb)	
		médiane	moyenne
homme-souris	vrai	50 670	128 300
	faux	1 314	7 602
homme-chien	vrai	33 140	86 070
	faux	312	3 552
homme-chimpanzé	vrai	21 600	99 390
	faux	3	1 899

TAB. 5.1: Comparaison des tailles des régions affinées entre les vraies et les fausses cassures pour les trois couples d'espèces comparées.

Les différences sont très importantes et ne sont pas dues à un petit nombre de points de cassure qui seraient très grands, comme le montre la différence importante entre les médianes. Par exemple, 65 % des points de cassure homme-souris sont plus grands que le quantile à 95 % de la distribution de longueur des fausses cassures (29 Kb).

On remarque que cette différence est encore plus grande pour la comparaison homme-chimpanzé, alors qu'on aurait pu s'attendre à la tendance inverse puisque les évènements de réarrangements sont plus récents que dans les autres comparaisons.

b. Similarité avec d'autres régions génomiques

Qu'est-ce qui pourrait expliquer que certains points de cassure soient si grands ? Du point de vue de l'affinement, les régions délimitées par la segmentation contiennent, soit très peu de hits d'alignement avec les deux séquences de l'autre génome, soit des hits provenant de ces deux séquences en alternance ou bien se chevauchant. Dans ce dernier cas, la région de cassure est orthologue aux deux séquences S_{oA} et S_{oB} ; le plus souvent il s'agit d'une duplication dans le génome G_o. Nous nous intéressons à ce cas plus précisément dans la Section 5.3. Nous verrons que les cas de duplications impliquant les deux séquences S_{oA} et S_{oB} sont facilement détectables et ne sont pas majoritaires. Nous nous intéressons ici principalement aux autres cas.

Si la région du point de cassure contient très peu de hits avec les séquences S_{oA} et S_{oB}, deux explications sont possibles. La première est que la séquence du point de cassure est orthologue avec une partie du génome G_o autre que S_{oA} et S_{oB}. Dans ce cas, il faut envisager au moins un autre réarrangement ayant eu lieu dans cette région que l'on n'aurait pas détecté (bloc de synténie manquant). Si cette région n'est similaire à aucune région du génome G_o, alors on peut envisager trois hypothèses : les séquences ont tellement divergé qu'elles ne sont plus alignables, il s'agit d'une "nouvelle" séquence insérée dans le génome G_r, ou bien c'est une séquence qui a été délétée dans le génome G_o.

Nous voulons vérifier ici que la séquence du point de cassure n'est pas orthologue à une autre région génomique. Pour cela, nous avons mesuré dans les séquences des points de cassure et leurs séquences adjacentes, la couverture en alignements génomiques provenant de comparaisons de génomes complets (la définition précise des séquences adjacentes est décrite dans la Section 5.2.1). La présence (et la quantité) d'alignements dans une séquence humaine est une indication de la similarité de cette séquence avec n'importe quelle région de l'autre génome.

Nous présentons ici les résultats pour les points de cassure homme-souris dont la taille est supérieure à 10 Kb. Nous avons utilisé les alignements entre le génome humain et celui de la souris obtenus avec la méthode Chainnet (voir la description de la méthode dans le Chapitre 2, Section 2.3). Ils sont disponibles sur le navigateur de génome de l'UCSC (Karolchik *et al.*, 2008). Ils couvrent environ 37 % du génome humain ; chaque position du génome humain est couverte par au plus un alignement avec la souris, mais la réciproque n'est pas vraie. Les alignements ont été chaînés mais ils ne font pas forcément partie d'un bloc de synténie.

On observe une différence importante de couverture des séquences en alignements génomiques entre les séquences de points de cassure et leurs séquences adjacentes. Les alignements génomiques couvrent en moyenne 14.0 % des séquences de points de cassure, contre 28.4 % dans leurs séquences adjacentes (voir les distributions dans la Figure 5.3).

On observe surtout que presque la moitié des points de cassure de plus de 10 Kb (soit 114 points de cassure) sont couverts à moins de 5 % par des alignements génomiques avec la souris. Ces points de cassures sont donc, soit des séquences uniques au génome de l'homme, soit elles ont trop divergé et on ne peut plus les aligner avec le génome de la souris, soit elles ont été perdues chez la souris.

Par contre, 47 séquences de points de cassure sont couvertes à plus de 30 % par des alignements génomiques. On pourrait penser que, pour ceux-là, nous avons raté un bloc de synténie dans le point de cassure. Cependant, 42 de ces séquences correspondent, au moins en partie, à des séquences dupliquées, soit dans le génome de l'homme, soit dans le génome de la souris. En effet, 33 sont également couvertes à plus de 30 % par des duplications segmentaires humaines

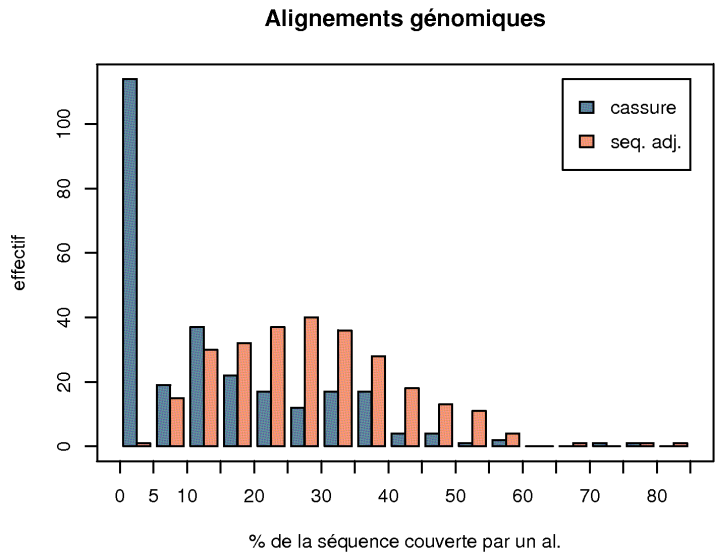


FIG. 5.3: Couverture des séquences de points de cassure (en bleu) et des séquences adjacentes (en orange) en alignements génomiques avec le génome de la souris.

(voir Section 5.2.3). Et 13 correspondent à des régions que nous avons réussi à aligner, mais avec les deux séquences orthologues S_{oA} et S_{oB} (voir Section 5.3).

Finalement, très peu de points de cassure montrent des similarités de séquence avec d'autres régions du génome de la souris (autres que des duplications). On ne peut donc pas expliquer la grande taille des points de cassure par des blocs de synténie manqués ou erronés.

5.1.3 Alignements des séquences

Nous nous intéressons ici aux résultats des alignements effectués pour la segmentation (l'alignement de S_r avec S_{oA}, donnant les hits rouges, et l'alignement de S_r avec S_{oB}, donnant les hits verts). On considère la distribution des hits dans la séquence S_r entière, ou bien la séquence S_r privée de la région affinée (correspondant au point de cassure). En effet, de part et d'autre de la région affinée, les séquences sont considérés orthologues respectivement aux séquences S_{oA} et S_{oB} (voir Figure 5.4).

Nous nous demandons si les vraies cassures s'alignent de la même façon que les fausses cassures.

a. Similarité des séquences alignées

Nous comparons tout d'abord la quantité et la distribution des hits le long de la séquence S_r entre les vraies et les fausses cassures. Nous mesurons la couverture de la séquence S_r en hits (indépendamment de leur couleur). Nous pouvons également différencier les deux types de hits en regardant les valeurs ajustées de la fonction constante par morceaux des segments 1 et 3 de la segmentation (s_1 et s_3 respectivement) (voir Chapitre 4, Section 4.3, et Figure 5.4). Brièvement, la valeur s_1 du segment 1 (représentant l'homologie avec la séquence S_{oA})

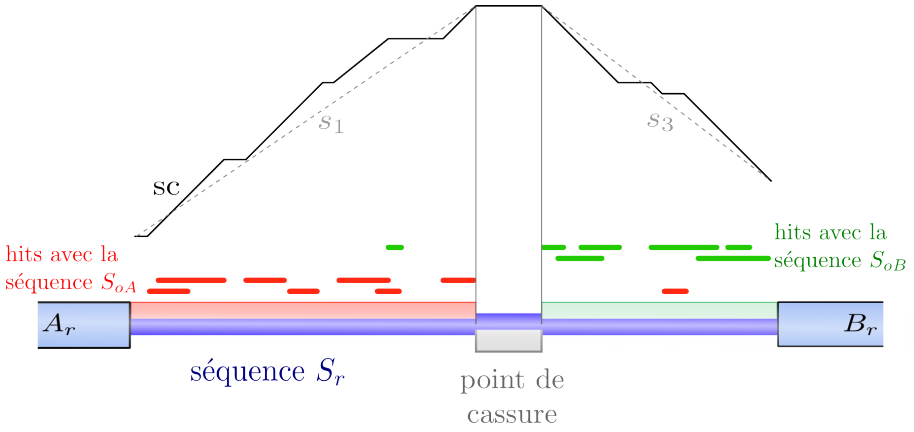


FIG. 5.4: Représentation schématique du résultat de la méthode d'affinement. Les blocs de synténie A_r et B_r ont été étendus sur le génome G_r jusqu'au point de cassure, grâce au processus de segmentation (représenté par la fonction de score) des hits d'alignement. On a représenté également les valeurs s_1 et s_3 du processus de segmentation, comme les pentes de la fonction de score de part et d'autre du point de cassure.

est la couverture sur ce segment en hits rouges moins la couverture sur ce segment en hits verts. C'est une valeur comprise entre 0 et 1 pour s_1 et -1 et 0 pour s_3.

Les résultats des comparaisons de ces différentes valeurs sont résumés dans le Tableau 5.2.

		Couverture en hits (%)		s_1	s_3
		S_r entière	S_r sans seg. 2		
homme-souris	vrai	42	50	0.49	-0.50
	faux	58	59	0.59	-0.59
homme-chien	vrai	54	62	0.61	-0.62
	faux	72	73	0.73	-0.73
homme-chimpanzé	vrai	83	85	0.80	-0.82
	faux	95	96	0.94	-0.94

TAB. 5.2: Comparaison des distributions de hits sur la séquence S_r entre les vraies et les fausses cassures pour les trois couples d'espèces comparées. La couverture en hits est le nombre de positions couvertes par un hit (rouge ou vert) divisé par la longueur de la séquence concernée ; les positions masquées ne sont pas comptées. La valeur s_1 (resp. s_3) est celle de la fonction constante par morceaux dans le segment 1 (resp. 3), c'est-à-dire la différence de couverture des hits rouges et des hits verts sur le segment 1 (resp. 3).

Nous observons tout d'abord que les séquences S_r des vraies régions de cassures sont globalement moins couvertes par des hits que les séquences des fausses cassures. Pour la comparaison homme-souris, en moyenne 42 % de la séquence masquée (sans compter les positions masquées) est couverte par au moins un hit (rouge ou vert) pour les vraies cassures contre 58 % pour les fausses (voir aussi la Figure 5.5).

Cela pourrait être dû au segment central (le point de cassure affiné) qui contient très peu de hits et qui est beaucoup plus grand pour les vraies cassures que pour les fausses. Mais même en éliminant le segment central, la couverture en hits reste significativement plus faible

pour les vraies cassures que pour les fausses (50 % contre 59 % en moyenne). De même, la valeur s_1 du premier segment est plus faible pour les vraies cassures que pour les fausses (0.49 contre 0.59 pour la comparaison homme-souris). On notera que la valeur de s_1 est très proche (si on la ramène à un pourcentage) de la couverture en hits pour les deux types de séquences. Cela montre que les hits rouges et verts sont bien séparés dans leurs segments respectifs, et c'est seulement la densité en hits qui fait la différence des valeurs s_1 entre les vraies et les fausses cassures (voir Tableau 5.2).

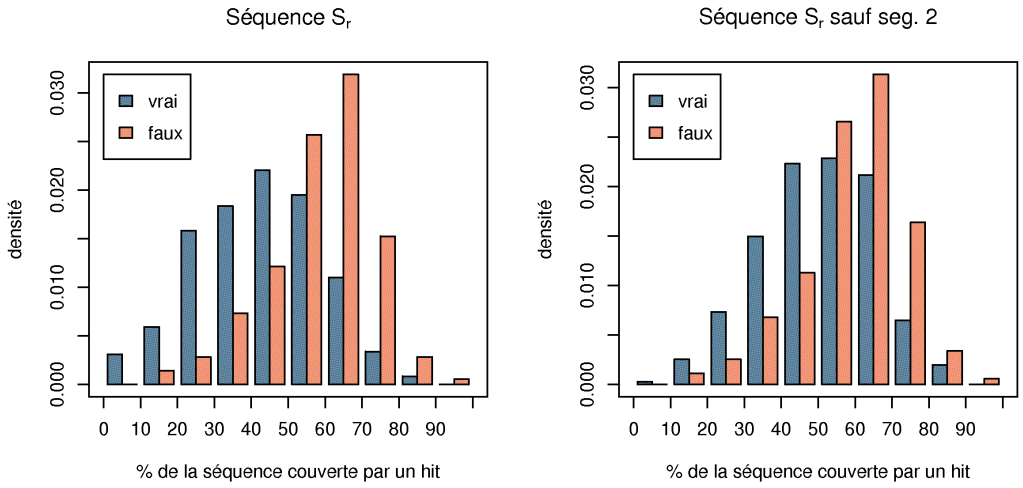


FIG. 5.5: Comparaison de la densité en hits des séquences S_r entre les vraies (en bleu) et les fausses cassures (en orange) pour la comparaison homme-souris. A gauche est représentée la densité en hits sur la séquence S_r en entier (sans compter les positions masquées), à droite le segment du milieu n'est pas compté.

Ces résultats montrent que les séquences voisines d'un point de cassure ont plus divergé que des séquences de régions colinéaires.

b. Colinéarité des alignements

La méthode d'affinement utilise la distribution des hits le long de la séquence S_r indépendamment de leurs positions dans les séquences de l'autre génome (S_{oA} et S_{oB}). On peut se demander comment ces hits sont distribués dans ces séquences, et plus exactement s'ils sont colinéaires entre eux. S'il n'y a aucun faux positif et aucun réarrangement dans ces séquences, ils devraient être colinéaires et former une chaîne. Le degré de colinéarité des alignements peut donc être un indice du degré de désordre dans les séquences alignées. On peut alors se demander si cette caractéristique est différente entre les vraies et les fausses cassures.

Il reste à définir une mesure du "degré de colinéarité". On pourrait envisager d'aligner les séquences avec des algorithmes permettant d'identifier des petits réarrangements, ou bien comparer l'ordre et l'orientation des hits dans les deux séquences alignées et compter le nombre de points de cassure. Plus simplement, nous avons utilisé une option du programme Blastz qui calcule, à partir des hits obtenus, la chaîne de hits colinéaires de meilleur score entre les deux séquences. On peut donc comparer la couverture de la séquence par les hits de la chaîne

et la couverture de la séquence par tous les hits. On définit l'indice de colinéarité, $r_{col}(X)$, de S_r avec la séquence X, par le rapport suivant :

$$r_{col}(X) = \frac{\sum_{i=1}^{n} couv(i, chain(X))}{\sum_{i=1}^{n} couv(i, hits(X))}$$

avec

$$couv(i, L) = \begin{cases} 1 \text{ si la position } i \text{ de } S_r \text{ est couverte par au moins un hit de l'ensemble L} \\ 0 \text{ sinon} \end{cases}$$

$hits(X)$ l'ensemble des hits obtenus entre S_r et X, et $chain(X)$ l'ensemble des hits appartenant à la meilleure chaîne entre S_r et X.

Lorsque tous les hits détectés forment une seule chaîne de hits colinéaires, la valeur de r_{col} atteint la valeur maximale 1.

Nous avons calculé cet indice pour chaque point de cassure homme-souris et chaque fausse cassure pour les séquences S_{oA} et S_{oB} ($r_{col}(S_{oA})$ et $r_{col}(S_{oB})$ (voir la distribution de $r_{col}(S_{oA})$ dans la Figure 5.6).

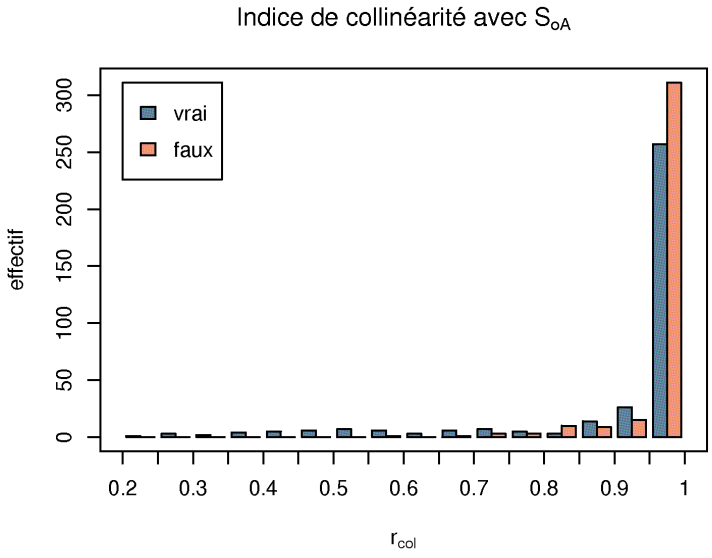


FIG. 5.6: Distribution de l'indice de colinéarité, $r_{col}(S_{oA})$, des alignements entre S_r et S_{oA} pour les vraies cassures (en bleu) et les fausses cassure (en orange) entre les génomes de l'homme et de la souris.

La très grande majorité des points de cassure et des fausses cassures ont leurs deux indices de colinéarité très proche de 1 : $r_{col}(S_{oA})$ est supérieur à 0.95 pour 77 % des points de cassure et 88 % des fausses cassures. Cela signifie que, pour ces séquences, presque tous les hits obtenus sont colinéaires deux à deux et forment une chaîne. Cela montre qu'il y a très peu de faux positifs dans les résultats d'alignement.

Par contre, on décompte 43 points de cassure pour lesquels $r_{col}(S_{oA})$ est inférieur à 0.7, alors que seulement 2 fausses cassures sont dans ce cas. Pour ces cas là, plus de 30 % de la

séquence S_r est couverte par des hits qui n'appartiennent pas à la meilleure chaîne, ce qui signifie que ces parties de S_r s'alignent, soit sur l'autre brin, soit sur le même brin mais sont non colinéaires. Dans le premier cas, il s'agit vraisemblablement de petites inversions, dans le deuxième, ce peut être des transpositions ou des duplications sur la séquence S_r.

Ces résultats montrent d'une part que les alignements de séquences utilisés pour l'affinement présentent très peu de faux positifs. D'autre part, si la majorité des points de cassure ne semblent pas avoir de petits réarrangements au sein des séquences S_{oA} et S_{oB}, d'autres présentent un fort taux de désordre dans les hits obtenus, indiquant la présence de micro-réarrangements. Ce résultat est en accord avec le modèle des régions fragiles (voir Chapitre 1, Section 1.4.2), qui propose que certaines régions du génome concentrent plus de réarrangements que d'autres.

Cela ne devrait pas, en théorie, affecter le processus de segmentation, ni la densité en hits de la séquence S_r, ni les valeurs s_1 et s_3 de la fonction constante par morceaux, car l'ordre et l'orientation des hits dans les séquences de G_o ne sont pas pris en compte. Cependant, on observe que pour les 43 points de cassure dont $r_{cor}(S_{oA})$ est inférieur à 0.7, la valeur moyenne de s_1 est seulement de 0.35, contre 0.51 pour les autres points de cassure. Il semble donc que la présence de micro-réarrangements dans les séquences réduise la densité en hits. Cependant, cela ne concerne qu'un petit nombre de points de cassure et n'est pas suffisant pour expliquer les différences de similarité entre les vraies et les fausses cassures.

c. Autres facteurs liés à la similarité des séquences

Nous avons testé si d'autres facteurs, non liés aux réarrangements, pouvaient expliquer cette plus faible similarité à proximité du point de cassure.

Les séquences des régions cassées contiennent globalement plus d'éléments répétés (détectés par RepeatMasker) que les séquences des régions colinéaires (le pourcentage moyen de la séquence S_r masquée par RepeatMasker est de 47.6 % pour les vraies cassures contre 43.7 % pour les fausses). Malgré le fait que les positions masquées par RepeatMasker ne sont pas comptées dans les calculs précédents de densité en hits, nous observons que cela joue un rôle dans les résultats observés. En fait, la densité en hits des séquences masquées est négativement corrélée avec leurs proportions de bases masquées, le coefficient de corrélation est de -0.39 et -0.44 pour les vraies et les fausses cassures respectivement (p-values du test de rang de Spearman inférieures à 10^{-14}). Ainsi, plus les séquences sont masquées, moins elles contiennent de hits en dehors des positions masquées. Cela peut être dû au fait que les intervalles entre deux positions masquées sont de plus en plus petits et qu'il est difficile d'obtenir des hits d'alignement courts qui soient significatifs.

Ainsi, la différence de contenu en éléments répétés pourrait expliquer la différence de similarité entre les vraies et les fausses cassures. Or, à contenu égal en éléments répétés, on observe encore des différences significatives de densités en hits entre les deux types de "cassures" (par exemple, pour un contenu en éléments répétés compris entre 40 et 45 %, la densité en hits (sur S_r sauf segment central) est de 53.7 % pour les vraies contre 59.5 % pour les fausses cassures, p-value de 0.001 au test de Wilcoxon).

Les différences de contenus en éléments répétés ne semblent donc pas expliquer entièrement les différences de similarité des séquences entre les vraies et les fausses cassure.

Il est connu que la vitesse d'évolution des séquences n'est pas constante le long du génome,

elle varie notamment entre les parties codantes et non codantes des séquences. Les parties codantes, étant fonctionnelles, sont généralement plus contraintes et évoluent moins vite que le reste du génome. Ainsi, on s'attend à ce que, plus les séquences contiennent de gènes (et d'exons), mieux elles s'alignent avec leurs séquences orthologues. Nous avons comparé la densité en gènes et la densité en codant dans les séquences S_r des vraies et des fausses cassures. On rappelle que les séquences sont définies entre deux gènes orthologues consécutifs, mais peuvent contenir des gènes sans orthologue 1-1 assigné.

Les séquences des points de cassure contiennent plus de gènes et d'exons que les séquences des régions colinéaires. Les gènes couvrent en moyenne 22.7 % des séquences S_r des points de cassure homme-souris contre 16.8% pour les fausses cassures. En ce qui concerne seulement les régions codantes, les exons couvrent 1.7 % des séquences de points de cassure contre 1.0 % pour les séquences des fausses cassures (différence significative au test de Wilcoxon avec une p-value de $3e-14$). Ainsi, si les séquences adjacentes aux points de cassure sont moins similaires que les séquences non cassées, ce n'est pas à cause d'un déficit en régions codantes ; au contraire, elles en contiennent plus.

5.1.4 Discussion

Les comparaisons avec des séquences colinéaires montrent deux caractéristiques spécifiques des points de cassure. D'une part, le point de cassure semble être une région étendue en général sur plusieurs dizaines de kilobases, voire plusieurs centaines de kilobases, contrairement aux cassures artificiellement créées dans les régions colinéaires. D'autre part, les séquences adjacentes au point de cassure sont plus difficilement alignables que des séquences de régions colinéaires.

Ces différences ne sont pas dues à des cas isolés de points de cassure. En effet, il peut rester des erreurs d'assignation d'orthologie qui engendrent des points de cassure pour lesquels les séquences que nous essayons d'aligner ne sont pas orthologues. Cependant, il est peu probable qu'il reste des erreurs d'assignation d'orthologie étant donné que presque tous les points de cassure alignés ont pu être segmentés de façon significative. De plus, même si on élimine les plus "mauvais" points de cassure, les différences restent significatives.

a. Artefact de la méthode ?

Si on s'attendait à ce que les points de cassure ne soient pas réduits à quelques paires de bases, le deuxième résultat est plus surprenant. On peut se demander si la différence de densité en hits dans les séquences adjacentes à la région affinée pourrait être due à un artefact de la méthode d'affinement. Si la méthode a tendance à raccourcir le segment central (le point de cassure), alors il peut rester des parties du point de cassure dans les séquences adjacentes. Or le point de cassure contient en général peu de hits. Cela peut réduire la valeur globale sur les séquences adjacentes de la densité en hits. Cet artefact devrait s'appliquer sur les deux types de régions : les cassures et les régions colinéaires. Cependant, comme le point de cassure s'étend sur une région assez grande (contrairement aux régions colinéaires), l'artefact aurait un effet plus important sur les vraies que sur les fausses cassures.

Nous avons effectué les mêmes comparaisons avec une méthode d'affinement légèrement différente. Celle-ci prend en compte les positions qui ne contiennent aucun hit (voir Chapitre 4, Section 4.3.4) et, contrairement à la méthode actuelle, a plutôt tendance à produire des segments centraux plus grands. Avec cette méthode, nous estimons peu probable que les segments extrêmes (1 et 3) contiennent une partie du point de cassure. Même dans ce cas, les différences de densité en hits entre les vraies et les fausses cassures restent significatives (la

couverture moyenne en hits sans compter le segment central est de 54 % pour les points de cassure homme-souris, contre 60 % pour les fausses cassures, p-value de 3e-9 au test de Wilcoxon).

Si les séquences voisines du point de cassure présentent une faible conservation, on peut se demander jusqu'à quelle distance du point de cassure cet effet est visible. Une perspective de ce travail serait d'analyser plus en détail la distribution des hits d'alignement autour du point de cassure. On étudierait par exemple si la densité de hits est fonction de la distance au point de cassure.

Ces résultats montrent également qu'il n'est pas aisé de définir une région de cassure. Si la cassure influence les régions voisines, quelle région faut-il considérer pour analyser ce qu'on appelle les points de cassure ?

b. Modèle de cassure

Les fausses cassures modélisent un processus de cassure où celle-ci se produit entre deux nucléotides successifs sans affecter les nucléotides voisins, et où les séquences voisines de la cassure évoluent comme si elles n'avaient pas été cassées (séquences colinéaires). De plus, le point de cassure est choisi aléatoirement dans les séquences colinéaires. Les différences observées entre les vraies et les fausses cassures montrent que ce modèle n'est pas valable pour les points de cassure des mammifères.

La taille des points de cassure suggère que la cassure affecte plus qu'un seul nucléotide. Une hypothèse serait que la région affectée par le réarrangement devient plus fragile et subit d'autres réarrangements comme des délétions, des insertions ou des petites inversions locales. Trinh *et al.* (2004) avaient proposé qu'à la suite d'un réarrangement, les séquences aux points de cassure évoluent plus vite que les autres à cause du mauvais appariement des chromosomes homologues dans les hétérocaryotypes. Le fait que les deux chromosomes homologues n'ont pas la même structure peut générer, aux points de cassure, d'autres réarrangements pendant la méïose ou durant les processus de réparation de l'ADN. Cependant, avec cette hypothèse, la divergence des séquences aux points de cassure n'est accélérée que durant la période où le réarrangement n'est pas encore fixé dans la population.

On peut également envisager que la faible similarité des séquences n'est pas une conséquence du réarrangement, mais une cause, c'est-à-dire que cette caractéristique est présente dans le génome avant l'apparition du réarrangement. On trouverait plus de réarrangements dans les régions du génome qui évoluent plus vite. D'un côté, ces régions pourraient être plus fragiles et subir plus de réarrangements que les autres. De l'autre, les réarrangements que nous observons sont ceux qui ont pu se propager et se fixer dans la population. Peut-être nous n'observons que les réarrangements se produisant dans des régions peu contraintes car ce sont les seuls à être sélectionnés au cours de l'évolution. La question de la distribution des points de cassure le long du génome sera analysée dans le Chapitre 6.

5.2 Etude du contenu des séquences de points de cassure

Nous nous concentrons ici sur les séquences des points de cassure à l'échelle du nucléotide. L'objectif de ce travail est d'identifier si les points de cassure présentent des caractéristiques spécifiques en terme de séquence et qui pourraient être reliées aux mécanismes moléculaires des réarrangements. Certaines caractéristiques ont déjà été identifiées dans la littérature, mais

les analyses ont rarement été effectuées à l'échelle d'un génome ou bien avec des points de cassure suffisamment précis.

Nous étudions le contenu des séquences de points de cassure et nous les comparons à leurs séquences adjacentes. Nous étudions ainsi des différences locales. Cela permet de s'affranchir de caractéristiques ou de biais qui seraient dus à une organisation du génome à plus grande échelle. Dans le chapitre suivant, nous verrons que les points de cassure ne sont pas uniformément répartis dans le génome et leurs positions ne sont pas indépendantes de l'organisation du génome à grande échelle.

5.2.1 Définition des séquences adjacentes

Pour chaque point de cassure, la méthode d'affinement a segmenté la séquence S_r en trois segments, le segment du milieu étant le point de cassure affiné (Figure 5.7). Les séquences adjacentes, qui serviront de point de comparaison, sont définies par les deux segments extrêmes de la segmentation. Plus exactement, elles sont délimitées par les extrémités des blocs de synténie et les limites du point de cassure affiné (Figure 5.7). Notons que les gènes orthologues aux extrémités des blocs, qui ont servi à la segmentation, ne sont pas inclus dans les séquences adjacentes.

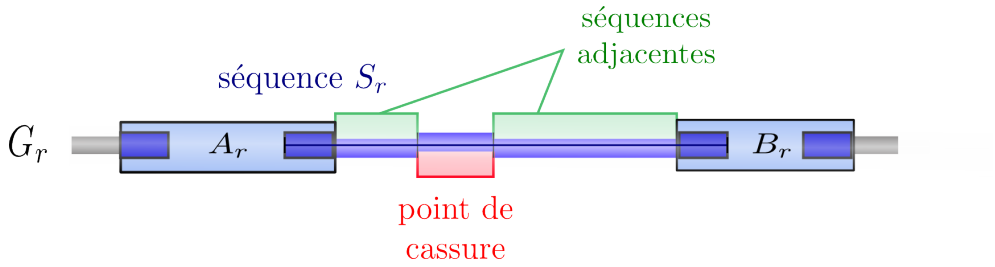


FIG. 5.7: Schéma des séquences utilisées pour analyser les caractéristiques des points de cassure : la séquence du point de cassure affinée et ses séquences adjacentes.

Les résultats sont présentés principalement pour les points de cassure entre l'homme et la souris, mais les analyses ont été effectuées sur les deux autres couples d'espèces, homme-chien et homme-chimpanzé, et ont donné des résultats similaires.

5.2.2 Composition des séquences

Nous avons calculé la composition en bases des séquences de points de cassure et de leurs séquences adjacentes. Notamment, nous nous sommes intéressés au contenu en bases G+C qui est très variable le long du génome humain.

Nous n'observons aucune différence significative de composition en base entre les séquences des points de cassure et leurs séquences adjacentes. En ce qui concerne le contenu en GC, il est de 43 % en moyenne dans les points de cassure, contre 43.5 % dans les séquences adjacentes, la différence n'est pas significative (p-value de 0.31 au test de Wilcoxon pour des données appariées).

5.2.3 Duplications segmentaires

La caractéristique des points de cassure la plus fréquemment reportée est la présence de duplications segmentaires dans les séquences de points de cassure ou à proximité (Armengol

et al., 2003; Bailey *et al.*, 2004; Murphy *et al.*, 2005). Les duplications segmentaires constituent environ 5 % du génome humain. Ce sont des séquences dupliquées en un faible nombre de copies (contrairement aux éléments transposables), très similaires (> 95 % d'identité) et assez grandes (> 1 Kb) (voir une description plus complète dans le Chapitre 1, Section 1.4.3). Les données ont été récupérées à partir du navigateur de génome UCSC (Bailey *et al.*, 2001).

Nous avons mesuré dans les séquences de points de cassure et dans leurs séquences adjacentes la proportion de bases couvertes par ces duplications segmentaires humaines (voir les distributions dans la Figure 5.8). La majorité de ces séquences ne possède aucune duplication segmentaire. Cependant, parmi les séquences possédant des duplications segmentaires, on observe une sur-représentation des points de cassure par rapport aux séquences adjacentes. Ainsi, 66 séquences de points de cassure correspondent presque entièrement à une ou plusieurs duplications segmentaires (séquences couvertes à plus de 80 % par des duplications segmentaires).

On observe que parmi ces 66 points de cassure, 82 % sont également présents dans la comparaison homme-chien, alors que sur l'ensemble des points de cassure homme-souris seulement 36 % sont communs à la comparaison homme-chien. Cela suggère que ces réarrangements se sont produits dans la lignée humaine et renforce l'hypothèse d'un lien de cause à effet entre réarrangements et duplications. Notons cependant que pour mettre en cause le mécanisme de recombinaison homologue entre copies d'une duplication, il serait nécessaire d'analyser la localisation des différentes copies des duplications par rapport aux réarrangements.

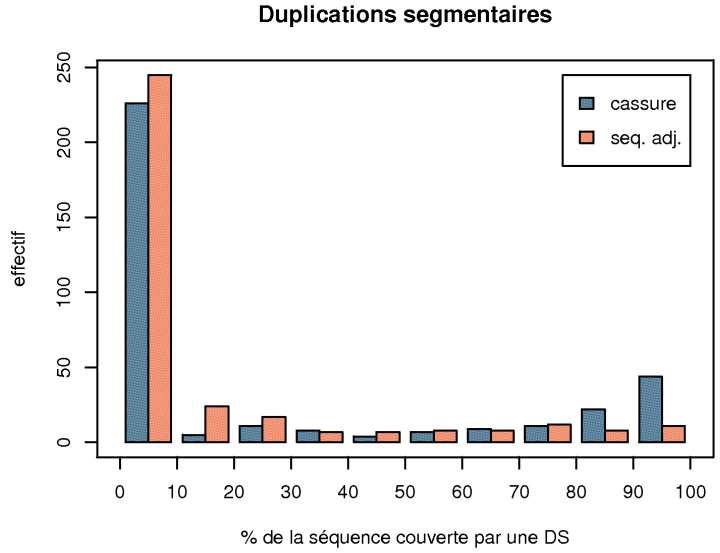


FIG. 5.8: Comparaison du contenu en duplications segmentaires des séquences de points de cassure (en bleu) et de leurs séquences adjacentes (en orange).

Ces résultats confirment ceux de la littérature en montrant qu'un certain nombre de points de cassure sont associés à des duplications segmentaires. Toutefois, nous ne pouvons comparer nos chiffres avec ceux de la littérature car les seuils de détection utilisées (taille des duplications segmentaires considérées, taille de la région de cassure considérée, etc.) ne sont pas les mêmes.

Par rapport aux résultats publiés, on observe ici que localement les points de cassure

possèdent plus de duplications segmentaires que leurs séquences adjacentes. Si on met bout à bout les séquences de points de cassure d'une part, et leurs séquences adjacentes d'autre part, les séquences de points de cassure sont couvertes à 38 % par des duplications segmentaires, contre 16 % pour les séquences adjacentes. Nous rappelons qu'il existe principalement deux hypothèses pour expliquer cette association : soit les duplications sont impliquées dans le mécanisme de réarrangement, par recombinaison ectopique entre différentes copies (NAHR), soit les deux évènements sont indépendants et ces régions sont des régions fragiles pour les réarrangements et pour l'insertion des duplications segmentaires. Dans la première hypothèse, il est nécessaire de trouver les duplications exactement aux points de cassure, alors que dans la deuxième cela dépend de l'étendue de la zone de fragilité. Ainsi, les résultats que nous obtenons sont plutôt en faveur de la première hypothèse, ou bien impliquent que la fragilité des régions est limitée en distance. Cependant, on remarque aussi que les séquences adjacentes contiennent plus de duplications segmentaires que la moyenne du génome (5.4 %). Cela implique que la première hypothèse (NAHR) n'est pas la seule explication de cette association. Une troisième hypothèse expliquant la présence de duplications au voisinage des points de cassure suggère que les duplications permettent de mettre à proximité dans le noyau les régions qui vont interagir dans le réarrangement (Kehrer-Sawatzki et Cooper, 2008).

5.2.4 Eléments répétés

Les éléments répétés sont des éléments majeurs du génome humain puisqu'ils constituent presque la moitié de celui-ci. De plus, ces séquences peuvent être impliqués dans des cassures de l'ADN et des réarrangements. Nous distinguons les éléments transposables des répétitions simples.

a. Eléments transposables

Les éléments transposables peuvent être associés aux réarrangements chromosomiques pour plusieurs raisons. La première est qu'ils peuvent représenter de bons substrats pour la recombinaison ectopique, puisqu'ils sont présents en un très grand nombre de copies dans le génome. Ensuite, leur mécanisme de transposition peut générer des réarrangements (principalement des délétions et des transpositions de régions flanquantes, voir Chapitre 1) ; ils génèrent également des cassures double brins qui peuvent être "mal" réparées. Enfin, s'il existe des régions fragiles dans le génome, on peut imaginer qu'elles sont plus propices à l'insertion d'éléments transposables.

Des éléments transposables ont été identifiés dans des points de cassure de manière ponctuelle (des éléments de type LINE (Goidts *et al.*, 2005) ou SINE (Dennehey *et al.*, 2004; Kehrer-Sawatzki *et al.*, 2002; Goidts *et al.*, 2005), ou LTR (Kehrer-Sawatzki *et al.*, 2002, 2005a)). Des analyses sur un plus grand nombre de points de cassure ont également montré certaines tendances : un enrichissement en éléments de types LINE et LTR dans 15 points de cassure du chromosome 19 de l'homme (Dehal *et al.*, 2001), l'analyse de 198 points de cassure entre les génomes de l'homme et de la vache a montré un enrichissement en éléments transposables spécifiques de l'espèce ayant subit le réarrangement (Schibler *et al.*, 2006).

Grâce au programme RepeatMasker, on peut identifier et classer les éléments transposables présents dans les séquences nucléiques.

Dans le Tableau 5.3 sont résumées les couvertures moyennes en éléments transposables des séquences de points de cassure et de leurs séquences adjacentes. Globalement, les séquences de points de cassure possèdent plus d'éléments transposables que leurs séquences adjacentes

(cette différence est significative au test de Wilcoxon pour des données appariées). Cependant, la différence moyenne deux à deux est faible (0.077) et si une majorité de points de cassure possèdent plus d'éléments transposables que leurs séquences adjacentes respectives (66 %), cela n'est pas systématique (voir la Figure 5.9 graphique de droite). Ainsi, il semblerait que la différence observée soit essentiellement due à quelques points de cassures très enrichis en éléments transposables. On observe, par exemple, sur la Figure 5.9 (graphique de gauche) que 47 points de cassure sont couverts à plus de 70 % par des éléments transposables.

Si on distingue les grandes classes d'éléments transposables, ce sont les éléments LTRs qui montrent la plus forte différence entre les deux types de séquences (voir Tableau 5.3).

	Eléments transposables				
	total	SINEs	LINEs	LTRs	transposons
Points de cassure	0.520	0.195	0.180	0.112	0.024
Seq. adjacentes	0.443	0.177	0.161	0.076	0.025
p-value	9e-12	0.06	0.015	8e-8	0.13

TAB. 5.3: Couverture moyenne des séquences en éléments transposables. Un test de Wilcoxon pour des données appariées a été effectué pour chaque classe d'ET.

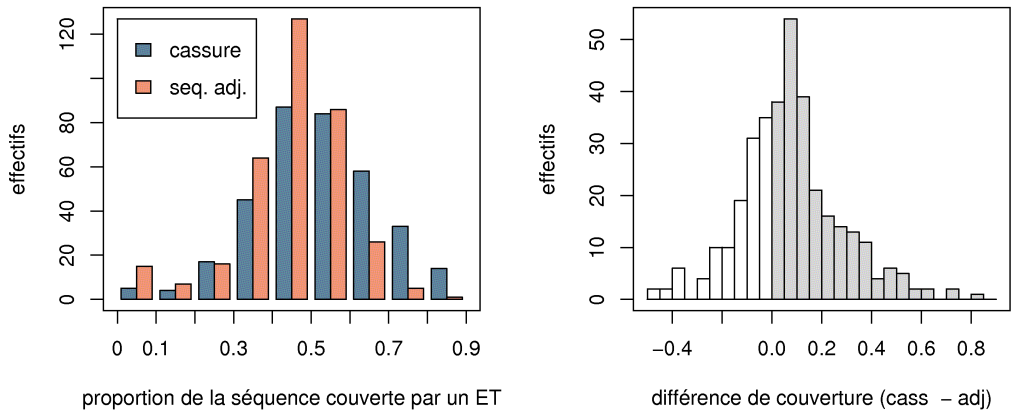


FIG. 5.9: Distribution des couvertures des séquences en éléments transposables pour les deux types de séquences, les points de cassure (en bleu) et leurs séquences adjacentes (en orange). A droite est représentée la distribution de la différence de couverture deux à deux, les barres grisées correspondent aux points de cassure possèdant plus d'éléments transposables que leurs séquences adjacentes respectives.

b. Répétitions simples

Les répétitions en tandem de petites séquences sont des répétitions assez courantes dans le génome humain et sont des facteurs de variabilité. Elles pourraient également être responsables de fragilités locales de l'ADN et ainsi être liées aux cassures et aux réarrangements

(lire la revue (Usdin, 2008)). Les palindromes, quant à eux, sont capables de former des structures secondaires de l'ADN. Ces structures pourraient également fragiliser localement l'ADN, notamment au moment de la réplication engendrant des cassures double brin (Aguilera et Gómez-González, 2008). Plusieurs points de cassure de maladie ou de cancer ont également été localisés très précisément au centre de palindromes riches en AT (voir Section 1.4.3d.).

Ces deux types de répétitions sont facilement détectables dans les séquences. Le programme Tandem Repeat Finder (Benson, 1999) permet d'identifier les répétitions en tandem. Les résultats sur tout le génome humain sont disponibles sous forme de tables dans le navigateur de génome de l'UCSC (Karolchik *et al.*, 2008). On peut donc analyser le contenu des séquences en répétitions en tandem. Pour identifier des séquences palindromiques nous avons utilisé l'outil PALINDROME de la suite EMBOSS (Rice *et al.*, 2000).

Pour ces deux types de séquences, nous calculons leur couverture dans les régions étudiées (le nombre de bases qu'elles couvrent rapporté à la longueur de la région étudiée). Nous comparons ces valeurs entre les séquences de point de cassure et leurs séquences adjacentes.

	Couverture moyenne des séquences en	
	rep. en tandem	palindromes
Points de cassure	0.0247	0.0090
Seq. adjacentes	0.0206	0.0083
p-value	0.053	0.77

TAB. 5.4: Couvertures moyenne des séquences en répétitions en tandem et en séquences palindromiques, pour les points de cassure et pour leurs séquences adjacentes.

Les résultats sont présentés dans le Tableau 5.4. Les différences de moyenne sont très faibles et ne sont pas significatives au test de Wilcoxon pour des données appariées (p-value de 0.05 pour les répétitions en tandem).

5.2.5 Discussion et perspectives

Grâce à la précision des points de cassure, on a pu distinguer les séquences des points de cassure de leurs séquences immédiatement adjacentes et étudier localement des différences de contenu entre ces séquences. Nous avons ainsi confirmé plusieurs tendances déjà mentionnées dans la littérature : la présence de duplications segmentaires et un enrichissement en éléments transposables.

En ce qui concerne des caractéristiques de composition des séquences ou de courtes répétitions simples ou palindromiques, nous n'avons pas mis en évidence de différence significative. Il serait peut-être plus judicieux de chercher plus spécifiquement des motifs connus qui pourraient jouer un rôle dans les mécanismes de réarrangements ou dans la stabilité de la molécule d'ADN. Par exemple, on connaît certains motifs associés aux hotspots de recombinaison, des motifs de fixation de protéines associées à la structure de l'ADN (topo-isomérases) ou des motifs spécifiques des sites fragiles (voir Section 1.4.3d.). Enfin, il est probable que de telles traces (répétitions courtes et palindromes), si elles existaient au moment du réarrangement, ont été perdues avec l'évolution des séquences et ne sont plus détectables à l'heure actuelle. En effet, la plupart des évènements que nous regardons sont anciens de plusieurs millions d'années, voire dizaines de millions d'années.

5.3 Duplications aux points de cassure

Nous nous intéressons ici aux cas où le point de cassure correspond à une duplication dans le génome G_o dont les deux copies se trouvent dans les séquences S_{oA} et S_{oB} (voir Figure 5.10).

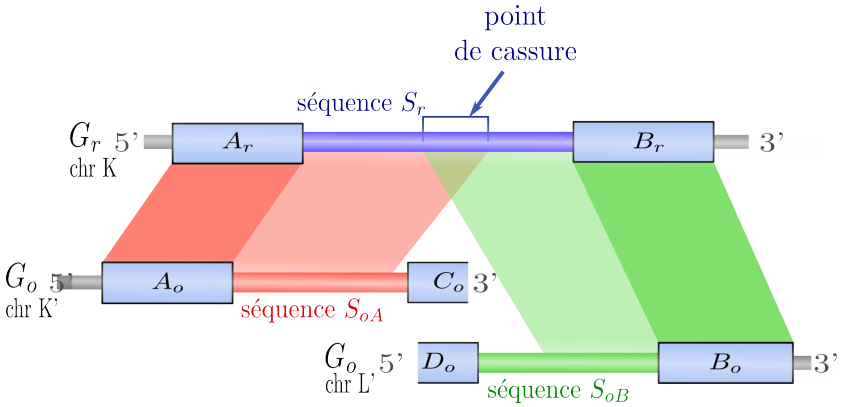


FIG. 5.10: Exemple schématique de point de cassure qui s'aligne avec les deux séquences S_{oA} et S_{oB}. Le point de cassure est donc une séquence qui est dupliquée dans le génome G_o.

Ces duplications sont intéressantes pour deux raisons. La première est qu'elles expliquent, pour ces points de cassure, leur taille. Si les parties orthologues sur S_r avec les deux autres séquences se chevauchent, on ne peut affiner le point de cassure dans la zone de chevauchement (la duplication).

La deuxième raison est qu'on peut potentiellement faire un lien avec le mécanisme de réarrangement qui a eu lieu. Nous avons vu dans le Chapitre 1 que les duplications peuvent causer des réarrangements par le mécanisme de recombinaison homologue non allélique (NAHR), ou bien en être la conséquence par le mécanisme de complétion des bouts collant de cassures double brin (voir Section 1.4.3a. et Figure 1.6). Les duplications que nous observons au point de cassure sont des bons candidats pour ces deux mécanismes car elles proviennent des deux autres points de cassure orthologues (les séquences S_{oA} et S_{oB} sont des points de cassure dans le génome G_o) impliqués dans le réarrangement. Elles constituent donc de meilleurs indices que les analyses précédentes sur les duplications segmentaires humaines. Cependant, pour mettre en évidence ces liens, il faut tout de même s'assurer que le réarrangement s'est produit dans la lignée qui possède les duplications. Si ce n'est pas le cas, la présence de duplications peut suggérer un point de cassure ré-utilisé dans différentes lignées, dans l'une pour le réarrangement étudié et dans l'autre pour la duplication.

5.3.1 Détection des duplications

a. Avec la méthode d'affinement

La méthode d'affinement permet de détecter ce type de duplication puisque nous disposons de la répartition des hits rouges et verts (respectivement avec les séquences S_{oA} et S_{oB}) le long de la séquence S_r. Il est possible alors de compter les positions sur S_r qui sont couvertes à la fois par des hits rouges et des hits verts. On pourra suggérer la présence d'une duplication au point de cassure si ce dernier contient un grand nombre de ces positions. Cependant, un

ensemble de positions proches sur S_r qui sont couvertes par les deux types de hits ne signifie pas forcément la présence d'une séquence dupliquée dans les deux séquences S_{oA} et S_{oB}. En effet, comme on ne tient pas compte de l'ordre et de l'orientation des hits dans les séquences de G_o, on ne peut pas être sûr que toutes ces positions correspondent à une même région dans ces séquences. Par exemple, une région très répétée ou de faible complexité qui n'a pas été masquée par RepeatMasker peut être couverte par les deux types de hits, mais les hits ne sont pas colinéaires et peuvent correspondre à des séquences non orthologues.

Ainsi, les données issues du processus d'affinement peuvent être utilisées pour détecter automatiquement des duplications potentielles. Il est ensuite nécessaire de vérifier la présence des duplications. On peut par exemple, représenter graphiquement les hits d'alignement dans les séquences, sous forme de dotplot (voir l'exemple de la Figure 5.11), ou bien aligner directement les deux copies potentielles sur G_o.

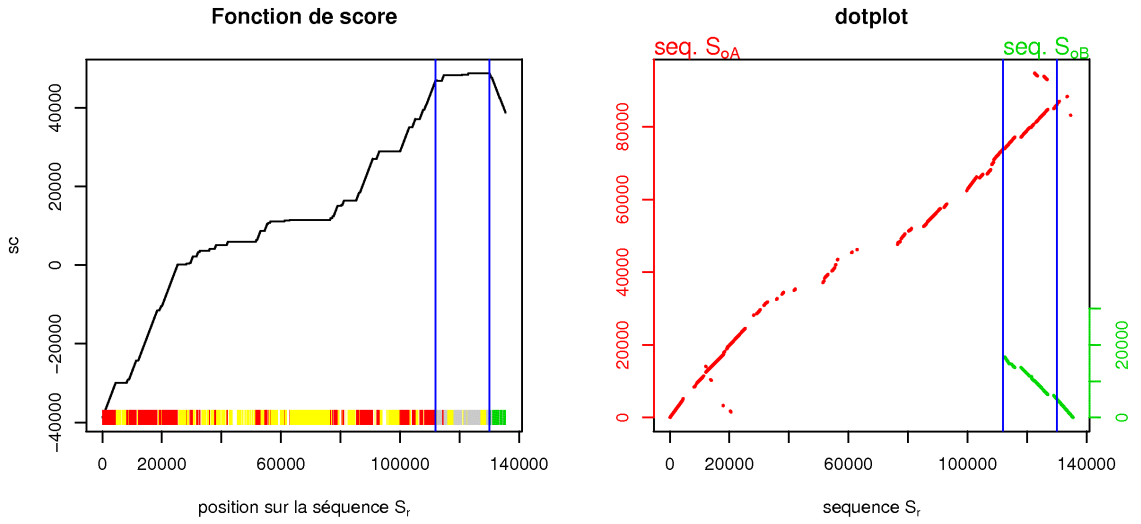


FIG. 5.11: Exemple de point de cassure entre l'homme et la souris. A gauche, la représentation de la fonction de score *sc* montre que le point de cassure affiné (délimité par des droites verticales bleues) contient un grand nombre de positions couvertes par les deux types de hits (barres grises en dessous de la courbe). A droite, la représentation en dotplot confirme la duplication. Les hits sont représentés sous forme de dotplot, en rouge les hits entre S_r et S_{oA} et en vert entre S_r et S_{oB}.

Pour identifier les candidats de duplication, nous mesurons dans le point de cassure affiné la proportion de positions couvertes par les deux types de hits à la fois, c'est-à-dire le nombre de positions couvertes par un hit rouge et un hit vert divisé par le nombre de positions couvertes par l'un ou l'autre type de hit. Lorsque ce chiffre dépasse un certain seuil, nous considérons le point de cassure comme candidat de duplication. Le seuil a été fixé à 0.3 pour la comparaison homme-souris, en analysant manuellement les points de cassure et leurs dotplots.

b. Duplications dans les points de cassure homme-souris

Parmi les 354 points de cassure homme-souris, nous avons identifié 57 candidats de duplication au point de cassure. L'analyse graphique des dotplots suggère qu'au moins 60 % d'entre eux correspondent à des séquences dupliquées dans les séquences S_{oA} er S_{oB}, et non à des répétitions manquées par RepeatMasker.

La taille de la duplication est variable. La taille moyenne des points de cassure ayant une duplication est de 82 Kb, mais si on ne considère que les positions du point de cassure couvertes par les deux types de hits, ce chiffre varie de 44 pb à 83 Kb avec une valeur moyenne de 8.5 Kb. Cela ne tient pas compte des positions masquées ou ne contenant aucun hit qui sont comprises dans la séquence dupliquée. Les tailles observées sont vraisemblablement sous-estimées, car si l'une des copies (sur S_{oA} ou sur S_{oB}) s'aligne mieux que l'autre sur S_r, alors il arrive souvent que la segmentation assigne certaines portions de la duplication dans les segments extrêmes plutôt que dans le point de cassure.

Notons que 95 % de ces points de cassure (54 points) sont absents de la comparaison homme-chien, ce qui suggère que les réarrangements impliqués se sont produits dans la lignée de la souris plutôt que dans celle de l'homme (si on considère la divergence homme-chien plus ancienne que celle avec les rongeurs). Les réarrangements et les duplications se sont donc produits dans la même lignée. Il est possible, dans ce cas, qu'il y ait un lien de cause à effet entre les duplications et les réarrangements.

Si ces cas de duplications sont assez marginaux dans la comparaison homme-souris, ils sont plus fréquents dans les points de cassure homme-chimpanzé. En étudiant manuellement les 36 points de cassure homme-chimpanzé, nous avons identifié une duplication d'au moins 1 Kb pour au moins 15 d'entre eux (soit 42 %). On peut supposer que les cas de duplications sont d'autant plus faciles à détecter que les évènements sont récents.

c. Perspectives

La méthode d'affinement permet d'identifier des candidats de points cassure correspondant à des duplications dans l'autre génome, mais il reste à confirmer ces candidats et à mieux délimiter la région dupliquée. On pourrait donc envisager une méthode similaire à la méthode d'affinement mais dédiée à la détection de ces duplications. Par exemple, le modèle multinomial présenté à la Section 4.3.4c. du Chapitre 4 pourrait être plus efficace dans ce but, puisqu'il permet de distinguer les positions couvertes par aucun hit de celles couvertes par les deux types de hit.

Enfin, pour établir des liens de cause à effet entre les duplications et les réarrangements, voire même déterminer les mécanismes responsables, des comparaisons avec d'autres espèces sont nécessaires pour identifier la lignée dans laquelle se sont produits les réarrangements ainsi que l'âge des duplications.

5.3.2 Le cas des chromosomes XY

Nous avons appliqué la méthode d'affinement des points de cassure dans le cadre de l'étude des réarrangements différenciant les chromosomes sexuels humains. Nous avons notamment identifié des duplications dans certains points de cassure qui apportent des arguments en faveur d'une hypothèse controversée sur l'évolution des chromosomes sexuels : celle de la formation du chromosome Y par inversions successives.

Le travail que nous présentons ici fait partie d'une étude plus complète des chromosomes X et Y, incluant notamment l'analyse des scénarios d'inversions produits par des algorithmes de tri de permutations. L'ensemble du travail a été effectué en collaboration avec Marília Braga et Gabriel Marais et fait l'objet d'un article soumis (Lemaitre *et al.*, 2008a).

a. Les strates du chromosome X

Il y a environ 300 millions d'années, notre ancêtre ne possédait vraisemblablement pas de chromosomes sexuels. On suppose que les chromosomes X et Y proviennent d'une paire d'autosomes qui se sont différenciés et ont arrêté de recombiner. En fait, les chromosomes sexuels actuels ne peuvent recombiner que sur une petite région appelée région pseudo-autosomale.

La différenciation des deux chromosomes sexuels s'est faite petit à petit. En 1999, Lahn et Page observent que la divergence des gènes homologues entre les chromosomes X et Y est corrélée à leur position sur le chromosome X avec une forme en escalier (voir Figure 5.12b). Cela suggère que l'arrêt de la recombinaison entre les deux chromosomes s'est fait par plusieurs grandes étapes successives, formant des **strates évolutives** sur le chromosome X (une strate est un groupe de gènes qui ont commencé à diverger au même moment). Observant que l'ordre des gènes homologues sur les chromosomes X et Y est très différent, ils proposent que les strates évolutives ont été formées par des inversions successives sur le chromosome Y (voir Figure 5.12). Notons que la partie du chromosome X concernée est colinéaire à son chromosome orthologue chez le poulet, suggérant que les réarrangements qui différencient X et Y se sont produits sur le chromosome Y.

(a) Histoire évolutive des chromosomes X et Y.

(b) Divergence des gènes homologues entre X et Y (Ks) en fonction de leur position sur le chromosome X, définissant 4 strates évolutives (dénotées "group").

FIG. 5.12: Hypothèse de la différenciation du chromosome Y par des inversions successives, images tirées de (Lahn et Page, 1999).

Cette hypothèse a été ensuite remise en question, notamment par d'autres analyses de la divergence de marqueurs homologues ne présentant plus la forme en escalier de Lahn et Page (Iwase *et al.*, 2003; Skaletsky *et al.*, 2003). Les limites des strates n'étant plus claires, on a alors proposé que l'arrêt de la recombinaison avait été progressif et n'était pas dû aux inversions, ces dernières étant postérieures. Plus récemment, avec le séquençage complet du chromosome

X humain, le débat a été relancé. Ross *et al.* (2005) ont analysé l'ordre et l'orientation de marqueurs homologues sur les chromosomes X et Y et ont identifié notamment un scénario d'inversions parcimonieux sur le chromosome Y qui explique l'ordre des marqueurs orthologues et confirme la formation des trois dernières strates (strates 3, 4 et 5).

b. Duplications aux bornes des strates

Nous avons appliqué la méthode d'affinement sur les points de cassure du chromosome X comparé au chromosome Y chez l'homme. Nous nous sommes notamment intéressés aux points de cassure correspondant aux limites des strates sur le chromosome X.

Pour deux limites de strates, celle entre la région pseudo-autosomale (PAR) et la strate 5 et celle entre la strate 5 et la strate 4, nous avons identifié des duplications présentes dans les séquences orthologues du chromosome Y (voir le schéma de la Figure 5.13).

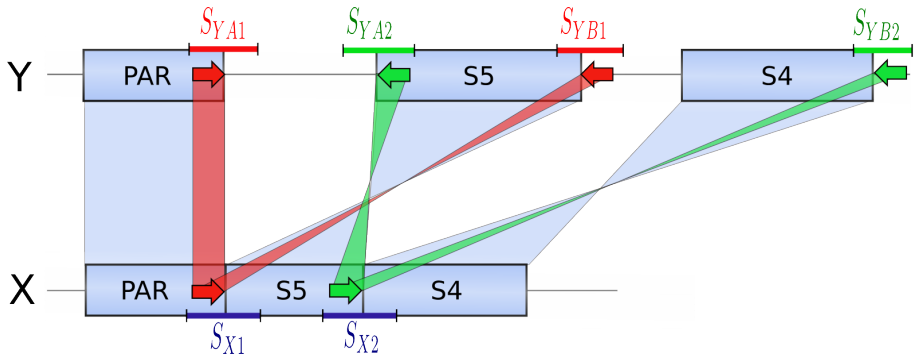


FIG. 5.13: Schéma représentant les trois régions PAR, strate 5 et strate 4 sur les chromosomes X et Y (rectangles bleus), ansi que les séquences utilisées pour affiner deux points de cassure sur le chromosome X (segments horizontaux) et les duplications détectées (flèches épaisses). La séquence S_{X1} correspond à la limite entre PAR et la strate 5 sur le chromosome X, elle a été affinée avec les séquences S_{YA1} et S_{YB1} sur le chromosome Y. La séquence S_{X2} correspond à la limite entre les strates 4 et 5 sur le chromosome X, elle a été affinée avec les séquences S_{YA2} et S_{YB2} sur le chromosome Y. Le schéma n'est pas à l'echelle.

Après avoir identifié les duplications avec la méthode d'affinement, nous avons aligné les séquences dupliquées sur le chromosome Y deux à deux avec un algorithme d'alignement exact afin de mesurer leur niveau de similarité. Nous avons utilisé les programmes WATER et MATCHER, qui sont des implémentations de l'algorithme de Smith et Waterman distribués dans la suite EMBOSS (Rice *et al.*, 2000). Nous avons également aligné la séquence correspondant à la duplication sur le chromosome X contre tout le chromosome Y pour s'assurer que les copies sur le Y ne sont présentes qu'aux points de cassure (voir les dotplots obtenus dans la Figure 5.14). Enfin, nous avons vérifié que ces séquences n'étaient pas dupliquées sur le chromosome X. Ces alignements appliqués sur les chromosomes entiers ont été effectués avec le programme Blastz sur les séquences masquées de leurs éléments répétés par RepeatMasker.

La duplication correspondant à la limite PAR/strate 5 sur le chromosome X est une séquence de 45 Kb environ dont les deux copies sur le chromosome Y sont sur des brins opposés (voir Figure 5.14 graphique de gauche). L'alignement exact de ces deux copies est long de 44 Kb et présente 70 % d'identité et 20 % d'indels.

La duplication correspondant à la limite entre les strates 4 et 5 sur le chromosome X est une séquence de 110 Kb environ. Les deux copies sur le chromosome Y sont sur le même brin et leur alignement exact présente 55 % de bases identiques et 19 % d'indels sur une longueur de 103 Kb (voir Figure 5.14 graphique de droite).

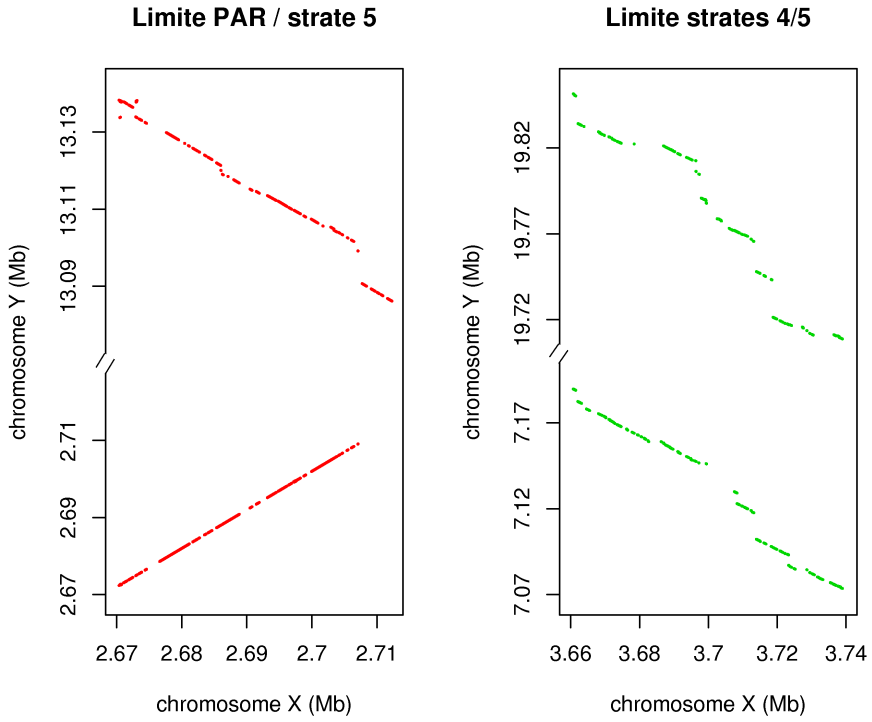


FIG. 5.14: Dotplots des alignements des limites de strates sur le chromosome X avec le chromosome Y. A gauche la séquence à la limite de PAR et de la strate 5 sur le chromosome X a été alignée contre tout le chromosome Y. Cette séquence s'aligne avec deux régions sur le chromosome Y. A droite, la limite entre les strates 4 et 5 sur le chromosome X s'aligne également avec deux régions sur le chromosome Y. Les alignements ont été obtenus avec le programme Blastz sur les séquences masquées de leurs éléments répétés.

c. Implications dans l'évolution des chromosomes sexuels

Les séquences dupliquées sur le chromosome Y, correspondant à la limite PAR/strate 5 sur le chromosome X, se trouvent aux points de cassure de l'inversion supposée ayant créé la strate 5. De plus, elles sont sur des brins opposés. Elles pourraient donc constituer des traces de cette inversion. En effet, les inversions sont souvent trouvées associées à des duplications inversées aux points de cassure. Deux modèles ont été proposés, le premier fait appel au mécanisme de recombinaison homologue entre deux copies d'une duplication antérieure à l'inversion. Le deuxième propose que les duplications sont créées au moment du réarrangement par complétion des brins d'ADN simples résultant de cassures double brin à bouts collants. Dans notre cas, les duplications sont présentes seulement sur le chromosome ayant subi l'inversion. Cela pourrait suggérer que les duplications ne sont pas antérieures à l'inver-

sion et serait plutôt en faveur du deuxième modèle. Cependant, on peut aussi imaginer que la duplication est antérieure à l'inversion, mais n'est pas présente sur le chromosome X car la deuxième copie s'est produite dans la région non recombinante du chromosome Y, c'est-à-dire après la strate 5 (voir schéma Figure 5.15). Ainsi, les duplications identifiées sont compatibles avec les deux modèles d'inversion, mais ne permettent pas de trancher.

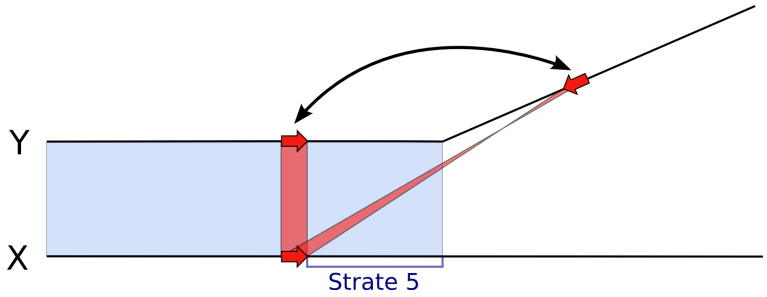


FIG. 5.15: Schéma illustrant l'hypothèse où la duplication (flèches rouges) est antérieure à l'inversion de la strate 5. L'ombre bleue délimite la région pseudo-autosomale (recombinante). La deuxième copie de la duplication n'est présente que sur le chromosome Y, car elle se trouve dans une région qui ne recombine plus avec le chromosome X. Il peut ensuite se produire de la recombinaison ectopique entre les deux copies sur le chromosome Y, résultant en l'inversion de la strate 5 (double flèche noire).

En ce qui concerne la duplication de la limite entre les strates 4 et 5, on observe que les deux copies sont sur le même brin. Cela semble incompatible avec un modèle d'inversion associant le réarrangement et les duplications. Cependant, les copies sont sur le même brin sur le chromosome Y actuel, mais si on suppose que l'inversion de la strate 5 est postérieure à celle de la strate 4, les duplications se retrouvent sur des brins opposés au moment de l'inversion de la strate 4 (voir le scénario d'inversions proposé dans la Figure 5.16). Ainsi, ces duplications pourraient être des traces de l'inversion de la strate 4, et elles confirment également l'ordre dans lequel se sont produites les deux inversions de la strate 4 et de la strate 5. Les niveaux de divergence observés entre les copies dupliquées sont également en faveur de cet ordre (strate 4 avant la strate 5), puisque les copies de la limite PAR/strate 5 sont plus similaires entre elles que celles de la limite strates 4/5. Ces dernières semblent donc être plus anciennes.

En conclusion, nous avons identifié des duplications constituant des traces potentielles d'inversions. Il est donc très probable que ces deux inversions ont eu lieu sur le chromosome Y dans l'ordre suivant : d'abord l'inversion incluant la strate 4 puis celle incluant la strate 5. Celles-ci sont en accord avec le modèle de formation des strates 4 et 5 par deux inversions successives, puisque les points de cassure de ces inversions se trouvent exactement à la limite des strates. On ne peut cependant pas affirmer avec certitude que ces inversions sont responsables de l'arrêt de la recombinaison dans ces strates. Il est clair que, après de telles inversions, la recombinaison est fortement réduite, voire nulle dans les segments inversés. Mais il est possible que la recombinaison ait été stoppée avant l'inversion. Cependant, dans ce cas, on ne sait expliquer pourquoi les points de cassure des inversions se trouvent justement aux limites des strates, c'est-à-dire de l'arrêt de la recombinaison.

Ainsi, il est très probable que ces deux dernières strates (la strate 4 et la strate 5) aient été formées grâce (ou à cause) de deux inversions sur le chromosome Y. En ce qui concerne les autres strates plus anciennes, nous n'avons pas détecté de traces laissées par des inversions, telles que des duplications. Cela ne veut pas dire qu'elles n'ont pas eu lieu, puisque toutes les

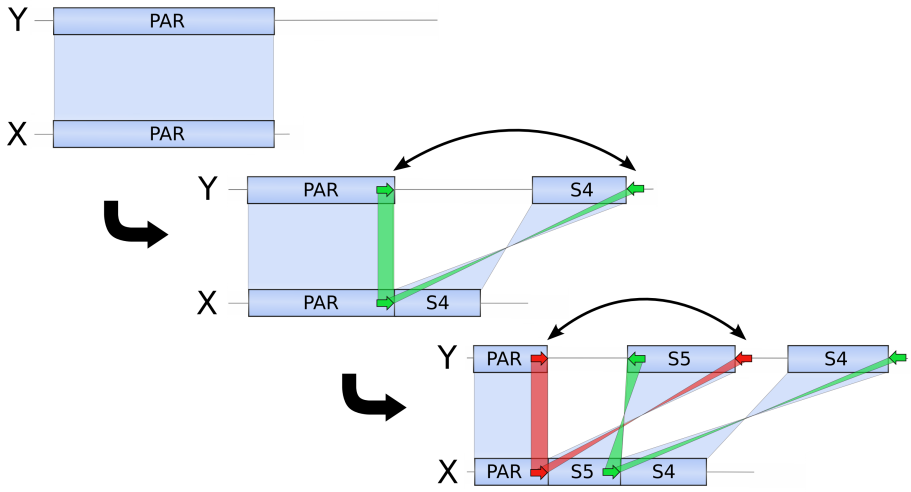


FIG. 5.16: Représentation schématique du scénario d'inversions proposé ayant formé les strates 4 et 5. Les deux inversions sont représentées par les doubles flèches noires et les duplications sont montrées par des flèches épaisses, en vert pour celles de la limite strates 4/5 et en rouge pour celles de PAR/strate5.

inversions ne génèrent pas, ou ne sont pas dues à des duplications. De plus, les séquences sont de moins en moins similaires lorsqu'on s'éloigne de la PAR sur le chromosome X et ont subi de nombreux réarrangements sur le chromosome Y, ce qui rend la détection des réarrangements et des similarités très difficiles.

5.4 Conclusion et perspectives

Dans ce chapitre, nous avons montré que les points de cassure ne sont pas des régions du génome comme les autres qui simplement auraient été "cassées" entre deux nucléotides successifs puis recollées ailleurs. En effet, ces régions sont plus grandes qu'attendu et leurs séquences voisines présentent elles aussi une conservation plus faible que le reste du génome. En comparant systématiquement les séquences des points de cassure avec leurs séquences adjacentes, nous avons pu mettre en évidence certaines caractéristiques locales qui pourraient être reliées aux évènements de réarrangements. Notamment, nous observons que certains points de cassure sont enrichis en séquences dupliquées dans l'un ou l'autre génome comparé, ainsi qu'en éléments transposables.

Cependant, pour établir des liens de cause à effet entre ces caractéristiques de séquence et les évènements de réarrangement, il apparaît indispensable de tenir compte de l'origine évolutive de ces derniers. Dans le cas des inversions du chromosome Y, c'est cette connaissance qui a permis d'émettre des hypothèses sur les mécanismes de réarrangement. Ainsi, il serait intéressant de distinguer, dans les analyses systématiques des points de cassures, ceux qui correspondent à des régions qui ont effectivement été cassées dans la lignée étudiée, de ceux qui sont seulement orthologues à de telles régions. Ces deux types de séquences sont intéressantes à analyser, dans un cas on se place après l'évènement de cassure, dans l'autre avant. En les distinguant, on pourrait peut-être identifier des caractéristiques différentes. Les premières seraient la conséquence des réarrangements et les deuxièmes la cause. Par exemple, dans le cas de la perte de similarité autour des points de cassure, l'étude de la conservation de ces

séquences avec d'autres génomes et en fonction de l'origine des réarrangements pourrait peut-être permettre d'établir si cette caractéristique est une conséquence du réarrangement, où si les réarrangements se produisent préférentiellement dans des régions qui évoluent plus vite.

Pour éviter d'autres effets de moyenne, il pourrait être intéressant de distinguer les différents types de réarrangements, comme les inversions, les translocations, les transpositions, etc. Si des caractéristiques différentes étaient mises en évidence, elles pourraient nous éclairer sur les mécanismes moléculaires de ces différents évènements. Réciproquement, des caractéristiques spécifiques pourraient permettre d'inférer les différents types de réarrangements simplement par l'analyse de la séquence. Enfin, d'autres distinctions pourraient être envisagées, en fonction de l'âge des réarrangements par exemple, ou bien du niveau de ré-utilisation des régions de cassure. On peut ainsi imaginer que les régions subissant fréquemment des réarrangements dans différentes lignées possèdent des particularités de séquence qui seraient conservées au cours de l'évolution et qui seraient absentes des autres points de cassure.

Chapitre 6

Réarrangements et organisation génomique

Sommaire

Dans ce chapitre, nous nous intéressons à la distribution des points de cassure le long du génome. Le modèle de cassures aléatoires proposé par Nadeau et Taylor dans les années 80 suppose que les points de cassure sont distribués uniformément et indépendamment dans le génome. Cependant, il a été critiqué à de nombreuses reprises. Notamment, l'observation de régions concentrant un grand nombre de points de cassure a conduit à proposer un autre modèle de réarrangement : le modèle des régions fragiles. Dans ce modèle, il existerait le long du génome des régions plus fragiles que d'autres qui concentreraient plus de points de cassure. Si on observe effectivement une distribution non uniforme des points de cassure, on ne sait toujours pas ce qui détermine cette distribution et quels facteurs génomiques influencent la position de ces points.

Dans cette problématique, des corrélations entre la distribution des points de cassure et d'autres structures génomiques ont été recherchées. La corrélation la plus forte mise en évidence porte sur les duplications segmentaires mais elle ne concerne pas tous les points de

cassure et elle ne permet pas d'expliquer la localisation de ces derniers. En fait, il semblerait que les régions fragiles pour les réarrangements soient également des régions fragiles pour l'insertion des duplications segmentaires (voir Chapitre 1, Section 1.4.3). D'autres caractéristiques comme les sites fragiles, divers éléments répétés, le contenu en GC, la densité en gènes ont été étudiées dans la littérature, mais ces études sont limitées, soit à certains points de cassure ou certaines régions génomiques (Gordon *et al.*, 2007; Webber et Ponting, 2005), soit par la résolution des points de cassure considérés (Murphy *et al.*, 2005; Schibler *et al.*, 2006; Ruiz-Herrera *et al.*, 2006).

Nous espérons, grâce à la meilleure résolution de nos points de cassure, pouvoir apporter de nouveaux éléments de réponse à ces questions. Nous nous proposons donc d'étudier la distribution des points de cassure le long du génome humain en fonction de plusieurs structures génomiques. Nous nous intéressons tout d'abord à la distribution des gènes et des régions codantes, puis, à plus grande échelle, à la structuration du génome en isochores et en domaines de réplication.

6.1 Jeux de données

6.1.1 Les points de cassure

Les points de cassure deux à deux entre l'homme et une autre espèce sont limités, soit en nombre, soit en précision. Entre l'homme et la souris, on dispose au plus de 350 points de cassure dont la moitié seulement ont une taille inférieure à 50 Kb. Les comparaisons avec des génomes primates donnent des points de cassure plus précis mais beaucoup moins nombreux.

Afin de disposer d'un jeu de données plus conséquent et plus précis, nous avons utilisé les points de cassure de plusieurs comparaisons deux à deux. Le génome de l'homme reste le génome de référence et nous étudierons la localisation des points de cassure sur ce génome (assemblage de Mai 2004, NCBI35 ou hg17). Il a été comparé avec 5 autres génomes de mammifères :

- le chimpanzé (assemblage de mars 2006 ou panTro2.1),
- le macaque (assemblage de février 2006 ou rheMac2),
- la souris (assemblage de décembre 2005, NCBIm35 ou mm7),
- le rat (assemblage de décembre 2004 ou rn4),
- le chien (assemblage de mai 2005 ou canFam2).

Nous ne pouvons simplement concaténer les 5 jeux de coordonnées de points de cassure sur le génome humain, car certains points de cassure correspondent à un même évènement de réarrangement qui est visible dans plusieurs comparaisons deux à deux. Par exemple, si l'ancêtre de la souris et du rat a subi un réarrangement, les points de cassure correspondants devraient être présents dans le jeu homme-souris et dans le jeu homme-rat.

Il est donc nécessaire d'analyser et de comparer plus en détail ces 5 jeux de coordonnées de points de cassure.

a. Assignation des points de cassure sur une branche de l'arbre des espèces

Nous avons utilisé une méthode basée sur le principe de parcimonie qui permet d'estimer si plusieurs points de cassure sont issus d'un même évènement de réarrangement et sur quelle branche de l'arbre des 6 espèces étudiées il a eu lieu (voir l'arbre utilisé Figure 6.1). Le résultat de cette méthode est une liste de coordonnées de régions de cassure non chevauchantes sur le génome humain avec une assignation sur une branche de l'arbre, ou bien une étiquette

"ré-utilisé" si l'assignation d'un seul évènement n'a pas été possible.

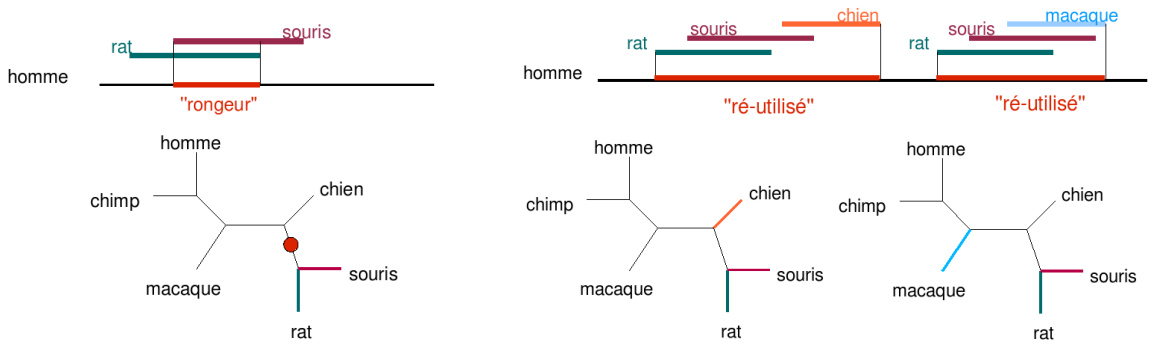
Tout d'abord, les points de cassure qui sont communs à plusieurs comparaisons deux à deux sont identifiés : ce sont des points de cassure dont les régions se chevauchent sur le génome de l'homme. Les autres points de cassure, ceux qui ne chevauchent aucun autre point, sont assignés sur la branche de l'espèce non humaine avec laquelle ils ont été identifiés et leurs coordonnées ne sont pas modifiées.

Pour les points de cassure non "uniques", on construit un graphe G où les noeuds sont les points de cassure et deux points de cassure sont reliés par une arête si leurs régions se chevauchent sur le génome humain. Un groupe d'espèces S est considéré **monophylétique** dans un arbre non raciné si toutes les feuilles du sous-arbre minimal contenant toutes les espèces de S appartiennent à S. Pour chaque composante connexe du graphe G, on teste si :

1. il existe une unique intersection commune à toutes les régions de cassure ;
2. les espèces (non-humaines) impliquées dans les points de cassure de la composante constituent un groupe monophylétique.

Si ces deux conditions sont respectées, alors tous les points de cassure de la composante peuvent être expliqués par un seul évènement de réarrangement. Sinon, si une seule de ces conditions n'est pas respectée, il faut invoquer au moins deux évènements de réarrangements pour expliquer les points de cassure de la composante.

Dans le premier cas, une seule région de cassure est construite à partir de l'ensemble des régions de la composante : c'est leur intersection commune. On assigne ce point de cassure à la branche parent du sous-arbre comprenant toutes les espèces monophylétiques. Dans le deuxième cas, une seule région de cassure est également construite, mais cette fois on considère l'union de l'ensemble des régions de cassure de la composante, et l'étiquette de ce point de cassure est "ré-utilisé" (voir des exemples dans la Figure 6.1).



(a) Exemple avec un unique évènement de réarrangement dans la branche des rongeurs.

(b) Exemples où l'une des deux conditions n'est pas respectée.

FIG. 6.1: Illustration de la méthode d'assignation des points de cassure à une branche de l'arbre, à travers trois exemples de points de cassure se chevauchant sur le génome humain.

Cette méthode a l'avantage d'être simple et intuitive. Cependant la fiabilité des assignations des points de cassure à une branche de l'arbre peut être mise en doute. La raison principale est que les points de cassure ne sont pas détectés de la même manière pour les 5 comparaisons : ce ne sont pas les mêmes ensembles de gènes orthologues qui sont utilisés

pour chaque comparaison. Ainsi, un point de cassure absent d'une comparaison deux à deux peut simplement refléter l'absence d'assignation d'orthologie du gène concerné pour ce couple d'espèces, alors qu'il sera présent dans une autre comparaison deux à deux.

Malgré tout, nous utilisons cette méthode, non pas pour les assignations des réarrangements sur les branches de l'arbre, mais principalement pour identifier les points de cassure issus d'un même évènement de réarrangement afin de ne pas le compter plusieurs fois dans nos analyses.

b. Description du jeu de données

On obtient 622 points de cassure non chevauchant localisés sur le génome de l'homme (sur les 22 autosomes et le chromosome X). Parmi ces 622 points, seulement 40 n'ont pas pu être assignés à une branche de l'arbre, dont la moitié par manque d'intersection commune et l'autre moitié à cause du caractère non monophylétique du groupe d'espèces concernées.

La distribution des tailles des points de cassure est représentée dans la Figure 6.2. Leur taille varie de 2 paires de bases à 2.8 Mb, avec une taille moyenne de 104 Kb et la taille médiane est de 26 Kb. Sur les 12 points de cassure dont la taille est supérieure à 1 Mb, 9 sont classés comme "ré-utilisé" car les régions de cassure ne possèdent aucune intersection commune.

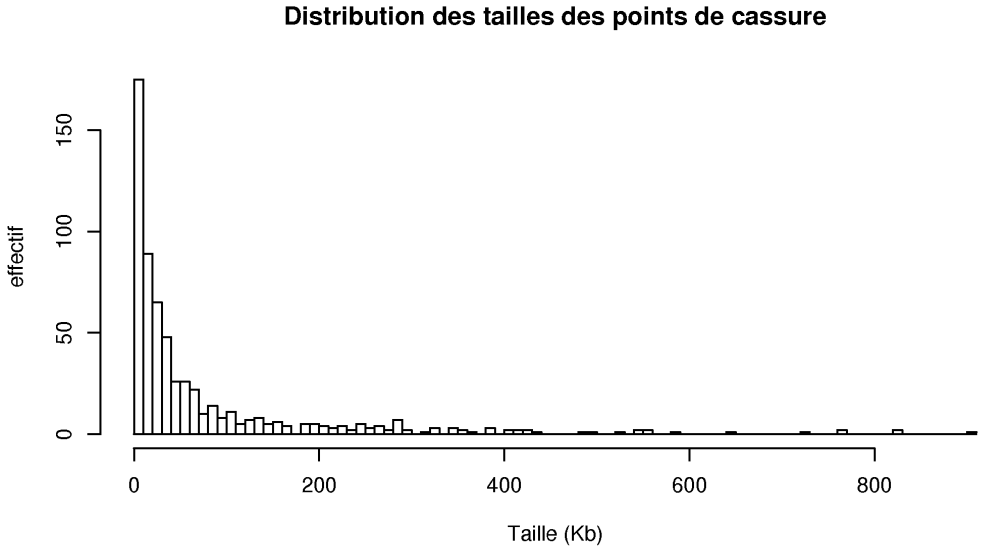


Fig. 6.2: Distribution des tailles du jeu de points de cassure (622 points de cassure). Douze points de cassure ont une taille comprise entre 1 et 3 Mb et n'ont pas été représentés sur cette figure pour des raisons de clarté.

6.1.2 Randomisation des points de cassure

Afin d'évaluer si les caractéristiques des points de cassure sont particulières, nous effectuons des simulations dans lesquelles les régions de cassure sont redistribuées de façon uniforme dans le génome. Ces simulations représentent le modèle de cassures aléatoires pour lequel chaque position du génome à la même probabilité de contenir un point de cassure.

Pour chaque simulation, nous obtenons 622 régions non chevauchantes sur le génome humain, ayant les mêmes tailles que les régions de cassure. Nous les utilisons alors exactement de la même manière que les régions de cassure. Par la suite, nous appellerons ces données les "points simulés".

Enfin, pour estimer la probabilité d'obtenir les caractéristiques observées avec ce modèle (p-value), nous créons 1000 jeux de données de ce type. Cela permet d'obtenir la distribution des caractéristiques testées sous le modèle de cassures aléatoires et d'y comparer la valeur observée pour les points de cassure.

6.2 Réarrangements et distribution des gènes

Les gènes protéiques sont les séquences fonctionnelles du génome que l'on connaît le mieux. Le génome humain possède de 20 à 25,000 gènes protéiques qui couvrent environ 38 % du génome. Ils ne sont pas distribués aléatoirement et leur densité est très hétérogène le long du génome humain. On peut alors se demander si la distribution des gènes et celle des points de cassure dans le génome humain sont liées.

Les annotations des gènes protéiques du génome humain, que nous avons utilisées par la suite, proviennent de la table "known genes" du navigateur de génome de l'UCSC (Hsu *et al.*, 2006). Celle-ci contient environ 39 000 entrées, qui correspondent à 20 181 gènes lorsque les redondances dues aux transcrits alternatifs sont éliminées et 18 669 intervalles géniques sur le génome. Les exons couvrent environ 2 % du génome, alors que les parties géniques en couvrent 38 %.

6.2.1 Le modèle intergénique

Les réarrangements peuvent avoir des impacts importants sur l'expression et la fonction des gènes (voir Chapitre 1, Section 1.3.2). On pense que ces impacts sont le plus souvent négatifs et notamment si le point de cassure (la cassure double brin) se trouve à l'intérieur d'un gène. Ainsi, le **modèle intergénique des réarrangements** suppose qu'un réarrangement, s'il se produit dans un gène, sera contre-sélectionné. On s'attend donc avec ce modèle à observer très peu de points de cassure dans des régions géniques. Ce modèle s'oppose au modèle de cassures aléatoires en proposant certaines régions "interdites" de cassures, des régions **solides** (par opposition à fragile). Il a été proposé notamment par Peng *et al.* (2006) afin d'expliquer le taux élevé de ré-utilisation des points de cassure. Même si la signification de ce taux est controversée, ce modèle n'est pas impossible.

Pour tester ce modèle, nous avons mesuré, pour chaque point de cassure, la proportion de la séquence du point de cassure couverte pas des gènes (considérant comme gène l'unité de transcription, avec les régions 5' et 3' UTR, les exons et les introns). Nous avons procédé de la même manière pour les régions issues du modèle de cassures aléatoires (c'est-à-dire, ces mêmes régions redistribuées uniformément dans le génome).

Globalement, les points de cassures possèdent moins de gènes que les points simulés ; la couverture moyenne des séquences de points de cassure par des gènes est de 0.22, contre 0.38 pour les points simulés (cela correspond à la couverture du génome en gènes) (voir les distributions de la Figure 6.3). Plus de la moitié des points de cassure sont entièrement intergéniques (319 sur 622, soit 51 %), alors que ce n'est le cas que pour 280 (42 %) des points simulés. Si on effectue 1000 tirages indépendants de régions aléatoires, on obtient seulement 3 tirages qui

ont au moins 319 régions entièrement intergéniques. Cela signifie qu'il est très peu probable (probabilité de 0.003) d'obtenir autant de points de cassures intergéniques avec un modèle de cassures aléatoires.

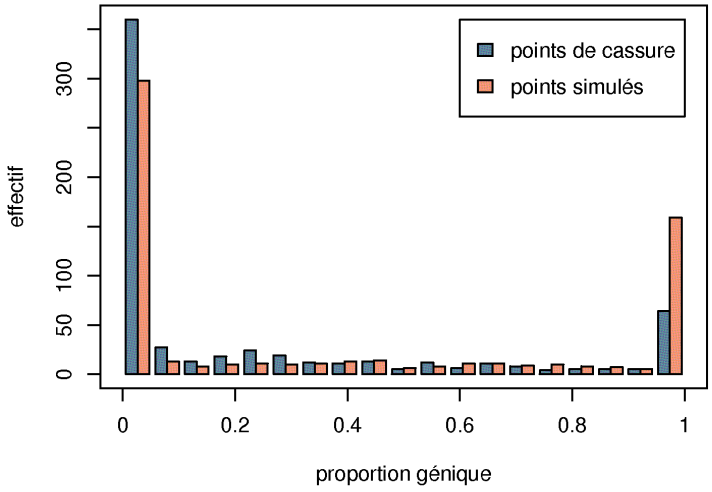


Fig. 6.3: Distribution de la proportion de la séquence couverte par des gènes, pour les 622 points de cassures (en bleu) et pour des régions uniformément redistribuées dans le génome humain (points simulés).

On note cependant que la différence n'est pas très marquée et que le nombre de points de cassure entièrement couverts par des gènes n'est pas négligeable (63 points de cassure sont couverts à plus de 95 % par des régions géniques). La plupart de ces points de cassure ont été assignés sur des branches n'appartenant pas à la lignée humaine, ce qui suggère que la cassure n'a pas eu lieu dans le gène humain. Il est donc possible que ces points de cassure n'aient pas endommagé de gènes. Certains de ces points de cassure peuvent également correspondre à des séquences dupliquées comme nous l'avons vu dans le chapitre précédent (Section 5.3). Il serait intéressant d'analyser plus en détail ces points de cassure et leurs régions orthologues dans les génomes concernés, afin notamment d'étudier les conséquences fonctionnelles du réarrangements sur les gènes concernés.

6.2.2 Taille des inter-gènes

On peut étendre le modèle intergénique au delà des gènes et inclure les promoteurs et les régions régulatrices des gènes. En effet, ces régions jouent un rôle important dans la régulation de l'expression des gènes et peuvent donc également constituer des régions solides pour les réarrangements. C'est ce que proposent Peng *et al.* (2006). Une conséquence de ce modèle est que les petites séquences intergéniques constituent des régions solides pour les réarrangements. On s'attend donc à observer un évitement des petits inter-gènes dans les points de cassure. Nous appelons inter-gène, la région séparant deux gènes successifs sur le génome humain.

Or, ce n'est pas ce qu'on observe dans les points de cassure. Au contraire, si on compare

la distribution de tailles des inter-gènes possédant un point de cassure avec celles des inter-gènes possédant un point simulé, on s'aperçoit que les petits inter-gènes possédant un point de cassure sont sur-représentés (voir Figure 6.4). La taille moyenne des inter-gènes possédant un point de cassure est de 221 Kb contre 908 Kb pour les points simulés (voir Table 6.1).

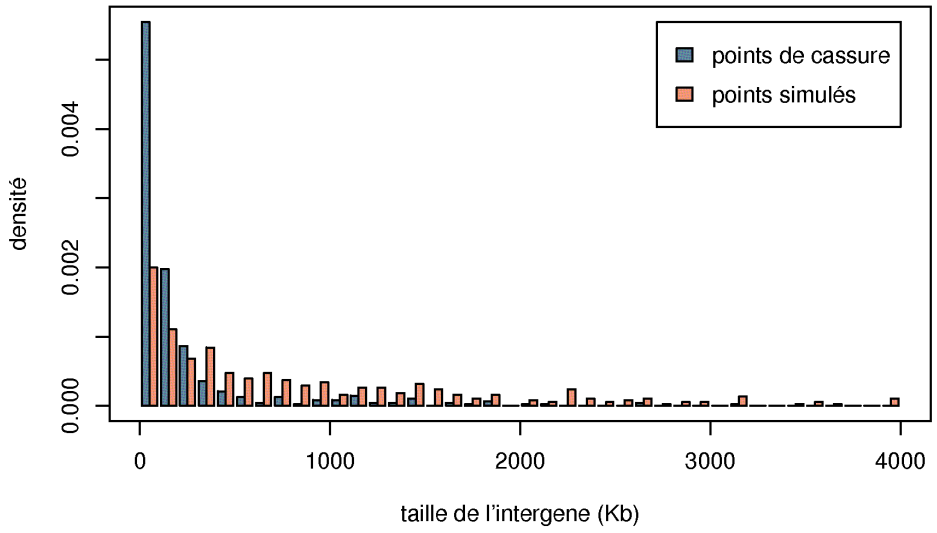


FIG. 6.4: Distributions des tailles des inter-gènes contenant le point milieu d'une région de cassure (en bleu) et de ceux contenant le point milieu d'une région simulée (en orange). La dernière valeur à 4 Mb regroupe 7 points simulés dont la taille de l'inter-gène est supérieure à 4 Mb.

	Taille des inter-gènes (Kb)		% avec inter-gène
	médiane	moyenne	< 100 Kb
Points de cassure	86.5	221.2	56.6 %
Points simulés	480.5	819.1	20.0 %

TAB. 6.1: Caractéristiques des inter-gènes possédant un point de cassure ou un point simulé.

Ainsi, il ne semble pas y avoir d'évitement des régions régulatrices des gènes. Pour nous en assurer, nous avons également calculé la distance de chaque point en amont d'un gène à son site d'initiation de la transcription (TSS). En accord avec les résultats précédents, les points de cassure sont significativement plus proches des TSS que les points simulés (les points de cassure inter-géniques en amont d'un gène sont à 95 Kb de leur TSS contre 354 Kb pour les points simulés, p-value $< 2e-16$ au test de Wilcoxon). Notamment, 77 points de cassures (soit 16 % des points de cassure inter-géniques) sont à moins de 10 Kb d'un TSS, alors que cela n'est le cas que pour 28 points simulés (7 % des points simulés inter-géniques).

6.2.3 Densité en gènes et en régions codantes

Les points de cassure se trouvent dans des inter-gènes plus petits qu'attendu sous un modèle de cassures aléatoires, mais cela ne signifie pas nécessairement que les points de cassure sont dans des régions plus denses en gènes. Nous avons donc mesuré la densité en gènes à plus grande échelle, dans des fenêtres de 500 Kb, autour des points de cassure et autour des points simulés. Nous effectuons trois mesures distinctes :

- la densité en bases géniques, c'est la proportion de la séquence considérée couverte par des gènes (unités de transcription) ;
- la densité en gènes, c'est le nombre de gènes chevauchant la séquence considérée (exprimé en nombre de gènes par Mb) ;
- la densité en bases codantes, c'est la proportion de la séquence considérée couverte par des exons de gènes.

Il est important de distinguer ces trois mesures, car il existe une grande diversité de structure des gènes. Certains gènes sont très longs mais possèdent en fait beaucoup de séquences introniques et donc peu de séquences codantes. De plus, nous verrons dans la section suivante que ces caractéristiques sont corrélées. Ainsi, les petits gènes très compacts (avec peu de séquences introniques et un CDS court) sont souvent regroupés dans des régions denses en gènes.

En accord avec la taille des inter-gènes, les points de cassure se trouvent dans des régions plus denses en gènes et en codant que les régions issues du modèle aléatoire (différences significatives au test de Wilcoxon, voir Tableau 6.2).

	Densité moyenne en bases		Nombre moyen de
	géniques	codantes	gènes par Mb
Points de cassure	0.451	0.033	14.9
Points simulés	0.376	0.020	8.3
p-value	1.6e-07	$< 2.2e\text{-}16$	$< 2.2e\text{-}16$

TAB. 6.2: Comparaison de la densité en gènes et en codant dans des fenêtres de 500 Kb autour des points de cassures et des points simulés. La dernière ligne indique la p-value du test de Wilcoxon de comparaison des distributions pour chaque mesure de la densité en gènes.

Les différences sont plus marquées en ce qui concerne le nombre de gènes et la densité en bases codantes, ce qui indique que les points de cassure sont dans des régions où les gènes sont nombreux et compacts (on peut vérifier que la taille des gènes est significativement plus faible pour les points de cassure que pour les points simulés).

On peut noter que ce résultat renforce le modèle intergénique (évitement des gènes), car plus les régions sont denses en gènes, plus il est probable de "tomber" dans un gène avec un modèle aléatoire. Ainsi, si on redistribuait les points de cassure en préservant la densité en gènes, le nombre de points simulés dans un gène serait encore plus grand que celui obtenu avec le modèle de cassures aléatoires, et donc encore plus différent de celui des points de cassure.

6.2.4 Artefact de la méthode de détection des points de cassure ?

On ne peut identifier, avec notre méthode (Chapitre 3, Section 3.2), qu'un seul point de cassure entre deux gènes successifs, et qui plus est, entre deux gènes orthologues successifs

(pour une comparaison deux à deux). Ainsi, on peut se demander si le fait d'observer plus de points de cassure dans des régions riches en gènes n'est pas simplement dû au fait qu'il est plus facile de les détecter dans ces régions.

a. Simulations en contrôlant le nombre de points par inter-gène

La redistribution des points de cassure uniformément dans le génome ne tient pas compte de cette contrainte. Ainsi, sur les 380 régions simulées dont le point milieu tombe dans un inter-gène, 46 partagent un même inter-gène avec une autre région simulée. Si on enlève les "dupliqués", la distribution de taille des inter-gènes est légèrement déplacée vers de plus petits inter-gènes (la moyenne passe de 819 Kb à 774 Kb), mais reste significativement différente des inter-gènes des points de cassure. Il en est de même pour les différentes mesures de la densité en gènes. Cependant, pour simuler correctement le modèle aléatoire, il faudrait tirer plus de points que de cassures observées et éliminer ceux qui tombent dans un même inter-gène afin d'obtenir à la fin le même nombre de points simulés que de points de cassure.

Si on tire, par exemple, 1000 points uniformément dans le génome mais qu'on ne garde qu'un seul point par inter-gène ou par gène, il en reste 910 (ce qui est beaucoup plus que le nombre de points de cassure détectés). Les caractéristiques des points simulés sont très peu changées (taille des inter-gènes, densité en bases codantes et nombre de gènes par Mb), sauf la densité en bases géniques qui augmente (la moyenne passe de 0.36 à 0.40), mais reste significativement plus faible que celle des points de cassure (p-value de 0.00018 au test de Wilcoxon).

Ainsi, on peut conclure que les points de cassure détectés se trouvent dans des régions plus riches en gènes que des points émis par un modèle de cassures aléatoires, en ne retenant qu'un seul point par inter-gène et par gène.

Mais qu'en est-il de tous les points de cassure, pas seulement ceux que nous avons détecté ? Car il est possible que des points de cassure aient été ratés dans les grands inter-gènes et cela pourrait biaiser les résultats. Deux points de cassure proches sur le génome mais n'étant séparés par aucun gène (avec orthologue) peuvent ne pas être détectés ou bien groupés en un seul point de cassure. Le fait d'utiliser les gènes comme marqueurs pour construire les blocs de synténie peut apparaître inapproprié dans ce cas, mais en fait les méthodes basées sur l'alignement de génomes complets ont les mêmes problèmes, puisque la conservation des séquences est fortement corrélée à la densité de gènes, en particulier pour les espèces distantes comme l'homme et la souris.

b. Etude des points de cassure Homme-Chimpanzé

L'homme et le chimpanzé sont des espèces très proches qui ont très peu divergé, notamment au niveau de la séquence. Le nombre de gènes orthologues est donc plus important et s'il y a un biais de détection des points de cassure en faveur des régions plus denses en gènes, on s'attend à ce qu'il soit moindre dans la comparaison d'espèces proches.

Nous ne disposons que de 36 points de cassure pour lesquels nous avons calculé la taille des inter-gènes et la densité en gènes (avec les trois mesures décrites précédemment). Les points de cassure ont été redistribués uniformément dans le génome pour obtenir un jeu similaire de points simulés. Les résultats sont résumés dans le Tableau 6.3. A part pour la densité en bases géniques, les caractéristiques des points de cassure Homme-Chimpanzé sont similaires aux caractéristiques de l'ensemble des 622 points de cassure, et significativement différentes de celles des points simulés.

	Taille des inter-gènes (Kb)		Densité moyenne en gènes		
	médiane	moyenne	bases géniques	bases codantes	nb / Mb
Points de cassure	74.2	249.2	0.360	0.032	14.6
Points simulés	356.3	682.3	0.388	0.019	7.2
p-value	0.002		0.87	0.003	0.0002

TAB. 6.3: Comparaison des caractéristiques des régions de cassure homme-chimpanzé et des mêmes régions redistribuées uniformément dans le génome (points simulés). Les p-values sont indiquées pour le test de Wilcoxon de comparaison des distributions de chaque caractéristique entre les deux types de régions.

Ces résultats ont été également confirmés avec des points de cassure obtenus par alignement des génomes complets de l'homme et du chimpanzé, méthode *a priori* non dépendante des gènes, étant donné le niveau de similarité des deux génomes (presque 90 % du génome humain est aligné avec celui du chimpanzé, avec environ 1.2 % de divergence). Les données sont composées de 56 points de cassures, issues d'Ensembl, version 49 sur l'assemblage NCBI 36 du génome humain.

c. Simulations de ré-utilisation des points de cassure

Un autre contrôle possible pour vérifier que les données ne sont pas biaisées est de simuler plusieurs points de cassure dans les grandes régions de cassure. Les régions de cassure susceptibles de contenir plusieurs points de cassure sont les plus grandes. Bien sûr, nous n'avons pas détecté tous les points de cassure de réarrangements ; les petits réarrangements (ou micro-réarrangements) ont vraisemblablement été manqués. Par contre, nous aimerions avoir un ensemble non biaisé des macro-réarrangements. Si on suppose qu'on a détecté tous les réarrangements entre les blocs de synténie d'au moins 100 Kb, on peut imaginer que les régions de cassure dont la taille est supérieure à 100 Kb peuvent contenir au moins un bloc de synténie de plus de 100 Kb manqué. Ainsi, ces régions pourraient contenir plusieurs points de cassure au lieu d'un. On effectue alors des simulations dans lesquelles on tire uniformément n points dans les régions de cassure, n étant égal à 1 pour les régions inférieures à 200 Kb et à $\lfloor \frac{L}{100000} \rfloor$ pour les autres (L étant la taille de la région de cassure). On obtient ainsi 946 points de cassure et on tire le même nombre de points uniformément dans le génome pour obtenir le jeu de points simulés.

Les résultats sont similaires aux résultats précédents et restent significatifs (voir Tableau 6.4).

	Taille des inter-gènes (Kb)		Densité moyenne en gènes		
	médiane	moyenne	bases géniques	bases codantes	nb / Mb
Points de cassure	104	270	0.41	0.030	13.8
Points simulés	389	765	0.39	0.020	8.2
p-value	$<$ 2.2e-16		0.014	$<$ 2.2e-16	$<$ 2.2e-16

TAB. 6.4: Comparaison des caractéristiques des points de cassure avec re-utilisation et des points simulés. Les p-values sont indiquées pour le test de Wilcoxon de comparaison des distributions de chaque caractéristique entre les deux types de régions.

d. Conclusion

En conclusion, il apparaît très peu probable que les résultats soient dus à un biais de la méthode de détection des points de cassure. On peut donc rejeter le modèle de cassures aléatoires et le modèle intergénique seul et conclure que les points de cassure se trouvent préférentiellement dans des inter-gènes de régions denses en gènes et en bases codantes. Notons que des résultats similaires relatifs à la densité en gènes des points de cassure avaient été obtenus avec des méthodes de détection différentes, basées sur des alignements de génomes complets, des cartes génétiques, ou des méthodes de cytogénétique chez l'homme (Murphy *et al.*, 2005; Ma *et al.*, 2006; Roberto *et al.*, 2006), et chez le poulet (Gordon *et al.*, 2007). Une étude approfondie des déserts de gènes dans le génome humain (inter-gènes dont la taille est supérieure à 640 Kb, couvrant plus de 25 % du génome) a également montré qu'un grand nombre d'entre eux sont extrêmement conservés à l'échelle des mammifères et des vertébrés (Ovcharenko *et al.*, 2005). Ces auteurs observent notamment que 170 déserts ne possèdent aucun point de cassure lorsqu'ils sont comparés avec le génome du poulet ou celui de la souris. Ces régions contiendraient des éléments régulateurs distants jouant un rôle dans les phases de développement et seraient donc des régions solides "interdites" de cassure.

6.3 Réarrangements et isochores

L'organisation des gènes dans le génome est corrélée avec d'autres caractéristiques génomiques, et notamment avec le contenu en bases G+C et ce qu'on appelle les **isochores**.

Les isochores ont été découverts dans les années 1970s. Par des expériences de centrifugation en gradient de densité de l'ADN, on a mis en évidence une très grande variabilité de composition en bases G+C dans les génomes de mammifères et d'oiseaux (Macaya *et al.*, 1976; Thiery *et al.*, 1976). Bernardi et collaborateurs ont alors décrit ces génomes comme des mosaïques de grandes régions ($>$ 300 Kb) ayant un contenu en bases G+C homogène et présentant des limites nettes avec les régions voisines ; ces régions ont été appelées isochores (Bernardi, 2000).

Ces variations ont ensuite été confirmées par des analyses de séquences à plus grande échelle (Fukagawa *et al.*, 1995). Cependant, avec le séquençage de génomes complets, il s'est avéré que les régions n'étaient pas si homogènes et les limites pas si nettes, remettant en question l'existence des isochores (Lander *et al.*, 2001). Finalement, bien que le débat ne soit pas clos, il semble accepté aujourd'hui que les isochores existent dans les génomes de mammifères et d'oiseaux, même si l'homogénéité compositionnelle est relative et les limites entre isochores floues (Eyre-Walker et Hurst, 2001).

Ce qui est intéressant dans ce découpage des génomes, c'est qu'il est fortement lié à d'autres propriétés biologiques, telles que l'organisation et la taille des gènes (Mouchiroud *et al.*, 1991; Duret *et al.*, 1995), la distribution des éléments transposables (Soriano *et al.*, 1983), le taux de recombinaison (Eyre-Walker, 1993; Fullerton *et al.*, 2001), les bandes chromosomiques (Saccone *et al.*, 1993) et le timing de réplication (Costantini et Bernardi, 2008).

Dans cette partie, nous étudions la distribution des points de cassure par rapport au découpage du génome en isochores, et par rapport aux autres caractéristiques liées à ces derniers.

6.3.1 Contenu en GC et isochores

a. Composition des séquences

Tout d'abord, nous observons que les séquences des points de cassure sont plus riches en bases G et C que les séquences des points simulés. Le taux de GC moyen des points de cassure est de 44 % alors qu'il est de 41 % pour les points simulés. Cette différence est significative au test de Wilcoxon, avec une p-value inférieure à 10^{-12} (voir les distributions dans la Figure 6.5).

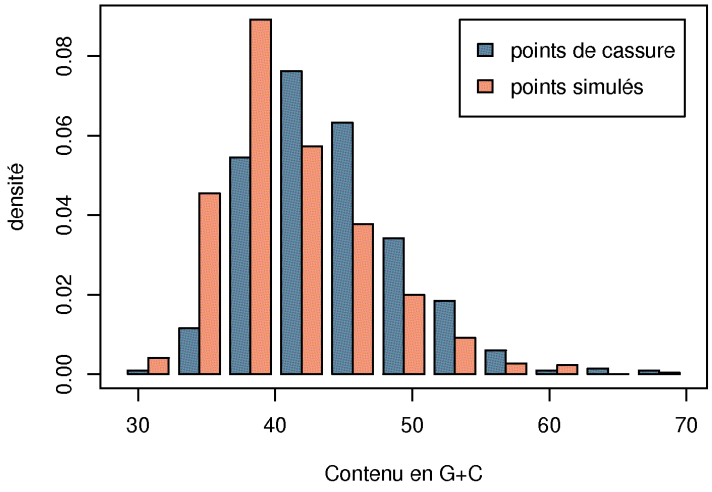


FIG. 6.5: Distributions des compositions en bases G+C des points de cassure (en bleu) et des points simulés (en orange).

b. Répartition dans les classes d'isochores

En ce qui concerne les isochores, nous avons utilisé les données produites par l'équipe de Bernardi (Costantini *et al.*, 2006). Il existe de nombreuses méthodes pour détecter les isochores dans les génomes (certaines sont revues et comparées dans (Schmidt et Frishman, 2008)). Elles sont très diverses et donnent des résultats très différents. Cela est dû au fait qu'il n'existe pas de définition consensus des isochores. Les différences concernent principalement le nombre et la taille des isochores (allant de 1200 isochores jusqu'à plus de 76000, de taille moyenne variant de 40 Kb à 2.4 Mb (Schmidt et Frishman, 2008)). Par contre, environ deux tiers du génome humain sont prédits appartenir à la même classe d'isochore pour les quatres méthodes comparées par Schmidt et Frishman (2008) et les couvertures des séquences par les 5 classes d'isochores sont également très similaires.

Nous avons choisi les données de Costantini *et al.* (2006), bien que la méthode utilisée ne soit pas automatique, car ce sont celles qui ont les caractéristiques les plus similaires de celles des isochores identifiés expérimentalement dans les années 70. Néanmoins, nous avons effectué les analyses avec d'autres données d'isochores, grâce à la base de données IsoBase (Schmidt et Frishman, 2008) et nous avons obtenu des résultats similaires. Les données de Costantini

regroupent 3159 isochores parmi les 5 classes généralement définies : L1, L2, H1, H2 et H3, de contenu en G+C croissant. Leur taille moyenne est de 900 Kb. Pour clarifier les résultats, nous avons regroupé les isochores en 3 classes : L1-L2, H1-H2 et H3 ; leurs caractéristiques sont présentées dans le Tableau 6.5.

	L1-L2	H1-H2	H3
Domaine de GC (%)	33.8−41.0	41.1−53.0	53.1−58.0
Couverture du génome	57.1 %	39.2 %	2.3 %
Taille moyenne (médiane)	2.4 Mb (700 Kb)	1.4 Mb (800 Kb)	685 Kb (30 Kb)

TAB. 6.5: Description des données d'isochores sur le génome humain de Costantini *et al.* (2006).

La répartition des points de cassure dans les trois classes d'isochores est significativement différente de celle des points simulés (p-value du test de Chi2 inférieure à $2e-16$). Les points de cassure se trouvent dans des isochores plus lourds que les points simulés (voir Tableau 6.6). Cette tendance est également significative au test de tendance avec une p-value de 0 (voir Section 6.4.2 a. pour une description de ce test non paramétrique).

	Nombre de points par classe d'isochore			
	L1-L2	H1-H2	H3	indéfini
Points de cassure	191 (31 %)	391 (63 %)	23 (4 %)	13 (2 %)
Points simulés	347 (56 %)	252 (41 %)	10 (2 %)	9 (1 %)

TAB. 6.6: Comparaison des classes d'isochores contenant des points de cassure et des points simulés. L'assignation d'une région à une classe d'isochore est effectuée si la région est couverte à plus de 70 % par des isochores de cette classe, sinon la région est classée "indéfinie".

c. Distance à une frontière d'isochore

Si les points de cassure se trouvent préférentiellement dans certaines classes d'isochores, on peut également se demander comment ils se distribuent à l'intérieur des isochores, et notamment par rapport à leurs frontières. Les isochores sont liés au banding des chromosomes et au timing de réplication, les isochores riches en GC correspondraient à des bandes R qui se répliqueraient précocément (Wolfe *et al.*, 1989; Saccone *et al.*, 1993; Costantini et Bernardi, 2008). Ainsi, certaines frontières d'isochores pourraient correspondre à des transitions entre des régions à réplication précoce et d'autres à réplication tardive (Schmegner *et al.*, 2007). Or, ces régions de transition pourraient constituer des régions fragiles, notamment lors de la réplication lorsque la fourche initiée précocément est en attente de celle initiée tardivement (Rothstein *et al.*, 2000). Plusieurs études ont mis en évidence de telles fragilités. Par exemple, dans une étude de duplications segmentaires associées à des points chauds de réarrangements cancéreux, 11 copies sur 15 ont été trouvées à la frontière d'isochores (Darai-Ramqvist *et al.*, 2008). D'autre part, certains sites fragiles [1] sont trouvés très fréquemment à la frontière de bandes chromosomiques au timing de réplication différent (Achkar *et al.*, 2005).

[1] Régions du génome fréquemment endommagées lorsque la cellule est cultivée dans certaines conditions particulières (voir Section 1.4.3 c.). Il ne faut pas les confondre avec les régions fragiles suggérées par le modèle non aléatoire des points de cassure.

Nous avons donc cherché s'il existe un lien entre la localisation des points de cassure et les frontières d'isochores. Avec les données d'isochores de Costantini *et al.* (2006), et en utilisant les 5 classes d'isochores, nous avons calculé la distance des points de cassure et des points simulés à une frontière d'isochore (2849 frontières). La distance moyenne des points de cassure à une frontière est légèrement plus faible (503 Kb) que celle des points simulés (584 Kb). Cette différence est peu significative au test de Wilcoxon (p-value de 0.0465) et est seulement due au fait que les points de cassure sont préférentiellement dans des petits isochores (les isochores lourds sont plus petits en moyenne que les isochores légers). Si on compare les distances aux frontières dans chaque classe d'isochore, les différences ne sont pas significatives.

Ces résultats ne permettent pas de conclure sur un lien entre les points de cassure et les frontières d'isochores. Cela pourrait provenir de la qualité des données de frontières d'isochores. En effet, les données utilisées ici ont une précision de 100 Kb. De plus, si on utilise des isochores obtenus avec d'autres méthodes, le nombre de frontières est radicalement différent (un ordre de grandeur de différence : 2 900 frontières pour les données de Costantini *et al.* (2006), 25 000 pour les données issues de la méthode consensus de Schmidt et Frishman (2008)) et les résultats, bien que démontrant la même tendance, présentent des significativités différentes. Enfin, l'existence même de frontières nettes entre isochores de différentes classes n'est pas établie à l'heure actuelle.

Toutes les frontières d'isochores ne correspondent pas forcément à des transitions dans le timing de réplication. Ainsi, il serait plus intéressant d'analyser les données de points de cassure avec des données de timing de réplication directement. Il serait nécessaire alors de disposer de données à haute résolution sur l'ensemble du génome humain, et obtenues avec des cellules germinales, car ce sont dans ces cellules que se produisent les réarrangements évolutifs. Nous étudierons dans la Section 6.4 d'autres caractéristiques liées à la réplication : les origines de réplication.

6.3.2 Classes d'éléments transposables

Les éléments transposables ne sont pas distribués aléatoirement dans le génome par rapport aux isochores. Les éléments de type SINEs sont plus riches en GC et sont trouvés préférentiellement dans les isochores lourds. C'est l'inverse pour les éléments de type LINEs.

Nous avons comparé la couverture des séquences en éléments transposables entre les régions de cassure et les régions simulées. Nous obtenons des résultats en accord avec les résultats précédents concernant les isochores (voir Tableau 6.7). Les régions de cassure sont significativement plus riches en éléments de type SINEs et LTRs.

	Couverture des séquences (%) en ETs				
	total	LINEs	SINEs (Alu)	LTRs	DNA
Points de cassure	49.6	17.1	19.2 (17.3)	10.1	2.3
Points simulés	41.8	18.4	12.6 (10.3)	7.9	2.7
p-value		0.3	5e-13 (2e-16)	0	0.007

TAB. 6.7: Couverture des séquences en éléments transposables (ETs). Pour chaque classe d'éléments est donnée la p-value du test de Wilcoxon entre les deux distributions.

6.3.3 Recombinaison

Le taux de recombinaison est également corrélé avec le contenu en GC et avec les isochores. En fait, la recombinaison à travers la conversion génique est supposée être le mécanisme à l'origine des isochores (Eyre-Walker, 1993; Galtier *et al.*, 2001; Galtier, 2003; Marais, 2003). La conversion génique est le remplacement d'un allèle par un autre au moment de la recombinaison : la recombinaison génère des hétéroduplex (entre ADN paternel et maternel) qui peuvent contenir des mésappariements, ces derniers sont ensuite réparés. Or la réparation est biaisée vers les bases GC. Ainsi les régions à fort taux de recombinaison s'enrichissent en GC.

Une autre raison de s'intéresser à la recombinaison vient des mécanismes de réarrangements. L'un de ces mécanismes est la recombinaison homologue non allélique (NAHR), qui comme son nom l'indique, implique la recombinaison homologue. De plus, la recombinaison méiotique est initiée par une cassure double brin, tout comme les réarrangements. On peut donc se demander si les variations des taux de recombinaison sont corrélées avec les données de points de cassure.

La distribution des taux de recombinaison le long du génome humain est très hétérogène, et ce à différentes échelles, à l'échelle des chromosomes, du mégabase mais également du kilobase. C'est à petite échelle qu'on observe les plus grandes hétérogénéités. Des analyses à très grande résolution ont montré qu'il y aurait environ 25 000 points chauds ("hotspots") de recombinaison dans le génome humain (Myers *et al.*, 2005). Ce sont des régions de 1 à 2 Kb qui présentent de très forts taux de recombinaison par rapport aux séquences voisines.

Cependant, les positions et intensités de ces hotspots semblent très variables, même entre individus d'une même espèce (Coop *et al.*, 2008). Ils ne sont pas du tout conservés entre l'homme et le chimpanzé par exemple (Ptak *et al.*, 2005; Winckler *et al.*, 2005). Ainsi, il ne fait pas de sens de corréler ces données à l'échelle du Kb puisque les points de cassure sont le plus souvent des évènements très anciens. Cependant, si les positions des hotspots varient à l'échelle du Kb, la densité de hotspots, comme les taux de recombinaison, dans des fenêtres de 1 à 5 Mb semblent plus conservés (Myers *et al.*, 2005).

Nous avons donc comparé les taux de recombinaison et les densités en hotspots dans des fenêtres de 1 Mb autour des points de cassure et des points simulés. Deux types de données de recombinaison ont été utilisées : les données Hapmap et les données deCODE accessibles depuis le navigateur de génome de l'UCSC (Karolchik *et al.*, 2008). Les données Hapmap sont issues d'études de patrons de variation génétique (SNP) (McVean *et al.*, 2004; Myers *et al.*, 2005) (taux de recombinaison dans des fenêtres de 1 Mb et localisations des hotspots). Les données deCODE sont inférées à partir des patrons de transmission de marqueurs polymorphes dans des grands pedigree (Kong *et al.*, 2002) (taux de recombinaison dans des fenêtres de 1 Mb).

Si le taux de recombinaison moyen des points de cassure est plus élevé que celui des points simulés, la différence n'est pas significative (voir Tableau 6.8). Par contre, on observe la tendance inverse concernant le nombre de hotspots par Mb, il est légèrement supérieur dans les points simulés que dans les points de cassure.

Ces non-résultats sont surprenants au vu de ceux obtenus précédemment. Nous avons observé que les points de cassure se trouvent préférentiellement dans des isochores lourds, et le taux de recombinaison augmente avec le contenu en GC des isochores. Cette association est visible avec les taux de recombinaison utilisés ici, puisqu'ils sont de 1.26, 1.45, 1.69 pour

	Taux de recombinaison moyen		Densité de
	DeCODE	Hapmap	hotspots (/Mb)
Points de cassure	1.32	1.37	7.3
Points simulés	1.25	1.31	7.9
p-value (Wilcoxon)	0.20	0.25	0.003

TAB. 6.8: Description des résultats de la comparaison des taux de recombinaison et de la densité de hotspots entre les points de cassure et les points simulés. Les taux et densités sont calculées dans des fenêtres de 1Mb autour des régions concernées.

les isochores L1-L2, H1-H2 et H3 respectivement (différences significatives deux à deux au test de Wilcoxon). Une des raisons pour laquelle l'association entre les points de cassure et la recombinaison n'est pas visible, alors qu'elle l'est avec les isochores, pourrait être que ces caractéristiques évoluent à des échelles de temps différentes. La recombinaison varie à des échelles de temps très courtes, alors que le contenu en GC, bien qu'influencé par cette dernière, varie beaucoup plus lentement ; le taux de recombinaison est en fait beaucoup plus corrélé au GC futur (à l'équilibre) qu'au GC actuel (Meunier et Duret, 2004). Or, la majorité des points de cassure que nous analysons correspondent à des évènements anciens. Il n'est donc pas surprenant que les corrélations soient plus fortes avec des co-variables évoluant plus lentement.

6.3.4 Conclusion sur les isochores

Les résultats de ces différentes analyses sont en accord avec une sur-représentation des points de cassure dans les isochores lourds. Les points de cassure se trouvent donc dans des régions génomiques à forte densité en gènes et en codant (les gènes y sont courts avec peu de séquences introniques), à forte composition en bases G+C, avec un enrichissement en éléments transposables de types SINEs et LTR. Notons que ces résultats sont en accord avec ceux de la littérature, notamment chez l'homme (Ma *et al.*, 2006), et chez le poulet (Gordon *et al.*, 2007).

Ainsi, les points de cassure ne sont pas distribués aléatoirement par rapport à l'organisation du génome à grande échelle (plusieurs centaines de kilobases). Parmi les différents facteurs étudiés ici (les gènes, les éléments transposables, le contenu en GC), nous ne pouvons établir de liens de cause à effet avec les réarrangements. Nous avons étudié ces différents aspects de manière indépendante, mais il serait intéressant de pouvoir séparer les co-variables, si cela est possible, afin d'identifier celles qui sont le plus liées aux points de cassure. De plus, il reste encore d'autres facteurs probablement corrélés aux isochores que nous n'avons pas étudiés, comme par exemple la structure de la chromatine et les caractéristiques de réplication.

6.4 Réarrangements et origines de réplication

On appelle **origines de réplication** des sites spécifiques sur le génome où est initiée la réplication de l'ADN. Le génome des procaryotes est généralement formé d'un seul chromosome circulaire qui présente une origine de réplication et un terminus à l'opposé de celle-ci. Il est intéressant de noter que la majorité des réarrangements observés chez les procaryotes sont localisés de façon symétrique par rapport à cet axe origine-terminus (lire la revue (Rocha, 2004)). On observe également une organisation des gènes en fonction de leur position

et de leur orientation par rapport à cet axe : de nombreux gènes sont co-orientés avec le sens de progression de la fourche de réplication (Frank et Lobry, 1999). Cette co-orientation procure un avantage sélectif, puisqu'elle permet de limiter les collisions frontales entre la machinerie de réplication et celle de transcription. Ainsi, l'hypothèse avancée pour expliquer le patron particulier de réarrangements autour de l'origine fait appel à la sélection naturelle : les réarrangements cassant l'organisation des gènes par rapport à la réplication seraient contre-sélectionnés.

Si le lien entre réarrangements et réplication est très étudié et bien établi chez les procaryotes, il n'en est rien chez les eucaryotes et encore moins chez les mammifères. Le problème est plus complexe chez ces espèces car il existe un grand nombre d'origines de réplication sur le génome, dont la détection n'est pas triviale. Récemment, une nouvelle méthode *in silico* a permis de proposer environ un millier d'origines de réplication localisées sur le génome humain (Huvet *et al.*, 2007). C'est avec ce jeu de données que nous avons comparé les points de cassure de réarrangements. Nous avons travaillé en collaboration avec Lamia Zaghloul, Benjamin Audit et Alain Arnéodo du laboratoire Joliot-Curie de l'ENS de Lyon, qui produisent et étudient ces données de réplication. Ce travail fait l'objet d'un article soumis (Lemaitre *et al.*, 2008c).

6.4.1 Les domaines de réplication

La méthode de détection des origines de réplication de Huvet *et al.* (2007) est basée sur la composition en bases des séquences d'ADN. Si les deux brins de la séquence d'ADN sont soumis aux mêmes pressions de mutation et de réparation, on s'attend à ce que les compositions en nucléotides C et G sur un même brin soient égales, et de même pour les nucléotides A et T. Cependant, la réplication de l'ADN n'est pas un processus qui traîte de la même façon les deux brins de l'ADN : le brin lagging reste plus longtemps sous forme simple brin que le brin leading. Cela introduit une asymétrie de composition en bases en fonction du brin, leading ou lagging (Frank et Lobry, 1999). Ainsi, le brin leading accumule plus de G que de C et plus de T que de A, et inversement pour le brin lagging. Si le (ou les) origines de réplication sont fixes, c'est-à-dire si elles sont à la même position à chaque cycle de réplication, alors les différences de compositions s'accumulent dans les séquences. Si on calcule ce biais le long de génomes bactériens, on voit clairement qu'il change abruptement de signe au niveau de l'origine de réplication (Lobry, 1996). Cependant, chez les eucaryotes on ne connaît pas *a priori* le nombre d'origines et les changements de biais sont plus confus. De plus, la transcription rend les choses encore moins claires puisqu'elle peut elle aussi introduire des biais de composition en fonction du brin sur lequel sont codés les gènes.

Huvet *et al.* (2007) ont développé une méthode multi-échelle basée sur une transformation en ondelettes du signal de biais. Cela permet de distinguer les biais présents à des échelles différentes, notamment ceux dus à la transcription de ceux dus à la réplication. Plus précisément, la méthode estime des domaines de réplication, appelés *N-domaines* : ce sont des régions dont le biais présente un saut à chaque extrémité et décroit de façon linéaire entre ces deux dernières (le biais a une forme en N). Cette forme est interprétée par le modèle suivant : un N-domaine serait une région contenant exactement deux origines de réplication fixes situées à ses extrémités, et le terminus de réplication, quant à lui, est variable dans le domaine. Cette méthode a été validée sur 10 origines humaines prédites expérimentalement, puis appliquée à tout le génome. Elle a permis d'obtenir 678 N-domaines bordés par 1060 origines de réplication potentielles et couvrant environ 28.3 % du génome. Il faut noter que l'efficacité de la méthode est fortement dépendante du contenu en G+C et de la densité en

gènes. Ainsi, elle n'est pas capable d'identifier des domaines dans des régions trop riches en gènes, car les domaines de réplication seraient plus petits et l'échelle devient trop proche de celle des gènes (biais de transcription).

La taille des N-domaines varie de 300 Kb à 2.8 Mb et le contenu en G+C moyen est de 39.4 % (voir les distributions de la Figure 6.6).

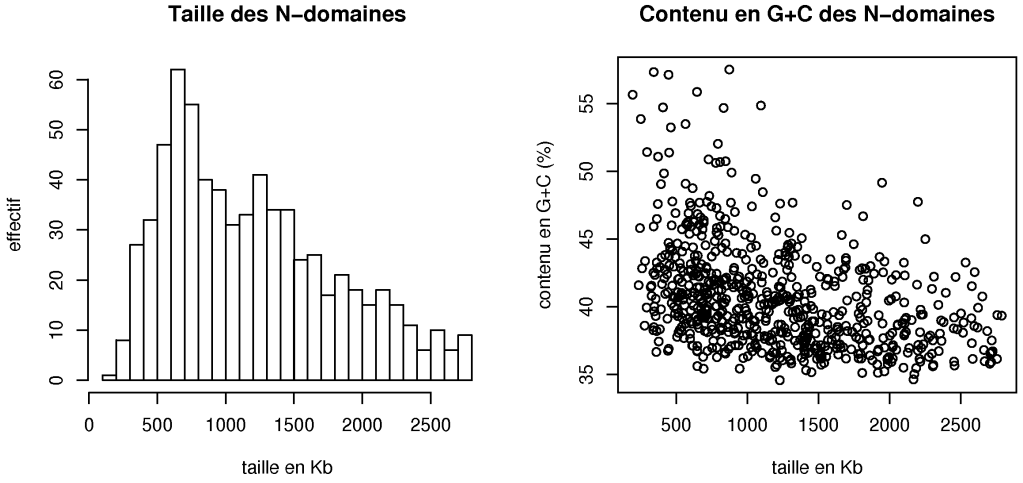


FIG. 6.6: Taille et contenu en G+C des 678 N-domaines.

L'avantage de cette méthode est qu'elle prédit les origines de réplication fixes qui sont actives dans la lignée germinale ; c'est la lignée qui nous intéresse puisque les réarrangements évolutifs se produisent dans celle-ci. De plus, elle a été appliquée sur le génome de la souris également et cela a permis de montrer que les positions des origines de réplication détectées sont relativement bien conservées à l'échelle des mammifères (données non publiées). On peut donc comparer ces données avec celles des réarrangements évolutifs des mammifères.

Il faut cependant rester prudent, car ces données sont des prédictions et n'ont pas été vérifiées expérimentalement. De plus, la méthode prédit un petit nombre d'origines et cet échantillon peut présenter des biais par rapport à l'ensemble des origines de réplication du génome. Néanmoins, nous appellerons les extrémités des N-domaines des origines de réplication.

6.4.2 Les points de cassure aux origines de réplication

Nous avons localisé les régions de cassure par rapport aux N-domaines sur le génome humain. Plus précisément, nous avons sélectionné les régions de cassure dont le point milieu se trouve dans un N-domaine, et pour celles-là, nous avons calculé la distance de ce point à l'extrémité de N-domaine la plus proche (origine de réplication potentielle). On obtient 111 régions de cassure qui "tombent" dans un domaine, soit 18 % du jeu de données. L'histogramme des distances à l'origine est représenté dans la Figure 6.7 ; nous y avons également représenté les effectifs attendus sous l'hypothèse que les points sont répartis uniformément dans les N-domaines (croix rouges). On observe que les régions de cassure ne sont pas réparties uniformément dans les N-domaines, et qu'un grand nombre d'entre elles sont localisées

très proches des origines : plus de la moitié des points (51 %) sont à moins de 200 Kb d'une origine prédite.

On notera également que, même sous l'hypothèse de répartition uniforme, les effectifs diminuent avec la distance à l'origine (croix rouges de la Figure 6.7). Cela est dû au fait que tous les domaines considérés n'ont pas la même taille et que plus on s'éloigne de l'origine, moins il y a de domaines concernés. On observe, cependant, que la tendance est beaucoup plus forte pour les points de cassure que pour les points uniformément répartis. Cela n'est pas dû au fait que les points de cassure tomberaient préférentiellement dans des petits domaines, puisque les N-domaines contenant un point de cassure sont globalement plus grands (taille médiane de 1.3 Mb) que les autres (taille médiane de 1.1 Mb).

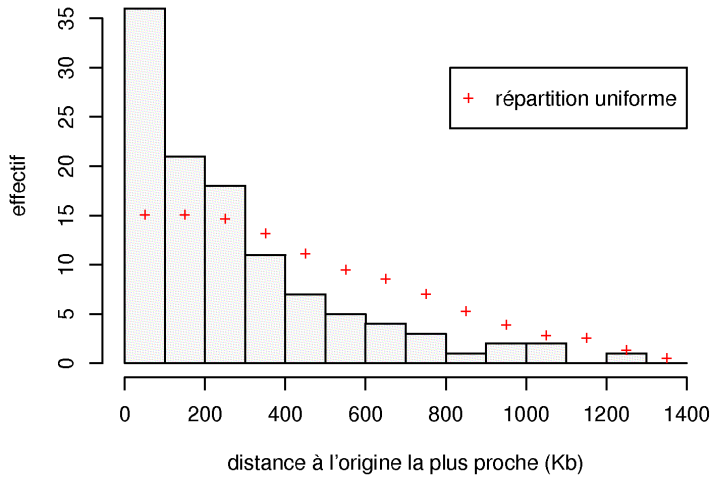


FIG. 6.7: Répartition des points de cassure dans les demi-domaines (111 points milieux des régions de cassure qui tombent dans un N-domaine). Les croix rouges représentent les effectifs attendus sous l'hypothèse que les 111 points de cassure sont répartis uniformément dans les N-domaines.

Pour tester si la tendance observée pour les points de cassure est significative, on utilise un test non paramétrique, appelé le test de tendance. Notons que le test statistique n'est pas nécessaire dans ce cas, il ne fait aucun doute que le profil de distances ne peut être obtenu avec une répartition uniforme des points de cassure. Nous introduisons cependant ce test, car il nous resservira plus tard.

a. Test de tendance

Nous voulons tester si les points de cassure se répartissent dans les domaines indépendamment de leur distance à l'origine (hypothèse nulle H_0) ou bien si le nombre de points de cassure augmente lorsqu'on se rapproche de l'origine (hypothèse alternative H_1). Nous établissons K classes de distances à l'origine et nous voulons donc tester si les effectifs de points de cassure dans chaque classe sont indépendants les uns des autres. Le test de χ^2 permet de

tester cette hypothèse, mais il ne prend pas en compte le fait que les classes sont ordonnées et son hypothèse alternative est seulement la non indépendance des effectifs. Par contre, le test de tendance, prend en compte cette information et son hypothèse alternative est qu'il y a une tendance vers les classes d'ordre inférieur ou supérieur (Chessel, 1978).

Les demi-domaines sont découpés en segments de longueur 50 Kb, définissant ainsi $K = 28$ classes de distances (le plus grand demi-domaine fait 1.39 Mb), numérotées de 1 à K par distance croissante à l'origine ($C_1 = [0..50kb[, C_2 = [50kb..100kb[, ..$). Pour chaque classe C_i, on calcule le nombre de demi-domaines possédant un segment appartenant à cette classe, X_i, et parmi ceux-ci le nombre de ces segments contenant un point de cassure, Y_i. On note N, le nombre total de segments ($N = \sum_{i=1}^{K} X_i$), et M le nombre total de segments contenant un point de cassure ($M = \sum_{i=1}^{K} Y_i$). On définit alors la variable aléatoire $ST = \sum_{i=1}^{K} iY_i$. On sait calculer, sous l'hypothèse d'indépendance H_0, l'espérance et la variance de ST, et on peut approcher la distribution de la valeur centrée réduite de ST ($ST^* = \frac{ST - E(ST)}{\sqrt{V(ST)}}$), sous H_0, par la loi normale de moyenne 0 et de variance 1. Si on pose $m = \frac{1}{N}\sum_{i=1}^{K} iX_i$ et $\sigma^2 = \frac{1}{N}\sum_{i=1}^{K} i^2X_i - m^2$, on a les formules suivantes pour l'espérance et la variance de ST :

$$E(ST) = Mm \quad \text{et} \quad V(ST) = \frac{M(N-M)}{N-1}\sigma^2$$

Pour tester s'il y a une tendance pour les points de cassures à être proches de l'origine, on compare la valeur de ST^* à la loi normale $N(0,1)$; la p-value est la probabilité sous cette loi d'avoir une valeur inférieure à ST^*.

b. Résultats du test de tendance

La valeur ST^* des points de cassure est -4.35, correspondant à une p-value de $6.6e-6$ avec l'approximation normale. Ainsi, on rejette l'hypothèse d'indépendance et on peut dire que plus on se rapproche de l'origine, plus on trouve de points de cassure. Nous avons appliqué ce test aux points simulés (redistribués uniformément dans le génome). La valeur de ST^* obtenue est de -0.04, donnant une p-value de 0.48. Il est rassurant d'accepter l'hypothèse d'indépendance ici puisque les points simulés ont été uniformément tirés dans le génome.

c. En dehors des domaines

On remarque que seulement 111 régions de cassure, soit 18 % des régions, ont leur point milieu dans un N-domaine, alors que ces derniers couvrent environ 28 % du génome. Pour les points simulés, on a un pourcentage plus proche de la couverture du génome par les N-domaines : 148 points simulés, soit 24 % se trouvent dans un domaine.

Le faible pourcentage de points de cassure dans les N-domaines peut s'expliquer par la distribution des ces derniers par rapport aux isochores. Les N-domaines dont nous disposons sont un sous-ensemble non représentatif du génome. Ils se trouvent majoritairement dans des isochores légers : les N-domaines sont couverts à 68 % par des isochores L1-L2 (isochores prédits par Costantini *et al.* (2006)). Ce biais est dû à la méthode de détection. En effet, dans les isochores lourds, la densité en gènes est plus importante, les domaines sont supposés plus petit et la méthodologie n'est pas capable de différencier les asymétries de transcription des asymétries de réplication. Or nous avons vu dans la section précédente que, au contraire, les points de cassure se trouvent préférentiellement dans des isochores lourds.

On peut également expliquer ce faible pourcentage par les origines de réplication. Si, comme on vient de l'observer, les points de cassure se trouvent préférentiellement proches des

origines de réplication, alors on s'attend à ce que le pourcentage de points de cassure dans un N-domaine reflète le pourcentage d'origines de réplication couvertes par les N-domaines. Or, on pense qu'à l'extérieur des N-domaines, et notamment dans les isochores lourds, les domaines de réplication sont plus petits et la densité d'origines est plus importante (Huvet *et al.*, 2007). Ainsi, si les N-domaines couvrent 28 % du génome, on pense qu'ils représentent beaucoup moins de 28 % des origines de réplication humaines.

6.4.3 Liens avec l'organisation des gènes

La densité des origines de réplication humaines semble être liée à d'autres structures génomiques, telles que les isochores et la densité en gènes. Même à l'intérieur des N-domaines, ces caractéristiques montrent un patron particulier. Notamment, la densité en gènes et le contenu en G+C augmentent lorsqu'on se rapproche des extrémités des domaines, c'est-à-dire des origines de réplication prédites.

a. Distribution des gènes dans les N-domaines

Huvet *et al.* (2007) ont étudié l'organisation des gènes le long des N-domaines détectés. Ils ont observé que les gènes ne sont pas distribués uniformément dans les domaines et que cela dépend de leur orientation. Ils observent, proche des origines, une organisation similaire à celle des procaryotes, où les gènes ont tendance à être orientés dans la direction de progression de la fourche de réplication (voir Figure 6.8). Ils observent également que la densité en gènes décroît lorsqu'on s'éloigne de l'origine.

Si la densité en gènes est plus importante à proximité des origines, on observe la même tendance en ce qui concerne la taille des inter-gènes : les inter-gènes sont plus petits proche des origines prédites (voir la Figure 6.9 graphique de gauche). En conséquence, on observe que le contenu en bases G+C décroît avec la distance à l'origine (Figure 6.9 graphique de droite). De même, si les centres des N-domaines sont majoritairement dans des isochores légers (65 % sont dans des isochores L1-L2), la tendance est beaucoup plus faible pour les extrémités (53 % sont dans des isochores L1-L2 et 45 % dans H1-H2).

Enfin, la Figure 6.10 représente conjointement les points de cassure et les gènes dans les 71 demi-domaines contenant un point de cassure de moins de 50 Kb. On observe alors que les points de cassure sont souvent inter-géniques et dans des petits inter-gènes proches de l'origine.

b. Test de tendance avec les inter-gènes

Nous avons montré dans la Section 6.2 que les points de cassure se trouvent préférentiellement dans les régions riches en gènes et dans les petits inter-gènes. Ainsi, on peut se demander si le profil de points de cassure le long des N-domaines ne reflète pas simplement leur corrélation avec la densité en gènes. Ou bien inversement, la préférence pour les régions riches en gènes pourrait être seulement une conséquence de la co-localisation des points de cassure avec les origines de réplication. Afin d'essayer de séparer ces deux co-variables, l'organisation des gènes et les origines de réplication, nous avons effectué des simulations dans lesquelles nous contrôlons l'un des deux facteurs.

Nous avons tiré des points aléatoirement dans le génome en respectant la distribution des longueurs des inter-gènes des points de cassure ; nous appelons cette simulation "inter-gènes". Ayant contrôlé le facteur inter-gène, on peut alors s'intéresser au facteur origine.

FIG. 6.8: Image tirée de (Huvet *et al.*, 2007) représentant l'organisation des gènes dans les 678 N-domaines. En abscisse, on considère les N-domaines alignés sur leurs extrémités (origines potentielles) (en 0), d est la distance à l'extrémité la plus proche. Le code couleur rouge/bleu distingue les gènes en fonction du brin sur lequel ils sont codés (rouge brin +, bleu brin −). A) les flèches R+ indiquent la direction de progression la plus fréquente des fourches de réplication. B) La densité en gènes (nombre de gènes par fenêtre de 50 Kb et par domaine) ; C) La taille moyenne des gènes en Kb ; D) le pourcentage de bases transcrites dans les deux orientations (rouge et bleu) et en noir le pourcentage de bases non transcrites.

Nous avons appliqué le test de tendance pour les simulations "inter-gènes". Nous avons effectué 1000 simulations et obtenu l'histogramme de la statistique ST^* de la Figure 6.11. Pour ces simulations, nous avons dû restreindre le jeu de points de cassure à ceux dont le point milieu est intergénique, la valeur de ST^* pour ces points est de -3.31. On obtient 312 simulations pour lesquelles la valeur de ST^* est inférieure à -3.31. On conclut donc que le profil des points de cassure intergéniques dans les N-domaines peut être obtenu avec un modèle où seule la distance inter-gène est contrôlée (p-value 0.312).

On peut tout de même observer une différence significative de la proportion de points tombant dans un N-domaine entre les points de cassure et les simulations. Pour les simulations, cette proportion est toujours proche de la couverture du génome par les domaines (23.2 % en moyenne), alors que pour les points intergéniques elle est significativement plus faible (20.0 %)

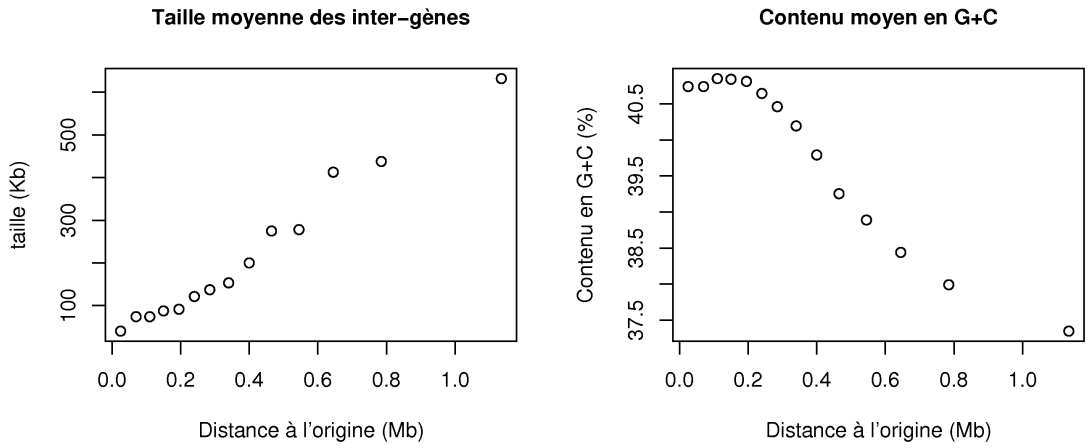


FIG. 6.9: Taille moyenne des inter-gènes et contenu moyen en G+C le long des demi-domaines.

(seulement 45 simulations ont moins de 20 % de points tombant dans un domaine, soit une p-value de 0.045). Cela signifie que si la taille des inter-gènes peut expliquer la tendance des points de cassure vers les extrémités des N-domaines, elle ne permet pas d'expliquer la sur-représentation des points de cassure en dehors des N-domaines.

Enfin, le fait que le test de tendance ne soit pas significatif ne signifie pas nécessairement que seule la taille des inter-gènes est responsable du profil observé dans les N-domaines. Cela peut également indiquer qu'on n'est pas capable de séparer ces deux co-variables.

6.5 Discussion et perspectives

Tous les résultats présentés ici semblent indiquer que les points de cassure se trouvent préférentiellement dans des régions riches en gènes (même si les parties codantes sont évitées), riches en GC et proches des origines de réplication prédites. Ces trois caractéristiques sont en fait fortement corrélées entre elles et nous n'avons pas pu déterminer si l'une d'entre elles joue un rôle prépondérant dans la distribution des points de cassure. Ce résultat s'oppose au modèle de cassures aléatoires. Même en ajoutant au processus neutre de cassures aléatoires un processus sélectif, on ne sait pas expliquer pourquoi les points de cassure proches des gènes seraient plus sélectionnés que les autres.

Ce qui semble caractériser ces régions est une forte activité au niveau de l'ADN, qu'elle soit transcriptionnelle ou réplicative. Ainsi, une hypothèse neutre permettant d'expliquer ces trois associations serait l'état de la chromatine dans lequel l'ADN serait accessible (chromatine dite "ouverte") dans ces régions à forte activité (Gilbert *et al.*, 2004). L'ADN serait alors plus sujet aux cassures et aux réarrangements. Ces zones correspondraient donc à des régions plus fragiles du génome.

Ces régions à forte activité transcriptionnelle sont généralement moins méthylées que les autres. L'étude des dinucléotides CpG peut donner des indications quant au niveau de méthylation de l'ADN dans les cellules germinales. Lorsque les nucléotides C sont méthylés, il sont hypermutables en bases T [2]. Or, les bases C sont très souvent méthylées lorsqu'elles

[2]En fait, les bases C peuvent être dé-aminées, donnant une base U. Cette base est alors reconnue comme

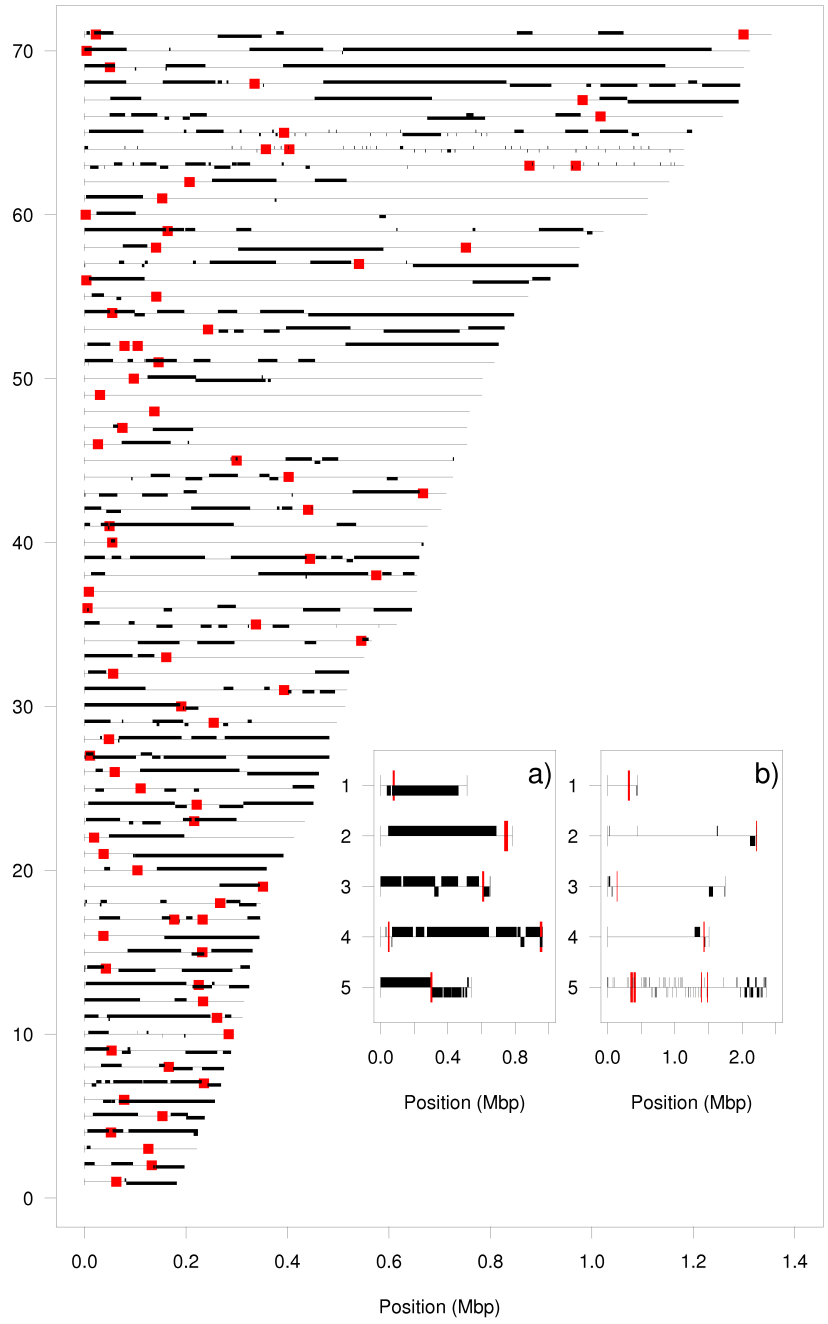


FIG. 6.10: Représentation des demi-domaines possédant une région de cassure de moins de 50 Kb (en rouge). Les origines sont alignées à gauche et les demi-domaines sont ordonnés en fonction de leur taille le long de l'axe vertical. L'axe des abscisses indique la distance à l'origine. Les gènes sont représentés sur les demi-domaines par des segments épais noirs (au dessus de la ligne si l'orientation est positive, en dessous si négative). Dans les cadres a) et b), les domaines sont représentés en entier avec une origine à chaque extrémité ; a) représente les 5 domaines les plus riches en gènes, b) les 5 domaines les plus pauvres en gènes.

une erreur et remplacée par un C par les systèmes de réparation. Il n'y a donc pas de substitution. Le problème, c'est que lorsque la base C est méthylée, la dé-amination donne un T, conduisant à un mésappariement T-G qui peut être réparé en C-G ou T-A (Bird, 1980)

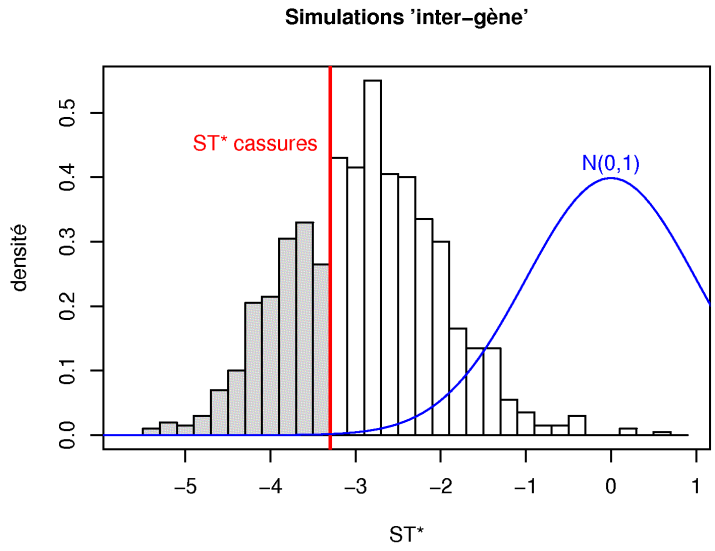


FIG. 6.11: Histogramme des valeurs de la statistique du test de tendance ST^* pour les 1000 simulations 'inter-gène'. La courbe bleue est la densité de la loi normale de moyenne 0 et de variance 1. La droite rouge verticale positionne la valeur de ST^* obtenue pour les 95 points de cassure intergéniques tombant dans les N-domaines.

sont suivies d'un G. Cela implique une sous-représentation des dinucléotides CpG dans les séquences d'ADN, mesurable par le rapport, appelé $CpG_{o/e}$, du nombre de dinucléotides CpG observé sur le nombre de CpG attendu sous un modèle d'indépendance (nombre de C $\times$ nombre de G divisé par la longueur de la séquence). Si on compare ce rapport entre les points de cassure et les points simulés, on obtient que les premiers ont un rapport $CpG_{o/e}$ moyen significativement plus élevé (0.13) que les seconds (0.10) (p-value de 3e-13 au test de Wilcoxon), indiquant que les points de cassure seraient moins méthylés que les points simulés. On peut également déterminer le contenu de ces séquences en îlots CpG. Ce sont des séquences d'au moins 200 pb très riches en CpG et qui échappent vraisemblablement à la méthylation (nous les avons calculées avec le programme CpGproD (Ponger et Mouchiroud, 2002)). De même, les régions de cassure contiennent significativement plus d'îlots CpG (7.4 % des séquences en moyenne) que les régions simulées (4.8 %) (p-value de 5e-11 au test de Wilcoxon). Cependant, si ces deux mesures sont liées au niveau de méthylation de l'ADN, elles le sont également avec la composition en bases G+C, les isochores et la densité en gènes (Duret et Galtier, 2000; Han *et al.*, 2008). Ces résultats ne contredisent donc pas l'hypothèse avancée, mais ne permettent pas de l'affirmer.

Pour vérifier cette hypothèse, il faudrait utiliser directement des données d'expression des gènes, de méthylation de l'ADN et de compaction de la chromatine. Il est important de les mesurer dans les cellules de la lignée germinale, car ces caractéristiques varient d'un tissu à un autre. On pourrait également s'intéresser à des données de timing de réplication, car les origines utilisées ici sont supposées être répliquées précocément.

Ainsi, les résultats obtenus nous permettent seulement de proposer un modèle de réarrangement mettant en jeu l'état de la chromatine. Les régions à forte activité transcriptionnelle dont la chromatine est dans un état ouvert, non condensé, seraient plus exposées que les autres

aux cassures double brin et aux réarrangements. Bien sûr, ce n'est peut-être pas le seul facteur influençant la susceptibilité de l'ADN aux cassures et on peut imaginer que, parmi ces régions, des zones plus localisées sont encore plus fragiles et concentrent plus de points de cassure. Par exemple, certaines séquences répétées ou motifs particuliers pourraient expliquer les points chauds de réarrangements sur des zones plus localisées que les domaines de chromatine. Enfin, ce modèle est neutre et ne fait pas intervenir la sélection. Or, la sélection naturelle intervient probablement dans la distribution des points de cassure, notamment lorsque les cassures se produisent dans des régions fonctionnelles (gènes et déserts de gènes fonctionnels).

Conclusion et perspectives

Dans cette thèse, nous nous sommes intéressés à des régions particulières des génomes, les points de cassure. Nous avons privilégié une approche systématique à l'échelle du génome de l'homme, comparé avec les autres génomes de mammifères entièrement séquencés.

Nous avons, dans un premier temps, développé une méthode qui permet de localiser ces régions sur un génome par comparaison avec le génome d'une autre espèce. L'objectif était d'identifier des points de cassure à la fois "sûrs" et précis. Nous avons pu atteindre ce but en procédant en deux étapes. Dans la première, des blocs de synténie sont construits entre deux génomes ; ils définissent des points de cassure qui seront "affinés" dans la deuxième étape.

La fiabilité des blocs de synténie est assurée par l'utilisation de gènes comme marqueurs orthologues et par le développement d'un algorithme de comparaison de l'ordre de ces marqueurs permettant de tolérer des erreurs isolées. Un avantage de cet algorithme par rapport aux autres méthodes existantes est qu'il produit des blocs de synténie sans conflit interne, ni chevauchements.

La précision des points de cassure est recherchée dans la deuxième étape, celle de l'affinement. Alors que les méthodes existantes s'arrêtent une fois les blocs de synténie identifiés, nous nous intéressons plus en détail aux points de cassure et à leurs séquences orthologues. Chaque séquence de point de cassure est alignée avec ses séquences orthologues définies par les blocs de synténie. Un algorithme de segmentation permet ensuite d'identifier les nouvelles limites du point de cassure sur la base des alignements.

Cette méthode permet d'améliorer considérablement la précision des points de cassure par rapport aux autres méthodes existantes. Cela permet d'analyser plus efficacement les réarrangements évolutifs. Par exemple, cette méthode pourrait servir dans l'analyse des co-localisations de points de cassure (des réarrangements évolutifs ou autres) et ainsi on pourrait ré-évaluer les taux de ré-utilisation des points de cassure dans des lignées différentes et mieux définir la taille de ces régions ré-utilisées. Dans ce travail, cela nous a permis de caractériser plus finement les séquences de ces régions génomiques et d'analyser leur distribution par rapport à des éléments génomiques de résolution comparable.

La première caractéristique que nous avons mise en évidence est la perte de similarité au niveau des points de cassure. Cela avait déjà été proposé dans la littérature, mais il n'était alors pas possible d'évaluer dans quelle mesure cela reflétait des artefacts méthodologiques de délimitation des blocs de synténie. Ici, nous avons pu clairement délimiter ce qui appartient aux blocs de synténie et ce qui fait partie du point de cassure. Nous avons donc pu confirmer que les points de cassure sont en général des régions assez grandes présentant peu, voire aucune, similarité avec le génome comparé. Mais nous avons également montré que cette faible similarité s'étend au delà des points de cassure, dans les régions adjacentes à l'intérieur des blocs de synténie. Une perspective immédiate de ce travail est de définir plus précisément

la zone autour des points de cassure présentant cette caractéristique. L'utilisation de méthodes de segmentation des données de hits d'alignement similaires à celle de l'étape d'affinement pourrait être envisagée.

Il reste également à établir la signification biologique de ces régions et peut-être à revoir la définition même des points de cassure. On pourrait tout d'abord rechercher si la perte de similarité est ancestrale au réarrangement ou bien une conséquence de celui-ci. Plusieurs comparaisons deux à deux permettraient de définir l'origine évolutive des évènements de réarrangements et on pourrait alors comparer la divergence de ces séquences entre génomes réarrangés et non réarrangés. La recherche de micro-réarrangements dans ces zones est également une piste à explorer, puisque l'existence de régions fragiles concentrant de nombreux réarrangements est le modèle accepté à l'heure actuelle.

Enfin, nous avons également mis en évidence la présence de nombreux éléments transposables et duplications segmentaires dans les points de cassure et leurs séquences adjacentes. Dans une étude plus approfondie des inversions du chromosome Y, nous avons démontré le lien de cause à effet entre des duplications de séquences et les réarrangements en proposant deux mécanismes possibles pour ces inversions : la recombinaison ectopique entre copies de séquences dupliquées (ou NAHR), ou bien la formation des séquences dupliquées coïncidant avec l'évènement d'inversion à cause de cassures double brin à bouts collants. Cependant, ce ne sont pas les seuls liens qui existent entre duplications et réarrangements. La co-localisation de ces divers éléments peut également simplement souligner le caractère fragile de ces régions. Ainsi, de manière plus systématique, l'identification de l'origine évolutive de ces différents évènements de réarrangements et d'insertions d'éléments transposables ou de duplications permettra peut-être de trancher entre ces deux hypothèses, et également d'évaluer la taille de ces régions et l'étendue de ces fragilités.

La recherche d'autres éléments responsables de fragilité de l'ADN a été entamée dans ce travail. Elle pourrait être poursuivie, soit en ciblant les éléments et motifs déjà connus, soit, au contraire, en recherchant de manière exploratoire des mots ou motifs de séquence sur- ou sous-représentés dans ces régions. Enfin, il serait intéressant d'effectuer ces analyses en distinguant les différents types de réarrangements et de régions fragiles, afin de rechercher des caractéristiques spécifiques.

Nous avons également analysé la répartition des points de cassure dans le génome humain par rapport à plusieurs niveaux d'organisation de celui-ci. Nous avons tout d'abord rejeté le modèle uniforme de distribution des points de cassure, en montrant notamment que certaines régions comme les gènes sont évitées. Ce résultat était attendu et révèle le rôle de la sélection naturelle dans la répartition observée des points de cassure évolutifs. Cependant, de ce point de vue, un résultat plus surprenant est la sur-représentation des points de cassure dans les régions riches en gènes. Nous avons en effet montré une relation forte entre les réarrangements et l'organisation du génome en isochores. Les isochores à fort GC et à forte densité en gènes seraient ainsi enrichis en points de cassure. Nous observons également un enrichissement à proximité d'origines de réplication prédites *in silico*, qui présentent elles aussi les mêmes caractéristiques de forte densité en gènes. Une hypothèse possible pour expliquer toutes ces relations est une hypothèse neutraliste faisant intervenir la transcription et l'ouverture de la chromatine. Les régions à forte activité transcriptionnelle et à chromatine ouverte seraient plus accessibles à des cassures de l'ADN et donc à des réarrangements. Bien sûr, la perspective à court terme de ce travail est de tester cette hypothèse avec des données d'expression, de structure de la chromatine, de méthylation, etc.

Une autre approche pour étudier le lien entre réarrangements et organisation génomique, que nous n'avons pas abordée ici, est d'étudier l'impact de ces remaniements sur l'organisation fonctionnelle du génome. On pourrait rechercher si les réarrangements observés tendent à préserver la structure en isochores et en réplicons, comme c'est le cas chez les bactéries pour les réplicons. En effet, la co-orientation des gènes avec le sens de progression de la fourche réplicative est également observée chez l'homme. Elle pourrait refléter un avantage sélectif et les réarrangements "cassant" cette organisation devraient être contre-sélectionnés. On pourrait également s'intéresser aux cas des rongeurs dont la structure en isochore a beaucoup évolué (notamment avec des variations de GC moins marquées) et qui ont subit un grand nombre de réarrangements. On peut se demander si l'érosion des isochores est due à ces remaniements.

Enfin, une perspective à plus long terme consisterait à analyser les points de cassure au regard de l'organisation du génome en 3D dans le noyau. Nous nous sommes intéressés ici exclusivement à la structuration du génome le long des chromosomes de façon linéaire. Or, celui-ci est également organisé en espace dans le noyau. Les chromosomes sont répartis dans des territoires chromosomiques individuels, mais il existe des chevauchements entre territoires de chromosomes différents. Ces chevauchements ne seraient pas aléatoires et seraient dus à des interactions spécifiques entre certaines régions génomiques. Des études ont déjà rapporté que la fréquence de translocations entre deux chromosomes est corrélée avec la fréquence de leurs interactions dans le noyau (Branco et Pombo, 2006b). Des données de localisation 3D des chromosomes dans le noyau et d'interactions à longue distance pourraient permettre non seulement de proposer des régions fragiles mais également des partenaires privilégiés pour former des réarrangements.

Références bibliographiques

Abeysinghe, S. S., Chuzhanova, N., Krawczak, M., Ball, E. V. et Cooper, D. N. (2003). Translocation and gross deletion breakpoints in human inherited disease and cancer I : Nucleotide composition and recombination-associated motifs. *Hum Mutat*, 22(3):229–244.

Achaz, G., Boyer, F., Rocha, E. P. C., Viari, A. et Coissac, E. (2007). Repseek, a tool to retrieve approximate repeats from large DNA sequences. *Bioinformatics*, 23(1):119–121.

Achkar, E. E., Gerbault-Seureau, M., Muleris, M., Dutrillaux, B. et Debatisse, M. (2005). Premature condensation induces breaks at the interface of early and late replicating chromosome bands bearing common fragile sites. *Proc Natl Acad Sci U S A*, 102(50):18069–18074.

Aguilera, A. et Gómez-González, B. (2008). Genome instability : a mechanistic view of its causes and consequences. *Nat Rev Genet*, 9:204–217.

Alekseyev, M. et Pevzner, P. (2007). Are there rearrangement hotspots in the human genome ? *PLoS Comput Biol*, 3(11):e209.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. et Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol*, 215(3):403–410.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. et Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST : a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–3402.

Alves, C. E. R., do Lago, A. P. et Vellozo, A. F. (2005). Alignment with non-overlapping inversions in $O(n^3 \log n)$-time. *Electronic Notes in Discrete Mathematics*, 19:365–371.

Andersson, L., Archibald, A., Ashburner, M., Audun, S., Barendse, W., Bitgood, J., Bottema, C., Broad, T., Brown, S., Burt, D. *et al.* (1996). Comparative genome organization of vertebrates. The First International Workshop on Comparative Genome Organization. *Mamm Genome*, 7(10):717–734.

Arlt, M. F., Durkin, S. G., Ragland, R. L. et Glover, T. W. (2006). Common fragile sites as targets for chromosome rearrangements. *DNA Repair (Amst)*, 5(9-10):1126–1135.

Armengol, L., Marquès-Bonet, T., Cheung, J., Khaja, R., González, J. R., Scherer, S. W., Navarro, A. et Estivill, X. (2005). Murine segmental duplications are hot spots for chromosome and gene evolution. *Genomics*, 86(6):692–700.

Armengol, L., Pujana, M. A., Cheung, J., Scherer, S. W. et Estivill, X. (2003). Enrichment of segmental duplications in regions of breaks of synteny between the human and mouse genomes suggest their involvement in evolutionary rearrangements. *Hum Mol Genet*, 12(17):2201–2208.

Aten, J. A., Stap, J., Krawczyk, P. M., van Oven, C. H., Hoebe, R. A., Essers, J. et Kanaar, R. (2004). Dynamics of DNA double-strand breaks revealed by clustering of damaged chromosome domains. *Science*, 303(5654):92–95.

Auger, I. et Lawrence, C. (1989). Algorithms for the optimal identification of segments neighborhoods. *Bull. Math. Biol.*, 51:39–54.

Babcock, M., Yatsenko, S., Stankiewicz, P., Lupski, J. R. et Morrow, B. E. (2007). AT-rich repeats associated with chromosome 22q11.2 rearrangement disorders shape human genome architecture on Yq12. *Genome Res*, 17:451–460.

Bacolla, A., Jaworski, A., Larson, J. E., Jakupciak, J. P., Chuzhanova, N., Abeysinghe, S. S., O'Connell, C. D., Cooper, D. N. et Wells, R. D. (2004). Breakpoints of gross deletions coincide with non-B DNA conformations. *Proc Natl Acad Sci U S A*, 101(39):14162–14167.

Bailey, J. A., Baertsch, R., Kent, W. J., Haussler, D. et Eichler, E. E. (2004). Hotspots of mammalian chromosomal evolution. *Genome Biol*, 5(4):R23.

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

Bailey, J. A., Yavor, A. M., Massa, H. F., Trask, B. J. et Eichler, E. E. (2001). Segmental duplications : organization and impact within the current human genome project assembly. *Genome Res*, 11(6):1005–1017.

Bashir, A., Volik, S., Collins, C., Bafna, V. et Raphael, B. J. (2008). Evaluation of paired-end sequencing strategies for detection of genome rearrangements in cancer. *PLoS Comput Biol*, 4(4):e1000051.

Bellman, R. et Dreyfus, S. (1962). *Applied dynamic programming*. Princeton University Press.

Benson, G. (1999). Tandem repeats finder : a program to analyze DNA sequences. *Nucleic Acids Res*, 27(2):573–580.

Bergeron, A., Corteel, S. et Raffinot, M. (2002). The algorithmic of gene teams. *In Workshop on Algorithms in Bioinformatics (WABI)*, numéro 2452, pages 464–476. Springer-Verlag, Berlin.

Bernardi, G. (2000). Isochores and the evolutionary genomics of vertebrates. *Gene*, 241(1): 3–17.

Bird, A. P. (1980). DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res*, 8(7):1499–1504.

Bourque, G., Pevzner, P. A. et Tesler, G. (2004). Reconstructing the genomic architecture of ancestral mammals : lessons from human, mouse, and rat genomes. *Genome Res*, 14(4):507–516.

Bourque, G., Zdobnov, E. M., Bork, P., Pevzner, P. A. et Tesler, G. (2005). Comparative architectures of mammalian and chicken genomes reveal highly variable rates of genomic rearrangements across different lineages. *Genome Res*, 15(1):98–110.

Branco, M. R. et Pombo, A. (2006a). Chromosome organization : new facts, new models. *Trends Cell Biol*, 17(3):127–134.

Branco, M. R. et Pombo, A. (2006b). Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. *PLoS Biol*, 4(5): e138.

Bray, N., Dubchak, I. et Pachter, L. (2003). AVID : A global alignment program. *Genome Res*, 13(1):97–102.

Britton-Davidian, J., Catalan, J., da Graça Ramalhinho, M., Ganem, G., Auffray, J. C., Capela, R., Biscoito, M., Searle, J. B. et da Luz Mathias, M. (2000). Rapid chromosomal evolution in island mice. *Nature*, 403(6766):158.

Brudno, M., Chapman, M., Göttgens, B., Batzoglou, S. et Morgenstern, B. (2003a). Fast and sensitive multiple alignment of large genomic sequences. *BMC Bioinformatics*, 4:66.

Brudno, M., Malde, S., Poliakov, A., Do, C. B., Couronne, O., Dubchak, I. et Batzoglou, S. (2003b). Glocal alignment : finding rearrangements during alignment. *Bioinformatics*, 19 Suppl 1:i54–i62.

Brudno, M., Poliakov, A., Salamov, A., Cooper, G. M., Sidow, A., Rubin, E. M., Solovyev, V., Batzoglou, S. et Dubchak, I. (2004). Automated whole-genome multiple alignment of rat, mouse, and human. *Genome Res*, 14(4):685–692.

Burgetz, I., Shariff, S., Pang, A. et Tillier, E. (2006). Positional homology in bacterial genomes. *Evol Bioinform Online*, 2:77–90.

Burt, D. W., Bruley, C., Dunn, I. C., Jones, C. T., Ramage, A., Law, A. S., Morrice, D. R., Paton, I. R., Smith, J., Windsor, D. *et al.* (1999). The dynamics of chromosome evolution in birds and mammals. *Nature*, 402(6760):411–413.

Calabrese, P. P., Chakravarty, S. et Vision, T. J. (2003). Fast identification and statistical evaluation of segmental homologies in comparative maps. *Bioinformatics*, 19 Suppl 1:i74–i80.

Cannon, S. B., Kozik, A., Chan, B., Michelmore, R. et Young, N. D. (2003). DiagHunter and GenoPix2D : programs for genomic comparisons, large-scale homology discovery and visualization. *Genome Biol*, 4(10):R68.

Carbone, L., Vessere, G. M., ten Hallers, B. F. H., Zhu, B., Osoegawa, K., Mootnick, A., Kofler, A., Wienberg, J., Rogers, J., Humphray, S. *et al.* (2006). A high-resolution map of synteny disruptions in gibbon and human genomes. *PLoS Genet*, 2(12):e223.

Casals, F. et Navarro, A. (2007). Chromosomal evolution : inversions : the chicken or the egg ? *Heredity*, 99(5):479–480.

Chessel, D. (1978). *Biométrie et Ecologie*, chapitre Description non paramétrique de la dispersion spatiale des individus d'une espèce, pages 45–135. Soc. Fr. Biométrie.

Chimpanzee Sequencing and Analysis Consortium (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 437(7055):69–87.

Choi, V., Zheng, C., Zhu, Q. et Sankoff, D. (2007). Algorithms for the extraction of synteny blocks from comparative maps. *In Workshop on Algorithms in Bioinformatics (WABI)*, volume 4645, pages 277–288. Springer Berlin / Heidelberg.

Chuzhanova, N., Abeysinghe, S. S., Krawczak, M. et Cooper, D. N. (2003). Translocation and gross deletion breakpoints in human inherited disease and cancer II : Potential involvement of repetitive sequence elements in secondary structure formation between DNA ends. *Hum Mutat*, 22(3):245–251.

Coghlan, A., Eichler, E. E., Oliver, S. G., Paterson, A. H. et Stein, L. (2005). Chromosome evolution in eukaryotes : a multi-kingdom perspective. *Trends Genet*, 21(12):673–682.

Coop, G., Wen, X., Ober, C., Pritchard, J. K. et Przeworski, M. (2008). High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. *Science*, 319(5868):1395–1398.

Costantini, M. et Bernardi, G. (2008). Replication timing, chromosomal bands, and isochores. *Proc Natl Acad Sci U S A*, 105:3433–3437.

Costantini, M., Clay, O., Auletta, F. et Bernardi, G. (2006). An isochore map of human chromosomes. *Genome Res*, 16(4):536–541.

Couronne, O., Poliakov, A., Bray, N., Ishkhanov, T., Ryaboy, D., Rubin, E., Pachter, L. et Dubchak, I. (2003). Strategies and tools for whole-genome alignments. *Genome Res*, 13(1):73–80.

Cáceres, M., of Health Intramural Sequencing Center Comparative Sequencing Program, N. I., Sullivan, R. T. et Thomas, J. W. (2007). A recurrent inversion on the eutherian X chromosome. *Proc Natl Acad Sci U S A*, 104:18571–18576.

D'Anjou, H., Chabot, C. et Chartrand, P. (2004). Preferential accessibility to specific genomic loci for the repair of double-strand breaks in human cells. *Nucleic Acids Res*, 32(20):6136–6143.

Darai-Ramqvist, E., Sandlund, A., Müller, S., Klein, G., Imreh, S. et Kost-Alimova, M. (2008). Segmental duplications and evolutionary plasticity at tumor chromosome break-prone regions. *Genome Res*, 18:370–379.

Darling, A. C. E., Mau, B., Blattner, F. R. et Perna, N. T. (2004). Mauve : multiple alignment of conserved genomic sequence with rearrangements. *Genome Res*, 14(7):1394–1403.

DeBry, R. W. et Seldin, M. F. (1996). Human/mouse homology relationships. *Genomics*, 33(3):337–351.

Dehal, P., Predki, P., Olsen, A. S., Kobayashi, A., Folta, P., Lucas, S., Land, M., Terry, A., Zhou, C. L. E., Rash, S. *et al.* (2001). Human chromosome 19 and related regions in mouse : conservative and lineage-specific evolution. *Science*, 293(5527):104–111.

Dennehey, B. K., Gutches, D. G., McConkey, E. H. et Krauter, K. S. (2004). Inversion, duplication, and changes in gene context are associated with human chromosome 18 evolution. *Genomics*, 83(3):493–501.

Derrien, T., André, C., Galibert, F. et Hitte, C. (2007). AutoGRAPH : an interactive web server for automating and visualizing comparative genome maps. *Bioinformatics*, 23(4): 498–499.

Dewey, C. N. et Pachter, L. (2006). Evolution at the nucleotide level : the problem of multiple whole-genome alignment. *Hum Mol Genet*, 15(Review Issue 1):R51–R56.

Dobigny, G., Ozouf-Costaz, C., Waters, P. D., Bonillo, C., Coutanceau, J.-P. et Volobouev, V. (2004). LINE-1 amplification accompanies explosive genome repatterning in rodents. *Chromosome Res*, 12(8):787–793.

Dufayard, J.-F., Duret, L., Penel, S., Gouy, M., Rechenmann, F. et Perrière, G. (2005). Tree pattern matching in phylogenetic trees : automatic search for orthologs or paralogs in homologous gene sequence databases. *Bioinformatics*, 21(11):2596–2603.

Duret, L. et Galtier, N. (2000). The covariation between TpA deficiency, CpG deficiency, and G+C content of human isochores is due to a mathematical artifact. *Mol Biol Evol*, 17(11):1620–1625.

Duret, L., Mouchiroud, D. et Gautier, C. (1995). Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores. *J Mol Evol*, 40(3):308–317.

Dutrillaux, B. (1997). Comment évoluent les chromosomes de mammifères. *La Recherche*, 296:70–75.

Ehrlich, J., Sankoff, D. et Nadeau, J. H. (1997). Synteny conservation and chromosome rearrangements during mammalian evolution. *Genetics*, 147(1):289–296.

Eichler, E. E. et Sankoff, D. (2003). Structural dynamics of eukaryotic chromosome evolution. *Science*, 301(5634):793–797.

Eyre-Walker, A. (1993). Recombination and mammalian genome evolution. *Proc Biol Sci*, 252(1335):237–243.

Eyre-Walker, A. et Hurst, L. D. (2001). The evolution of isochores. *Nat Rev Genet*, 2(7):549–555.

Ferguson-Smith, M. A. et Trifonov, V. (2007). Mammalian karyotype evolution. *Nat Rev Genet*, 8(12):950–962.

Feuk, L., Carson, A. R. et Scherer, S. W. (2006). Structural variation in the human genome. *Nat Rev Genet*, 7(2):85–97.

Frank, A. C. et Lobry, J. R. (1999). Asymmetric substitution patterns : a review of possible underlying mutational or selective mechanisms. *Gene*, 238(1):65–77.

Freeman, J. L., Perry, G. H., Feuk, L., Redon, R., McCarroll, S. A., Altshuler, D. M., Aburatani, H., Jones, K. W., Tyler-Smith, C., Hurles, M. E. *et al.* (2006). Copy number variation : New insights in genome diversity. *Genome Res*, 16(8):949–961.

Froenicke, L. (2005). Origins of primate chromosomes - as delineated by Zoo-FISH and alignments of human and mouse draft genome sequences. *Cytogenet Genome Res*, 108(1-3):122–138.

Fukagawa, T., Sugaya, K., Matsumoto, K., Okumura, K., Ando, A., Inoko, H. et Ikemura, T. (1995). A boundary of long-range G + C % mosaic domains in the human MHC locus : pseudoautosomal boundary-like sequence exists near the boundary. *Genomics*, 25(1):184–191.

Fullerton, S. M., Carvalho, A. B. et Clark, A. G. (2001). Local rates of recombination are positively correlated with GC content in the human genome. *Mol Biol Evol*, 18(6):1139–1142.

Galtier, N. (2003). Gene conversion drives GC content evolution in mammalian histones. *Trends Genet*, 19(2):65–68.

Galtier, N., Piganeau, G., Mouchiroud, D. et Duret, L. (2001). GC-content evolution in mammalian genomes : the biased gene conversion hypothesis. *Genetics*, 159(2):907–911.

Gibbs, R. A., Weinstock, G. M., Metzker, M. L., Muzny, D. M., Sodergren, E. J., Scherer, S., Scott, G., Steffen, D., Worley, K. C., Burch, P. E. *et al.* (2004). Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature*, 428(6982):493–521.

Gilbert, N., Boyle, S., Fiegler, H., Woodfine, K., Carter, N. P. et Bickmore, W. A. (2004). Chromatin architecture of the human genome : gene-rich domains are enriched in open chromatin fibers. *Cell*, 118(5):555–566.

Goidts, V., Szamalek, J. M., de Jong, P. J., Cooper, D. N., Chuzhanova, N., Hameister, H. et Kehrer-Sawatzki, H. (2005). Independent intrachromosomal recombination events underlie the pericentric inversions of chimpanzee and gorilla chromosomes homologous to human chromosome 16. *Genome Res*, 15(9):1232–1242.

Goidts, V., Szamalek, J. M., Hameister, H. et Kehrer-Sawatzki, H. (2004). Segmental duplication associated with the human-specific inversion of chromosome 18 : a further example of the impact of segmental duplications on karyotype and genome evolution in primates. *Hum Genet*, 115(2):116–122.

Gordon, L., Yang, S., Tran-Gyamfi, M., Baggott, D., Christensen, M., Hamilton, A., Crooijmans, R., Groenen, M., Lucas, S., Ovcharenko, I. *et al.* (2007). Comparative analysis of chicken chromosome 28 provides new clues to the evolutionary fragility of gene-rich vertebrate regions. *Genome Res*, 17:1603–1613.

Gotter, A. L., Nimmakayalu, M. A., Jalali, G. R., Hacker, A. M., Vorstman, J., Duffy, D. C., Medne, L. et Emanuel, B. S. (2007). A palindrome-driven complex rearrangement of 22q11.2 and 8q24.1 elucidated using novel technologies. *Genome Res*, 17:470–481.

Gray, Y. H. (2000). It takes two transposons to tango : transposable-element-mediated chromosomal rearrangements. *Trends Genet*, 16(10):461–468.

Haas, B. J., Delcher, A. L., Wortman, J. R. et Salzberg, S. L. (2004). DAGchainer : a tool for mining segmental genome duplications and synteny. *Bioinformatics*, 20(18):3643–3646.

Hampson, S., McLysaght, A., Gaut, B. et Baldi, P. (2003). LineUp : statistical detection of chromosomal homology with application to plant comparative genomics. *Genome Res*, 13(5):999–1010.

Han, K., Sen, S. K., Wang, J., Callinan, P. A., Lee, J., Cordaux, R., Liang, P. et Batzer, M. A. (2005). Genomic rearrangements by LINE-1 insertion-mediated deletion in the human and chimpanzee lineages. *Nucleic Acids Res*, 33(13):4040–4052.

Han, L., Su, B., Li, W.-H. et Zhao, Z. (2008). Cpg island density and its correlations with genomic features in mammalian genomes. *Genome Biol*, 9:R79.

Handt, O., Sutherland, G. R. et Richards, R. I. (2000). Fragile sites and minisatellite repeat instability. *Mol Genet Metab*, 70(2):99–105.

Hinsch, H. et Hannenhalli, S. (2006). Recurring genomic breaks in independent lineages support genomic fragility. *BMC Evol Biol*, 6:90.

Hoberman, R., Sankoff, D. et Durand, D. (2005). The statistical analysis of spatially clustered genes under the maximum gap criterion. *J Comput Biol*, 12(8):1083–1102.

Hsu, F., Kent, W. J., Clawson, H., Kuhn, R. M., Diekhans, M. et Haussler, D. (2006). The UCSC Known Genes. *Bioinformatics*, 22(9):1036–1046.

Hubbard, T. J. P., Aken, B. L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T. *et al.* (2007). Ensembl 2007. *Nucleic Acids Res*, 35(Database issue):D610–D617.

Hurst, L. D., Pál, C. et Lercher, M. J. (2004). The evolutionary dynamics of eukaryotic gene order. *Nat Rev Genet*, 5(4):299–310.

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(9):1278–1285.

Iwase, M., Satta, Y., Hirai, Y., Hirai, H., Imai, H. et Takahata, N. (2003). The amelogenin loci span an ancient pseudoautosomal boundary in diverse mammalian species. *Proc Natl Acad Sci U S A*, 100(9):5258–5263.

Ji, Y., Eichler, E. E., Schwartz, S. et Nicholls, R. D. (2000). Structure of chromosomal duplicons and their role in mediating human genomic disorders. *Genome Res*, 10(5):597–610.

Karolchik, D., Kuhn, R. M., Baertsch, R., Barber, G. P., Clawson, H., Diekhans, M., Giardine, B., Harte, R. A., Hinrichs, A. S., Hsu, F. *et al.* (2008). The UCSC Genome Browser Database : 2008 update. *Nucleic Acids Res*, 36(Database issue):D773–D779.

Kato, T., Inagaki, H., Kogo, H., Ohye, T., Yamada, K., Emanuel, B. S. et Kurahashi, H. (2008). Two different forms of palindrome resolution in the human genome : deletion or translocation. *Hum Mol Genet*, 17(8):1184–1191.

Kehrer-Sawatzki, H. et Cooper, D. N. (2008). Molecular mechanisms of chromosomal rearrangement during primate evolution. *Chromosome Res*, 16(1):41–56.

Kehrer-Sawatzki, H., Sandig, C., Chuzhanova, N., Goidts, V., Szamalek, J. M., Tänzer, S., Müller, S., Platzer, M., Cooper, D. N. et Hameister, H. *et al.* (2005a). Breakpoint analysis of the pericentric inversion distinguishing human chromosome 4 from the homologous chromosome in the chimpanzee (Pan troglodytes). *Hum Mutat*, 25(1):45–55.

Kehrer-Sawatzki, H., Sandig, C. A., Goidts, V. et Hameister, H. (2005b). Breakpoint analysis of the pericentric inversion between chimpanzee chromosome 10 and the homologous chromosome 12 in humans. *Cytogenet Genome Res*, 108(1-3):91–97.

Kehrer-Sawatzki, H., Schreiner, B., Tänzer, S., Platzer, M., Müller, S. et Hameister, H. (2002). Molecular characterization of the pericentric inversion that causes differences between chimpanzee chromosome 19 and human chromosome 17. *Am J Hum Genet*, 71(2):375–388.

Kent, W. J. (2002). BLAT–the BLAST-like alignment tool. *Genome Res*, 12(4):656–664.

Kent, W. J., Baertsch, R., Hinrichs, A., Miller, W. et Haussler, D. (2003). Evolution's cauldron : duplication, deletion, and rearrangement in the mouse and human genomes. *Proc Natl Acad Sci U S A*, 100(20):11484–11489.

Kidd, J. M., Cooper, G. M., Donahue, W. F., Hayden, H. S., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., Antonacci, F. *et al.* (2008). Mapping and sequencing of structural variation from eight human genomes. *Nature*, 453(7191):56–64.

Kleinjan, D. J. et van Heyningen, V. (1998). Position effect in human genetic disease. *Hum Mol Genet*, 7(10):1611–1618.

Kong, A., Gudbjartsson, D. F., Sainz, J., Jonsdottir, G. M., Gudjonsson, S. A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G. *et al.* (2002). A high-resolution recombination map of the human genome. *Nat Genet*, 31(3):241–247.

Korbel, J. O., Urban, A. E., Affourtit, J. P., Godwin, B., Grubert, F., Simons, J. F., Kim, P. M., Palejev, D., Carriero, N. J., Du, L. *et al.* (2007). Paired-end mapping reveals extensive structural variation in the human genome. *Science*, 318(5849):420–426.

Kurahashi, H., Inagaki, H., Hosoba, E., Kato, T., Ohye, T., Kogo, H. et Emanuel, B. S. (2007). Molecular cloning of a translocation breakpoint hotspot in 22q11. *Genome Res*, 17:461–469.

Kurahashi, H., Shaikh, T., Takata, M., Toda, T. et Emanuel, B. S. (2003). The constitutional t(17;22) : another translocation mediated by palindromic AT-rich repeats. *Am J Hum Genet*, 72(3):733–738.

Lahn, B. T. et Page, D. C. (1999). Four evolutionary strata on the human X chromosome. *Science*, 286(5441):964–967.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W. *et al.* (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

Lemaitre, C., Braga, M., Gautier, C., Sagot, M.-F., Tannier, E. et Marais, G. A. B. (2008a). Footprints of inversions at present and past pseudoautosomal boundaries in human sex chromosomes. *soumis*.

Lemaitre, C. et Sagot, M.-F. (2008). A small trip in the untranquil world of genomes : A survey on the detection and analysis of genome rearrangement breakpoints. *Theor Comput Sci*, 395(2-3):171–192.

Lemaitre, C., Tannier, E., Gautier, C. et Sagot, M.-F. (2008b). Precise detection of rearrangement breakpoints in mammalian chromosomes. *BMC Bioinformatics*, 9(1):286.

Lemaitre, C., Zaghloul, L., Sagot, M.-F., Gautier, C., Arneodo, A., Tannier, E. et Audit, B. (2008c). Analysis of fine-scale mammalian evolutionary breakpoints provides new insight into their relations to genome organisation and open chromatin. *soumis*.

Lieber, M. R., Ma, Y., Pannicke, U. et Schwarz, K. (2003). Mechanism and regulation of human non-homologous DNA end-joining. *Nat Rev Mol Cell Biol*, 4(9):712–720.

Lobry, J. R. (1996). Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol Biol Evol*, 13(5):660–665.

Ma, B., Tromp, J. et Li, M. (2002). Pattern Hunter : faster and more sensitive homology search. *Bioinformatics*, 18:440–445.

Ma, J., Zhang, L., Suh, B. B., Raney, B. J., Burhans, R. C., Kent, W. J., Blanchette, M., Haussler, D. et Miller, W. (2006). Reconstructing contiguous regions of an ancestral genome. *Genome Res*, 16:1557–1565.

Macaya, G., Thiery, J. P. et Bernardi, G. (1976). An approach to the organization of eukaryotic genomes at a macromolecular level. *J Mol Biol*, 108(1):237–254.

Machado, C. A., Kliman, R. M., Markert, J. A. et Hey, J. (2002). Inferring the history of speciation from multilocus DNA sequence data : the case of Drosophila pseudoobscura and close relatives. *Mol Biol Evol*, 19(4):472–488.

Marais, G. (2003). Biased gene conversion : implications for genome and sex evolution. *Trends Genet*, 19(6):330–338.

Martin, S. et Pombo, A. (2003). Transcription factories : quantitative studies of nanostructures in the mammalian nucleus. *Chromosome Res*, 11(5):461–470.

McVean, G. A. T., Myers, S. R., Hunt, S., Deloukas, P., Bentley, D. R. et Donnelly, P. (2004). The fine-scale structure of recombination rate variation in the human genome. *Science*, 304(5670):581–584.

Meunier, J. et Duret, L. (2004). Recombination drives the evolution of GC-content in the human genome. *Mol Biol Evol*, 21(6):984–990.

Mouchiroud, D., D'Onofrio, G., Aïssani, B., Macaya, G., Gautier, C. et Bernardi, G. (1991). The distribution of genes in the human genome. *Gene*, 100:181–187.

Murphy, W. J., Larkin, D. M., van der Wind, A. E., Bourque, G., Tesler, G., Auvil, L., Beever, J. E., Chowdhary, B. P., Galibert, F., Gatzke, L. *et al.* (2005). Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science*, 309(5734):613–617.

Myers, S., Bottolo, L., Freeman, C., McVean, G. et Donnelly, P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. *Science*, 310(5746):321–324.

Nadeau, J. H. et Sankoff, D. (1998). The lengths of undiscovered conserved segments in comparative maps. *Mamm Genome*, 9(6):491–495.

Nadeau, J. H. et Taylor, B. A. (1984). Lengths of chromosomal segments conserved since divergence of man and mouse. *Proc Natl Acad Sci U S A*, 81(3):814–818.

Navarro, A. et Barton, N. H. (2003). Accumulating postzygotic isolation genes in parapatry : a new twist on chromosomal speciation. *Evolution Int J Org Evolution*, 57(3):447–459.

Navratil, V. (2005). *Modélisation des connaissances en génomique par une approche de cartographie comparée : Application à la détection et à l'analyse des SNPs exoniques chez les vertébrés.* Thèse de doctorat, Université Claude Bernard Lyon 1.

Needleman, S. B. et Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, 48(3):443–453.

Nergadze, S. G., Rocchi, M., Azzalin, C. M., Mondello, C. et Giulotto, E. (2004). Insertion of telomeric repeats at intrachromosomal break sites during primate evolution. *Genome Res*, 14(9):1704–1710.

Newman, T. L., Tuzun, E., Morrison, V. A., Hayden, K. E., Ventura, M., McGrath, S. D., Rocchi, M. et Eichler, E. E. (2005). A genome-wide survey of structural variation between human and chimpanzee. *Genome Res*, 15(10):1344–1356.

Noor, M. A., Grams, K. L., Bertucci, L. A. et Reiland, J. (2001). Chromosomal inversions and the reproductive isolation of species. *Proc Natl Acad Sci U S A*, 98(21):12084–12088.

Notebaart, R. A., Huynen, M. A., Teusink, B., Siezen, R. J. et Snel, B. (2005). Correlation between sequence conservation and the genomic context after gene duplication. *Nucleic Acids Res*, 33(19):6164–6171.

O'Brien, K. P., Remm, M. et Sonnhammer, E. L. L. (2005). Inparanoid : a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res*, 33(Database issue):D476–D480.

O'driscoll, M. et Jeggo, P. A. (2006). The role of double-strand break repair - insights from human genetics. *Nat Rev Genet*, 7(1):45–54.

Ohno, S. (1973). Ancient linkage groups and frozen accidents. *Nature*, 244:259–262.

Ovcharenko, I., Loots, G. G., Nobrega, M. A., Hardison, R. C., Miller, W. et Stubbs, L. (2005). Evolution and functional classification of vertebrate gene deserts. *Genome Res*, 15(1):137–145.

Page, R. D. et Charleston, M. A. (1997). From gene to organismal phylogeny : reconciled trees and the gene tree/species tree problem. *Mol Phylogenet Evol*, 7(2):231–240.

Pan, X., Stein, L. et Brendel, V. (2005). SynBrowse : a synteny browser for comparative sequence analysis. *Bioinformatics*, 21(17):3461–3468.

Passarge, E., Horsthemke, B. et Farber, R. A. (1999). Incorrect use of the term synteny. *Nat Genet*, 23(4):387.

Pavesi, G., Mauri, G., Iannelli, F., Gissi, C. et Pesole, G. (2004). GeneSyn : a tool for detecting conserved gene order across genomes. *Bioinformatics*, 20(9):1472–1474.

Peng, Q., Pevzner, P. A. et Tesler, G. (2006). The fragile breakage versus random breakage models of chromosome evolution. *PLoS Comput Biol*, 2:e14.

Peterlongo, P. (2006). *Filtrage de séquences d'ADN pour la recherche de longues répétitions multiples*. Thèse de doctorat, Université de Marne-la-Vallée.

Peterlongo, P., Pisanti, N., Boyer, F., Pereira do Lago, A. et Sagot, M.-F. (2008). Lossless filter for multiple repetitions with hamming distance. *Journal of Discrete Algorithms*, 6(3):497–509.

Pevzner, P. et Tesler, G. (2003a). Genome rearrangements in mammalian evolution : lessons from human and mouse genomes. *Genome Res*, 13(1):37–45.

Pevzner, P. et Tesler, G. (2003b). Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proc Natl Acad Sci U S A*, 100(13):7672–7677.

Pfeiffer, P., Goedecke, W. et Obe, G. (2000). Mechanisms of DNA double-strand break repair and their potential to induce chromosomal aberrations. *Mutagenesis*, 15(4):289–302.

Piálek, J., Hauffe, H. C., Rodríguez-Clark, K. M. et Searle, J. B. (2001). Raciation and speciation in house mice from the Alps : the role of chromosomes. *Mol Ecol*, 10(3):613–625.

Ponger, L. et Mouchiroud, D. (2002). CpGProD : identifying CpG islands associated with transcription start sites in large genomic mammalian sequences. *Bioinformatics*, 18(4):631–633.

Ptak, S. E., Hinds, D. A., Koehler, K., Nickel, B., Patil, N., Ballinger, D. G., Przeworski, M., Frazer, K. A. et Pääbo, S. (2005). Fine-scale recombination patterns differ between chimpanzees and humans. *Nat Genet*, 37(4):429–434.

Ranz, J. M., Maurin, D., Chan, Y. S., von Grotthuss, M., Hillier, L. W., Roote, J., Ashburner, M. et Bergman, C. M. (2007). Principles of Genome Evolution in the Drosophila melanogaster Species Group. *PLoS Biol*, 5(6):e152.

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., Shapero, M. H., Carson, A. R., Chen, W. *et al.* (2006). Global variation in copy number in the human genome. *Nature*, 444(7118):444–454.

Rice, P., Longden, I. et Bleasby, A. (2000). EMBOSS : the European Molecular Biology Open Software Suite. *Trends Genet*, 16(6):276–277.

Richard, F., Lombard, M. et Dutrillaux, B. (2003). Reconstruction of the ancestral karyotype of eutherian mammals. *Chromosome Res*, 11(6):605–618.

Rieseberg, L. (2001). Chromosomal rearrangements and speciation. *Trends Ecol Evol*, 16(7):351–358.

Roberto, R., Capozzi, O., Wilson, R. K., Mardis, E. R., Lomiento, M., Tuzun, E., Cheng, Z., Mootnick, A. R., Archidiacono, N., Rocchi, M. *et al.* (2006). Molecular refinement of gibbon genome rearrangement. *Genome Res*, 17:249–257.

Rocha, E. P. C. (2004). Order and disorder in bacterial genomes. *Curr Opin Microbiol*, 7(5):519–527.

Ross, M. T., Grafham, D. V., Coffey, A. J., Scherer, S., McLay, K., Muzny, D., Platzer, M., Howell, G. R., Burrows, C., Bird, C. P. *et al.* (2005). The DNA sequence of the human X chromosome. *Nature*, 434(7031):325–337.

Rothstein, R., Michel, B. et Gangloff, S. (2000). Replication fork pausing and recombination or "gimme a break". *Genes Dev*, 14(1):1–10.

Ruiz-Herrera, A., Castresana, J. et Robinson, T. J. (2006). Is mammalian chromosomal evolution driven by regions of genome fragility ? *Genome Biol*, 7(12):R115.

Ruiz-Herrera, A., García, F., Giulotto, E., Attolini, C., Egozcue, J., Ponsà, M. et Garcia, M. (2005a). Evolutionary breakpoints are co-localized with fragile sites and intrachromosomal telomeric sequences in primates. *Cytogenet Genome Res*, 108(1-3):234–247.

Ruiz-Herrera, A., García, F., Mora, L., Egozcue, J., Ponsà, M. et Garcia, M. (2005b). Evolutionary conserved chromosomal segments in the human karyotype are bounded by unstable chromosome bands. *Cytogenet Genome Res*, 108(1-3):161–174.

Ruiz-Herrera, A., Ponsà, M., García, F., Egozcue, J. et García, M. (2002). Fragile sites in human and Macaca fascicularis chromosomes are breakpoints in chromosome evolution. *Chromosome Res*, 10(1):33–44.

Ruiz-Herrera, A. et Robinson, T. (2007). Chromosomal instability in Afrotheria : fragile sites, evolutionary breakpoints and phylogenetic inference from genome sequence assemblies. *BMC Evol Biol*, 7(1):199.

Saccone, S., Sario, A. D., Wiegant, J., Raap, A. K., Valle, G. D. et Bernardi, G. (1993). Correlations between isochores and chromosomal bands in the human genome. *Proc Natl Acad Sci U S A*, 90(24):11929–11933.

Sankoff, D. (2006). The signal in the genomes. *PLoS Comput Biol*, 2(4):e35.

Sankoff, D. et Nadeau, J. H. (2003). Chromosome rearrangements in evolution : From gene order to genome sequence and back. *Proc Natl Acad Sci U S A*, 100(20):11188–11189.

Sankoff, D. et Trinh, P. (2005). Chromosomal breakpoint reuse in genome sequence rearrangement. *J Comput Biol*, 12(6):812–821.

Savage, J. R. (1998). A brief survey of aberration origin theories. *Mutat Res*, 404(1-2):139–147.

Schibler, L., Roig, A., Mahe, M.-F., Laurent, P., Hayes, H., Rodolphe, F. et Cribiu, E. (2006). High-resolution comparative mapping among man, cattle and mouse suggests a role for repeat sequences in mammalian genome evolution. *BMC Genomics*, 7(1):194.

Schmegner, C., Hameister, H., Vogel, W. et Assum, G. (2007). Isochores and replication time zones : a perfect match. *Cytogenet Genome Res*, 116(3):167–172.

Schmidt, T. et Frishman, D. (2008). Assignment of isochores for all completely sequenced vertebrate genomes using a consensus. *Genome Biol*, 9:R104.

Schwartz, M., Zlotorynski, E. et Kerem, B. (2006). The molecular basis of common and rare fragile sites. *Cancer Lett*, 232(1):13–26.

Schwartz, S., Kent, W. J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R. C., Haussler, D. et Miller, W. (2003). Human-mouse alignments with BLASTZ. *Genome Res*, 13(1):103–107.

Sharp, A. J., Locke, D. P., McGrath, S. D., Cheng, Z., Bailey, J. A., Vallente, R. U., Pertz, L. M., Clark, R. A., Schwartz, S., Segraves, R. *et al.* (2005). Segmental duplications and copy-number variation in the human genome. *Am J Hum Genet*, 77(1):78–88.

Sinha, A. U. et Meller, J. (2007). Cinteny : flexible analysis and visualization of synteny and genome rearrangements in multiple organisms. *BMC Bioinformatics*, 8:82.

Skaletsky, H., Kuroda-Kawaguchi, T., Minx, P. J., Cordum, H. S., Hillier, L., Brown, L. G., Repping, S., Pyntikova, T., Ali, J., Bieri, T. *et al.* (2003). The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature*, 423(6942):825–837.

Smith, T. F. et Waterman, M. S. (1981). Identification of common molecular subsequences. *J Mol Biol*, 147(1):195–197.

Soriano, P., Meunier-Rotival, M. et Bernardi, G. (1983). The distribution of interspersed repeats is nonuniform and conserved in the mouse and human genomes. *Proc Natl Acad Sci U S A*, 80(7):1816–1820.

Sproul, D., Gilbert, N. et Bickmore, W. A. (2005). The role of chromatin structure in regulating the expression of clustered genes. *Nat Rev Genet*, 6(10):775–781.

Stankiewicz, P. et Lupski, J. R. (2002). Genome architecture, rearrangements and genomic disorders. *Trends Genet*, 18(2):74–82.

Stankiewicz, P., Park, S. S., Inoue, K. et Lupski, J. R. (2001). The evolutionary chromosome translocation 4 ;19 in Gorilla gorilla is associated with microduplication of the chromosome fragment syntenic to sequences surrounding the human proximal CMT1A-REP. *Genome Res*, 11(7):1205–1210.

Sturtevant, A. H. (1921). A Case of Rearrangement of Genes in Drosophila. *Proc Natl Acad Sci U S A*, 7(8):235–237.

Sun, Y. et Buhler, J. (2006). Choosing the best heuristic for seeded alignment of DNA sequences. *BMC Bioinformatics*, 7(1):133.

Swidan, F., Rocha, E. P. C., Shmoish, M. et Pinter, R. Y. (2006). An integrative method for accurate comparative genome mapping. *PLoS Comput Biol*, 2(8):e75.

Szamalek, J. M., Goidts, V., Chuzhanova, N., Hameister, H., Cooper, D. N. et Kehrer-Sawatzki, H. (2005). Molecular characterisation of the pericentric inversion that distinguishes human chromosome 5 from the homologous chimpanzee chromosome. *Hum Genet*, 117(2-3):168–176.

Szamalek, J. M., Goidts, V., Searle, J. B., Cooper, D. N., Hameister, H. et Kehrer-Sawatzki, H. (2006). The chimpanzee-specific pericentric inversions that distinguish humans and chimpanzees have identical breakpoints in Pan troglodytes and Pan paniscus. *Genomics*, 87(1):39–45.

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

Sémon, M. et Duret, L. (2006). Evolutionary origin and maintenance of coexpressed gene clusters in mammals. *Mol Biol Evol*, 23(9):1715–1723.

Tatusov, R. L., Koonin, E. V. et Lipman, D. J. (1997). A genomic perspective on protein families. *Science*, 278(5338):631–637.

Thiery, J. P., Macaya, G. et Bernardi, G. (1976). An analysis of eukaryotic genomes by density gradient centrifugation. *J Mol Biol*, 108(1):219–235.

Trinh, P., McLysaght, A. et Sankoff, D. (2004). Genomic features in the breakpoint regions between syntenic blocks. *Bioinformatics*, 20 Suppl 1:I318–I325.

Tuzun, E., Sharp, A. J., Bailey, J. A., Kaul, R., Morrison, V. A., Pertz, L. M., Haugen, E., Hayden, H., Albertson, D., Pinkel, D. *et al.* (2005). Fine-scale structural variation of the human genome. *Nat Genet*, 37(7):727–732.

Usdin, K. (2008). The biological effects of simple tandem repeats : Lessons from the repeat expansion diseases. *Genome Res*, 18(7):1011–1019.

van der Wind, A. E., Kata, S. R., Band, M. R., Rebeiz, M., Larkin, D. M., Everts, R. E., Green, C. A., Liu, L., Natarajan, S., Goldammer, T. *et al.* (2004). A 1463 gene cattle-human comparative map with anchor points defined by human genome sequence coordinates. *Genome Res*, 14(7):1424–1437.

Vandepoele, K., Saeys, Y., Simillion, C., Raes, J. et Peer, Y. V. D. (2002). The automatic detection of homologous regions (ADHoRe) and its application to microcolinearity between Arabidopsis and rice. *Genome Res*, 12(11):1792–1801.

Verschure, P. J. (2006). Chromosome organization and gene control : it is difficult to see the picture when you are inside the frame. *J Cell Biochem*, 99(1):23–34.

Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P. *et al.* (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–562.

Webber, C. et Ponting, C. P. (2005). Hotspots of mutation and breakage in dog and human chromosomes. *Genome Res*, 15:1787–1797.

White, M. J. D. (1978). *Modes of speciation*. W.H. Freeman and company, San Francisco.

Winckler, W., Myers, S. R., Richter, D. J., Onofrio, R. C., McDonald, G. J., Bontrop, R. E., McVean, G. A. T., Gabriel, S. B., Reich, D., Donnelly, P. *et al.* (2005). Comparison of fine-scale recombination rates in humans and chimpanzees. *Science*, 308(5718):107–111.

Wolfe, K. H., Sharp, P. M. et Li, W. H. (1989). Mutation rates differ among regions of the mammalian genome. *Nature*, 337(6204):283–285.

Yue, Y., Grossmann, B., Ferguson-Smith, M., Yang, F. et Haaf, T. (2005). Comparative cytogenetics of human chromosome 3q21.3 reveals a hot spot for ectopic recombination in hominoid evolution. *Genomics*, 85(1):36–47.

Yunis, J. J., Sawyer, J. R. et Dunham, K. (1980). The striking resemblance of high-resolution g-banded chromosomes of man and chimpanzee. *Science*, 208(4448):1145–1148.

Zdobnov, E. M., von Mering, C., Letunic, I., Torrents, D., Suyama, M., Copley, R. R., Christophides, G. K., Thomasova, D., Holt, R. A., Subramanian, G. M. *et al.* (2002). Comparative genome and proteome analysis of Anopheles gambiae and Drosophila melanogaster. *Science*, 298(5591):149–159.

Zhang, Z., Raghavachari, B., Hardison, R. C. et Miller, W. (1994). Chaining multiple-alignment blocks. *J Comput Biol*, 1(3):217–226.

Zheng, C. et Sankoff, D. (2006). Rearrangement of noisy genomes. *In International Conference on Computational Science (2)*, volume 3992, pages 791–798. Springer Berlin / Heidelberg.

TITRE en français

Réarrangements chromosomiques dans les génomes de mammifères : caractérisation des points de cassure

RÉSUMÉ en français

Les réarrangements chromosomiques sont des mutations qui modifient la structure et l'organisation des génomes. Ils sont ici étudiés dans le cadre de l'évolution des génomes de mammifères. L'objectif de ces travaux est de caractériser les régions du génome qui ont subi de tels évènements ; elles sont appelées des points de cassure. Dans un premier temps, nous avons développé une méthode permettant d'identifier précisément ces régions sur un génome par comparaison avec un génome d'espèce différente. Nous montrons qu'elle améliore nettement la résolution par rapport aux méthodes existantes. Cela permet, dans un deuxième temps, d'analyser le contenu des séquences de points de cassure et leur répartition le long du génome. Plusieurs caractéristiques de séquences ont ainsi été identifiées dans les points de cassure chez l'homme, comme la perte de similarité avec les génomes comparés et la présence de duplications et d'éléments transposables. Enfin, nous montrons que les points de cassure ne sont pas répartis uniformément le long du génome, mais leur localisation serait fortement influencée par l'organisation des gènes et la structuration du génome en isochores.

MOTS-CLEFS en français

évolution des génomes ; génomique comparée ; réarrangements chromosomiques ; points de cassure ; blocs de synténie ; analyse de séquences

TITRE en anglais

Chromosomal rearrangements in mammalian genomes : characterising the breakpoints

RÉSUMÉ en anglais

Chromosomal rearrangements are large scale mutations that alter the structure and organisation of genomes. They are studied here in the scope of the evolution of the mammalian genomes. The aim of this work is to characterise the genomic regions which have undergone such events ; the latter are called breakpoints. We first developed a method to precisely localise these regions on a genome by comparison with the genome of another species. We showed that it markedly improves their resolution with respect to other published methods. This enables then to analyse the breakpoint sequences and their distribution along the genomes. Human breakpoints thus display several characteristics, such as the loss of similarity with related genomes and the presence of duplications and of transposable elements. Eventually, we argue that breakpoints are not randomly distributed along the genome, but instead their localisation seems to be linked with the gene organisation and the isochore landscape.

MOTS-CLEFS en anglais

genome evolution ; comparative genomics ; chromosomal rearrangements ; breakpoints ; synteny blocks ; sequence analysis

DISCIPLINE : Bioinformatique

INTITULE ET ADRESSE DE L'U.F.R. OU DU LABORATOIRE :

Laboratoire de Biométrie et Biologie Évolutive - UMR 5558 CNRS
Bâtiment Gregor Mendel - Université Claude Bernard Lyon1
43, bv du 11 novembre 1918 - 69622 Villeurbanne cedex