Etude des mécanismes évolutifs perturbant l'organisation des gènes dans les génomes de vertébrés

Camille Berthelot

UNIVERSITÉ PARIS SUD

Comprendre le monde, construire l'avenir®

UNIVERSITE PARIS-SUD

ÉCOLE DOCTORALE GÈNES, GÉNOMES, CELLULES
ÉCOLE DOCTORALE COMPLEXITÉ DU VIVANT
Laboratoire DYOGEN, Institut de Biologie de l'Ecole normale supérieure

DISCIPLINE : Génomique Comparative & Bioinformatique

THÈSE DE DOCTORAT

soutenue le 28/09/2012

par

Camille BERTHELOT

Etude des mécanismes évolutifs perturbant l'organisation des gènes dans les génomes de vertébrés

Directeur de thèse :	Hugues ROEST CROLLIUS	Directeur de recherche (IBENS)
Composition du jury :		
Président du jury :	Daniel GAUTHERET	Professeur (Université Paris-Sud)
Rapporteurs :	Nadia EL-MABROUK	Professeur (Université de Montréal)
	Hervé ISAMBERT	Directeur de recherche (Institut Curie)
Examinateurs :	Aoife McLYSAGHT	Directeur de recherche (University of Dublin)
	Eric TANNIER	Chargé de recherche (Université Lyon 1)

Remerciements

Cette thèse ne serait pas complète sans remercier tous ceux qui ont joué un rôle dans cette grande aventure qui m'aura vue passer d'Indiana Jones de la pipette à bioinformaticienne…

Je tiens à remercier sincèrement les membres de mon jury de thèse : Daniel Gautheret, Aoife McLysaght, Eric Tannier, et en particulier Nadia El-Mabrouk et Hervé Isambert, mes rapporteurs, pour avoir accepté d'évaluer mon travail. C'est un grand plaisir pour moi de soutenir cette thèse devant vous, j'espère que vous y aurez trouvé un intérêt également.

Un grand merci, bien sûr, à Hugues Roest Crollius, encadrant de thèse toujours enthousiaste et optimiste. Merci de m'avoir fait confiance dans des circonstances pas vraiment simples au début, j'espère que tu estimes que le pari est réussi (moi oui) ! Ces presque quatre ans sont passés bien vite, au cours desquels j'aurai beaucoup appris, entre liberté d'explorer toutes les pistes et discussions constructives à bâtons rompus pour prendre un peu de recul.

Je remercie chaleureusement tous les membres de l'équipe Dyogen : les Dyogen Girls, Alexandra, Marlène, Leila, Magali, Lilian, Céline et Amélie, et les Dyogen Boys, Charles, David, Joseph, et surtout Matthieu, co-bureau de l'extrême et Grand Pourvoyeur de Chocolat. Votre bonne humeur au quotidien et nos débats du midi vont me manquer ! Merci aussi à Stéphane Le Crom et Denis Thieffry, pour votre aide et vos conseils.

Je remercie particulièrement l'équipe du service informatique dirigée par Pierre Vincens pour l'environnement de travail de qualité dont nous bénéficions grâce à leur travail et leur patience : Arnaud, Catherine, Clarisse, Edouard et Jean-Pierre, merci ! Promis, un jour on arrivera à ne pas faire planter Heimdall le vendredi à 18h.

Je souhaite également remercier Brigitte Arnaud, Martine Duponchelle et Anne Comera-Grande. S'occuper des tâches administratives du laboratoire n'est pas de tout repos, merci pour votre efficacité et votre patience à toute épreuve.

Enfin, ces années de thèse auront été enrichies de belles rencontres parmi les personnalités du laboratoire, au gré des Monday Cakes, Happy Hours et séminaires étudiants : Thomas, Laurent, Sophie, Maria, Baptiste, Mathilde, Sébastien, Nina, Adrien, et j'en oublie sûrement…!

Au-delà du laboratoire, d'autres personnes auront contribué à maintenir mon équilibre mental pendant la thèse : tout d'abord, Sophie, Marie et Bénédicte, dites « les Cracottes ». Elles savent pourquoi. Ensuite, la bande des Lyonnais : Claire, François, Amélie, Charles, Stéphanie, Marion, Jérôme, Audrey, Arnaud, Thibaut, Soumaya et Pierre, toujours prêts lorsqu'il s'agit de me faire rire. Evidemment, la fine équipe de la prépa agreg et consorts, Estelle, Simon, Clara,

Guillaume, Marie, Fanny et Laurent, fiers adeptes de PéTho et de la pièce montée géante. Et n'oublions pas les artistes, surtout Julia et Alexandra, et leurs questions qui remettent les choses en perspective (« Mais au fait, c'est *quoi* exactement un génome ? ») !

Enfin, un immense merci à mes proches : mes parents, leurs conjoints, Léa, Alex, et aussi Jacques, mon grand-père. Merci pour votre enthousiasme et votre soutien jusqu'au bout de ce projet qui me tenait à cœur, merci d'avoir été là (même parfois à distance) et d'y avoir cru avec moi !

Et, peut-être le plus important… merci à Martin, qui a été à mes côtés pendant toutes ces années, dans les bons moments et les moins bons, et qui ne m'aura jamais laissée tomber une seule seconde. Heureusement que tu étais là !

Table des matières

Première Partie
Introduction

Chapitre 1. Préambule

1.1. Résumé de la thèse

Si la fonction des gènes comme unités de stockage, codage et expression de l'information génétique est désormais bien comprise, l'importance de leur organisation dans les génomes de vertébrés est encore mal connue. Cette organisation peut être affectée par différents processus mutationnels : le contenu en gènes du génome peut être modifié par des pertes et des gains de gènes, et leur ordre peut être modifié par des réarrangements chromosomiques. Ces mutations se produisent au hasard mais ne peuvent passer le filtre de l'évolution que si elles ne sont pas délétères pour le fonctionnement de l'organisme. Ainsi, l'étude des mécanismes évolutifs qui affectent l'organisation des gènes nous renseigne sur les contraintes fonctionnelles et structurelles qui s'exercent sur les génomes de vertébrés.

Ce travail de thèse s'organise autour de deux axes : le premier aborde la distribution des points de cassure de réarrangements évolutifs, et le second porte sur l'étude d'un cas de duplication complète du génome, la duplication 3R dans le génome du poisson zèbre. Ces deux projets font appel aux mêmes techniques d'analyse (étude des synténies conservées et/ou perdues) et s'intéressent tous deux à la même question biologique : quelles sont les forces qui gouvernent l'organisation des gènes dans les génomes de vertébrés ? Les résultats de ces travaux montrent que l'organisation des gènes est en réalité très peu soumise à sélection dans les génomes de vertébrés. Son évolution reflète majoritairement l'apparition et la fixation au hasard de mutations plutôt que des contraintes fonctionnelles. D'une part, les réarrangements chromosomiques sont un phénomène essentiellement neutre dont la probabilité d'apparition est liée à la structure du génome, et dont la fixation semble majoritairement aléatoire. D'autre part, le cas du poisson zèbre montre qu'après une duplication complète du génome, une sélection s'exerce sur les catégories de gènes retenus ou non sous forme de duplicats, mais pas sur l'organisation des gènes dans le génome.

1.2. Plan de la thèse

Le manuscrit est organisé en quatre parties. La première partie présente l'état actuel des connaissances sur les deux processus évolutifs étudiés au cours de cette thèse, les réarrangements chromosomiques et les duplications complètes du génome, ainsi que l'état de l'art sur l'utilisation des synténies conservées et des reconstructions de génomes ancestraux, qui sont à la base des analyses présentées ici.

La deuxième partie est une présentation synthétique des données et méthodes utilisées au cours de la thèse. Elle contient notamment un descriptif des méthodes d'analyse des synténies

conservées, des reconstructions de génomes ancestraux, et des méthodes statistiques utilisées (en particulier la régression de Poisson).

La troisième partie présente les résultats obtenus sur la distribution des points de cassure de réarrangements évolutifs dans les génomes d'un groupe de mammifères, les boreoeuthériens, par rapport à leur ancêtre commun. Cette partie propose le premier modèle statistique à rendre entièrement compte de la distribution des points de cassure de réarrangements dans les génomes de mammifères. La probabilité d'apparition d'un point de cassure entre deux gènes voisins est essentiellement une fonction de la distance qui sépare les gènes, avec une très faible déviation imputable à la sélection négative. Cette dernière préserve probablement certaines interactions importantes entre gènes et séquences régulatrices non codantes. Les résultats suggèrent que la compaction de la chromatine pourrait jouer un rôle primordial dans la probabilité d'apparition d'un réarrangement, mais que ce phénomène semble être neutre d'un point de vue évolutif dans la grande majorité des cas.

Dans la quatrième partie, nous présentons les résultats obtenus dans le cadre du consortium d'analyse du génome du poisson zèbre, pour lequel nous étions chargés d'analyser l'organisation du génome du poisson zèbre et l'impact de la duplication 3R. Le génome du poisson zèbre contient de nombreuses paires de gènes paralogues issus de cette duplication complète du génome, et des gènes revenus à l'état singleton suite à la perte d'une copie par délétion ou pseudogénisation. Les résultats montrent en particulier que la rétention des copies dupliquées s'est faite sur des critères de fonctions des protéines codées par les gènes, et non sur des critères d'organisation du génome. Nos observations confirment que l'organisation des gènes est plastique et est principalement gouvernée par le hasard plutôt que par des contraintes d'ordre fonctionnel. Le contenu en gènes, en revanche, évolue selon un patron qui est en partie reproductible d'un événement de duplication complète à l'autre.

La cinquième partie conclut le manuscrit par une discussion des principaux résultats obtenus, de leur importance dans le cadre de la génomique fonctionnelle, et des perspectives ouvertes par ce travail de thèse.

Chapitre 2. Notion de synténie conservée

Les notions de synténie et de synténie conservée sont sous-jacentes à toutes les études de génomique comparative s'intéressant à l'organisation du génome. Elles seront transversales à l'ensemble de ce travail de thèse, tant dans la partie sur l'étude de points de cassure de réarrangements évolutifs que dans celle portant sur l'analyse de l'organisation du génome du poisson zèbre. Or, ces notions anciennes ont évolué au fil des ans avec l'amélioration de la résolution des méthodes d'études des génomes. Ce chapitre revient rapidement sur ces notions afin de clarifier les définitions qui seront utilisées dans la suite du manuscrit, et aborde le lien entre synténie conservée, organisation ancestrale et évolution du génome.

2.1. Synténie au sens large

Initialement, le terme « synténie » est un synonyme de « liaison », pour désigner des gènes se trouvant sur le même chromosome dans un génome. Apparue dans les années 1970, la notion de synténie trouve son origine lors de la création de cartes d'hybrides d'irradiation. Cette méthode, basée sur la fusion d'une cellule irradiée au génome fractionné avec une cellule saine d'une autre espèce, a permis d'identifier des gènes portés par le même fragment de chromosome, qui sont donc en synténie au sens initial du terme (van Someren et al. 1974; Kucherlapati et al. 1975; Drillon and Fischer 2011).

Cette notion de synténie comme localisation des gènes sur le même chromosome, sans tenir compte de la distance qui les sépare ni de leur organisation relative, est ce que nous nommerons par la suite « synténie au sens large ». Ainsi, on peut tester la conservation de la synténie au sens large entre deux espèces, c'est-à-dire le degré auquel deux gènes qui se trouvent sur un même chromosome chez une espèce sont également sur un même chromosome dans une autre. Cette définition de la synténie peut sembler désuète aujourd'hui au vu de la résolution des données apportées par le séquençage massif de génomes amorcé depuis plusieurs années ; cependant, elle trouve encore toute sa place dans la comparaison de génomes très distants, ou très réarrangés.

2.2. Synténie au sens strict

Le concept de synténie a progressivement changé de sens dans les années 1990 avec la transition depuis la génétique classique vers la génomique, et en particulier la génomique comparative (Passarge et al. 1999; Drillon and Fischer 2011). L'établissement des premières

cartes génétiques, à la résolution bien plus fine que les cartes d'hybrides d'irradiation utilisées jusqu'alors, a permis d'évaluer la distance génétique séparant les marqueurs d'un même chromosome, et donc de déduire leur ordre relatif dans le génome. Ainsi, le terme de « synténie » s'est déplacé vers le sens de « ordre de gènes », et la notion de synténie conservée a commencé à être utilisée pour décrire la conservation de l'organisation (ordre et orientation) des gènes dans différents génomes. Par abus de langage, « synténie » est souvent utilisé aujourd'hui comme synonyme de « synténie conservée », c'est-à-dire pour désigner l'ordre des gènes uniquement dans le cas où il est conservé par rapport à une autre espèce, et donc d'origine ancestrale. Dans la suite du manuscrit, nous utiliserons le terme synténie dans cette acceptation stricte correspondant à l'ordre des gènes dans les génomes.

2.3. La synténie conservée comme vestige de l'organisation ancestrale

L'ordre des gènes dans les génomes n'est pas immuable : il est modifié par plusieurs types d'événements qui affectent soit le contenu en gènes du génome (pertes et gains de gènes), soit leur organisation les uns par rapport aux autres (réarrangements). Etant donné le grand nombre de gènes dans les génomes, en particulier de vertébrés (entre 20000 et 25000 gènes pour la plupart des espèces), il est très improbable que les gènes de deux espèces soient arrangés par hasard dans le même ordre de manière indépendante. Les régions où la synténie est conservée entre deux génomes sont donc considérées comme des vestiges de l'organisation des gènes héritée de leur dernier ancêtre commun, qui n'ont pas encore été remodelés par les processus évolutifs en œuvre dans les génomes.

L'étude des synténies conservées est donc un préalable nécessaire pour comprendre l'organisation ancestrale des gènes, son évolution dans différentes lignées apparentées, et les mécanismes qui sous-tendent cette dynamique. Les résultats présentés ici s'appuient de manière générale sur cette notion de synténie conservée comme vestige de l'organisation ancestrale du génome.

Chapitre 3. Les réarrangements chromosomiques évolutifs

3.1. Les réarrangements évolutifs : quelques définitions

Les réarrangements chromosomiques sont des événements mutationnels qui remanient la structure des chromosomes par cassure de l'ADN et réorganisation des blocs obtenus dans un ordre ou une orientation différents de ceux d'origine. On distingue classiquement deux classes de réarrangements : les réarrangements intrachromosomiques, qui ne font intervenir qu'une seule molécule d'ADN (inversions, fissions de chromosomes), et les réarrangements interchromosomiques qui en font intervenir au moins deux (translocations, transpositions et fusions de chromosomes). La taille du bloc dont la localisation ou l'orientation est modifiée peut être très variable, de quelques dizaines de bases à un bras chromosomique entier ; on les considère cependant comme distincts des indels, qui sont des insertions et délétions de séquences très courtes (quelques bases) dont l'origine n'est ni connue ni traçable.

Ces remaniements se font généralement sans modification de la séquence sous-jacente, sauf très localement au niveau des points de cassure. Les réarrangements peuvent en revanche amener à la duplication ou à la délétion d'une partie ou de l'ensemble du bloc réarrangé : on parle alors de réarrangements déséquilibrés, par opposition aux réarrangements équilibrés. Un réarrangement équilibré modifie le caryotype mais pas le contenu total en ADN, et en particulier en gènes, du génome. Les réarrangements, équilibrés ou non, peuvent être fonctionnellement neutres ou avoir des conséquences phénotypiques plus ou moins importantes que l'on regroupe chez l'homme sous le terme de désordres génomiques.

Les réarrangements évolutifs désignent les différences détectées entre les caryotypes de deux ou plusieurs espèces, c'est-à-dire les réarrangements qui se sont produits dans chacune des lignées depuis leur ancêtre commun. Ils sont généralement considérés comme fixés dans le génome de l'espèce, par opposition aux réarrangements polymorphes qui relèvent de la variation structurale au sein de la population. Ce terme de « réarrangement évolutif » sous-entend donc que les réarrangements observés ont passé avec succès le filtre de la sélection naturelle, et qu'il s'agit d'événements évolutivement neutres ou avantageux. Cependant, il faut noter que pour de nombreuses espèces, un seul génome de référence est disponible, provenant d'un individu sain : les réarrangements « évolutifs » ainsi observés au niveau de la séquence génomique sont alors un mélange de réarrangement fixés et de polymorphisme non ou peu délétère.

Les réarrangements les plus fréquents dans les génomes de vertébrés sont les événements équilibrés, qui ne modifient pas le contenu en gènes mais seulement leur ordre (Baptista et al. 2008). Parmi eux, la grande majorité sont des inversions, tant parmi les réarrangements évolutifs (Zhao and Bourque 2009) que parmi les variations structurales polymorphiques présentes dans les populations modernes (Kidd et al. 2010). Ainsi, le génome de référence de l'humain et celui du chimpanzé diffèrent par environ 1500 inversions, pour la plupart de l'ordre de quelques centaines de bases et dont certaines sont polymorphes dans le génome humain (Figure 3.1)(Feuk et al. 2005).

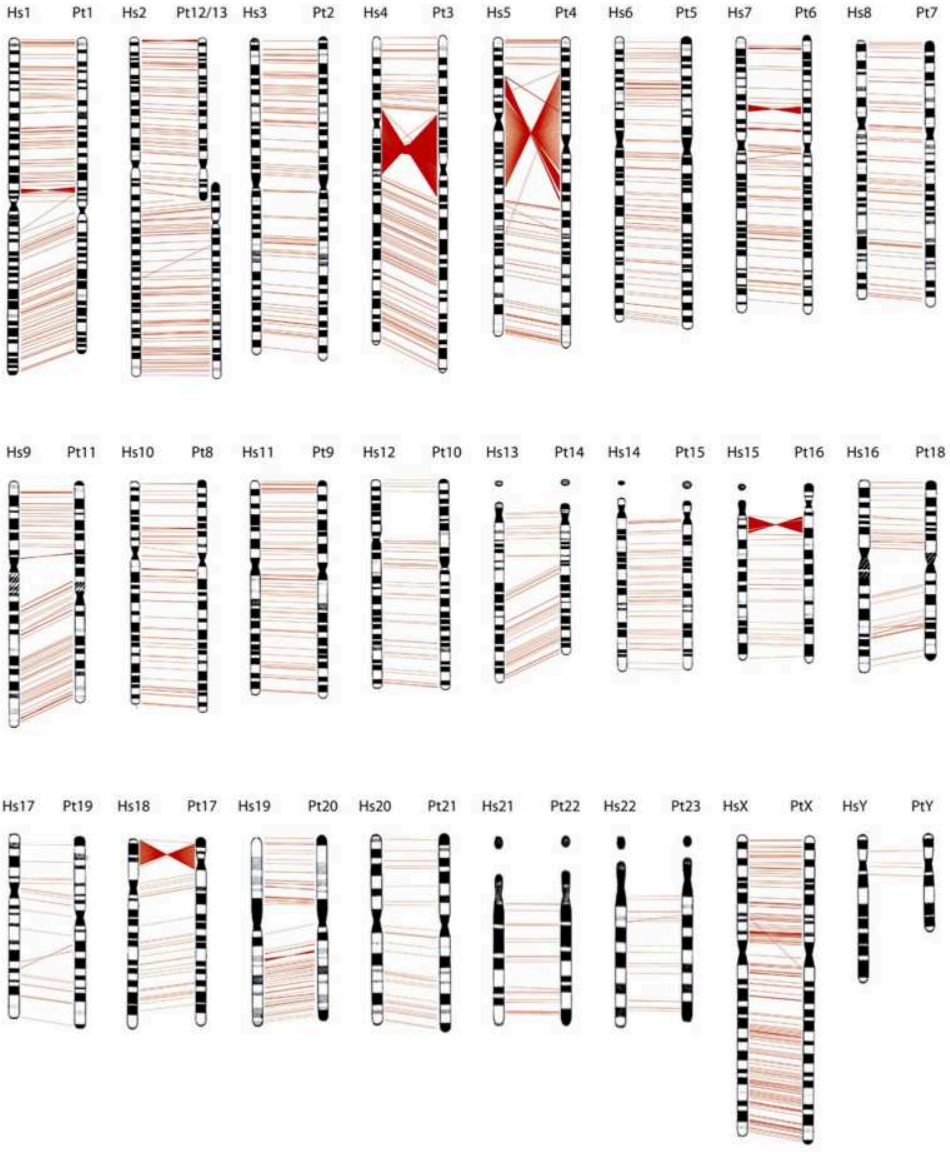


Figure 3.1. Inversions détectées entre les chromosomes orthologues de l'homme (Hs) et du chimpanzé (Pt). Chaque ligne rouge correspond à une inversion, les inversions les plus grandes (> 100 kb) étant représentées par un groupe de plusieurs lignes. Figure tirée de (Feuk et al. 2005).

3.2. Mécanismes moléculaires impliqués

Au niveau moléculaire, trois mécanismes principaux ont été identifiés comme causes possibles de réarrangements dans les génomes de vertébrés modernes, et particulièrement le génome humain où ils sont largement étudiés en raisons de leurs implications médicales. Comme tout autre type de mutation interspécifique, les réarrangements évolutifs représentent un échantillon de la variation existante dans la population qui a pu atteindre la fixation au fil des générations, par sélection ou par dérive. Les mécanismes biologiques qui causent les réarrangements polymorphiques dans les génomes modernes sont donc très certainement ceux qui ont également provoqué au départ les réarrangements évolutifs. Dans les génomes modernes, l'apparition de réarrangements est principalement attribuée à des erreurs lors de la réparation des cassures double-brin de l'ADN (Shaw and Lupski 2004; Liu et al. 2012). Les cassures double-brins sont fréquentes dans les cellules : leur taux moyen est estimé à 10 cassures double-brin par jour et par cellule (Lieber 2010). Leurs causes peuvent être endogènes (présence de nucléases, production d'espèces réactives de l'oxygène par la respiration cellulaire, fourches de réplication avortées, etc.) ou exogènes (radiations ionisantes, molécules chimiques exogènes). Alternativement, les réarrangements peuvent également être causés par des erreurs lors de la réplication.

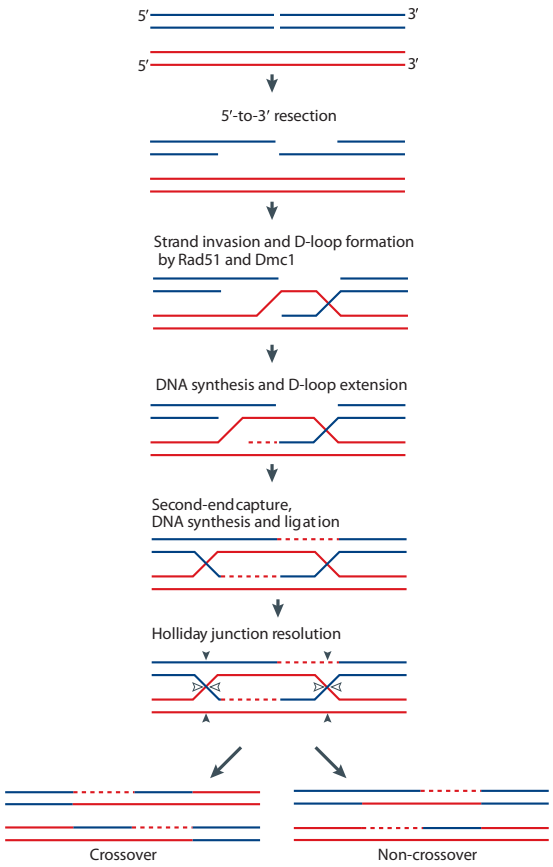


Figure 3.2. Réparation d'une cassure double-brin par recombinaison homologue. La résolution des jonctions de Holliday peut donner naissance à un crossover (réarrangement si la recombinaison est non-allélique). Figure adaptée de (Sasaki et al. 2010).

3.2.1. La recombinaison homologue illégitime

La recombinaison est le mécanisme de réparation des cassures double-brins le mieux étudié, car elle est notamment mise en jeu au cours de la méiose lors de la formation des cross-overs, nécessaires à la séparation des tétrades de chromosomes homologues. Dans le cas d'une méiose normale, la cassure double-brin est réparée par hybridation des brins d'ADN cassés avec le second allèle se trouvant sur le chromosome homologue, dont la séquence sert de guide pour réparer correctement la lésion (Figure 3.2). Les deux jonctions de Holliday sont alors résolues avec ou sans échange des bras chromosomiques entre les deux chromosomes homologues. Dans le cas de cassures accidentelles en dehors de la méiose, la recombinaison intervient dans la réparation essentiellement aux phases S et G2 du cycle, où la chromatide sœur issue de la réplication peut servir de guide à la réparation des cassures (Zierhut and Diffley 2008).

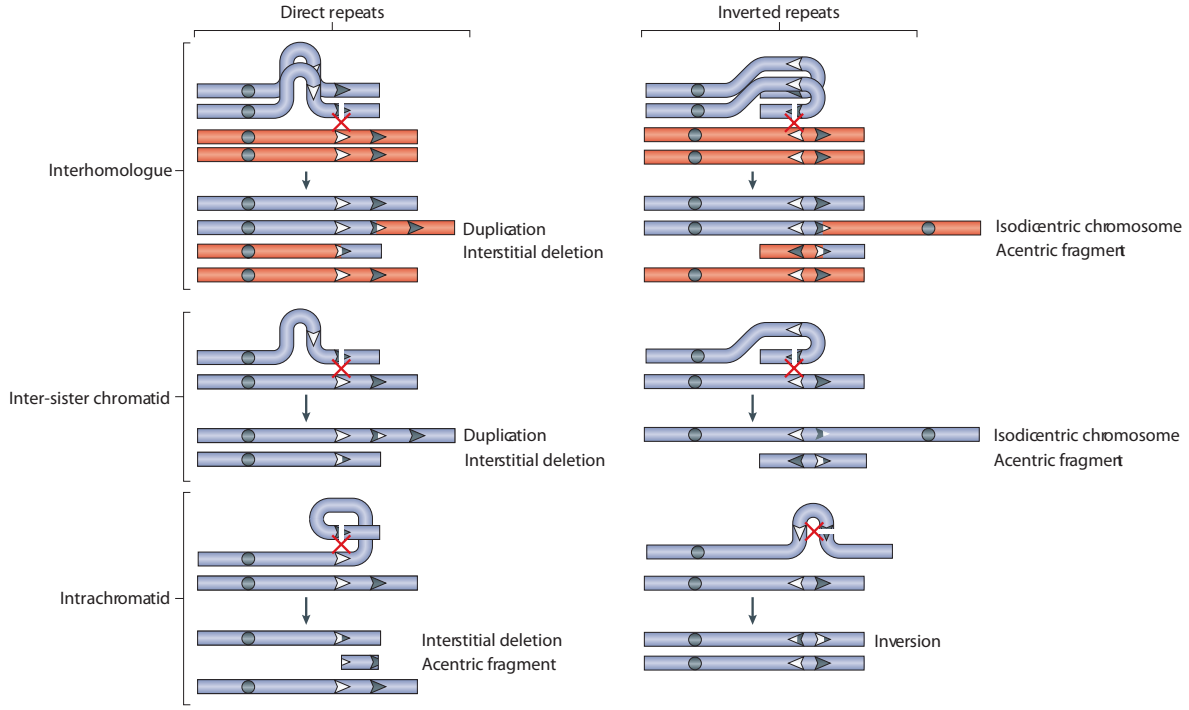


Figure 3.3. Réarrangements causés par recombinaison homologue non-allélique (NAHR). Un événement de recombinaison avec crossover peut résulter en une délétion, une duplication, une inversion ou un chromosome isodicentrique (à deux centromères) et un autre acentrique (sans centromère), selon l'orientation des séquences répétées ayant servi de substrat à la NAHR. Deux chromosomes homologues sont représentés en bleu et rouge ; les chromatides sœurs sont représentées de la même couleur. Les flèches représentent des séquences répétées. Figure tirée de (Sasaki et al. 2010).

Dans certains cas, cependant, la recombinaison peut avoir lieu par erreur entre deux séquences homologues mais non alléliques, c'est-à-dire des séquences identiques ou presque qui se trouvent à des endroits différents du génome et qui vont pouvoir s'hybrider. On parle alors de recombinaison ectopique ou de recombinaison homologue illégitime (Non-Allelic Homologous Recombination, ou NAHR). La résolution des jonctions de Holliday peut alors donner lieu à des

duplications, délétions ou réarrangements (Figure 3.3) : ainsi, la NAHR a été fréquemment invoquée comme un mécanisme majeur impliqué dans les variations chromosomiques (Stankiewicz and Lupski 2002; Stankiewicz and Lupski 2006; Sasaki et al. 2010; Ou et al. 2011). Dans la majorité des cas, la NAHR provoque des délétions ou des duplications qui modifient le nombre de copies d'un ou plusieurs gènes et se traduisent par des syndromes de type monosomie ou trisomie (modification de l'équilibre stœchiométrique de l'expression des gènes ; par exemple, le syndrome de Williams-Beuren, lié à une délétion au locus 7q11.23 dans le génome humain (Adams and Schmaier 2012)). Mais ce mécanisme pourrait également expliquer le cas des réarrangements sans modification du contenu en gènes se produisant de manière récurrente et indépendante dans les génomes, et en particulier le génome humain. La NAHR nécessite en effet un minimum de 300 à 500 pb de séquence parfaitement identique entre les deux brins d'ADN mis en jeu pour se produire (Reiter et al. 1998), et son efficacité dépend de la longueur, l'identité et la proximité physique des séquences homologues impliquées (Lupski 1998; Stankiewicz and Lupski 2002; Sharp et al. 2005). Les éléments transposables, duplications segmentales (régions de plus de 1 kb identiques à 90% ou plus dans le génome) et régions de faible complexité ont donc été proposés comme autant de séquences promouvant les réarrangements lors de la réparation de cassures double-brin. Ainsi, plusieurs cas de translocations récurrentes dans le génome humain s'expliquent par la présence de longues régions très répétées au niveau de leurs points de cassure typiques : par exemple, la translocation t(4 ;8)(p16 ;p23) résulte de recombinaisons entre deux clusters de gènes de récepteurs olfactifs, qui sont largement dupliqués dans le génome humain (Giglio et al. 2002; Maas et al. 2007).

Des cartes des grandes régions répétées du génome humain ont été établies pour prédire les régions de susceptibilité aux réarrangements récurrents (Sharp et al. 2005; Ou et al. 2011). Ces cartes mettent en évidence que si certains réarrangements récurrents connus ont en effet lieu entre régions de forte homologie, la grande majorité des régions de forte homologie ne donnent pas lieu à des réarrangements récurrents connus (Figure 3.4). La forte homologie ne semble donc pas être une condition suffisante pour provoquer l'apparition de réarrangements. Par ailleurs, les réarrangements récurrents ne représentent d'une petite partie des réarrangements observés dans les études de jeux de génomes récemment publiées : ils ne représentent que 11% des cas dans l'article de Ou et al. (2011), qui s'intéresse spécifiquement à l'implication de la NAHR dans les translocations. La majorité des réarrangements semblent être non récurrents, et leurs points de cassure se produisent en dehors des régions de forte homologie prédites par les cartes. En conclusion, s'il existe un ensemble cohérent d'indices montrant que la NAHR est probablement le mécanisme expliquant les réarrangements récurrents entre un petit nombre de régions dupliquées du génome, son implication dans la grande majorité des cas de réarrangements reste à démontrer.

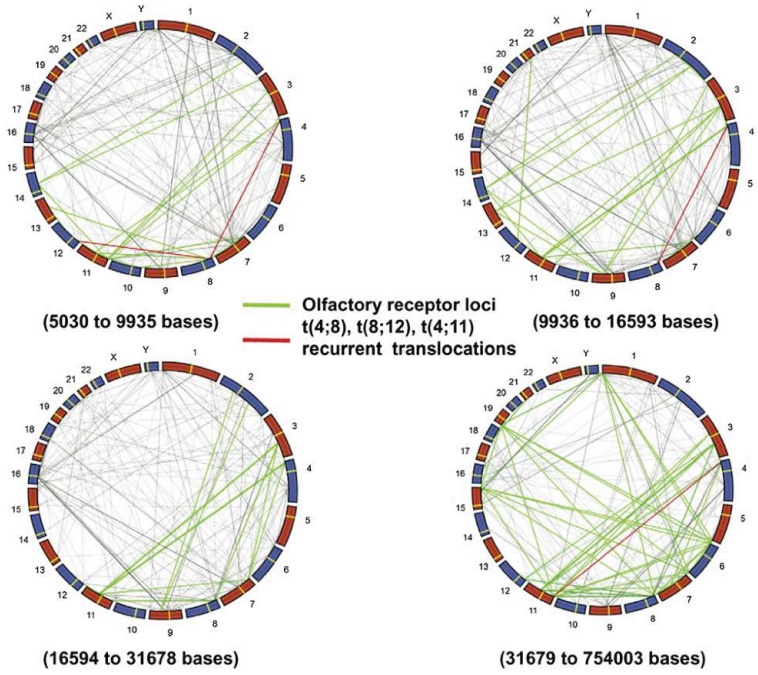


Figure 3.4. Cartographie des duplications segmentales interchromosomiques (> 5kb, > 94% d'identité) dans le génome humain. Les schémas représentent les quatre quartiles de la distribution des duplications segmentales, des plus courtes aux plus longues. Les duplications segmentales homologues sont reliées par un trait gris, et les locus des récepteurs olfactifs, très dupliqués dans le génome humain, sont reliés par des traits verts. Les duplications donnant lieu à des translocations récurrentes connues sont reliées par des traits rouges. La grande majorité des duplications segmentales du génome humain ne donnent pas naissance à des réarrangements récurrents connus. Figure tirée de (Ou et al. 2011).

3.2.2. La ligature d'extrémités non homologues

Dans de nombreux cas de réarrangements non récurrents et équilibrés, aucune homologie probante ne peut être détectée entre les régions de points de cassure d'un réarrangement (Baptista et al. 2008). Dans ces cas, le mécanisme explicatif le plus probable est que le réarrangement résulte de la ligature d'extrémités non homologues (Non Homologous End Joining, NHEJ). Il s'agit d'un mécanisme de réparation des cassures double-brins qui ne fait pas appel à une séquence homologue pour guider la réparation, mais recolle les bouts libres soit directement (NHEJ au sens strict), soit en se basant sur des microhomologies locales typiquement de 2 à 10 bases ; on parle alors parfois de NHEJ alternatif ou ligature conduite par les microhomologies (Microhomology Mediated End Joining, MMEJ). Cette voie de réparation existe chez la plupart des organismes, des bactéries aux mammifères, même si certaines protéines impliquées sont différentes (Symington and Gautier 2011). Ce mécanisme de réparation est flexible et permet de réparer rapidement une cassure en évitant la perte d'une partie du chromosome, en contrepartie du risque d'introduire des mutations à la cassure. Il est majoritaire pendant les phases G1 et M du cycle cellulaire, alors que la recombinaison homologue est essentiellement mise en jeu aux phases G2 et S (Zierhut and Diffley 2008).

Les microhomologies sur lesquelles se base le NHEJ pour réparer l'ADN étant très peu discriminantes, ce mécanisme peut conduire à des réarrangements dans le cas où plusieurs

cassures double-brin se trouvent à proximité physique dans le noyau (Soutoglou et al. 2007). Le NHEJ est d'ailleurs considéré comme étant le mécanisme principal conduisant aux aberrations chromosomiques dans les cellules tumorales (Lieber et al. 2008). Le NHEJ laisse une signature typique au niveau de la cassure réparée, puisqu'il introduit une édition des bouts libres en clivant ou ajoutant quelques bases afin de pouvoir lier les deux extrémités. Cette signature peut être détectée au niveau des points de cassure des réarrangements polymorphiques dans les génomes modernes sous la forme de petits indels ; en revanche, dans le cas des réarrangements évolutifs, la séquence autour des points de cassure est généralement trop divergente pour aligner les séquences et, a fortiori, détecter des traces de NHEJ (Lemaitre et al. 2008), à moins que les réarrangements ne soient très récents (Carbone et al. 2009; Girirajan et al. 2009). Lorsqu'aucune région d'homologie ne permettant d'attribuer le réarrangement à la NAHR, c'est généralement le NHEJ qui est invoqué pour expliquer l'événement même si sa signature n'est pas détectable en raison de la divergence des séquences (Carbone et al. 2009).

3.2.3. Le FoSTeS

Enfin, le troisième mécanisme de réarrangement proposé n'est pas lié aux cassures double-brin, mais à la réplication de l'ADN. En effet, certaines cassures seraient dues à des changements de brin matrice au cours de la réplication lorsque la fourche de réplication bute et que la machinerie moléculaire se décroche : le brin en cours de synthèse pourrait alors se réhybrider à une nouvelle matrice au niveau d'une autre fourche en utilisent des microhomologies locales de quelques bases, et reprendre la synthèse (Fork Stalling and Template Switching, FoSTeS ; Figure 3.5)(Lee et al. 2007).

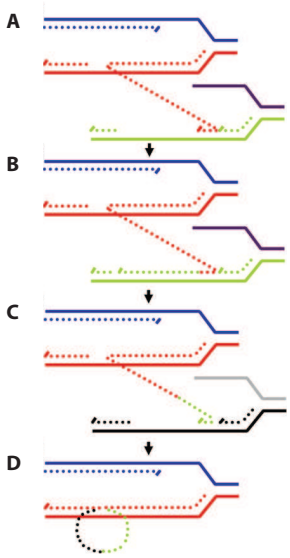


Figure 3.5. Mécanisme du FoSTeS. Le brin tardif de la fourche de réplication figurée en bleu et rouge se décroche et vient envahir une seconde fourche de réplication (verte et violette ; A). La réplication se poursuit (B), puis le brin se désengage et peut soit envahir une autre fourche de réplication (C), soit reprendre l'élongation dans sa fourche d'origine (D). Les traits pointillés représentent les brins d'ADN en cours de synthèse. Figure tirée de (Lee et al. 2007).

Ce mécanisme causerait essentiellement des duplications et pertes de segments d'ADN, mais pourrait être également impliqué dans certains réarrangements non récurrents en juxtaposant des segments d'ADN qui ne sont pas à l'origine proches l'un de l'autre sur le chromosome. Contrairement à la NAHR et au NHEJ, le FoSTeS n'est pas lié à l'existence d'une cassure double-brin : il s'agit d'un mécanisme dépendant des erreurs du processus de réplication. Il pourrait être dû dans certains cas à l'existence de palindromes formant des structures cruciformes dans l'ADN (Gu et al. 2008; Carvalho et al. 2009).

3.2.4. Les réarrangements : un résultat, plusieurs causes ?

Il est probable que les réarrangements évolutifs sont les conséquences non pas d'un seul mécanisme, mais d'une mosaïque d'erreurs variées dans différentes voies de maintenance et de traitement de l'ADN. Les contributions relatives de l'un ou l'autre de ces mécanismes sont encore largement débattues : en effet, nombre d'études se sont focalisées sur les réarrangements se produisant de manière récurrente dans la même région du génome et ayant des conséquences délétères chez l'homme (Stankiewicz and Lupski 2002; Liu et al. 2012). Cette double spécificité en fait de bons sujets d'étude d'un point de vue médical, mais introduit des limites quand il s'agit de rendre compte du mécanisme général d'apparition des réarrangements. La littérature portant sur les mécanismes de réarrangement est donc largement biaisée vers les réarrangements récurrents et déséquilibrés causés par la NAHR, et ne rend pas forcément compte de l'importance réelle de chaque type de mécanisme dans l'apparition de nouveaux événements dans la population générale. Avec la démocratisation des techniques d'analyse du génome et de séquençage à grande échelle, plusieurs études ont récemment cherché à caractériser et quantifier les points de cassure de réarrangements dans un grand nombre de génomes : plusieurs analyses rapportent que les réarrangements équilibrés sont les plus fréquents et portent généralement la marque du NHEJ (Korbel et al. 2007; Lam et al. 2010; Chiang et al. 2012), mais ce point est encore débattu, d'autres études attribuant une large part des réarrangements équilibrés à la NAHR (Kidd et al. 2010). De manière générale, les cas de réarrangements formellement identifiés comme résultant du FoSTeS sont assez rares dans toutes les études.

La multiplicité des mécanismes a une conséquence peu abordée dans la littérature : il est difficile d'avoir un réel attendu *a priori* sur la probabilité de réarrangement à un endroit donné du génome. En effet, les probabilités de cassure et d'erreur dans les différentes voies dépendent de trop de paramètres mal connus pour être modélisées à ce jour, et l'hypothèse d'une probabilité uniforme sur l'ensemble du génome est peu vraisemblable d'un point de vue biologique. La NAHR et le NHEJ sont probablement plus fréquents dans les zones du génome où le taux de cassures double-brins est élevé ; en revanche, le FoSTeS correspond à une erreur lors de la réplication et est donc potentiellement sous l'influence de paramètres différents, plus dépendants de la séquence et de la structure de la double-hélice de l'ADN (pour revue, voir (Gu et al. 2008)).

3.3. Conséquences biologiques des réarrangements

Comme abordé précédemment, les réarrangements déséquilibrés ont fréquemment des conséquences délétères sur le phénotype parce qu'ils modifient la stœchiométrie des taux d'expression des gènes, ce qui résulte en des syndromes de trisomie. Les réarrangements déséquilibrés sont rares dans les populations ; mais même les réarrangements équilibrés ont également des conséquences à différents stades de la vie de la cellule et de l'organisme.

3.3.1. Effets délétères sur l'expression des gènes

Les réarrangements chromosomiques peuvent avoir deux types d'effets délétères sur l'expression des gènes. Un point de cassure dans la séquence codante d'un gène cause généralement la production de transcrits aberrants et annihile l'expression du gène. De tels réarrangements sont impliqués dans de plusieurs désordres génétiques chez l'homme, où la perte d'une protéine importante ou la formation d'une nouvelle protéine par chimérisation de deux séquences a des conséquences sur la physiologie de l'organisme (par exemple, la translocation t(9 ;22)(q34 ;q11) responsable de la leucémie myéloïde chronique (Prakash and Yunis 1984)). Par ailleurs, de nombreux syndromes sont causés par des réarrangements qui ne perturbent la séquence d'aucun gène connu. Ainsi, Baptista et al. (2008) rapportent que parmi une cohorte d'individus présentant un réarrangement génomique responsable d'un phénotype observable, les points de cassure sont identifiés dans une région ne contenant aucun gène annoté dans 33% des cas (et dans 46% des cas, la résolution n'est pas suffisante pour dire avec certitude si les points de cassure affectent une séquence codante). Ces cas suggèrent que dans certaines régions du génome, l'organisation des gènes est elle-même nécessaire à l'expression correcte de l'information génétique. En effet, de nombreux gènes sont régulés en *cis* par des éléments non-codants qui contrôlent leur patron d'expression temporel et spatial (Pennacchio et al. 2006). Les réarrangements causant ainsi la séparation des gènes et de leur environnement de régulation, ou mettant en contact un gène avec des éléments de régulation qui vont modifier son patron d'expression, peuvent avoir des effets délétères graves. La séquence de Pierre Robin est un exemple de syndrome qui peut être causé, entre autres mutations, par des translocations séparant une séquence de régulation clé de son gène cible, le gène du développement *Sox9* situé à plus d'1 Mb de distance (Benko et al. 2009).

Les réarrangements délétères sont largement étudiés et très présents dans la littérature, où les études de désordres génomiques abondent en raison de leur intérêt médical d'une part, et des informations qu'elles apportent sur la fonction des différentes séquences codantes et non-codantes d'autre part. Cette surreprésentation par rapport aux réarrangements sans phénotype détectable a pu donner l'impression que la majorité des réarrangements a des conséquences délétères pour l'organisme. Or, peu d'études se sont intéressées aux réarrangements présents dans des génomes sains jusqu'à une période récente. On estime aujourd'hui qu'environ 94% des réarrangements ne causent aucun phénotype particulier (Warburton 1991; Baptista et al. 2008). De plus, les réarrangements liés à un phénotype sont généralement non équilibrés, c'est-à-dire qu'ils sont couplés à des délétions ou duplications de séquence au niveau des points de cassure

(Baptista et al. 2008). Il est probable que le phénotype soit alors lié à un déséquilibre dans les taux d'expressions de gènes, et non à la réorganisation des gènes en soi. Ces résultats récents apportent un éclairage nouveau sur cette question et révèlent que les réarrangements sans phénotype sont en réalité fréquents et que le génome est plus plastique que considéré jusqu'alors.

Les réarrangements évolutifs qui scindent un gène en deux ou réorganisent une séquence codante sont rares. Les événements de cassure identifiés au sein d'une séquence codante représentent de l'ordre de 3 à 9% des points de cassure selon les études et les espèces considérées (Lemaitre et al. 2009; Mongin et al. 2009), et une partie de ces points de cassure intragéniques sont probablement des artéfacts dus à des erreurs d'annotation des bornes de gènes ou à des erreurs d'assemblage de la séquence. Il est probable que ces cassures ont presque toujours à terme des conséquences négatives, et sont éliminées par sélection. Par conséquent, les points de cassures des réarrangements évolutifs sont considérés comme se produisant quasi-exclusivement dans les espaces intergéniques (Peng et al. 2006). C'est le point de vue qui sera adopté dans ce manuscrit, où nous nous focaliserons sur les caractéristiques des espaces intergéniques qui sous-tendent l'apparition de points de cassures de réarrangements en occultant le cas particulier des points de cassure intragéniques.

3.3.2. Suppression de la recombinaison et gamètes non équilibrés

Même lorsqu'ils ne sont pas délétères pour l'expression des gènes, les réarrangements chromosomiques ont des conséquences directes au moment de la méiose. En premier lieu, les deux chromosomes d'une paire n'étant en partie plus homologues, la recombinaison est essentiellement supprimée dans cette région chromosomique chez les hétérozygotes (Hoffmann and Rieseberg 2008). Le mécanisme de suppression de la recombinaison est mal connu, mais est probablement dû au fait que les séquences non homologues ne peuvent pas former de synapses et de chiasmas, nécessaires à la recombinaison méiotique.

Dans le cas d'une inversion relativement courte, la région réarrangée est simplement exclue du brassage génétique sans causer de conséquences graves. En revanche, si le réarrangement est de grande taille (inversion péricentromérique, etc.) ou implique plusieurs chromosomes, la méiose peut donner naissance à des gamètes non équilibrés (Figure 3.6). De plus, lorsque le réarrangement met en jeu une région suffisamment longue, la recombinaison peut également s'établir de façon anormale entre les différentes régions des chromosomes homologues, et résulter après résolution des crossovers en un chromosome sans centromère et un autre avec deux centromères, ce qui perturbe la disjonction des chromosomes. Ainsi, même lorsque les réarrangements sont équilibrés et sans phénotype à la génération où ils apparaissent, ils peuvent causer des aneuploïdies délétères à la génération suivante.

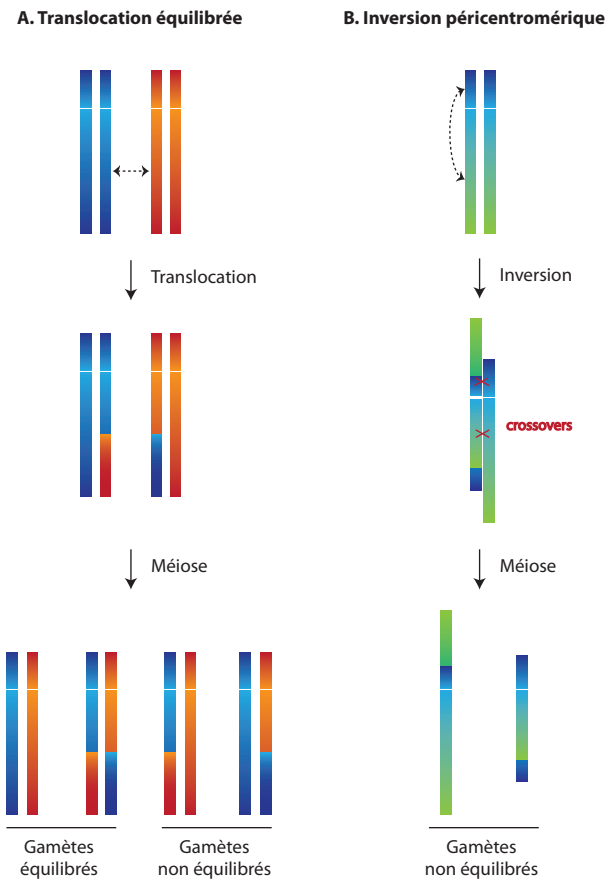


Figure 3.6. Conséquences de deux réarrangements équilibrés (A : translocation, B : inversion péricentrique) sur les gamètes produits à la méiose. Dans les deux cas, une partie des gamètes au moins est non équilibrée.

3.3.3. Spéciation

Depuis plus de vingt ans, les réarrangements ont été proposés comme l'un des mécanismes évolutifs majeurs favorisant la spéciation parallèlement au modèle classique de Bateson-Dobzhansky-Muller basé sur les incompatibilités d'allèles (pour revue, voir (Faria and Navarro 2010)). En empêchant la recombinaison dans les régions réarrangées chez les hétérozygotes, les réarrangements créent une barrière qui pourrait faciliter la divergence de la région réarrangée entre deux populations en limitant le flux génétique entre elles (Trickett and Butlin 1994). Ainsi, il a été démontré chez les drosophiles que les inversions fixées entre espèces portent fréquemment des gènes impliqués dans la stérilité hybride des mâles et le succès reproductif (Noor et al. 2001; Machado et al. 2002), suggérant que ces réarrangements, bien que n'influençant pas la physiologie de l'organisme lui-même, ont favorisé l'isolation des populations en réduisant la fréquence des croisements (isolation prézygotique) ou la fitness des hybrides (isolation postzygotique)(Navarro and Barton 2003a). De manière plus générale, plusieurs cas de réarrangements facilitant l'accumulation de gènes impliqués dans des adaptations locales ou dans l'isolation reproductive ont été décrits chez les plantes (Lai et al. 2005; Lowry and Willis 2010) ou les insectes (Ayala and Coluzzi 2005).

Chez les vertébrés en revanche, les données sur l'influence des réarrangements dans la spéciation sont plus floues. Certains éléments indirects initialement apportés en faveur d'un rôle des réarrangements dans la spéciation, comme l'observation d'une plus grande divergence des gènes dans les chromosomes ayant subi de larges réarrangements entre l'homme et le chimpanzé (Navarro and Barton 2003b), n'ont pas résisté aux données plus récentes obtenues à partir des séquences de génomes complets (Zhang et al. 2004; Marques-Bonet et al. 2007). Les études récentes menées chez la musaraigne (*Sorex araneus*) et la souris (*Mus musculus*), deux espèces présentant des polymorphismes caryotypiques entre populations, ont obtenu des résultats mitigés et contradictoires quant à l'existence de barrières aux flux de gènes dues aux réarrangements entre les différentes populations (Yannic et al. 2009; Franchini et al. 2010; Horn et al. 2012). On peut cependant noter qu'une corrélation a été relevée entre le taux de réarrangement chromosomique et le taux de spéciation dans certains groupes taxonomiques, comme les gibbons, qui se caractérisent par un grand nombre d'espèces et une grande variabilité caryotypique (Carbone et al. 2009; Girirajan et al. 2009); cependant, la validité à grande échelle de cette observation n'a, à notre connaissance, pas été testée.

3.4. Identification des points de cassure

Du fait de leur intérêt pour comprendre l'évolution des génomes et l'importance fonctionnelle de leur organisation, les différences caryotypiques entre espèces ont fait l'objet de nombreuses études de génomique comparative. Au fil des avancées technologiques, il est devenu possible de détecter un nombre croissant d'événements de réarrangements à une résolution de plus en plus fine, et d'identifier leurs points de cassure avec une précision également croissante. Nous détaillons ici les méthodes utilisées historiquement pour détecter les points de cassure de réarrangements évolutifs dans les génomes de vertébrés.

3.4.1. Cytogénétique

Les premières identifications de réarrangements datent des années 1970 et sont basées sur des observations directes de la morphologie des chromosomes en métaphase. La méthode la plus simple se base sur la comparaison des alternances de bandes G et R après coloration des chromosomes au Giemsa. Le patron des bandes chromosomiques est suffisamment conservé entre espèces proches pour détecter des régions inversées, transloquées ou des fusions et fissions de chromosomes. Les modifications du caryotype ainsi détectées ont été utilisées pour tenter de reconstruire l'histoire évolutive de certains phylums à des fins phylogénétiques, par exemple chez les primates (Dutrillaux 1979), les félins (Wurster-Hill and Gray 1975) ou les carnivores (Dutrillaux and Couturier 1983). Pour des espèces plus éloignées, comme l'homme et la souris, la conservation du patron des bandes chromosomiques n'est pas suffisante à l'échelle du génome entier pour décrire précisément l'évolution du caryotype ; cependant, on peut tout de même détecter des régions où ce patron est conservé, ce qui a suggéré que de larges régions des génomes de mammifères ont conservé leur organisation ancestrale sur des dizaines de millions d'années (Nash and O'Brien 1982; Dutrillaux and Couturier 1983; Sawyer and Hozier 1986).

A partir des années 1990, les techniques d'hybridation fluorescente *in situ* (FISH) ont permis de comparer l'architecture des chromosomes et de détecter des points de cassure de réarrangements entre espèces plus éloignées sans les limites de reconnaissance visuelle posées par les bandes chromosomiques. Les différentes techniques de FISH reposent sur l'utilisation de sondes ADN couplées à des fluorophores que l'on hybride à un génome afin de mettre en évidence la localisation de séquences d'intérêt dans les chromosomes. La méthode la plus utilisée en génomique comparative est le Zoo-FISH, ou « chromosome painting » (Jauch et al. 1992; Scherthan et al. 1994) : des sondes fluorescentes sont générées par PCR dégénérée à partir de différents chromosomes d'une espèce de référence, et les chromosomes en métaphase d'une espèce cible sont « coloriés » par hybridation des sondes afin de mettre en évidence les blocs de gènes ayant la même origine ancestrale dans les deux espèces (Figure 3.7). Cette technique a permis de mettre en évidence l'existence de grands blocs, voire de chromosomes entiers dont le contenu génétique est conservé entre différentes espèces : elle a été utilisée pour reconstruire l'histoire des chromosomes dans de nombreux phylums, notamment les primates (Muller et al. 1999; Muller and Wienberg 2001; Stanyon et al. 2001) et de manière plus générale les mammifères (Wienberg and Stanyon 1995; Chowdhary et al. 1996; Fronicke et al. 1996; Raudsepp et al. 1996; Wienberg and Stanyon 1997; Chowdhary et al. 1998; Grutzner et al. 1999; Fronicke et al. 2003), en prenant généralement le génome humain comme référence.

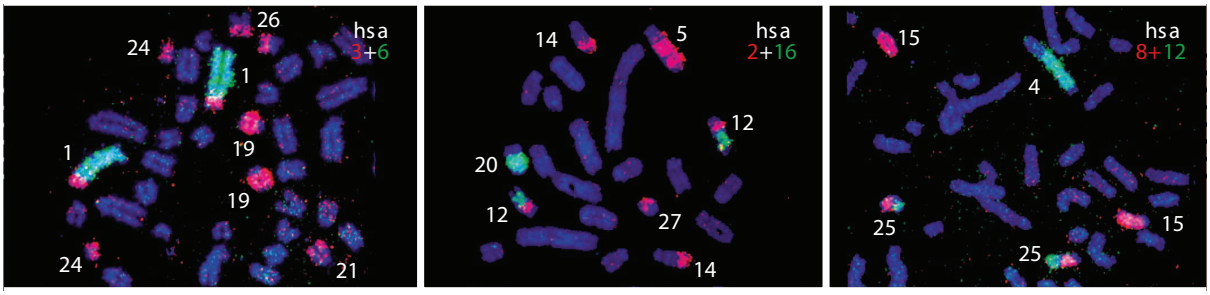


Figure 3.7. Exemple d'hybridations par Zoo-FISH entre le génome de l'éléphant d'Afrique et le génome humain. Les chromosomes de l'éléphant sont colorés au DAPI ; les sondes obtenues à partir de différents chromosomes humains sont indiquées avec la couleur correspondante en haut à droite de chaque image. Pour exemple, le chromosome 1 de l'éléphant est homologue au chromosome 6 et à une partie du chromosome 3 humains (premier panneau). Figure tirée de (Fronicke et al. 2003).

L'intérêt majeur du Zoo-FISH est de mettre en évidence les grands réarrangements ayant affecté l'évolution d'un groupe d'espèces de manière rapide et peu coûteuse en ressources. Son principal défaut est qu'il ne peut mettre en évidence que les réarrangements interchromosomiques, et pas les modifications de l'ordre des gènes qui se produisent au sein d'un bloc de gènes conservés entre les deux espèces. Pour palier à cette difficulté, le Zoo-FISH a été adapté en utilisant des jeux de sondes correspondant non plus à un chromosome entier mais à des régions sub-chromosomiques, avec une résolution allant jusqu'à quelques Mb (Muller et al. 2000). La méthode devient alors plus résolutive mais est nécessairement limitée à un nombre de régions génomiques restreint pour des raisons de faisabilité. Par ailleurs, la méthode suppose que les séquences des génomes comparés soient suffisamment proches pour que les sondes générées à partir d'un génome s'hybrident correctement sur l'autre : ainsi, le Zoo-FISH fonctionne à l'échelle des mammifères et des euthériens en général, mais ne permet pas de

comparer les génomes de marsupiaux et ceux des euthériens, trop divergents sauf au niveau du chromosome X (Glas et al. 1999; Wienberg 2004). Avec la diminution des coûts de séquençage de génomes, qui permettent d'étudier l'architecture des génomes avec une résolution bien plus fine, le Zoo-FISH et les approches cytogénétiques en général ont tendance à tomber en désuétude pour l'étude des réarrangements chromosomiques.

3.4.2. Comparaison de l'ordre de marqueurs dans les génomes

Avant même la mise à disposition de la communauté scientifique des séquences complètes de génomes de différentes espèces, la comparaison de l'ordre des gènes orthologues a été une méthode de choix pour détecter des réarrangements et leurs points de cassure entre différents génomes. Les analyses de fréquence de recombinaison méiotique ont été utilisées pour calculer les distances génétiques entre plusieurs centaines de gènes marqueurs sur les chromosomes de différentes espèces, afin d'établir des cartes génétiques permettant de comparer les génomes entre eux (Lalley et al. 1978; O'Brien et al. 1993). A partir de ces cartes génétiques, il est possible d'explorer la conservation de l'ordre des marqueurs orthologues entre les espèces (Nadeau and Taylor 1984; Eppig and Nadeau 1995; O'Brien et al. 1999; Murphy et al. 2005; Carbone et al. 2009). Toute différence dans l'ordre de ces marqueurs est caractéristique d'une région de cassure. Jugées au départ plus laborieuses et moins efficaces que le Zoo-FISH, les cartes génétiques lui ont pourtant survécu d'une part à cause de leur résolution de plus en plus fine et de leur capacité à rendre compte des réarrangements intrachromosomiques autant qu'interchromosomiques, et d'autre part grâce à leur capacité à comparer plusieurs génomes en parallèle avec un jeu de marqueurs communs. Ces cartes ont pavé la voie aux analyses utilisant les génomes complets et, si elles sont progressivement remplacées par les génomes complets eux-mêmes pour étudier les points de cassure, elles restent très utiles pour étudier l'architecture générale des chromosomes des espèces non séquencées, et ancrer et ordonner les fragments de séquence (scaffolds) sur les chromosomes lors de l'assemblage des génomes.

Avec les génomes séquencés et assemblés à des degrés divers mis à disposition de la communauté, l'analyse des réarrangements a pris un essor nouveau depuis le début des années 2000. Ces analyses sont la continuité logique des comparaisons de cartes génétiques, dont elles reproduisent le principe, mais avec un nombre de marqueurs orthologues beaucoup plus grand (quelques centaines pour les cartes génétiques, plusieurs milliers entre deux génomes de vertébrés). La comparaison des génomes et l'identification des points de cassure entre plusieurs espèces se base sur l'alignement de leurs séquences, afin d'obtenir un jeu de marqueurs orthologues dans les génomes. Ces marqueurs peuvent être des éléments fonctionnels identifiés comme orthologues comme des gènes ou des séquences conservées non-codantes (Bourque et al. 2005; Gordon et al. 2007; Kemkemer et al. 2009; Mongin et al. 2009; Baudet et al. 2010), ou être des blocs de séquence directement tirés de l'alignement multiple des génomes et contenant de nombreux éléments fonctionnels dont l'organisation est globalement conservée (Pevzner and Tesler 2003a; Bailey et al. 2004; Bourque et al. 2004; Zhao et al. 2004; Bourque et al. 2005; Ovcharenko et al. 2005; Ma et al. 2006; Peng et al. 2006; Girirajan et al. 2009; Zhao and Bourque 2009; Volker et al. 2010). Les génomes à comparer sont alors considérés comme une suite de marqueurs discrets et généralement orientés, dont l'ordre peut varier : toute région entre deux

marqueurs qui sont adjacents (directement l'un à côté de l'autre) dans un génome et pas dans un autre est une région de cassure (Figure 3.8). La comparaison de plusieurs génomes permet généralement de déduire l'état ancestral et la lignée dans laquelle le réarrangement a eu lieu.

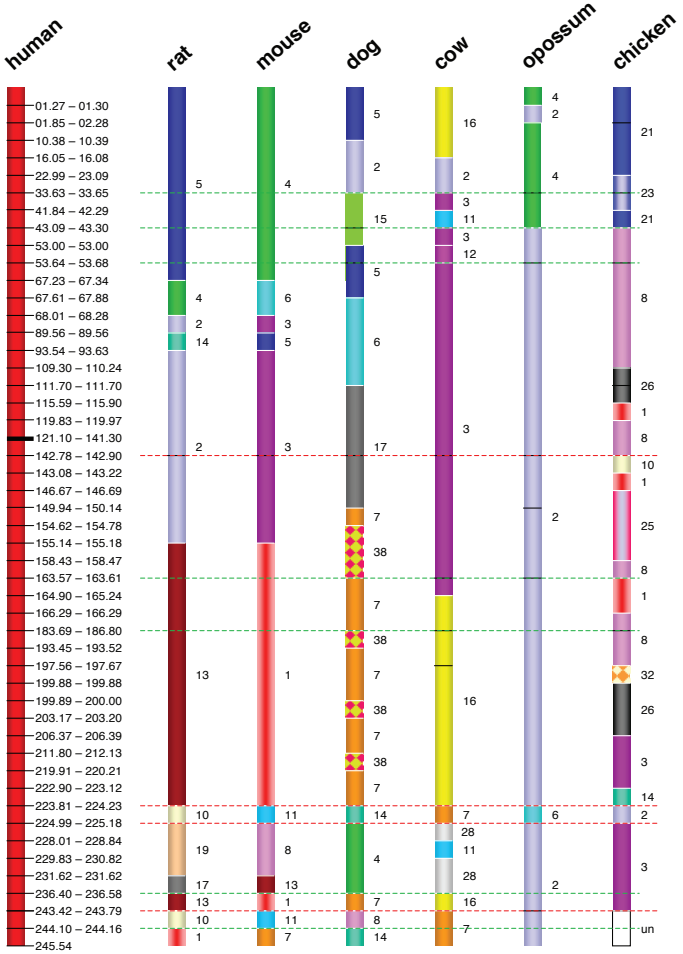


Figure 3.8. Régions de synténie conservée détectées entre le chromosome 1 de l'homme et les chromosomes de six autres mammifères. Chaque repère sur le chromosome humain correspond à un point de cassure dans l'une de lignées, avec ses coordonnées sur le chromosome humain (en Mb) ; les repères sont placés à intervalles réguliers pour des questions de lisibilité, mais la distance d'un repère à l'autre sur le schéma ne reflète pas leur distance réelle sur le chromosome. Les points de cassure dans la lignée humaine sont marqués par des pointillés rouges. Les régions de cassure réutilisées indépendamment dans différentes lignées sont figurées par des pointillés verts. Figure tirée de (Kemkemer et al. 2009).

Le niveau de résolution apporté par les séquences génomiques est beaucoup plus fin que celui des études cytogénétiques, ce qui est attendu étant donné le grand nombre de marqueurs considéré. A titre d'exemple, les études cytogénétiques avaient détecté 9 inversions de grande taille entre les génomes de l'homme et du chimpanzé, alors que la comparaison de leurs séquences en révèle environ 1500 de toutes tailles (Feuk et al. 2005). La limite principale des comparaisons de génomes pour détecter les points de cassure est qu'elles sont extrêmement dépendantes de la qualité de l'assemblage des génomes, ainsi que de leur annotation et de l'identification des orthologues entre espèces dans le cas où les marqueurs utilisés sont les gènes orthologues. Comme la méthode se base sur la détection de marqueurs dont l'ordre ou

l'orientation est modifiée, toute erreur à ces étapes sera détectée comme un réarrangement si elle modifie l'ordre apparent des marqueurs. Ainsi, la majorité des études notamment basées sur les blocs de séquence alignées s'accordent à ne considérer comme marqueurs que les régions alignées dépassant une certaine taille (quelques dizaines de kb à plusieurs Mb, selon les études, en passant outre les petites régions non alignables au sein du bloc ; Figure 3.9). Ce compromis nécessaire a pour conséquence de faire baisser la sensibilité de ces méthodes, en excluant d'office la détection de réarrangements de blocs de taille inférieure à la limite minimale autorisée.

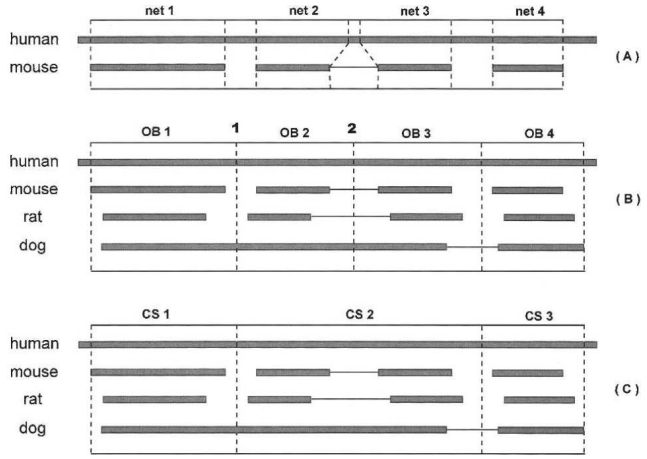


Figure 3.9. Définition de marqueurs conservés à partir d'un alignement de séquences, comme proposé par Ma 2006. (A) Les génomes sont alignés individuellement avec une espèce de référence (ici, le génome humain) pour définir des blocs de séquence alignée (« nets »). Un trait épais figure un bloc de séquence alignée d'une longueur supérieure au seuil minimal fixé ; un trait fin figure une partie de la séquence qui n'est pas alignée mais qui se trouve entre deux blocs alignés dans la même orientation. (B) Les différentes espèces sont regroupées pour définir des blocs d'orthologie (OB), c'est-à-dire des blocs dans les différentes espèces qui s'alignent sur la même région du genome de référence. Ces blocs sont séparés par des pointillés sur la figure. (C) Les blocs d'orthologie sont fusionnés en segments conservés (CS) s'ils sont dans le même ordre et la même orientation dans toutes les espèces (c'est le cas des OB2 et OB3), même si une partie de la séquence entre eux n'est pas alignable. Figure tirée de (Ma et al. 2006).

3.5. Méthodes d'étude des régions de cassures évolutives

Aujourd'hui, la grande majorité des analyses portant sur les points de cassure de réarrangement évolutifs se basent sur les cassures détectées à partir des alignements de génomes. Plusieurs types d'approches ont été proposées pour comprendre où, comment et pourquoi les réarrangements se produisent dans les génomes de vertébrés : chacune aborde le problème par un angle différent, et nous présentons ici les contributions et conclusions des trois grands axes de recherche sur les points de cassure évolutifs.

3.5.1. Distribution des points de cassure dans les génomes

La distribution des points de cassure dans les génomes a été abordée par Nadeau et Taylor (1984) à partir de la comparaison des cartes génétiques de l'homme et de la souris. Cet article, considéré comme l'une des pierres angulaires du domaine, montrait qu'à cette résolution les

deux génomes partagent environ 180 segments conservés (ordre des marqueurs similaire). La longueur de ces segments conservés suit alors une distribution exponentielle négative compatible avec celle attendue si les points de cassure de réarrangements étaient distribués aléatoirement dans les génomes. Plusieurs études postérieures portant sur des jeux de données plus larges ont conforté ce point par la suite (Copeland et al. 1993; O'Brien et al. 1999; Lander et al. 2001; Mural et al. 2002). Ce résultat a cependant été remis en question par les comparaisons à haute résolution des génomes complets, qui ont révélé de nombreux points de cassure en dessous de la résolution des cartes génétiques utilisées auparavant. Les analyses de longueur des segments conservés basées sur les génomes complets ainsi que les approches combinatoires (voir ci-dessous au paragraphe 3.5.3) ont toutes deux, de manière indépendante, mis en évidence un excès de petits blocs de synténie qui ne peuvent s'expliquer que si les points de cassure se produisent de manière regroupée dans certaines régions du génome (Figure 3.10) (Kent et al. 2003; Pevzner and Tesler 2003b; Bourque et al. 2004; Zhao et al. 2004; Webber and Ponting 2005). Par ailleurs, les analyses des points de cassure dans plusieurs lignées différentes ont montré que les cassures tendent à se reproduire indépendamment dans les mêmes régions du génome plus souvent qu'attendu au hasard, menant à la notion de récurrence des points de cassure et de l'existence de « points chauds » de cassure évolutivement conservés (Pevzner and Tesler 2003b; Murphy et al. 2005; Hinsch and Hannenhalli 2006; Gordon et al. 2007; Kemkemer et al. 2009).

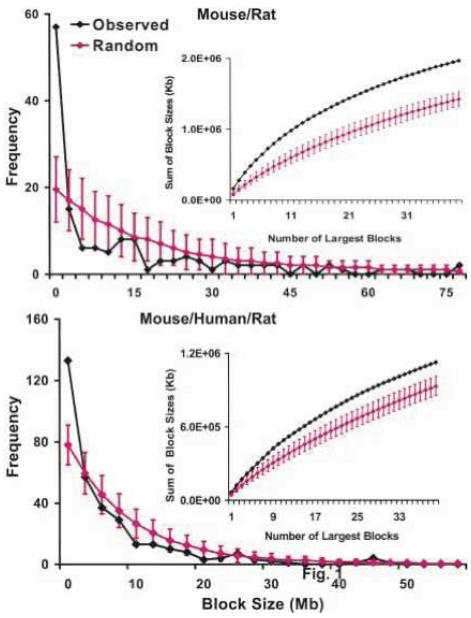


Figure 3.10. Distribution des longueurs des blocs de synténie conservée entre les génomes de la souris et du rat (en haut), et de la souris, du rat et de l'homme (en bas). Les nombres de blocs de longueur supérieure à 100 kb sont figurés en noir, et comparés à la distribution attendue sous le modèle aléatoire en rose (avec l'intervalle de confiance à 95%). En encart est représentée la longueur cumulée des blocs de synténie les plus longs, également comparée à l'attendu sous le modèle aléatoire. Figure tirée de (Zhao et al. 2004).

3.5.2. Analyses statistiques des caractéristiques des points de cassure

Parallèlement à la distribution des blocs conservés entre les génomes, un grand nombre d'études s'est attaché à explorer la composition des régions de cassure, afin de tester si les points de cassure se produisent préférentiellement dans des régions biaisées par rapport à l'ensemble du génome. Le principe général est de comparer les caractéristiques d'une fenêtre génomique autour d'un point de cassure identifié avec soit des fenêtres sans points de cassure (régions stables), soit des fenêtres échantillonnées aléatoirement dans le génome. L'objectif est de mettre en évidence des caractéristiques statistiquement associées aux cassures (ou aux régions stables, le cas échéant). Les méthodes exactes et les paramètres testés variant d'une analyse à l'autre, il est difficile de comparer les résultats obtenus ; cependant, plusieurs caractéristiques ressortent de manière plus ou moins récurrente comme étant associées aux points de cassure. Ces caractéristiques sont synthétisées dans la Table 3.1 (Armengol et al. 2003; Bailey et al. 2004; Murphy et al. 2005; Ovcharenko et al. 2005; Ma et al. 2006; Schibler et al. 2006; Gordon et al. 2007; Kikuta et al. 2007; Carbone et al. 2009; Hufton et al. 2009; Kemkemer et al. 2009; Lemaitre et al. 2009; Mongin et al. 2009; Zhao and Bourque 2009; Volker et al. 2010; Skinner and Griffin 2012).

	Régions de cassure	**Régions stables**
Taux de GC	GC-riches[5,13]	AT-riches[5,13]
Densité en gènes	Elevée[3,5,7,11,12,13]	Faible[3,5,7,11,12,13,14]
Densité en séquences répétées	Elevée[4,5,6,9,17]	Faible[4,5,6,9,17]
dont SINEs	Elevée[5,6,9,17]	Faible[5,6,9,17]
dont LINEs	Faible[5] ou Elevée[6,7,17] selon les études. Présence de paires d'éléments aux bornes de certains réarrangements[15]	Faible[6,7,17] ou Elevée[5] selon les études
Densité en îlots CpG	Elevée[7,9,12]	Faible[7,9,12]
Duplications segmentales	Plus nombreuses[1,2,3,5,7,9,11,15] Mais pourraient être causées par les réarrangements, non l'inverse[2,9]	Plus rares[1,2,5,7,9,11,15]
Densité en élements conservés non-codants	Faible[8,10,12,14]	Elevée[8,10,12,14]
CNV (Copy Number variants)	Plus nombreux[11,12,14,16]	Plus rares[11,12,14,16]
SNP (Single nucleotide polymorphism)	Plus nombreux[12]	-
Sites de fragilité	Associées[14] ou non[11] selon les études	Non associées[11]
Sites de cassures récurrentes dans les lignées cancéreuses	Associées[3] ou non[11] selon les études	Non associées[3]
Taux de recombinaison	Faible[12] ou Elevé[16] selon les études	Elevé[12] ou Faible[16] selon les études
Fonctions des gènes	Enrichies en facteurs à doigts de zinc[12], en gènes impliqués dans l'immunité[14]	Enrichies en gènes du développement[8,14], en gènes tissus-spécifiques[14]

Table 3.1. Caractéristiques des points de cassure de réarrangements évolutifs relevées dans la littérature. Les numéros en exposants font référence aux travaux suivants : [1] (Armengol et al. 2003) ; [2] (Bailey et al. 2004) ; [3] (Murphy et al. 2005) ; [4] (Ovcharenko et al. 2005) ; [5] (Ma et al. 2006) ; [6] (Schibler et al. 2006) ; [7] (Gordon et al. 2007) ; [8] (Kikuta et al. 2007) ; [9] (Carbone et al. 2009) ; [10] (Hufton et al. 2009) ; [11] (Kemkemer et al. 2009) ; [12] (Lemaitre et al. 2009) ; [13] (Mongin et al. 2009) ; [14] (Zhao and Bourque 2009) ; [15] (Volker et al. 2010) ; [16] (Skinner and Griffin 2012).

En fonction des caractéristiques génomiques testées et des résultats obtenus, nombre d'interprétations différentes ont été proposées au fil des analyses, rendant telle ou telle caractéristique responsable des cassures chromosomiques. La difficulté majeure rencontrée par ces analyses est que la plupart des paramètres statistiquement associés aux points de cassure sont également intercorrélés entre eux : par exemple, les génomes de vertébrés sont structurés en isochores (Bernardi 2000), et les régions denses en gènes sont statistiquement corrélées à des taux de GC élevés. La densité d'îlots CpG est également connue pour augmenter avec la densité en gènes (Lander et al. 2001), alors que le taux de recombinaison et la densité en SINEs sont plus élevés dans les régions riches en GC (Lander et al. 2001; Kong et al. 2002) ; à l'inverse, les LINEs sont plutôt associés aux régions pauvres en gènes et riches en AT. Les mécanismes qui sous-tendent cette structuration des génomes sont mal connus : il est probable qu'elle découle d'un équilibre qui s'établit entre divers processus (taux de mutations biaisés, taux de recombinaison, taux d'insertions, etc.)(Petrov et al. 2000; Galtier and Duret 2007; Nam and Ellegren 2012). Comme toutes les propriétés génomiques associées aux points de cassure sont liées entre elles, il est particulièrement difficile de mettre en évidence laquelle ou lesquelles sont réellement causatives des cassures, et lesquelles sont de simples corrélations secondaires. L'interprétation de ces résultats est donc en grande part spéculative, et se fait à la lumière des mécanismes moléculaires plausibles proposés pour les réarrangements qui, on l'a vu au paragraphe 3.2, sont eux-mêmes débattus.

L'interprétation la plus répandue dans la littérature est que les réarrangements évolutifs sont majoritairement dus à des événements de recombinaison non homologue (NAHR) au cours de la méiose dans la lignée germinale, causés par les duplications segmentales ou les éléments transposables, qui sont surreprésentés autour des points de cassure (Table 3.1). Cette hypothèse, séduisante au premier abord, n'est pourtant pas satisfaisante à plusieurs points de vue : premièrement, comme abordé au paragraphe 3.2, les études des séquences flanquant les points de cassure ont montré que la NAHR ne semble impliquée que dans une partie minoritaire des réarrangements, qu'il s'agisse des événements se produisant dans le génomes modernes ou des événements évolutifs. Ensuite, la corrélation statistique entre ces éléments et les cassures ne signifie pas qu'on observe forcément des paires de duplications segmentales ou d'éléments transposables identiques précisément aux bornes des réarrangements (Feuk et al. 2005); et lorsque des duplications sont effectivement trouvées de part et d'autre d'un réarrangement, ces duplications sont souvent spécifiques à la lignée où le réarrangement a eu lieu, et peuvent tout autant s'expliquer comme des conséquences que des causes des cassures (Ranz et al. 2007; Carbone et al. 2009; Girirajan et al. 2009). Par ailleurs, les cassures survenues dans une lignée sont également corrélées statistiquement aux duplications segmentales survenues indépendamment dans une autre lignée (Bailey et al. 2004; Marques-Bonet et al. 2007). Ces résultats suggèrent que réarrangements et duplications segmentales sont deux résultats (non-exclusifs) des mêmes mécanismes d'erreurs (NAHR et/ou NHEJ), qui eux-mêmes se produisent plus souvent dans certaines régions instables du génome. Ces régions instables pourraient correspondre aux régions où la transcription est élevée (Carbone et al. 2009; Lemaitre et al. 2009), au niveau des origines de réplication (Lemaitre et al. 2009), ou encore dans les régions de forte recombinaison méiotique (Volker et al. 2010). Les régions de cassure évolutives correspondraient à des domaines plastiques du génome où peuvent se concentrer divers

événements évolutifs se produisant de manière indépendante (cassures, duplications, insertions d'éléments transposables, etc.) sans avoir de conséquences délétères.

3.5.3. Approches combinatoires

Il existe enfin un troisième type d'analyse des points de cassure de réarrangements, plus théorique cette fois. A partir de la comparaison de leurs séquences, deux génomes peuvent en effet être considérés comme une succession de blocs identiques mais dont l'ordre ou l'orientation a été modifié, suivant un nombre limité d'événements possibles : inversions, translocations, fissions et fusions de blocs. Il est donc possible, en théorie, de reconstruire la combinaison d'événements optimale qui permet de passer d'un génome à un autre. On ne s'intéresse donc pas directement ici aux caractéristiques des points de cassure, mais bien au scénario évolutif le plus parcimonieux. Plusieurs algorithmes ont été développés pour permettre de calculer le nombre d'événements nécessaires pour transformer un génome en un autre, appelé « distance génomique ». Ces algorithmes utilisent des blocs conservés déterminés avec une résolution plus ou moins précise, sachant que le nombre de blocs utilisés impacte fortement à la fois le temps de calcul et le nombre de solutions optimales possibles au problème. GRIMM (*Genome Rearrangements In Man and Mouse*)(Tesler 2002), le premier programme développé, comme son nom l'indique, pour s'attaquer au problème de la comparaison du génome de l'homme et de la souris, a apporté un éclairage inattendu sur les points de cassure dans un article célèbre de Pevzner et Tesler (2003) : la seule solution pour transformer le génome de l'homme en celui de la souris est d'inférer que certains points de cassure seront réutilisés plusieurs fois, avec un taux de réutilisation moyen de 1,9 (Figure 3.11).

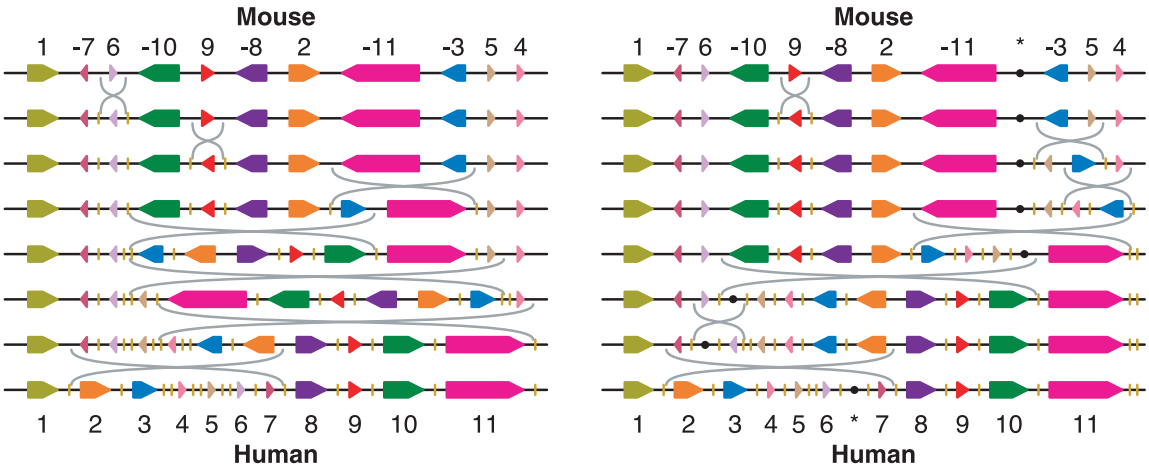


Figure 3.11. Deux scénarios d'évolution possibles du chromosome X entre le génome humain et celui de la souris, aussi parcimonieux l'un que l'autre. Les chromosomes X de l'homme et de la souris sont constitués de 11 blocs de synténie conservée. Les traits jaunes figurent des points de cassure entre deux blocs de synténie. Le point noir dans le deuxième scénario représente un éventuel bloc de synténie non détecté à la résolution utilisée par l'étude. Tout scénario évolutif basé sur ces 11 blocs de synténie conservée implique au moins deux ou trois cas de points de cassure réutilisés, en incluant les extrémités de chromosomes. Figure tirée de (Pevzner and Tesler 2003b).

Les auteurs en ont déduit que de nombreuses régions de cassure considérées contiennent non pas un mais plusieurs points de cassure, entre lesquels ils ont prédit l'existence de blocs conservés mais trop courts pour être détectés à la résolution utilisée dans l'étude (taille minimum des blocs : 1 Mb). Ce phénomène de réutilisation des régions de cassure et l'existence d'un très grand nombre de petits blocs de synténie ne sont pas compatibles avec une distribution aléatoire des cassures. Ce point a été fortement débattu dans la littérature, donnant lieu à un mémorable débat par articles interposés entre Pavel Pevzner et David Sankoff pour trancher s'il s'agissait d'un artefact dû à la résolution des blocs conservés utilisés ou s'il relevait d'une réalité biologique (Bourque et al. 2004; Trinh et al. 2004; Sankoff and Trinh 2005; Peng et al. 2006; Alekseyev and Pevzner 2007). Ce résultat théorique a cependant été rapidement corroboré par les observations de réutilisations indépendantes des régions de cassure dans différentes lignées et par les données montrant que les régions de cassure sont statistiquement biaisées (voir paragraphes précédents). D'autres algorithmes comme EMRAE (Zhao and Bourque 2009) arrivent également aux mêmes conclusions. Ainsi, il est clair aujourd'hui que les points de cassure sont effectivement plus regroupés qu'attendu sous une distribution aléatoire dans le génome. Les valeurs de taux de réutilisation des régions de cassure, par contre, varient largement d'une étude à l'autre étant donné qu'elles dépendent directement de la résolution utilisée pour définir les blocs conservés.

3.6. Modèles évolutifs

L'ensemble des données expérimentales accumulées sur les régions de cassure, et notamment leur distribution dans les génomes, a mené à l'établissement de trois grands modèles évolutifs que nous synthétisons ici. Ces modèles diffèrent largement sur leurs interprétations des données et les hypothèses qu'ils proposent quant aux mécanismes biologiques sous-jacents, et sont la base des hypothèses que nous avons considérées comme plausibles au cours de ce travail de thèse.

3.6.1. Modèle aléatoire

Historiquement, le premier modèle proposé pour décrire l'apparition des points de cassure dans les génomes a été proposé par Susumu Ohno (Ohno 1973). Ohno proposait alors que les génomes se cassent et se réarrangent de manière totalement aléatoire, ce qui suppose une probabilité de cassure uniforme et indépendante sur l'ensemble du génome. Cette hypothèse a été soutenue par plusieurs études menées à relativement faible résolution (Nadeau and Taylor 1984; Copeland et al. 1993; O'Brien et al. 1999; Lander et al. 2001). La conclusion que sous-tend ce modèle est que les blocs de gènes dont l'ordre est conservé dans les génomes modernes ne sont que des vestiges de leur organisation ancestrale commune en cours de dégradation, et ne reflètent pas une organisation préservée en raison de son importance fonctionnelle. Il s'agit donc d'un modèle d'évolution profondément neutraliste suggérant qu'il n'existe pas de pression de sélection sur l'organisation des gènes dans les génomes.

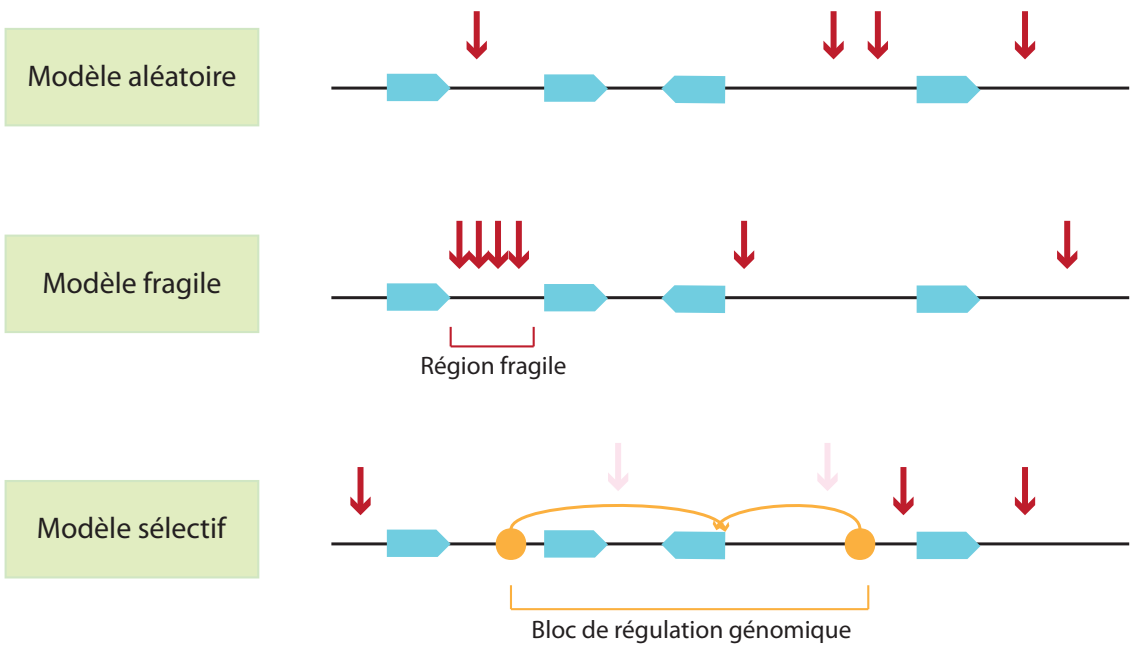


Figure 3.12. Trois modèles d'évolution du génome. Dans le cas du modèle aléatoire, les points de cassure (flèches) sont répartis aléatoirement entre les gènes. Dans le cas du modèle fragile, certaines régions du génome concentrent les points de cassure, et donnent l'impression d'une réutilisation indépendante des points de cassure à différentes étapes de l'évolution. Dans le cas du modèle sélectif, les points de cassure se produisant entre les gènes et leurs séquences de régulation (en jaune) sont contre-sélectionnés et éliminés (flèches pâles). Les points de cassure qui atteignent la fixation sont restreints aux zones où ils ne perturbent pas de circuits de régulation.

Ce modèle n'a cependant pas résisté aux analyses à haute résolution rendues possibles par les séquences complètes des génomes depuis les années 2000, qui montrent que les points de cassure de réarrangements évolutifs se concentrent dans un nombre de régions trop restreint pour refléter une distribution aléatoire, malgré le fait que ces régions sont elles-mêmes réparties aléatoirement dans le génome. Certains auteurs ont ainsi avancé que le modèle aléatoire est valable à grande échelle quand il s'agit de décrire l'organisation d'un génome par rapport à un autre, mais qu'il ne décrit pas la probabilité d'apparition des points de cassure sous-jacents aux événements de réarrangements (Becker and Lenhard 2007). Ce modèle aléatoire est largement utilisé comme hypothèse nulle considérée pour mettre en évidence des propriétés statistiquement associées aux cassures, à ceci près qu'il est aujourd'hui assez clair que les points de cassure évolutifs ne se trouvent presque jamais à l'intérieur des gènes (Peng et al. 2006; Lemaitre et al. 2009; Mongin et al. 2009). Il est probable que de telles cassures se produisent, mais qu'elles sont fortement contre-sélectionnées et qu'elles n'atteignent presque jamais la fixation dans les génomes modernes. Par « modèle aléatoire » ou « hypothèse nulle », on entend donc que la probabilité de cassure est uniforme sur toutes les bases ou sur toutes les bases non-codantes, selon les études (Figure 3.12).

3.6.2. Modèle fragile

Le premier modèle proposé pour expliquer la concentration des points de cassure dans des régions apparemment limitées du génome est celui énoncé initialement par Pevzner et Tesler dans leur démonstration théorique de la réutilisation des régions de cassure (2003) : les points de cassure se regrouperaient dans des régions « fragiles » du génome (Figure 3.12). Ces zones de fragilité seraient des régions où la probabilité qu'une cassure se produise est plus élevée qu'ailleurs, ou encore où lorsqu'une cassure se produit, la probabilité qu'elle soit réparée de façon erronée est plus élevée. Cette fragilité peut donc être d'ordre mécanique ou liée aux propriétés locales de la séquence, comme par exemple le contenu en séquences répétées qui peuvent servir de support à des recombinaisons ectopiques. Dans les deux cas, ce modèle fragile est à nouveau une interprétation neutraliste des caractéristiques des régions de cassure, puisqu'il sous-entend que la distribution non-aléatoire des points de cassure reflète principalement leurs modalités d'apparition, pas de fixation. Ce modèle s'oppose donc au modèle aléatoire sur l'aspect uniforme de la probabilité de réarrangement, mais les deux modèles s'accordent sur le fait que l'organisation des gènes dans le génome n'est pas spécialement sous contrainte de sélection.

3.6.3. Modèle sélectif

Plus récemment, un second modèle a été proposé pour expliquer la distribution non-aléatoire des points de cassure dans les génomes de vertébrés. Ce modèle s'appuie sur un ensemble d'observations montrant que les blocs de gènes dont l'ordre est conservé sur de grandes distances évolutives (à l'échelle des amniotes ou des vertébrés) contiennent de nombreux éléments conservés non-codants, alors que ceux-ci sont sous-représentés dans les régions de cassure (Nobrega et al. 2003; Sandelin et al. 2004; Kikuta et al. 2007; Hufton et al. 2009; Mongin et al. 2009). Une grande partie de ces éléments conservés non-codants sont des sites de fixation de facteurs de transcription dont le gène-cible peut se trouver à une grande distance sur le chromosome (jusqu'à plusieurs Mb (Poulin et al. 2005; Woolfe et al. 2005; Pennacchio et al. 2006; Kikuta et al. 2007)). Lorsqu'ils ont pu être identifiés, les gènes-cibles de ces séquences de régulation ultraconservées sont fréquemment des gènes impliqués dans le développement de l'organisme, qui sont eux-mêmes surreprésentés parmi les régions les plus stables du génome. Plusieurs articles ont donc proposé que les interactions à longue distance entre gènes et séquences de régulation forment des « blocs de régulation génomique » (GRB, Genomic Regulatory Blocks) et sont responsables en grande partie du maintien de l'ordre des gènes dans les blocs conservés entre espèces, parce qu'elles exercent une contrainte sélective sur l'organisation du génome (Becker and Lenhard 2007; Kikuta et al. 2007; Hufton et al. 2009; Mongin et al. 2009). Les réarrangements qui perturbent les interactions de régulations seraient en général délétères, et ne pourraient atteindre la fixation que lorsque leurs points de cassure se trouvent dans la portion réduite du génome où ils n'ont pas de conséquences fonctionnelles. Ainsi, ce modèle propose que la distribution des points de cassure n'est pas gouvernée par la probabilité qu'une cassure se produise à un endroit donné du génome (qu'elle soit uniforme ou

non), mais par les contraintes sélectives qui permettent ou non sa fixation (Figure 3.12) : il s'agit donc d'un modèle fondamentalement sélectif, contrairement aux deux précédents.

Les modèles sélectifs et fragiles peuvent l'un et l'autre rendre compte des observations de la littérature, et il est probable que la réalité se trouve quelque part entre les deux : la distribution des cassures s'explique dans doute en partie par des variations de la probabilité de cassure, et en partie par des contraintes sélectives sur certaines régions. A notre connaissance, aucune étude n'a pu explorer dans quelle mesure chacun des deux modèles contribue au modelage des génomes de vertébrés tels que nous les connaissons aujourd'hui, et si l'un des deux modèles domine en importance sur l'autre.

Chapitre 4. Les duplications complètes du génome

L'un des projets menés au cours de cette thèse a consisté à donner un aperçu global de l'organisation du génome d'une espèce modèle, le génome du poisson zèbre. Les génomes de poissons osseux (téléostéens), dont le poisson zèbre fait partie, ont la particularité d'avoir subi une duplication complète datée à environ 350 Ma (Jaillon et al. 2004; Meyer and Van de Peer 2005; Kasahara et al. 2007). C'est autour de cet événement et de ses conséquences sur le génome du poisson zèbre que nous avons centré notre étude. Les duplications complètes du génome sont des événements de duplication massive où le nombre de chromosomes dans le génome est doublé, chaque chromosome et chaque gène initial étant ensuite présent en deux copies. On les oppose classiquement aux duplications segmentales, où un segment de chromosome de taille variable est dupliqué, si bien qu'une petite partie des gènes seulement se trouve alors en deux copies paralogues (Conant and Wolfe 2008). Dans le cas des organismes diploïdes, comme la grande majorité des vertébrés, le génome initial contient donc 2n chromosomes avant une duplication complète, et 4n après. Il s'agit d'une forme particulière de polyploïdisation, un ensemble d'événements plus vaste qui englobe toutes les modifications du caryotype se traduisant par une augmentation du nombre de copies du génome haploïde : triploïdisations (3n), hexaploïdisations (6n), etc. Les duplications complètes du génome sont généralement suivies d'une phase de retour à l'état diploïde, appelée rediploïdisation, où une grande partie des duplicats de gènes sont perdus pour revenir à un nombre de gènes du même ordre de grandeur qu'une espèce au génome non dupliqué. Dans ce chapitre, nous décrivons l'état des connaissances sur les duplications complètes de génomes chez les eucaryotes de manière générale, et sur la duplication 3R des poissons téléostéens en particulier.

4.1. Duplications complètes dans l'arbre des eucaryotes

L'hypothèse selon laquelle certains génomes seraient des polyploïdes dégénérés issus de duplications complètes du génome plus ou moins récentes a initialement été proposée par Susumu Ohno dans son livre célèbre *Evolution by Gene Duplication* (Ohno 1970). L'idée générale énoncée par Ohno était que la création de nouveaux gènes est un processus nécessaire à l'évolution, et qu'il est plus simple de créer de nouveaux gènes en dupliquant ceux qui existent que *de novo*. Ainsi, Ohno avait proposé que les duplications complètes pourraient être un processus majeur de création de nouveau matériel génétique et donc d'innovation dans les génomes (Ohno 1970; Ohno 1973). Ses observations, quoique très parcellaires en raison des moyens techniques de l'époque, suggéraient en particulier l'existence de deux duplications complètes du génome à la base de l'arbre des vertébrés. Bien que son hypothèse se basait sur des interprétations que l'on sait aujourd'hui incorrectes (notamment sur la taille totale des génomes, qui n'est en réalité pas corrélée au nombre de gènes dans un génome, un phénomène

connu sous le nom de « paradoxe de la valeur C »)(Skrabanek and Wolfe 1998), la possibilité que certains génomes soient issus de duplications complètes a trouvé un regain d'intérêt lorsqu'il a été découvert que les gènes *Hox*, contrôlant en partie la mise en place du plan d'organisation chez les métazoaires, se présentent sous la forme de quatre clusters chez les mammifères similaires à un unique cluster dans le génome de la drosophile (Schughart et al. 1989) et dans celui de l'amphioxus (Garcia-Fernandez and Holland 1994). Ces clusters ont été interprétés comme des traces potentielles de deux événements de duplication complète du génome à la base de la radiation des vertébrés, nommées duplications 1R et 2R, comme initialement proposé par Ohno. Cette hypothèse a été extrêmement débattue par la suite (Skrabanek and Wolfe 1998; Makalowski 2001; Taylor and Brinkmann 2001; Wolfe 2001; Spring 2002; Durand 2003) : la question n'a été définitivement tranchée qu'avec la mise à disposition de la communauté des séquences de plusieurs génomes de vertébrés et la découverte de nombreux blocs de gènes paralogues cohérents avec deux duplications complètes du génome (McLysaght et al. 2002; Dehal and Boore 2005). Une revue complète des arguments ayant été proposés en faveur ou défaveur de l'existence des duplications 1R et 2R n'est pas le sujet de ce manuscrit, mais ces événements ont été historiquement les premiers cas de duplications complètes ancestrales proposés dans la littérature.

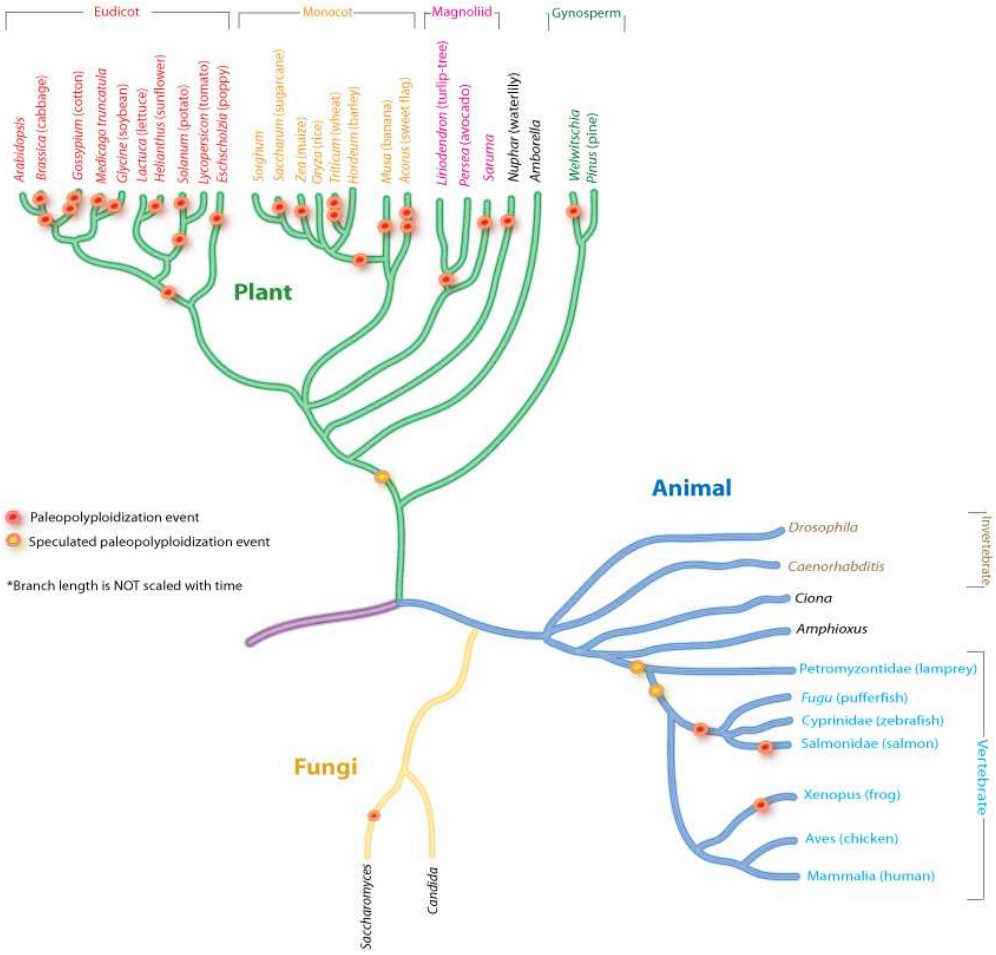


Figure 4.1. Localisation des duplications complètes du génome connues dans l'arbre phylogénétique des eucaryotes. Figure par P. Zhang à partir de données de (Wolfe 2001; Adams and Wendel 2005; Cui et al. 2006).

Depuis ces premières découvertes, il est apparu que les duplications complètes du génome sont un phénomène relativement courant dans l'histoire des eucaryotes (Figure 4.1)(Wolfe 2001; Van de Peer et al. 2009b). La fréquence des événements de duplication complète varie largement selon les taxons : elles sont assez rares dans les génomes de métazoaires en général, et en particulier chez les mammifères, alors qu'elles sont très communes chez les plantes où la plupart des groupes contiennent des espèces au génome dupliqué, voire présentant des niveaux de polyploïdie encore supérieurs (Masterson 1994; Otto and Whitton 2000; Wendel 2000; Blanc and Wolfe 2004). Les exemples les plus connus et les mieux étudiés sont les duplications 1R et 2R à la base du phylum des vertébrés, la duplication 3R de l'arbre des téléostéens (Jaillon et al. 2004), mais aussi la duplication complète à la base des levures saccharomycètes (Wolfe and Shields 1997; Kellis et al. 2004) et les duplications ayant affecté des plantes modèles, comme *Arabidopsis thaliana* (Blanc et al. 2003), ou d'intérêt agronomique, comme le maïs ou le coton (Blanc and Wolfe 2004).

4.2. Apparition des duplications complètes

La grande majorité de nos connaissances sur l'établissement des duplications complètes viennent des génomes de plantes, où les duplications complètes sont non seulement fréquentes mais où il est possible d'obtenir des polyploïdes viables de façon artificielle, contrairement aux génomes de métazoaires (Otto and Whitton 2000; Hufton and Panopoulou 2009). Les polyploïdies peuvent avoir des origines diverses : certaines résultent de la fusion des génomes de deux espèces proches mais distinctes (allopolyploïdisation, ou polyploïdisation hybride), ou de génomes provenant de la même espèce, voire du même individu (autopolyploïdisation)(Otto and Whitton 2000; Van de Peer et al. 2009b). Il n'est pas clair, à ce jour, si les duplications complètes sont généralement des phénomènes d'autopolyploïdisation ou d'allopolyploïdisation : dans les génomes de plantes, l'allopolyploïdisation semble favorisée (Levy and Feldman 2002), mais la duplication des levures saccharomycètes, par exemple, semble plus compatible avec une autopolyploïdisation en raison de la grande proximité initiale inférée pour les séquences des paralogues (Scannell et al. 2007).

Les mécanismes sous-jacents à l'apparition de duplications complètes naturelles sont mal connus. On considère cependant qu'il existe deux phases clés où les polyploïdisations peuvent se mettre en place : la mitose et la méiose (Otto and Whitton 2000; Comai 2005). Dans les deux cas, la division cellulaire échoue, si bien que la cellule se retrouve avec un double jeu de chromosomes. Si l'erreur a lieu lors de la mitose, on parle de doublement génomique ; si elle a lieu au cours de la méiose, on parle de non-réduction gamétique, puisqu'il s'agit généralement d'un problème de séparation des tétrades lors de la méiose I. La non-réduction gamétique est relativement fréquente dans les gamètes femelles chez les vertébrés (Otto and Whitton 2000) ; elle est plus rare dans les gamètes mâles, mais ceux-ci peuvent en revanche être impliqués dans des phénomènes de polyspermie (fécondation d'un ovocyte par plusieurs gamètes mâles), conduisant aux mêmes effets (Uchida and Freeman 1985). On estime ainsi que chez l'homme, 1 à 3% des fertilisations naturelles conduisent à un embryon triploïde (non viable)(McFadden et al. 1993).

L'apparition d'une duplication complète du génome et, surtout, sa fixation évolutive restent assez mystérieuses chez les espèces sexuées, car elles impliquent généralement le passage par des individus triploïdes à une étape ou à une autre : dans le cas d'un doublement génomique aux premières divisions de la vie de l'embryon, l'individu tétraploïde produira des gamètes diploïdes qui après fécondation par un gamète « normal » donneront des individus triploïdes ; dans les cas d'une non-réduction gamétique ou d'une polyspermie, c'est directement à l'étape de la fécondation que les individus triploïdes apparaissent. Or, les individus triploïdes sont considérés comme une impasse évolutive même lorsqu'ils sont viables, car leur fertilité est très basse en raison des méioses erratiques causées par leur caryotype impair (Otto and Whitton 2000; Comai 2005). Les triploïdes peuvent cependant produire des gamètes euploïdes (n ou 2n) à faible fréquence (Ramsey 1998) ; ainsi, les triploïdes pourraient permettre aux tétraploïdes de se reproduire, bien qu'avec un taux de succès faible, jusqu'au moment où la population de tétraploïdes est assez large pour grandir d'elle-même. Il a également été proposé que les duplications complètes affectent essentiellement des espèces pouvant pratiquer la reproduction asexuée ou l'autofécondation (Ramsey 1998; Otto and Whitton 2000), ce qui évite dans un premier temps la dépendance aux triploïdes, ou bien les espèces pérennes (Otto and Whitton 2000), dont le cycle de vie est assez long pour que la faible probabilité de produire une descendance viable soit compensée par une longue période de reproduction. Par ailleurs, la fréquence de production de gamètes diploïdes et d'individus tétraploïdes est très dépendante des conditions environnementales et du génotype parental (Bloom 1972; Bogart et al. 1989; Otto and Whitton 2000), suggérant que certaines circonstances pourraient donner lieu à une production locale élevée de tétraploïdes féconds entre eux.

4.3. Rediploïdisation

Dans la majorité des cas, les duplications complètes du génome ne produisent pas des génomes tétraploïdes stables sur de grandes distances évolutives. Ces événements sont suivis d'une érosion massive du contenu en gènes, où une grande partie des duplicats sont perdus, si bien qu'une seule copie sur les deux est retenue. Ce phénomène est appelé rediploïdisation et a été mis en évidence dans les génomes dupliqués de plantes (Blanc and Wolfe 2004; Paterson et al. 2004), de levures (Kellis et al. 2004; Scannell et al. 2006) et de vertébrés (Jaillon et al. 2004). Il s'agit d'un retour vers le nombre de gènes initial par pertes de gènes, et non d'un retour au nombre initial de chromosomes, qui peut rester constant. Ainsi, on considère que seuls environ 20% des gènes sont des paralogues issus de la duplication complète dans les génomes de levures saccharomycètes, les autres duplicats ayant été perdus (Byrne and Wolfe 2005).

Les mécanismes gouvernant la perte de gènes sont mal connus car ce processus commence dès l'événement de polyploïdisation et, à nouveau, l'essentiel de nos connaissances viennent de polyploïdes synthétiques qui ne reflètent pas forcément les processus se produisant dans la nature ou permettant à terme une fixation évolutive. Deux mécanismes semblent être mis en jeu : la délétion de segments d'ADN et la pseudogénisation (Hufton and Panopoulou 2009). Les polyploïdes synthétiques de blé et de levures ont montré qu'en quelques générations seulement, des segments d'ADN voire des chromosomes entiers sont perdus (Shaked et al. 2001; Levy and

Feldman 2002; Gerstein et al. 2006), dont certains de manière spécifique et reproductible d'une expérience à l'autre. Ces délétions se font par recombinaisons homologues non alléliques causées par des appariements impropres des chromosomes au cours de la méiose. Ces pertes de segments d'ADN sont peut-être favorisées parce qu'elles permettent la différenciation des double-paires de chromosomes homologues issues de la duplication en deux paires distinctes, restaurant progressivement l'établissement de bivalents normaux au cours de la méiose (Feldman and Levy 2005; Hufton and Panopoulou 2009). D'autres gènes sont perdus par pseudogénisation, c'est-à-dire qu'ils sont non-fonctionnalisés puis dégradés par accumulation de mutations ponctuelles (Wolfe 2001). La proportion de gènes perdus par l'un ou par l'autre de ces phénomènes n'est, à notre connaissance, pas connue à ce jour, mais on considère que la séquence d'événements de rediploïdisation passe par une phase rapide et courte de délétions, à laquelle succède une phase de pseudogénisation plus lente par mutations (Feldman and Levy 2005).

La demi-vie d'un gène dupliqué dans un génome eucaryote est estimée de 3 à 7 Ma seulement (Lynch and Conery 2000). La rediploïdisation de la majorité des gènes se produit donc probablement rapidement après la duplication complète, mais peut se poursuive, avec un taux plus faible, sur un temps assez long. Différentes espèces ayant divergé plusieurs dizaines de millions d'années après l'événement de duplication complète peuvent avoir retenu des copies différentes de certains gènes, preuve que la rediploïdisation n'était pas achevée lors de la divergence des lignées : par exemple, environ 1700 gènes (~8%) dans les génomes du poisson zèbre et du tétraodon sont des paralogues issus de pertes réciproques, alors que les deux lignées se sont séparées environ 30 Ma après la duplication complète du génome (Semon and Wolfe 2007b). La raison pour laquelle les génome dupliqués retournent progressivement à l'état diploïde au cours de leur évolution n'est pas claire : l'interprétation généralement retenue est qu'une grande partie des duplicats ne sont pas essentiels puisqu'une seule copie suffisait à l'organisme ancestral, et que la perte d'une des copies se fait de manière neutre, à défaut de pression de sélection pour retenir les gènes. Tous les gènes ne sont cependant pas affectés par la rediploïdisation : certains gènes sont préservés en deux copies dupliquées. Ces paralogues issus d'une duplication complète du génome sont appelés ohnologues, en référence à Susumu Ohno (Wolfe 2001). Ce sont eux qui constituent, à terme, les traces plus ou moins évidentes de duplications complètes du génome sur lesquelles nous reviendrons aux paragraphes suivants. Ces duplications anciennes, dégradées par la rediploïdisation et les événements ultérieurs affectant l'organisation du génome, sont alors désignées par le terme de « paléopolyploïdies ».

4.4. Conséquences des duplications complètes

Les duplications complètes sont des événements soudains : comme la plupart des événements de grande ampleur dans le génome, elles posent des problèmes immédiats qui peuvent se révéler délétères pour l'organisme, mais elles sont également des sources d'innovations suffisamment importantes pour apporter un avantage et être parfois retenues dans l'évolution.

4.4.1. Caractéristiques des organismes au génome dupliqué

L'une des principales caractéristiques morphologiques des organismes tétraploïdes récents ou artificiels est la taille de leurs cellules, qui sont plus grosses que celles des organismes diploïdes : ce phénomène a été observé chez les plantes et les métazoaires, en particulier les amphibiens tétraploïdes (Masterson 1994; Otto and Whitton 2000; Comai 2005). Cette augmentation de la taille se traduit par une modification du rapport surface/volume de la cellule, et a des conséquences sur les processus cellulaires impliquant notamment des membranes (interactions chromatine/lamina nucléaire, etc.)(Comai 2005; Otto 2007). L'impact de ces modifications n'est que peu documenté à ce jour : intuitivement, on peut imaginer qu'elles provoquent des déséquilibres dans l'organisation de la cellule et soient délétères, mais l'augmentation du volume cellulaire peut être avantageux dans certains cas, comme les cellules ayant un taux métabolique élevé (Comai 2005).

Ces organismes eux-mêmes peuvent également être plus gros, par exemple chez les plantes où les cultivars sont souvent des polyploïdes sélectionnés sur ce critère. Mais cela n'est pas toujours vrai, notamment chez les métazoaires où la taille des cellules tétraploïdes est globalement compensée par un nombre de cellules moins important au cours du développement (Otto and Whitton 2000; Comai 2005; Otto 2007).

4.4.2. Espèces sexuées et sex ratio

Les duplications complètes ont une conséquence importante sur les organismes dont la détermination du sexe est génétique et liée à des chromosomes sexuels différenciés. Deux mécanismes principaux de détermination du sexe faisant intervenir des chromosomes sexuels sont connus, notamment chez les animaux : la détermination par le ratio de chromosomes sexuels par rapport aux autosomes (mécanisme en place chez les drosophiles), et l'hétérogamétie, où le sexe est déterminé par la présence ou l'absence d'un chromosome dominant (Y donnant le sexe mâle, ou W donnant le sexe femelle, selon les espèces), qui est le mécanisme majoritaire chez les vertébrés. Dans les deux cas, la duplication complète va avoir des conséquences au moment de la reproduction (Otto and Whitton 2000).

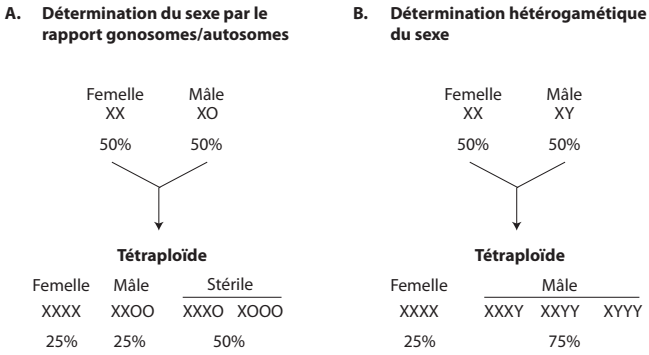


Figure 4.2. Déséquilibres du sex ratio causés par une duplication complète du génome dans les espèces à détermination du sexe liée au rapport gonosomes/autosomes (A), ou hétérogamétique (B).

Lorsque la détermination du sexe est basée sur le ratio gonosomes/autosomes, les croisements peuvent donner des individus pour lesquels le rapport n'est ni 1:1, ni 1:2, et qui sont alors stériles : ce phénomène est intuitivement délétère puisqu'il réduit la fitness de la descendance (Figure 4.2).

Lorsque la détermination du sexe est hétérogamétique, la duplication complète va déséquilibrer la descendance vers des descendants mâles (Figure 4.2). Ceci a pour conséquence de réduire de façon importante la taille effective de la population ; des mécanismes de sélection se mettent alors rapidement en place pour rééquilibrer le sex ratio (Basolo 1994). Par ailleurs, dans le cas où l'un des deux chromosomes sexuels est dégénéré et qu'il existe une compensation de dosage qui équilibre le taux d'expression des gènes des gonosomes par rapport à ceux des autosomes, comme chez les mammifères, la compensation de dosage n'est plus effective dans une partie de la descendance, ce qui peut également être délétère (Figure 4.2).

Ces difficultés ont été initialement proposées par Müller, en 1925, comme la principale raison pour laquelle les duplications complètes sont rares dans les espèces dont la détermination du sexe est liée à des chromosomes sexuels, et en particulier les amniotes. Il semblerait cependant que cette explication seule ne suffise pas à expliquer l'absence d'espèces polyploïdes récentes dans ce taxon : en effet, les polyploïdes naturels n'y sont pas viables, ce qui suggère également de graves problèmes développementaux causés par les duplications complètes (Otto 2007).

4.4.3. Création de nouveaux gènes

Comme nous l'avons évoqué plus haut, tous les gènes dupliqués ne sont pas conservés après une polyploïdisation : la duplication complète est généralement suivie par une période de perte massives de gènes pour revenir à un état diploïde. Dans certains cas, cependant, les deux gènes paralogues sont retenus, qu'on désigne alors sous le nom d'ohnologues pour les distinguer des paralogues issus de duplications segmentales (en hommage à Susumu Ohno). Ces gènes constituent une source importante de nouveau matériel génétique. Ainsi, les duplications complètes pourraient être sélectionnées sous certaines conditions parce que leur pouvoir d'innovation compense les conséquences délétères éventuelles causées par la tétraploïdie, que nous avons évoquées plus haut. Une attention particulière a été apportée à ces ohnologues dans la littérature, afin de comprendre pourquoi l'apparition de nouveaux gènes par duplication complète apporte un avantage à l'organisme, et les raisons pour lesquelles certains gènes sont maintenus sous forme de duplicats alors que d'autres sont éliminés. Les ohnologues issus de différents événements de tétraploïdisation présentent un ensemble de caractéristiques communes qui apportent un éclairage sur les forces évolutives en jeu après une duplication complète du génome.

4.4.3.1. Redondance

Chez les plantes et les levures comme chez les métazoaires, les ohnologues sont enrichis en gènes impliqués dans le développement et dans la régulation de l'expression des gènes, connus pour être fortement délétères voire létaux lorsqu'ils sont non-fonctionnels (Blomme et al. 2006;

Conant and Wolfe 2008; Huminiecki and Heldin 2010). Les duplications du génome pourraient être avantageuses parce qu'elles augmentent la redondance du génome : ce doublement des gènes de l'ensemble du génome fournit en quelque sorte un jeu de gènes « de rechange » qui pourrait masquer les mutations délétères (Cooke et al. 1997; Wolfe 2001; Comai 2005). Ainsi, certains gènes particulièrement essentiels seraient maintenus en deux copies après la duplication parce que la présence d'une seconde copie du gène permettrait de compenser les mutations de l'autre, rendant moins probable l'apparition de phénotypes délétères. Mais cette interprétation est sujette à controverse, car cette redondance peut également mener au maintien d'allèles mutés délétères à une fréquence assez importante dans les populations, ce qui n'est pas forcément avantageux (Otto and Whitton 2000; Otto 2007).

L'augmentation de la redondance du génome pourrait également présenter un avantage lorsque certains gènes présentent une situation d'hétérosis, soit un avantage sélectif aux hétérozygotes. Une duplication offre alors la possibilité de fixer les deux allèles parentaux dans le même génome, l'un à chacun des deux locus paralogues ; ainsi, tous les individus profitent de l'avantage sélectif (Comai 2005). Alternativement, les deux copies paralogues d'un gène peuvent également être retenues si l'augmentation du taux d'expression du gène est sélectionnée (Kondrashov et al. 2002; Conant and Wolfe 2008).

4.4.3.2. Subfonctionnalisation

Une autre interprétation est que les duplications de gènes permettraient à des gènes remplissant plusieurs fonctions de se spécialiser, soit par différenciation des fonctions remplies par les deux duplicats de la protéine initiale, soit par la séparation des patrons d'expression des deux gènes. Cette interprétation est soutenue par le fait que les gènes fortement exprimés (Seoighe and Wolfe 1999; Chain et al. 2011), pléiotropiques (Chain et al. 2011; Satake et al. 2012) ou comprenant un grand nombre de domaines protéiques différents (Gibson and Spring 1998) sont surreprésentés parmi les ohnologues. Le modèle désormais classique de duplication-dégénération-complémentation (DDC) proposé par Force et al. (1999) est présenté dans la Figure 4.3. Les deux copies ohnologues, se compensant mutuellement, pourraient absorber l'une comme l'autre des mutations qui les privent d'une partie de leurs fonctions sans conséquences pour l'organisme, jusqu'à ce que les deux copies soient fixées car remplissant chacune des fonctions que l'autre a perdues. Les mutations causant la divergence et la subfonctionnalisation peuvent affecter la séquence codante de la protéine, ou bien, plus fréquemment, les séquences de régulation contrôlant son expression (Wapinski et al. 2007b; Semon and Wolfe 2008). Les gènes avec un taux d'évolution lent (donc sous pression de sélection forte) sont plus susceptibles d'être subfonctionnalisés après une duplication complète du génome (Brunet et al. 2006; Semon and Wolfe 2008; Chain et al. 2011) ; par ailleurs, on note une accélération du taux d'évolution de ces gènes après la duplication, ce qui est compatible avec une levée partielle de la pression de sélection et une fixation par accumulation de mutations dans les deux copies (Brunet et al. 2006; Hellsten et al. 2007).

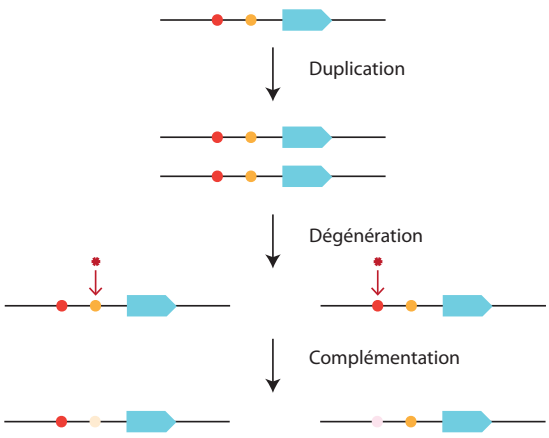


Figure 4.3. Modèle de Duplication-Dégénation-Complémentation (DDC) tel que proposé par Force et al. (1999). Les cercles rouge et jaune représentent des séquences de régulation du gène figuré en bleu. Les astérisques rouges représentent des mutations perte de fonction.

Parmi les gènes ohnologues ayant été retenus par subfonctionnalisation, on peut citer l'exemple des gènes *Sir3* et *Orc1* chez *Saccharomyces cerevisiae.* Ces gènes ont des fonctions différentes, l'un dans la répression de l'expression des gènes et l'autre dans le complexe de reconnaissance des origines de réplication, mais ils sont tous deux orthologues d'un seul gène qui remplit les deux fonctions chez *Lachancea kluyveri*, une levure proche au génome non dupliqué (Kellis et al. 2004; van Hoof 2005).

4.4.3.3. Néofonctionnalisation

Une troisième possibilité est que les duplicats issus de la polyploïdisation sont une source de gènes peu contraints puisque surnuméraires, dont l'une des copies aurait la possibilité de muter et d'explorer de nouvelles pistes évolutives (Ohno 1970). Parmi les ohnologues, on observe fréquemment une copie dont le taux d'évolution est plus important que l'autre, suggérant que l'on n'est pas face à deux copies partiellement dégénérées par rapport à l'état ancestral (subfonctionnalisation) mais à l'apparition de mutations préférentiellement sur une des deux copies, moins contrainte que l'autre et ayant potentiellement acquis de nouvelles adaptations (Kellis et al. 2004; Brunet et al. 2006; Byrne and Wolfe 2007). Ainsi, certains duplicats pourraient évoluer librement par mutations, formant une source de nouvelles adaptations en offrant des possibilités d'innovations avantageuses, pendant que la deuxième copie conserve la fonction ancestrale qui reste nécessaire au fonctionnement de l'organisme (Figure 4.4). Il faut cependant noter que les cas de néofonctionnalisations (où la nouvelle fonction est différente de la fonction ancestrale) démontrées dans la littérature sont rares, sauf chez *Saccharomyces cerevisiae* où ce mécanisme de rétention des copies pourrait être prévalent (Kellis et al. 2004; Byrne and Wolfe 2007). Les limites entre subfonctionnalisation, acquisition de fonctions proches de la fonction ancestrale et néofonctionnalisation sont de fait assez floues : certains auteurs ont donc proposé la notion de « subnéofonctionnalisation », où un gène dupliqué est initialement retenu par subfonctionnalisation puis peut également acquérir de nouvelles fonctions proches par mutation tout en conservant une partie de ses fonctions ancestrales (He and Zhang 2005). Certains auteurs ont également proposé que la rétention de gènes dupliqués préférentiellement

par subfonctionnalisation ou par néofonctionnalisation soit liée à la taille de la population. Dans les populations de grande taille, comme les levures, la rétention de deux copies ohnologues serait plus probable si l'une d'entre elle acquiert une innovation fonctionnelle que par subfonctionnalisation, qui sous-entend que des mutations perte de fonction se fixent par dérive (Lynch and Conery 2000; Byrne and Wolfe 2007).

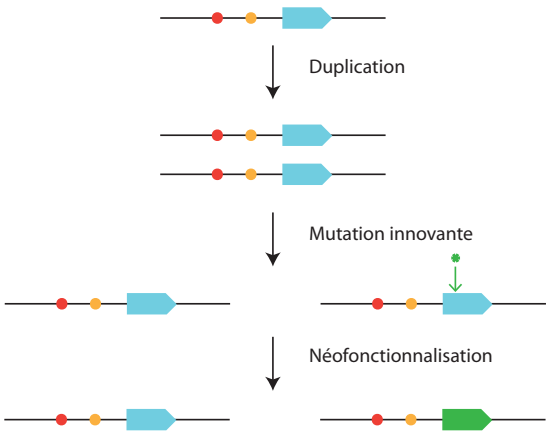


Figure 4.4. Conservation de gènes dupliqués par néofonctionnalisation. Les cercles rouge et jaune représentent des séquences de régulation du gène figuré en bleu. L'astérisque verte représente une mutation gain de fonction.

4.4.3.4. Equilibre de dosage

Une quatrième interprétation a été proposée pour expliquer la rétention d'une partie des gènes formés par duplication complète du génome, sans se baser sur la différenciation des fonctions du duplicats ou de leurs patrons d'expressions (Van de Peer et al. 2009b). Les ohnologues sont enrichis en gènes impliqués dans des complexes protéiques, dans la transduction du signal, en protéines interagissant avec de multiples partenaires et en facteurs de transcription, à l'inverse des paralogues issus de duplications segmentales parmi lesquels ces fonctions sont sous-représentées (Maere et al. 2005; Blomme et al. 2006; Guan et al. 2007; Hakes et al. 2007; Conant and Wolfe 2008; Makino and McLysaght 2010). Or, ces gènes présentent généralement des taux d'expression finement régulés qui permettent de maintenir la stœchiométrie des différents composants du complexe ou de la voie de signalisation. Suite à une duplication complète, la stœchiométrie est respectée puisque tous les taux d'expression sont doublés. En revanche, si une copie de l'un des gènes est perdue, ce gène sera deux fois moins exprimé que les autres, déséquilibrant le complexe. Ainsi, il existerait une pression de sélection pour maintenir l'équilibre de dosage entre les gènes impliqués dans des complexes protéiques, qui favoriserait la conservation de cette classe d'ohnologues ; par ailleurs, toute perte d'un des gènes du complexe entraînerait la rediploïdisation rapide de tous les gènes du complexe. Plus de 90% des expansions récentes de gènes impliqués dans la régulation de l'expression des gènes et dans la transduction du signal chez *Arabidopsis thaliana* sont dues à des duplications complètes du génome exclusivement (Maere et al. 2005) ; 75% des gènes du chromosome 21 du génome humain sont des ohnologues issus des duplications complètes 1R et 2R et sont potentiellement

impliqués dans les déséquilibres de dosage causant le syndrome de Down en cas de trisomie 21 (Makino and McLysaght 2010).

4.4.3.5. Dominants négatifs

Enfin, certains gènes pourraient être retenus en deux copies paralogues parce que les mutations ponctuelles dans ces gènes ont tendance à produire des dominants négatifs, qui résultent en des phénotypes délétères (Cooke et al. 1997; Gibson and Spring 1998; Wolfe 2001). Ainsi, ces ohnologues seraient conservés par défaut car les mutations pouvant mener à la pseudogénisation sont contre-sélectionnées. Ce phénomène a été proposé pour expliquer pourquoi certaines protéines, notamment dans la famille des tyrosines kinases, sont maintenues en huit copies paralogues pratiquement identiques chez les mammifères alors que leurs patrons d'expression sont redondants et que l'élimination de l'une ou l'autre d'entre elles par délétion complète du gène ne cause dans la plupart des cas pas de phénotype (Gibson and Spring 1998). Les protéines comprenant plusieurs domaines protéiques ou interagissant avec de nombreux partenaires, plus susceptibles de former des dominants négatifs en perdant une partie seulement de leurs fonctions par mutation ponctuelle ou troncage (Veitia 2010), seraient ainsi plus souvent retenues sous forme d'ohnologues sans pour autant présenter un avantage adaptatif.

4.4.4. Remodelage épigénétique

Au délà des nouveaux gènes créés par la duplication, les organismes tétraploïdes présentent également des différences au niveau des marques épigénétiques de leur génome. Dans les tétraploïdes synthétiques de plantes, entre 10% et 30% des CpG présentent des différences de méthylation entre le tétraploïde et le diploïde parental (Shaked et al. 2001; Otto 2007). Ces changements des marques épigénétiques ont pour conséquence des modifications des taux d'expression de certains gènes et une activation des éléments transposables (Comai 2005; Feldman and Levy 2005; Salmon et al. 2005; Chen and Ni 2006; Bento et al. 2008) ; elles peuvent également altérer les patrons d'empreinte parentale dans les gamètes (Comai 2005; Udall et al. 2006). Ces modifications pourraient être délétères en déstabilisant les patrons d'expression de gènes établis par sélection ; mais elles pourraient également jouer un rôle dans l'acquisition de nouvelles fonctions par les gènes créés lors des duplications, en augmentant la diversité et la plasticité (Comai 2005).

4.4.5. Taux de réarrangements

Plusieurs études ont observé que les génomes dupliqués, notamment chez les plantes, semblent plus réarrangés que les génomes diploïdes proches (Comai 2005; Chen and Ni 2006; Otto 2007; Semon and Wolfe 2007a). Cette accélération pourrait être causée soit par la forte similarité des chromosomes issus de la duplication, qui provoquerait une augmentation des recombinaisons aberrantes au cours de la méiose (Chen and Ni 2006), soit par l'activation des éléments transposables liée au remodelage épigénétique, qui déstabiliserait le génome (Feldman

and Levy 2005). Alternativement, comme dit plus haut pour les délétions, il est possible que les événements qui différencient les chromosomes homologues issus de la duplication soient fortement sélectionnés dès les premières générations parce qu'ils contribuent au rétablissement de bivalents normaux au cours de la méiose.

Cependant, il n'est pas évident qu'une augmentation de la fréquence des réarrangements soit réellement une caractéristique générale des duplications complètes du génome. En effet, certains génomes paléopolyploïdes ont gardé une organisation remarquablement colinéaire à leurs phylums apparentés diploïdes, comme par exemple chez les levures (Fischer et al. 2006; Gordon et al. 2009). Dans le cas du medaka, le caryotype est stable pendant près de 300 Ma, bien que plusieurs réarrangements majeurs se soient mis en place juste après la duplication (Kasahara et al. 2007). Chez la paramécie, qui a subi plusieurs duplications complètes successives, la structure des chromosomes est également bien conservée, malgré un fort taux de rétention des copies dupliquées (Aury et al. 2006). Le taux de réarrangements dans les lignées dupliquées est particulièrement complexe à estimer en raison de la mauvaise conservation de la synténie liée aux pertes massives de gènes (voir paragraphe 4.5 ci-dessous), si bien que la question n'est pas tranchée : certains résultats sont cohérents avec une accélération des taux de réarrangements (Semon and Wolfe 2007a), d'autres n'en détectent pas (Fischer et al. 2006). Une étude a par ailleurs suggéré qu'une augmentation des taux de réarrangements aurait en réalité précédé les deux événements de duplication complète dans le cas des vertébrés, et propose que duplications complètes et accélération des réarrangements ne seraient pas des conséquences l'un de l'autre mais deux types d'événements pouvant se produire lors de périodes d'instabilité du génome (Hufton et al. 2008).

4.4.6. Spéciations et extinctions

Les duplications complètes eucaryotes se placent fréquemment à la base de taxons ayant connu un taux de spéciation particulièrement important, si bien que de nombreux auteurs ont proposé que le nouveau matériel génétique produit par la duplication pourrait servir de support à l'adaptation à de nouvelles niches évolutives, en augmentant la vigueur de l'individu et les opportunités d'innovations et de spéciation (Wittbrodt et al. 1998; Comai 2005; Otto 2007). En effet, plusieurs études ont montré que, chez les plantes et les amphibiens, les polyploïdes ont tendance à être plus vigoureux que les parents diploïdes, à pouvoir croître dans une gamme plus vaste d'environnements, voire à se montrer invasifs (Thompson and Lumaret 1992; Ellstrand and Schierenbeck 2000). Paradoxalement, même dans les taxons où les duplications complètes sont très fréquentes, comme les plantes, les paléopolyploïdies ayant survécu restent assez peu nombreuses comparées au grand nombre d'événements récents connus : ainsi, il a été proposé que les duplications complètes se maintiennent rarement à long terme, mais que lorsqu'elles le font, la lignée évolutive qui en descend connaît un grand succès évolutif et se diversifie largement (Van de Peer et al. 2009b). Plusieurs paléopolyploïdies ayant donné naissance à de grands taxons modernes coïncident avec les dates de grandes transitions géologiques et écologiques : les duplications 1R et 2R des vertébrés sont concomitantes de la transition cambrienne, la duplication 3R des téléostéens de la crise Permien-Trias et de très nombreuses lignées de plantes ont subi une duplication complète il y environ 70 Ma, soit au moment de la

crise Crétacé-Tertiaire (Crow and Wagner 2006; Fawcett et al. 2009; Van de Peer et al. 2009b). Les duplications complètes pourraient être favorisées par les périodes de modifications profondes des habitats, où la mise en jeu de nouveaux gènes ouvrirait des possibilités d'adaptation nouvelles permettant de répondre aux modifications de l'environnement, évitant ainsi l'extinction de la lignée.

La raison pour laquelle une duplication avantageuse augmenterait ensuite la probabilité de spéciation, et conduirait à des explosions évolutives, mettrait en cause plusieurs mécanismes menant à l'isolation reproductive. La perte différentielle et la subfonctionnalisation des duplicats peuvent toutes deux mettre en place des barrières à la reproduction entre individus en suivant une variation du modèle classique de Bateson-Dobzhansky-Müller (Figure 4.5)(Lynch and Force 2000; Scannell et al. 2006; Semon and Wolfe 2007b; Van de Peer et al. 2009b).

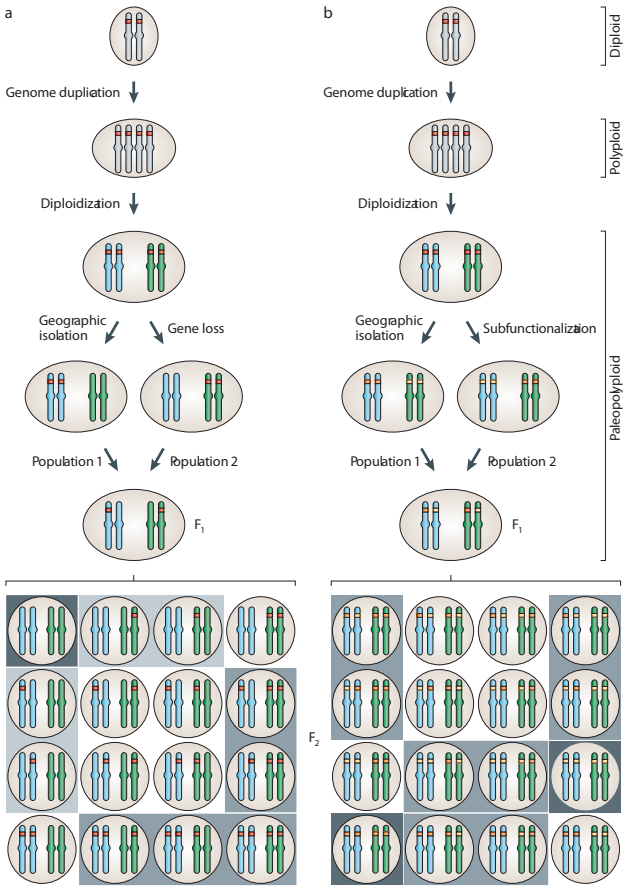


Figure 4.5. Facilitation de la spéciation par pertes réciproques (a) ou subfonctionnalisation (b) dans deux populations. Les bandes rouges représentent un locus dupliqué particulier sur les chromosomes issus de la duplication complète. Dans le cas (a), une des copies dupliquées est perdue sur un chromosome homologue différent dans chaque population. Dans le cas (b), chaque copie est subfonctionnalisée (gène orange ou jaune) de manière différente dans les deux populations. Lorsque les deux populations se croisent, la descendance hétérozygote donne à la génération F2 des individus auxquels manquent au moins une partie des fonctions ancestrales du gène (représentés sur fond gris foncé), et des individus pouvant avoir une fitness réduite pour des raisons de dosage ou d'haploinsuffisance (représentés sur fond gris clair). Figure tirée de (Van de Peer et al. 2009b).

Par exemple, avec une paire de paralogues revenus à l'état diploïde de manière différentielle, le croisement entre de tels individus donne à la génération F2 1/16ème d'individus auxquels manquent un gène ancestral ; plus les pertes réciproques de gènes sont nombreuses, en particulier au niveau de gènes essentiels, et plus la probabilité d'obtenir une descendance non viable augmente, ce qui conduirait, à terme, à l'isolation des deux populations et à la spéciation.

4.5. Organisation des génomes dupliqués

En raison du doublement de leur matériel chromosomique et de la perte massive de gènes qui suit, les génomes paléopolyploïdes ont une organisation particulière qui les différencie des génomes diploïdes. Ces caractéristiques ont d'ailleurs été largement exploitées pour mettre en évidence l'existence de duplications complètes ancestrales dans différents taxons.

4.5.1. Grandes régions de paralogie

Théoriquement, les paires d'ohnologues subsistant dans le génome dupliqué après la rediploïdisation permettent de détecter immédiatement les chromosomes dupliqués correspondant à un unique chromosome ancestral (Wolfe 2001). Dans une représentation circulaire du génome où l'on relie les positions des ohnologues, on s'attend à observer des faisceaux de liens reliant des chromosomes ou des régions de chromosomes de même origine ancestrale, à condition que le caryotype n'ait pas été trop remanié par les réarrangements. Ces relations de paralogies quasi-exclusives entre grandes régions du génome ont permis de mettre en évidence des duplications complètes ancestrales mêmes anciennes (Figure 4.6)(Jaillon et al. 2004; Dehal and Boore 2005; Aury et al. 2006). Ce signal diminue avec les réarrangements, les pertes de gènes et les duplications ultérieures qui viennent se surimposer au fil de l'évolution.

4.5.2. Dégradation de la synténie

Une caractéristique récurrente des génomes paléopolyploïdes est la disruption massive de l'ordre des gènes par rapport aux espèces non dupliquées les plus proches. Cette dégradation de la synténie au sens strict est visible dans de nombreux génomes dupliqués, où les adjacences gène à gène ancestrales ne sont que rarement conservées (Woods et al. 2000; Jaillon et al. 2004; Van de Peer 2004). Ces pertes d'adjacences sont liées aux délétions de gènes lors de la rediploïdisation (Figure 4.7) combinées, dans certains cas, à l'augmentation du taux de réarrangements (voir paragraphe 4.4.5). Si la rediploïdisation se fait essentiellement par pertes de segments entiers d'ADN contenant plusieurs gènes, la perte de synténie peut être plus limitée (Figure 4.7), mais le rôle joué par ce mécanisme dans la rediploïdisation est mal connu, comme nous l'avons vu au paragraphe 4.3.

La synténie au sens large est elle aussi perturbée par les mêmes mécanismes, puisque une région ancestrale voit son contenu en gènes réparti sur deux chromosomes différents après la rediploïdisation. De nombreux gènes qui se trouvaient initialement sur le même chromosome se retrouvent sur des chromosomes différents, et perdent les liaisons qui les unissaient dans le génome ancestral.

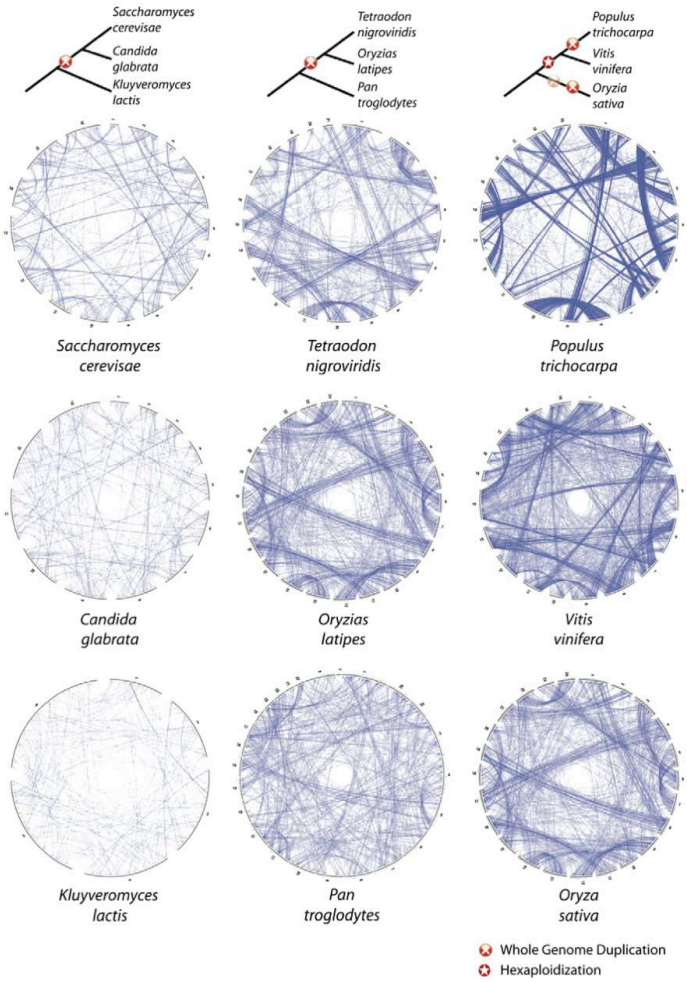


Figure 4.6. Représentation des liens de paralogie entre gènes dans différents génomes paléopolyploïdes ou non. Les arbres représentent la localisation des événements de duplication complète dans les différentes lignées. Les chromosomes de chaque espèce sont représentées sous forme d'un cercle, et les gènes paralogues sont reliés par des traits. Les génomes dupliqués ont une organisation nettement différente de celle des génomes non dupliqués, avec des faisceaux de liens reliant les grandes régions de paralogie issues de la duplication complète. Figure tirée de (Jaillon et al. 2009).

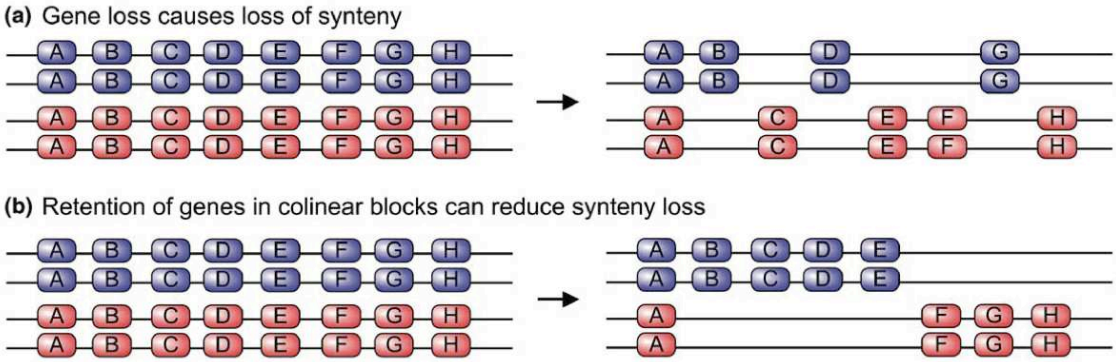


Figure 4.7. Perte de la synténie conservée au sens strict causée par la rediploïdisation qui suit une duplication complète. La perte de gènes par délétions de larges blocs peut limiter la perte de synténie. Figure tirée de (Hufton and Panopoulou 2009).

4.5.3. Synténies doubles conservées

Le dédoublement de chaque région ancestrale sur deux chromosomes différents dans le génome paléopolyploïde, s'il brouille le signal de synténie classique, donne cependant une signature particulière aux génomes paléopolyploïdes. En effet, chaque région d'un génome non dupliqué voisin correspond à deux régions du génome dupliqué, qui reconstituent son contenu en gènes. Les gènes de l'espèce non-dupliquée se retrouvent alternativement sur l'un ou l'autre des chromosomes de l'espèce dupliquée, créant un motif d'alternance nommé « synténie double-conservée » (Figure 4.8). Comme pour la synténie au sens classique, selon la distance phylogénétique entre les espèces, la synténie double-conservée peut être stricte, c'est-à-dire que le motif d'alternance respecte l'ordre des gènes ancestral (comme représenté sur la Figure 4.8) ; ou bien elle peut être prise au sens large : les gènes d'une région unique dans une espèce non dupliquée sont bien répartis sur deux chromosomes clairement identifiables mais l'ordre ne correspond pas à l'ordre ancestral.

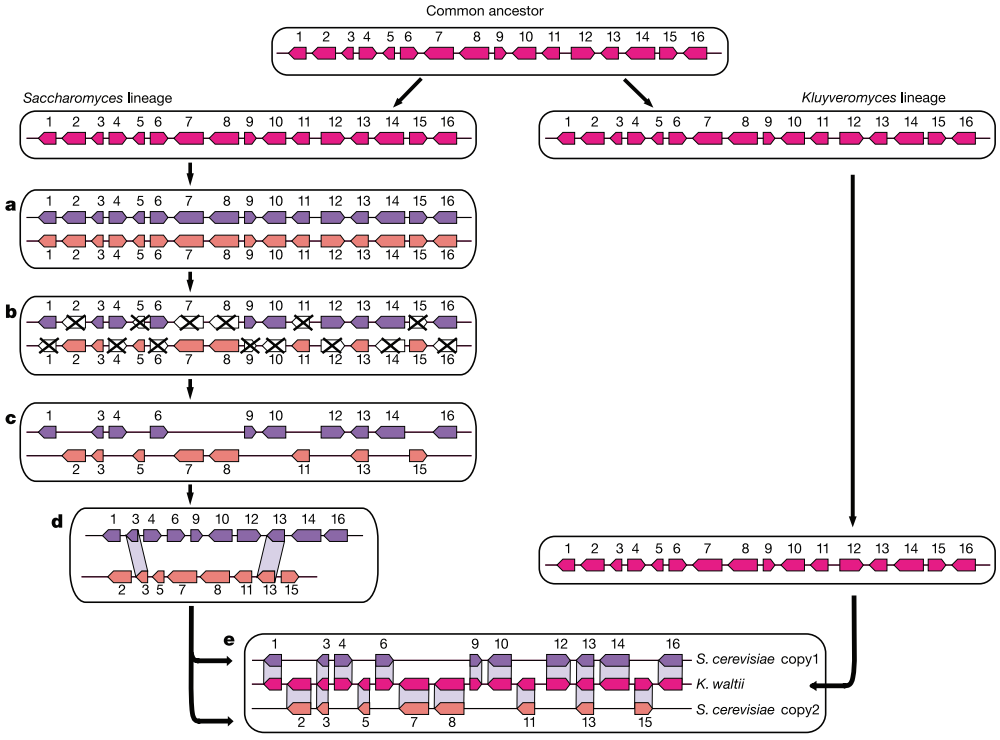


Figure 4.8. Blocs de synténie double-conservée entre le génome de la levure *K. waltii* (non dupliqué) et celui de *S. cerevisiae* (dupliqué). Suite à la duplication complète et aux pertes de gènes massives dues à la rediploïdisation, une région du génome de *K. waltii* correspond à deux régions du génome de *S. cerevisiae* avec un patron caractéristique d'alternance des gènes entre les deux régions. Figure tirée de (Kellis et al. 2004).

L'existence des synténies double conservées a été mise en évidence chez les poissons téléostéens, les levures et les plantes (Jaillon et al. 2004; Byrne and Wolfe 2005; Gordon et al. 2009; Van de Peer et al. 2009a). Elle est en revanche plus difficile à déceler pour les duplications

complètes très anciennes, comme les duplications 1R ou 2R, pour lesquelles les espèces non-dupliquées sont nécessairement très éloignées et le signal de synténie quasiment absent.

4.6. Le cas particulier de la duplication 3R

La duplication 3R se place à la base de l'arbre des téléostéens, postérieurement à la bifurcation des lépisostéiformes (lignée du lépisosté tacheté) mais antérieurement à la séparation entre la lignée du poisson zèbre et celle des poissons percomorphes (Figure 4.9). Elle est actuellement datée à environ 350 Ma (Van de Peer 2004). Elle concerne la très grande majorité des poissons osseux, puisque seules une quarantaine d'espèces d'actinoptérygiens dont le génome n'est pas dupliqué sont connues (Meyer and Van de Peer 2005). Il est possible que cette duplication ait joué un rôle moteur dans la radiation évolutive des téléostéens, qui comprennent plus de 25000 espèces différentes (Postlethwait et al. 2004; Vandepoele et al. 2004; Nelson 2006) ; cependant, cette hypothèse a été remise en question en raison de l'incertitude sur la date de la radiation des téléostéens, tantôt placée immédiatement après la duplication, vers 320 Ma, sur la base des estimations moléculaires, et tantôt bien plus tard, vers 200 Ma, sur la base du registre fossile (Meyer and Van de Peer 2005; Santini et al. 2009).

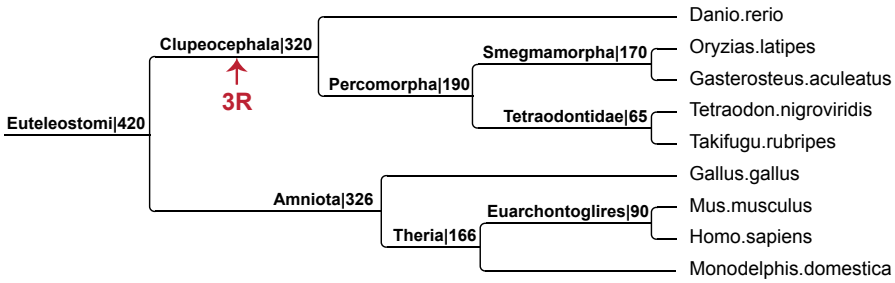


Figure 4.9. Arbre phylogénétique des génomes de poissons téléostéens séquencés et de quatre génomes amniotes de référence. La localisation de la duplication complète 3R est marquée dans l'arbre par une flèche rouge. Les principaux ancêtres de l'arbre sont indiqués avec leur âge consensus tel que fourni par la base de données Ensembl. Les longueurs de branches ne sont pas proportionnelles à l'âge.

Comme pour les duplications 1R et 2R des vertébrés, les premiers indices indiquant une duplication complète potentielle dans la lignée des téléostéens sont venus de leurs clusters de gènes *Hox*, qui sont au nombre de sept ou huit chez les téléostéens (Amores et al. 1998; Naruse et al. 2000). Postérieurement, un grand nombre d'autres gènes dupliqués ont été identifiés dans les génomes de poissons, dont la datation par analyse phylogénétique ou par horloge moléculaire place l'événement de duplication à la base de la radiation des téléostéens (Figure 4.10 ; (Taylor et al. 2003; Vandepoele et al. 2004; Meyer and Van de Peer 2005). Sans que ces observations soient une preuve absolue de l'existence d'une duplication complète, l'apparition conjointe d'un grand nombre de gènes dupliqués sur une période de temps restreinte est plus simplement expliquée par un événement unique de duplication complète que par de nombreux événements de duplications segmentales. La preuve finale d'une paléopolyploïdie spécifique aux téléostéens a été apportée par l'analyse des génomes du fugu (*Takifugu rubripes ;* (Christoffels et al. 2004) et surtout du tétraodon (*Tetraodon nigroviridis ;* (Jaillon et al. 2004), qui présente une

organisation typique d'un génome dupliqué avec des liens de paralogie chromosome à chromosome et une double-synténie conservée avec le génome humain.

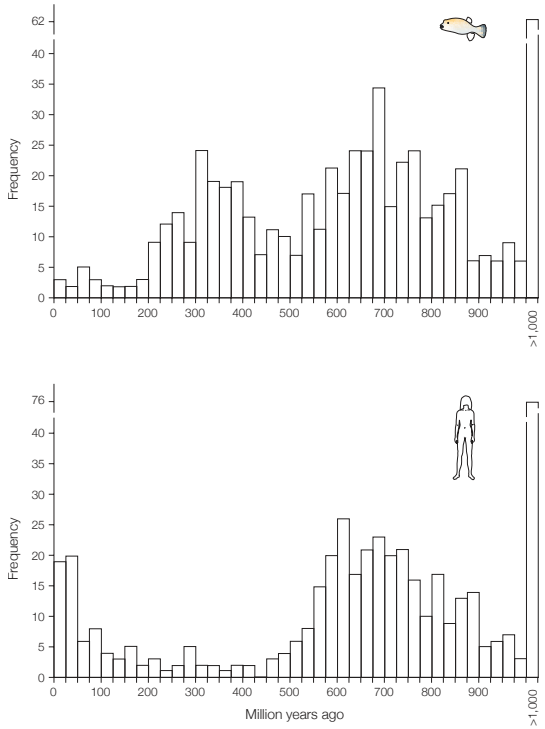


Figure 4.10. Distribution des âges des gènes paralogues dans le génome du fugu et de l'homme. Les âges sont obtenus par datation moléculaire basée sur la longueur des branches des arbres de gènes (sous une hypothèse d'horloge moléculaire). Dans le génome humain existe un grand nombre de paralogues datés d'environ 650 Ma correspondant aux duplications 1R et 2R. Dans le génome du fugu, on observe l'existence des ohnologues 1R et 2R mais également une deuxième vague de duplications vers 320 Ma correspondant aux ohnologues issus de la duplication 3R. Figure tirée de (Van de Peer 2004).

La duplication 3R est un cas d'étude intéressant, dans la mesure où il s'agit d'une duplication suffisamment récente pour être encore détectable dans les génomes qui en descendent, et en même temps assez ancienne pour pouvoir en observer les conséquences évolutives à long terme. Plusieurs génomes de téléostéens ont été séquencés et sont disponibles pour étudier cette duplication complète : le medaka (*Oryzias latipes*), l'épinoche (*Gasterosteus aculeatus*), le tétraodon (*Tetraodon nigroviridis*), le fugu (*Takifugu rubripes*) et le poisson zèbre (*Danio rerio*), sur lequel nous avons travaillé dans le cadre d'un consortium international. Cette grande masse de données fait de la duplication 3R un modèle de paléopolyploïdie se prêtant bien à l'analyse, et renseigne sur l'histoire évolutive des génomes de vertébrés suite aux duplications complètes.

Chapitre 5. Reconstructions de génomes ancestraux

Comprendre l'évolution de l'organisation des génomes et reconstruire la situation ancestrale sont souvent deux faces d'un même problème. En effet, les régions conservées entre espèces sont les traces de l'organisation du génome chez leur ancêtre ; les régions qui diffèrent sont celles qui ont été affectées par les réarrangements, duplications et délétions que l'on cherche à étudier. Reconstructions de l'information ancestrale et études des mécanismes d'évolution des génomes vont donc souvent de pair dans la littérature, les avancées de l'un ouvrant de nouvelles pistes pour l'autre. Bien que ce travail de thèse ne porte pas directement sur les reconstructions de génomes ancestraux, il s'appuie en grande partie sur l'exploitation des génomes ancestraux reconstruits par une méthode développée au laboratoire. Afin de remettre en contexte les méthodes et résultats détaillés dans ce manuscrit, nous présentons un bref état de l'art sur les reconstructions de génomes ancestraux dans la lignée des vertébrés, ainsi que les avantages et les limites de différentes méthodes proposées dans le cadre de l'étude de l'évolution de l'organisation des génomes depuis un ancêtre. Nous présentons ici uniquement les méthodes à jour s'appuyant sur les génomes entièrement séquencés, et leurs implications dans un cadre biologique : pour un comparatif complet des reconstructions de génomes sous un angle méthodologique et algorithmique, les lecteurs peuvent se référer à (Muffato and Roest Crollius 2008) et la thèse de Matthieu Muffato (2010).

5.1. Reconstructions de l'ordre ancestral de marqueurs

Dans la majorité des cas, on entend par « reconstruction du génome ancestral » l'inférence de l'organisation des gènes dans les chromosomes du génome de l'ancêtre. Les différentes approches utilisées pour étudier les points de cassure de réarrangements (cytogénétique, analyse de la synténie, optimisation combinatoire ; voir chapitre 3) ont toutes été historiquement utilisées pour reconstruire la configuration de ces marqueurs dans le génome ancestral. Les méthodes les plus performantes sont basées sur deux approches principales : l'approche combinatoire et l'analyse des adjacences conservées.

5.1.1. Approche combinatoire

Les approches combinatoires cherchant à trouver le scénario optimal de réarrangements transformant un génome en un autre, comme GRIMM (Tesler 2002)(voir paragraphe 3.5.3), permettent essentiellement d'étudier l'histoire évolutive écoulée entre deux génomes. Le génome ancestral correspond à l'une des étapes intermédiaires sur le chemin des

réarrangements séparant les génomes, mais il n'est pas possible de savoir laquelle avec uniquement deux génomes.

La méthode MGR (Bourque and Pevzner 2002) est un développement de GRIMM qui permet de prendre trois génomes modernes ou plus, et d'en reconstituer les génomes ancestraux : l'algorithme fait converger les deux génomes les plus proches en sélectionnant les réarrangements qui les rapprochent à la fois l'un de l'autre et des autres génomes, jusqu'à obtenir le « génome médian » considéré comme leur ancêtre commun, puis réitère le processus pour reconstruire le génome ancestral à chaque nœud de l'arbre. Le programme fournit ainsi en sortie les génomes ancestraux et un arbre phylogénétique des espèces, basé sur le nombre de réarrangements nécessaires sur chaque branche. MGR fonctionne via des heuristiques de parcimonie, et ne teste pas toutes les combinaisons possibles, si bien que la ou les solutions produites ne sont pas forcément optimales. Cette méthode a été utilisée pour étudier les taux de réarrangements entre lignées et, en creux, les régions de cassure (Bourque et al. 2004; Bourque et al. 2005; Murphy et al. 2005; Larkin et al. 2009). Plus récemment, le programme MGRA (Alekseyev and Pevzner 2009) a été développé pour proposer une solution optimale au problème de reconstruction du scénario de réarrangements et du génome ancestral pour un groupe de génomes modernes. Cette méthode se base sur la résolution de graphes d'adjacences intégrant les données de synténie de tous les génomes considérés à la fois, et est plus puissante que MGR. Cependant, cette méthode a pour l'instant été très peu utilisée pour étudier l'évolution de l'organisation du génome. Une troisième méthode, EMRAE (Zhao and Bourque 2009), a également été proposée : contrairement aux deux autres, elle se base sur un arbre phylogénétique fourni en entrée qui guide la reconstruction du scénario de réarrangements et des génomes ancestraux.

Ces méthodes ont l'avantage de produire à la fois un génome ancestral mais également toute la suite de réarrangements menant à chaque lignée moderne. Les différents événements (translocations, inversions, etc.) sont bien identifiés et les points de cassure de chaque événement successif sont connus. Par ailleurs, MGR et MGRA ne se basent pas sur une phylogénie fournie en entrée : au contraire, elles fournissent en sortie un arbre des espèces basé sur la distance en termes de réarrangements entre les génomes, calculé indépendamment de l'arbre des espèces connu. Cet arbre peut ensuite être comparé à l'arbre réel pour juger de la pertinence de la reconstruction.

En revanche, ces méthodes se heurtent à un certain nombre de problèmes. Le premier est qu'elles sont très sensibles à la définition correcte des blocs de synténie, et que leurs résultats peuvent être perturbés par une mauvaise délimitation des blocs due à des micro-réarrangements réels ou à des artefacts d'annotation ou d'assemblage dans les génomes modernes. L'algorithme GRIMM-Synteny (Pevzner and Tesler 2003b) a été spécialement développé pour délimiter des blocs de synténie robustes, correspondant à des segments entièrement conservés entre plusieurs génomes si l'on passe outre les microréarrangements à l'intérieur de ces blocs. Initialement, GRIMM-Synteny a été conçu pour une utilisation avec GRIMM, mais il améliore également sensiblement les performances de MGR. Un deuxième problème soulevé par ces approches combinatoires est que dans certains cas, les génomes ancestraux obtenus par combinatoire ne sont pas congruents avec les génomes ancestraux

inférés par cytogénétique, moins précis du point de vue de la résolution des marqueurs mais bénéficiant de données de conservation très robustes sur des dizaines d'espèces (plus de 80 dans le cas des mammifères euthériens)(Froenicke et al. 2006). Les différences sont en partie liées au nombre d'espèces réduit pouvant être pris en compte dans les méthodes de combinatoire pour des raisons de faisabilité du calcul : augmenter le nombre d'espèces pris en compte permet de régler la plupart des incohérences entre les résultats des deux approches (Figure 5.1)(Bourque et al. 2006). Mais une partie du problème pourrait également être lié au modèle d'évolution implémenté dans le programme : ce modèle n'autorise qu'un certain nombre d'événements qui contraignent la solution offerte par la reconstruction. Modifier le modèle d'évolution peut donc, en principe, donner une reconstruction du génome ancestral différente. Enfin, ces méthodes ne fonctionnent qu'avec des marqueurs uniques présents en une copie dans chaque génome considéré. Ainsi, ces algorithmes sont inapplicables dans les faits pour reconstruire les ancêtres de génomes ayant subi une duplication complète récente, et pour tenir compte des très nombreuses duplications segmentales impliquant des gènes.

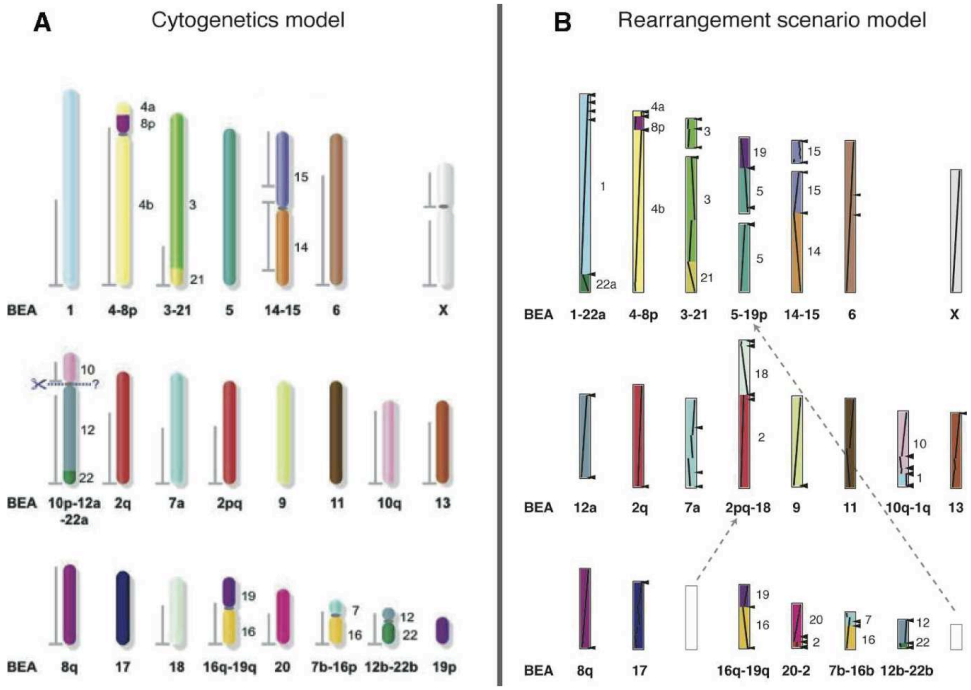


Figure 5.1. Architecture génomique putative de l'ancêtre des mammifères boreoeuthériens reconstruite (A) à partie des données cytogénétiques (Froenicke et al. 2006), ou (B) à partir du scénario de réarrangements entre les génomes humain, souris, rat, chien, vache et cochon (Murphy et al. 2005) en utilisant le programme MGR. Les deux modèles présentent deux différences majeures, notamment l'existence de deux chromosomes inférés par cytogénétique qui sont reconstruits comme part d'un autre chromosome ancestral par le scénario de réarrangements (marqués par des pointillés à droite). On peut noter d'autres différences plus mineures, comme la fusion des chromosomes 1 et 22a en un seul chromosome dans le modèle (B). Figure tirée de (Bourque et al. 2006).

5.1.2. Analyse des adjacences conservées

Dans ce second type d'approches, il s'agit d'identifier des marqueurs qui sont dans la même configuration (adjacents et dans la même orientation) dans deux génomes ou plus. Cette configuration est alors supposée ancestrale ; l'ordre des marqueurs peut être reconstruit de proche en proche en utilisant les informations de conservation à travers les génomes de plusieurs espèces.

Deux méthodes principales implémentent ce type de reconstructions ancestrales à partir des données génomiques : la méthode inferCARs (Ma et al. 2006), et celle proposée par Chauve et Tannier (Chauve and Tannier 2008). Les deux méthodes diffèrent sur les données qu'elles prennent en entrée : inferCars se base sur un alignement multiple de génomes modernes pour en extraire un ensemble de blocs de séquences alignées qui sont dans le même ordre et la même orientation dans tous les génomes, alors que la méthode de Chauve et Tannier peut utiliser des marqueurs présentant des « caractéristiques conservées » plus ou moins relaxées : blocs de gènes dans le même ordre et la même orientation entre les génomes, blocs génomiques ayant le même contenu en gènes, sans que l'ordre de ceux-ci soit préservé, etc. Dans les deux méthodes, la configuration ancestrale de ces blocs entre eux est déduite en tenant compte du nombre d'espèces soutenant chaque adjacence de blocs comme étant d'origine ancestrale, et de la phylogénie des espèces, supposée connue : les deux méthodes calculent alors une probabilité d'adjacence ancestrale pour chaque configuration possible. En dernière étape, inferCARs choisit les adjacences comme étant correctes en fonction de leur score, des plus probables aux moins probables, pour rabouter les blocs en pseudo-chromosomes. La méthode de Chauve et Tannier, en revanche, implémente une approche tirée de la théorie des graphes (résolution du « problème des 1 consécutifs »), qui cherche à regrouper les marqueurs d'une manière qui concilie le maximum d'adjacences, en supprimant si nécessaire des adjacences en commençant par celles de plus faible score (considérées comme ambigües) : ce regroupement ne fournit pas forcément l'ordre des blocs dans les pseudo-chromosomes, simplement le contenu en blocs de ceux-ci.

L'avantage de ces méthodes par rapport aux approches combinatoires comme MGR ou MGRA est qu'elles ne se basent pas sur un modèle d'évolution prédéfini pour remonter à l'organisation ancestrale (taux d'inversions, translocations, etc.) : elles évitent ainsi une certaine circularité de raisonnement, où le génome ancestral reconstruit dépend en partie du modèle d'évolution inféré pour le phylum. Par ailleurs, les résultats obtenus par ces approches sont plus cohérents avec les liaisons ancestrales prédites par cytogénétique que les approches combinatoires. En contrepartie, ces méthodes proposent uniquement une prédiction du génome ancestral, et pas du scénario évolutif entre l'ancêtre et les espèces modernes.

Les deux méthodes les plus à jour présentées ici ont cependant un certain nombre de limites : la première est qu'elles ne peuvent considérer qu'un nombre d'espèces assez limité, sans quoi les ambigüités et la complexité du problème augmentent rapidement et le problème devient insoluble. Ensuite, dans le cas d'inferCARs, le génome ancestral reconstruit dépend de l'arbre des espèces fourni en entrée, ce qui a longtemps été sujet de débats dans le cas des mammifères notamment, même si la phylogénie est essentiellement résolue aujourd'hui pour les

grands taxons (Alekseyev and Pevzner 2009). Enfin, comme les méthodes basées sur la combinatoire, ces approches sont confrontées à un problème d'équilibre entre résolution des marqueurs utilisés, complexité du problème, et continuité des pseudo-chromosomes obtenus dans la reconstruction. Plus les marqueurs sont nombreux et précis, plus la probabilité d'observer des incohérences réelles ou artéfactuelles (erreurs d'assemblage ou d'annotation dans les génomes) est importante, et plus il est nécessaire d'éliminer de l'information non fiable pour résoudre le problème : cette élimination se fait au détriment d'adjacences au score relativement faible mais pourtant vraies, qui se trouvent justement dans les régions les plus difficiles à reconstruire, ce qui limite la possibilité de regrouper des blocs en pseudo-chromosomes. Or, pour étudier les mécanismes perturbant l'organisation des gènes, il est nécessaire d'utiliser des reconstructions de génomes ancestraux dont la résolution est de l'échelle du gène, et idéalement fournissant l'ordre de ces gènes et non uniquement le contenu en gènes des pseudo-chromosomes. Enfin, à nouveau comme les méthodes combinatoires, ces analyses d'adjacences ne considèrent que les marqueurs présents en une copie dans chaque génome, si bien qu'elles ne sont pas applicables aux génomes dupliqués récemment.

AGORA, la méthode développée au laboratoire et utilisée pour produire les données initiales exploitées dans ce travail de thèse, est également une méthode basée sur les adjacences conservées entre espèces, qui répond à une partie de ces problèmes. Cette méthode est présentée au chapitre 9.

5.2. Reconstructions de séquences ancestrales

En parallèle des méthodes de reconstruction de la structure des génomes ancestraux, d'autres méthodes ont été développées pour reconstruire la séquence du génome de l'ancêtre. Les différentes méthodes fonctionnent toutes sur le même principe général : la séquence d'une région du génome ancestral est déduite en utilisant un alignement multiple des séquences modernes, et un arbre phylogénétique qui guide la reconstruction de la séquence ancestrale (soit fourni en entrée, soit lui-même déduit de l'alignement multiple). Selon les méthodes, l'algorithme reconstruit la séquence pour chaque ancêtre dans l'arbre phylogénétique en remontant progressivement dans l'arbre depuis les espèces modernes (par exemple, Paten et al. 2008), ou reconstruisent la séquence seulement pour un ancêtre donné (comme Blanchette et al. 2004).

Reconstruire l'état ancestral à partir d'un alignement des séquences modernes pose deux difficultés majeures. La première, la plus évidente, est de déterminer la base nucléotidique ancestrale dans le cas fréquent où des substitutions de bases ont eu lieu dans certaines lignées. La plupart des algorithmes récents infèrent la séquence ancestrale en utilisant des méthodes probabilistes de type maximum de vraisemblance ou bayésiennes, qui sont plus puissantes que les approches plus simples comme le maximum de parcimonie (Zhang and Nei 1997; Krishnan et al. 2004) et qui fournissent en outre un score de confiance pour chaque base ancestrale reconstruite. La seconde difficulté est de gérer les indels. Idéalement, ceux-ci doivent expressément être annotés comme une insertion ou une délétion dans une lignée et la méthode doit être capable de traiter des indels nichés ou superposés, un problème non trivial à résoudre

à partir d'un alignement multiple. Les approches pour résoudre ce problème diffèrent : certaines méthodes appliquent un raisonnement de parcimonie qui cherche à minimiser le nombre d'événements (Blanchette et al. 2004; Snir and Pachter 2011), alors que d'autres implémentent des modèles probabilistes (Kim and Sinha 2007; Paten et al. 2008).

Si la plupart des méthodes de reconstruction de séquences ancestrales s'intéressent à la séquence de protéines, plusieurs ont été proposées pour reconstruire des séquences nucléotidiques, en général assez courtes (séquence d'un gène en particulier, d'un élément transposable, etc.) (Huelsenbeck and Bollback 2001; Blanchette et al. 2004; Ashkenazy et al. 2012). En revanche, seules deux d'entre elles ont été à notre connaissance implémentées et testées pour la reconstruction de larges régions au-delà d'un seul gène, voire même du génome dans son ensemble (Blanchette et al. 2004; Paten et al. 2008). Ces travaux sont particulièrement intéressants pour nous, puisqu'ils se sont tous deux intéressés à la reconstruction de la séquence du génome ancestral Boreoeutheria, qui est l'ancêtre des primates, rongeurs, carnivores et ongulés sur lequel notre étude porte également. La méthode proposée par Blanchette et al. (2004) tout comme Ortheus (Paten et al. 2008) ont été validées en reconstruisant des génomes « ancestraux » à partir de simulations d'évolution, et parviennent à reconstruire des séquences ancestrales avec moins de 3% de bases erronées (et moins de 1% dans les régions sans éléments répétés) à partir des génomes modernes entièrement séquencés disponibles dans les bases de données à l'époque de l'étude (20 pour Blanchette et al., 33 pour Paten et al.). Ainsi, il est possible de reconstruire les séquences ancestrales avec une très grande fiabilité, ce qui permet d'inférer des caractéristiques locales de la séquence ancestrale (%GC, contenu en éléments répétés, etc.) ou des informations fonctionnelles (fonction putative des gènes ancestraux, structure des protéines, etc.).

Cependant, ces méthodes ont deux désavantages majeurs : tout d'abord, elles sont assez dépendantes de la phylogénie des espèces utilisée. La reconstruction est plus puissante sur les nœuds de l'arbre suivis par des radiations rapides, car les nombreuses lignées qui émergent après le nœud représentent autant de comparaisons susceptibles d'éclairer la situation ancestrale. Mais si les branches courtes à la base de l'arbre sont mal résolues, la séquence ancestrale reconstruite peut contenir des erreurs, comme relevé par les auteurs des deux méthodes. Enfin, et c'est sans doute le point le plus important, ces méthodes ne fonctionnent que dans les régions où il est possible d'obtenir un alignement multiple des séquences de bonne qualité. La fiabilité de la reconstruction décroit très vite quand le nombre de séquences dans l'alignement diminue (Figure 5.2) ; ainsi, les reconstructions de génomes ancestraux ne sont fiables (ou même n'existent) que dans les régions alignables entre plus d'une dizaine d'espèces. Or, si les gènes sont généralement alignables entre espèces de manière relativement aisée, il n'en est pas toujours de même pour les régions non codantes. Une reconstruction du génome ancestral de Boreoeutheria, reconstruit avec Ortheus, est disponible sur la base de données Ensembl, mais la plus grande partie des intergènes est manquante dans ce génome ancestral. Il n'est donc pas possible d'utiliser ce type de données pour inférer des caractéristiques ancestrales dans la majorité des régions non-codantes du génome.

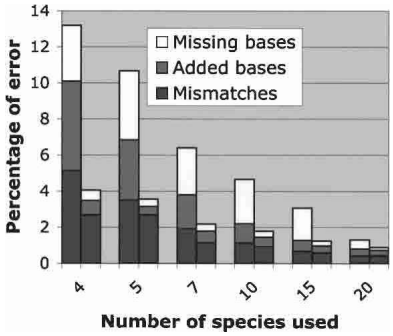


Figure 5.2. Fraction de bases du génome de l'ancêtre des mammifères boreoeuthériens simulé incorrectement reconstruites par la méthode de Blanchette et al. (2004), en fonction du nombre d'espèces modernes utilisées pour la reconstruction. Pour chaque reconstruction avec un nombre d'espèce donné (abscisse), le nombre d'erreurs est donné pour toutes les bases (colonne de gauche) ou seulement pour les bases dans des régions non répétées du génome (colonne de droite). Figure tirée de (Blanchette et al. 2004).

Dans le cadre de notre étude, nous avons utilisé les reconstructions de génomes ancestraux en particulier pour étudier l'apparition de points de cassure de réarrangements. L'une des questions les plus intéressantes et informatives est : y a-t-il des caractéristiques locales de la séquence, dans l'état ancestral, qui sont associées à une augmentation du taux de cassure ? Les cassures étant essentiellement intergéniques, les reconstructions de séquences ancestrales ne permettent pas de répondre à cette question d'un manière globale et quantifiée sur l'ensemble du génome : il nous a donc fallu développer de nouvelles approches pour inférer certaines caractéristiques génomiques locales chez l'ancêtre.

Chapitre 6. Problématique

L'évolution de l'organisation des génomes de vertébrés est le produit d'événements qui modifient le contenu en éléments fonctionnels du génome (duplications, délétions, créations de nouveaux éléments) et d'autres qui réarrangent et réordonnent ces éléments. Les forces qui gouvernent cette évolution sont encore mal comprises ; l'organisation des génomes évolue-t-elle majoritairement de façon aléatoire et neutre, ou est-elle au contraire soumise à de fortes pressions fonctionnelles et sélectives qui contraignent les événements pouvant être tolérés et fixés ? Ces événements se produisent-ils au hasard, ou leur probabilité varie-t-elle selon les phylums et les régions du génome ? Peut-on détecter les signatures de ces dynamiques évolutives dans les génomes modernes, et en tirer une information pertinente pour notre compréhension du fonctionnement du génome ? Autant de questions qui restent encore en suspens à ce jour.

La génomique comparative est un outil de choix pour tenter de répondre à ces questions : grâce aux outils informatiques désormais disponibles pour comparer de nombreux génomes à haute résolution, il est possible d'inférer avec précision l'histoire des événements évolutifs qui modèlent les génomes, et leurs conséquences sur l'organisation actuelle.

Au cours de cette thèse, nous nous sommes intéressés à deux questions. La première est la distribution des points de cassure de réarrangements évolutifs dans les génomes de mammifères. Les méthodes existantes pour étudier ces cassures ont mené à différents modèles d'évolution du génome. Certains proposent que le taux de réarrangements évolutifs varie le long du génome parce que la probabilité qu'un événement de réarrangement se produise n'est pas homogène ; d'autres proposent que les variations du taux de réarrangements évolutifs reflètent en creux l'intensité d'une contrainte sélective qui s'exerce sur l'organisation locale des éléments fonctionnels du génome. Or, ces deux possibilités traduisent des conceptions entièrement différentes de l'évolution des génomes, l'une neutraliste et l'autre sélective. Nous avons donc cherché à aborder ce problème par un nouvel angle, en se plaçant du point de vue de l'organisation du génome ancestral. Modéliser quelles propriétés ont influencé significativement la probabilité qu'un réarrangement se produise et soit maintenu permettrait de renseigner sur les forces mises en jeu lors de la formation et de la fixation évolutive des réarrangements.

Dans un second temps, nous nous sommes penchés sur l'impact des duplications complètes du génome sur les génomes de vertébrés, dans le cadre du projet de séquençage du génome du poisson zèbre porté par D. Stemple et K. Howe au Sanger Institute (Royaume-Uni). Le but de ce travail, mené en parallèle du projet principal, était de donner une perspective générale sur le génome de cette espèce modèle sous l'angle de la génomique comparative. Nous avons centré notre étude sur la duplication 3R se trouvant à la base de la radiation des téléostéens, puisque le génome du poisson zèbre est le génome ayant divergé le plus précocement après cette

duplication complète parmi les téléostéens actuellement séquencés. Nous avons cherché à documenter les conséquences de cette duplication sur le contenu en gènes et leur organisation dans le génome du poisson zèbre par rapport à la fois aux génomes amniotes non dupliqués, mais également aux autres génomes téléostéens séquencés, afin de détecter des caractéristiques spécifiques du poisson zèbre.

Les résultats obtenus au cours de cette thèse seront donc organisés en deux parties distinctes, l'une sur les mécanismes de réarrangements évolutifs dans les génomes de vertébrés (Partie III), et l'autre sur l'impact de la duplication 3R dans le génome du poisson zèbre (Partie IV). Ces deux parties seront suivies d'une discussion à la fin de ce manuscrit.

Deuxième partie
Matériel et Méthodes

Chapitre 7. Origine des données

7.1. Génomes et arbres de gènes

Les contenus en gènes des différents génomes utilisés ont été téléchargés sur la base de données Ensembl (http://www.ensembl.org; (Flicek et al. 2010)). Cette base met à disposition de la communauté scientifique des annotations de génomes séquencés obtenues grâce à une suite de programmes automatisés, et vérifiées manuellement dans certains cas (Wilming et al. 2008) ; Ensembl fournit également une intégration de nombreuses caractéristiques des gènes, qui peuvent être aisément téléchargées via l'interface d'interrogation Biomart, l'interface de programmation en Perl API, ou directement sur un serveur ftp. Pour les analyses de réarrangements, nous avons utilisé les génomes de la version 57 de la base datant de mars 2010 ; cette version contient 46 génomes de vertébrés ainsi que 5 invertébrés modèles. Pour les analyses sur le génome du poisson zèbre, conduites ultérieurement, nous avons mis à jour les données et utilisé la version 63 de la base, mise à disposition de la communauté en juin 2011. Les gènes utilisés sont les gènes codants pour des protéines.

Les relations d'orthologie entre gènes ont également été téléchargées sur Ensembl. La base utilise une suite d'algorithmes nommée Compara pour construire les arbres de gènes à travers toutes les espèces de la base (Vilella et al. 2009). Compara prend en entrée des familles de protéines homologues entre les espèces obtenues par BlastP sur le critère de leur similarité de séquence. L'approche consiste à construire un alignement multiple des protéines de chaque famille en utilisant une combinaison d'aligneurs (MCoffee2 (Wallace et al. 2006)), qui est ensuite transformé en un alignement des séquences nucléotidiques codantes (CDS). L'alignement multiple est utilisé pour construire un arbre phylogénétique des gènes homologues en utilisant l'algorithme TreeBeST, qui calcule un arbre consensus à partir de cinq méthodes différentes. L'arbre de gènes est ensuite réconcilié avec la phylogénie des espèces fournie en entrée, en introduisant des nœuds de duplication et des pertes de gènes dans l'arbre, avec un score de robustesse à chaque nœud. Un exemple d'arbre de gènes réconcilié avec la phylogénie des espèces est fourni en Figure 7.1.

La version 57 d'Ensembl contient 886 547 gènes codant pour des protéines dans 51 espèces, pour une moyenne de 17 383 gènes par espèce. Parmi ceux-ci, 96,8% sont inclus dans des arbres de gènes reconstruits par Compara ; ces arbres définissent 43 ancêtres aux différents nœuds, qui comptent en moyenne 19 361 gènes.

La version 63 quant à elle contient 927 950 gènes codant pour des protéines dans 53 espèces, pour une moyenne de 17 508 par espèce. Parmi ces gènes, 95,5% sont inclus dans des arbres de gènes reconstruits par Compara, qui définissent 44 ancêtres aux différents nœuds contenant en moyenne 25 653 gènes.

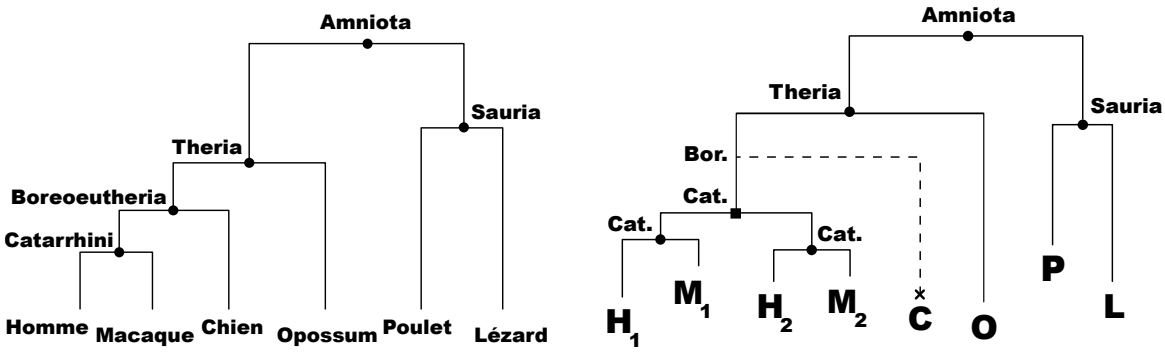


Figure 7.1. Exemple d'arbre des gènes (à droite) réconcilié avec l'arbre des espèces (à gauche). Les cercles indiquent les nœuds de spéciation, les carrés les duplications. Les pertes de gènes n'apparaissent pas dans les phylogénies de gènes, mais peuvent être déduites de la phylogénie des espèces (ici, la perte du gène chez le chien est figurée par une branche en pointillés). Figure tirée de (Muffato 2010).

7.2. Caractéristiques de séquence

7.2.1. Taux de GC

Le taux de GC des séquences intergéniques a été calculé à partir des séquences complètes de génomes téléchargées sur le serveur ftp d'Ensembl. Nous utilisons un script en Python qui croise une liste d'intervalles génomiques (gènes, intergènes, etc.) avec la séquence et compte le pourcentage de bases G et C sur le brin direct. Ce script peut retourner le taux de GC complet ou le taux de GC après soustraction des régions répétées si la séquence fournie en entrée différencie ces derniers par un masquage « soft » (les éléments répétés sont indiquées en lettres minuscules, les éléments non répétés en majuscules).

7.2.2. Îlots CpG

Les îlots CpG sont des régions du génome où le taux de dinucléotides CG est particulièrement élevé par rapport au reste du génome : en effet, les cytosines des dinucléotides CG sont fréquemment méthylées, et ont alors tendance à muter vers une thymine par désamination spontanée. Les dinucléotides CG sont donc rares dans le génome, sauf aux endroits où la méthylation joue un rôle et où les transitions C vers T sont contre-sélectionnées, ou bien dans les régions où la méthylation est absente pour des raisons fonctionnelles. La méthylation des îlots CpG est une marque épigénétique de chromatine compactée et inactive : les régions promotrices des gènes sont fréquemment riches en îlots CpG, qui sont méthylés lorsque le gène est non exprimé et déméthylés lorsqu'il est actif (Deaton and Bird 2011).

La liste des îlots CpG détectés dans le génome humain (version h19) a été téléchargée sur le Genome Browser de l'Université de Californie Santa-Cruz (UCSC ; http://genome.ucsc.edu/), qui comporte une large base de données génomiques comprenant des tables de positions d'éléments caractéristiques pour une cinquantaine d'espèces, ainsi que des informations de similarité entre les génomes (alignements multiples des séquences, etc.). Les îlots CpG ont été identifiés en

parcourant le génome humain pour trouver des régions qui satisfont les conditions suivantes (Gardiner-Garden and Frommer 1987) :

- taux de GC supérieur à 50%
- longueur supérieure à 200 pb
- ratio CpG observés/attendus supérieur à 0,6 ; le ratio est calculé avec la formule suivante, où N est la longueur de la région, CpG_{obs} le nombre de CpG observés, C_{obs} le nombre de C dans la séquence et G_{obs} le nombre de G :

$$R = \frac{CpG_{obs} * N}{C_{obs} * G_{obs}}$$

Les îlots CpG ainsi détectés sont au nombre de 28 691 dans la version hg19 du génome humain ; ils mesurent en moyenne 761 bp, et couvrent au total 0,75% du génome.

7.2.3. Eléments répétés et duplications segmentales

De même que les îlots CpG, la liste des positions des éléments répétés non-codants du génome humain a été téléchargée sur le site de l'UCSC. Ces éléments répétés ont été identifiés par RepeatMasker (A. Smit et al. ; http://www.repeatmasker.org/), un algorithme qui utilise des alignements du génome contre lui-même et des modèles d'éléments transposables pour annoter les séquences répétées du génome. RepeatMasker détecte dans le génome les différentes variations des familles d'éléments transposables et des régions de faible complexité à partir de la base de modèles qui lui est fournie, puis cherche à étendre ces régions répétées à partir d'alignements locaux avec d'autres régions du génome (algorithme de Smith-Waterman). Il faut noter que RepeatMasker ne peut pas annoter un élément répété sans modèle correspondant dans la base, uniquement sur le critère d'alignement : les duplications segmentales ne sont donc pas annotées par RepeatMasker (voir ci-dessous).

Le génome humain est constitué à près de 51% d'éléments répétés, correspondant pour 20% à des rétrotransposons de la famille des LINEs, 13% à des rétrotransposons de la famille des SINEs, 8% à des rétrotranposons à LTR et 3% à des transposons à ADN (Lander et al. 2001). Dans la version hg19 du génome humain, RepeatMasker détecte 5 298 130 éléments répétés, d'une longueur moyenne de 277 pb, représentant 50,59% des bases du génome.

Les duplications segmentales, ou LCR (Low-Copy Repeats) sont traditionnellement définies comme des régions du génome d'une longueur supérieure à 1 kb et identiques à plus de 90% (Bailey et al. 2002). Ces séquences répétées ne correspondant pas à des éléments transposables et ne peuvent pas être détectées par RepeatMasker, qui nécessite un modèle pour détecter les occurrences de chaque élément. Nous avons utilisé les annotations de duplications segmentales disponibles sur le site de l'UCSC, détectées par la méthode dite de « fuguisation » (un terme rappelant la compaction du génome de *Takifugu rubripes*, très pauvre en éléments répétés)(Bailey et al. 2001) : cette méthode consiste à éliminer les éléments répétés annotés par RepeatMasker de la séquence en « épissant » virtuellement les séquences de part et d'autre, puis à aligner ce génome « sans éléments répétés » contre lui-même pour identifier des séquences de forte identité. Grâce à cette méthode, l'alignement peut se faire sans que le signal ne soit brouillé

par les nombreux éléments transposables qui perturbent les alignements, et permet de détecter des grandes régions d'homologie qui correspondent aux duplications segmentales. Les éléments répétés sont ensuite réinsérés dans la séquence à leur place d'origine. Cette méthode détecte 51 599 duplications segmentales dans le génome humain (hg19), d'une longueur moyenne de 13 192 pb, et couvrant au total 5,73% du génome.

7.2.4. Taux d'évolution des gènes

Les taux de mutations synonymes (dS) et non synonymes (dN) entre les gènes orthologues de l'homme et de la souris ont été obtenus sur la base de données Ensembl. Pour chaque paire de gènes orthologues, les séquences protéiques ont été alignées puis l'alignement est rétrotraduit en alignement nucléotidique. Les mutations de bases qui modifient l'acide aminé intégré à la chaine polypeptidique sont comptées comme non synonymes, et leur nombre est ramené au nombre de mutations non synonymes possibles dans la séquence, ce qui donne le ratio dN (également noté Ka) ; de même, le nombre de différences observées qui ne changent pas la séquence protéiques est ramené au nombre possible de mutations synonymes, pour donner le ratio dS (ou Ks). Le ratio dS est considéré comme une estimation du taux de mutations neutres se produisant dans une séquence codante ; il permet d'estimer si une séquence évolue rapidement (point chaud de mutations) ou lentement (point froid) par rapport à une autre. Le ratio dN/dS renseigne sur la proportion de mutations ayant des conséquences au niveau de la séquence de la protéine par rapport au taux neutre attendu dans la région : si le ratio est proche de 1, le gène évolue de manière neutre (mutations non synonymes aussi probables que les mutations synonymes) ; plus le ratio est petit, et plus le gène est sous pression de sélection négative (les mutations non synonymes sont très peu tolérées par rapport aux mutations synonymes) ; si le ratio est supérieur à 1, on considère généralement qu'il s'agit d'une signature de sélection positive (les mutations modifiant la séquence codante ont plus de chances d'être fixées que les mutations neutres).

Le modèle de calcul des ratios dN et dS utilisé par Ensembl est codeml, un modèle de maximum de vraisemblance intégré dans la librairie d'outils PAML (Yang 1997), avec les paramètres *model* à 0 et *NSsites* à 0 également.

7.3. Taux de recombinaison dans le génome humain

Les taux de recombinaison utilisés dans cette étude proviennent de la carte génétique du génome humain créée à partir des données de la phase II du projet HapMap (Frazer et al. 2007), disponible sur le site du projet HapMap (http://hapmap.ncbi.nlm.nih.gov/downloads/). Cette carte a été construite en dénombrant le nombre moyen de crossovers se produisant au cours de la méiose entre des marqueurs connus, conduisant à la transmission conjointe ou non des marqueurs à la descendance. Ce nombre de crossovers permet de calculer la distance génétique en centiMorgans entre deux marqueurs. En ramenant la distance génétique entre les marqueurs à la distance physique en paires de bases qui les séparent dans le génome, il est possible de calculer le taux de recombinaison local moyen en cM/Mb. Cette carte génétique se base sur plus

de 3,1 millions de SNP (Single Nucleotide Polymorphisms, polymorphismes d'une seule base) et a initialement été établie pour la version hg18 du génome humain. Nous avons mis à jour cette carte du taux de recombinaison en transférant les coordonnées des marqueurs vers la version hg19 du génome humain (« liftover »).

7.4. Eléments conservés non-codants

Les éléments conservés non-codants utilisés au cours de ce travail correspondent aux éléments détectés par la méthode GERP (Cooper et al. 2005). GERP (Genomic Evolutionary Rate Profiling) se base sur un alignement multiple de génomes qu'il parcourt colonne par colonne pour détecter des colonnes qui contiennent moins de substitutions qu'attendu au hasard, et donc potentiellement sous sélection négative. Le nombre de substitutions attendu au hasard dans une colonne de l'alignement est calculé en sommant le taux moyen de substitutions neutres par site dans chaque espèce représentée dans la colonne (les gaps ne sont donc pas pris en compte ; Figure 7.2). Un élément conservé est détecté lorsque la somme du déficit de substitutions sur plusieurs colonnes consécutives dépasse un seuil. Ce seuil est choisi à partir d'une permutation des colonnes de l'alignement, qui contient donc des colonnes avec un défaut de substitutions mais qui ne sont consécutives que par hasard : cette permutation permet de calculer, pour chaque seuil possible, le nombre de colonnes détectées comme se trouvant dans un élément conservé. On choisit comme seuil le déficit de substitutions nécessaire pour que le nombre de colonnes détectées comme conservées dans la permutation correspondent à 5% du nombre de colonnes détectées comme conservées dans l'alignement réel (soit un taux de faux positifs de 0,05). Les éléments conservés ainsi détectés sont ensuite fusionnés s'ils sont séparés par moins de 10 colonnes.

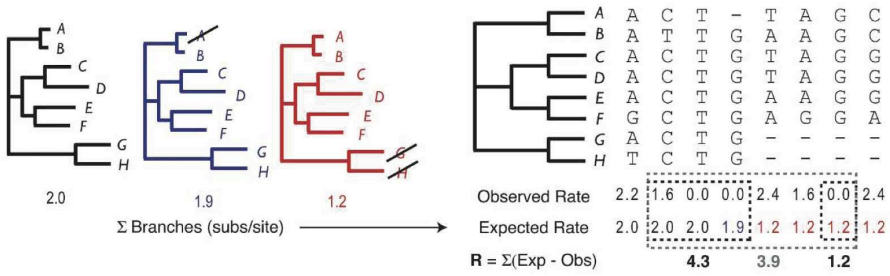


Figure 7.2. Fonctionnement de la méthode GERP. Pour chaque colonne de l'alignement multiple prise indépendamment, on calcule le nombre de substitutions neutres attendues en sommant la longueur des branches de l'arbre uniquement pour les espèces représentées dans la colonne (arbre, noir, bleu ou rouge dans l'exemple, selon les espèces manquantes à chaque colonne). On compare alors le taux de substitution observé à celui attendu, pour détecter des colonnes avec un taux plus faible qu'attendu. Si la somme du déficit de substitutions sur plusieurs colonnes successives dépasse un seuil fixé, un élément conservé est détecté (encadré par des pointillés gras). Les éléments proches sont ensuite fusionnés (pointillés fins). Figure tirée de (Cooper et al. 2005).

Les éléments GERP disponibles sur la base de données Ensembl version 57 sont calculés sur la base d'un alignement des génomes de 33 génomes euthériens, comprenant les 28 génomes utilisés pour l'estimation des caractères ancestraux du génome de l'ancêtre Boreoeutheria (voir

Partie III). Nous avons croisé la liste de ces éléments avec les annotations de gènes des différentes espèces de l'alignement afin de retirer tous les éléments conservés chevauchant un gène annoté dans au moins une espèce, et de ne conserver que les éléments intergéniques non-codants. Ces éléments ont ensuite été reportés dans chaque génome pour calculer le taux de séquence conservée pour chaque intergène.

Dans le génome humain, les séquences conservées intergéniques couvrent 88 358 657 pb, soit 5,1% de l'ensemble de la séquence intergénique. Le pourcentage de bases conservées détectées dépend de la profondeur de séquençage et la qualité de l'assemblage du génome : ainsi, dans le génome de l'écureuil (*Spermophilus tridecemlineatus*) séquencé avec une couverture de 2x, seules 3 626 120 pb sont détectées comme conservées, soit 2,3% de la séquence intergénique disponible.

7.5. Blocs de régulation génomique dans le génome humain

Au cours de ce travail, nous avons utilisé des prédictions de gènes-cible de blocs de régulation génomique (GRB) dans le génome humain communiquées par B. Lenhard et A. Akalin. Ces blocs correspondent à des régions du génome maintenues en synténie dans différentes espèces par l'action de la sélection négative qui préserve l'organisation en cis d'un gène cible et de ses séquences de régulation, qui peuvent se trouver à grande distance du gène (Figure 7.3). Les GRB connus sont caractérisés par une haute densité en séquences conservées non-codantes, sont conservés en synténie sur de longues distances évolutives (homme-poisson zèbre), contiennent des marques de séquences promotrices (îlots CpG, etc.) et ont fréquemment pour gène-cible des gènes du développement ou des facteurs de transcription de manière générale (Woolfe et al. 2005; Pennacchio et al. 2006; Kikuta et al. 2007). La méthode de prédiction des gènes-cible de GRB est basée sur le raisonnement détaillé dans (Engstrom et al. 2008) : l'algorithme recherche, au sein de blocs de gènes conservés en synténie entre l'homme et le poisson zèbre, le meilleur gène candidat qui serait la cible du GRB. Les critères de détection du meilleur candidat se font sur la densité locale en éléments conservés non-codants, en îlots CpG, et sur le termes Gene Ontology attachés aux différents gènes dans la région conservée ; l'algorithme affine ses paramètres de détection par apprentissage à partir d'une liste de gènes-cible connus, puis fournit une liste de gènes-cible de GRB putatifs. Cette méthode détecte 864 gènes-cible putatifs dans le génome humain, soit 3,8% des gènes.

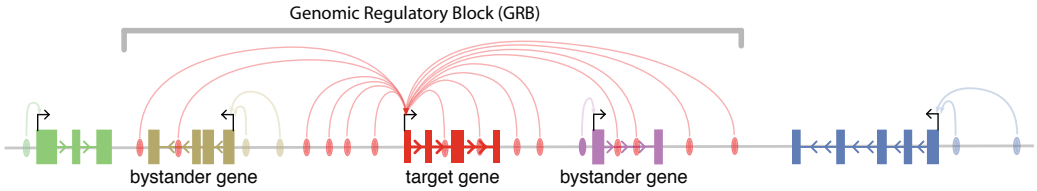


Figure 7.3. Organisation d'un bloc de régulation génomique (GRB). Les interactions entre un gène-cible et ses séquences de régulation, sous pression de sélection, maintiennent en synténie le gène-cible et des gènes voisins (bystanders) enchevêtrés avec les séquences de régulation. Figure tirée de (Becker and Lenhard 2007).

7.6. Origines de réplication prédites dans le génome humain

Les origines de réplication existant dans le génome humain ne sont pas exhaustivement connues, en raison de la difficulté à les détecter (Touchon et al. 2005). Une partie d'entre elles, correspondant aux origines évolutivement conservées et désignées sous le terme d' « origines maîtres », peuvent cependant être détectées en raison de la signature particulière qu'elles laissent sur la séquence. La réplication de l'ADN introduit un biais mutationnel qui se traduit par une asymétrie des nucléotides présents sur chacun des deux brins (Touchon et al. 2005). Si une origine de réplication est maintenue pendant un temps suffisamment long, cette accumulation asymétrique de mutations peut être détectée au niveau de la séquence et permet de déterminer la localisation de l'origine de réplication. Le rapport S, défini ci-dessous, s'inverse de part et d'autre d'une origine de réplication (très fort d'un côté, très faible de l'autre) :

$$S = \frac{G - C}{G + C} + \frac{T - A}{T + A}$$

Plus on s'éloigne de l'origine de réplication, plus l'asymétrie des deux brins s'atténue, car la séquence a une probabilité de plus en plus élevée d'être répliquée par une fourche de réplication arrivant par l'autre côté, si bien que le biais mutationnel a tendance à se compenser. Ainsi, le rapport S suit un profil en « toits d'usine », où il augmente de manière très brusque au niveau des origines de réplication puis diminue progressivement entre deux origines de réplication (Figure 7.4). Chaque région entre deux origines s'appelle un N-domaine (nommés ainsi à cause de l'aspect du profil de S) : la détection de ces N-domaines à partir du calcul de S permet de prédire la localisation des origines de réplication (Touchon et al. 2005; Huvet et al. 2007).

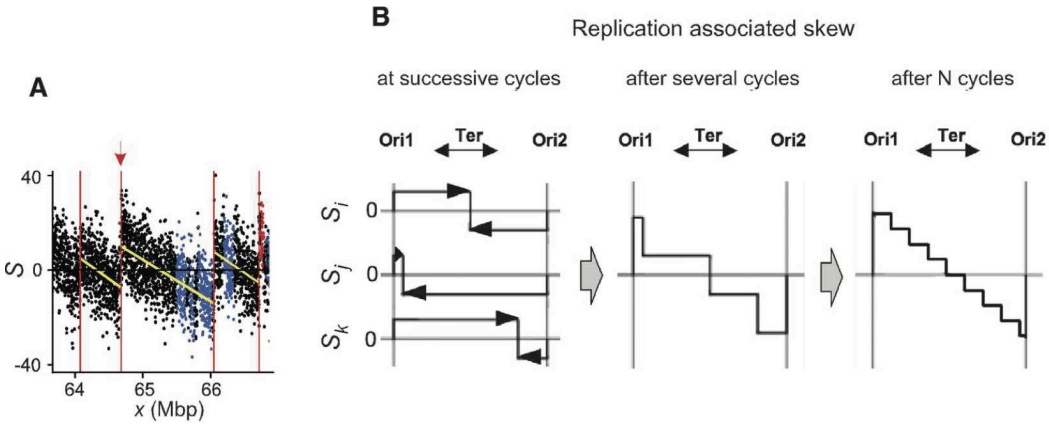


Figure 7.4. Structure et mise en place des N-domaines. A. Profil d'asymétrie des nucléotides dans la région du gène *MYC* dans le génome humain. Les lignes rouges marquent les origines de réplication putatives, la flèche marque une origine validée expérimentalement. B. Modèle de mise en place de l'asymétrie des brins entre deux origines de réplication. Le profil en « toits d'usine » reflèterait la superposition du biais mutationnel sur de nombreux cycles de réplication successifs où les fourches démarrent toujours au même endroit (aux origines de réplication) mais peuvent se terminer de manière moins stricte, en fonction de la vitesse de parcours de chaque fourche à chaque cycle. Figure tirée de (Huvet et al. 2007).

Le jeu d'origines de réplication utilisé dans cette étude est une mise à jour des origines publiées dans (Huvet et al. 2007) communiquée par les auteurs. Ce jeu contient 1 546 origines de réplication maîtres prédites aux bornes de N-domaines dans la version hg19 du génome humain ; ces N-domaines couvrent approximativement deux tiers du génome. Parmi ces origines, 874 sont intergéniques et utilisées dans nos analyses.

Chapitre 8. Analyses de synténies conservées

Les analyses de conservation de la synténie ont été réalisées en faisant appel à des programmes de comparaison de génomes initialement développés au laboratoire dans le cadre du projet de reconstruction des génomes ancestraux mené par Matthieu Muffato pendant sa thèse. Nous présentons ici rapidement leur principe de fonctionnement.

8.1. Détection de synténies conservées simples

La détection des synténies conservées au sens strict entre deux génomes, c'est-à-dire les blocs de gènes orthologues dans le même ordre et la même orientation dans les deux génomes, se fait en deux étapes. On fournit en entrée au programme d'une part la liste des gènes des deux génomes et leurs positions, qui permet de déduire l'ordre relatif des gènes dans chaque génome, et d'autre part, les arbres phylogénétiques des gènes, qui fournissent les liens d'orthologie entre les gènes des deux génomes.

En première étape, l'algorithme recherche des paires de gènes a_1 et b_1 adjacents dans le génome 1, telles que les gènes orthologues a_2 et b_2 soient également voisins et dans la même orientation dans le génome 2. Un paramètre permet de spécifier les gènes à considérer à cette étape :

-option *all* : tous les gènes des deux génomes sont considérés, même s'il n'existe pas d'orthologue dans l'autre génome (auquel cas ils ne peuvent jamais former de paire conservée en synténie).

- option *inEitherSpecies* : on considère tous les gènes qui, d'après l'arbre des gènes, existaient dans l'ancêtre commun et existent toujours dans au moins un des deux génomes, même s'ils sont perdus dans l'autre (à nouveau, ces derniers ne peuvent pas former de paire conservée).

- option *inBothSpecies* : on considère uniquement les gènes qui existent dans les deux espèces, c'est-à-dire qu'on ignore les pertes et gains de gènes espèce-spécifiques. Un gène qui n'existe pas dans l'une ou l'autre espèce est « invisible », et si les gènes de part et d'autre sont voisins dans l'autre génome, ils sont considérés comme une paire de gène conservée en synténie.

C'est l'option *inBothSpecies* qui sera systématiquement utilisée ici, puisqu'elle permet la détection la plus sensible des blocs de synténie conservée.

En deuxième étape, l'algorithme fusionne les paires conservées en blocs de synténie. Les paires sont fusionnées si elles sont dans le même ordre et la même orientation dans les deux

génomes, et séparées dans chacun par moins de *k* gènes. Le paramètre de stringence *k* est défini par l'utilisateur.

8.2. Détection de synténies double-conservées

Les synténies double-conservées sont détectées entre le génome d'une espèce ayant subi une duplication complète du génome (dit « génome dupliqué ») et le génome d'une espèce d'un groupe externe n'ayant pas subi la duplication (dit « génome non-dupliqué »). Un segment du génome non-dupliqué est synténique, en théorie, avec deux segments du génome dupliqué sur lesquels se trouvent alternativement les gènes orthologues. En pratique, avec les réarrangements chromosomiques qui ont pu affecter le génome dupliqué depuis la duplication complète, un segment dans le génome non-dupliqué peut partager des orthologues avec *k* segments du génome dupliquée, *k* étant supérieur ou égal à 2. En raison de la dégradation massive de la synténie au sens strict observée dans la plupart des génomes dupliqués (voir paragraphe 4.5.2), on détecte les synténies double-conservées en utilisant une définition de synténie beaucoup plus souple que l'ordre strict des gènes et proche de la définition large de la synténie conservée (voir chapitre 2).

On recherche des segments du génome dupliqué qui contiennent plusieurs gènes orthologues au même segment non-dupliqué, et reliés entre eux par des gènes ohnologues (Figure 8.1). Les blocs de synténie double-conservée sont construits en parcourant le génome non-dupliqué, et en testant la position du ou des gènes orthologues dans le génome dupliqué. A l'étape *n* du parcours du génome non-dupliqué, on étend le bloc de synténie double-conservée au gène *n* et à ses orthologues si une des conditions suivantes est vérifiée :

- au moins l'un des orthologues du gène *n* est sur le même chromosome du génome dupliqué que l'un des orthologues du gène *n-1* (synténie au sens large conservée entre les gènes *n-1* et *n* dans les deux génomes)
- au moins l'un des orthologues du gène *n* est dans le voisinage d'un gène déjà inclus dans le bloc de synténie. La notion de « voisinage » est paramétrée par l'utilisateur : il s'agit du nombre de gènes autour de chaque gène du génome dupliqué faisant déjà partie du bloc de synténie double-conservée parmi lesquels on va rechercher des gènes permettant d'étendre le bloc de synténie.
- au moins l'un des orthologues du gène *n* est dans le voisinage d'un ohnologue d'un gène soit déjà inclus dans le bloc de synténie, soit lui-même dans le voisinage d'un gène déjà inclus (cette condition permet de passer d'un segment du génome dupliqué à un autre en utilisant les liens d'ohnologie locaux entre gènes)

Si le gène *n* ne vérifie pas ces conditions, on interrompt le bloc de synténie courant et on amorce un nouveau bloc d'alternance, qu'on cherche à étendre en continuant le parcours du génome non-dupliqué. Ainsi, on partitionne le génome non-dupliqué en segments qui sont en synténie (au sens large) avec deux ou plusieurs segments du génome dupliqué, même si l'ordre des gènes parmi ces segments peut être très différent de l'ordre dans le segment non dupliqué.

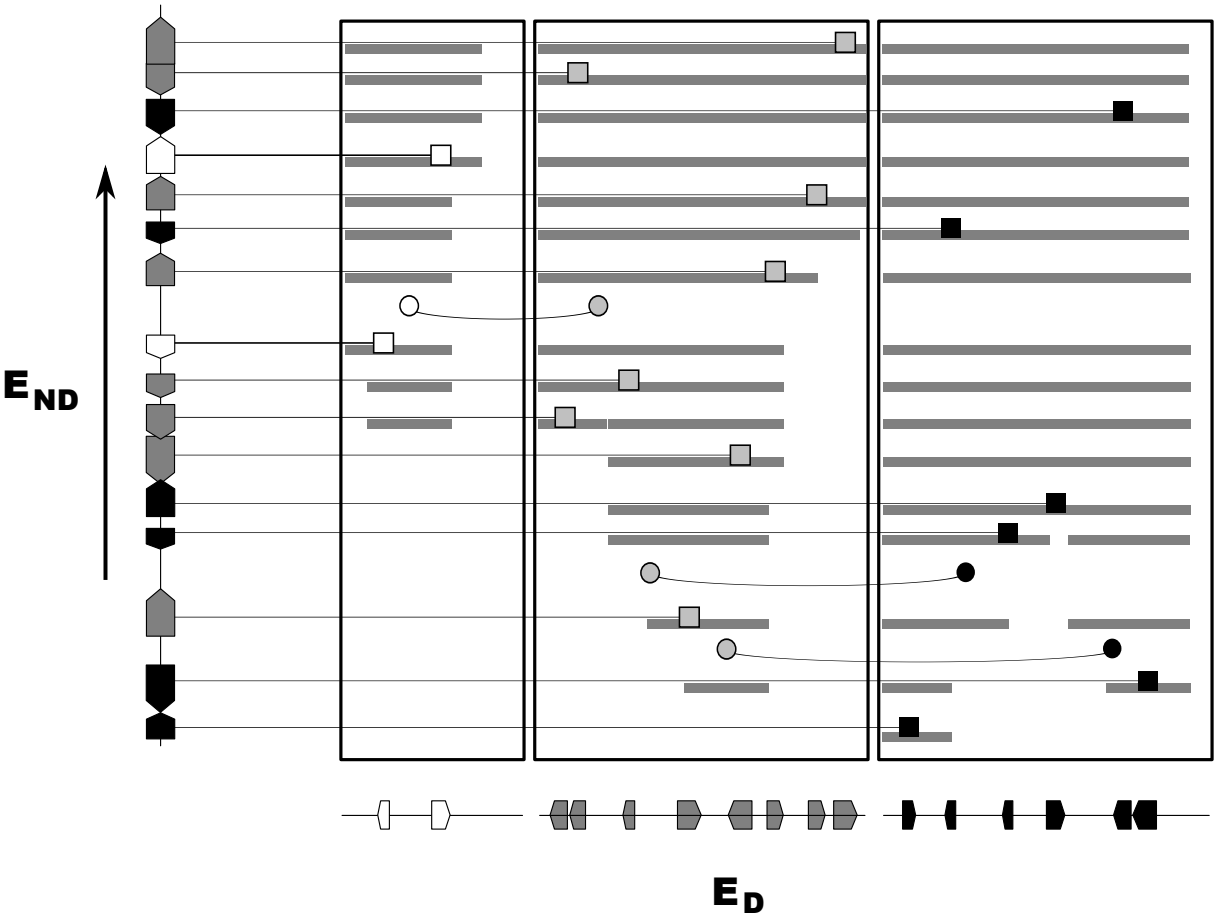


Figure 8.1. Extraction d'un bloc de synténie double-conservée. Le chromosome du génome non dupliqué est représenté verticalement à gauche, et les chromosomes du génome dupliqué sont disposés horizontalement. Un carré représente des gènes orthologues, les cercles représentent des gènes ohnologues. On parcourt les gènes du génome non dupliqué de bas en haut : à chaque nouveau gène, on étend le bloc si le ou les orthologues du gène se trouvent (i) sur le même chromosome que celui du gène précédent, (ii) dans le voisinage d'un gène déjà inclus (régions autorisées : zones grises), (iii) dans le voisinage d'un ohnologue d'un gène dans une région autorisée. Dans cet exemple, le 2ème gène est ajouté au bloc car son orthologue est sur le même chromosome que celui du 1er gène. Le 3ème gène est rajouté car son orthologue est dans le voisinage de l'ohnologue d'un gène qui lui-même se trouve dans le voisinage du 2ème gène. Figure tirée de (Muffato 2010).

Chapitre 9. Reconstruction des génomes ancestraux avec AGORA

AGORA (Algorithms for Gene Order Reconstruction in Ancestors) est une méthode développée au laboratoire pour la reconstruction de génomes ancestraux (Muffato et al., *en préparation*). L'ordre des gènes dans le génome ancestral reconstitué par AGORA est la base du projet portant sur les facteurs gouvernant l'occurrence des réarrangements mené au cours de cette thèse. Dans cette partie, nous présentons brièvement le fonctionnement d'AGORA et ses caractéristiques par rapport à d'autres méthodes de reconstructions ancestrales ; pour une description approfondie d'AGORA et de ses performances, nous référons les lecteurs à la thèse de Matthieu Muffato (Muffato 2010), qui a développé la méthode.

AGORA est une méthode fondée sur un raisonnement de parcimonie. Son principe peut se résumer assez simplement en une phrase : si deux gènes sont adjacents (l'un à côté de l'autre) dans deux génomes, alors ces gènes sont adjacents depuis leur dernier ancêtre commun et dans tous les ancêtres intermédiaires. AGORA intègre les comparaisons de l'ordre des gènes dans un grand nombre d'espèces pour inférer l'ordre des gènes dans tous les ancêtres de l'arbre phylogénétique. L'algorithme peut être utilisé de deux manières différentes : l'une, dite « approche en une passe », utilise de manière compréhensive et exhaustive toutes les adjacences de gènes dans tous les génomes modernes (haute spécificité), et interrompt la continuité des blocs de gènes ancestraux si l'adjacence ancestrale d'un des gènes est inconnue ou ambiguë (faible sensibilité) ; l'autre, dite « approche multi-passes », favorise la continuité maximale du génome ancestral jusqu'à obtenir des chromosomes ancestraux entiers (haute sensibilité) quitte à inférer localement des adjacences de gènes peu soutenues par les génomes modernes (faible spécificité), en se basant prioritairement sur un petit nombre de gènes dont l'ordre relatif est non ambigu entre les génomes puis en rajoutant progressivement d'autres gènes dans ce squelette de reconstruction. Nous ne présenterons ici que le principe de l'approche en une passe, qui est celle qui a été utilisée pour produire les données de génomes ancestraux exploitées au cours de cette thèse.

9.1. Comparaison des génomes deux à deux

Chaque nœud de l'arbre phylogénétique des espèces modernes considérées définit un ancêtre à reconstruire. AGORA prend en entrée deux types de données : d'une part, la liste des gènes dans les génomes modernes, avec leurs positions, et d'autre part, les arbres phylogénétiques des gènes reconstruits par TreeBeST (voir paragraphe 7.1). Les arbres de gènes fournissent les relations d'orthologie entre les gènes de toutes les espèces, et les génomes ancestraux dans lesquels chaque gène existe (pour rappel, voir Figure 7.1). AGORA compare tous

les génomes modernes deux à deux, et détecte les paires de gènes orthologues qui sont à la fois adjacents et dans la même orientation dans les deux génomes (Figure 9.1.A). Ainsi, si les gènes a_1 et b_1 sont voisins dans un génome, et que leurs orthologues respectifs a_2 et b_2 sont également adjacents et dans la même orientation dans un autre génome, alors ces gènes ont potentiellement retenu leur configuration relative ancestrale.

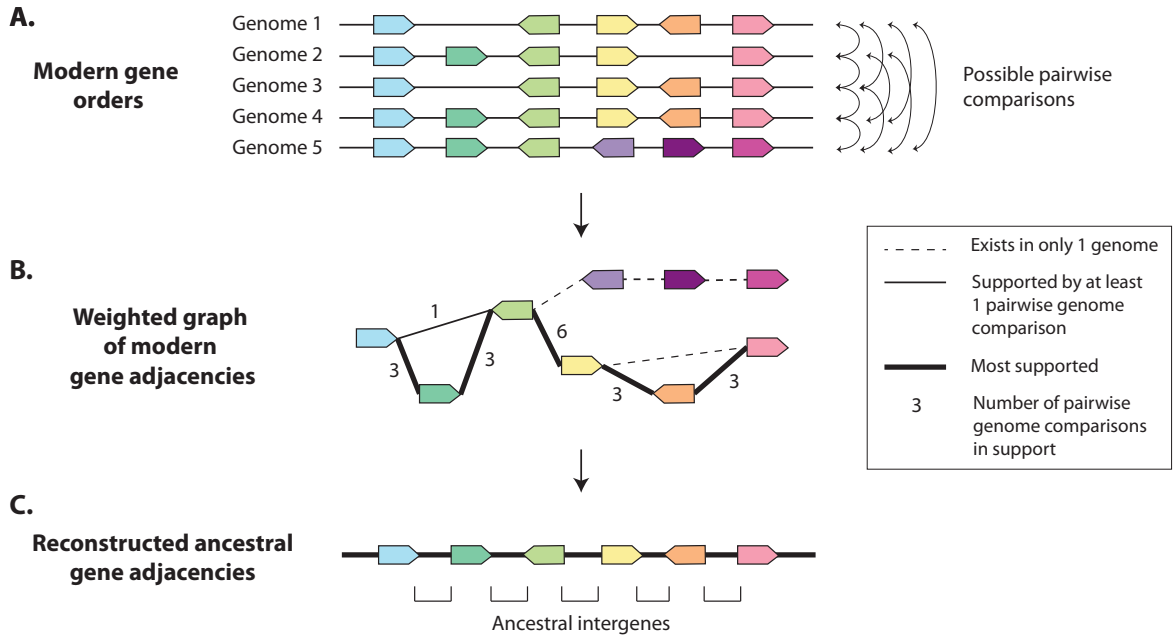


Figure 9.1. Fonctionnement d'AGORA. A. Comparaison deux à deux des génomes informatifs pour l'ordre des gènes au nœud ancestral. B. Construction du graphe d'adjacences, pondéré par le nombre de comparaisons deux à deux soutenant chaque arête. C. Linéarisation du graphe et extraction de l'ordre ancestral des gènes.

9.2. Construction d'un graphe d'adjacences

Pour chaque ancêtre, on ne considère que le sous-jeu de comparaisons de génomes deux à deux informatives pour cet ancêtre, c'est-à-dire que l'ancêtre se trouve sur le chemin dans l'arbre phylogénétique qui permet d'aller d'une espèce à l'autre. AGORA construit un graphe des adjacences de gènes possibles dans le génome de l'ancêtre : toute adjacence moderne de gènes est représentée comme une arête sur le graphe qui relie les gènes du génome ancestral, et chaque arête se voit attribuer un score qui correspond au nombre de comparaisons où cette adjacence est observée dans les deux génomes modernes (Figure 9.1.B). Ce score peut aller de 1 (adjacence de gènes vue dans une seule comparaison de génomes deux à deux) à un score maximal qui correspond au nombre de comparaisons deux à deux considérées (l'adjacence existe dans tous les génomes modernes, et est donc soutenue comme ancestrale par toutes les comparaisons). On obtient alors un graphe des adjacences de gènes possibles dans le génome de l'ancêtre qui doit être ensuite linéarisé.

9.3. Linéarisation du graphe et extraction de l'ordre ancestral des gènes

L'extraction de l'ordre le plus probable des gènes dans le génome de l'ancêtre se fait en linéarisant le graphe d'adjacences. En effet, le graphe n'est pas forcément linéaire, comme attendu dans le cas idéal : dans certains cas, plusieurs comparaisons de génomes peuvent rapporter des voisins différents pour un même gène (à cause de la perte indépendante d'un gène *n* dans plusieurs lignées, par exemple : les gènes *n-1* et *n+1* se retrouvent alors adjacents dans ces lignées). L'algorithme teste successivement toutes les arêtes du graphe, en commençant par celles de plus haut score, et conserve les arêtes tant qu'elles ne sont pas incohérentes avec les arêtes déjà sélectionnées. Les arêtes incohérentes sont supprimées. On obtient ainsi une liste de chemins maximaux et non chevauchants liant les gènes ancestraux, représentant l'ordre des gènes reconstruit dans le génome ancestral (Figure 9.1.C).

9.4. Performances d'AGORA

Dans son principe, AGORA est proche des autres méthodes de reconstruction des génomes ancestraux basées sur les adjacences de marqueurs (voir chapitre 5), puisqu'elle se base sur des synténies conservées entre marqueurs pour construire de proche en proche le chemin correspondant à l'ordre des marqueurs ancestraux. AGORA a été confrontée à trois autres méthodes, MGR, MGRA et inferCARs, pour tester la performance de la méthode dans un cas théorique et idéal. Les quatre méthodes ont été appliquées pour reconstruire un génome ancestral théorique, à partir de « génomes modernes » simulés en réarrangeant ce génome ancestral théorique. Les réarrangements se font en suivant un modèle d'évolution réaliste implémenté dans le programme MagSimus (pour plus de détails, consulter la thèse de Matthieu Muffato), suivant un taux de réarrangements fixé par l'utilisateur. Le génome ancestral contient un nombre de marqueurs fixé par l'utilisateur (ici, entre 100 et 20 000 marqueurs), tous uniques : il s'agit d'un cas théorique idéal sans pertes ni duplications de gènes. MGR et MGRA échouent toutes les deux à produire une reconstruction du génome ancestral pour un grand nombre de gènes dans le génome ou un taux de réarrangements élevé ; seuls inferCARs et AGORA produisent un résultat lorsque le génome contient 20 000 marqueurs, soit le nombre de gènes dans un génome de vertébrés typique. InferCARs et AGORA donnent des résultats très similaires, avec des spécificités et des sensibilités au-delà de 99,9%. AGORA se révèle légèrement plus spécifique et légèrement moins sensible qu'inferCARs pour les taux de réarrangements élevés, mais les différences sont du niveau de quelques adjacences non ou mal reconstruites parmi les 20 000 marqueurs considérés. AGORA est donc très performant, et peut reconstruire des génomes caractérisés par un grand nombre de marqueurs et un taux de réarrangements élevés, tout comme inferCARs.

La supériorité majeure d'AGORA sur les méthodes précédemment proposées, et notamment inferCARs, réside dans le type de marqueurs pouvant être pris en compte dans l'analyse sur des données réelles et imparfaites : les autres méthodes ne peuvent considérer que des marqueurs

uniques et existants dans tous les génomes modernes considérés. Il ne leur est pas possible de considérer les gènes dans leur ensemble comme des marqueurs, puisque dans le cas réel, certains gènes sont perdus et d'autres dupliqués au fil de l'évolution, si bien que la plupart ne constituent pas des marqueurs en exactement une copie dans chaque génome. Les autres méthodes utilisent des marqueurs basés sur les alignements de génomes, qui ne prennent pas en compte les régions (et les gènes) dupliquées ou éliminées dans certains génomes ; ces marqueurs passent également outre les régions non alignées, inversées ou transloquées trop courtes pour être prises en compte. Ces méthodes permettent d'obtenir une bonne continuité des génomes ancestraux et une architecture générale à l'échelle du chromosome qui est correcte, mais au prix d'une reconstruction où seul un nombre restreint de gènes ancestraux sera réellement et correctement placé dans le génome de l'ancêtre. AGORA, en revanche, construit des blocs ancestraux qui n'atteignent pas la longueur d'un chromosome avec l'approche en une passe, car la linéarisation du graphe d'adjacences interrompt le bloc si elle se heurte à des impasses ou à des ambiguïtés (le N50[1] des blocs de gènes ancestraux pour l'ancêtre Boreoeutheria est de 54 gènes, alors que le N50 des chromosomes d'un vertébré typique est de 950 gènes) ; mais au sein de ces blocs l'agencement des gènes est intégralement reconstruit, y compris pour des gènes dupliqués ou perdus ultérieurement dans la phylogénie. AGORA est donc beaucoup plus compréhensif et résolutif que les autres méthodes sur les données réelles, bien que le génome ancestral reconstruit soit moins continu.

Au cours de ce travail de thèse, nous avons étudié la distribution des points de cassure de réarrangement dans les espaces intergéniques qui séparent les gènes dans le génome de l'ancêtre. Connaître l'ordre exact des gènes, et donc les intergènes existants dans le génome ancestral, est impératif pour répondre à la question posée. Dans notre cas, l'utilisation des reconstructions de génomes ancestraux fournies par AGORA s'impose donc comme le choix logique.

[1] N50 : longueur du bloc de gènes tel que 50% des gènes se trouvent dans un bloc au moins aussi long.

Chapitre 10. Modélisations par régression de Poisson

Au cours de ce travail, nous utilisons la méthode de la régression de Poisson pour modéliser la distribution des points de cassure de réarrangements évolutifs depuis un génome ancestral. Dans ce chapitre, nous présentons les bases statistiques de la distribution de Poisson et de la régression de Poisson multivariée. Nous décrirons simplement les aspects nécessaires pour comprendre les analyses menées dans ce travail ; la régression de Poisson est une technique de régression multivariée classique dont les tenants et aboutissants détaillés sont disponibles dans la plupart des manuels de statistiques.

10.1. Distribution de Poisson

La distribution de Poisson est utilisée pour décrire la probabilité d'occurrence d'événements rares dans un intervalle d'une longueur définie (intervalle spatial ou intervalle de temps). Les événements sont supposés aléatoires et indépendants ; pour un taux moyen t d'événements par unité spatiale ou temporelle, la distribution de Poisson décrit la probabilité d'observer k événements (0, 1, 2, …) dans un intervalle de taille L. Cette probabilité est :

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \qquad avec \qquad \lambda = tL$$

L'espérance $E(X)$ d'une distribution correspond à la valeur moyenne attendue pour X (plus exactement, à la limite vers laquelle tend la moyenne de X quand le nombre d'essais tend vers l'infini). Pour la distribution de Poisson, il s'agit du nombre moyen d'événements attendus dans un intervalle de taille L, soit le taux moyen d'événements par unité de longueur multiplié par la longueur de l'intervalle :

$$E(X) = tL = \lambda$$

L'espérance est donc égale au paramètre λ de la distribution. Connaître $E(X)$ suffit à calculer la probabilité $P(X = k)$ pour tout $k > 0$.

L'une des propriétés intéressantes de la distribution de Poisson est son additivité : si X_1 et X_2 sont deux variables indépendantes suivant des distributions de Poisson, alors le total des événements X_1+X_2 suit également une distribution de Poisson dont l'espérance est $\lambda_1+\lambda_2$. Si on compte les événements se produisant dans un même intervalle de taille L, l'espérance de X_1+X_2 est $(t_1+t_2)L$. Réciproquement, l'espérance de X dans un intervalle de taille nL est ntL soit $nE(X)$: intuitivement, on attend effectivement deux fois plus d'événements dans un intervalle deux fois plus long.

10.2. Modélisation par régression de Poisson

La régression de Poisson est une technique de régression multivariée par modèles linéaires généralisés, c'est-à-dire une généralisation de la régression linéaire qui permet de modéliser les variations d'une variable aléatoire *X* en fonction d'autres variables explicatives *Y, Z, W, …* dans le cas où *X* ne suit pas une distribution gaussienne (seule condition pour laquelle la régression linéaire classique est valide). La régression de Poisson permet de modéliser comment varie le nombre discret *X* d'événements qui se produisent dans un intervalle en fonction des caractéristiques *Y, Z, W, …* de cet intervalle. Elle est donc utilisée pour modéliser des comptages ou des taux d'événements.

10.2.1. Principe de la méthode

La majorité des comptages d'événements se produisant dans des intervalles peuvent être modélisées par des fonctions linéaires d'autres variables après une transformation logarithmique : on dit que le logarithme[2] est la « fonction de lien usuelle » de la régression de Poisson. La régression de Poisson estime les paramètres *a, b, c, … , d* tels que :

$$\log\big(E(X)\big) = a\,Y + b\,Z + c\,W + \cdots + d$$

Ces coefficients sont estimés par une méthode de maximum de vraisemblance, implémentés dans R dans la fonction *glm()*. Le but est de trouver le jeu de paramètres tels que la variable réponse *X* se rapproche au mieux du comportement d'une variable de Poisson de paramètre *E(X)* quand *E(X)* est adaptée aux caractéristiques de chaque intervalle du jeu. Ces caractéristiques peuvent indifféremment être des variables continues, catégorielles ou binomiales.

En pratique, comme *E(X)* est la moyenne vers laquelle tend *X* sur un grand nombre d'essais, on peut lisser les données en triant les intervalles par classes homogènes, puis régresser directement le nombre total d'événements sur un nombre *n* d'intervalles en fonction de la valeur moyenne de la classe pour chaque caractéristique *Y, Z, W, …* , dite valeur représentante. Si les intervalles ne sont pas homogènes en taille, la taille moyenne des intervalles est prise en compte parmi les variables explicatives. Le nombre d'intervalles *n* dans chaque classe est également pris en compte dans la régression sous la forme d'une variable d'exposition *log(n)*, dont le coefficient est forcément 1, afin de refléter la relation suivante :

$$\log\big(E(nX)\big) = \log\big(nE(X)\big) = \log\big(E(X)\big) + \log(n)$$

$$donc \quad \log\big(E(nX)\big) = a\,Y + b\,Z + c\,W + \cdots + d + \log(n)$$

[2] On utilise dans ce manuscrit la notation log pour le logarithme népérien (notation anglo-saxonne).

10.2.2. Variables explicatives significatives

Lorsque l'un ou l'autre des coefficients de régression *a, b, c* est significativement différent de 0, il existe une corrélation significative entre la variable explicative et la variable réponse que l'on étudie : *X* augmente quand *Y* augmente si a > 0 et *X* diminue quand *Y* augmente si a < 0. La significativité du paramètre est testée par un z-test à partir de la valeur et de l'écart-type estimé du paramètre. Les variables dont les paramètres sont significatifs sont retenues dans le modèle, les autres sont rejetées. Certaines variables peuvent être significatives quand elles sont considérées comme seule variable explicative, mais pas lorsqu'elles sont intégrées dans une régression multivariée avec d'autres variables. Ceci signifie que lorsque les autres paramètres sont égaux par ailleurs, cette variable n'influence pas la variable réponse : la corrélation initialement observée entre les deux est une conséquence secondaire de la corrélation indépendante des deux variables avec l'une des variables explicatives significatives intégrée dans la régression.

10.2.3. Statistiques de validation de la régression

Lorsque des variables explicatives ont été sélectionnées, il faut pouvoir tester la qualité de la régression, c'est-à-dire à quel point les variables choisies permettent de rendre compte des fluctuations de la variable réponse. Il faut alors pouvoir répondre à deux questions : premièrement, le modèle est-il complet, c'est-à-dire qu'il n'y a pas de différence significative entre les valeurs attendues sous le modèle et les valeurs observées ? Et deuxièmement, quelle part de la variation de la variable réponse est expliquée par les variations des variables explicatives ?

Le premier point est testé par un test de Chi^2 classique sur la déviance résiduelle du modèle par rapport au nombre de degrés de liberté restants, qui est égal au nombre de classes moins le nombre de variables explicatives retenues. La déviance est une mesure de log-vraisemblance issue de la modélisation par maximum de vraisemblance, qui s'apparente dans son principe à la variance dans une modélisation linéaire classique : elle mesure la déviation entre les valeurs observées et les valeurs du modèle, et approxime les valeurs du Chi^2. Si le test est non significatif, alors le modèle est complet : la déviance résiduelle entre les données et le modèle est attribuable au bruit statistique.

Le deuxième point est plus difficile à tester parce qu'il n'existe pas pour les modèles linéaires généralisés de statistique réellement comparable au R^2 de la régression linéaire classique. Plusieurs statistiques ont été proposées, parmi lesquelles nous avons choisi d'utiliser le pseudo R^2 de McFadden, qui est le plus simple et qui est relativement conservatif par rapport à d'autres statistiques comme le pseudo R^2 de Cox & Snell ou celui de Nagelkerke. Le pseudo R^2 de McFadden correspond au pourcentage de déviance expliqué par le modèle, c'est-à-dire qu'il compare la déviance initiale existant dans les données (modèle nul) à la déviance résiduelle du modèle. Il donne ainsi une impression de la qualité du modèle, mais n'est pas aussi directement interprétable que le R^2 de la régression linéaire puisqu'il correspond à un rapport de log-vraisemblances, pas de variances.

10.3. Régression multivariée progressive

Enfin, la régression multivariée progressive consiste à trouver une combinaison de variables explicatives « nécessaires et suffisantes » en ajoutant progressivement des variables afin d'aboutir à un modèle complet mais simple. Il s'agit ici d'affiner et de mieux documenter le modèle multivarié qu'on obtient en faisant une régression multivariée avec toutes les variables explicatives possibles. La procédure « vers l'avant », qui est celle que nous avons utilisée, fonctionne avec les étapes suivantes :

- on fait un modèle avec chaque variable individuellement
- on ne garde que les variables qui, individuellement, sont significatives, et on les classe en fonction de leur significativité (du maximum au minimum)
- on part d'un modèle avec uniquement la variable la plus explicative, puis on fait des régressions successives en introduisant les variables une à une : à chaque étape, on conserve les variables qui sont significatives si elles améliorent le modèle par rapport à l'étape précédente
- on procède ainsi jusqu'à ce que toutes les variables soient épuisées

La régression est faite indépendamment à toutes les étapes, si bien que les variables obtenues en fin de régression doivent être celles qui seraient sélectionnées si la régression multivariée était faite en une seule passe en considérant toutes les variables à la fois. Sinon, c'est qu'il y a de nombreuses variables intercorrélées dans le jeu et que l'ordre d'introduction des variables n'est pas le bon et est à revoir (mais nous n'avons pas rencontré cette situation dans les analyses). Cette méthode permet, à chaque étape, de distinguer le gain apporté par l'introduction d'une nouvelle variable.

Les étapes successives sont comparées à l'aide de deux statistiques. A nouveau, on utilise le test de Chi^2 pour comparer la déviance résiduelle entre deux étapes successives (1 degré de liberté) car la différence des déviances approxime le Chi^2. Le nouveau modèle est meilleur que l'ancien si le test de Chi^2 est significatif, c'est-à-dire qu'on réduit la déviance de manière significative par l'ajout d'une nouvelle variable. Par ailleurs, l'introduction de variables faiblement significatives peut introduire une différence de déviance significative par rapport à l'étape précédente, mais avec un coût jugé trop important au vu l'amélioration du modèle. Ce coût est mesuré par le critère d'information d'Akaike (AIC) : on ne sélectionne que les étapes qui diminuent l'AIC du modèle, sans quoi on estime qu'il y a sur-adéquation du modèle (trop de variables pour la déviance expliquée par le modèle).

Troisième Partie

Etude par modélisation des points de cassure de réarrangements évolutifs

Introduction

La première partie de ce travail de thèse porte sur l'étude des points de cassure de réarrangements évolutifs dans les génomes de mammifères. Ce projet s'inscrit dans la continuité des travaux réalisés au laboratoire sur la reconstruction de génomes ancestraux, utilisant les nombreux génomes mis à disposition de la communauté scientifique notamment pour le phylum des vertébrés. Les reconstructions de génomes s'attachent à reconstruire l'organisation des gènes dans l'état ancestral. Ces reconstructions fournissent une perspective historique sur l'évolution des génomes, et permettent de replacer, dans chaque lignée descendante, les événements qui ont modifié l'organisation ancestrale pour amener aux génomes modernes. La question qui découle logiquement de cette ressource est : peut-on utiliser l'état ancestral pour comprendre comment se produisent ces réarrangements ? Modéliser l'apparition de ces événements apporterait des indications sur les propriétés génomiques et les contraintes fonctionnelles qui gouvernent l'évolution de l'organisation des génomes.

La méthode d'étude des points de cassure que nous développons est assez différente de celles publiées dans la littérature. Nous proposons en effet de partir de l'ensemble du génome dans l'état ancestral afin de mesurer la probabilité que chaque région ancestrale ait subi une cassure en fonction de ses propriétés. Le travail réalisé sur ce projet s'articule en trois grandes parties : tout d'abord, nous avons développé une approche qui permet d'estimer des caractéristiques du génome dans l'état ancestral plus poussées que le simple ordre des gènes (longueur des intergènes, taux de GC, etc.) sans avoir à recourir à des reconstructions de la séquence ancestrale peu fiables à l'échelle évolutive que nous étudions. La distribution des ruptures de synténie observées entre le génome ancestral et les génomes modernes a ensuite été modélisée en fonction des caractéristiques locales du génome dans l'état ancestral, ce qui nous a permis d'identifier les paramètres influençant significativement la cassure d'une part, et leur importance relative de l'autre. Enfin, nous avons validé le modèle obtenu par simulations, au cours desquelles nous démontrons que ce modèle simple suffit à rendre compte de manière adéquate des observations de la littérature liées à la répartition non aléatoire des points de cassure dans les génomes de vertébrés.

Le modèle obtenu initialement dans les génomes de mammifères a été confronté à des données similaires dans les génomes de levures : ceux-ci présentent les traces d'une évolution similaire, où les points de cassure sont dépendants des mêmes propriétés génomiques. Les résultats obtenus dans cette partie éclairent donc vraisemblablement les bases des mécanismes modifiant l'organisation des génomes eucaryotes de manière générale. L'ensemble de ce travail fait l'objet d'une publication, fournie en Annexe.

Chapitre 11. Reconstruction du génome ancestral des mammifères boreoeuthériens

L'approche de modélisation que nous adoptons dans cette étude nécessite d'avoir pour base un génome ancestral dont plusieurs propriétés sont connues : tout d'abord, l'ordre des gènes, qui nous servent de marqueurs référents pour identifier les réarrangements, et d'autre part, des propriétés génomiques dans l'état ancestral dont la corrélation avec les points de cassure pourra être testée par la suite. Boreoeutheria est le nom utilisé pour désigner le dernier ancêtre commun aux groupes des euarchontoglires, comprenant les primates et les rongeurs, et des laurasiathériens, qui comprennent en autre les ongulés et les carnivores. Boreoeutheria est donc l'ancêtre d'une grande partie des mammifères, en excluant les atlantogénates (éléphants, tatou, etc.), les marsupiaux et l'ornithorynque (Figure 11.1).

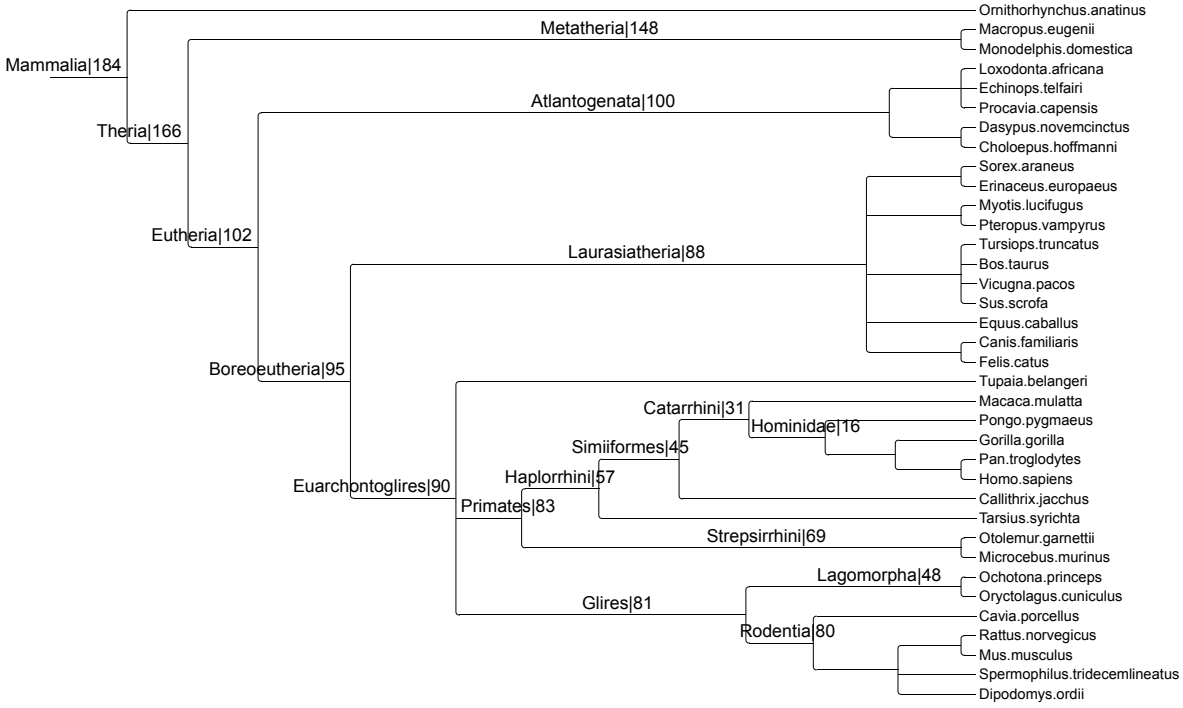


Figure 11.1. Arbre phylogénétique des génomes de mammifères séquencés. Les nombres correspondent aux âges consensus des nœuds estimés à partir des données d'évolution moléculaire (fournies par la base de données Ensembl ou estimées d'après TimeTree (Hedges et al. 2006)).

Avec 28 espèces descendantes et plusieurs groupes externes proches séquencés lors de la réalisation de cette étude, cet ancêtre est idéalement placé dans la phylogénie des mammifères

pour les reconstructions de génomes ancestraux (Blanchette et al. 2004). Dans cette partie est décrite la reconstruction des adjacences gène-gène existant dans le génome de Boreoeutheria, ainsi que de trois caractéristiques ancestrales de ces espaces intergéniques : la longueur, le taux de GC et la présence de contraintes fonctionnelles sur des couples gène/élément régulateur.

11.1. Reconstruction des adjacences ancestrales avec AGORA

Comme évoqué en introduction, les points de cassure de réarrangements évolutifs sont presque exclusivement intergéniques. Cette caractéristique est probablement due au fait que les cassures au sein d'un gène sont presque toujours délétères et sont à terme éliminées par sélection. Nous nous sommes donc concentrés dans cette étude sur les points de cassure intergéniques, en considérant chaque intergène comme une unité biologique continue et susceptible d'être cassée au cours de l'évolution.

Les gènes présents chez l'ancêtre Boreoeutheria ont été documentés à partir des arbres de gènes disponibles dans la version 57 de la base Ensembl (voir paragraphe 7.1). Brièvement, tout gène existant en deux copies orthologues chez deux espèces dont le dernier ancêtre commun est soit Boreoeutheria, soit un ancêtre plus ancien, est un gène qui existait dans le génome de Boreoeutheria. L'ordre de ces gènes dans le génome de Boreoeutheria a été reconstruit en utilisant la méthode AGORA, développée au sein du laboratoire (Muffato et al., *en préparation*). Cette méthode, décrite au chapitre 9 de ce manuscrit, est une méthode s'appuyant sur un parcours de graphes selon des contraintes de parcimonie. Tout comme pour le contenu en gènes du génome ancestral, le raisonnement sous-jacent est que toute paire de gènes qui se trouvent côte-à-côte dans deux génomes est une paire de gènes qui a retenu sa configuration ancestrale, et définit donc un intergène orthologue entre les deux espèces. AGORA compare ainsi tous les génomes susceptibles d'être informatifs sur l'ordre des gènes chez l'ancêtre Boreoeutheria, et construit un graphe pondéré des adjacences gène-gène possibles dans le génome de l'ancêtre dont elle extrait l'ordre des gènes le plus probable.

La méthode AGORA a permis de reconstruire 18436 adjacences de gènes dans le génome de l'ancêtre Boreoeutheria. En moyenne, ces intergènes existent encore dans 13,7 espèces modernes, parmi les 28 descendants séquencés utilisés dans cette étude (écart-type = 6,2; Figure 11.2.A). Parmi ces intergènes, 73,5% existent dans au moins 10 descendants, et plus de 90% sont soutenus par au moins 5 génomes modernes, dont au moins un euarchontoglire et un laurasiathérien. Ces chiffres peuvent paraître faibles, mais il faut noter que 16 des 28 génomes boreoeuthériens disponibles à la date de cette étude sont des séquençages à faible couverture (2 à 3x) dans un état d'assemblage très fragmenté. Ainsi, les génomes à faible couverture contribuent moins que les génomes à haute couverture à la reconstruction ancestrale et on y observe moins d'adjacences gène-gène conservées simplement du fait de la fragmentation de l'assemblage (Figure 11.2.B). Dans ce cadre, les statistiques de reconstruction des intergènes ancestraux sont, de fait, la preuve d'un bon soutien général par les génomes modernes.

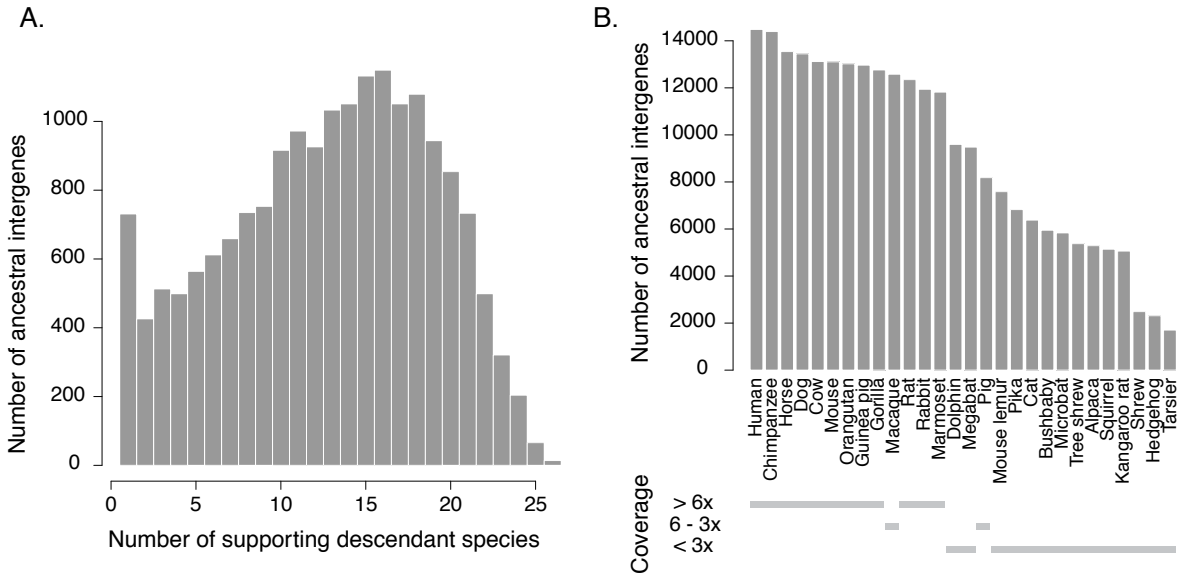


Figure 11.2. Soutien des adjacences de gènes dans le génome de Boreoeutheria par les données modernes. A. Distribution du nombre de génomes boreoeuthériens modernes soutenant chaque adjacence ancestrale. B. Contribution de chaque génome moderne à la reconstruction ancestrale. L'histogramme représente le nombre d'adjacences ancestrales conservées dans chaque génome. Les génomes séquencés à faible couverture (< 3x) sont disponibles sous forme d'assemblages très fragmentés, et sont moins informatifs que les génomes séquencés à haute couverture pour la reconstruction ancestrale.

11.2. Distances intergéniques ancestrales

La suite de ce travail repose sur une modélisation de la distribution des points de cassure apparus depuis l'état ancestral, c'est-à-dire le génome de Boreoeutheria, jusqu'aux génomes de ses descendants. Cette modélisation a pour but de mettre en évidence les propriétés locales du génome qui ont soit favorisé, soit limité l'apparition des cassures. Il est donc nécessaire de reconstruire un certain nombre de propriétés des intergènes dans l'état ancestral à corréler aux points de cassure, telles que leur longueur, leur taux de GC, etc.

A ce jour, deux types de méthodes de reconstruction des génomes ancestraux ont été proposées (voir chapitre 5). Certaines, comme AGORA, s'attachent à reconstruire l'organisation générale du génome via l'ordre d'un certain nombre de marqueurs. Ces méthodes n'ont pas vocation à renseigner sur les propriétés locales de la séquence, mais uniquement sur l'agencement des éléments fonctionnels tels que les gènes les uns par rapport aux autres. D'autres méthodes s'intéressent à un niveau de résolution plus fin et cherchent à reconstruire la séquence nucléotidique du génome ancestral. Dans ce cas, la séquence ancestrale reconstruite permet de déduire certaines propriétés locales du génome, mais l'information n'est disponible que dans les régions où la séquence ancestrale peut effectivement être reconstruite, ce qui suppose que les séquences modernes soient suffisamment conservées pour être alignées. De fait, pour l'ancêtre Boreoeutheria, les reconstructions de séquence ancestrale se limitent essentiellement aux régions géniques. Nous proposons ici une méthode intermédiaire, partant de l'hypothèse que le détail de la séquence nucléotidique n'est pas nécessaire pour estimer un

certain nombre de propriétés locales du génome ancestral pourvu que ces propriétés soient suffisamment conservées dans les génomes modernes descendants de cet ancêtre.

11.2.1. Corrélation des longueurs intergéniques orthologues modernes

La première propriété que nous avons cherché à reconstruire est la longueur des intergènes. En effet, sous l'hypothèse la plus simple d'une distribution aléatoire, les points de cassure devraient être distribués dans les intergènes en suivant une loi de Poisson, c'est-à-dire que le nombre de points de cassure par intergène devrait être proportionnel, en moyenne, à la longueur de l'intergène. Afin d'estimer la longueur des intergènes ancestraux, nous nous sommes intéressés à la corrélation qui existe entre les longueurs d'un même intergène orthologue chez plusieurs espèces. En effet, si la longueur d'un même intergène est proche chez les différentes espèces qui descendent de l'ancêtre Boreoeutheria, l'explication la plus parcimonieuse est que la taille de cet intergène a peu varié depuis l'état ancestral. Dans ce cas, la médiane des valeurs modernes serait une bonne approximation de la longueur de l'intergène ancestral.

Afin de tester si cette approximation est valide sur l'ensemble du génome, nous avons représenté la dispersion des longueurs des intergènes orthologues modernes (en y) autour de leur valeur médiane (en x) pour l'ensemble des intergènes ancestraux reconstruits chez Boreoeutheria (Figure 11.3.A). Les médianes étant ordonnées sur l'axe des abscisses, on s'attend à observer un nuage autour d'une droite de pente 1.

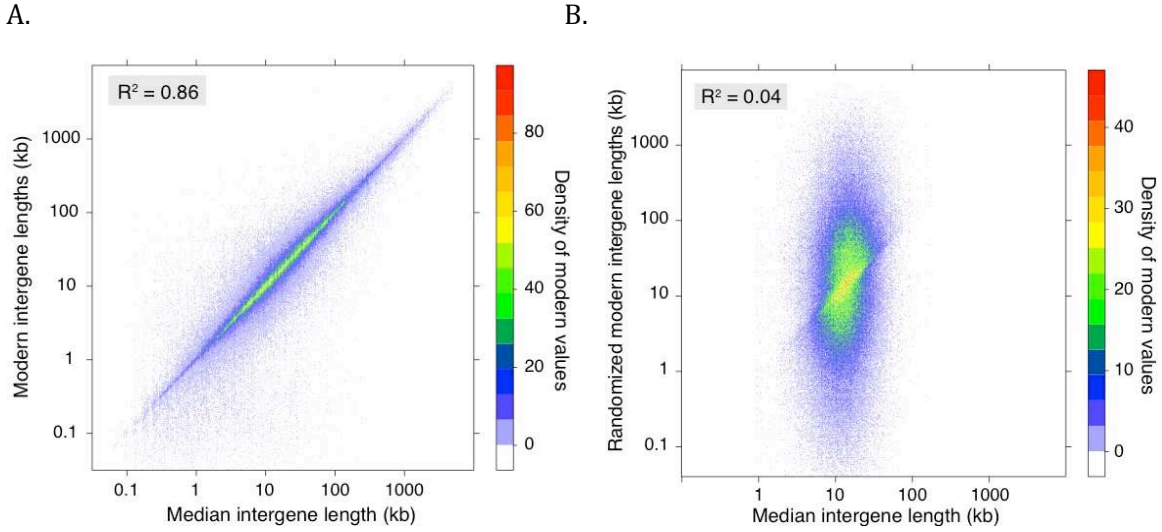


Figure 11.3. Estimation des longueurs d'intergènes ancestrales. A. Corrélation des longueurs d'intergènes orthologues à travers les génomes modernes. Pour chaque intergène ancestral, les différentes valeurs de longueurs modernes orthologues sont représentées en ordonnées contre leur valeur médiane en abscisse, utilisée comme un estimateur de la valeur ancestrale. Les points ont été groupés par classes de 0.01 en échelle log sur les deux axes, et la densité des données est représentée par un code couleur à droite. B. Corrélation observée avec les valeurs modernes randomisées.

La dispersion des valeurs autour de cette droite, décrite par le coefficient de détermination R^2, nous renseigne sur la qualité de la corrélation entre les longueurs modernes d'un intergène ancestral et la valeur médiane. Dans ce cas, la corrélation observée est remarquablement bonne, avec un R^2 de 0,86. Pour vérifier que cette corrélation n'est pas un artefact dû au fait que la valeur utilisée en abscisse est la médiane des valeurs en ordonnée, la même analyse a été réalisée en redistribuant de manière aléatoire les longueurs d'intergènes au sein de chaque espèce descendant de Boreoeutheria. Ainsi, chaque intergène se voit attribuer une longueur tirée aléatoirement de la distribution des longueurs intergéniques du génome dont il provient, abolissant tout lien évolutif entre les valeurs attribuées aux intergènes orthologues d'une espèce à l'autre. Après cette randomisation des valeurs, la corrélation observée entre les différentes longueurs pour chaque intergène et la valeur médiane est très faible, avec un R^2 de 0,05 (Figure 11.3.B).

La forte corrélation observée entre les longueurs des intergènes qui ont la même origine ancestrale suggère donc que la longueur moderne médiane est une estimation raisonnable de la valeur ancestrale. Cette observation s'explique bien dans le cadre le plus probable d'une évolution de la taille des intergènes par insertions et délétions aléatoires de segments d'ADN. Alternativement, elle pourrait également refléter une expansion globale de la taille des intergènes (plus d'insertions que de délétions) ou une contraction globale (plus de délétions que d'insertions) de manière indépendante dans tous les génomes de boreoeuthériens. Dans ce cas, les longueurs ancestrales estimées seraient légèrement sur- ou sous-évaluées par rapport à leurs valeurs réelles, mais resteraient globalement proportionnées les unes par rapport aux autres.

11.2.2. Filtrage des distances intergéniques ancestrales non fiables

Parmi les 18436 intergènes ancestraux reconstruits, tous ne présentent pas une estimation de longueur dans l'état ancestral aussi fiable les uns que les autres. Tout d'abord, dans certains cas, les longueurs modernes de l'intergène sont très variables d'une espèce à l'autre. De telles situations se produisent quand l'intergène n'existe plus que dans peu d'espèces, ou dans une majorité d'espèces dont le génome est très fragmenté et comporte de nombreux trous de séquence qui induisent des erreurs dans les longueurs d'intergènes. Par ailleurs, certains intergènes n'existent plus dans une large partie de l'arbre phylogénétique, parce qu'ils ont subi un réarrangement ou un événement génique (perte ou gain de gène) peu après la bifurcation depuis l'ancêtre Boreoeutheria. Dans ce cas, l'estimation ancestrale s'appuie essentiellement sur des espèces dont l'ancêtre commun est plus récent que Boreoeutheria : l'estimation est donc pertinente pour cet ancêtre mais pas nécessairement pour l'ancêtre Boreoeutheria.

Afin d'éliminer les estimations ancestrales non fiables, nous avons mis en place un système de filtres. Les estimations de longueur ancestrale sont retenues uniquement si elles satisfont les conditions suivantes :

- l'intergène existe dans au moins deux espèces dont le génome a été séquencé à haute couverture, et qui se trouvent dans deux des trois branches majeures de l'arbre phylogénétique (primates, rongeurs et laurasiathériens).

- l'écart interquartile des longueurs modernes (la gamme de valeurs comprises entre le 1er et le 3ème quartile de la distribution) est inférieur à 1,5x la valeur médiane. Ce filtre permet d'éliminer les estimations pour lesquelles les valeurs modernes sont manifestement très différentes entre elles, résultant en une distribution très étalée.

Après ce filtrage, nous obtenons une estimation de la longueur ancestrale pour 16115 intergènes parmi les 18436 reconstruits dans le génome de Boreoeutheria, soit 87,4% (les 12,6% restants ayant alors une longueur notée comme indéterminée).

11.2.3. Distribution des longueurs d'intergènes ancestraux

Un bon indicateur de la qualité à la fois de la reconstruction du génome de Boreoeutheria, et des estimations des longueurs d'intergènes ancestraux, est de comparer la distribution de ces longueurs par rapport à celle d'un génome moderne de bonne qualité, comme présenté à la Figure 11.4.

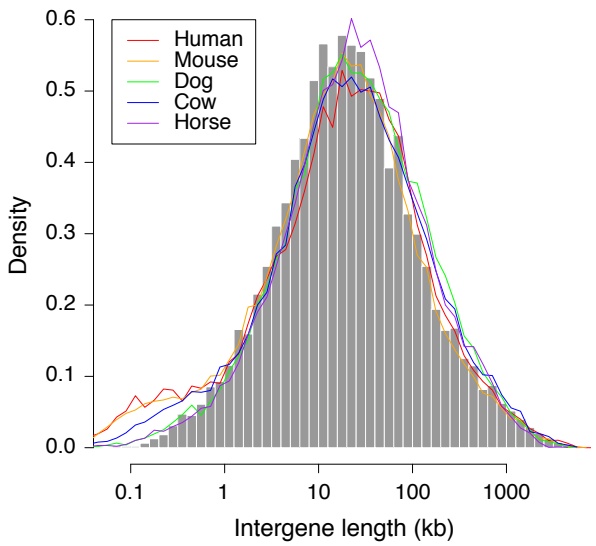


Figure 11.4. Distribution des estimations de longueur des intergènes ancestraux (en gris), et les intergènes de cinq génomes modernes.

	Taille des intergènes		
	Médiane	**Moyenne**	**Ecart-type**
Humain	13,9 kb	77,7 kb	337,8 kb
Souris	14,1 kb	69,1 kb	240,6 kb
Chien	22,3 kb	91,4 kb	242,3 kb
Vache	19,9 kb	95,5 kb	278,8 kb
Cheval	22,6 kb	79,8 kb	196,2 kb
Boreoeutheria	**19,5 kb**	**91,2 kb**	**270,2 kb**

Table 11.1. Statistiques de la distribution des tailles d'intergènes dans cinq génomes modernes et le génome ancestral de Boreoeutheria reconstruit.

La distribution des longueurs ancestrales reconstruites pour l'ancêtre Boreoeutheria est comparable à celle d'un génome moderne séquencé à haute couverture. On constate notamment que la distribution est bien log-normale, avec une médiane et une moyenne qui sont proches de

celles des cinq génomes testés (Table 11.1). Il n'y a pas de biais de reconstruction flagrant concernant des catégories spécifiques d'intergènes au niveau de la taille (par exemple les grands intergènes, qui sont moins souvent observés dans les génomes fragmentés et qui auraient pu, par conséquent, être plus difficilement reconstruits dans le génome ou plus souvent éliminés lors du filtrage des estimations de longueur). Les estimations de longueurs ancestrales semblent donc correctes. On note cependant une déplétion parmi les intergènes très courts (< 1 kb) comparé à l'homme et la souris, et dans une moindre mesure la vache. Ces trois espèces sont les seules pour lesquelles les UTR (Untranslated Regions, extrémités de transcrits non traduites) des gènes sont largement annotées dans le génome, ce qui se traduit par un excès de très courts intergènes par rapport aux autres génomes séquencés. Notre reconstruction ancestrale surestime donc probablement la longueur des intergènes les plus courts, en y incluant des séquences qui sont en réalité géniques. A l'exception de son extrémité inférieure, cependant, la distribution des longueurs intergéniques chez l'ancêtre Boreoeutheria reconstruit est très proche de celles des cinq génomes présentés.

11.3. Taux de GC intergéniques ancestraux

Parmi les propriétés du génome corrélées à la probabilité de cassure, le taux local de GC revient régulièrement dans la littérature. En effet, les points de cassure sont plus fréquents dans les régions riches en GC (voir paragraphe 3.5.2). Nous avons donc cherché à estimer le contenu en GC des intergènes dans le génome de l'ancêtre Boreoeutheria.

L'estimation du taux de GC des intergènes ancestraux repose sur les mêmes hypothèses et la même méthode que celle de leur longueur. Lorsque l'on s'intéresse à la distribution des taux de GC intergéniques orthologues (calculés sur les séquences génomiques masquées) autour de leur valeur médiane, on observe que cette distribution est en moyenne peu étalée, avec un R^2 de 0,81 (Figure 11.5.A). Comme pour la longueur des intergènes, cette corrélation est très supérieure à celle observée après une randomisation des taux de GC au sein des différents génomes modernes (Figure 11.5.B). Ce résultat suggère que le taux de GC des intergènes orthologues a peu varié depuis l'ancêtre Boreoeutheria, et que la valeur moderne médiane est une bonne approximation du taux ancestral. Il faut noter que la même expérience conduite en utilisant les valeurs de GC calculées sur les séquences non masquées (prenant donc en compte les éléments transposables) donne des résultats très similaires, avec un R^2 de 0,77. Ce résultat est en accord avec le fait que les éléments répétés tendent à adopter par mutations un contenu en GC proche de celui de leur environnement local. Cependant, les valeurs de GC calculées sur les séquences non masquées prenant nécessairement en compte des éléments transposables récents qui n'ont pas lieu d'être considérés pour une estimation de la valeur ancestrale, nous retenons par la suite les estimations obtenues avec les taux de GC calculés sur les séquences génomiques masquées.

A. B.

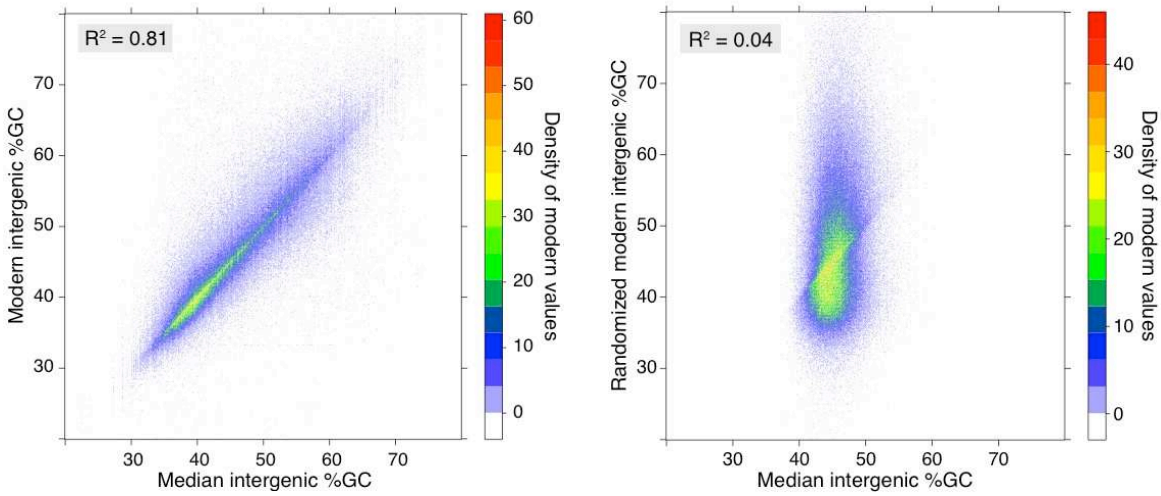


Figure 11.5. Estimation du taux de GC dans les intergènes ancestraux. A. Corrélation des taux de GC modernes orthologues (calculés sur la séquence masquée). Pour chaque intergène ancestral, les différentes valeurs de GC modernes orthologues sont représentées en abscisse contre leur valeur médiane, utilisée comme une estimation de la valeur ancestrale. Les données ont été groupées en classes de 0,1% sur chaque axe, et la densité des données est représentée par le code couleur à droite. B. Corrélation après randomisation des valeurs modernes.

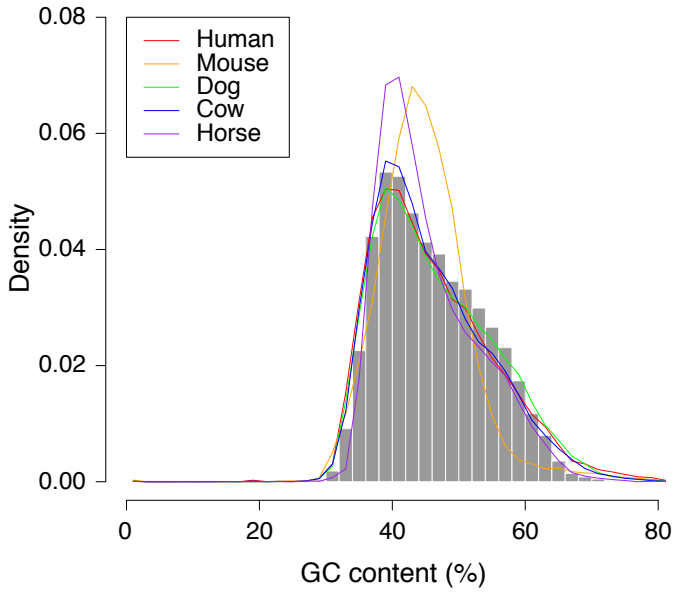


Figure 11.6. Distribution des taux de GC intergéniques estimés dans le génome de l'ancêtre Boreoeutheria (en gris) et de cinq génomes modernes.

Afin d'éliminer les estimations de taux de GC ancestrales peu soutenues par les données modernes, les filtres appliqués sont les suivants :

- de même que pour la longueur ancestrale, l'intergène doit exister dans au moins deux espèces dont le génome a été séquencé à haute couverture, dans deux des trois branches majeures de l'arbre phylogénétique.
- l'écart interquartile des taux de GC modernes doit être inférieur à 10%.

Après cette étape de filtrage, nous obtenons une estimation fiable du taux de GC pour 15856 intergènes ancestraux parmi les 18436 reconstruits dans le génome de Boreoeutheria, soit 86,0%. La distribution des valeurs ancestrales s'approche à nouveau de la distribution observée dans un génome de boreoeuthérien séquencé à haute couverture (Figure 11.6).

11.4. Pression de sélection sur les interactions de régulation

Nous avons également souhaité intégrer à nos analyses une troisième propriété des intergènes ancestraux mesurant la contrainte exercée sur l'organisation du génome par le contenu en éléments de régulation. De précédents travaux dans la littérature démontrent en effet que dans certaines régions du génome s'exerce une pression de sélection négative contre les réarrangements qui maintient l'organisation des gènes et de leurs éléments de régulation à longue distance, résultant en des blocs de régulation génomique (GRB) réfractaires à la cassure (Engstrom et al. 2007; Kikuta et al. 2007)(voir paragraphe 3.6.3). Cependant, l'existence de cette pression de sélection n'a pu être démontrée que dans quelques exemples précis ; la proportion du génome incluse dans de tels GRB n'est pas connue, et l'on ne sait pas à ce jour si cette sélection négative a un impact significatif sur la distribution des points de cassure à l'échelle du génome ou s'il s'agit d'un phénomène anecdotique qui ne concerne que quelques gènes très importants et fortement régulés. Afin d'étudier l'impact des éléments de régulation sur le maintien de l'organisation des gènes, nous avons utilisé deux approches.

11.4.1. Contenu ancestral en éléments conservés non-codants

L'identification exhaustive des éléments de régulation et de leurs gènes cibles dans les génomes de mammifères est un problème notoirement compliqué, qui n'est pas encore résolu. Comme il n'existe pas de carte à l'échelle du génome traçant les liens entre éléments de régulation et gènes cibles, nous nous sommes servis du contenu local en éléments conservés non-codants comme première approximation de la contrainte sélective s'exerçant contre les réarrangements. Les éléments conservés non-codants (CNE, Conserved Non-coding Elements) que l'on peut retrouver dans les génomes de mammifères correspondent souvent à des enhancers ou plus généralement à des sites de fixation de facteurs de transcription (Woolfe et al. 2005; Pennacchio et al. 2006; Kikuta et al. 2007). Par ailleurs, les gènes cibles de blocs de régulation génomique (GRB) connus se trouvent généralement dans des régions très denses en éléments conservés non-codants (Nobrega et al. 2003; Sandelin et al. 2004; Becker and Lenhard 2007; Dong et al. 2009). Ces caractéristiques suggèrent que la densité en éléments conservés non-codants pourrait être utilisée comme un proxy reflétant la complexité et la force des

interactions locales de régulation. On pourrait objecter à cette approche que les éléments conservés non-codants sont loin d'être les seuls éléments de régulation existant dans les génomes de mammifères, la plupart des sites de fixation de facteurs de transcription ayant une durée de vie courte et faisant l'objet d'un turn-over rapide (Schmidt et al. 2010). Cependant, les éléments de régulation à turn-over rapide sont souvent redondants et peu susceptibles d'induire une contrainte forte et prolongée sur l'organisation des gènes, de par leur nature transitoire. Les éléments conservés, en revanche, sont potentiellement des sites de régulation dont la localisation par rapport à leur gène cible doit être conservée afin de préserver leur fonction, induisant une contrainte sélective sur la région concernée.

Pour reconstruire le contenu ancestral en éléments contraints, nous avons utilisé les CNEs identifiés par la méthode GERP (Cooper et al. 2005) dans les génomes des 28 boreoeuthériens descendants de Boreoeutheria à partir de l'alignement multiple de 35 mammifères euthériens disponible via Ensembl (voir paragraphe 7.4). La reconstitution des éléments non-codants précis existants dans chaque intergène de Boreoeutheria à partir de ces données revient à étendre la méthode de reconstruction des génomes ancestraux à des données incluant gènes et éléments non-codants orthologues ; il s'agit d'un problème complexe qui est au-delà des objectifs de ce travail de thèse. Nous avons opté pour une stratégie moins résolutive mais plus simple qui se base sur le même principe que les estimations de longueur et de GC des intergènes ancestraux. Nous nous sommes intéressés à la longueur totale couverte par des CNEs dans chaque espèce pour un même intergène orthologue : sous l'hypothèse que les CNEs détectés dans des intergènes orthologues sont les mêmes éléments dans toutes les espèces, on s'attend à ce que la longueur totale des CNEs soit proche dans les différentes espèces (plus ou moins les éléments perdus ou dupliqués). En effet, lorsqu'on regarde la distribution des longueurs totales de CNEs dans chaque espèce par rapport à la médiane de ces longueurs parmi les intergènes orthologues, on constate que la distribution est peu dispersée et que les valeurs sont bien corrélées entre elles à l'échelle du génome, avec un R^2 de 0,82 (Figure 11.7.A). Cette corrélation est très supérieure à celle observée après une randomisation des valeurs modernes au sein de chaque génome (Figure 11.7.B). Nous avons donc utilisé la valeur moderne médiane comme une estimation de la longueur totale des éléments sous contrainte chez l'ancêtre.

Une inspection des estimations montre que 34% des intergènes ancestraux ne contiennent aucun CNE (taille totale estimée de 0 pb), et que 12,5% supplémentaires présentent de fortes disparités dans les longueurs totales de CNEs modernes (écart interquartile > 1,5x la médiane). Il est probable que dans une partie de ces régions, soit les séquences intergéniques modernes sont incomplètes dans certains génomes (trous de séquence), soit la qualité de l'alignement est insuffisante, ce qui empêche la détection correcte des CNEs.

A. B.

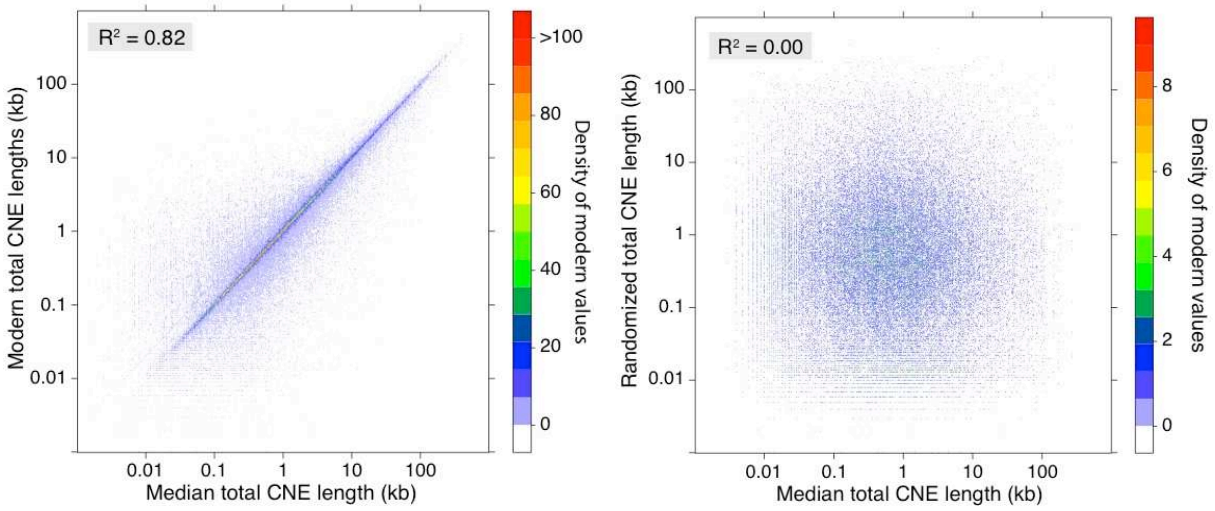


Figure 11.7. Estimation de la proportion de séquence conservée non-codante dans les intergènes ancestraux. A. Corrélation des proportions de CNE modernes orthologues. Pour chaque intergène ancestral, les différentes proportions de CNE modernes orthologues sont représentées en abscisse contre leur valeur médiane, utilisée comme une estimation de la valeur ancestrale. Les données ont été groupées en classes de 0.01 en échelle log sur chaque axe, et la densité des données est représentée par le code couleur à droite. B. Corrélation après randomisation des valeurs modernes.

Nous avons pourtant choisi de garder les estimations dans tous les intergènes pour lesquelles la longueur totale de l'intergène a pu être reconstruite de façon fiable, en nous basant sur les observations suivantes :

- le coefficient de détermination de la distribution est élevé (0,82), suggérant que globalement la corrélation est bonne.
- retirer les intergènes ne vérifiant pas la condition « écart interquartile des valeurs modernes < 1,5x la médiane » biaise largement le jeu d'intergènes ancestraux avec une estimation vers les intergènes les plus longs (Figure 11.8). Ce filtre retire en effet de nombreux intergènes courts avec une taille totale de CNEs très petite ou nulle mais qui contiennent tout de même des CNEs identifiés entre certaines espèces, résultant en un écart interquartile relativement important dans la distribution moderne. Dans la majorité des cas, ces CNEs sont identifiés entre un sous-groupe d'espèces plus proches que les boreoeuthériens : ces CNEs sont probablement apparus postérieurement à l'ancêtre Boreoeutheria, si bien que la valeur faible ou nulle obtenue par l'estimation par la médiane est probablement correcte au nœud ancestral Boreoeutheria.
- si l'on filtre les intergènes sur les mêmes critères que ceux appliqués pour la longueur de l'intergène (écart interquartile < 1.5x la médiane, et valeur ancestrale soutenue par des génomes de haute qualité dans au moins deux grands clades descendants de l'ancêtre), l'analyse portant sur les 53,5% des intergènes restants donne des résultats très similaires à ceux obtenus sur l'ensemble du jeu (mais avec un pouvoir statistique plus limité).

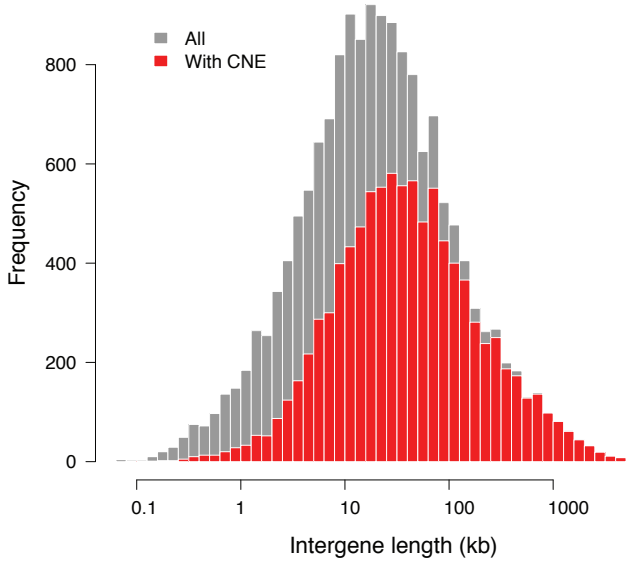


Figure 11.8. Biais sur les tailles d'intergènes introduit par l'exclusion des intergènes ancestraux dont la proportion de CNE estimée est nulle.

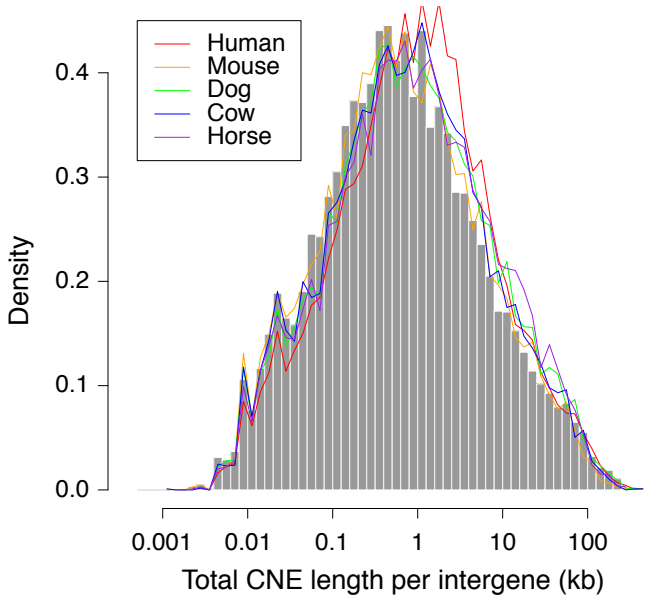


Figure 11.9. Distribution des proportions intergéniques de séquences conservées non-codantes estimées dans le génome de l'ancêtre Boreoeutheria (en gris) et de cinq génomes modernes.

La distribution des longueurs totales de CNE dans les intergènes de Boreoeutheria s'approche ici aussi de la distribution que l'on observe dans un génome de boreoeuthérien séquencé à haute couverture (Figure 11.9). On constate que pour les intergènes riches en séquences conservées, les estimations chez l'ancêtre sont globalement biaisées vers des valeurs un peu plus petites que celles des génomes modernes, bien que ce biais ne se ressente pas sur les moyennes (3,8 kb de séquence conservée en moyenne chez l'ancêtre contre 3,9 kb pour

l'humain, 2,3 kb pour la souris, 3,5 kb pour le chien, 3,0 kb pour la vache, 3,2 kb pour le cheval). Cependant, comme les estimations ancestrales ne seront utilisées dans la suite de l'analyse que pour définir deux groupes d'intergènes (fort et faible pourcentage de séquence conservée), cette légère sous-estimation générale des valeurs ancestrales réelles n'est pas un problème majeur.

11.4.2. Utilisation de gènes cibles de blocs de régulation génomique

Dans le cadre de ce projet, nous avons également eu accès à des données de prédictions informatiques de gènes cibles de GRB dans le génome humain (B. Lenhard et A. Akalin, communication personnelle ; voir paragraphe 7.5). Ces prédictions sont basées notamment sur la conservation de synténie de la région orthologue entre le génome humain et du poisson zèbre, ainsi que sur la densité en éléments conservés non-codants détectés entre ces deux espèces. Les données se présentent sous la forme d'une liste de gènes humains qui sont probablement la cible d'un ensemble d'éléments non-codants intercalés parmi les gènes du voisinage. Etant donné que l'ancêtre Boroeutheria descend de l'ancêtre Euteleostomi, le dernier ancêtre commun entre l'humain et le poisson zèbre, tout gène qui serait cible d'un groupe de séquences de régulation chez ces deux espèces l'est également dans le génome de Boreoeutheria. Nous pouvons donc directement transposer les prédictions dans le génome humain sur celui de Boreoeutheria, et faire la supposition que les intergènes flanquant ces gènes cibles sont soumis à une pression de sélection négative contre les réarrangements, qui perturberaient le circuit de régulation de ces gènes.

Il est important de noter que ces prédictions font intervenir la conservation de l'ordre des gènes entre l'homme et le poisson zèbre, a fortiori donc entre Boreoeutheria et l'humain. Utiliser ces gènes cibles pour modéliser la distribution des points de cassure survenus entre l'ancêtre Boreoeutheria et l'homme pose donc un problème de circularité : on infère que l'intergène est sous pression de sélection parce qu'il n'a pas été réarrangé chez l'homme, puis on fait l'hypothèse qu'il est sous pression de sélection pour expliquer l'absence de cassure. Il sera donc nécessaire d'exclure les points de cassure identifiés dans la lignée de l'homme pour utiliser ces données. En revanche, avec les autres lignées de boreoeuthériens, le problème de circularité n'existe pas. Si l'hypothèse de départ est fausse (ces intergènes ne sont pas spécialement sous sélection, ils auraient été conservés entre l'homme et le poisson zèbre par hasard), alors ces intergènes ne devraient pas avoir de comportement particulier dans les autres espèces. S'ils se révèlent moins souvent cassés toutes choses égales par ailleurs, alors on pourra en conclure que l'hypothèse de départ était valide et que ces intergènes sont sous une contrainte de sélection.

Chapitre 12. Points de cassure évolutifs dans cinq génomes de mammifères

L'objectif de ce projet est la modélisation de la distribution des points de cassure depuis le génome ancestral jusqu'aux génomes modernes. Dans ce chapitre, nous abordons la méthode développée afin d'identifier les points de cassure de réarrangements évolutifs dans cinq génomes de mammifères qui descendent de l'ancêtre Boreoeutheria, dont l'ordre des gènes et certaines caractéristiques génomiques locales ont été reconstruits au chapitre précédent. Nous montrons ici que les points de cassure ainsi trouvés représentent un jeu plus complet que de précédentes études, mais présentent les mêmes caractéristiques que celles fréquemment rapportées dans la littérature.

12.1. Identification des points de cassure

Afin d'obtenir un jeu de points de cassure de réarrangements évolutifs, nous nous sommes intéressés à cinq génomes de mammifères descendants de l'ancêtre Boreoeutheria : l'homme (*Homo sapiens*), la souris (*Mus musculus*), le chien (*Canis familiaris*), la vache (*Bos taurus*) et le cheval (*Equus caballus*). Ces génomes ont été choisis pour la qualité de leur assemblage, ainsi que pour leur radiation rapide après l'ancêtre Boreoeutheria (Nery et al. 2012), si bien que les événements de cassure identifiés dans ces génomes sont essentiellement indépendants.

La méthode d'identification des points de réarrangements a été développée au laboratoire par Matthieu Muffato. Brièvement, l'identification des points de cassure se fait en comparant l'ordre et l'orientation des gènes dans le génome de l'ancêtre Boreoeutheria avec celui de chaque espèce moderne pour identifier des régions où un gène moderne (ou plusieurs gènes consécutifs) a un voisin différent par rapport à l'état ancestral. Ces régions peuvent correspondre à un intergène unique ou un train d'intergènes successifs. Dans certaines de ces régions, la perturbation est due à des pertes ou des gains de gènes mais pas à une réorganisation des gènes existants, ce que nous ne considérons pas comme un réarrangement au sens strict (pas de perte de la synténie). Afin d'exclure ces régions de l'analyse, les génomes de l'ancêtre et de l'espèce moderne ont été réduits aux gènes présents en exactement une copie dans chacun des génomes. Les régions où les adjacences de gènes restent perturbées entre l'ancêtre et le moderne sont des régions qui contiennent un point de cassure de réarrangement. Dans certains cas, cependant, les régions de cassure sont affectées à la fois par un réarrangement et par des événements de pertes ou de gains de gènes. Comme nous ne sommes pas en mesure de dater si l'événement de perte (ou de gain) est antérieur, concomitant ou postérieur au réarrangement, nous ne connaissons pas l'état ancestral exact au moment du réarrangement. Ainsi, bien que ces points de cassure soient probablement valides, ils n'ont pas été retenus par la suite car ils ne

peuvent pas être analysés par la méthode de régression que nous proposons par la suite. Nous reviendrons sur cette limite de la méthode en fin de chapitre.

Après inspection manuelle d'une partie des réarrangements identifiés par cette méthode, nous avons constaté la nécessité de filtrer les résultats pour éliminer des cas de faux positifs probables apparaissant régulièrement. La méthode d'identification est en effet particulièrement sensible aux erreurs dans les génomes ancestraux et modernes, qu'elle détecte systématiquement comme un point de cassure de réarrangement. Les points de cassure ont été rejetés lorsqu'ils présentent l'une des caractéristiques suivantes :

- point de cassure dans un intergène ancestral qui n'existe dans aucun génome non boreoeuthérien (groupe externe) : dans ce cas, une erreur au niveau de la reconstruction de l'état ancestral est probable.
- de même, point de cassure dans un intergène ancestral qui ne présente qu'un support moderne médiocre (cet intergène n'existe pas dans un groupe d'espèces de référence choisies parmi les génomes boreoeuthériens séquencés à haute couverture).
- deux points de cassure identifiés dans deux intergènes consécutifs (déplacement d'un gène unique) : dans la très grande majorité des cas, l'inspection manuelle révèle qu'il s'agit d'une erreur d'assemblage ou d'annotation dans le génome moderne.
- point de cassure entre deux gènes dont l'un est monoexonique dans le génome moderne, ou bien se trouve dans la même orientation à proximité immédiate dans le génome ancestral : à nouveau, ces points de cassure sont en général des erreurs d'annotation dans le génome moderne.

Par cette méthode, nous avons identifié un total de 779 points de cassure dans un intergène ancestral dans les cinq lignées considérées. Parmi eux, 100 se trouvent sur la branche de l'homme, 176 sur celle de la souris, 116 sur celle du chien, 305 sur celle de la vache et 82 sur celle du cheval (Figure 12.1). Parmi eux, 20 intergènes ont été cassés au moins deux fois de façon indépendante. En revanche, 24 sont identifiés comme cassés dans deux ou trois espèces qui présentent une courte branche commune dans l'arbre et sont probablement des événements uniques qui se sont produits avant la spéciation : ils sont considérés comme tels dans la suite de l'analyse, ce qui donne un total de 751 points de cassure réels. Les comptages observés sont du même ordre de grandeur que ceux obtenus dans d'autres analyses avec des méthodes différentes (Larkin et al. 2009; Zhao and Bourque 2009).

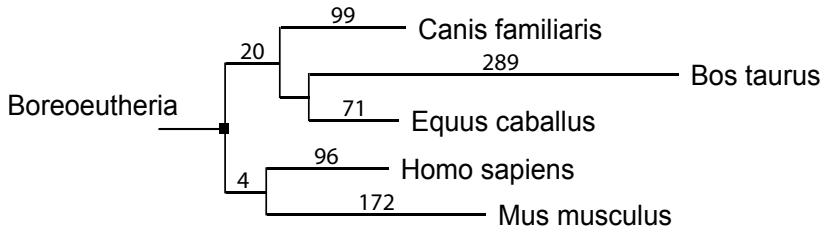


Figure 12.1. Arbre phylogénétique des cinq génomes boreoeuthériens utilisés dans l'analyse des points de cassure de réarrangements. Les longueurs de branche correspondent au nombre de points de cassure par branche.

12.2. Comparaison aux données de Larkin et al. 2009

Nous avons comparé la liste des points de cassure identifiés par notre méthode avec ceux publiés dans (Larkin et al. 2009). La méthode de Larkin et al. est différente de la nôtre : elle ne fait pas intervenir de reconstruction du génome ancestral, et les régions de cassures (notées LBR ci-dessous, pour Larkin's Breakpoint Regions) y sont définies par défaut par rapport aux régions conservées entre les différents génomes. Les auteurs se basent sur la comparaison de l'ordre de marqueurs orthologues dans huit génomes entièrement séquencés (homme, chimpanzé, macaque, souris, rat, chien, opossum et poulet) et deux cartes génétiques à haute résolution (vache et porc). Une région est conservée entre deux génomes lorsque l'ordre et l'orientation des marqueurs est conservé sur une distance d'au moins 500 kb. Toute région où les marqueurs ne répondent pas à ces critères est une région de cassure, quelque soit sa taille ou le nombre de marqueurs qu'elle contient. Le génome ou le phylum dans lequel s'est produite la cassure peut être déduit des différentes comparaisons des génomes entre eux. Ainsi, 433 points de cassure sont identifiés depuis l'ancêtre Boreoeutheria dans les génomes de l'homme, de la souris, du chien et de la vache, ainsi que dans la branche commune à la vache et au cheval (notée « laurasiathériens » dans la comparaison puisqu'ils correspondent à cette branche dans notre arbre). Les régions de cassure sont fournies sous la forme de coordonnées reportées dans le génome humain.

Après un transfert des coordonnées des LBR vers la version hg19 du génome humain (liftover), les LBR ont été comparées aux points de cassure identifiés par notre méthode. Nos points de cassure étant identifiés dans le génome ancestral, la comparaison n'est pas directe. Les gènes humains se trouvant dans ou de part et d'autre d'une LBR sont comparés aux gènes ancestraux flanquant les intergènes cassés dans chaque lignée : si certains des gènes de LBR descendent de ces gènes ancestraux, alors on a recouvrement entre une LBR et l'un de nos points de cassure. Soixante pourcents des LBR sont correctement identifiés comme une région de cassure par notre méthode (Table 12.1).

	Human	Mouse	Dog	Cow	Laurasiatherians	**Total**
Number of LBR	81	166	75	99	12	**433**
After liftover to hg19	60	165	75	96	12	**408**
LBR overlapping breakpoints						
Counts	45	98	39	57	7	**246**
Percents	75%	59%	52%	59%	58%	**60%**
Number of breakpoints in LBR						
Counts	58	114	41	62	13	**288**
Percents	58%	65%	41%	21%	65%	**38%**
Average per LBR	1.29	1.16	1.05	1.09	1.86	**1.17**

Table 12.1. Intersection entre les points de cassure de notre jeu de données et de celui de Larkin et al. (2009), notés LBR.

Une inspection manuelle d'une partie des 40% restants révèle qu'ils correspondent approximativement pour moitié à des régions complexes comprenant des pertes et gain de gènes, éliminées de notre analyse, et pour l'autre moitié à des faux positifs parmi les LBR, dus à des erreurs d'assemblage dans les versions des génomes utilisées, qui ont été corrigées depuis. Plus de 60% de nos points de cassure sont des événements nouveaux par rapport aux LBR, dont la méthode d'identification a une résolution beaucoup moins fine que la nôtre et ne peut pas identifier les événements locaux (notamment les inversions inférieures à la résolution des blocs utilisés).

12.3. Caractéristiques des points de cassure

En termes de fiabilité de la reconstruction ancestrale, les intergènes contenant une cassure sont soutenus en moyenne par 11,2 génomes boreoeuthériens, contre 13,8 pour les intergènes ne contenant aucun point de cassure. Cette différence est attendue, puisque ces intergènes sont réarrangés : ils n'existent donc plus dans au moins un génome, et dans plusieurs si le réarrangement n'a pas eu lieu sur la branche terminale de l'arbre. Cependant, la grande majorité présente un indice de confiance comparable au reste du génome ancestral, puisque 88% sont préservés dans au moins 5 génomes boreoeuthériens modernes (90% pour l'ensemble du génome ; test binomial : P = 0,34). Parmi les 751 cassures identifiées, 682 sont localisées dans des intergènes pour lesquels nous avons reconstruit une longueur dans l'état ancestral, soit autant que l'ensemble du génome (91% contre 87% ; test binomial : P = 0,56). Seuls ces 682 intergènes sont accessibles à la suite de l'étude de modélisation par régression de Poisson.

Les points de cassure ont été associés de manière récurrente à différentes caractéristiques génomiques dans la littérature, comme présenté au paragraphe 3.5.2. Les régions de cassure sont notamment statistiquement plus riches en GC et plus denses en gènes qu'attendu au hasard (Murphy et al. 2005; Ma et al. 2006; Gordon et al. 2007; Larkin et al. 2009; Lemaitre et al. 2009). Plus précisément, les points de cassure tendent à se produire dans des intergènes plus longs que la moyenne des intergènes du génome (Peng et al. 2006), mais plus courts que ceux attendus sous une distribution aléatoire des cassures, les tailles d'intergènes ne suivant pas une distribution normale (Lemaitre et al. 2009). Par ailleurs, les points de cassure se produisent dans les régions contenant moins d'éléments de séquence conservés qu'attendu au hasard (Larkin et al. 2009). Nous avons testé si les cassures identifiées dans le génome de Boreoeutheria vérifient également ces caractéristiques (Table 12.2). Les points de cassure concordent et confirment effectivement les observations de la littérature.

	Distribution Average	Random expectation	Breakpoints	Significance
GC	45.9 %	40.7 %	**44.5 %**	***
Intergene length	91179 bp	881727 bp	**179352 bp**	***
Conserved sequence	2.4 %	4.4 %	**2.4 %**	***

Table 12.2. Caractéristiques des intergènes affectés par des points de cassure dans le génome de Boreoeutheria.

Comme précisé au paragraphe 12.1, l'une des limites principales de notre approche d'étude des points de cassure est que ces ruptures de synténie doivent être identifiées à l'intergène ancestral près : une région de cassure de plusieurs intergènes ne peut pas être incluse dans l'analyse par régression de Poisson. Ainsi, tous les événements complexes impliquant des gains ou pertes de gènes flanquant un événement de cassure ont été éliminés car non analysables par la suite. Nous constatons cependant que les points de cassure retenus recouvrent la majorité et présentent les mêmes caractéristiques que ceux publiés précédemment dans la littérature. Ainsi, il ne semble pas y avoir de biais dans notre jeu de points de cassure : les événements complexes semblent concerner les régions de cassure de manière aléatoire, et les éliminer ne perturbe donc pas la suite des analyses.

Chapitre 13. Identification des facteurs influant sur la cassure

Cette partie présente la modélisation de la distribution des points de cassure depuis l'ancêtre Boreoeutheria dans les cinq lignées modernes analysées. Afin de tester si les caractéristiques génomiques reconstruites dans chaque intergène ancestral influence sa probabilité de cassure, nous avons utilisé une modélisation par régression de Poisson. La régression de Poisson, présentée plus en détails dans le chapitre 10, est une méthode de régression suivant un modèle linéaire généralisé qui permet de décrire la probabilité d'occurrence d'un événement rare (ici, une cassure) dans un groupe d'intervalles (ici, les intergènes ancestraux) en fonction des caractéristiques de ces intervalles. Dans la plupart des cas, le taux d'apparition d'un événement rare par intervalle peut en effet être modélisé comme une somme pondérée de différents paramètres après une transformation logarithmique.

13.1. Influence majeure de la longueur des intergènes

Comme évoqué au chapitre 11, sous l'hypothèse nulle d'une distribution aléatoire des points de cassure dans le génome, on s'attend à ce que le taux de cassure par intergène soit proportionnel, en moyenne, à la longueur des intergènes (distribution de Poisson). Ainsi, pour chaque classe d'intergènes de longueur moyenne L, le taux de cassure moyen attendu r, qui correspond à l'espérance de la distribution du nombre de cassures par intergène $E(X|L)$, est :

$$r = E(X|L) = a * L$$

Nous avons calculé le taux de cassure par Mb attendu sous une distribution aléatoire, c'est-à-dire où chaque base du génome a la même probabilité de cassure, en divisant le nombre total de cassures par la longueur cumulée totale des intergènes du génome ancestral. Ainsi, on peut calculer le taux de cassure r attendu au hasard pour toute classe d'intergènes de longueur moyenne L :

$$r = \frac{N\ cassures}{\sum longueurs} * L = 4{,}61.10^{-7} * L$$

C'est-à-dire, après une transformation logarithmique [3] (nécessaire à la régression de Poisson) :

$$\log(r) = \log(L) - 14{,}59$$

Ainsi, dans une représentation logarithmique du taux de cassure par intergène en fonction de la longueur des intergènes, on s'attend à observer une droite de pente 1. Si d'autres facteurs

3 Pour rappel, on utilise dans ce manuscrit la notation log pour le logarithme népérien.

influencent la cassure, sous les hypothèses de la régression de Poisson, *log(r)* sera une fonction linéaire de *log(L)* et de ces autres paramètres :

$$\log(r) = \alpha * \log(L) + \beta * M + \gamma * N + \cdots + \delta$$

Les intergènes du génome de l'ancêtre Boreoeutheria ont été segmentés par classes de longueur de 0,5 en 0,5 sur une échelle log, et le taux de cassure moyen par intergène a été calculé pour chaque classe. Après régression de Poisson, nous trouvons que le taux de cassure est en effet hautement corrélé à la longueur moyenne des intergènes, et augmente régulièrement avec la longueur (z-test : $P < 2.10^{-16}$; Figure 13.1). Cependant, la pente de la régression n'est pas 1 comme attendu sous l'hypothèse nulle, mais 0,28. L'équation de régression obtenue est :

$$\log(r) = 0{,}28 * \log(L) - 6{,}04 \quad \leftrightarrow \quad r = 2{,}4\,.\,10^{-3} * L^{0{,}28}$$

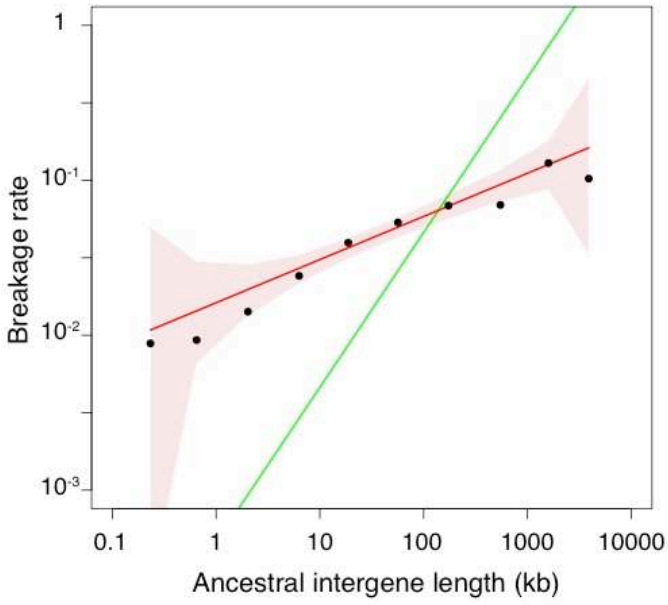


Figure 13.1. Corrélation entre la longueur des intergènes ancestraux (abscisse) et leur taux de cassure moyen dans les lignées descendantes (ordonnée). Après une transformation logarithmique, le taux de cassure est linéairement corrélé à la longueur de l'intergène. Le modèle de régression obtenu est différent de l'attendu sous le modèle aléatoire classique (ligne rouge : équation de régression ; zone colorée : intervalle de confiance à 95% du modèle ; ligne verte : attendu sous une distribution aléatoire pure).

L'excellente corrélation entre la longueur des intergènes et leur taux de cassure suffit à expliquer l'essentiel de la variation des taux de cassure dans le génome de l'ancêtre (Pseudo R^2 de McFadden = 0,93). Les différences entre les taux observés et ceux du modèle ne sont pas significatives d'après un test de Chi^2 sur la déviance résiduelle et peuvent donc être attribuées à du bruit statistique (P = 0,19). Malgré cette corrélation très forte, la distribution s'éloigne de la distribution purement aléatoire attendue sous l'hypothèse nulle : bien que les longs intergènes soient plus souvent cassés que les petits en nombre absolu de cassures, les intergènes courts ont été plus souvent cassés qu'attendu au hasard, et les intergènes longs l'ont moins été. Ce résultat

est cohérent avec les observations paradoxales de la littérature, où les points de cassure sont décrits en même temps comme affectant des intergènes plus longs que la moyenne (Peng et al. 2006) mais également comme plus fréquents dans les régions denses en gènes (donc flanqués de petits intergènes)(Murphy et al. 2005; Ma et al. 2006; Larkin et al. 2009; Lemaitre et al. 2009). L'augmentation du taux de cassure est ici proportionnelle à $L^{0,28}$, soit moins rapide que l'augmentation de la longueur des intergènes. Ainsi, l'espacement entre les gènes semble conditionner la probabilité de cassure, mais d'une manière plus complexe qu'une simple relation de proportionnalité entre les deux valeurs.

13.2. Influence du taux de GC

Les génomes de mammifères sont organisés en isochores, c'est-à-dire en une alternance de régions denses en gènes et riches en GC avec des régions moins denses en gènes et riches en AT (Bernardi 2000). Le taux de GC intergénique est alors corrélé à la longueur des intergènes, une propriété également vérifiée par le génome ancestral Boreoeutheria (Figure 13.2). Nous avons donc testé si les variations du taux de GC dans le génome ancestral peuvent expliquer les variations du taux de cassure d'une manière plus simple que la longueur des intergènes.

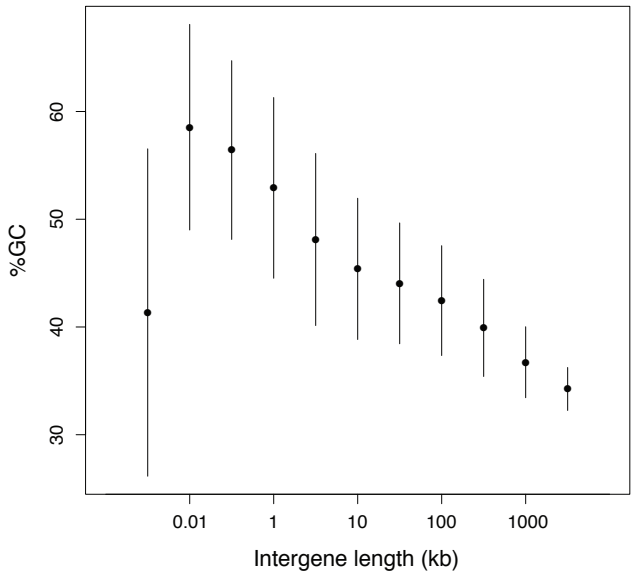


Figure 13.2. Corrélation entre le taux de GC et la longueur des intergènes dans le génome ancestral Boreoeutheria (point : valeur moyenne de la classe d'intergènes ; barre : écart-type).

Tout d'abord, la corrélation entre le taux de cassure et le taux de GC seul a été mesurée. Comme pour la longueur des intergènes, nous avons segmenté les intergènes du génome ancestral en classes de GC par tranches de 5%, et calculé le taux de cassure moyen par intergène pour chaque classe. Le taux de cassure est bien corrélé au taux de GC (z-test : $P = 1.10^{-7}$) : il diminue quand le taux de GC augmente (Figure 13.3). Le taux de GC étant plus élevé dans les petits intergènes, ce résultat est cohérent avec ceux obtenus sur les longueurs des intergènes. La corrélation est légèrement moins bonne que celle obtenue entre le taux de cassure et la longueur des intergènes, mais le modèle basé sur le taux de GC suffit lui aussi à expliquer la variabilité du

taux de cassure (Pseudo R^2 de McFadden = 0,68 ; test de Chi^2 sur la déviance résiduelle : P = 0,28). L'équation obtenue est :

$$\log(r) = -0{,}028 * \%GC - 1{,}90 \quad \leftrightarrow \quad r = \frac{0{,}97^{\,\%GC}}{6{,}69}$$

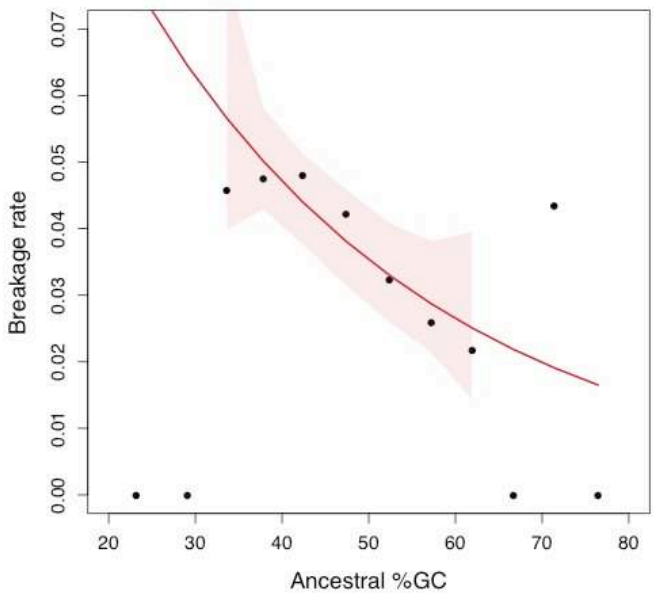


Figure 13.3. Corrélation entre le taux de GC des intergènes ancestraux (abscisse) et leur taux de cassure moyen dans les lignées descendantes (ordonnée). Ligne rouge : équation de régression ; zone colorée : intervalle de confiance à 95% du modèle (non représenté pour les extrêmes, où l'intervalle de confiance est très vaste en raison du nombre restreint d'intergènes). Les données ne sont pas représentées dans un graphe logarithmique afin de visualiser les points à 0.

Afin de tester si la corrélation observée entre le taux de GC et le taux de cassure est une conséquence de la corrélation entre la longueur des intergènes et le taux de GC, nous avons mené une régression de Poisson multivariée par procédure progressive (voir paragraphe 10.4). Cette méthode consiste à faire une première régression du taux de cassure sur l'une des variables explicatives, puis d'ajouter progressivement des variables supplémentaires une à une. On compare les modèles à chaque étape : si l'ajout d'une nouvelle variable améliore le modèle, on conserve cette variable, sinon on la rejette et on s'en tient au modèle plus simple. Pour cette analyse, les intergènes ont été segmentés en classes de longueur et de GC homogène (de 0,5 en 0,5 sur une échelle log pour les longueurs, puis par tranches de 20% sur le taux de GC). Les classes choisies sont plus larges pour le GC que lors de la modélisation précédente, pour préserver un bon équilibre entre le nombre de classes d'intergènes (pouvoir résolutif de la régression) et le nombre d'intergènes par classe (puissance statistique). Cependant, la même procédure avec des segmentations différentes des données amène à des résultats similaires.

La régression progressive donne les résultats résumés dans la table 13.1. Le coefficient de régression obtenu dans la régression simple basée sur la longueur des intergènes est légèrement différent de celui obtenu dans l'analyse précédente en raison de la segmentation différente des données. On constate que lorsqu'on introduit le taux de GC conjointement à la longueur des

intergènes, le taux de GC devient non significatif dans la régression. Cet ajout n'améliore pas le modèle puisqu'un test de Chi² sur la différence des déviances résiduelles entre deux modèles n'est pas significatif (P = 0,42). Ceci est confirmé par le critère d'information d'Akaike (AIC), qui augmente entre les deux modèles. Ainsi, ces résultats suggèrent que c'est bien la longueur des intergènes qui détermine le taux de cassure des intergènes, et que la corrélation entre taux de GC et cassures n'est qu'une conséquence secondaire de la corrélation entre taux de GC et longueur des intergènes. Ces résultats confirment des observations précédemment publiées dans la littérature, liant les cassures à la densité en gènes plutôt qu'au taux de GC (Lemaitre et al. 2009).

	Coefficients					Goodness of fit		
	Simple regression	Stepwise regression	P(>|z|)	Null deviance (df)	Residual Deviance (df)	χ^2 P-value	Stepwise χ^2 P-value	Pseudo R^2
Model 2 : length + %GC								
Intergene length	0.26	0.27	**< 2.10^{-16}**	137.8 (28)	25.7 (27)	**0.53**	-	**0.81**
%GC	-	0.003	0.44	137.8 (28)	25.1 (26)	0.52	0.42	0.82

Table 13.1. Résultats de la procédure de régression progressive du taux de cassure sur la longueur des intergènes et le taux de GC.

La régression multivariée pourrait cependant être mise en défaut par cette même corrélation entre taux de GC et longueur des intergènes. En effet, lorsque deux variables fortement corrélées sont utilisées dans une régression multivariée, des problèmes de multicollinéarité peuvent se poser et amener à éliminer la variable causative au profit de la variable qui lui est corrélée. Afin d'éliminer formellement le taux de GC comme facteur influençant la cassure, nous avons donc comparé, dans des classes de GC homogène, la longueur des intergènes ancestraux avec ou sans cassure (Figure 13.4.A), et réciproquement, dans des classes de longueur homogène, le taux de GC des intergènes avec ou sans cassure (Figure 13.4.B). On constate qu'à GC constant, les intergènes cassés sont significativement plus longs que les autres, alors qu'à longueur constante, il n'y a pas de différence de taux de GC entre intergènes cassés ou non. C'est donc bien la longueur des intergènes qui influence la probabilité de cassure, et non le taux de GC, que l'on peut éliminer des candidats.

A. B.

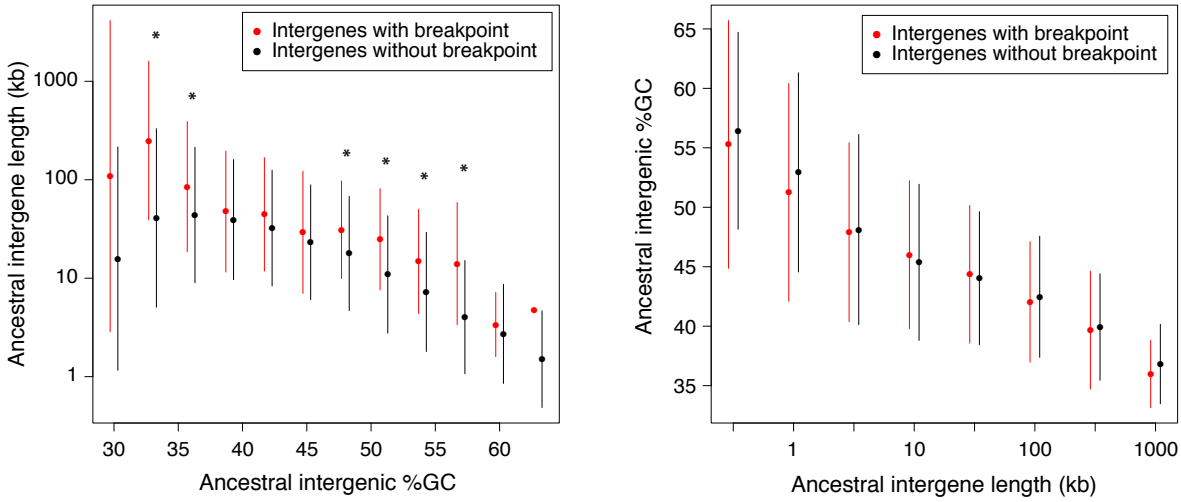


Figure 13.4. A. Longueur moyenne des intergènes ancestraux contenant ou non un point de cassure dans l'une des lignées descendantes, triés par classes de GC homogène. Les intergènes contenant un point de cassure sont plus longs en moyenne que les autres à GC équivalent. B. Taux de GC moyen des intergènes ancestraux contenant ou non un point de cassure dans l'une des lignées descendantes, triés par classes de longueur homogène. On ne constate pas de différence entre les deux catégories. Barres : écart-type. Astérisques : différences significatives après correction de Bonferroni pour les tests multiples.

13.3. Influence mineure du contenu en éléments non-codants

13.3.1. Eléments conservés non-codants

Nous avons ensuite testé si la proportion de séquences conservées non-codantes dans les intergènes ancestraux a influencé la probabilité de cassure. On s'attend en effet à ce que les régions qui contiennent une large proportion de séquence non-codante conservée soient riches en interactions fonctionnelles entre gènes et éléments de régulation, et donc potentiellement sous contrainte de sélection. Lorsque l'on segmente les intergènes par classes de proportion de séquence conservée, on constate que le taux de cassure a tendance à diminuer lorsque la proportion de séquence conservée augmente, mais la corrélation semble pauvre (Figure 13.5 ; la représentation n'est pas logarithmique afin de visualiser les taux de cassure de 0). Cette observation est confirmée par la régression : la proportion de séquence conservée est significativement corrélée au taux de cassure (z-test : P = 0,01), mais la corrélation est faible (Pseudo R^2 de McFadden = 0,11) et la déviance résiduelle est très importante (test de Chi^2 : $P \sim 0$). L'équation obtenue est :

$$\log(r) = -62{,}12 * \log(PropCNE) + 29{,}62 \quad \leftrightarrow \quad r = 7{,}31\,.\,10^{12} * PropCNE^{\,-62{,}12}$$

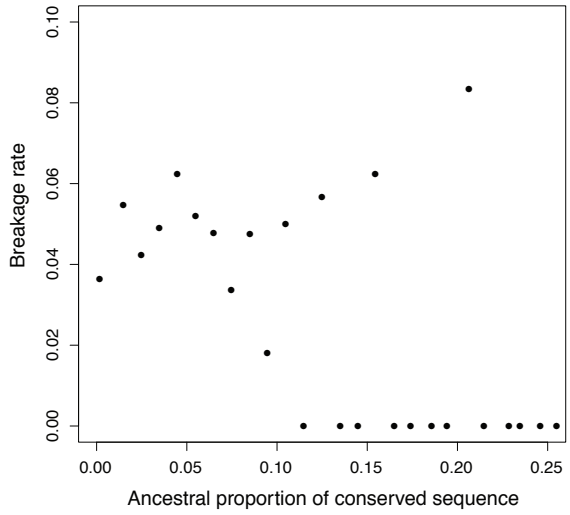


Figure 13.5. Corrélation entre la proportion de séquence conservée estimée des intergènes ancestraux (abscisse) et leur taux de cassure moyen dans les lignées descendantes (ordonnée).

Nous avons donc introduit la proportion de séquence conservée non-codante dans la procédure de régression multivariée. Les intergènes ont été segmentés par classes de longueur comme précédemment (de 0,5 en 0,5 en échelle log), puis sur la base de leur proportion de séquence conservée. Une segmentation en plusieurs classes de bornes fixes n'a pas de sens pour cette variable, car la gamme couverte par la proportion de séquence conservée varie en fonction de la longueur des intergènes : pour les plus courts, la gamme va de 0 à 100% alors que pour les plus grands, les valeurs ne dépassent pas quelques pourcents de séquence conservée. Pour obtenir des groupes d'intergènes aux comptages aussi bien répartis que possible et optimiser le pouvoir statistique de la régression, nous avons trié chaque classe d'intergènes en fonction de leur proportion de séquence conservée, qui ont ensuite été séparés en deux sous-classes de part et d'autre de la valeur médiane. La proportion de séquence conservée non-codante utilisée dans la régression pour chaque groupe d'intergènes est la moyenne des valeurs du groupe. Les résultats de la régression multivariée progressive sont résumés dans le tableau 13.2.

	Coefficients					Goodness of fit		
	Simple regression	Stepwise regression	P(>|z|)	Null deviance (df)	Residual Deviance (df)	χ^2 P-value	Stepwise χ^2 P-value	Pseudo R^2
Model 3 : length + %CNE								
Intergene length	0.28	0.30	**< 2.10^{-16}**	179.2 (19)	26.3 (18)	**0.09**	-	**0.85**
%CNE	-	- 4.55	**0.01**	179.2 (19)	20.7 (17)	**0.24**	**0.02**	**0.88**

Table 13.2. Résultats de la procédure de régression progressive du taux de cassure sur la longueur des intergènes et la proportion de séquence conservée.

Les intergènes ancestraux cassés sont en moyenne moins riches en séquences conservées que les intergènes sans cassure, à longueur équivalente, bien qu'aucune des différences ne soit statistiquement significative prise individuellement (Figure 13.6). Sur l'ensemble des données, la proportion de séquence conservée dans l'intergène est un facteur influençant significativement le taux de cassure, même lorsque la longueur de l'intergène est prise en compte (z-test : P = 0,02). L'ajout de cette variable améliore significativement le modèle (test de

Chi² sur la différence des déviances résiduelles : P = 0,02), mais cette amélioration est ténue puisque l'incrément du pseudo R^2 de McFadden n'est que de 3%, passant de 0,85 à 0,88. L'effet de la proportion de séquence conservée non-codante est donc réel mais presque négligeable devant l'effet très fort de la longueur de intergènes.

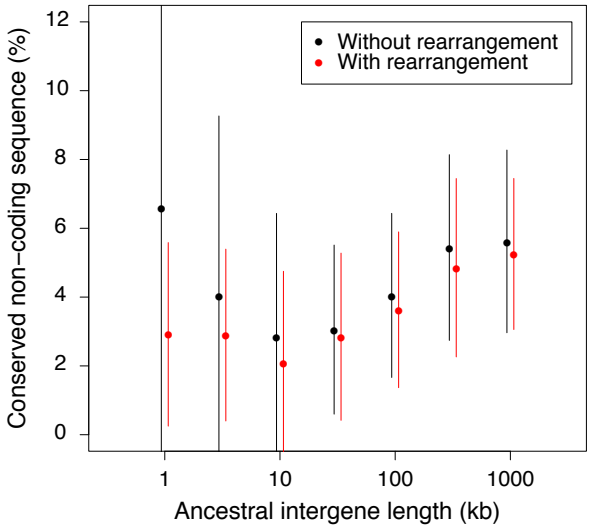


Figure 13.6. Proportion de séquence conservée dans les intergènes ancestraux affectés ou non par un point de cassure dans l'une des lignées descendantes.

Ces résultats suggèrent que des contraintes sélectives existent effectivement sur certaines régions du génome et tendent à en préserver l'organisation, mais qu'elles ne sont pas un facteur d'influence majeure sur la réorganisation des gènes à l'échelle du génome entier. Les points de cassure apparus depuis l'ancêtre Boreoeutheria suivent la distribution d'une variable aléatoire de Poisson : ils se sont produits et ont atteint la fixation avec une probabilité presque entièrement déterminée par l'espacement entre les gènes. Le modèle mathématique obtenu, basé sur la longueur des intergènes et la proportion de séquences codantes, est un modèle complet qui suffit à expliquer l'ensemble de la distribution des points de cassure depuis l'ancêtre (test de Chi² sur la déviance résiduelle : P = 0,24). Ce modèle est étonnamment simple et suggère que les cassures sont un phénomène aléatoire et essentiellement neutre, lié aux caractéristiques physiques du génome.

13.3.2. Blocs de régulation génomique

En dernière étape de la modélisation, nous avons cherché à remplacer la proportion de séquence conservée non-codante par une variable moins indirecte, en utilisant les prédictions de gènes cibles de blocs de régulation génomique (GRB ; blocs de gènes maintenus en synténie par des interactions à longue distance entre un gène et ses séquences de régulation en *cis* ; (Kikuta et al. 2007)). Nous faisons l'hypothèse que ce réseau de régulation exerce une contrainte de sélection négative contre les réarrangements sur les intergènes flanquant les gènes cibles putatifs. Comme décrit aux paragraphes 7.5 et 11.4.2, ces prédictions sont basées sur un

algorithme utilisant entre autres prédicteurs la conservation locale de la synténie entre le génome de l'homme et celui du poisson zèbre. Pour éviter une circularité de raisonnement, nous avons donc éliminé de l'analyse tous les points de cassure identifiés entre l'ancêtre Boreoeutheria et l'homme, puisque nous savons que ces intergènes borderont rarement un gène cible du fait de l'algorithme utilisé. L'analyse se base donc sur 598 points de cassure dans des intergènes pour lesquels une longueur ancestrale a pu être reconstruite.

Les intergènes sont segmentés sur la base de leur longueur comme précédemment (classes de 0,5 en 0,5 en échelle log), puis en deux sous-groupes : ceux bordant un gène cible putatif de GRB, et les autres. Border ou non un gène cible est représenté par une variable binaire qui prend la valeur 1 (oui) ou 0 (non). On constate que les intergènes ne bordant pas un gène cible de GRB sont en moyenne plus souvent cassés que les autres à longueur équivalente (Figure 13.7). La régression confirme cette tendance : le taux de cassure est significativement plus faible dans les intergènes bordant un gène cible putatif (z-test : P = 0,004 ; Table 13.3). Ajouter cette variable améliore significativement le modèle (test de Chi^2 sur la différence de déviance résiduelle : P = 0,002), mais seulement à la marge, comme précédemment, en augmentant le pseudo R^2 de McFadden de 5%.

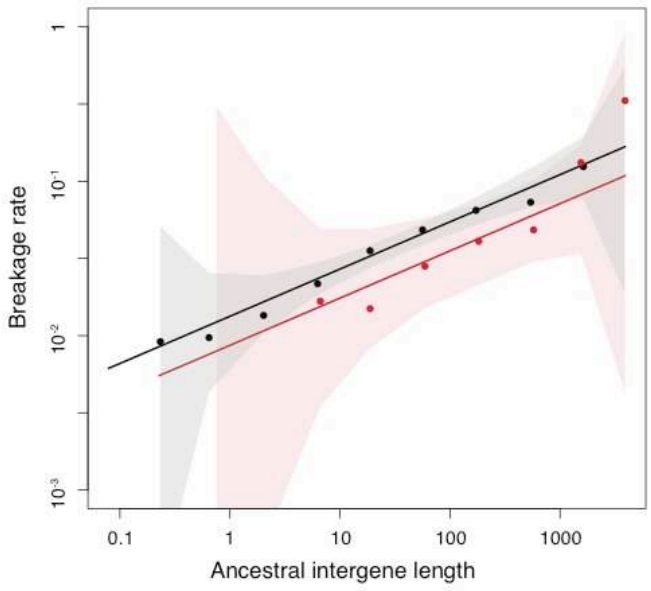


Figure 13.7. Corrélation entre la longueur des intergènes ancestraux et leur taux de cassure dans les lignées descendantes pour les intergènes bordant un gène cible de GRB (rouge) ou non (noir). Les droites représentent l'équation de régression pour chaque catégorie, et les zones colorées les intervalles de confiance du modèle.

	Coefficients					Goodness of fit		
	Simple regression	Stepwise regression	P(>|z|)	Null deviance (df)	Residual Deviance (df)	χ^2 P-value	Stepwise χ^2 P-value	Pseudo R^2
Model 4 : length + GRB								
Intergene length	0.28	0.29	**< 2.10^{-16}**	208.8 (31)	58.5 (30)	**5.10^{-3}**	-	**0.74**
GRB	-	-0.45	**1.10^{-3}**	208.8 (31)	41.9 (29)	**0.06**	**5.10^{-5}**	**0.80**

Table 13.3. Résultats de la procédure de régression progressive du taux de cassure sur la longueur des intergènes et les cibles de GRB.

Ce résultat confirme les observations obtenues avec la mesure plus brute de la proportion de séquence conservée dans les intergènes : il existe des régions du génome où les réarrangements ont des effets délétères sur les réseaux de régulation génique et où s'exerce donc une pression de sélection négative, mais cet effet est mineur. La probabilité d'observer un point de cassure fixé est gouvernée presque entièrement par les longueurs intergéniques, donc a priori par les propriétés physiques du génome. Ceci suggère que les réarrangements sont un phénomène majoritairement neutre lié à des contraintes mécaniques plutôt que sélectives.

Par ailleurs, ce résultat confirme a posteriori la validité de l'approche de prédiction utilisée pour établir la liste des gènes cibles de GRB putatifs dans le génome humain. En effet, les prédictions ne font pas intervenir les quatre génomes de boreoeuthériens dans lesquels ont été identifiés les points de cassure utilisés dans cette analyse (souris, chien, vache et cheval). Le fait d'observer que les gènes identifiés comme cibles de GRB dans le génome humain sont indépendamment moins souvent séparés de leur environnement que les autres dans ces quatre génomes confirme que les régions identifiées sont effectivement enrichies en régions sous contrainte, et non un échantillonnage aléatoire.

13.4. Effet des événements gène unique

Le modèle de régression du taux de cassure montre que la probabilité de cassure est fortement corrélée à la longueur des intergènes, mais pas de manière directe : le taux de cassure est proportionnel à une racine de la longueur des intergènes, sans explication évidente. Nous avons vérifié, en première approche, que ce résultat n'est pas un artefact lié aux événements gène unique que nous avons exclus du jeu de données. Comme abordé au chapitre 12, nous avons exclu du jeu les paires de points de cassure qui se trouvent de part et d'autre d'un même gène dans la même lignée : le jeu contenait un nombre suspect de tels événements et l'inspection manuelle a révélé que dans de nombreux cas, ces « réarrangements » sont en réalité des scaffolds mal placés dans l'assemblage ou des gènes mal annotés. Ce filtre a donc pour but de retirer ces faux positifs, mais il est probable qu'il exclut également des points de cassure *bona fide*, surtout si les réarrangements impliquent souvent des régions relativements courtes pouvant ne comporter qu'un seul gène. Nous avons donc testé si la corrélation forte entre le taux de cassure et la longueur des intergènes obtenue précédemment est reproductible lorsque ces points de cassure sont inclus dans le jeu.

Nous avons reproduit à l'identique la procédure décrite au paragraphe 13.1 en utilisant cette fois le jeu de 833 points de cassure incluant à la fois les 682 points de cassure utilisés précédemment (jeu fiable) et les 151 points de cassure correspondant à des événements gène unique douteux. Le modèle obtenu est remarquablement similaire à celui obtenu sur le jeu fiable : le taux de cassure reste fortement corrélé à la longueur des intergènes après une transformation logarithmique (Figure 13.8).

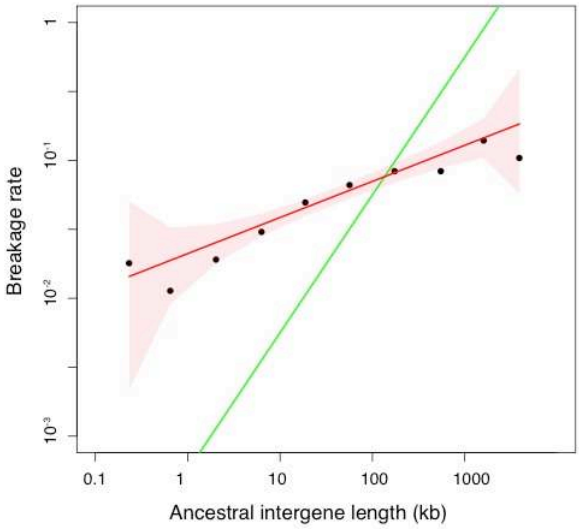


Figure 13.8. Corrélation entre la longueur des intergènes ancestraux (abscisse) et leur taux de cassure moyen dans les lignées descendantes (ordonnée), lorsque les événements gène unique sont inclus dans le jeu de données.

Le coefficient de la régression est également remarquablement proche dans les deux cas (0,26 comparé à 0,28 précédemment). Au niveau des statistiques de validation du modèle, la déviance résiduelle est supérieure à celle observée sur le jeu fiable (15,9 contre 12,4 précédemment), ce qui est cohérent avec un signal un peu bruité par l'inclusion de faux positifs ; cependant, le test de Chi^2 reste non significatif (P = 0,07, comparé à P = 0,19 précédemment). La corrélation entre le taux de cassure et une racine de la longueur des intergènes n'est donc pas un artefact lié à l'exclusion des événements gène unique. Au contraire, le signal observé se maintient malgré l'inclusion probable de faux positifs dans le jeu de données, et est donc particulièrement robuste.

13.5. Elimination de potentiels facteurs confondants

Le modèle mathématique obtenu ci-dessus est étonnamment simple, puisqu'il montre que la longueur des intergènes seule suffit à expliquer la variation du taux de cassure dans les intergènes du génome ancestral, variation qui est modulée par des contraintes sélectives sur certaines régions. Ce résultat peut sembler contradictoire avec les diverses observations publiées dans la littérature, où les points de cassure ont été fréquemment et significativement associés à des propriétés génomiques comme la densité en éléments transposables, en îlots CpG, ou encore en duplications segmentales. De plus, notre modèle montre que le taux de cassure n'est pas directement proportionnel à la longueur des intergènes mais à une racine de cette longueur ($L^{0,28}$). D'un point de vue biologique, ce résultat n'a pas d'explication intuitive évidente : on peut donc légitimement se demander si le réel déterminant de la probabilité de cassure ne serait pas une autre propriété génomique, elle-même corrélée à la longueur des intergènes qui tiendrait alors le rôle de proxy dans l'analyse de régression.

Parmi les propriétés du génome, seules quelques unes ont pu être reconstruites dans l'état ancestral en raison de leur bonne conservation entre les différentes espèces descendantes. Nous avons recherché dans le génome humain des facteurs confondants potentiels parmi les propriétés susceptibles de promouvoir les cassures, mais qui n'ont pas pu être reconstruites dans le génome ancestral et testés dans le modèle. L'excellente relation de proportionnalité entre le taux de cassure et $L^{0,28}$ nous renseigne sur le comportement des candidats : un facteur confondant potentiel est une propriété qui se comporte comme $L^{0,28}$, c'est-à-dire qui augmente en nombre absolu avec la longueur des intergènes L mais dont la densité dans les intergènes diminue quand L augmente (tout comme $L^{0,28}/L$ diminue). Nous avons utilisé cette condition simple pour tester si les propriétés génomiques ayant été précédemment associées aux points de cassure peuvent être des facteurs confondants dans notre modèle de régression. Il s'agit ici d'éliminer rapidement des facteurs statistiquement enrichis ou appauvris dans les régions de cassure, mais qui ne présentent pas une distribution susceptible d'expliquer nos données. Dans le chapitre suivant, nous testerons par simulations la proposition inverse, c'est-à-dire que ces variables montrent des corrélations avec les points de cassure parce qu'elles varient entre régions riches et pauvres en gènes, bien qu'elles ne soient pas causatives des réarrangements.

13.5.1. Eléments transposables

Les séquences répétées, et notamment les éléments transposables, ont souvent été invoqués comme promoteurs de réarrangements en favorisant la recombinaison illégitime (Gray 2000; Shaw and Lupski 2004; Liu et al. 2012). Les points de cassure ont notamment été fréquemment associés à des densités en SINEs plus élevées que la moyenne du génome (Ma et al. 2006; Schibler et al. 2006; Carbone et al. 2009). Lorsque l'on s'intéresse à l'évolution de la densité moyenne en éléments transposables toutes catégories confondues dans les intergènes de longueur croissante, on constate que cette densité augmente rapidement dans les intergènes inférieurs à 10 kb pour ensuite se stabiliser autour de 50% pour les intergènes plus grands (qui représentent 56% des intergènes ; Figure 13.9). Ce comportement ne correspond pas à la diminution continue en densité attendue si les éléments transposables devaient expliquer les variations du taux de cassure observées à la place de la longueur des intergènes.

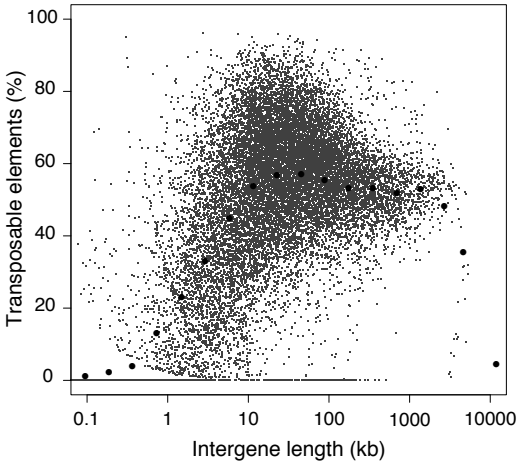


Figure 13.9. Variations de la densité en éléments transposables des intergènes du génome humain en fonction de leur longueur. Le nuage correspond aux mesures individuelles des intergènes, les cercles pleins à la moyenne par classe de taille homogène.

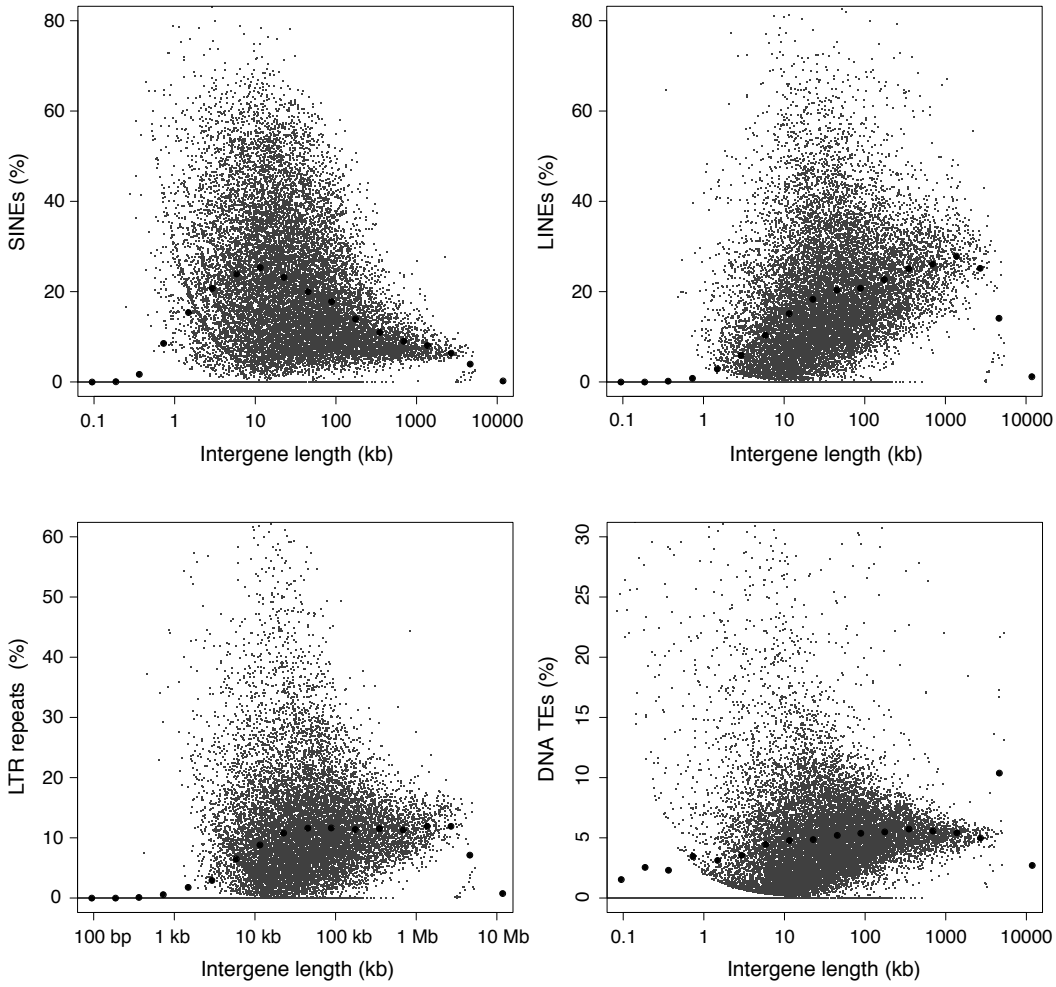


Figure 13.10. Variations de la densité des quatre principales classes d'éléments transposables dans les intergènes du génome humain en fonction de la longueur de l'intergène, prises séparément. Symboles comme précédemment.

Quand on considère les différentes catégories d'éléments transposables séparément, aucune des quatre grandes catégories de transposons (SINEs, LINEs, transposons à LTR et à ADN) ne présente le profil attendu (Figure 13.10). Aucun type de transposons ne voit sa densité moyenne diminuer lorsque la longueur des intergènes augmente sur l'ensemble de la gamme de longueurs des intergènes humains. Seuls les SINEs décroissent en densité sur une partie seulement de la gamme de longueurs, les intergènes plus longs que 10 kb ; leur densité augmente par contre dans les intergènes plus courts. Ce comportement est typique des SINEs de boreoeuthériens, qui présentent le même profil chez la souris et le chien (Figure 13.11). Le taux de cassure, lui, ne présente pas de point d'inflexion et son augmentation est régulière sur l'ensemble de la gamme, ce qui ne peut pas être expliqué par les variations de la densité en SINEs. Ainsi, les éléments transposables ne sont pas des facteurs confondants plausibles dans notre analyse de régression.

A. B.

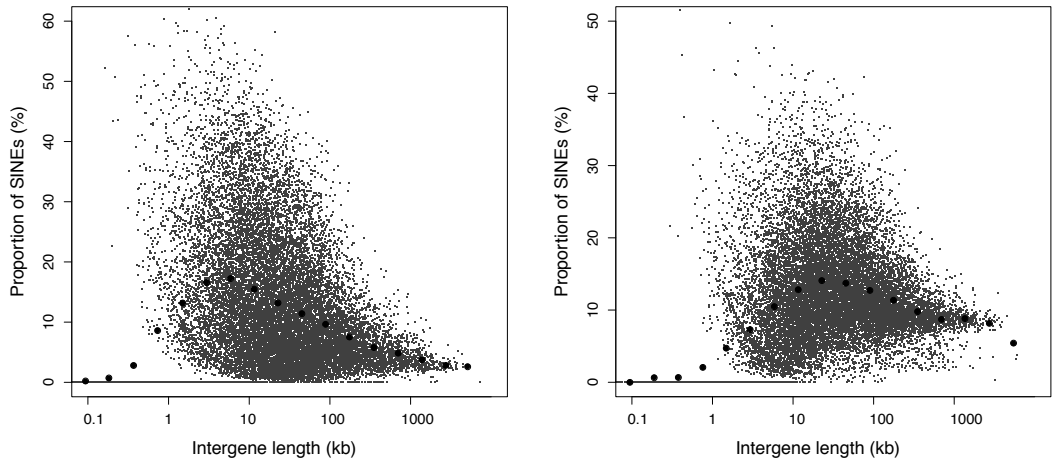


Figure 13.11. Variations de la densité en SINEs des intergènes du génome de la souris (A) et du chien (B) en fonction de leur longueur. Symboles comme précédemment.

13.5.2. Duplications segmentales

Les duplications segmentales (définies comme des régions du génome de plus de 1 kb identiques à plus de 90%) ont également été associées positivement aux points de cassure (Bailey et al. 2004; Ma et al. 2006; Zhao and Bourque 2009). Tout comme pour les éléments transposables, il a été proposé qu'elles pourraient servir de substrat à des recombinaisons illégitimes, favorisant les réarrangements. La proportion moyenne de séquence contenue dans une duplication segmentale varie avec la longueur des intergènes, mais de manière différente selon la gamme de longueurs considérée (Figure 13.12) : elle augmente jusqu'à 100 kb, et diminue au-delà. Le taux de cassure, lui, ne change pas de comportement sur l'ensemble de la gamme de longueurs d'intergènes. Les variations de densité de duplications segmentales ne peuvent donc pas expliquer celles du taux de cassure.

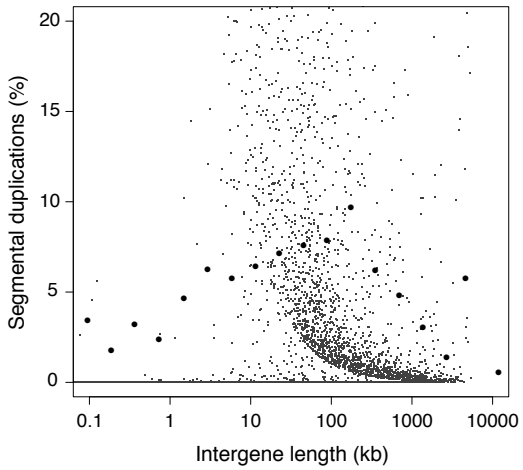


Figure 13.12. Variations de la densité en séquence incluse dans une duplication segmentale dans les intergènes du génome humain, en fonction de la longueur des intergènes. Symboles comme précédemment.

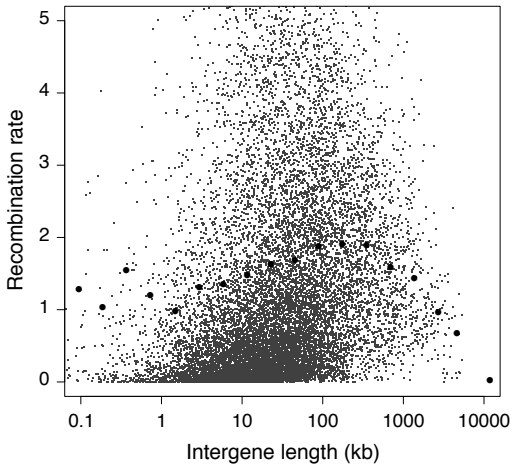


Figure 13.13. Variations du taux de recombinaison dans les intergènes du génome humain, en fonction de la longueur des intergènes. Symboles comme précédemment.

13.5.3. Taux de recombinaison

Le taux de recombinaison a été proposé comme explication aux variations du taux de réarrangements, sous l'hypothèse que les réarrangements seraient causés par des erreurs lors de la recombinaison méiotique, et seraient donc plus fréquents dans les régions à fort taux de recombinaison (Volker et al. 2010). Nous avons donc testé cette possibilité, bien que le taux de recombinaison ne soit que faiblement corrélé à la densité locale en gènes et plutôt déterminé par la composition en base et notamment le taux de GC (Kong et al. 2002) ; or, nous avons éliminé le taux de GC des déterminants potentiels de la cassure au paragraphe 13.2. Comme attendu, la corrélation entre le taux de recombinaison et la longueur des intergènes est pauvre ; par ailleurs le taux de recombinaison moyen augmente dans une partie de la gamme de longueurs, et diminue dans l'autre, ce qui à nouveau l'élimine comme facteur confondant potentiel dans la régression (Figure 13.13).

13.5.4. Origines de réplication

Enfin, nous avons testé si les origines de réplication, qui se trouvent fréquemment dans les régions denses en gènes et sont associées avec des marques de chromatine ouverte (Mechali 2010), seraient susceptibles de promouvoir les cassures et donc être associées aux points de cassure, comme suggéré dans une étude précédente chez la levure (Gordon et al. 2009). Nous avons utilisé 874 origines de réplications conservées prédites dans le génome humain sur la base des biais compositionnels de la séquence (Huvet et al. 2007). Lorsque l'on s'intéresse à la distribution de ces origines de réplication dans les intergènes, on observe une remarquable corrélation entre le taux d'origines de réplication par intergène et leur longueur (Figure 13.14). Les origines de réplication sont en effet plus nombreuses qu'attendu au hasard dans les intergènes courts et moins nombreuses dans les intergènes longs, ce qui est cohérent avec la littérature ; de plus, cette corrélation est linéaire en représentation logarithmique et rappelle celle des points de cassure. Les origines de réplication se comportent également comme une variable aléatoire de Poisson, et leur taux augmente avec une racine de la longueur des intergènes.

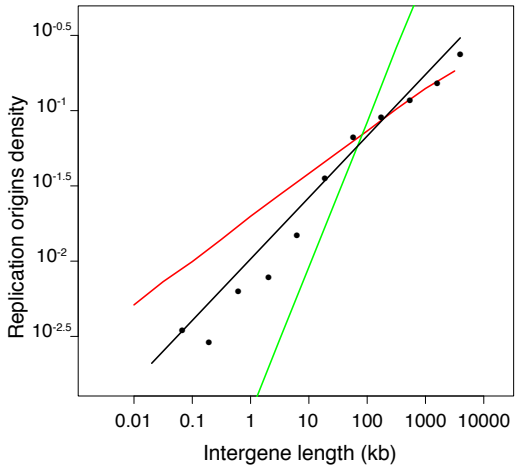


Figure 13.14. Corrélation entre la longueur des intergènes et leur densité en origines de réplication dans le génome humain. L'équation obtenue par régression de Poisson est figurée en noir ; la droite attendue sous une distribution aléatoire est figurée en vert, et la droite observée pour les points de cassure de réarrangements est figurée en rouge à titre de comparaison.

Les origines et le timing de réplication sont des propriétés conservées dans les génomes de mammifères (Ryba et al. 2010), si bien que l'on peut raisonnablement faire l'hypothèse que les origines de réplication du génome humain existaient également dans les intergènes de l'ancêtre Boreoeutheria. Le faible nombre d'origines prédites ne permet cependant pas de les intégrer dans l'analyse de régression. On peut tout de même mesurer la contingence entre les origines de réplication et les points de cassure : les intergènes ancestraux contenant à la fois un point de cassure et une origine de réplication sont plus nombreux qu'attendus au hasard (test exact de Fisher : P = 0,01). Ce résultat est attendu a priori, puisque les deux propriétés sont enrichies dans les intergènes courts. En revanche, lorsque l'on sépare le génome en deux groupes

d'intergènes de tailles plus homogènes de part et d'autre de la médiane dans le génome, le nombre d'intergènes présentant les deux propriétés dans chaque catégorie n'est pas différent de celui attendu au hasard (intergènes < 20 kb : $P = 0,29$; intergènes > 20kb : $P = 0,13$). Points de cassure et origines de réplication sont donc indépendants l'un de l'autre. Le fait que leurs distributions présentent des caractéristiques identiques, avec un taux moyen corrélé à une racine de la longueur, laisse supposer que les deux propriétés sont des conséquences d'une même cause.

Chapitre 14. Validation du modèle de distribution des points de cassure par simulations

La reconstruction du génome ancestral nous a permis d'établir un modèle mathématique décrivant la distribution des points de cassure de réarrangements évolutifs. Ce modèle simple montre que la probabilité de cassure d'un intergène dépend presque exclusivement de sa longueur, mais d'une façon non strictement linéaire. Aucune des propriétés génomiques connues pour être corrélées aux points de cassure ne présente les caractéristiques d'un facteur confondant qui suffirait à expliquer cette relation entre la longueur des intergènes et le taux de cassure. En revanche, dans ce chapitre, nous démontrons que des cassures simulées suivant notre modèle suffisent à récapituler les corrélations observées dans la littérature.

14.1. Simulations de cassures suivant le modèle de régression

14.1.1. Composition des régions de cassure

La majorité des études précédentes sur les points de cassure de réarrangements utilise une approche différente de la nôtre : elles ne se basent pas sur les intergènes comme unités pouvant être cassées ou non, mais comparent la composition des « régions de cassure » au reste du génome. Les régions de cassure sont alors des fenêtres de taille arbitraire autour des points de cassure, qui sont comparées à des fenêtres de même taille échantillonnées aléatoirement pour évaluer si les points de cassure se produisent dans des régions statistiquement biaisées. Afin de tester si notre modèle suffit à expliquer les observations publiées dans la littérature, nous avons mené une expérience d'évolution simulée en distribuant des « points de cassure » dans les intergènes du génome humain en suivant l'équation de régression obtenue dans le modèle, basée sur la longueur des intergènes et leur proportion en éléments conservés non-codants. Nous avons ensuite mesuré un certain nombre de propriétés dans des fenêtres de 100 kb centrées sur les points de cassure. Ces fenêtres ont ensuite été comparées à deux jeux de fenêtres de 100 kb contrôles. Le premier contrôle est un jeu de fenêtres tirées entièrement aléatoirement, centrées sur n'importe quelle base du génome humain avec la même probabilité. Pour le second contrôle, le centre de la fenêtre doit être une base intergénique (toutes les bases intergéniques ayant la même probabilité d'être tirées) : ce jeu permet de contrôler que les différences vues entre le premier jeu contrôle et les simulations ne sont pas dues au fait que nos points de cassure simulés ne peuvent être qu'intergéniques. Chaque groupe de simulations (points de cassure et deux contrôles) a été répété 100 fois, donnant 100 mesures de comparaison (Figure 14.1).

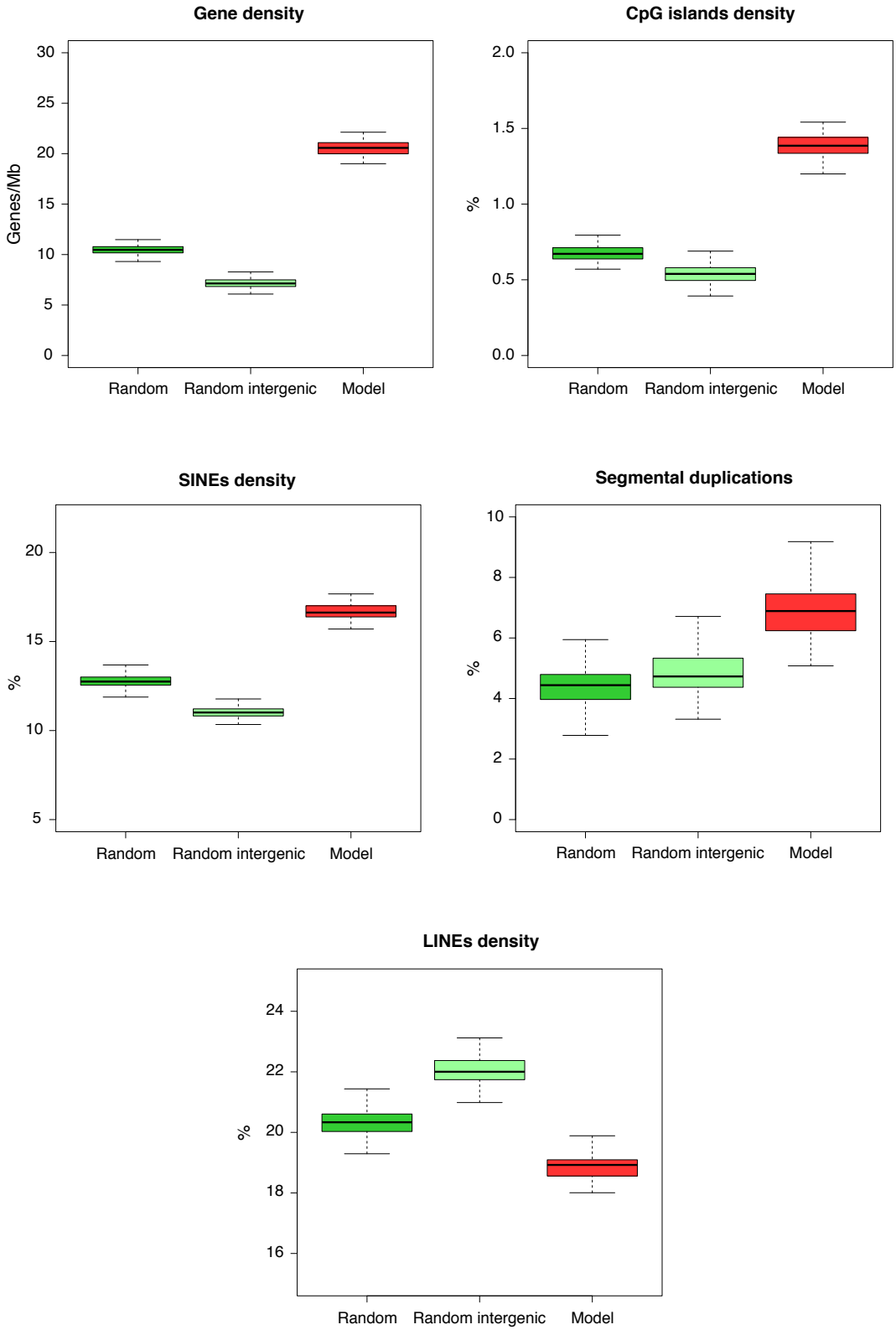


Figure 14.1. Composition des régions de cassure simulées avec le modèle de régression (longueur + proportion de séquence conservée ; en rouge), comparées à des régions échantillonnées aléatoirement dans tout le génome (vert foncé) ou aléatoirement mais centrées dans un intergène (vert clair). Les boxplots représentent la distribution des valeurs moyennes obtenues sur 100 simulations.

On observe alors que les points de cassure simulés sont associés à une densité en gènes en moyenne deux fois plus importante que les contrôles (test de Wilcoxon : $P < 1.10^{-13}$ pour toutes les comparaisons), ce qui est cohérent avec leur probabilité d'occurrence plus élevée qu'attendue dans les petits intergènes, et plus faible qu'attendue dans les grands (pour rappel, Figure 13.1). Les points de cassure simulés sont également associés avec un doublement de la densité en îlots CpG ($P < 2.10^{-8}$ pour toutes les comparaisons). Toutes les simulations montrent que les points de cassure sont significativement associés à une forte densité en SINEs, avec une moyenne de 16,7% contre 12,8% et 11,0% dans chacun des contrôles (aléatoire pur et aléatoire intergénique, respectivement). Dans 96% des simulations, on trouve que les points de cassure sont significativement associés aux duplications segmentales, qui présentent une densité moyenne de 6,9% contre 4,4% et 4,9% dans chacun des contrôles. En ce qui concerne les LINEs, ceux-ci sont appauvris autour des points de cassure simulés par rapport aux contrôles ($P < 5.10^{-5}$ pour toutes les comparaisons). Ce point fait débat dans les littérature, puisque selon les espèces et les méthodes utilisées, certains auteurs trouvent que les LINEs sont appauvris autour des points de cassure (Carbone et al. 2009), d'autres observent une surreprésentation des LINEs (Zhao and Bourque 2009), et d'autres enfin trouvent des résultats contradictoires entre eux lorsque les points de cassure sont subdivisés en différentes catégories (Ma et al. 2006).

De manière générale, les points de cassure simulés selon le modèle de régression récapitulent bien les différentes caractéristiques des régions de cassure qui apparaissent de manière récurrente dans la littérature. Les différents enrichissements et appauvrissements sont par ailleurs du même ordre de grandeur que ceux observés dans la littérature (Ma et al. 2006; Gordon et al. 2007; Larkin et al. 2009), à l'exception des duplications segmentales, qui sont moins enrichies autour des points de cassure simulés que reporté dans certaines études (Ma et al. 2006; Zhao and Bourque 2009). Ce dernier résultat s'explique à la lumière de plusieurs travaux de la littérature, qui suggèrent que les duplications segmentales sont générées au cours des événements de réarrangements plutôt qu'elles ne les causent (Bailey et al. 2004; Bailey and Eichler 2006; Ranz et al. 2007; Girirajan et al. 2009). Comme nos simulations ne modélisent pas l'apparition de duplications au niveau des points de cassure, les points de cassure simulés apparaissent moins enrichis en duplications segmentales que ceux observés dans les génomes modernes.

14.1.2. Longueur des blocs de synténie simulés

Historiquement, l'une des observations majeures ayant permis de réfuter la distribution aléatoire des points de cassure se base sur la longueur des blocs de synténie conservés entre les différentes espèces, c'est-à-dire les distances qui séparent les points de cassure consécutifs (voir Introduction, chapitre 3). La distribution de ces longueurs de blocs n'est pas celle attendue si les points de cassure étaient distribués au hasard : les approches théoriques (Pevzner and Tesler 2003b) et expérimentales (Kent et al. 2003; Bourque et al. 2004; Zhao et al. 2004) mettent en évidence un excès de blocs courts, signe que les points de cassure se produisent de façon plus regroupée qu'attendu. Nous avons testé si les cassures simulées selon notre modèle dans le génome humain vérifient également cette propriété. On constate qu'en effet, les points de cassure simulés définissent un excès de blocs courts par rapport aux deux contrôles aléatoires,

reproduisant l'effet observé sur les données réelles (Figure 14.2). Ainsi, les simulations d'évolution montrent que le modèle de régression suffit à reproduire les principales caractéristiques des points de cassure dans les génomes de vertébrés publiées dans la littérature.

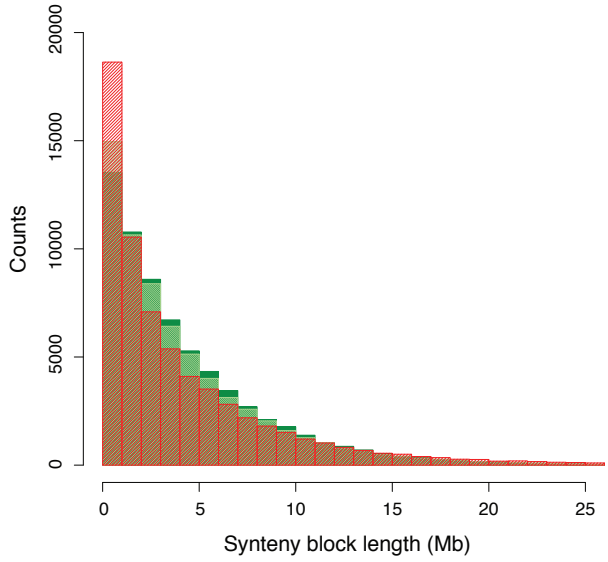


Figure 14.2. Distribution des longueurs de blocs de synténie simulés avec le modèle de régression (longueur + proportion de séquence conservée ; en rouge), par comparaison à l'attendu si les régions sont échantillonnées aléatoirement dans tout le génome (vert foncé) ou aléatoirement mais centrées dans un intergène (vert clair). La distribution correspond à la moyenne des fréquences observées sur 100 simulations.

14.2. Simulations de points de cassure dépendants (inversions)

Bien que notre modèle rende compte de l'association entre les points de cassures et certaines propriétés du génome, la raison pour laquelle le taux de cassure par intergène n'augmente pas proportionnellement à la longueur des intergènes reste inconnue. Dans les génomes de vertébrés, la majorité des événements de réarrangement sont des inversions, où un chromosome est cassé en deux points et le segment intermédiaire réinséré en sens inverse (Bourque et al. 2004; Zhao et al. 2004). Dans ce cas, il est possible que la localisation des deux points de cassure ne soit pas indépendante : cette dépendance, si elle existe, n'est pas prise en compte dans notre modèle, qui considère chaque point de cassure comme un événement indépendant. Comme notre méthode ne peut identifier un point de cassure que lorsque l'ordre des gène est modifié, il est plausible que l'excès de points de cassure observés dans les petits intergènes soit dû non pas à une plus forte probabilité de cassure, mais à une meilleure sensibilité : ainsi, une inversion courte pourra être détectée dans les régions denses en gènes, alors qu'une inversion de même taille dans une région pauvre en gènes se produira fréquemment au sein d'un même intergène et ne sera pas identifiée. Ce biais pourrait être particulièrement sévère si les petites inversions sont beaucoup plus fréquentes que les grandes.

Tester cette hypothèse est un problème difficile, car la distribution précise de la longueur des inversions dans les génomes de mammifères n'est pas connue. S'il semble clair que les inversions courtes sont plus fréquentes que les longues (Feuk et al. 2005), établir la distribution des longueurs d'inversions entre deux génomes reviendrait à identifier tous les points de cassure au niveau de la séquence nucléotidique (avec tous les problèmes de faux positifs que cela suppose), puis à reconstruire précisément l'historique de ces réarrangements, afin de résoudre tous les cas d'inversions chevauchantes. Il s'agit d'un travail complexe qui est au-delà des objectifs de cette thèse ; n'ayant pas accès à cette distribution, nous avons simulé des inversions dans le génome humain en faisant deux hypothèses. La première est qu'un contact physique est nécessaire pour qu'un réarrangement puisse se produire, et que par conséquence, la probabilité qu'un réarrangement se produise entre deux régions du génome est liée à la probabilité de contact entre ces deux régions. Nous avons utilisé les données de la carte Hi-C du génome humain (Lieberman-Aiden et al. 2009), qui comptabilise les interactions entre les différentes régions du génome dans le noyau en interphase. Cette carte montre qu'en moyenne, la probabilité de contact entre deux locus du même chromosome diminue régulièrement quand la distance entre les locus augmente. Ainsi, en utilisant la probabilité de contact comme un proxy de la probabilité de réarrangement entre deux régions, nous favorisons effectivement les inversions courtes. Nous présentons les simulations de cassure réalisées à partir de ces données au paragraphe 14.2.1. Nous testerons ensuite l'hypothèse que les éléments transposables peuvent favoriser l'appariement de séquences non homologues, et favorisent les réarrangements. Bien que nous ayons exclu que les éléments transposables seuls puissent expliquer la distribution des points de cassure pris individuellement (paragraphe 13.5.1), nous testerons tout de même cette possibilité dans le cadre de points de cassures dépendants deux à deux au paragraphe 14.2.2.

14.2.1. Cassures basées sur la distance uniquement

Dans le premier cas, nous sélectionnons des points de cassure dans le génome humain en deux temps : le premier point de cassure est tiré au hasard dans un intergène, puis un second est tiré à une distance d en fonction de la probabilité de contact tirée de la carte Hi-C. Deux cas peuvent se produire : si l'intervalle entre des deux points de cassure inclut au moins un gène, l'inversion est considérée comme détectable et les points de cassure sont retenus. En revanche, si les deux points de cassure se trouvent dans le même intergène, l'inversion est considérée comme invisible. On procède à 100 simulations : chaque simulation se poursuit jusqu'au ce que l'on obtienne le même nombre de points de cassure visibles qu'entre l'ancêtre Boreoeutheria et les cinq génomes modernes analysés, soit 682. La distribution moyenne des longueurs d'inversions simulées, visibles ou non, est présentée dans la Figure 14.3.

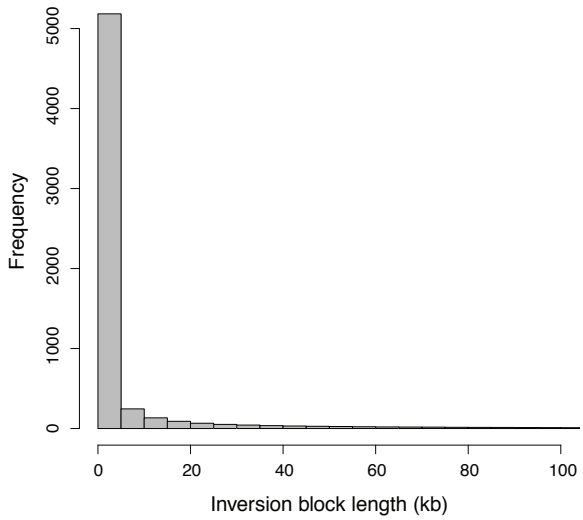


Figure 14.3. Distribution des longueurs de blocs « inversés » dans la simulation (moyenne sur 100 simulations). La distance entre les points de cassure simulés sélectionnés est fréquemment très courte (< 5kb).

On mesure ensuite le taux de cassure observé dans chaque catégorie d'intergènes en fonction de leur taille : si les simulations reflètent effectivement l'occurrence des réarrangements et notre capacité à les détecter, alors en représentation logarithmique le taux de cassure devrait être une fonction linéaire de la longueur de intergènes avec une pente proche de 0,28, comme observé dans les données réelles (Figure 13.1). Les résultats de la simulation montrent que le taux de cassure reste effectivement une fonction linéaire de la longueur des intergènes en représentation logarithmique, mais la régression donne une pente de 0,74 (Figure 14.4). Les simulations de points de cassure dépendants sont donc loin d'expliquer la déviation observée dans les données réelles, où les points de cassure dans les intergènes courts sont beaucoup plus fréquents.

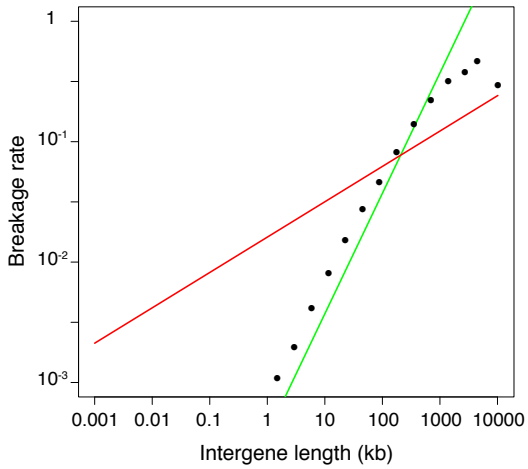


Figure 14.4. Corrélation entre la longueur des intergènes et leur taux de cassure simulé (cercles pleins). La relation attendue sous une distribution aléatoire est représentée en vert ; la corrélation observée dans les données réelles est représentée en rouge.

14.2.2. Cassures dépendantes des éléments transposables

Les éléments transposables ont fréquemment été invoqués comme facilitant les réarrangements par recombinaison non homologue ; si leurs variations de densité dans le génome ne suffisent pas à elles seules à expliquer la répartition des points de cassure, il est possible que le contact entre séquences similaires ou identique facilite leur hybridation et par conséquent favorise les réarrangements. Afin de tenir compte de l'influence possible des éléments transposables dans les simulations, nous avons réitéré les simulations en autorisant la sélection des paires de points de cassure uniquement si les deux se produisent dans des éléments transposables de la même classe au sens large (SINEs, LINEs, etc.) ou strictement du même type (AluY, MIRb, L1M4, etc.). Chaque simulation n'a été réalisée qu'une seule fois, en raison du long temps de calcul nécessaire pour obtenir une itération complète : en effet, la faible probabilité de tomber dans deux éléments transposables de même catégorie s'ajoute alors à la forte probabilité d'obtenir deux points de cassure dans un même intergène (donc « invisibles »), si bien que les paires de points de cassure tirées au hasard ne sont que très rarement validées. Dans les deux cas, le taux de cassure reste une fonction linéaire de la longueur des intergènes après une transformation logarithmique, et la pente de la corrélation est proche de celle obtenue avec la simulation simple basée uniquement sur la probabilité de contact, voire plus proche encore de l'aléatoire (0,75 avec les éléments transposables de même classe ; 0,88 avec les éléments transposables de même type strictement ; Figure 14.5). Cette contrainte d'homologie supplémentaire ne rapproche donc pas la simulation des observations réelles, et ne permet pas d'expliquer l'excès de points de cassure observé dans les petits intergènes aux dépens des plus grands.

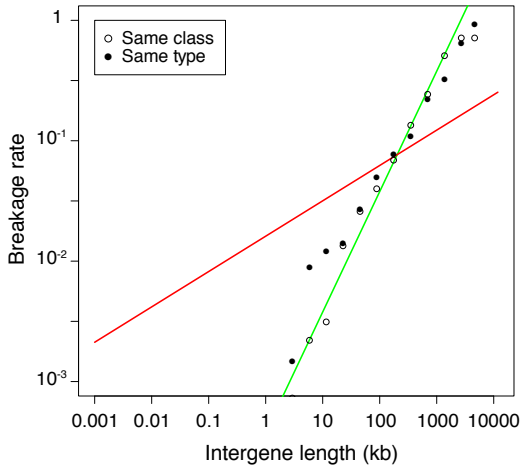


Figure 14.5. Corrélation entre la longueur des intergènes et leur taux de cassure simulé lorsque les points de cassure sont obligatoirement sélectionnés dans des éléments tranposables de même classe (cercles vides) ou de même type (cercles pleins). La relation attendue sous une distribution aléatoire est représentée en vert ; la corrélation observée dans les données réelles est représentée en rouge.

Ainsi, ces résultats montrent que l'aspect et l'équation de la courbe liant le taux de cassure et la longueur des intergènes ne peut s'expliquer uniquement par l'apparition conjointe de paires de points de cassure, dont une partie resterait non détectable par notre méthode d'analyse. Par ailleurs, la probabilité de contact entre séquences dans le noyau n'est pas non plus à elle seule une explication satisfaisante au lien entre taux de cassure et longueur des intergènes que nous observons. Il semble donc que ce soit bien la quantité d'ADN non-codant d'un gène à l'autre qui conditionne la probabilité de cassure, ce qui suggère un phénomène mécanique, lié à la structure du génome. Nous proposons que c'est la probabilité qu'une cassure double-brin apparaisse qui gouverne la probabilité qu'un réarrangement se produise ; cette probabilité pourrait être conditionnée par l'accessibilité de l'ADN en fonction de l'état de compaction de la chromatine entre éléments non-codants fonctionnels ou non (voir discussion).

Chapitre 15. Extension du modèle de distribution des points de cassure au phylum des levures

La distribution des points de cassure observés dans différentes lignées de mammifères s'explique comme une fonction de la longueur des intergènes dans le génome de leur ancêtre. Cette distribution permet d'expliquer simplement les diverses propriétés corrélées aux points de cassure de génomes de vertébrés rapportées dans la littérature. La probabilité d'apparition d'un réarrangement dans ces génomes semble dépendante de l'organisation du génome en gènes et intergènes, potentiellement à cause de différences dans l'organisation et l'accessibilité de l'ADN. Si c'est le cas, il est possible que cette propriété se vérifie plus largement dans tout le domaine des eucaryotes, où l'organisation de l'ADN et de la chromatine suit globalement les mêmes règles. Nous avons testé cette prédiction en étudiant la distribution des points de cassure de réarrangements évolutifs dans un phylum très distant des vertébrés : les levures. Ces génomes présentent une différence majeure par rapport aux génomes de mammifères : la gamme de taille des intergènes y est beaucoup plus réduite, si bien que le génome est beaucoup plus compact.

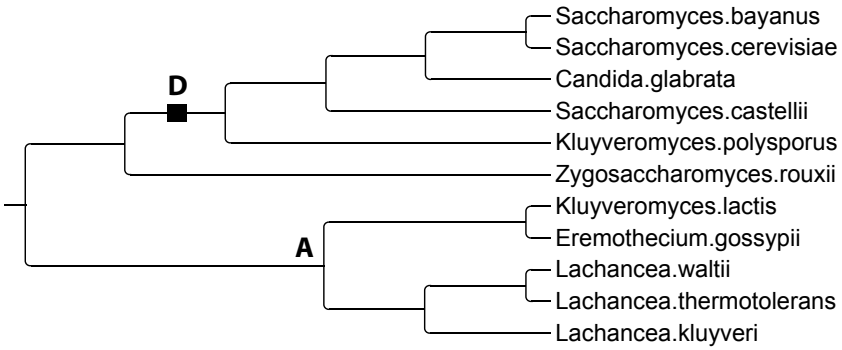


Figure 15.1. Arbre phylogénétique des génomes de levures utilisés. A : nœud du génome ancestral d'intérêt ; D : événement de duplication complète du génome. Les longueurs de branches ne sont pas proportionnelles à l'âge.

15.1. Reconstruction du génome ancestral

Nous nous sommes intéressés pour cette étude au groupe de levures qui comprend les genres *Klyuveromyces* et *Lachancea*. Ce groupe est moins étudié que le groupe des saccharomycètes, mais a l'avantage de ne pas contenir d'espèces ayant subi une duplication complète du génome récente, ce qui facilite la reconstruction du génome ancestral. Pour reconstruire ce génome ancestral, nous avons utilisé des données d'ordre des gènes dans onze espèces de levures disponibles dans la base de données Génolevures (Figure 15.1)(Sherman et al. 2009).

Les arbres phylogénétiques des gènes ont été reconstruits à l'aide du programme TreeBest (Vilella et al. 2009), puis la méthode AGORA a été appliquée afin d'obtenir l'ordre des gènes dans le génome ancestral. On obtient alors 4608 paires de gènes adjacents reconstruits dans l'ancêtre. Tout comme pour les mammifères boreoeuthériens, nous avons inféré la longueur des intergènes et leur contenu en GC à partir des propriétés de cinq génomes descendants dont l'histoire est majoritairement indépendante (*Lachancea thermotolerans*, *Lachancea kluyveri*, *Lachancea waltii*, *Kluyveromyces lactis* et *Eremothecium gossypii*). La longueur des intergènes est généralement bien corrélée à la longueur médiane de l'ensemble des orthologues (R^2 = 0,80, bien supérieur au R^2 obtenu avec des valeurs randomisées (0,05) ; Figure 15.2), suggérant que comme dans les génomes de mammifères, la longueur médiane peut être utilisée comme une approximation vraisemblable de la longueur ancestrale.

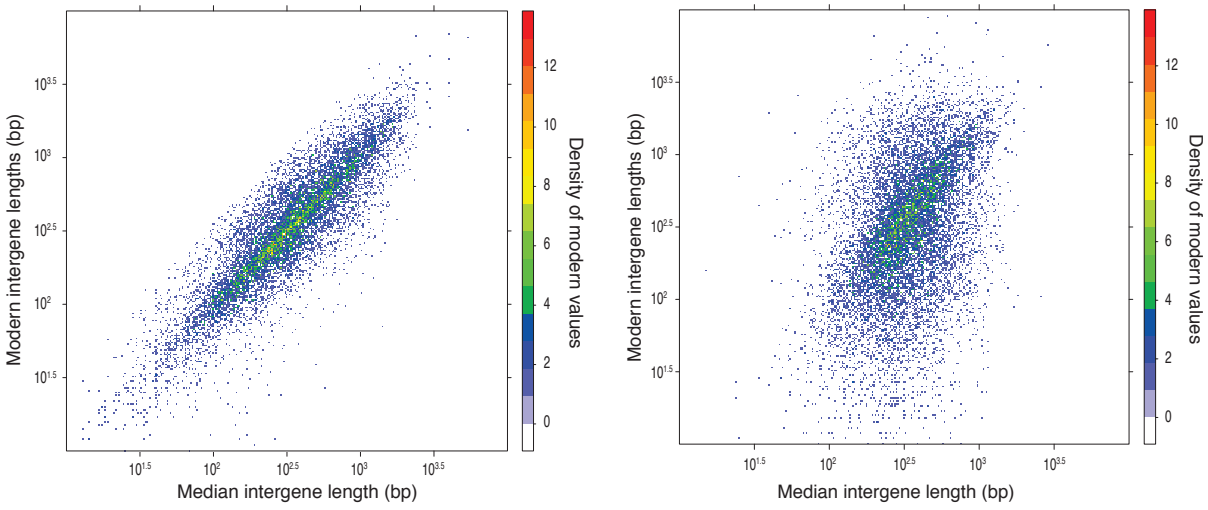


Figure 15.2. Estimation des longueurs d'intergènes ancestrales. A. Corrélation des longueurs d'intergènes orthologues à travers les génomes modernes. Pour chaque intergène ancestral, les différentes valeurs de longueurs modernes orthologues sont représentées en ordonnées contre leur valeur médiane en abscisse, utilisée comme un estimateur de la valeur ancestrale. Les points ont été groupés par classes de 0.01 en échelle log sur les deux axes, et la densité des données est représentée par un code couleur à droite. B. Corrélation observée avec les valeurs modernes randomisées.

15.2. Identification des points de cassure

En utilisant la même méthode d'identification des points de cassure que celle explicitée au paragraphe 12.1, nous avons identifié 505 réarrangements dans les lignées de *Lachancea kluyveri, Lachancea waltii* et *Kluyveromyces lactis* (Figure 15.3). La répartition des points de cassure dans l'arbre, avec notamment un grand nombre de réarrangements dans la lignée de *Kluyveromyces lactis* et un nombre restreint dans la lignée de *Lachancea kluyveri*, est cohérente avec les résultats rapportés par (Gordon et al. 2009).

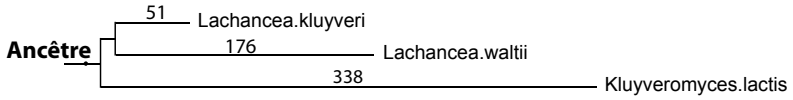


Figure 15.3. Répartition dans l'arbre phylogénétique des points de cassure de réarrangements identifiés dans les génomes de levures.

Les données publiées dans la littérature sur les caractéristiques des points de cassure de réarrangements dans les levures sont étonnamment rares. Les résultats rapportés montrent que les points de cassure sont corrélés avec le taux de recombinaison (élevé dans les régions riches en GC), avec les origines de réplication et les séquences répétées, et que le taux de cassure est plus faible autour des gènes essentiels (Pal and Hurst 2003; Di Rienzi et al. 2009; Gordon et al. 2009) ; ces caractéristiques sont concordantes avec celles des points de cassure de mammifères. Les points de cassure que nous obtenons chez les levures sont corrélés aux régions denses en gènes et riches en GC, comme observé également dans les génomes de mammifères.

15.3. Modélisation de la distribution des points de cassure

La distribution des points de cassure a été modélisée par régression de Poisson, comme décrit pour les mammifères (chapitre 13). Dans un premier temps, la variable utilisée pour modéliser le taux de cassure est la longueur des intergènes. De manière frappante, le taux de cassure intergénique chez les levures est également proportionnel à la longueur des intergènes après une transformation logarithmique, comme observé dans les génomes de mammifères (Figure 15.4). L'équation de régression obtenue est :

$$\log(r) = 0{,}43 * \log(L) - 4{,}90 \quad \leftrightarrow \quad r = 7{,}5 \,.\, 10^{-3} * L^{0{,}43}$$

L'adéquation du modèle est remarquablement bonne, puisque la longueur des intergènes suffit à expliquer la totalité de la variabilité du taux de cassure dans le génome (test de Chi^2 : P = 0,14 ; pseudo R^2 de McFadden = 0,81). Comme dans les génomes de mammifères, on a donc une très forte corrélation entre le taux de cassure et la longueur des intergènes, mais cette corrélation fait intervenir un exposant inférieur à 1 : le taux de cassure augmente moins vite que la longueur des intergènes.

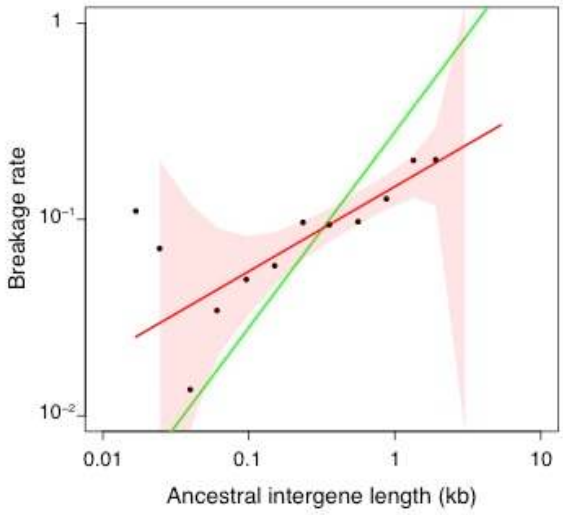


Figure 15.4. Corrélation entre la longueur des intergènes ancestraux (abscisse) et leur taux de cassure moyen dans les lignées descendantes (ordonnée). Le modèle de régression obtenu est différent de l'attendu sous le modèle aléatoire classique (ligne rouge : équation de régression ; zone colorée : intervalle de confiance à 95% du modèle ; ligne verte : attendu sous une distribution aléatoire pure).

Nous n'avons pas testé dans les génomes de levures l'importance des éléments de régulation, faute de données suffisamment complètes et facilement accessibles. Il est possible que, comme chez les mammifères, la présence d'éléments de régulation influe de manière significative en réduisant la probabilité de cassure ; mais, à nouveau, cette influence sera nécessairement mineure, la longueur des intergènes suffisant à elle seule à expliquer de manière adéquate la répartition des points de cassure.

Il faut noter que les résultats rapportés ici sont en contradiction partielle avec un précédent travail publié dans la littérature, qui cherchait à modéliser non pas la probabilité de cassure mais son inverse, la probabilité que deux gènes restent voisins entre deux espèces de levures (Poyatos and Hurst 2007). La méthode utilisée est proche de celle que nous proposons : les auteurs modélisent le fait qu'un intergène soit préservé entre deux espèces (ou non) par régression logistique, une méthode de régression multivariée qui permet de modéliser une variable binaire comme une fonction linéaire de prédicteurs. Cette étude montre que la longueur des intergènes est un prédicteur significatif mais faible de la probabilité de préservation d'un intergène, qui ne suffit pas à expliquer toute la variabilité observée. La différence fondamentale entre les deux méthodes se joue sur le fait que dans une régression logistique, l'hypothèse nulle est que tous les intergènes ont la même probabilité d'être préservés quelque soit leur taille. Dans ce contexte, il n'est pas intuitif d'appliquer la transformation logarithmique aux longueurs nécessaire à la linéarisation de la relation. Ainsi, le lien fort entre taux de cassure et longueur des intergènes n'est mis en évidence que dans le cadre d'une régression de Poisson (voir discussion).

A la lumière de ces résultats, le lien entre longueur des intergènes et taux de cassure est donc une propriété commune aux génomes de mammifères et à ceux des levures : on peut raisonnablement suggérer qu'il s'agisse d'une propriété générale des génomes eucaryotes. Bien

que les équations de régression obtenues soient en partie différentes, notamment les racines de la longueur obtenues (0,43 contre 0,28 pour les mammifères), ce résultat plaide en faveur d'un mécanisme d'origine physique, lié à la structure du génome et potentiellement à l'organisation de la chromatine, qui est l'une des propriétés partagées par tous les génomes eucaryotes (voir discussion).

Quatrième Partie

Etude d'un cas de duplication complète du génome : la duplication 3R dans le génome du poisson zèbre

Introduction

En parallèle du projet principal portant sur la distribution des points de cassure de réarrangement, un second projet annexe a été mené au cours de cette thèse. Notre équipe est en effet partenaire du consortium de séquençage du génome du poisson zèbre, un projet porté par le Sanger Institute et financé par le Wellcome Trust. Ce projet a débuté en 2001, et plusieurs versions du génome de poisson zèbre ont été mises à disposition de la communauté scientifique au fil des ans, en raison de l'intérêt majeur de ce poisson comme organisme modèle notamment dans le domaine du développement. Cependant, les assemblages du génome jusqu'ici n'étaient pas considérés comme des versions abouties et n'ont pas encore fait l'objet d'une publication globale décrivant les grandes caractéristiques du génomes du poisson zèbre, comme ont pu l'être la plupart des autres génomes séquencés de haute qualité. Dans l'objectif d'une telle publication, nous avons été chargés d'un certain nombre d'analyses de génomique comparative à réaliser sur le nouvel assemblage du génome du poisson zèbre (version Zv9, mise à disposition de la communauté scientifique en novembre 2010), afin de donner une vue d'ensemble des caractéristiques principales de ce génome.

Cette partie de la thèse est par essence beaucoup plus descriptive que le projet sur les réarrangements, et sans doute plus superficielle également pour deux raisons : tout d'abord, l'objectif était de donner un aperçu général de l'évolution du génome du poisson zèbre par rapport à la fois aux amniotes mais également aux autres poissons séquencés, et non à détailler précisément un mécanisme ou un point en particulier. Ensuite, ce projet a été réalisé sur quelques mois avec des contraintes de temps assez strictes. Nous avons néanmoins axé l'analyse sur les conséquences de la duplication 3R, qui s'est produite à la base de l'arbre des poissons téléostéens, sur l'organisation du génome du poisson zèbre. Les duplications complètes du génome sont en effet un phénomène relativement rare dans les génomes de vertébrés (à l'inverse des génomes de plantes), et qui ont pour conséquence de modifier profondément le génome à la fois du point de vue du contenu en gènes et de leur organisation les uns par rapport aux autres. Les génomes de poissons en général, et du poisson zèbre en particulier, sont un bon système d'étude pour comprendre les conséquences de ces duplications sur les génomes de vertébrés.

Si la question biologique de l'organisation des génomes est transversale aux deux parties de la thèse, ce projet recoupe surtout le travail principal sur les réarrangements par ses méthodes d'études, qui se basent largement sur la comparaison de l'ordre des gènes dans différents génomes et la composition en gènes du génome ancestral.

Chapitre 16. Analyse de la conservation de synténie dans le génome du poisson zèbre

Les duplications complètes de génomes, et la perte massive de gènes qui les suit (rediploïdisation), ont pour conséquence de modifier profondément l'ordre des gènes. Ainsi, les génomes de poissons téléostéens précédemment étudiés dans la littérature présentent une synténie dégradée avec les génomes amniotes, qui n'ont pas été affectés par la duplication 3R (Jaillon et al. 2004; Kasahara et al. 2007). Le génome du poisson zèbre est par ailleurs réputé pour être particulièrement réarrangé par rapport aux autres génomes de poissons téléostéens (Semon and Wolfe 2007a). Dans ce chapitre, nous avons exploré à quel degré la synténie reste conservée dans le génome du poisson zèbre par rapport à ceux des autres poissons et à ceux des amniotes. Ces statistiques de conservation de la synténie poisson zèbre/amniotes ont été comparées à celles d'autres poissons téléostéens afin de tester si la lignée du poisson zèbre présente effectivement une évolution accélérée.

16.1. Synténie conservée avec les téléostéens

En première approche, la conservation de la synténie au sens large (conservation des contenus géniques des chromosomes) a été inspectée entre le génome du poisson zèbre et ceux de trois autres poissons téléostéens dont l'assemblage est de bonne qualité : le medaka (*Oryzias latipes* ; Figure 16.1.A), le tétraodon (*Tetraodon nigroviridis* ; Figure 16.1.B) et l'épinoche (*Gasterosteus aculeatus* ; non représenté). La synténie entre les différents poissons est relativement élevée, en ce sens que la plupart des chromosomes du poisson zèbre ont exactement un chromosome orthologue dans chaque autre génome de poisson (une couleur dominante). On a donc eu peu de réarrangements interchromosomiques dans les différentes lignées de poissons. On note cependant quelques exceptions comme le chromosome 6, très réarrangé, ou une translocation bien visible au niveau du chromosome 18, qui est un mélange de deux chromosomes différents dans les autres génomes de poissons. La plupart des chromosomes contiennent des petites régions qui ne sont pas en synténie au sens large dans les autres génomes, notamment le chromosome 13 par exemple , qui est essentiellement orthologue au chromosome 17 du tétraodon mais contient également de nombreuses régions orthologues au chromosome 10. Ces régions sont dispersées le long du chromosome, suggérant une translocation dans la lignée du poisson zèbre suivie de réarrangements intrachromosomiques.

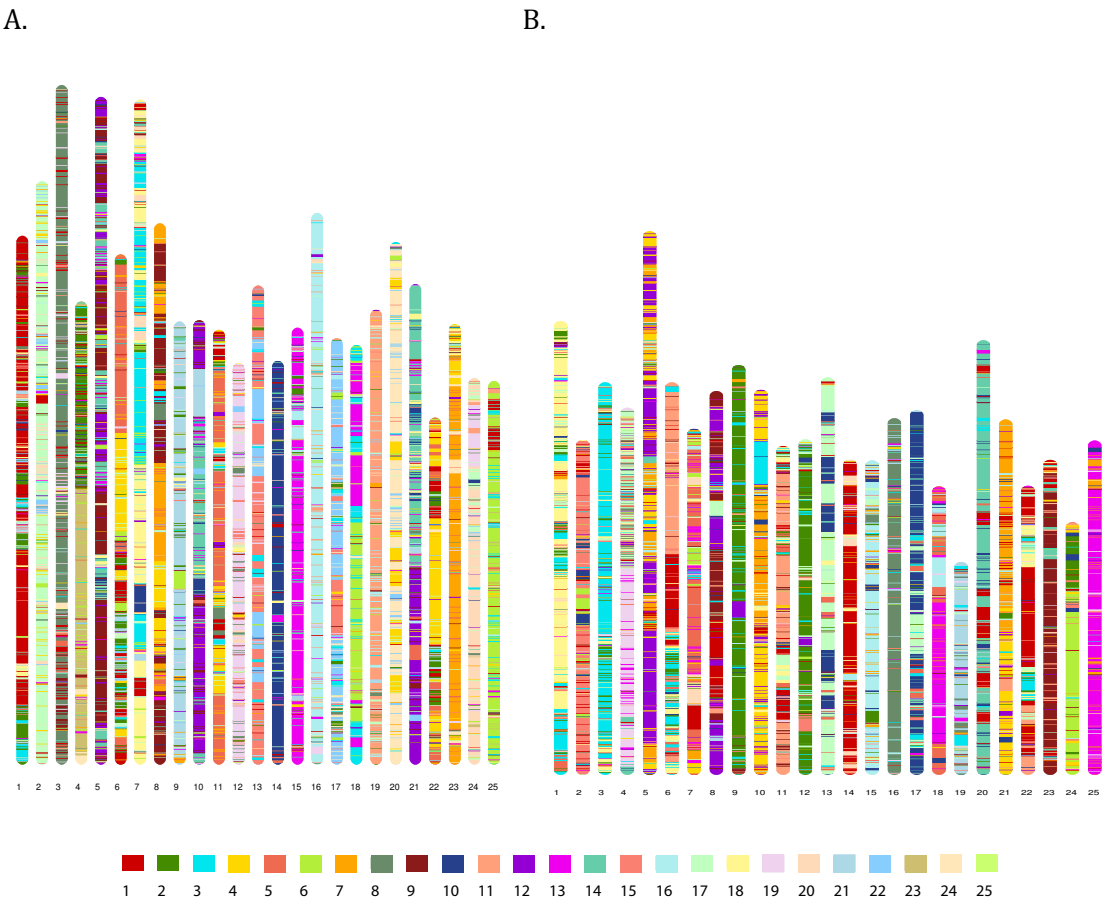


Figure 16.1. Caryotype du poisson zèbre colorisé en fonction de la position de chaque gène orthologue dans les chromosomes du medaka (A) et du tétraodon (B). La légende correspond à la couleur utilisée pour représenter chaque chromosome de l'espèce de comparaison. La taille des chromosomes n'est pas informative ; elle dépend du nombre d'orthologues dans le génome de comparaison.

Afin de quantifier plus précisément le degré de remodelage de l'ordre des gènes, nous avons calculé la longueur des blocs de synténie stricte entre le génome du poisson zèbre et ceux des autres poissons. On considère qu'un bloc de gènes est en synténie entre deux génomes si l'ordre et l'orientation de ces gènes est identique dans les deux génomes ; on tolère cependant jusqu'à *k* gènes intercalés entre deux gènes dont les orthologues sont colinéaires entre les deux génomes (les calculs ont été faits avec un paramètre de tolérance *k* de 2, 3, 5 et 8 gènes ; Table 16.1). Les blocs de synténie obtenus sont courts, avec une médiane à 3 gènes et une moyenne autour de 4 gènes dans toutes les comparaisons. Cette moyenne augmente très peu quand on relaxe le paramètre de tolérance, ce qui suggère qu'il ne s'agit pas d'un problème de blocs de synténie interrompus par des erreurs d'annotation ou d'assignation d'orthologues entre les espèces. A titre de comparaison, nous avons calculé la longueur des blocs de synténie existant entre les génomes de l'homme et du poulet, dont la distance phylogénétique est équivalente à celle qui sépare le poisson zèbre des trois poissons téléostéens testés (divergence homme-poulet datée de 326 Ma et divergence poisson zèbre-percomorphes datée de 320 Ma ; Figure 4.9). Les blocs de synténie homme-poulet sont nettement plus longs que ceux retrouvés dans les comparaisons

poisson zèbre-percomorphes, avec une médiane à 5 gènes et une moyenne autour de 10 gènes (Table 16.1).

Espèce 1	Espèce 2	*k*	Minimum	Maximum	Q25	Médiane	Q75	Moyenne	Ecart-type
Poisson zèbre (Zv9)	**Medaka**	2	2	32	2	**3**	4	**3.74**	2.97
Poisson zèbre (Zv9)	**Medaka**	3	2	34	2	**3**	4	**3.79**	3.07
Poisson zèbre (Zv9)	**Medaka**	5	2	34	2	**3**	4	**3.85**	3.24
Poisson zèbre (Zv9)	**Medaka**	8	2	52	2	**3**	4	**3.91**	3.40
Poisson zèbre (Zv9)	**Tetraodon**	2	2	30	2	**3**	5	**4.03**	3.30
Poisson zèbre (Zv9)	**Tetraodon**	3	2	30	2	**3**	5	**4.10**	3.41
Poisson zèbre (Zv9)	**Tetraodon**	5	2	47	2	**3**	5	**4.16**	3.61
Poisson zèbre (Zv9)	**Tetraodon**	8	2	47	2	**3**	5	**4.23**	3.73
Poisson zèbre (Zv9)	**Epinoche**	2	2	30	2	**3**	4	**3.87**	3.06
Poisson zèbre (Zv9)	**Epinoche**	3	2	30	2	**3**	5	**3.92**	3.16
Poisson zèbre (Zv9)	**Epinoche**	5	2	33	2	**3**	5	**4.00**	3.31
Poisson zèbre (Zv9)	**Epinoche**	8	2	46	2	**3**	5	**4.08**	3.46
Homme	**Poulet**	2	2	205	9	**5**	12	**10.42**	14.66
Homme	**Poulet**	3	2	271	9	**5**	12	**10.75**	16.91
Homme	**Poulet**	5	2	283	9	**5**	12	**11.00**	18.05
Homme	**Poulet**	8	2	281	9	**5**	12	**11.31**	18.64

Table 16.1. Statistiques de longueur des blocs de synténie entre le génome du poisson zèbre et ceux du medaka, du tétraodon et de l'épinoche. Les statistiques entre le génome de l'homme et du poulet sont fournies à titre de comparaison.

Il semble donc que la synténie soit particulièrement peu préservée entre le poisson zèbre et les autres poissons. Comme les trois autres poissons sont plus apparentés entre eux qu'avec le poisson zèbre, les comparaisons de leurs génomes ne sont pas informatifs pour trancher si cette rétention de synténie limitée est une caractéristique générale des génomes de poissons, peut-être liée à la duplication complète, ou s'il s'agit d'une conséquence d'un taux de réarrangements particulièrement élevé dans le génome du poisson zèbre. Pour tenter de répondre à cette question, nous nous sommes intéressés à la conservation de la synténie entre le génome du poisson zèbre et ceux des amniotes.

16.2. Synténie conservée avec les amniotes

Comme précédemment, nous avons inspecté visuellement la conservation de la synténie au sens large entre le poisson zèbre et deux amniotes, l'homme et le poulet (Figure 16.2). On constate immédiatement que la conservation existe mais est assez pauvre : chaque chromosome du poison zèbre correspond à une mosaïque de gènes qui se trouvent sur des chromosomes différents chez les amniotes.

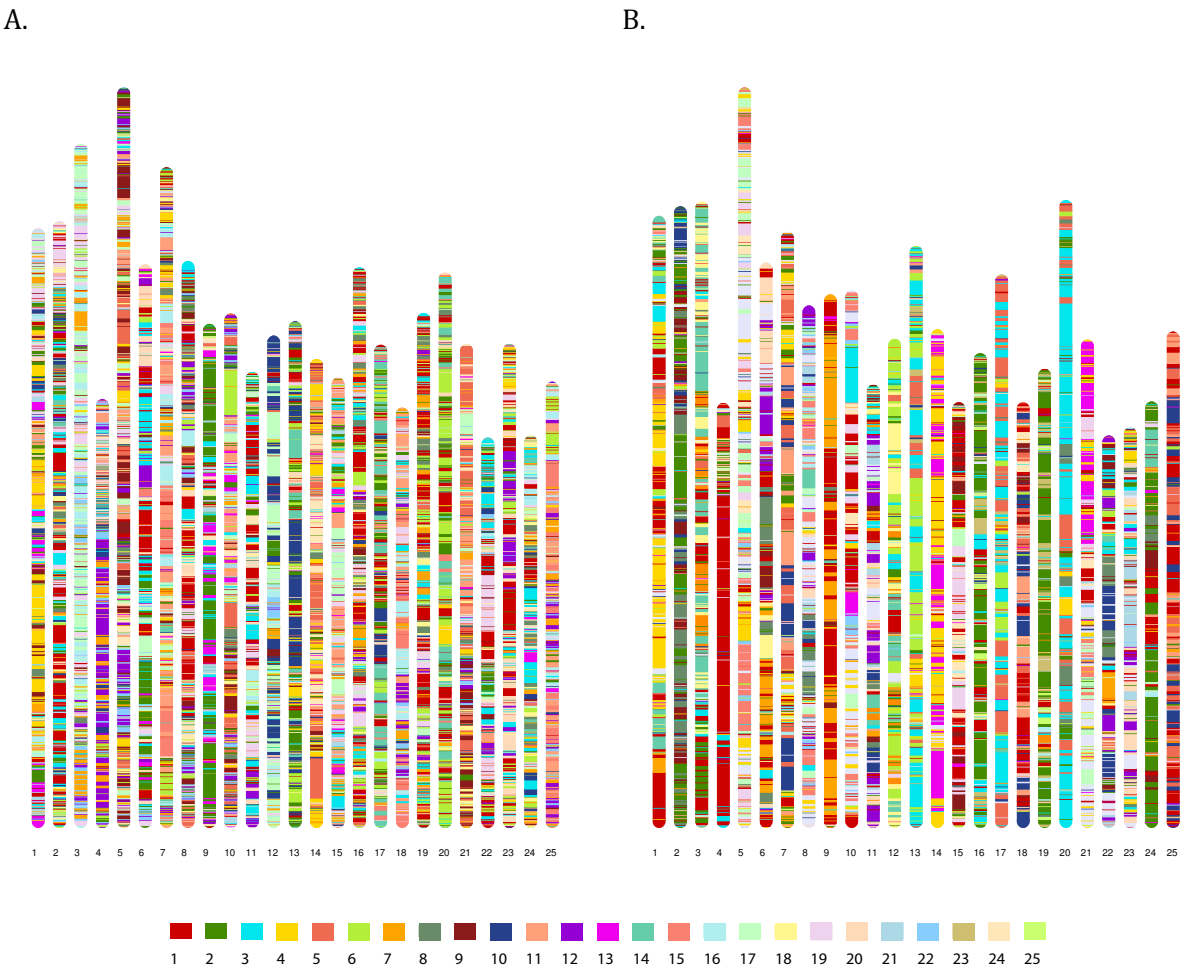


Figure 16.2. Caryotype du poisson zèbre colorisé en fonction de la position de chaque gène orthologue dans les chromosomes de l'humain (A) et du poulet (B). La légende correspond à la couleur utilisée pour représenter chaque chromosome de l'espèce de comparaison. La taille des chromosomes n'est pas informative ; elle dépend du nombre d'orthologues dans le génome de comparaison.

Nous avons ensuite calculé plus précisément la longueur des blocs de synténie stricte entre le génome du poisson zèbre et celui du poulet. Le génome du poulet a été choisi comme référence car il a conservé une structure proche de celle de l'ancêtre Euteleostomi (Nakatani et al. 2007), et permet donc une meilleure sensibilité dans les analyses de synténie conservée avec les génomes de poissons par rapport à d'autres génomes comme ceux de l'homme ou de la souris. Dans tous les cas, les blocs de synténie sont très courts, avec une moyenne qui dépasse juste les 3 gènes et une médiane à 2 (Table 16.2). Relaxer le paramètre de tolérance permet d'augmenter légèrement la taille de blocs, comme attendu, mais de manière très faible, la moyenne du nombre de gènes par blocs montant à 3,22 gènes pour k = 8. Ceci confirme que le degré de rétention de synténie stricte entre le poisson zèbre et les amniotes est très faible, ce qui est attendu puisque la synténie était déjà peu conservée au sein des poissons. Comme l'augmentation de la tolérance n'augmente que marginalement la sensibilité, ce sont les blocs de synténie obtenus avec le paramètre k = 2 qui ont été retenus par la suite pour les analyses.

Espèce 1	Espèce 2	*k*	Minimum	Maximum	Q25	Médiane	Q75	Moyenne	Ecart-type	Nombre de blocs
Poisson zèbre (Zv9)	Poulet	2	2	22	2	3	5	3,08	1,98	3071
Poisson zèbre (Zv9)	Poulet	3	2	22	2	3	5	3,13	2,04	3086
Poisson zèbre (Zv9)	Poulet	5	2	22	2	3	6	3,17	2,09	3131
Poisson zèbre (Zv9)	Poulet	8	2	29	2	3	6	3,22	2,19	3144
Medaka	Poulet	2	2	55	2	3	6	3,24	2,46	2433
Medaka	Poulet	3	2	56	2	3	6	3,3	2,54	2433
Medaka	Poulet	5	2	56	2	3	6	3,35	2,58	2458
Medaka	Poulet	8	2	67	2	4	6	3,38	2,75	2500
Tetraodon	Poulet	2	2	50	2	3	6	3,35	2,66	1954
Tetraodon	Poulet	3	2	50	2	3	6	3,38	2,72	1977
Tetraodon	Poulet	5	2	50	2	4	6	3,42	2,73	1999
Tetraodon	Poulet	8	2	62	2	4	7	3,49	2,89	2013
Epinoche	Poulet	2	2	35	2	3	6	3,28	2,41	2638
Epinoche	Poulet	3	2	35	2	3	6	3,32	2,5	2647
Epinoche	Poulet	5	2	35	2	3	6	3,35	2,53	2688
Epinoche	Poulet	8	2	35	2	4	6	3,4	2,56	2718

Table 16.2. Statistiques de longueur des blocs de synténie entre les génomes de poissons téléostéens et celui du poulet.

16.3. Comparaison aux autres génomes de poissons

16.3.1. Dégradation de la synténie dans les génomes de téléostéens

La conservation de l'ordre des gènes entre le génome du poisson zèbre et les génomes amniotes est pauvre ; nous avons testé si les autres poissons téléostéens présentent un degré de rétention de la synténie avec les amniotes supérieur à celui observé dans le génome du poisson zèbre. De la même façon, nous avons donc calculé les blocs de synténie stricte entre le génomes du poulet et ceux du medaka, de l'épinoche et du tétraodon (Table 16.2). Dans tous les génomes de téléostéens, les blocs de synténie avec le génome du poulet sont extrêmement courts, avec à nouveau une moyenne à peine supérieure à 3 gènes quelque soit le paramètre de tolérance utilisé. A titre comparatif, la même expérience a été faite entre le génome du poulet et celui du coelacanthe, dont la divergence avec les amniotes suit de près celle des poissons téléostéens dans l'arbre des espèces (420 et 400 millions d'années). Malgré la forte fragmentation de l'assemblage actuellement disponible pour le génome de coelacanthe, celui-ci présente de meilleures statistiques de conservation de la synténie avec le poulet qu'aucun des poissons téléostéens testés (Table 16.3).

Espèce 1	Espèce 2	*k*	Minimum	Maximum	Q25	Médiane	Q75	Moyenne	Ecart-type	Nombre de blocs
Coelacanthe	Poulet	2	2	41	3	**5**	9	**4,26**	3,65	**2126**
Coelacanthe	Poulet	3	2	41	3	**5**	9	**4,26**	3,65	**2135**
Coelacanthe	Poulet	5	2	41	3	**5**	9	**4,29**	3,7	**2130**
Coelacanthe	Poulet	8	2	41	3	**5**	9	**4,33**	3,73	**2118**

Table 16.3. Statistiques de longueur des blocs de synténie entre le génome du coelacanthe et celui du poulet.

Il semble donc que la synténie par rapport aux amniotes soit particulièrement dégradée dans l'ensemble des génomes de téléostéens, et non uniquement dans celui du poisson zèbre. Il manque actuellement un génome de bonne qualité ayant divergé après la séparation téléostéens/amniotes mais avant la duplication 3R pour pouvoir tester formellement si cette dégradation de la synténie est effectivement spécifique des génomes ayant subi la duplication complète du génome 3R et serait une conséquence de cette duplication, mais cette hypothèse est

vraisemblable au vu des connaissances notamment chez les plantes ayant subi des duplications complètes du génome.

16.3.2. Dégradation non spécifique au poisson zèbre

Bien que tous les génomes de poissons téléostéens présentent une synténie très dégradée par rapport aux amniotes, la distribution des tailles de blocs de synténie montre que les blocs du poisson zèbre sont malgré tout significativement plus courts en moyenne que ceux des autres poissons téléostéens (Figure 16.3).

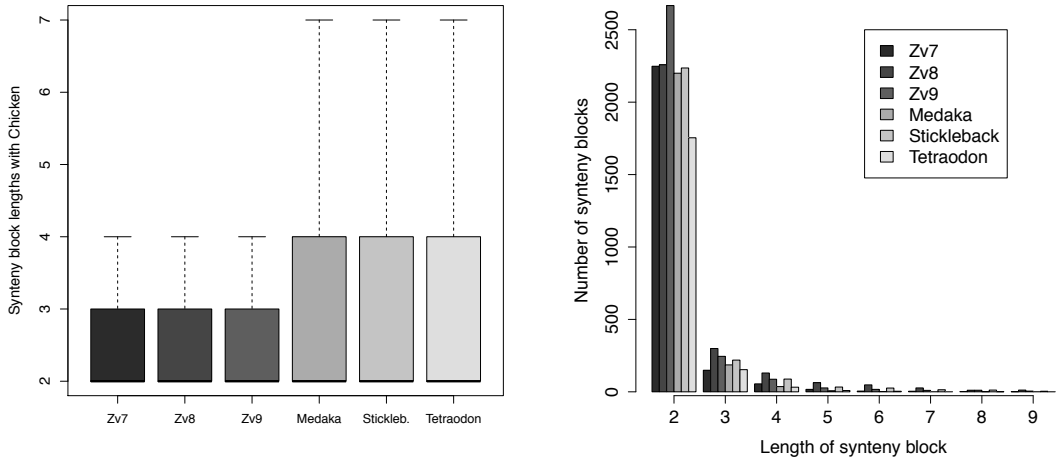


Figure 16.3. Distribution de la taille des blocs de synténie entre les poissons téléostéens et le poulet, représentés en boxplot et en histogramme. Pour le poisson zèbre, trois assemblages successifs du génome sont utilisés (Zv7, Zv8 et Zv9, la version utilisée dans ces analyses).

Cependant, ces blocs sont aussi nettement plus nombreux en nombre absolu dans le génome du poisson zèbre, avec un excès à la fois de blocs courts et de blocs longs par rapport aux autres génomes (Table 16.2). Ce résultat pourrait être une conséquence du plus grand nombre d'orthologues annotés entre le génome du poulet et celui du poisson zèbre qu'avec celui d'aucun des autres poissons, qui confèrerait à la comparaison poulet/poisson zèbre une granularité plus fine. Afin de comparer la conservation de synténie sur les mêmes bases entre toutes les espèces, nous avons recherché les blocs de synténie en ne considérant que l'ordre et l'orientation des gènes qui existent en exactement une copie dans les quatre génomes de poissons considérés (8937 gènes). De fait, lorsque la comparaison est faite à nombre d'orthologues égal entre les différents poissons, la différence entre la longueur moyenne des blocs de synténie du poisson zèbre et celle des autres poissons n'est plus significative après correction pour tests multiples, et la distribution des tailles de blocs de synténie dans le génome du poisson zèbre devient comparable à celle des autres espèces (Figure 16.4).

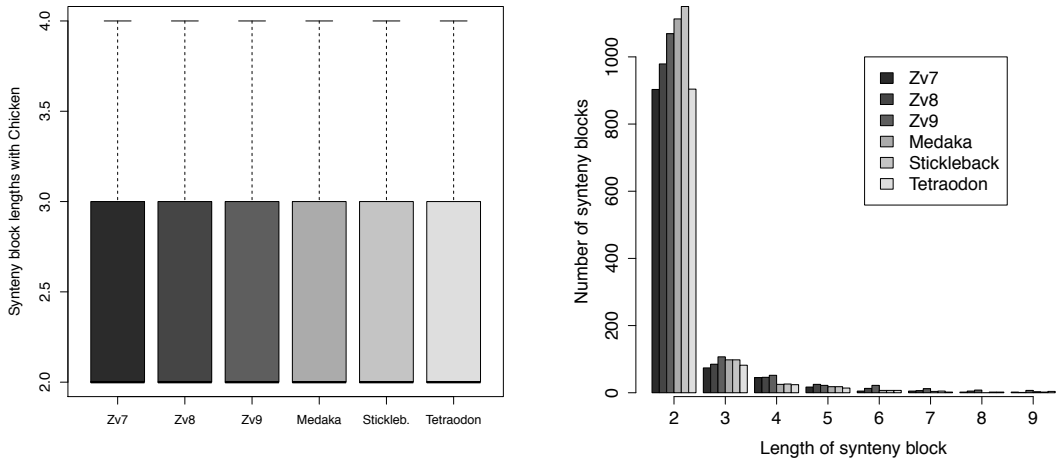


Figure 16.4. Distribution de la taille des blocs de synténie entre les poissons téléostéens et le poulet lorsque seuls les gènes présents en exactement une copie dans chaque génome sont considérés.

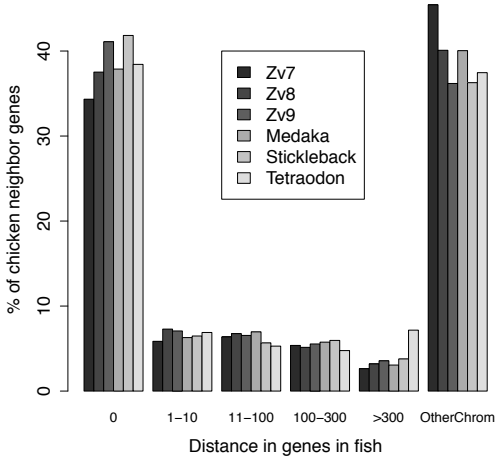


Figure 16.5. Distribution de la distance séparant les orthologues de gènes adjacents dans le génome du poulet, dans les différents génomes de poissons téléostéens. Pour le poisson zèbre, trois assemblages successifs du génome sont utilisés (Zv7, Zv8 et Zv9, la version utilisée dans ces analyses).

Bien que la conservation locale des adjacences de gènes soit équivalente entre le génome de poisson zèbre et ceux des autres poissons, celle-ci est très pauvre dans tous les cas et laisse donc peu de place à la détection de différences entre les génomes de poissons. Nous avons donc testé la possibilité que les liaisons à plus longue distance entre les gènes soient plus dégradées dans le cas du génome du poisson zèbre. Nous avons utilisé une mesure plus globale de la conservation de synténie proche de celle utilisée dans (Semon and Wolfe 2007a) : pour chaque paire de gènes adjacents dans le génome du poulet, nous avons mesuré le nombre de gènes qui séparent leurs gènes orthologues dans le génome de chaque poisson (en ne prenant en compte que les gènes ayant des orthologues dans les deux génomes pour chaque comparaison). Avec cette mesure

plus relaxée de la conservation de synténie, le profil du génome du poisson zèbre est très similaire à ceux des autres poissons (Figure 16.5). Ce résultat confirme que le poisson zèbre ne semble pas davantage réarrangé que les autres génomes de poissons, contrairement aux conclusions rapportées par (Semon and Wolfe 2007a).

16.4. Conséquences de la qualité de l'assemblage dans le génome du poisson zèbre

La contradiction entre nos résultats et ceux rapportés par les travaux précédents pourrait être une conséquence de la moindre qualité des assemblages précédemment disponibles pour le génome du poisson zèbre. En effet, l'assemblage utilisé dans (Semon and Wolfe 2007a) est la version Zv5 du génome du poisson zèbre, mise à disposition de la communauté en mai 2005. Les assemblages antérieurs à la version Zv8 n'étaient pas ancrés à la carte génétique du génome du poisson zèbre, mais à une carte d'hybrides de radiation, moins fiable au niveau des liaisons génétiques à longue distance. Afin d'éclaircir ces données contradictoires avec la littérature, nous avons comparé les distributions de longueur de blocs de synténie entre le génome du poulet et celui du poisson zèbre avec ceux des autres poissons, mais en utilisant cette fois deux versions antérieures du génome du poisson zèbre encore accessibles en ligne, les versions Zv7 (juillet 2007) et Zv8 (décembre 2008). Même lorsque seuls les gènes présents en une copie dans chacun des génomes sont considérés, les longueurs moyennes des blocs de synténie obtenus avec les anciennes versions du génome du poisson zèbre restent plus faibles que celles des autres poissons (test de Wilcoxon : $P < 8.10^{-3}$ pour toutes les comparaisons), contrairement aux résultats obtenus avec la version Zv9 (aucune comparaison significative après correction de Bonferroni pour les tests multiples). La version Zv7, en particulier, présente à la fois des blocs de synténie courts et en nombre réduit par rapport aux autres espèces (Figure 16.4). Ce résultat est confirmé par la mesure relaxée de la synténie présentée à la Figure 16.5 : le nombre de gènes voisins à la fois chez le poulet et le poisson zèbre version Zv7 est inhabituellement bas, et la proportion de gènes voisins chez le poulet et sur des chromosomes différents chez le poisson zèbre y est particulièrement importante. Ce profil recoupe les observations de (Semon and Wolfe 2007a), interprétées par la suite comme un taux plus élevé de réarrangements intra- et interchromosomiques dans le génome du poisson zèbre par rapport aux autres poissons téléostéens. Les nouveaux éléments apportés par nos travaux montrent donc que le poisson zèbre ne semble pas plus réarrangé que les autres génomes de poissons ; de fait, c'est la mauvaise qualité des versions précédentes de l'assemblage du génome qui a induit en erreur les analyses précédentes et donné la fausse impression d'un taux de réarrangement particulièrement élevé.

16.5. Analyse des plus longs blocs de synténie conservée

Les blocs de synténie conservée entre le génome du poisson zèbre et les génomes amniotes sont en moyenne très courts, notamment avec le génome de l'homme qui est plus réarrangé que celui du poulet. On peut néanmoins détecter plusieurs trains de gènes relativement longs dont l'ordre et l'orientation sont conservés entre le génome du poisson zèbre et celui de l'homme. Ces gènes sont d'un intérêt particulier, puisque cette préservation de la colinéarité des gènes peut être la marque de contraintes fonctionnelles sur la topologie des gènes préservées au cours de l'évolution des vertébrés. Les dix blocs les plus longs parmi les blocs de synténie conservée homme/poisson zèbre comptent entre 19 et 15 gènes, et couvrent au total 170 gènes (Figure 16.6). Parmi eux se trouve le cluster *HoxD*, connu pour être bien préservé dans les génomes de poissons (Lee et al. 2006). Lorsque l'on inspecte les régions orthologues dans les autres génomes de vertébrés à l'aide du serveur Genomicus (Muffato et al. 2010), on constate que la colinéarité des gènes y est également bien préservée en général, un indice d'une rétention probablement fonctionnelle plutôt que liée au hasard. Il faut cependant noter que certaines de ces régions sont peu ou non préservées dans certains génomes de poissons, peut-être parce que le doublement du contenu génique et régulateur dû à la duplication complète du génome a permis une relaxation des contraintes fonctionnelles et une réorganisation de ces régions.

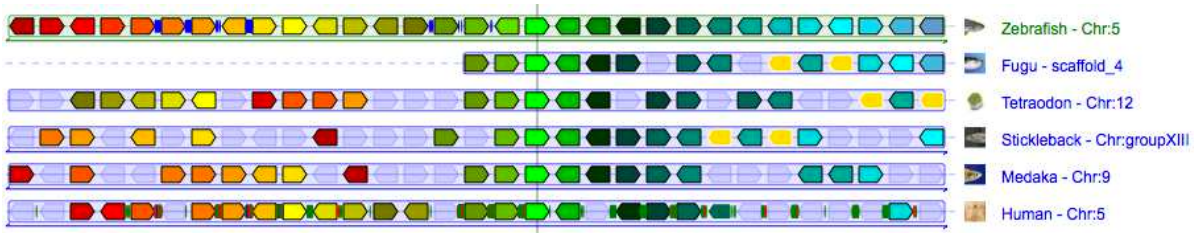


Figure 16.6. Visualisation du plus long bloc de synténie conservée entre les génomes du poisson zèbre et de l'homme (19 gènes), et conservation de la région dans les autres génomes de téléostéens. Chaque bloc représente un gène ; les blocs de même couleur sont les gènes orthologues d'un génome à l'autre, et les blocs grisés sont des gènes qui n'ont pas orthologues dans cette région du génome du poisson zèbre (utilisé comme espèce de référence).

Les blocs de régulation génomique (GRB), qui impliquent des enhancers à longue distance et leurs gènes cibles, ont été proposés comme déterminants de la conservation de la synténie dans les génomes de vertébrés. Les travaux de (Kikuta et al. 2007), notamment, ont mis en évidence que dans le génome du poisson zèbre, les gènes du développement se trouvent dans des blocs de synténie avec l'homme plus longs (en taille absolue) que les autres catégories fonctionnelles de gènes, et la majorité des plus grands blocs de synténie homme/poisson zèbre contiennent au moins un gène impliqué dans le développement et une forte concentration de séquences conservées non-codantes, interprétées comme des enhancers putatifs. Leur conclusion est que les interactions de régulation entre des séquences non-codantes se trouvant à longue distance de leurs gènes cibles contraignent certaines régions du génome, en raison des conséquences délétères des réarrangements qui dissocieraient le gène d'une partie de ses séquences de régulation. Les auteurs en déduisent notamment que la détection des gènes cibles de réseaux

complexes de régulation pourrait se faire en utilisant la densité en éléments conservés non-codants, une idée implémentée par la suite pour identifier les cibles de GRB que nous utilisons au chapitre 11 de ce manuscrit (Engstrom et al. 2008).

Ce constat est en partie vérifié par nos propres résultats sur les facteurs influençant la probabilité de réarrangement dans les génomes de vertébrés : en effet, nous avons mis en évidence au chapitre 13 que bien que la longueur des intergènes soit le déterminant majeur de la cassure, les régions contenant une forte proportion d'éléments conservés non-codants sont moins susceptibles d'être réarrangées que les régions moins conservées au niveau de la séquence ; par ailleurs, les gènes cibles proposés sont effectivement moins souvent trouvés au bord de points de cassure de réarrangements évolutifs. Nous avons donc testé si les cinquante plus longs blocs de synténie détectés entre le génome du poisson zèbre et celui de l'homme sont particulièrement riches en éléments conservés non-codants, ce qui viendrait appuyer l'hypothèse des GRB. Les éléments conservés non-codants utilisés ici sont les éléments détectés par PhastCons (Siepel et al. 2005). Bien qu'après inspection manuelle, on constate que de nombreux éléments conservés non codants parsèment ces plus grands blocs de synténie, la proportion de séquence conservée non-codante dans les cinquante plus grands blocs est tout-à-fait comparable à ce que l'on attend dans un échantillonnage aléatoire de blocs de gènes de même taille dans le génome du poisson zèbre (Figure 16.7). Ce résultat est cohérent avec nos résultats sur les points de cassure de réarrangement : la pression de sélection exercée par les relations gène/séquence de régulation existe et est visible à l'échelle évolutive, mais son effet est presque négligeable devant l'influence mécanique de l'organisation du génome. Ainsi, il n'est pas surprenant de ne pas observer d'enrichissement notable en séquences conservées non-codantes dans les grands blocs de synténie.

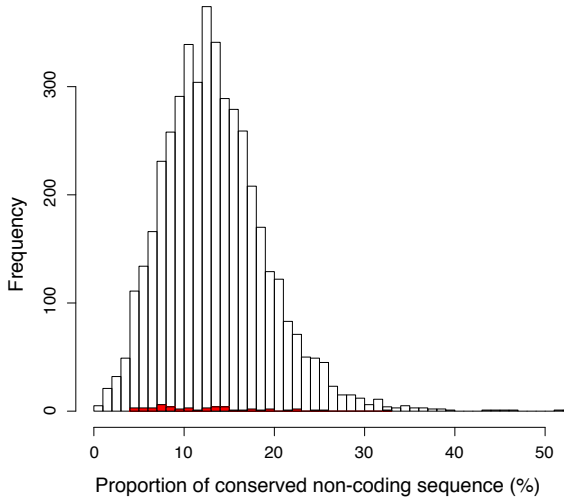


Figure 16.7. Comparaison des proportions de séquence conservée non-codante observés dans les 50 plus grands blocs de synténie conservée entre le génome du poisson zèbre et de l'humain (rouge), et de la distribution des proportions attendues dans des blocs de gènes de même taille échantillonnés aléatoirement.

Nous avons ensuite testé si les plus longs blocs de synténie sont enrichis en gènes du développement, comme suggéré par les résultats de (Kikuta et al. 2007). Une analyse des termes Gene Ontology associés aux gènes des cinquante plus longs blocs de synténie montre que ces blocs sont principalement enrichis en gènes impliqués dans l'adhésion cellulaire (cell-cell adhesion, $P = 8.10^{-11}$), et la liaison au calcium (calcium binding, $P = 3.10^{-7}$; Figure 16.8) par rapport au reste du génome. Aucun terme associé de façon claire aux processus développementaux ne ressort de l'analyse, contrairement à nos attentes. Ainsi, même si les grands blocs de synténie contiennent des gènes impliqués dans le développement, comme décrit par (Kikuta et al. 2007), ils ne comptent pas d'enrichissement particulier par rapport au reste du génome. De fait, la seule caractéristique notable que nous trouvons dans ces régions est que les intergènes y sont légèrement plus courts que l'ensemble du génome (test de Wilcoxon : $P = 7.10^{-5}$; Figure 16.9). La bonne conservation de la synténie dans ces régions à travers les différents génomes de vertébrés suggère pourtant qu'une pression de sélection existe effectivement et que ces blocs de synténie ne sont pas simplement retenus par défaut ; cependant, nous n'identifions pas de caractéristiques particulières dans ces régions, ce qui suggère que les contraintes fonctionnelles qui s'y exercent ne suivent pas un patron commun à tous les blocs mais sont le résultat d'interactions complexes.

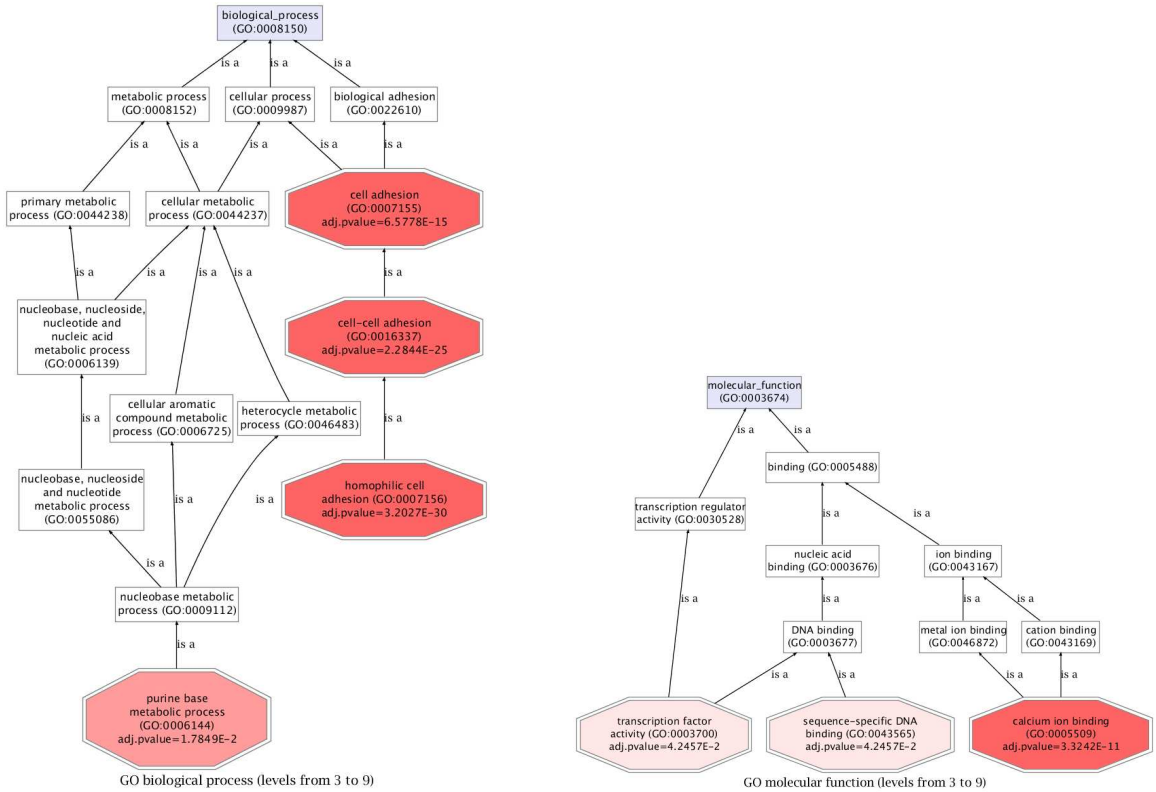


Figure 16.8. Arborescence des termes Gene Ontology significativement enrichis parmi les gènes inclus dans les 50 plus longs blocs de synténie conservés entre le poisson zèbre et l'humain, visualisée avec GOGraphViz (http://babelomics.bioinfo.cipf.es).

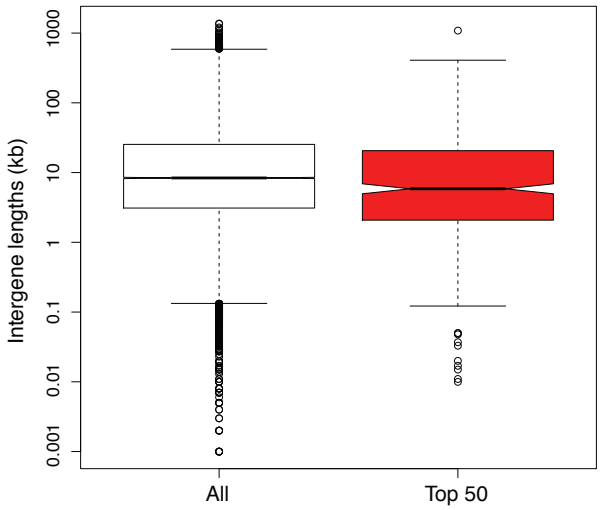


Figure 16.9. Distribution de la longueur des intergènes dans l'ensemble du génome du poisson zèbre (gauche) et dans les 50 plus longs blocs de synténie conservée avec l'humain.

Chapitre 17. Identification des blocs de synténie double-conservée dans le génome du poisson zèbre

Les duplications complètes du génome laissent une signature caractéristique sur l'organisation des génomes : en effet, chaque région d'un génome voisin non dupliqué est orthologue à deux régions différentes dans le génome dupliqué. Après la perte massive de gènes lors de la rediploïdisation du génome, les deux régions dupliquées conservent l'une et l'autre une partie du contenu génique initial et présentent donc un motif d'alternance des gènes par rapport à la région orthologue d'un génome voisin non dupliqué (pour rappel, Figure 4.8). Ces blocs de synténie qui s'imbriquent l'un dans l'autre pour reconstituer le contenu en gènes initial sont appelés blocs de synténie double-conservée ou blocs de synténie dédoublée. La détection de ces synténies double-conservées permet de distinguer les gènes retenus en deux copies après la duplication complète, c'est-à-dire les ohnologues, par rapport aux paralogues issus d'autres duplications précédant ou suivant la duplication complète. Nous avons cherché à caractériser les régions de synténie double-conservée dans le génome du poisson zèbre, et à identifier les ohnologues issus de la duplication complète 3R.

17.1. Détection des blocs de synténie double-conservée

Comme nous l'avons vu au chapitre précédent, l'ordre des gènes est très peu conservé entre les génomes de poissons et les génomes des amniotes non dupliqués, même en considérant l'ordre des gènes de façon assez relâchée. On utilise donc ici une définition large de la synténie double-conservée, en cherchant à détecter des régions dont les contenus en gènes sont complémentaires et reconstituent celui d'une seule région orthologue dans un génome non dupliqué, sans tenir compte de l'ordre des gènes au sein de ces régions. La méthode utilisée est décrite dans le Matériel et Méthodes au chapitre 8, en utilisant comme espèces de référence non dupliquées les 43 génomes d'amniotes disponibles dans la version 63 d'Ensembl. Les blocs de synténie double-conservée (DCS) ainsi identifiés avec toutes les espèces incluent 21926 gènes, soit 84% du génome du poisson zèbre. La Figure 17.1 représente les blocs de DCS identifiés entre le génome du poisson zèbre et ceux de l'homme et du poulet.

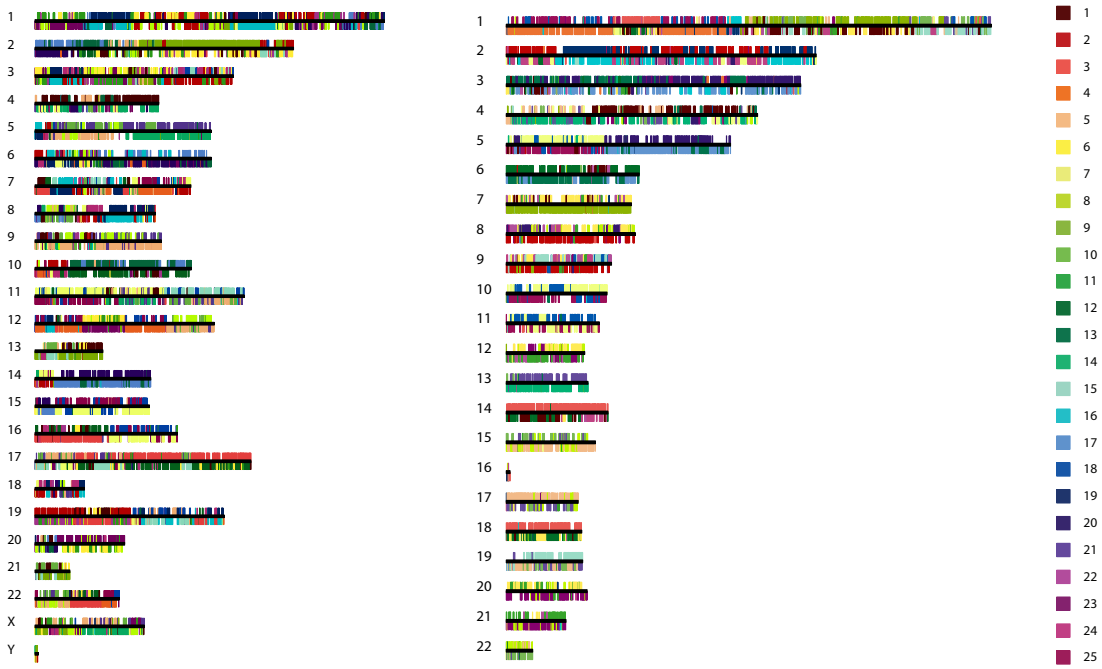


Figure 17.1. Synténies double-conservées entre le génome du poisson zèbre et les génomes de l'homme (gauche) et du poulet (droite). Chaque chromosome du génome amniote de référence est figuré par une barre noire : pour chaque gène, le ou les chromosomes portant les orthologues dans le génome du poisson zèbre sont représentés de part et d'autre suivant un code couleur (légende à droite).

17.2. Identification des ohnologues 3R

En utilisant les blocs de DCS identifiés, on peut distinguer les paralogues issus de la duplication complète du génome (ohnologues 3R) de ceux apparus lors de duplications locales. La majorité du génome étant restée telle que la synténie double-conservée est toujours détectable, on fait l'approximation que la plupart des ohnologues sont restés dans leur environnement de synténie : on identifie comme ohnologues une paire de paralogues faisant partie l'un et l'autre de blocs alternants identifiés parmi les blocs de DCS. Ainsi, nous identifions 3440 paires de gènes comme étant des ohnologues issus de la duplication 3R, pour un total de 8083 gènes en incluant les duplications postérieures de certains gènes : 31% du génome du poisson zèbre est donc constitué de gènes ohnologues issus de la duplications 3R, un taux cohérent avec les estimations précédentes sur un petit nombre de gènes (Postlethwait et al. 2000).

Lorsque l'on représente ces paires d'ohnologues dans le génome du poisson zèbre en utilisant une représentation circulaire avec le programme Circos (Krzywinski et al. 2009), on observe immédiatement que certains chromosomes descendent du même chromosome ancestral pré-duplication et partagent la grande majorité de leurs paires d'ohnologues, comme par exemple les chromosomes 3 et 12, 17 et 20, ou 16 et 19 (Figure 17.2). D'autres chromosomes en revanche, comme le chromosome 2, partagent des ohnologues avec plusieurs chromosomes, une signature de réarrangements interchromosomiques postérieurs à la duplication complète.

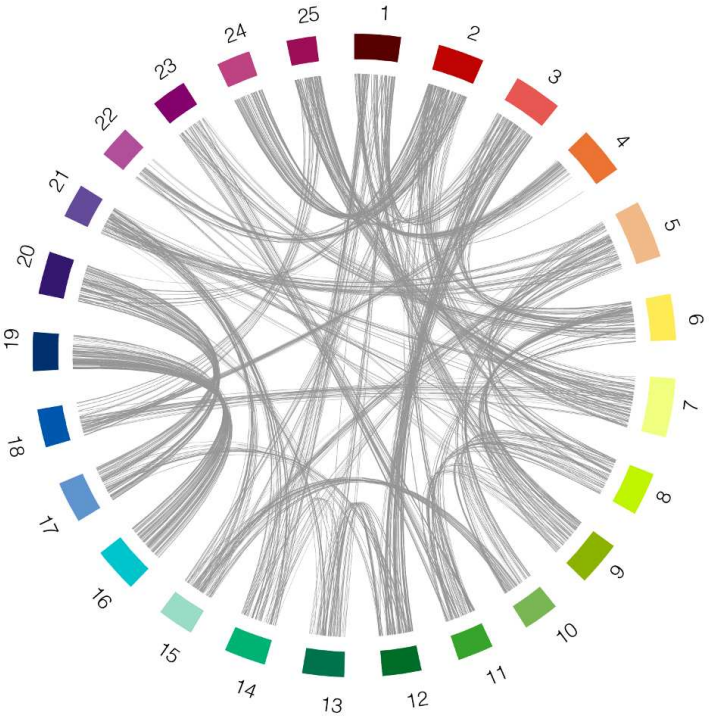


Figure 17.2. Paires de gènes ohnologues dans le génome du poisson zèbre. Chaque bloc représente un chromosome, et chaque lien relie les positions de deux gènes ohnologues. Les liens entre chromosomes partageant moins de 20 paires d'ohnologues ne sont pas représentés pour plus de clarté.

17.3. Comparaison aux autres génomes de poissons

A titre de comparaison, nous avons détecté les blocs de DCS dans les trois autres génomes de poissons (medaka, épinoche et tétraodon), et identifié les ohnologues encore présents dans ces génomes. Les génomes du medaka, de l'épinoche et du tétraodon contiennent respectivement 27%, 19% et 29% d'ohnologues. Afin de pouvoir comparer rigoureusement la rétention des ohnologues dans les différents génomes, il est nécessaire de raisonner en termes de gènes ancestraux retenus en une ou deux copies, pour s'affranchir des problèmes de duplications de gènes spécifiques à chaque lignée, qui modifient les comptages modernes. Les gènes ancestraux pré-duplication ont été déduits des arbres de gènes d'Ensembl, afin d'obtenir le comptage des gènes toujours présents en au moins une copie dans chaque génome de poisson. On peut ensuite comparer la proportion de gènes ancestraux retenus ou non comme ohnologues dans chaque génome (Table 17.1). Le génome du poisson zèbre contient plus de gènes ancestraux retenus sous forme d'ohnologues, à la fois en nombre et en proportion, qu'aucun des trois autres génomes de poisson (tests de Chi^2 : $P < 3.10^{-5}$ dans tous les cas).

Génome	Gènes ancestraux	Singletons	% Singletons	Ohnologues	% Ohnologues	Total
Poisson zèbre	14168	10728	75,7	3440	24,3	26039
Medaka	12792	10274	80,3	2518	19,7	19686
Epinoche	12873	10025	77,9	2848	22,1	20787
Tetraodon	13292	11579	87,1	1713	12,9	19602

Table 17.1. Proportions de gènes ancestraux retenus sous forme d'ohnologues dans les différents génomes de poissons téléostéens.

17.4. Architecture chromosomique et taux de réarrangements

Avoir à notre disposition des paires d'ohnologues permet d'explorer l'évolution de l'architecture chromosomique des génomes de poisson avec une vue d'ensemble que les études classiques de synténie poisson/tétrapode ne permettent pas d'atteindre, en raison de la forte dégradation de cette synténie. Immédiatement après la duplication du génome, les paires de chromosomes dupliqués partagent tous leurs ohnologues de manière exclusive. Ce motif va être progressivement dégradé au fil de l'évolution par des réarrangements interchromosomiques, si bien qu'un chromosome pourra partager des ohnologues avec plusieurs chromosomes, jusqu'à disparition totale du signal (Figure 17.3).

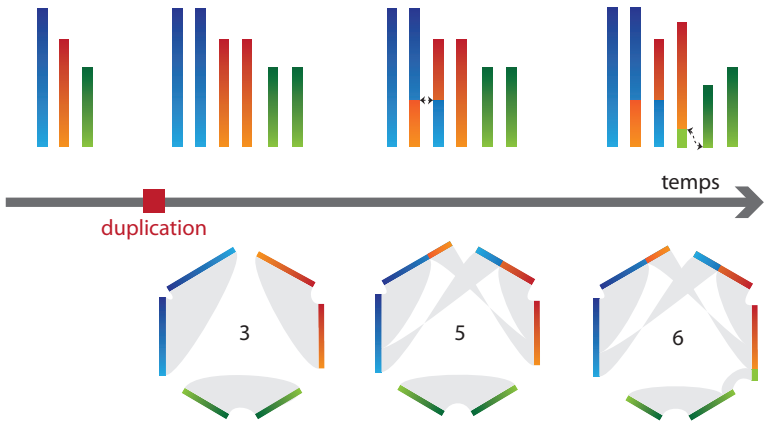


Figure 17.3. Dégradation des relations d'ohnologie entre chromosomes par les réarrangements interchromosomiques. Après la duplication, le nombre initial de chromosomes est doublé et chaque chromosome partage tous ses duplicats avec un seul autre chromosome. Au fur et à mesure des réarrangements, le nombre de paires de chromosomes partageant des gènes ohnologues augmente.

Ainsi, dénombrer le nombre de paires de chromosomes dans chaque génome qui partagent un nombre significatif d'ohnologues renseigne à grande échelle sur le nombre de réarrangements interchromosomiques qui ont perturbé l'organisation chromosomique initiale post-duplication. Le génome ancestral pré-duplication des téléostéens contenait un nombre de chromosomes estimé à 13 (Nakatani et al. 2007), donnant donc 13 paires de chromosomes ohnologues initiaux après la duplication. Dans les génomes du medaka, de l'épinoche et du tétraodon, nous trouvons entre 15 et 20 chromosomes partageant un nombre important de paires d'ohnologues (plus de 1% du total ; Figure 17.4). Peu de réarrangements

interchromosomiques se sont donc produits dans ces lignées depuis la duplication complète. En revanche, le génome du poisson zèbre contient 31 paires de chromosomes contenant chacune plus de 1% des paires d'ohnologues totales, soit presque deux fois plus que les autres poissons ; son profil est manifestement différent des autres poissons (Figure 17.4). Ce profil ne peut s'expliquer que par un taux plus élevé de réarrangements interchromosomiques dans la lignée du poisson zèbre.

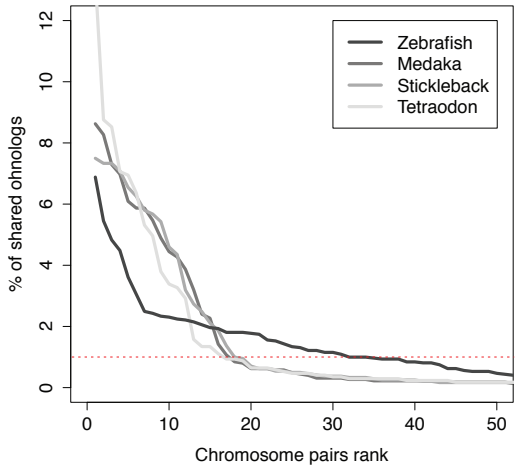


Figure 17.4. Fragmentation des relations d'ohnologie entre chromosomes dans les génomes de poissons téléostéens. Les combinaisons possibles de paires de chromosomes sont triées en fonction de la proportion des paires d'ohnologues qu'elles partagent. La ligne pointillée rouge marque la limite des paires partageant moins de 1% des ohnologues totaux dans le génome.

Chapitre 18. Devenir des gènes suite à la duplication complète du génome

Le génome du poisson zèbre contient 31% de gènes qui sont des ohnologues issus de la duplication complète du génome 3R. Les processus qui mènent à la rétention des deux copies ohnologues ou à l'élimination d'une des deux, et le choix de la copie éliminée le cas échéant, sont encore mal connus dans les génomes de vertébrés. Les principaux exemples de duplications complètes de génomes récentes se trouvent en effet dans les génomes de plantes et de levures, qui ont une dynamique assez différente des génomes de vertébrés. Afin de mieux comprendre le processus de rediploïdisation, nous avons recherché les biais éventuels qui affectent les jeux de gènes retenus ou non en deux copies, et leur localisation dans le génome du poisson zèbre.

18.1. Rétention des ohnologues sur les différents chromosomes du poisson zèbre

En première approche, nous avons examiné la possibilité que des contraintes structurelles ou fonctionnelles à l'échelle chromosomique aient pu introduire des biais ayant affecté le taux de pertes des copies ohnologues. Nous avons donc comparé le contenu en ohnologues et en singletons 3R des différents chromosomes du génome du poisson zèbre.

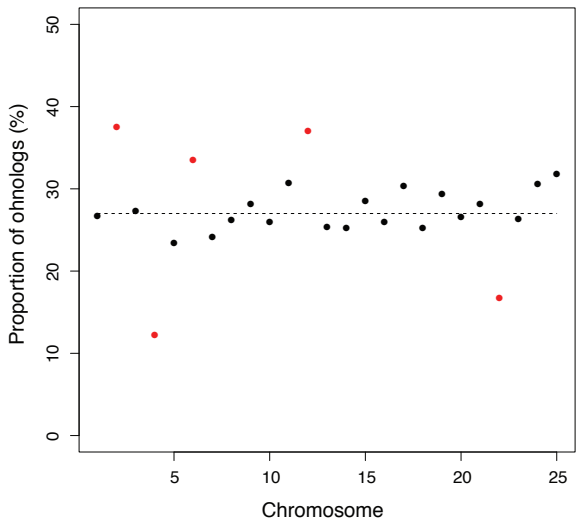


Figure 18.1. Proportion d'ohnologues parmi les gènes de chaque chromosome du génome du poisson zèbre. La ligne pointillée représente la moyenne globale du génome. En rouge sont figurés les chromosomes dont la proportion d'ohnologues est significativement différente de la moyenne après correction de Bonferroni pour tests multiples.

Afin de bien comparer la rétention des ohnologues sur chaque chromosome, on compte le nombre de gènes d'origine ancestrale présents en une copie post-3R ou en deux copies ohnologues dans le génome sur chacun des chromosomes du génome du poisson zèbre, sans tenir compte des éventuelles duplications de gènes postérieures à la duplication 3R. La distribution des ohnologues par chromosome montre que les chromosomes 2, 6 et 12 comptent plus d'ohnologues qu'attendu au hasard (test de proportions, $P < 10^{-3}$ dans tous les cas après correction de Bonferroni ; Figure 18.1) ; les chromosomes 4 et 22, en revanche, contiennent moins d'ohnologues qu'attendu ($P < 10^{-11}$ après correction de Bonferroni). Nous avons testé si le pourcentage de gènes ohnologues corrèle avec des paramètres génomiques pouvant suggérer une explication pour cette rétention différentielle (Figure 18.2).

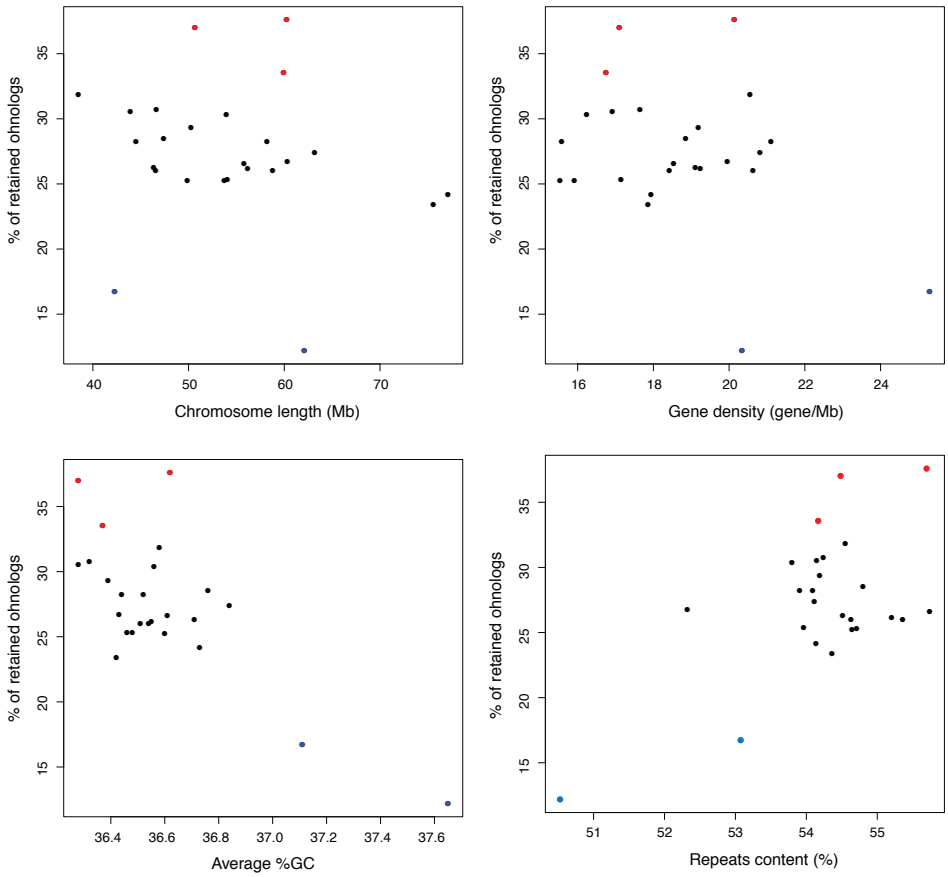


Figure 18.2. Corrélations entre le taux d'ohnologues dans les différents chromosomes du poisson zèbre, et les valeurs moyennes de différentes caractéristiques génomiques dans ces chromosomes : longueur du chromosome, densité en gènes, taux de GC et contenu en séquences répétées. Les chromosomes 2, 6, 12, enrichis en ohnologues, sont représentés en rouge. Les chromosomes 4 et 22, pauvres en ohnologues, sont représentés en bleu.

Aucune corrélation n'existe entre le taux d'ohnologues et la longueur du chromosome ou sa densité en gènes. Le taux de GC moyen et le contenu en séquences répétées sont tous deux corrélés au taux de rétention des ohnologues, avec respectivement un R^2 de 0,54 ($P < 10^{-4}$) et de 0,32 (P = 0,002). Cependant, ces corrélations sont largement dues à l'organisation très particulière du chromosome 4 : ce chromosome contient une large région d'environ 30 Mb avec

un taux très inhabituel de séquences répétées et un fort taux de GC, composée de gènes dupliqués de très nombreuses fois, aux fonctions essentiellement non identifiées et aux relations d'orthologie et de synténie très mal définies par rapport aux autres espèces (73% n'ont pas d'orthologues dans le génome humain). Le chromosome 22 porte également deux régions du même type, plus courtes cependant (environ 3 et 5 Mb). L'origine des gènes de ces régions très dupliquées n'étant pas résolue, ces gènes ne peuvent pas être inclus dans les analyses de DCS et n'ont donc pas d'ohnologues identifiés, bien que certains soient peut-être des copies ohnologues issues de la duplication complète 3R. Ces régions représentent respectivement 52,5% et 20,7% de leurs gènes, ce qui explique pourquoi les chromosomes 4 et 22 présentent un faible taux d'ohnologues par rapport au reste du génome. Lorsque l'on retire ces deux chromosomes outliers de la distribution, les corrélations entre taux de rétention des ohnologues et taux de GC ou de séquences répétées disparaissent, avec respectivement un R^2 de 0,05 (P = 0,15) et de 0,05 (P = 0,16). Dans tous les cas, les variations de taux de GC et de séquences répétées ne peuvent expliquer la rétention préférentielle d'ohnologues sur les chromosomes 2, 6 et 12.

18.2. Rétention des ohnologues dans les espèces voisines

Si l'organisation du génome ne semble pas biaiser la rétention des ohnologues dans le génome du poisson zèbre, nous avons cherché à savoir si c'est la fonction des gènes, et non leur organisation, qui influence la rétention ou la perte des copies ohnologues.

En première approche, nous nous sommes intéressés à la conservation des gènes ohnologues dans d'autres espèces. Nous avons testé si les gènes conservés en deux copies ohnologues dans le génome du poisson zèbre sont plus susceptibles d'avoir des orthologues dans les génomes amniotes que l'ensemble du génome. Nous avons comparé les génomes de deux mammifères (l'homme et la souris), d'un saurien (le poulet) et celui du poisson-zèbre afin de quantifier la proportion de gènes partagés ou espèce-spécifiques dans chaque génome (Figure 18.3.A). La même comparaison a ensuite été faite en prenant uniquement en compte les gènes identifiés comme ohnologues dans le génome du poisson zèbre (Figure 18.3.B). Afin de pouvoir interpréter les résultats, les comptages sont faits en termes de gènes ancestraux, et non en nombre de copies modernes : ainsi, un gène qui existe en une copie chez le poisson zèbre et en deux copies paralogues chez l'humain suite à une duplication dans la lignée des tétrapodes sera compté une seule fois dans l'intersection des deux génomes.

A. B.

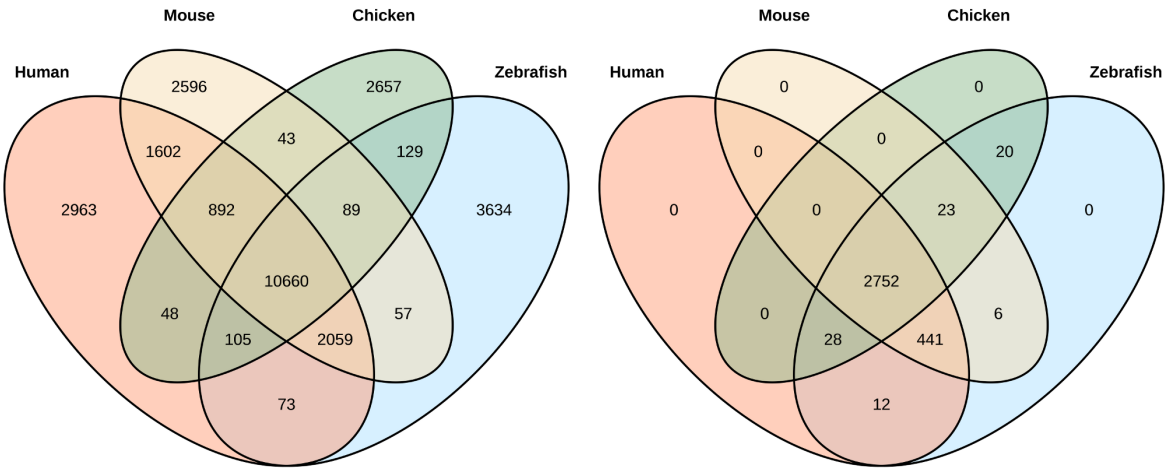


Figure 18.3. Gènes orthologues partagés entre le poisson zèbre et trois amniotes : l'homme, la souris et le poulet. A. Ensemble du génome. B. Gènes ohnologues issus de la duplication 3R dans le génome du poisson zèbre uniquement.

Les proportions d'ohnologues de poisson zèbre ayant des orthologues avec chacune des espèces respectent les proportions du génome global (si l'on excepte le fait que les ohnologues existent forcément dans le génome du poisson zèbre, ce qui explique les compartiments amniote-spécifiques vides). La seule différence se trouve au niveau de la proportion de gènes présents dans toutes les espèces sauf le génome du poulet, pour lesquels les ohnologues sont surreprésentés par rapport à l'ensemble du génome (test de Chi^2 : P = 0,001). Cependant, cette différence est probablement artéfactuelle : en effet, on note que sur l'ensemble du génome, le nombre de gènes inclus dans toutes les espèces sauf le poulet est de 2059, alors que les gènes inclus dans toutes les espèces sauf le poisson zèbre, bien que plus distant évolutivement que le poulet, n'est que de 892. Il est donc probable qu'il s'agisse d'un problème d'annotation ou de résolution des liens d'orthologie au niveau du génome du poulet, plutôt qu'un réel enrichissement en ohnologues parmi les gènes spécifiquement existant dans toutes les espèces sauf le poulet.

18.3. Taux d'évolution des ohnologues

De précédentes études ont montré que les orthologues des gènes ohnologues 3R du tétraodon existant dans les génomes de mammifères ont des taux d'évolution plus faibles que les orthologues des gènes singletons 3R (Brunet et al. 2006). Ces conclusions sont également supportées par les analyses de duplications complètes du génome chez les levures ou les nématodes (Davis and Petrov 2004). Nous avons comparé le rapport des taux de substitutions non-synonymes et synonymes (dN/dS) entre les génomes de l'homme et de la souris pour les gènes orthologues aux ohnologues 3R du génome du poisson zèbre, et pour les orthologues des gènes singletons. En effet, les gènes dupliqués et gardés en deux copies chez le poisson zèbre ont des taux d'évolution significativement plus faibles chez les mammifères que les gènes revenus à

l'état singleton (dN/dS = 0,101 ± 0.100 contre 0.132 ± 0.113 ; test de Wilcoxon : $P < 2.10^{-16}$; Figure 18.4). Les gènes dont les deux copies ont été retenues après la duplication 3R ont une vitesse d'évolution chez les mammifères 23% inférieure à ceux retenus en une seule copie ; ces gènes étaient probablement en moyenne sous plus forte contrainte évolutive dans l'état ancestral également.

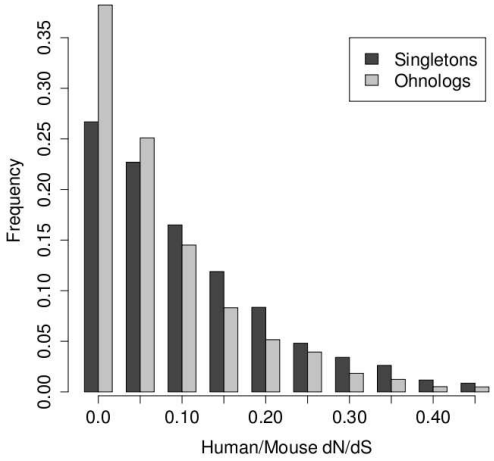


Figure 18.4. Taux d'évolution chez les mammifères des orthologues de gènes retenus en une ou deux copies après la duplication 3R dans le génome du poisson zèbre.

18.4. Catégories de gènes retenus en deux copies

Les deux événements de duplication complète du génome ayant eu lieu à la base de la lignée de vertébrés (Dehal and Boore 2005) sont à l'origine de 20 à 30% des gènes présents actuellement dans le génome humain, notamment les quatre clusters de gènes *Hox* (Nakatani et al. 2007; Makino and McLysaght 2010). Plusieurs études ont montré que ces ohnologues humains sont enrichis en gènes impliqués dans le développement, la transcription, la transduction du signal et en gènes faisant partie de complexes protéiques (Blomme et al. 2006; Brunet et al. 2006; Hufton et al. 2008; Makino and McLysaght 2010). Dans le génome du poisson zèbre, nous retrouvons ces tendances parmi les ohnologues retenus de la duplication 3R : une analyse Gene Ontology montre que ces gènes présentent un enrichissement en termes liés aux fonctions neurales, à la transduction du signal et au développement (Table 18.1). Plus généralement, on observe un enrichissement en protéines localisées au niveau de la membrane plasmique, impliquées dans l'adhésion cellulaire et dans le transport des cations (calcium, sodium, potassium). Ces biais fonctionnels peuvent peut-être être rapprochés des observations sur le taux d'évolution des ohnologues : les gènes impliqués dans le développement et la régulation de la transcription notamment sont généralement sous forte pression de sélection. Ce haut niveau de régulation pourrait en faire de bonnes cibles de rétention sous le modèle de (Force et al. 1999). De plus, l'existence de copies redondantes de gènes hautement régulés pourrait être favorable dans le sens où elles préviendraient et contrebalanceraient l'apparition d'allèles mutés délétères se produisant dans les premiers stades du développement (Clark 1994).

Gene Ontology term	p-value
calcium ion binding	$1{,}5.10^{-15}$
ion channel activity	$1{,}0.10^{-12}$
cation channel activity	$4{,}3.10^{-9}$
ion transmembrane transporter activity	$4{,}3.10^{-9}$
metal ion transembrane transporter activity	$5{,}5.10^{-8}$
GABA-B receptor activity	$4{,}4.10^{-7}$
voltage-gated channel activity	$5{,}9.10^{-7}$
sodium channel activity	$5{,}9.10^{-7}$
voltage-gated ion channel activity	$8{,}1.10^{-7}$
cation transmembrane transporter activity	$5{,}3.10^{-6}$
transcription factor activity	$9{,}4.10^{-6}$
potassium channel activity	$1{,}2.10^{-5}$
G-protein coupled receptor activity	$1{,}4.10^{-5}$
calcium channel activity	$2{,}2.10^{-5}$
sequence-specific DNA binding	$2{,}2.10^{-5}$
cytoskeletal protein binding	$3{,}1.10^{-5}$
cation:chloride symporter activity	$2{,}9.10^{-4}$
transmembrane receptor protein serine/threonine kinase activity	$3{,}3.10^{-4}$
ligand-gated ion channel activity	$4{,}3.10^{-4}$
neurotransmitter receptor activity	$1{,}6.10^{-3}$
adrenoceptor activity	$1{,}8.10^{-3}$
ferric iron binding	$1{,}9.10^{-3}$
protein domain specific binding	$1{,}9.10^{-3}$
extracellular ligand-gated ion channel activity	$3{,}0.10^{-3}$
neuropeptide receptor activity	$3{,}1.10^{-3}$
neuropeptide Y receptor activity	$3{,}1.10^{-3}$
retinoic acid receptor activity	$3{,}8.10^{-3}$
anion transmembrane transporter activity	$3{,}8.10^{-4}$
lipid binding	$4{,}2.10^{-3}$
small GTPase regulator activity	$8{,}1.10^{-3}$
purinergic nucleotide receptor activity	$8{,}6.10^{-3}$
gap junction channel activity	$9{,}9.10^{-3}$

Table 18.1. Termes Gene Ontology enrichis parmi les gènes ohnologues issus de la duplication 3R dans le génome du poisson zèbre (test de Fischer corrigé pour les tests multiples, $P < 0{,}01$). En gris : termes liés aux fonction neurales ; en bleu : termes liés à la transcription ; en vert : termes liés à la transduction du signal ; en orange : termes liés au développement.

18.5. Comparaison des ohnologues retenus aux duplications 2R et 3R

Enfin, nous avons examiné si les mêmes gènes tendent à être retenus en deux copies au fil des duplications complètes du génome s'étant succédées dans l'histoire des vertébrés. Plus spécifiquement, nous avons testé si les gènes ohnologues issus de la duplication 2R avaient plus de chance d'être également retenus comme ohnologues après la duplication 3R. Il est difficile de trouver des traces de la duplication 2R avec certitude dans le génome du poisson zèbre en raison de la très faible conservation de la synténie dans les génomes de poisson, la duplication 3R brouillant encore davantage le signal. Nous avons donc défini une liste d'ohnologues issus de la duplication 2R dans le dernier ancêtre commun de l'homme et du poisson zèbre (Euteleostomi) à partir d'une liste d'ohnologues 2R identifiés dans le génome humain (Makino and McLysaght 2010), en faisant l'hypothèse que la rediploïdisation était achevée ou presque au moment de la bifurcation des deux lignées. Nous avons alors dénombré combien de ces gènes ancestraux issus

de la duplication 2R sont également conservés en deux copies après la duplication 3R dans le génome du poisson zèbre, par rapport aux gènes singletons 2R. L'intersection des deux jeux est remarquablement haute : 33% des ohnologues 2R sont également conservés en deux copies après la duplication 3R, alors que seulement 18% des singletons 2R le sont (test de Chi^2 : $P < 2.10^{-16}$; Figure 18.5). Il semble donc que la probabilité pour un gène d'être retenu lors d'une duplication complète soit en partie une propriété conservée qui, au vu des résultats des deux paragraphes précédents ainsi que de la littérature, est liée à sa fonction et non à l'organisation des gènes dans le génome, qui est très peu contrainte et conservée.

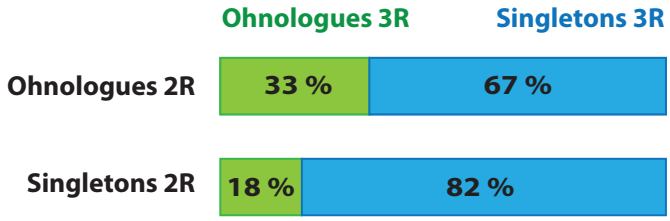


Figure 18.5. Rétention en deux copies des ohnologues et singletons 2R après la duplication 3R dans le génome du poisson zèbre.

Cinquième Partie
Discussion

Chapitre 19. Discussion et Perspectives

Au cours de cette thèse, deux projets ont été menés : le projet principal, portant sur les points de cassure de réarrangements évolutifs, a permis d'établir le premier modèle mathématique complet décrivant la distribution des points de cassure dans les génomes eucaryotes. Ce modèle a permis d'identifier les paramètres sous-tendant la probabilité de cassure et de distinguer entre l'importance des forces moléculaires et des forces d'ordre évolutif qui affectent l'organisation des génomes. En parallèle, le projet d'analyse comparative du génome du poisson zèbre a permis d'étudier l'importance et l'impact de la duplication complète 3R sur l'organisation des génomes de poissons téléostéens, confirmant ou précisant les observations existantes de la littérature. Nous discutons dans ce chapitre la portée de ces résultats, leurs limites et les pistes de recherche ouvertes à l'exploration.

19.1. Modalités des cassures chromosomiques dans les génomes eucaryotes

Les précédentes études menées sur les points de cassures de réarrangements évolutifs se sont principalement basées sur la comparaison des régions de cassure au reste du génome. A ce titre, elles ont rapporté à de nombreuses reprises que les points de cassure sont répartis dans le génome de manière non aléatoire et sont statistiquement associés aux régions riches en gènes, en GC, en duplications segmentales, en éléments transposables, etc. (Murphy et al. 2005; Ma et al. 2006; Larkin et al. 2009; Zhao and Bourque 2009). Ces paramètres étant largement intercorrélés entre eux, il n'a pas été possible jusqu'ici de distinguer entre ceux influençant réellement la probabilité de cassure et les corrélations secondaires qu'ils engendrent. L'hypothèse dominante dans la littérature est que la distribution des réarrangements évolutifs est gouvernée par deux facteurs principaux. Premièrement, les cassures seraient la conséquence de recombinaisons non-homologues entre séquences répétées dans le génome (éléments transposables et duplications segmentales) ; elles se produiraient donc de façon non uniforme en fonction de la densité locale en séquences répétées (Armengol et al. 2003; Schibler et al. 2006; Kemkemer et al. 2009; Zhao and Bourque 2009; Farre et al. 2011; Skinner and Griffin 2012). Deuxièmement, l'architecture des gènes et de leurs éléments de régulation serait sous pression de sélection dans certaines régions du génome, où les réarrangements seraient délétères et contre-sélectionnés ; ces contraintes sélectives joueraient un rôle majeur dans la distribution non-aléatoire des cassures (Kikuta et al. 2007; Hufton et al. 2009; Larkin et al. 2009; Mongin et al. 2009).

Les résultats obtenus au cours de ce travail remettent profondément en question cette vision de l'évolution de l'organisation du génome. En effet, nos observations montrent que, d'une part, les contraintes sur l'organisation du génome sont marginales et ne semblent pas jouer un rôle

majeur dans l'évolution de l'ordre des gènes. D'autre part, la distribution des cassures semble être conditionnée par la probabilité d'occurrence des cassures double-brin plutôt que la probabilité de recombinaison non-homologue due à des séquences répétées.

19.1.1. Les réarrangements, un phénomène essentiellement neutre du point de vue évolutif

A cours de cette thèse, nous avons utilisé une nouvelle méthode d'estimation des caractéristiques du génome ancestral pour décrire et modéliser comment les paramètres locaux (longueur des intergènes, taux de GC, contenu en séquences conservées non-codantes) influencent la probabilité de cassure. Nos résultats démontrent que la probabilité de cassure peut se décrire comme une fonction très simple de la longueur des intergènes, facteur prédicteur principal qui suffit à expliquer à lui seul l'essentiel de la variation du taux de cassure. Le modèle montre que les points de cassure se comportent comme une variable aléatoire de Poisson, dont le taux par intergène est proportionnel à une racine de la longueur de cet intergène. Cette distribution des points de cassure se retrouve dans les génomes des mammifères et des levures, et est donc potentiellement universelle aux génomes eucaryotes.

Il n'est pas nécessaire d'invoquer la présence d'arrangements de gènes sous contrainte fonctionnelle ou la présence d'interactions de régulation en *cis* pour expliquer la fréquence (ou l'absence) des points de cassure dans les différentes régions du génome : la longueur des intergènes seule permet d'expliquer le motif de cassures observées. L'introduction de variables approximant la présence de telles contraintes (densité en éléments conservés non-codants, blocs de régulation génomiques présumés) n'apportent qu'une amélioration marginale au modèle : leur importance est réelle mais très faible en comparaison de celle de la longueur des intergènes. Or, la taille des intergènes eucaryotes évolue de façon aléatoire et neutre, sans intervention de la sélection naturelle (Petrov et al. 2000; Nam and Ellegren 2012) : la taille des génomes est gouvernée par l'équilibre entre insertions (éléments transposables, duplications, etc.) et délétions (en grande partie liées à la recombinaison) dans les séquences non-codantes. Ce type de variations stochastiques de la taille en fonction des taux d'insertions et de délétions résulte théoriquement en une distribution log-normale, ce qui est effectivement observé pour la distribution des tailles d'intergènes dans les génomes eucaryotes. Ces observations suggèrent donc qu'en dehors des séquences géniques qui sont sous forte contrainte de sélection, les régions du génome dont l'organisation est sous contrainte sont rares et ne constituent pas une force majeure de l'évolution des génomes. Les cassures seraient un phénomène neutre du point de vue de l'organisation des gènes, et dont la probabilité de se produire dépend simplement de la quantité d'ADN non-codant présent entre deux gènes.

Par ailleurs, le modèle ne nécessite pas non plus d'invoquer l'existence de régions « fragiles » où les cassures s'accumuleraient plus qu'attendu au hasard. En effet, nous observons que les intergènes courts sont plus souvent cassés, et les grands intergènes réciproquement moins cassés qu'attendu dans la distribution aléatoire telle qu'on l'entend classiquement ; mais cette variation continue correspond bien à une distribution aléatoire dans les intergènes au vu de leur longueur, et ne s'explique pas en termes d'intergènes spécifiquement « fragiles » ou « robustes ».

Bien qu'ils soient à contre-courant de l'hypothèse la plus répandue dans la littérature sur les réarrangements évolutifs, ces résultats sont en accord avec d'autres observations publiées, montrant notamment que les gènes retenus sélectivement sous forme de clusters fonctionnels (gènes coexprimés) existent mais sont rares dans les génomes d'amniotes (Sémon and Duret 2006). Ainsi, dans la grande majorité du génome, l'ordre des gènes ne semble pas avoir de signification fonctionnelle, contrairement à ce qui a pu être avancé par le passé (Hurst et al. 2004).

19.1.2. Structure de la chromatine et probabilité de réarrangement

Bien que la longueur des intergènes explique d'un point de vue mathématique la distribution des points de cassure, la relation entre ces deux variables n'est pas directe, puisque le taux de cassure est proportionnel à une racine de la longueur plutôt qu'à la longueur elle-même. Aucun des paramètres génomiques testés ne permet d'expliquer correctement cet exposant dans l'équation obtenue (paragraphe 13.5) : notamment, les séquences répétées et les duplications segmentales augmentent en densité dans les grands intergènes alors que la densité des cassures, elle, diminue. Le modèle proposé ici n'est pas compatible avec l'hypothèse très répandue dans la littérature qui fait des éléments transposables et autres séquences répétées la cause majeure des réarrangements (Bailey and Eichler 2006; Schibler et al. 2006; Zhao and Bourque 2009). Nous n'excluons pas que des phénomènes de recombinaison non-homologue entre séquences de forte homologie puissent ponctuellement influencer le choix du site partenaire lors d'un réarrangement, ou être impliqués dans un sous-jeu de réarrangements récurrents entre régions particulièrement accessibles aux cassures, mais la distribution des séquences répétées dans les génomes montre qu'elles ne peuvent pas, à elles seules, expliquer la corrélation entre points de cassure et longueur des intergènes observée dans nos données. Les corrélations entre séquences répétées et points de cassure décrits dans la littérature sont plus probablement des corrélations secondaires dues d'une part à la répartition non homogène de ces éléments dans les génomes de mammifères, comme montré au chapitre 14, et d'autre part à des phénomènes de duplications aux bornes des segments réarrangés qui donnent l'illusion que deux éléments homologues ont causé le réarrangement.

A défaut d'avoir trouvé parmi les paramètres génomiques plausibles une explication à nos données, nos résultats suggèrent que c'est bien la longueur des intergènes elle-même qui détermine la probabilité de cassure, mais modulée par la présence de contraintes d'ordre moléculaire. L'explication la plus simple serait que la probabilité qu'une cassure double-brin se produise est liée à l'organisation locale de l'ADN, suivie d'une fixation aléatoire des événements par dérive génétique. Plusieurs études indépendantes ont récemment démontré l'existence d'un lien probable entre ouverture de la chromatine, accessibilité de l'ADN et augmentation des cassures dans différents contextes. Une étude portant sur des lignées de cellules cancéreuses (Lin et al. 2009) a montré que l'ouverture de la chromatine lors de la fixation de facteurs de transcription augmentait la probabilité locale de cassure et donc de réarrangements. Ce résultat pourrait potentiellement expliquer pourquoi certaines translocations sont spécifiques de certains cancers si elles sont causées indirectement par la fixation de facteurs de transcription tissus-spécifiques. Ce phénomène a été décrit comme le « danger de la transcription », postulant

que la transcription est un des talons d'Achille du génome parce que la décompaction de la chromatine nécessaire à l'expression des gènes et à la fixation des facteurs protéiques rend l'ADN accessible et donc vulnérable (Mathas and Misteli 2009). Par ailleurs, très récemment, une autre étude a montré que les variations structurales présentes dans le génome humain et les points de cassure de réarrangements spécifiques de la lignée humaine sont surreprésentés dans les régions hypométhylées du génome des cellules de la lignée germinale (Li et al. 2012) ; or, l'hypométhylation est une marque de chromatine ouverte et décompactée. Les auteurs soulignent que l'hypométhylation est plus corrélée aux points de cassures et aux variations structurales que ne le sont les séquences répétées, pointant vers un effet potentiellement fort de l'état de compaction de l'ADN sur sa mutabilité et, à terme, sur l'évolution du génome.

Ces observations suggèrent une explication simple et directe à nos résultats : le lien non linéaire entre longueur des intergènes et probabilité de cassure que nous observons pourrait refléter la proportion de chromatine ouverte et réellement accessible à la cassure dans les intergènes. Les régions les plus compactées (sans fonction) ne seraient paradoxalement cassées que très rarement, alors que les régions décompactées et accessibles (fonctionnelles) seraient plus susceptibles aux cassures double-brins (Figure 19.1). De manière générale, la proportion de chromatine ouverte diminuerait lorsque la distance entre les gènes augmente, les modules fonctionnels ouverts se faisant plus rares lorsque l'on s'éloigne d'un gène.

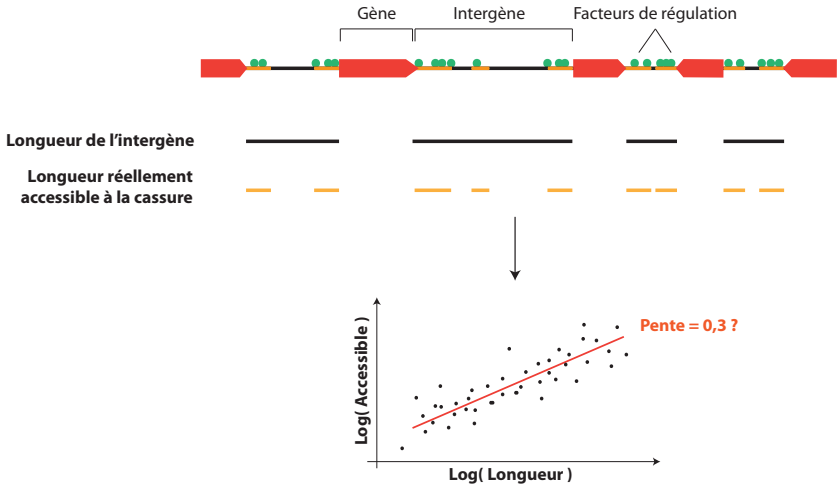


Figure 19.1. Proposition de modèle liant fixation de facteurs de régulation, ouverture de la chromatine et probabilité de cassure.

Cette hypothèse pourrait expliquer simplement à la fois nos résultats et certaines observations surprenantes dans la littérature sur les réarrangements évolutifs : le fait que les régions conservées du génome soient enrichies en gènes tissus-spécifiques (donc rarement exprimés)(Zhao and Bourque 2009), par exemple, ou que les régions denses en gènes, qui sont enrichies en gènes largement exprimés, concentrent plus de cassures qu'attendu au hasard (Murphy et al. 2005; Lemaitre et al. 2009). Cette hypothèse pourrait être testée en comparant les profils d'accessibilité de la chromatine dans les lignées germinales et embryonnaires avec les profils de réarrangements dans différentes lignées. Cette piste est actuellement à l'étude au laboratoire.

19.1.3. La régression de Poisson, un outil pour l'étude des phénomènes mutationnels

Le modèle obtenu au cours de ce travail est basé sur la régression de Poisson, une méthode qui n'a à notre connaissance jamais été utilisée jusqu'ici pour étudier les points de cassure de réarrangements évolutifs. Elle s'y prête pourtant particulièrement bien puisqu'elle traite des événements rares se produisant dans des intervalles : or, les intergènes répondent à cette définition et à cette échelle, la détection des points de cassure est à la fois relativement simple et résolutive. Les intervalles entre gènes sont d'ailleurs souvent utilisés comme « intervalles de base » dans les études de type combinatoire qui cherchent à trouver le scénario de réarrangements le plus probable permettant de transformer un génome en un autre (Bourque et al. 2005; Peng et al. 2006) : les relations d'orthologie étant en général bien résolues, les gènes sont de bons marqueurs pour détecter des modifications dans l'organisation du génome. Pourtant, dans les études précédentes s'intéressant aux caractéristiques des points de cassure, les « régions de cassure » considérées sont typiquement soit de larges fenêtres de taille fixe autour des points de cassure identifiés (Ma et al. 2006; Zhao and Bourque 2009), soit l'espace autour des points de cassure où les séquences entre différents espèces ne peuvent plus être alignées (Lemaitre et al. 2009). Dans les deux cas, la longueur des régions de cassure n'a pas de signification biologique : elle est arbitraire dans le cas des fenêtres de taille choisie, et déterminée par les limites techniques dans l'autre cas. Il est donc nécessaire de moyenner les paramètres d'intérêt (et notamment les longueurs d'intergènes) sur chaque région considérée pour étudier les corrélations entre points de cassure et caractéristiques locales du génome. Nos résultats éclairent deux points : tout d'abord, étudier les réarrangements à la résolution de l'intergène est non seulement pratique mais a un sens du point de vue biologique. Considérer les intergènes comme des unités continues et cassables séparant les gènes incassables semble valide, au vu de l'excellente corrélation qu'on observe en moyenne entre leur longueur et leur probabilité de cassure. Ensuite, moyenner les caractéristiques du génome dans une fenêtre autour des points de cassure part d'une idée raisonnable, en voulant traiter tous les points de cassure de manière homogène, mais se révèle en fait problématique : le taux de cassure augmentant avec $L^{0,28}$ et non L, la corrélation n'est plus détectable lorsqu'on moyenne les longueurs des intergènes sur une fenêtre.

La régression de Poisson est bien adaptée pour traiter les données de points de cassure puisqu'elle peut prendre en compte des intervalles ayant un sens biologique mais des tailles inhomogènes, tenir compte de leurs longueurs et éventuellement, comme dans notre cas, trouver des corrélations plus complexes qu'une simple relation de proportionnalité. Cette méthode de régression multivariée permet de dégager les variables influençant significativement la distribution des points de cassure, d'éliminer certaines corrélations secondaires, et de quantifier l'importance relative de chaque variable. Cette méthode est donc plus puissante que les études classiques des paramètres corrélés aux points de cassure. Une autre méthode de régression multivariée, la régression logistique, a été proposée dans une étude antérieure à la nôtre pour étudier la distribution des points de cassure dans un contexte similaire (Poyatos and Hurst 2007). La régression logistique modélise des variables binaires, ici

le fait qu'un intergène soit conservé ou cassé dans une comparaison de deux génomes. La régression de Poisson est plus adaptée aux données de points de cassure pour plusieurs raisons : premièrement, la régression logistique ne permet pas de prendre en compte des intergènes cassés plusieurs fois dans le contexte de points de cassure identifiés dans plusieurs lignées indépendantes. Deuxièmement, dans le cadre d'une régression logistique, l'hypothèse nulle est que tous les intergènes ont la même probabilité d'être retenus ou cassés : il n'y a pas de supposition a priori sur la longueur des intergènes. Il est pourtant nécessaire d'effectuer une transformation logarithmique sur la longueur pour mettre en évidence la corrélation presque parfaite entre le taux de cassure et la longueur des intergènes que nous observons dans ces travaux. Or, d'un point de vue biologique, cette transformation n'est pas intuitive et pourrait même paraître totalement arbitraire. L'attendu sous une distribution aléatoire est que le taux de cassure soit proportionnel à la longueur des intergènes, suivant une distribution de Poisson classique. La régression modélise le logarithme du taux de cassure : il est alors logique d'utiliser également le logarithme de la longueur comme variable explicative, plutôt que la longueur elle-même. Cette transformation n'est intuitive et rationnelle que dans le contexte d'une régression de Poisson : ceci explique sans doute pourquoi cette excellente corrélation était passée inaperçue jusqu'ici.

Les modélisations de Poisson basées sur la structure en gènes/intergènes du génome pourraient être adaptées à l'étude d'autres phénomènes génomiques rares et ponctuels : elles abordent en effet le problème par un angle différent des comparaisons de composition du génome traité par fenêtres, et permettent dans certains cas de découvrir les corrélations plus complexes que de simples relations de proportionnalité. L'approche que nous proposons ici pourrait être applicable à un grand nombre de situations : par exemple, l'étude de points de cassures dans d'autres contextes que celui évolutif (lignées cancéreuses, variation polymorphique, etc.), l'insertion d'éléments transposables, la distribution des origines de réplication, etc.

19.1.4. Valeur prédictive du modèle

L'un des aspects les plus intéressants du modèle que nous présentons ici réside dans sa valeur prédictive. En effet, il est tentant d'exploiter l'accumulation récente de ressources génomiques pour réaliser une cartographie de la densité en points de cassure et d'en déduire une répartition de la contrainte sélective sur l'organisation des gènes à l'échelle du génome. Nos résultats montrent que la probabilité de cassure de chaque intergène peut se déduire directement de sa longueur, modulée par des petites déviations attribuables à la sélection. Ces petites déviations, difficiles à détecter avec cinq génomes, deviendront plus manifestes avec un plus grand nombre de génomes et donc de points de cassure inclus dans l'analyse. Avec un nombre de cassures suffisant, il sera possible de tester si chaque intergène pris individuellement a été cassé significativement moins souvent qu'attendu au hasard au vu de sa longueur, et si c'est le cas, d'en déduire l'existence d'une contrainte fonctionnelle.

Etant donné que le taux de cassure des grands intergènes (> 100 kb) est d'environ 10% lorsque nous considérons cinq lignées boreoeuthériennes essentiellement indépendantes

cumulant environ 450 Ma d'évolution, nous estimons qu'une centaine de génomes amniotes bien choisis dans l'arbre phylogénétique afin de maximiser la longueur totale des branches devraient donner suffisamment de puissance statistique pour fournir des informations sur une large partie du génome. C'est moins du double du nombre de génomes amniotes déjà séquencés ; ces ressources seront probablement disponibles dans un futur proche. En fournissant le premier modèle décrivant de façon adéquate la distribution des points de cassure dans les génomes de vertébrés, nos résultats proposent un cadre d'étude pour cette cartographie des régions sous contrainte fonctionnelle, notamment dans le génome humain.

19.1.5. Extension aux réarrangements chromosomiques somatiques

L'étude que nous présentons ici ne s'attache qu'à décrire les réarrangements évolutifs, c'est-à-dire fixés, qui se produisent dans les génomes de mammifères. Plusieurs analyses publiées dans la littérature ont montré que les sites fragiles communs (régions instables du génome cassant préférentiellement lorsque les cellules sont soumises à des stress chimiques, irradiations, etc.) et les points de cassure récurrents dans les cellules cancéreuses sont plus corrélés qu'attendu au hasard avec les régions de cassure évolutives (Murphy et al. 2005; Darai-Ramqvist et al. 2008; Fungtammasan et al. 2012). Il serait particulièrement intéressant de tester si les points de cassure qui se produisent au cours de la vie des organismes présentent une distribution similaire à celle des points de cassure évolutifs, ce qui expliquerait pourquoi ils se chevauchent plus souvent qu'attendu. Cela n'est pas absolument évident, cependant, car les points de cassure évolutifs sont par définition transmis à la descendance et se produisent donc dans la lignée germinale, alors que les points de cassure qui se produisent dans les tissus sains ou cancéreux sont somatiques. Si la probabilité de cassure est liée à la transcription et à l'organisation de la chromatine dans le noyau, comme nous le supposons, le patron de cassures est peut-être très différent dans les cellules germinales, par exemple du fait de leur division active (dans le cas des spermatocytes), de leur compaction (spermatozoïdes) ou de leur physiologie particulière (ovocytes et embryon aux premiers stades).

Les jeux de points de cassure actuellement disponibles pour les lignées germinales sont généralement d'origine ou à visée médicale, et sont largement biaisés vers les mutations provoquant un phénotype, qu'ils proviennent d'individus présentant des tumeurs ou une maladie génétique. Les génomes sains, lorsqu'ils sont séquencés, sont rarement assemblés *de novo* mais en général projetés sur le génome de référence, ce qui empêche de détecter des réarrangements qui ne modifient pas le phénotype. Ainsi, il n'existe pas actuellement de ressource qui soit suffisamment abondante, complète et non biaisée pour tester la distribution des cassures chromosomiques telles qu'elles se produisent dans la population générale. Les avancées technologiques récentes sur le séquençage de génomes commencent cependant à fournir des séquences de génomes sains individuels assemblés. Il sera possible prochainement d'identifier des jeux de points de cassure somatiques non biaisés : ainsi, on pourra tester si les cassures évolutives et somatiques se produisent en suivant les mêmes contraintes d'une part, et d'autre part si les déviations par rapport à la densité de points de cassure attendue permettent d'identifier des régions du génome où les modifications de l'organisation des gènes sont associées à des modifications phénotypiques.

19.2. Duplications complètes et modification de la structure du génome

Le but du deuxième projet mené au cours de cette thèse a été de donner un tour d'horizon général de l'organisation et de l'évolution du génome du poisson zèbre dans le cadre de la première publication concernant le séquençage de ce génome de référence, en collaboration avec principalement D. Stemple et K. Howe au Sanger Institute (Royaume-Uni). Cette partie du travail était donc par essence beaucoup plus vaste et descriptive que le projet central sur les points de cassure de réarrangements ; il s'agissait de mener plusieurs analyses classiques de génomique comparative désormais attendues dans les articles portant sur un génome séquencé, ainsi que d'affiner ou infirmer certaines études publiées dans la littérature ayant porté sur des versions préliminaires du génome mises à disposition de la communauté scientifique. A ce titre, cette partie se prête moins à la discussion. Nous avons cependant axé la plupart des analyses autour de l'étude de la duplication 3R qui s'est produite dans la lignée des poissons téléostéens : il s'agit en effet du cas de duplication complète du génome le plus récent dans l'histoire des vertébrés pour lequel plusieurs descendants sont séquencés. Les poissons téléostéens sont donc le meilleur taxon que nous ayons actuellement à disposition pour étudier les duplications complètes de génome chez les vertébrés. La position du poisson zèbre dans l'arbre des téléostéens est particulièrement intéressante par rapport aux autres génomes de poissons séquencés, tous des percomorphes, puisque leur dernier ancêtre commun est plus proche de la duplication complète que l'ancêtre des percomorphes. Dans cette partie nous détaillons et discutons les points clés de cette analyse, notamment en regard des conclusions obtenues sur l'évolution de l'ordre des gènes dans la première partie du travail de thèse, ainsi que les perspectives ouvertes par cet aperçu général de l'organisation du génome du poisson zèbre.

19.2.1. Dégradation de la synténie

La dégradation de la synténie par rapport aux groupes externes est une des conséquences récurrentes des duplications complètes du génome : elle a été observée dans des génomes dupliqués de plantes (Shoemaker et al. 2006), de levures (Gordon et al. 2009), et avait déjà été relevée dans les génomes de poissons (Semon and Wolfe 2007a). Il n'est pas clair à ce jour si cette dégradation de la synténie est liée à la rediploïdisation qui suit la duplication complète, où l'ordre des gènes est modifié par des pertes massives de gènes, ou si la duplication complète entraîne une hausse du taux de réarrangements dans le génome. Cette deuxième hypothèse a été soulevée en particulier pour le phylum des poissons téléostéens (Semon and Wolfe 2007a) ; cependant, les calculs de taux de réarrangements sur lesquelles cette hypothèse se base sont assez indirects, et des dires mêmes des auteurs de l'étude, le résultat obtenu repose en grande partie sur le fort taux de réarrangement inféré pour la lignée du poisson zèbre. Or, comme nous l'avons montré dans ce travail, ce résultat était au moins en partie un artefact lié à la mauvaise qualité de l'assemblage du génome, et non une réalité biologique.

Nos résultats confirment que la synténie est très dégradée chez les poissons téléostéens en général par rapport aux espèces non dupliquées. Mais malgré le fait que cinq génomes de

poissons soient désormais séquencés en profondeur, il reste très difficile de mesurer si cette dégradation de la synténie est d'une part supérieure à celle attendue au vu de la distance phylogénétique entre les poissons et les amniotes, qui est le phylum de comparaison le plus proche actuellement disponible, et d'autre part, si cette dégradation est due à un taux de réarrangements accéléré après la duplication complète du génome. Le génome du coelacanthe présente une meilleure conservation de la synténie avec les génomes amniotes que ceux des poissons téléostéens malgré une distance évolutive à peine plus courte, ce qui va dans le sens d'une dégradation de la synténie liée à la duplication complète dans la lignée des téléostéens. On ne peut cependant pas exclure que ce résultat soit la marque d'une conservation de la synténie particulièrement bonne dans le génome du coelacanthe ; il manque actuellement un génome de téléostéen dont la divergence précède la duplication complète pour trancher définitivement la question de la dégradation de la synténie suite à la duplication 3R. Le génome du lépisosté tacheté (spotted gar, *Lepisostus oculatus*) est actuellement en cours de séquençage et permettra de répondre à cette question ; les premiers résultats obtenus à partir de cartes méiotiques suggèrent que la dégradation de la synténie semble bien être particulièrement prononcée dans les génomes de poissons dupliqués (Amores et al. 2011). Il sera peut-être également possible, si la conservation de la synténie entre le lépisosté et les autres poissons est suffisante, de tester si la duplication complète s'est accompagnée d'une accélération des réarrangements ou si la dégradation de la synténie est essentiellement imputable à la perte de gènes massive.

Par ailleurs, nos résultats attirent l'attention sur l'importance d'utiliser des assemblages de bonne qualité pour l'étude des synténies conservées sur une grande échelle évolutive. En effet, moins la synténie est conservée et plus la qualité de l'assemblage est déterminante lorsque l'on s'intéresse aux blocs conservés et aux taux de réarrangements. Ainsi, nous avons montré que certains résultats obtenus sur les versions préliminaires du génome du poisson zèbre, dont l'ancrage avait été réalisé à partir d'une carte d'hybrides de radiation, ne résistent pas à l'analyse avec la nouvelle version du génome ancrée sur la carte génétique, dont les informations de liaisons sont plus fiables sur de longues distances génomiques. Le poisson zèbre ne semble pas présenter un taux global de réarrangements particulièrement élevé par rapport aux autres poissons, le taux de liaison des gènes par rapport aux amniotes étant similaire aux autres poissons quelque soit la mesure utilisée. En revanche, les relations d'ohnologie entre les différents chromosomes confirment que l'histoire évolutive des chromosomes du poisson zèbre a été plus complexe que chez les autres poissons, avec un nombre de réarrangements interchromosomiques plus important.

19.2.2. Rétention préférentielle des gènes au fil des duplications successives dans l'histoire des vertébrés

Les mécanismes gouvernant la rediploïdisation qui suit une duplication complète du génome sont encore largement débattus dans la littérature, notamment en ce qui concerne le choix des gènes retenus en une ou deux copies. Chez les vertébrés, la majorité des études se concentrent sur les duplications 1R et 2R qui se sont produites à la base de l'arbre des vertébrés, et dans une moindre mesure sur la duplication 3R des poissons téléostéens. De précédents travaux ont

montré que les gènes retenus en deux copies ohnologues après une duplication complète ne correspondent pas à un échantillonnage au hasard. Les ohnologues dans les génomes de vertébrés sont enrichis en gènes codant pour des facteurs de transcription et/ou impliqués dans le développement, en gènes impliqués dans la transduction du signal, en gènes contenant un grand nombre de modules protéiques différents, et en gènes sous pression de sélection relativement forte avant la duplication (Blomme et al. 2006; Brunet et al. 2006; Hufton et al. 2009; Makino and McLysaght 2010; Satake et al. 2012). Ces résultats peuvent s'expliquer dans le cadre du modèle désormais classique de Force et al. (1999) qui propose que les duplications de gènes sont des sources d'innovation relâchant la contrainte sélective sur les gènes et leur permettant de se subfonctionnaliser ou d'acquérir de nouvelles fonctions par mutations, notamment pour les gènes sous forte pression de sélection pour lesquels l'innovation par mutations est habituellement très lente. Par ailleurs, les ohnologues 2R codent plus souvent qu'attendu pour des protéines impliquées dans des complexes protéiques (Hufton et al. 2009; Makino and McLysaght 2010). Cette rétention préférentielle découlerait d'un équilibre dans les taux d'expression des différents gènes du complexe : la perte d'un des gènes déséquilibrerait la stœchiométrie du complexe et serait contre-sélectionnée, favorisant la rétention de l'ensemble des duplicats pour les complexes protéiques. Les gènes dupliqués lors d'événements locaux (duplications segmentales, etc.) ont en effet un profil de rétention assez différent et sont notamment appauvris en gènes impliqués dans des complexes, sans doute pour les même raisons de stœchiométrie des différentes protéines d'un complexe (Makino and McLysaght 2010; Satake et al. 2012).

Nos résultats confirment les observations précédentes de la littérature ; les ohnologues retenus à la duplication 3R dans le génome du poisson zèbre présentent un enrichissement en gènes impliqués dans la transduction du signal, la régulation de la transcription et le développement, comme précédemment rapporté pour les ohnologues 2R. Les ohnologues 3R sont également des gènes qui, en moyenne, évoluaient moins vite que les autres avant la duplication. En raison du temps limité imparti à ces analyses, nous n'avons pas pu tester si les ohnologues 3R contiennent plus de modules protéiques que les singletons ou encore s'ils sont davantage impliqués dans des complexes protéiques, comme les ohnologues 2R. Cependant, nous avons montré que les gènes conservés en deux copies après la duplication 2R ont été deux fois plus souvent retenus également en deux copies à la duplication 3R que les autres. Ainsi, il semble que ce soient les mêmes caractéristiques fonctionnelles qui gouvernent la rétention des gènes en deux copies au fil de l'évolution, conduisant à l'expansion de certaines familles de gènes au fur et à mesure des duplications successives dans l'histoire des vertébrés. En revanche, la position des gènes dans le génome ou la composition locale du génome ne semble pas avoir d'influence sur la rétention ou non d'un gène en deux copies. Ce résultat est cohérent avec ceux obtenus dans le projet principal de la thèse, où nous montrons que l'organisation des gènes semble très peu contrainte dans les génomes de vertébrés: il est logique que l'organisation des gènes n'influence que peu la rediploïdisation qui suit une duplication complète du génome, puisque son rôle fonctionnel est sans doute limité.

19.2.3. Rétention différentielle des singletons entre le poisson zèbre et les percomorphes

L'une des perspectives les plus intéressantes ouvertes par le séquençage du génome du poisson zèbre est la possibilité d'examiner le processus de rediploïdisation du génome après une duplication complète, notamment la vitesse à laquelle elle se produit. En effet, nous disposons désormais de deux lignées ayant divergé après un temps relativement court après la duplication (estimé à 45 Ma (Kasahara et al. 2007)) : le poisson zèbre d'une part, et les percomorphes. Cependant, cette analyse pose un problème majeur : lorsque l'un des deux gènes ohnologues est perdu dans la lignée du poisson zèbre et l'autre copie est perdue dans la lignée des percomorphes (rétention différentielle des copies dans les deux lignées), l'arbre des gènes a exactement la même topologie que si une seule copie avait été perdue avant la séparation des deux lignées, ou indépendamment dans chaque lignée (Figure 19.2). TreeBeST, le logiciel utilisé dans notre analyse pour reconstruire les arbres, réconcilie l'arbre des gènes obtenu à partir de l'alignement de séquences avec l'arbre des espèces en introduisant des duplications de manière parcimonieuse : dans ce cas, il n'introduira pas de duplication au niveau de la duplication 3R, puisque l'arbre des gènes reproduit parfaitement l'arbre des espèces sans avoir besoin de recourir à des duplications. Ainsi, dans le cas d'une rétention différentielle des copies, les deux gènes paralogues seront en réalité annotés comme des orthologues. De manière plus générale, il arrive également que l'arbre contienne effectivement un nœud de duplication correspondant à la duplication 3R mais que les gènes soient mal classés en dessous du nœud (les deux groupes de gènes ne correspondent pas à des orthologues mais à un mélange d'orthologues et de paralogues selon les espèces).

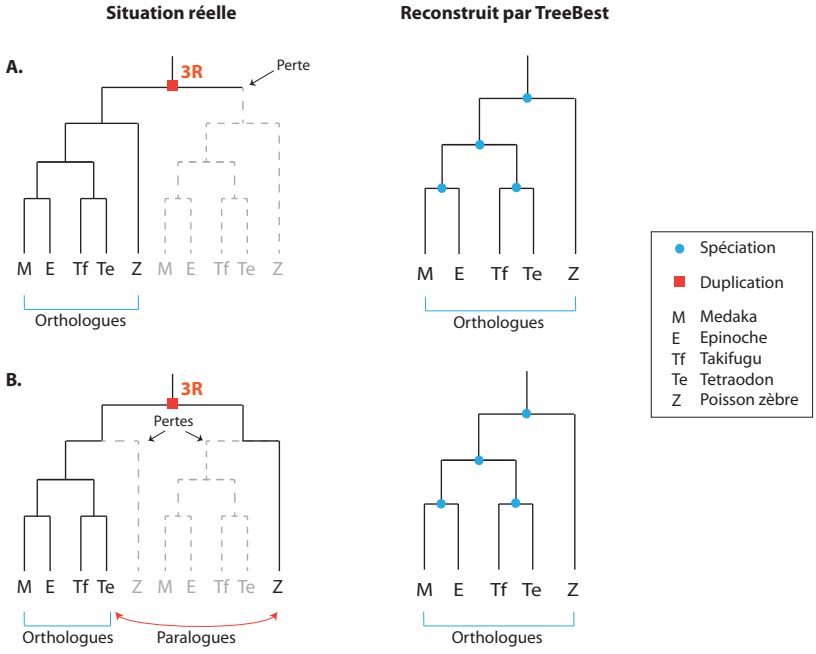


Figure 19.2. Comparaison des arbres de gènes reconstruits par TreeBeST dans le cas de la perte d'un paralogue avant la divergence des lignées (A) ou d'une perte différentielle des deux paralogues entre le poisson zèbre et les autres téléostéens (B).

Afin d'examiner à quel point la rétention différentielle est importante dans les génomes de poissons téléostéens et s'il est possible de déceler des biais dans la rétention des copies, nous avons cherché à corriger les arbres de gènes afin d'identifier correctement paralogues et orthologues, en nous basant sur la synténie double-conservée par rapport aux amniotes. L'idée d'utiliser la synténie pour améliorer les arbres de gènes n'est pas nouvelle : elle a d'ailleurs fait partie du pipeline de construction des arbres de gènes d'Ensembl jusqu'à la version 41, et est implémentée dans le programme de reconstruction d'arbres SYNERGY (Wapinski et al. 2007a). L'utilisation de la synténie repose en général sur l'idée que des gènes orthologues entre deux génomes se trouvent dans un environnement composé de gènes également orthologues entre eux, alors que des paralogues auront des environnements différents (Swidan et al. 2006; Vilella et al. 2009). Dans notre cas, le problème est un peu plus compliqué puisqu'il s'agit de considérer les blocs de synténie double-conservée, soit deux blocs de synténie interdigités, pour différencier les orthologues vrais des paralogues. Le principe consiste à réorganiser l'ensemble des gènes descendant d'un même gène ancestral avant la duplication, en utilisant leur environnement local de synténie dans chaque génome afin de retrouver les orthologues descendant du même gène post-duplication, puis de reconstruire la topologie correcte de l'arbre en suivant l'arbre des espèces en introduisant un nœud de duplication si nécessaire (Figure 19.3).

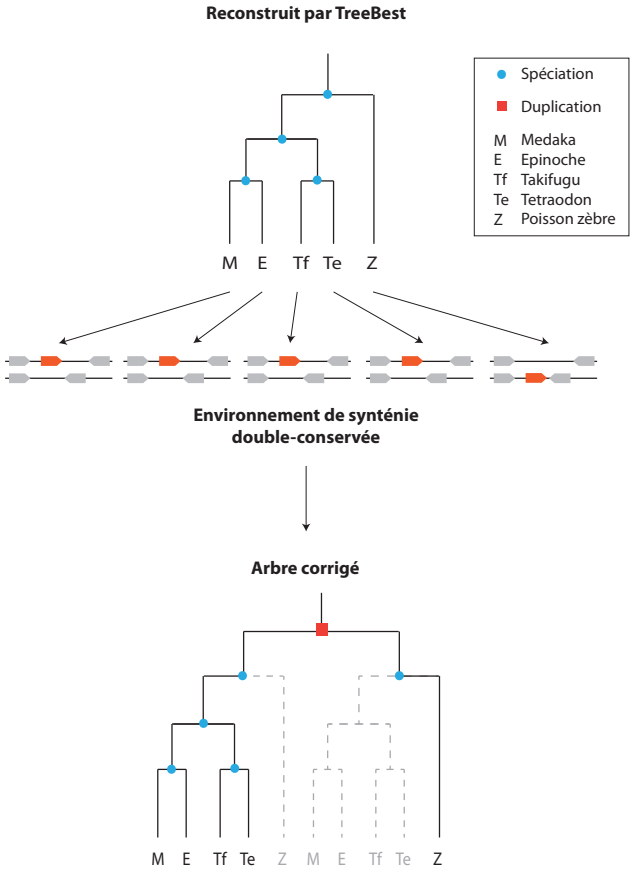


Figure 19.3. Correction de la topologie des arbres de gènes en utilisant les informations de synténie double-conservée.

Cependant, si le principe est assez direct en apparence, nous nous sommes rapidement rendus compte que le problème devient très complexe lorsque l'arbre de gènes diffère du cas relativement simple où chaque espèce contient un ou deux duplicats 3R. Les pertes de gènes et les duplications postérieures à la duplication complète brouillent le signal des blocs de synténie double-conservée, si bien que certains gènes ne peuvent pas être replacés dans l'arbre en se basant sur la synténie seule. Il faut alors se fier à l'information de proximité des séquences (l'arbre initial basé sur l'alignement des séquences) plutôt qu'à l'information de synténie : de fait, la méthode n'est d'une part pas toujours capable d'améliorer l'arbre, et d'autre part mène parfois à des incohérences. De plus, la méthode est en partie circulaire et repose sur le présupposé qu'une grande partie des arbres de gènes est correcte, sans quoi utiliser les relations d'orthologie des gènes voisins pour éditer l'arbre d'un gène risque d'introduire de nouvelles erreurs. Ce travail a donc été entamé mais nécessite encore de nombreuses optimisations avant de pouvoir être réellement utilisé ; nous espérons que l'utilisation de la synténie double-conservée permettra à terme d'obtenir des arbres de gènes post-duplication plus fiables, et d'examiner les contraintes qui s'exercent sur le processus de rediploïdisation, s'il en existe.

19.2.4. Analyse de la duplication 4R des salmonidés

La duplication 3R n'est pas la seule duplication complète du génome dans l'arbre des poissons téléostéens : il existe en effet une autre duplication, dite duplication 4R, à la base du phylum des salmonidés, qui comprend notamment les saumons, les truites, les ombres et les corégones. Cette duplication est relativement récente puisqu'elle est datée entre 25 et 100 Ma selon les estimations (Lien et al. 2011; Guyomard et al. 2012). Le phylum des salmonidés est donc doublement intéressant pour la biologie puisqu'il contient des espèces d'intérêt agronomique majeur (notamment la truite arc-en-ciel et le saumon atlantique) et au génome récemment dupliqué. Ces génomes font logiquement l'objet de nombreuses attentions depuis quelques années, avec plusieurs projets de séquençage en cours ou en vue à court terme (Davidson et al. 2010; Palti et al. 2011). Un effort conjoint au niveau national a été fait dans le cadre du projet ANR Genotrout pour séquencer et analyser le génome de la truite arc-en-ciel (*Oncorhynchus mykiss*), projet dont le laboratoire est partenaire.

Les premiers résultats de ce projet sont très encourageants. Le génome de la truite contient plus de 46000 gènes annotés et ayant des orthologues dans les autres espèces de poissons, contre environ 25000 gènes dans un génome de poisson typique n'ayant pas subi la duplication 4R ; de plus, une grande partie des gènes sont présents en deux copies paralogues pour un gène dans chaque génome non-4R. Il semble donc que la rediploïdisation ne soit qu'amorcée dans ce génome, fournissant une occasion rare d'observer le phénomène alors qu'il est en cours et non essentiellement achevé comme dans la plupart des cas de duplications complètes connus. Bien que le génome ne soit pas totalement assemblé, on peut d'ores et déjà détecter dans le génome de la truite de longs blocs de synténie double-conservée par rapports aux autres téléostéens, et retracer l'origine ancestrale de chaque chromosome, qui sont presque tous une fusion de deux duplicats de chromosomes ancestraux. Cette nouvelle ressource devrait permettre d'examiner plus en détail les conséquences immédiates d'une duplication sur le génome, notamment en ce

qui concerne deux aspects clés débattus dans la littérature : augmentation des taux de réarrangements, et déroulement temporel de la rediploïdisation.

Table des figures

Liste des tableaux

Références

Adams GN, Schmaier AH. 2012. The Williams-Beuren Syndrome-a window into genetic variants leading to the development of cardiovascular disease. *PLoS genetics* **8**(2): e1002479.

Adams KL, Wendel JF. 2005. Polyploidy and genome evolution in plants. *Current opinion in plant biology* **8**(2): 135-141.

Alekseyev MA, Pevzner PA. 2007. Are there rearrangement hotspots in the human genome? *PLoS computational biology* **3**(11): e209.

-. 2009. Breakpoint graphs and ancestral genome reconstructions. *Genome research* **19**(5): 943-957.

Amores A, Catchen J, Ferrara A, Fontenot Q, Postlethwait JH. 2011. Genome evolution and meiotic maps by massively parallel DNA sequencing: spotted gar, an outgroup for the teleost genome duplication. *Genetics* **188**(4): 799-808.

Amores A, Force A, Yan YL, Joly L, Amemiya C, Fritz A, Ho RK, Langeland J, Prince V, Wang YL et al. 1998. Zebrafish hox clusters and vertebrate genome evolution. *Science* **282**(5394): 1711-1714.

Armengol L, Pujana MA, Cheung J, Scherer SW, 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. *Human molecular genetics* **12**(17): 2201-2208.

Ashkenazy H, Penn O, Doron-Faigenboim A, Cohen O, Cannarozzi G, Zomer O, Pupko T. 2012. FastML: a web server for probabilistic reconstruction of ancestral sequences. *Nucleic acids research* **40**(Web Server issue): W580-584.

Aury JM, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, Segurens B, Daubin V, Anthouard V, Aiach N et al. 2006. Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. *Nature* **444**(7116): 171-178.

Ayala FJ, Coluzzi M. 2005. Chromosome speciation: humans, Drosophila, and mosquitoes. *Proceedings of the National Academy of Sciences of the United States of America* **102 Suppl 1**: 6535-6542.

Bailey JA, Baertsch R, Kent WJ, Haussler D, Eichler EE. 2004. Hotspots of mammalian chromosomal evolution. *Genome biology* **5**(4): R23.

Bailey JA, Eichler EE. 2006. Primate segmental duplications: crucibles of evolution, diversity and disease. *Nature reviews Genetics* **7**(7): 552-564.

Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE. 2002. Recent segmental duplications in the human genome. *Science* **297**(5583): 1003-1007.

Bailey JA, Yavor AM, Massa HF, Trask BJ, Eichler EE. 2001. Segmental duplications: organization and impact within the current human genome project assembly. *Genome research* **11**(6): 1005-1017.

Baptista J, Mercer C, Prigmore E, Gribble SM, Carter NP, Maloney V, Thomas NS, Jacobs PA, Crolla JA. 2008. Breakpoint mapping and array CGH in translocations: comparison of a phenotypically normal and an abnormal cohort. *American journal of human genetics* **82**(4): 927-936.

Basolo AL. 1994. The dynamics of Fisherian sex-ratio evolution: theoratical and experimental investigations. *Am Nat* **144**(3): 473-490.

Baudet C, Lemaitre C, Dias Z, Gautier C, Tannier E, Sagot MF. 2010. Cassis: detection of genomic rearrangement breakpoints. *Bioinformatics* **26**(15): 1897-1898.

Becker TS, Lenhard B. 2007. The random vs. fragile breakage model of chromosome evolution: a matter of resolution. *Molecular Genetics and Genomics* **in press.** .

Benko S, Fantes JA, Amiel J, Kleinjan DJ, Thomas S, Ramsay J, Jamshidi N, Essafi A, Heaney S, Gordon CT et al. 2009. Highly conserved non-coding elements on either side of SOX9 associated with Pierre Robin sequence. *Nature genetics* **41**(3): 359-364.

Bento M, Pereira HS, Rocheta M, Gustafson P, Viegas W, Silva M. 2008. Polyploidization as a retraction force in plant genome evolution: sequence rearrangements in triticale. *PloS one* **3**(1): e1402.

Bernardi G. 2000. Isochores and the evolutionary genomics of vertebrates. *Gene* **241**(1): 3-17.

Blanc G, Hokamp K, Wolfe KH. 2003. A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome. *Genome research* **13**(2): 137-144.

Blanc G, Wolfe KH. 2004. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *The Plant cell* **16**(7): 1667-1678.

Blanchette M, Green ED, Miller W, Haussler D. 2004. Reconstructing large regions of an ancestral mammalian genome in silico. *Genome research* **14**(12): 2412-2423.

Blomme T, Vandepoele K, De Bodt S, Simillion C, Maere S, Van de Peer Y. 2006. The gain and loss of genes during 600 million years of vertebrate evolution. *Genome biology* **7**(5): R43.

Bloom SE. 1972. Chromosome abnormalities in chicken (Gallus domesticus) embryos: types, frequencies and phenotypic effects. *Chromosoma* **37**(3): 309-326.

Bogart JP, Elinson RP, Licht LE. 1989. Temperature and sperm incorporation in polyploid salamanders. *Science* **246**(4933): 1032-1034.

Bourque G, Pevzner PA. 2002. Genome-scale evolution: reconstructing gene orders in the ancestral species. *Genome research* **12**(1): 26-36.

Bourque G, Pevzner PA, Tesler G. 2004. Reconstructing the genomic architecture of ancestral mammals: lessons from human, mouse, and rat genomes. *Genome research* **14**(4): 507-516.

Bourque G, Tesler G, Pevzner PA. 2006. The convergence of cytogenetics and rearrangement-based models for ancestral genome reconstruction. *Genome research* **16**(3): 311-313.

Bourque G, Zdobnov EM, Bork P, Pevzner PA, Tesler G. 2005. Comparative architectures of mammalian and chicken genomes reveal highly variable rates of genomic rearrangements across different lineages. *Genome research* **15**(1): 98-110.

Brunet FG, Roest Crollius H, Paris M, Aury JM, Gibert P, Jaillon O, Laudet V, Robinson-Rechavi M. 2006. Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Molecular biology and evolution* **23**(9): 1808-1816.

Byrne KP, Wolfe KH. 2005. The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome research* **15**(10): 1456-1461.

-. 2007. Consistent patterns of rate asymmetry and gene loss indicate widespread neofunctionalization of yeast genes after whole-genome duplication. *Genetics* **175**(3): 1341-1350.

Carbone L, Harris RA, Vessere GM, Mootnick AR, Humphray S, Rogers J, Kim SK, Wall JD, Martin D, Jurka J et al. 2009. Evolutionary breakpoints in the gibbon suggest association between cytosine methylation and karyotype evolution. *PLoS genetics* **5**(6): e1000538.

Carvalho CM, Zhang F, Liu P, Patel A, Sahoo T, Bacino CA, Shaw C, Peacock S, Pursley A, Tavyev YJ et al. 2009. Complex rearrangements in patients with duplications of MECP2 can occur by fork stalling and template switching. *Human molecular genetics* **18**(12): 2188-2203.

Chain FJ, Dushoff J, Evans BJ. 2011. The odds of duplicate gene persistence after polyploidization. *BMC genomics* **12**: 599.

Chauve C, Tannier E. 2008. A methodological framework for the reconstruction of contiguous regions of ancestral genomes and its application to Mammalian genomes. *PLoS Comp Biol* **4**(11): e1000234.

Chen ZJ, Ni Z. 2006. Mechanisms of genomic rearrangements and gene expression changes in plant polyploids. *BioEssays : news and reviews in molecular, cellular and developmental biology* **28**(3): 240-252.

Chiang C, Jacobsen JC, Ernst C, Hanscom C, Heilbut A, Blumenthal I, Mills RE, Kirby A, Lindgren AM, Rudiger SR et al. 2012. Complex reorganization and predominant non-homologous repair following chromosomal breakage in karyotypically balanced germline rearrangements and transgenic integration. *Nature genetics* **44**(4): 390-397.

Chowdhary BP, Fronicke L, Gustavsson I, Scherthan H. 1996. Comparative analysis of the cattle and human genomes: detection of ZOO-FISH and gene mapping-based chromosomal homologies. *Mammalian genome : official journal of the International Mammalian Genome Society* **7**(4): 297-302.

Chowdhary BP, Raudsepp T, Fronicke L, Scherthan H. 1998. Emerging patterns of comparative genome organization in some mammalian species as revealed by Zoo-FISH. *Genome research* **8**(6): 577-589.

Christoffels A, Koh EG, Chia JM, Brenner S, Aparicio S, Venkatesh B. 2004. Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes. *Molecular biology and evolution* **21**(6): 1146-1151.

Clark AG. 1994. Invasion and maintenance of a gene duplication. *Proceedings of the National Academy of Sciences of the United States of America* **91**(8): 2950-2954.

Comai L. 2005. The advantages and disadvantages of being polyploid. *Nature reviews Genetics* **6**(11): 836-846.

Conant GC, Wolfe KH. 2008. Turning a hobby into a job: how duplicated genes find new functions. *Nature reviews Genetics* **9**(12): 938-950.

Cooke J, Nowak MA, Boerlijst M, Maynard-Smith J. 1997. Evolutionary origins and maintenance of redundant gene expression during metazoan development. *Trends in genetics : TIG* **13**(9): 360-364.

Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A. 2005. Distribution and intensity of constraint in mammalian genomic sequence. *Genome research* **15**(7): 901-913.

Copeland NG, Jenkins NA, Gilbert DJ, Eppig JT, Maltais LJ, Miller JC, Dietrich WF, Weaver A, Lincoln SE, Steen RG et al. 1993. A genetic linkage map of the mouse: current applications and future prospects. *Science* **262**(5130): 57-66.

Crow KD, Wagner GP. 2006. Proceedings of the SMBE Tri-National Young Investigators' Workshop 2005. What is the role of genome duplication in the evolution of complexity and diversity? *Molecular biology and evolution* **23**(5): 887-892.

Cui L, Wall PK, Leebens-Mack JH, Lindsay BG, Soltis DE, Doyle JJ, Soltis PS, Carlson JE, Arumuganathan K, Barakat A et al. 2006. Widespread genome duplications throughout the history of flowering plants. *Genome research* **16**(6): 738-749.

Darai-Ramqvist E, Sandlund A, Muller S, Klein G, Imreh S, Kost-Alimova M. 2008. Segmental duplications and evolutionary plasticity at tumor chromosome break-prone regions. *Genome research* **18**(3): 370-379.

Davidson WS, Koop BF, Jones SJ, Iturra P, Vidal R, Maass A, Jonassen I, Lien S, Omholt SW. 2010. Sequencing the genome of the Atlantic salmon (Salmo salar). *Genome biology* **11**(9): 403.

Davis JC, Petrov DA. 2004. Preferential duplication of conserved proteins in eukaryotic genomes. *PLoS biology* **2**(3): E55.

Deaton AM, Bird A. 2011. CpG islands and the regulation of transcription. *Genes & development* **25**(10): 1010-1022.

Dehal P, Boore JL. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS biology* **3**(10): e314.

Di Rienzi SC, Collingwood D, Raghuraman MK, Brewer BJ. 2009. Fragile genomic sites are associated with origins of replication. *Genome biology and evolution* **1**: 350-363.

Dong X, Fredman D, Lenhard B. 2009. Synorth: exploring the evolution of synteny and long-range regulatory interactions in vertebrate genomes. *Genome biology* **10**(8): R86.

Drillon G, Fischer G. 2011. Comparative study on synteny between yeasts and vertebrates. *Comptes rendus biologies* **334**(8-9): 629-638.

Durand D. 2003. Vertebrate evolution: doubling and shuffling with a full deck. *Trends in genetics : TIG* **19**(1): 2-5.

Dutrillaux B. 1979. Chromosomal evolution in primates: tentative phylogeny from Microcebus murinus (Prosimian) to man. *Human genetics* **48**(3): 251-314.

Dutrillaux B, Couturier J. 1983. The ancestral karyotype of Carnivora: comparison with that of platyrrhine monkeys. *Cytogenetics and cell genetics* **35**(3): 200-208.

Ellstrand NC, Schierenbeck KA. 2000. Hybridization as a stimulus for the evolution of invasiveness in plants? *Proceedings of the National Academy of Sciences of the United States of America* **97**(13): 7043-7050.

Engstrom PG, Fredman D, Lenhard B. 2008. Ancora: a web resource for exploring highly conserved noncoding elements and their association with developmental regulatory genes. *Genome biology* **9**(2): R34.

Engstrom PG, Ho Sui SJ, Drivenes O, Becker TS, Lenhard B. 2007. Genomic regulatory blocks underlie extensive microsynteny conservation in insects. *Genome research* **17**(12): 1898-1908.

Eppig JT, Nadeau JH. 1995. Comparative maps: the mammalian jigsaw puzzle. *Current opinion in genetics & development* **5**(6): 709-716.

Faria R, Navarro A. 2010. Chromosomal speciation revisited: rearranging theory with pieces of evidence. *Trends in ecology & evolution* **25**(11): 660-669.

Farre M, Bosch M, Lopez-Giraldez F, Ponsa M, Ruiz-Herrera A. 2011. Assessing the role of tandem repeats in shaping the genomic architecture of great apes. *PloS one* **6**(11): e27239.

Fawcett JA, Maere S, Van de Peer Y. 2009. Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. *Proceedings of the National Academy of Sciences of the United States of America* **106**(14): 5737-5742.

Feldman M, Levy AA. 2005. Allopolyploidy--a shaping force in the evolution of wheat genomes. *Cytogenetic and genome research* **109**(1-3): 250-258.

Feuk L, MacDonald JR, Tang T, Carson AR, Li M, Rao G, Khaja R, Scherer SW. 2005. Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies. *PLoS genetics* **1**(4): e56.

Fischer G, Rocha EP, Brunet F, Vergassola M, Dujon B. 2006. Highly variable rates of genome rearrangements between hemiascomycetous yeast lineages. *PLoS genetics* **2**(3): e32.

Flicek P, Aken BL, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Coates G, Fairley S et al. 2010. Ensembl's 10th year. *Nucleic acids research* **38**(Database issue): D557-562.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**(4): 1531-1545.

Franchini P, Colangelo P, Solano E, Capanna E, Verheyen E, Castiglia R. 2010. Reduced gene flow at pericentromeric loci in a hybrid zone involving chromosomal races of the house mouse Mus musculus domesticus. *Evolution; international journal of organic evolution* **64**(7): 2020-2032.

Frazer KA Ballinger DG Cox DR Hinds DA Stuve LL Gibbs RA Belmont JW Boudreau A Hardenbol P Leal SM et al. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**(7164): 851-861.

Froenicke L, Caldes MG, Graphodatsky A, Muller S, Lyons LA, Robinson TJ, Volleth M, Yang F, Wienberg J. 2006. Are molecular cytogenetics and bioinformatics suggesting diverging models of ancestral mammalian genomes? *Genome research* **16**(3): 306-310.

Fronicke L, Chowdhary BP, Scherthan H, Gustavsson I. 1996. A comparative map of the porcine and human genomes demonstrates ZOO-FISH and gene mapping-based chromosomal homologies. *Mammalian genome : official journal of the International Mammalian Genome Society* **7**(4): 285-290.

Fronicke L, Wienberg J, Stone G, Adams L, Stanyon R. 2003. Towards the delineation of the ancestral eutherian genome organization: comparative genome maps of human and the African elephant (Loxodonta africana) generated by chromosome painting. *Proc Biol Sci* **270**(1522): 1331-1340.

Fungtammasan A, Walsh E, Chiaromonte F, Eckert KA, Makova KD. 2012. A genome-wide analysis of common fragile sites: What features determine chromosomal instability in the human genome? *Genome research*.

Galtier N, Duret L. 2007. Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. *Trends in genetics : TIG* **23**(6): 273-277.

Garcia-Fernandez J, Holland PW. 1994. Archetypal organization of the amphioxus Hox gene cluster. *Nature* **370**(6490): 563-566.

Gardiner-Garden M, Frommer M. 1987. CpG islands in vertebrate genomes. *Journal of molecular biology* **196**(2): 261-282.

Gerstein AC, Chun HJ, Grant A, Otto SP. 2006. Genomic convergence toward diploidy in Saccharomyces cerevisiae. *PLoS genetics* **2**(9): e145.

Gibson TJ, Spring J. 1998. Genetic redundancy in vertebrates: polyploidy and persistence of genes encoding multidomain proteins. *Trends in genetics : TIG* **14**(2): 46-49; discussion 49-50.

Giglio S, Calvari V, Gregato G, Gimelli G, Camanini S, Giorda R, Ragusa A, Guerneri S, Selicorni A, Stumm M et al. 2002. Heterozygous submicroscopic inversions involving olfactory

receptor-gene clusters mediate the recurrent t(4;8)(p16;p23) translocation. *American journal of human genetics* **71**(2): 276-285.

Girirajan S, Chen L, Graves T, Marques-Bonet T, Ventura M, Fronick C, Fulton L, Rocchi M, Fulton RS, Wilson RK et al. 2009. Sequencing human-gibbon breakpoints of synteny reveals mosaic new insertions at rearrangement sites. *Genome research* **19**(2): 178-190.

Glas R, Marshall Graves JA, Toder R, Ferguson-Smith M, O'Brien PC. 1999. Cross-species chromosome painting between human and marsupial directly demonstrates the ancient region of the mammalian X. *Mammalian genome : official journal of the International Mammalian Genome Society* **10**(11): 1115-1116.

Gordon JL, Byrne KP, Wolfe KH. 2009. Additions, losses, and rearrangements on the evolutionary route from a reconstructed ancestor to the modern Saccharomyces cerevisiae genome. *PLoS genetics* **5**(5): e1000485.

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 research* **17**(11): 1603-1613.

Gray YH. 2000. It takes two transposons to tango: transposable-element-mediated chromosomal rearrangements. *Trends in genetics : TIG* **16**(10): 461-468.

Grutzner F, Himmelbauer H, Paulsen M, Ropers HH, Haaf T. 1999. Comparative mapping of mouse and rat chromosomes by fluorescence in situ hybridization [In Process Citation]. *Genomics* **55**(3): 306-313.

Gu W, Zhang F, Lupski JR. 2008. Mechanisms for human genomic rearrangements. *PathoGenetics* **1**(1): 4.

Guan Y, Dunham MJ, Troyanskaya OG. 2007. Functional analysis of gene duplications in Saccharomyces cerevisiae. *Genetics* **175**(2): 933-943.

Guyomard R, Boussaha M, Krieg F, Hervet C, Quillet E. 2012. A synthetic rainbow trout linkage map provides new insights into the salmonid whole genome duplication and the conservation of synteny among teleosts. *BMC genetics* **13**(1): 15.

Hakes L, Pinney JW, Lovell SC, Oliver SG, Robertson DL. 2007. All duplicates are not equal: the difference between small-scale and genome duplication. *Genome biology* **8**(10): R209.

He X, Zhang J. 2005. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* **169**(2): 1157-1164.

Hedges SB, Dudley J, Kumar S. 2006. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics* **22**(23): 2971-2972.

Hellsten U, Khokha MK, Grammer TC, Harland RM, Richardson P, Rokhsar DS. 2007. Accelerated gene evolution and subfunctionalization in the pseudotetraploid frog Xenopus laevis. *BMC biology* **5**: 31.

Hinsch H, Hannenhalli S. 2006. Recurring genomic breaks in independent lineages support genomic fragility. *BMC evolutionary biology* **6**: 90.

Hoffmann AA, Rieseberg LH. 2008. Revisiting the Impact of Inversions in Evolution: From Population Genetic Markers to Drivers of Adaptive Shifts and Speciation? *Annual review of ecology, evolution, and systematics* **39**: 21-42.

Horn A, Basset P, Yannic G, Banaszek A, Borodin PM, Bulatova NS, Jadwiszczak K, Jones RM, Polyakov AV, Ratkiewicz M et al. 2012. Chromosomal rearrangements do not seem to affect the gene flow in hybrid zones between karyotypic races of the common shrew (Sorex araneus). *Evolution; international journal of organic evolution* **66**(3): 882-889.

Huelsenbeck JP, Bollback JP. 2001. Empirical and hierarchical Bayesian estimation of ancestral states. *Systematic biology* **50**(3): 351-366.

Hufton AL, Groth D, Vingron M, Lehrach H, Poustka AJ, Panopoulou G. 2008. Early vertebrate whole genome duplications were predated by a period of intense genome rearrangement. *Genome research* **18**(10): 1582-1591.

Hufton AL, Mathia S, Braun H, Georgi U, Lehrach H, Vingron M, Poustka AJ, Panopoulou G. 2009. Deeply conserved chordate noncoding sequences preserve genome synteny but do not drive gene duplicate retention. *Genome research* **19**(11): 2036-2051.

Hufton AL, Panopoulou G. 2009. Polyploidy and genome restructuring: a variety of outcomes. *Current opinion in genetics & development* **19**(6): 600-606.

Huminiecki L, Heldin CH. 2010. 2R and remodeling of vertebrate signal transduction engine. *BMC biology* **8**: 146.

Hurst LD, Pal C, Lercher MJ. 2004. The evolutionary dynamics of eukaryotic gene order. *Nature reviews Genetics* **5**(4): 299-310.

Huvet M, Nicolay S, Touchon M, Audit B, d'Aubenton-Carafa Y, Arneodo A, Thermes C. 2007. Human gene organization driven by the coordination of replication and transcription. *Genome research* **17**(9): 1278-1285.

Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, Bernot A et al. 2004. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature* **431**(7011): 946-957.

Jaillon O, Aury JM, Wincker P. 2009. "Changing by doubling", the impact of Whole Genome Duplications in the evolution of eukaryotes. *Comptes rendus biologies* **332**(2-3): 241-253.

Jauch A, Wienberg J, Stanyon R, Arnold N, Tofanelli S, Ishida T, Cremer T. 1992. Reconstruction of genomic rearrangements in great apes and gibbons by chromosome painting. *Proceedings of the National Academy of Sciences of the United States of America* **89**(18): 8611-8615.

Kasahara M, Naruse K, Sasaki S, Nakatani Y, Qu W, Ahsan B, Yamada T, Nagayasu Y, Doi K, Kasai Y et al. 2007. The medaka draft genome and insights into vertebrate genome evolution. *Nature* **447**(7145): 714-719.

Kellis M, Birren BW, Lander ES. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. *Nature* **428**(6983): 617-624.

Kemkemer C, Kohn M, Cooper DN, Froenicke L, Hogel J, Hameister H, Kehrer-Sawatzki H. 2009. Gene synteny comparisons between different vertebrates provide new insights into breakage and fusion events during mammalian karyotype evolution. *BMC evolutionary biology* **9**: 84.

Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D. 2003. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proceedings of the National Academy of Sciences of the United States of America* **100**(20): 11484-11489.

Kidd JM, Graves T, Newman TL, Fulton R, Hayden HS, Malig M, Kallicki J, Kaul R, Wilson RK, Eichler EE. 2010. A human genome structural variation sequencing resource reveals insights into mutational mechanisms. *Cell* **143**(5): 837-847.

Kikuta H, Laplante M, Navratilova P, Komisarczuk AZ, Engstrom PG, Fredman D, Akalin A, Caccamo M, Sealy I, Howe K et al. 2007. Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome research*.

Kim J, Sinha S. 2007. Indelign: a probabilistic framework for annotation of insertions and deletions in a multiple alignment. *Bioinformatics* **23**(3): 289-297.

Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV. 2002. Selection in the evolution of gene duplications. *Genome biology* **3**(2): RESEARCH0008.

Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G et al. 2002. A high-resolution recombination map of the human genome. *Nature genetics* **31**(3): 241-247.

Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L et al. 2007. Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**(5849): 420-426.

Krishnan NM, Seligmann H, Stewart CB, De Koning AP, Pollock DD. 2004. Ancestral sequence reconstruction in primate mitochondrial DNA: compositional bias and effect on functional inference. *Molecular biology and evolution* **21**(10): 1871-1883.

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: an information aesthetic for comparative genomics. *Genome research* **19**(9): 1639-1645.

Kucherlapati RS, Creagan RP, Nichols EA, Borgaonkar DS, Ruddle FH. 1975. Synteny relationships of four human genes: mannose phosphate isomerase to pyruvate kinase-3 and triose phophate isomerase to lactate dehydrogenase-B. *Cytogenetics and cell genetics* **14**(3-6): 364-367.

Lai Z, Nakazato T, Salmaso M, Burke JM, Tang S, Knapp SJ, Rieseberg LH. 2005. Extensive chromosomal repatterning and the evolution of sterility barriers in hybrid sunflower species. *Genetics* **171**(1): 291-303.

Lalley PA, Minna JD, Francke U. 1978. Conservation of autosomal gene synteny groups in mouse and man. *Nature* **274**(5667): 160-163.

Lam HY, Mu XJ, Stutz AM, Tanzer A, Cayting PD, Snyder M, Kim PM, Korbel JO, Gerstein MB. 2010. Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. *Nature biotechnology* **28**(1): 47-55.

Lander ES Linton LM Birren B Nusbaum C Zody MC 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.

Larkin DM, Pape G, Donthu R, Auvil L, Welge M, Lewin HA. 2009. Breakpoint regions and homologous synteny blocks in chromosomes have different evolutionary histories. *Genome research* **19**(5): 770-777.

Lee AP, Koh EG, Tay A, Brenner S, Venkatesh B. 2006. Highly conserved syntenic blocks at the vertebrate Hox loci and conserved regulatory elements within and outside Hox gene clusters. *Proceedings of the National Academy of Sciences of the United States of America* **103**(18): 6994-6999.

Lee JA, Carvalho CM, Lupski JR. 2007. A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell* **131**(7): 1235-1247.

Lemaitre C, Tannier E, Gautier C, Sagot MF. 2008. Precise detection of rearrangement breakpoints in mammalian chromosomes. *BMC bioinformatics* **9**: 286.

Lemaitre C, Zaghloul L, Sagot MF, Gautier C, Arneodo A, Tannier E, Audit B. 2009. Analysis of fine-scale mammalian evolutionary breakpoints provides new insight into their relation to genome organisation. *BMC genomics* **10**: 335.

Levy AA, Feldman M. 2002. The impact of polyploidy on grass genome evolution. *Plant physiology* **130**(4): 1587-1593.

Li J, Harris RA, Cheung SW, Coarfa C, Jeong M, Goodell MA, White LD, Patel A, Kang SH, Shaw C et al. 2012. Genomic hypomethylation in the human germline associates with selective structural mutability in the human genome. *PLoS genetics* **8**(5): e1002692.

Lieber MR. 2010. The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway. *Annual review of biochemistry* **79**: 181-211.

Lieber MR, Lu H, Gu J, Schwarz K. 2008. Flexibility in the order of action and in the enzymology of the nuclease, polymerases, and ligase of vertebrate non-homologous DNA end joining: relevance to cancer, aging, and the immune system. *Cell research* **18**(1): 125-133.

Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**(5950): 289-293.

Lien S, Gidskehaug L, Moen T, Hayes BJ, Berg PR, Davidson WS, Omholt SW, Kent MP. 2011. A dense SNP-based linkage map for Atlantic salmon (Salmo salar) reveals extended chromosome homeologies and striking differences in sex-specific recombination patterns. *BMC genomics* **12**: 615.

Lin C, Yang L, Tanasa B, Hutt K, Ju BG, Ohgi K, Zhang J, Rose DW, Fu XD, Glass CK et al. 2009. Nuclear receptor-induced chromosomal proximity and DNA breaks underlie specific translocations in cancer. *Cell* **139**(6): 1069-1083.

Liu P, Carvalho CM, Hastings P, Lupski JR. 2012. Mechanisms for recurrent and complex human genomic rearrangements. *Current opinion in genetics & development*.

Lowry DB, Willis JH. 2010. A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation. *PLoS biology* **8**(9).

Lupski JR. 1998. Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends in genetics : TIG* **14**(10): 417-422.

Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science* **290**(5494): 1151-1155.

Lynch M, Force A. 2000. The probability of duplicate gene preservation by subfunctionalization. *Genetics* **154**(1): 459-473.

Ma J, Zhang L, Suh BB, Raney BJ, Burhans RC, Kent WJ, Blanchette M, Haussler D, Miller W. 2006. Reconstructing contiguous regions of an ancestral genome. *Genome research* **16**(12): 1557-1565.

Maas NM, Van Vooren S, Hannes F, Van Buggenhout G, Mysliwiec M, Moreau Y, Fagan K, Midro A, Engiz O, Balci S et al. 2007. The t(4;8) is mediated by homologous recombination between olfactory receptor gene clusters, but other 4p16 translocations occur at random. *Genet Couns* **18**(4): 357-365.

Machado CA, Kliman RM, Markert JA, Hey J. 2002. Inferring the history of speciation from multilocus DNA sequence data: the case of Drosophila pseudoobscura and close relatives. *Molecular biology and evolution* **19**(4): 472-488.

Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, Van de Peer Y. 2005. Modeling gene and genome duplications in eukaryotes. *Proceedings of the National Academy of Sciences of the United States of America* **102**(15): 5454-5459.

Makalowski W. 2001. Are we polyploids? A brief history of one hypothesis. *Genome research* **11**(5): 667-670.

Makino T, McLysaght A. 2010. Ohnologs in the human genome are dosage balanced and frequently associated with disease. *Proceedings of the National Academy of Sciences of the United States of America* **107**(20): 9270-9274.

Marques-Bonet T, Sanchez-Ruiz J, Armengol L, Khaja R, Bertranpetit J, Lopez-Bigas N, Rocchi M, Gazave E, Navarro A. 2007. On the association between chromosomal rearrangements and genic evolution in humans and chimpanzees. *Genome biology* **8**(10): R230.

Masterson J. 1994. Stomatal size in fossil plants: evidence for polyploidy in majority of angiosperms. *Science* **264**(5157): 421-424.

Mathas S, Misteli T. 2009. The dangers of transcription. *Cell* **139**(6): 1047-1049.

McFadden DE, Kwong LC, Yam IY, Langlois S. 1993. Parental origin of triploidy in human fetuses: evidence for genomic imprinting. *Human genetics* **92**(5): 465-469.

McLysaght A, Hokamp K, Wolfe KH. 2002. Extensive genomic duplication during early chordate evolution. *Nature genetics* **31**(2): 200-204.

Mechali M. 2010. Eukaryotic DNA replication origins: many choices for appropriate answers. *Nature reviews Molecular cell biology* **11**(10): 728-738.

Meyer A, Van de Peer Y. 2005. From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). *BioEssays : news and reviews in molecular, cellular and developmental biology* **27**(9): 937-945.

Mongin E, Dewar K, Blanchette M. 2009. Long-range regulation is a major driving force in maintaining genome integrity. *BMC evolutionary biology* **9**: 203.

Muffato M. 2010. Reconstruction de génomes ancestraux chez les vertébrés.

Muffato M, Louis A, Poisnel CE, Roest Crollius H. 2010. Genomicus: a database and a browser to study gene synteny in modern and ancestral genomes. *Bioinformatics* **26**(8): 1119-1121.

Muffato M, Roest Crollius H. 2008. Paleogenomics in vertebrates, or the recovery of lost genomes from the mist of time. *BioEssays : news and reviews in molecular, cellular and developmental biology* **30**(2): 122-134.

Muller S, Stanyon R, Finelli P, Archidiacono N, Wienberg J. 2000. Molecular cytogenetic dissection of human chromosomes 3 and 21 evolution. *Proceedings of the National Academy of Sciences of the United States of America* **97**(1): 206-211.

Muller S, Stanyon R, O'Brien PC, Ferguson-Smith MA, Plesker R, Wienberg J. 1999. Defining the ancestral karyotype of all primates by multidirectional chromosome painting between tree shrews, lemurs and humans. *Chromosoma* **108**(6): 393-400.

Muller S, Wienberg J. 2001. "Bar-coding" primate chromosomes: molecular cytogenetic screening for the ancestral hominoid karyotype. *Human genetics* **109**(1): 85-94.

Mural RJ Adams MD Myers EW Smith HO Miklos GL Wides R Halpern A Li PW Sutton GG Nadeau J et al. 2002. A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. *Science* **296**(5573): 1661-1671.

Murphy WJ, Larkin DM, Everts-van der Wind A, Bourque G, Tesler G, Auvil L, Beever JE, Chowdhary BP, Galibert F, Gatzke L et al. 2005. Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science* **309**(5734): 613-617.

Nadeau JH, Taylor BA. 1984. Lengths of chromosomal segments conserved since divergence of man and mouse. *Proceedings of the National Academy of Sciences of the United States of America* **81**(3): 814-818.

Nakatani Y, Takeda H, Kohara Y, Morishita S. 2007. Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. *Genome research* **17**(9): 1254-1265.

Nam K, Ellegren H. 2012. Recombination drives vertebrate genome contraction. *PLoS genetics* **8**(5): e1002680.

Naruse K, Fukamachi S, Mitani H, Kondo M, Matsuoka T, Kondo S, Hanamura N, Morita Y, Hasegawa K, Nishigaki R et al. 2000. A detailed linkage map of medaka, Oryzias latipes: comparative genomics and genome evolution. *Genetics* **154**(4): 1773-1784.

Nash WG, O'Brien SJ. 1982. Conserved regions of homologous G-banded chromosomes between orders in mammalian evolution: carnivores and primates. *Proceedings of the National Academy of Sciences of the United States of America* **79**(21): 6631-6635.

Navarro A, Barton NH. 2003a. Accumulating postzygotic isolation genes in parapatry: a new twist on chromosomal speciation. *Evolution; international journal of organic evolution* **57**(3): 447-459.

-. 2003b. Chromosomal speciation and molecular divergence--accelerated evolution in rearranged chromosomes. *Science* **300**(5617): 321-324.

Nelson JS. 2006. *Fishes of the world*. John Wiley & Sons, Hoboken, New Jersey.

Nery MF, Gonzalez DJ, Hoffmann FG, Opazo JC. 2012. Resolution of the laurasiatherian phylogeny: Evidence from genomic data. *Molecular phylogenetics and evolution*.

Nobrega MA, Ovcharenko I, Afzal V, Rubin EM. 2003. Scanning human gene deserts for long-range enhancers. *Science* **302**(5644): 413.

Noor MA, Grams KL, Bertucci LA, Reiland J. 2001. Chromosomal inversions and the reproductive isolation of species. *Proceedings of the National Academy of Sciences of the United States of America* **98**(21): 12084-12088.

O'Brien SJ, Menotti-Raymond M, Murphy WJ, Nash WG, Wienberg J, Stanyon R, Copeland NG, Jenkins NA, Womack JE, Marshall Graves JA. 1999. The promise of comparative genomics in mammals. *Science* **286**(5439): 458-462, 479-481.

O'Brien SJ, Womack JE, Lyons LA, Moore KJ, Jenkins NA, Copeland NG. 1993. Anchored reference loci for comparative genome mapping in mammals. *Nature genetics* **3**(2): 103-112.

Ohno S. 1970. *Evolution by gene duplication*. Allen and Unwin, London.

-. 1973. Ancient linkage groups and frozen accidents. *Nature* **244**(5414): 259-262.

Otto SP. 2007. The evolutionary consequences of polyploidy. *Cell* **131**(3): 452-462.

Otto SP, Whitton J. 2000. Polyploid incidence and evolution. *Annual review of genetics* **34**: 401-437.

Ou Z, Stankiewicz P, Xia Z, Breman AM, Dawson B, Wiszniewska J, Szafranski P, Cooper ML, Rao M, Shao L et al. 2011. Observation and prediction of recurrent human translocations mediated by NAHR between nonhomologous chromosomes. *Genome research* **21**(1): 33-46.

Ovcharenko I, Loots GG, Nobrega MA, Hardison RC, Miller W, Stubbs L. 2005. Evolution and functional classification of vertebrate gene deserts. *Genome research* **15**(1): 137-145.

Pal C, Hurst LD. 2003. Evidence for co-evolution of gene order and recombination rate. *Nature genetics* **33**(3): 392-395.

Palti Y, Genet C, Luo MC, Charlet A, Gao G, Hu Y, Castano-Sanchez C, Tabet-Canale K, Krieg F, Yao J et al. 2011. A first generation integrated map of the rainbow trout genome. *BMC genomics* **12**: 180.

Passarge E, Horsthemke B, Farber RA. 1999. Incorrect use of the term synteny. *Nature genetics* **23**(4): 387.

Paten B, Herrero J, Fitzgerald S, Beal K, Flicek P, Holmes I, Birney E. 2008. Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome research* **18**(11): 1829-1843.

Paterson AH, Bowers JE, Chapman BA. 2004. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proceedings of the National Academy of Sciences of the United States of America* **101**(26): 9903-9908.

Peng Q, Pevzner PA, Tesler G. 2006. The fragile breakage versus random breakage models of chromosome evolution. *PLoS computational biology* **2**(2): e14.

Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD et al. 2006. In vivo enhancer analysis of human conserved non-coding sequences. *Nature* **444**(7118): 499-502.

Petrov DA, Sangster TA, Johnston JS, Hartl DL, Shaw KL. 2000. Evidence for DNA loss as a determinant of genome size. *Science* **287**(5455): 1060-1062.

Pevzner P, Tesler G. 2003a. Genome rearrangements in mammalian evolution: lessons from human and mouse genomes. *Genome research* **13**(1): 37-45.

-. 2003b. Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proceedings of the National Academy of Sciences of the United States of America* **100**(13): 7672-7677.

Postlethwait J, Amores A, Cresko W, Singer A, Yan YL. 2004. Subfunction partitioning, the teleost radiation and the annotation of the human genome. *Trends in genetics : TIG* **20**(10): 481-490.

Postlethwait JH, Woods IG, Ngo-Hazelett P, Yan YL, Kelly PD, Chu F, Huang H, Hill-Force A, Talbot WS. 2000. Zebrafish comparative genomics and the origins of vertebrate chromosomes. *Genome research* **10**(12): 1890-1902.

Poulin F, Nobrega MA, Plajzer-Frick I, Holt A, Afzal V, Rubin EM, Pennacchio LA. 2005. In vivo characterization of a vertebrate ultraconserved enhancer. *Genomics* **85**(6): 774-781.

Poyatos JF, Hurst LD. 2007. The determinants of gene order conservation in yeasts. *Genome biology* **8**(11): R233.

Prakash O, Yunis JJ. 1984. High resolution chromosomes of the t(9;22) positive leukemias. *Cancer genetics and cytogenetics* **11**(4): 361-367.

Ramsey J, Schemske, D.W. 1998. Pathways, mechanisms, and rates of polyploid formation in flowering plants. *Annu Rec Ecol Syst* **29**(467-501).

Ranz JM, Maurin D, Chan YS, von Grotthuss M, Hillier LW, Roote J, Ashburner M, Bergman CM. 2007. Principles of genome evolution in the Drosophila melanogaster species group. *PLoS biology* **5**(6): e152.

Raudsepp T, Fronicke L, Scherthan H, Gustavsson I, Chowdhary BP. 1996. Zoo-FISH delineates conserved chromosomal segments in horse and man. *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology* **4**(3): 218-225.

Reiter LT, Hastings PJ, Nelis E, De Jonghe P, Van Broeckhoven C, Lupski JR. 1998. Human meiotic recombination products revealed by sequencing a hotspot for homologous strand exchange in multiple HNPP deletion patients. *American journal of human genetics* **62**(5): 1023-1033.

Ryba T, Hiratani I, Lu J, Itoh M, Kulik M, Zhang J, Schulz TC, Robins AJ, Dalton S, Gilbert DM. 2010. Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome research* **20**(6): 761-770.

Salmon A, Ainouche ML, Wendel JF. 2005. Genetic and epigenetic consequences of recent hybridization and polyploidy in Spartina (Poaceae). *Molecular ecology* **14**(4): 1163-1175.

Sandelin A, Bailey P, Bruce S, Engstrom PG, Klos JM, Wasserman WW, Ericson J, Lenhard B. 2004. Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. *BMC genomics* **5**(1): 99.

Sankoff D, Trinh P. 2005. Chromosomal breakpoint reuse in genome sequence rearrangement. *Journal of computational biology : a journal of computational molecular cell biology* **12**(6): 812-821.

Santini F, Harmon LJ, Carnevale G, Alfaro ME. 2009. Did genome duplication drive the origin of teleosts? A comparative study of diversification in ray-finned fishes. *BMC evolutionary biology* **9**: 194.

Sasaki M, Lange J, Keeney S. 2010. Genome destabilization by homologous recombination in the germ line. *Nature reviews Molecular cell biology* **11**(3): 182-195.

Satake M, Kawata M, McLysaght A, Makino T. 2012. Evolution of Vertebrate Tissues Driven by Differential Modes of Gene Duplication. *DNA research : an international journal for rapid publication of reports on genes and genomes*.

Sawyer JR, Hozier JC. 1986. High resolution of mouse chromosomes: banding conservation between man and mouse. *Science* **232**(4758): 1632-1635.

Scannell DR, Byrne KP, Gordon JL, Wong S, Wolfe KH. 2006. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature* **440**(7082): 341-345.

Scannell DR, Frank AC, Conant GC, Byrne KP, Woolfit M, Wolfe KH. 2007. Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication. *Proceedings of the National Academy of Sciences of the United States of America* **104**(20): 8397-8402.

Scherthan H, Cremer T, Arnason U, Weier HU, Lima-de-Faria A, Fronicke L. 1994. Comparative chromosome painting discloses homologous segments in distantly related mammals. *Nature genetics* **6**(4): 342-347.

Schibler L, Roig A, Mahe MF, Laurent P, Hayes H, Rodolphe F, Cribiu EP. 2006. High-resolution comparative mapping among man, cattle and mouse suggests a role for repeat sequences in mammalian genome evolution. *BMC genomics* **7**: 194.

Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, Kutter C, Watt S, Martinez-Jimenez CP, Mackay S et al. 2010. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* **328**(5981): 1036-1040.

Schughart K, Kappen C, Ruddle FH. 1989. Duplication of large genomic regions during the evolution of vertebrate homeobox genes. *Proceedings of the National Academy of Sciences of the United States of America* **86**(18): 7067-7071.

Sémon M, Duret L. 2006. Evolutionary origin and maintenance of coexpressed gene clusters in mammals. *Molecular biology and evolution* **23**(9): 1715-1723.

Semon M, Wolfe KH. 2007a. Rearrangement Rate following the Whole-Genome Duplication in Teleosts. *Molecular biology and evolution* **24**(3): 860-867.

-. 2007b. Reciprocal gene loss between Tetraodon and zebrafish after whole genome duplication in their ancestor. *Trends in genetics : TIG* **23**(3): 108-112.

-. 2008. Preferential subfunctionalization of slow-evolving genes after allopolyploidization in Xenopus laevis. *Proceedings of the National Academy of Sciences of the United States of America* **105**(24): 8333-8338.

Seoighe C, Wolfe KH. 1999. Yeast genome evolution in the post-genome era. *Current opinion in microbiology* **2**(5): 548-554.

Shaked H, Kashkush K, Ozkan H, Feldman M, Levy AA. 2001. Sequence elimination and cytosine methylation are rapid and reproducible responses of the genome to wide hybridization and allopolyploidy in wheat. *The Plant cell* **13**(8): 1749-1759.

Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, Vallente RU, Pertz LM, Clark RA, Schwartz S, Segraves R et al. 2005. Segmental duplications and copy-number variation in the human genome. *American journal of human genetics* **77**(1): 78-88.

Shaw CJ, Lupski JR. 2004. Implications of human genome architecture for rearrangement-based disorders: the genomic basis of disease. *Human molecular genetics* **13 Spec No 1**: R57-64.

Sherman DJ, Martin T, Nikolski M, Cayla C, Souciet JL, Durrens P. 2009. Genolevures: protein families and synteny among complete hemiascomycetous yeast proteomes and genomes. *Nucleic acids research* **37**(Database issue): D550-554.

Shoemaker RC, Schlueter J, Doyle JJ. 2006. Paleopolyploidy and gene duplication in soybean and other legumes. *Current opinion in plant biology* **9**(2): 104-109.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research* **15**(8): 1034-1050.

Skinner BM, Griffin DK. 2012. Intrachromosomal rearrangements in avian genome evolution: evidence for regions prone to breakpoints. *Heredity* **108**(1): 37-41.

Skrabanek L, Wolfe KH. 1998. Eukaryote genome duplication - where's the evidence? *Current opinion in genetics & development* **8**(6): 694-700.

Snir S, Pachter L. 2011. Tracing the most parsimonious indel history. *Journal of computational biology : a journal of computational molecular cell biology* **18**(8): 967-986.

Soutoglou E, Dorn JF, Sengupta K, Jasin M, Nussenzweig A, Ried T, Danuser G, Misteli T. 2007. Positional stability of single double-strand breaks in mammalian cells. *Nature cell biology* **9**(6): 675-682.

Spring J. 2002. Genome duplication strikes back. *Nature genetics* **31**(2): 128-129.

Stankiewicz P, Lupski JR. 2002. Genome architecture, rearrangements and genomic disorders. *Trends in genetics : TIG* **18**(2): 74-82.

-. 2006. The genomic basis of disease, mechanisms and assays for genomic disorders. *Genome dynamics* **1**: 1-16.

Stanyon R, Consigliere S, Bigoni F, Ferguson-Smith M, O'Brien PC, Wienberg J. 2001. Reciprocal chromosome painting between a New World primate, the woolly monkey, and humans. *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology* **9**(2): 97-106.

Swidan F, Rocha EP, Shmoish M, Pinter RY. 2006. An integrative method for accurate comparative genome mapping. *PLoS computational biology* **2**(8): e75.

Symington LS, Gautier J. 2011. Double-strand break end resection and repair pathway choice. *Annual review of genetics* **45**: 247-271.

Taylor JS, Braasch I, Frickey T, Meyer A, Van de Peer Y. 2003. Genome duplication, a trait shared by 22000 species of ray-finned fish. *Genome research* **13**(3): 382-390.

Taylor JS, Brinkmann H. 2001. 2R or not 2R? *Trends in genetics : TIG* **17**(9): 488-489.

Tesler G. 2002. GRIMM: genome rearrangements web server. *Bioinformatics* **18**(3): 492-493.

Thompson JD, Lumaret R. 1992. The evolutionary dynamics of polyploid plants: origins, establishment and persistence. *Trends in ecology & evolution* **7**(9): 302-307.

Touchon M, Nicolay S, Audit B, Brodie of Brodie EB, d'Aubenton-Carafa Y, Arneodo A, Thermes C. 2005. Replication-associated strand asymmetries in mammalian genomes: toward detection of replication origins. *Proceedings of the National Academy of Sciences of the United States of America* **102**(28): 9836-9841.

Trickett AJ, Butlin RK. 1994. Recombination suppressors and the evolution of new species. *Heredity* **73 (Pt 4)**: 339-345.

Trinh P, McLysaght A, Sankoff D. 2004. Genomic features in the breakpoint regions between syntenic blocks. *Bioinformatics* **20 Suppl 1**: i318-325.

Uchida IA, Freeman VC. 1985. Triploidy and chromosomes. *American journal of obstetrics and gynecology* **151**(1): 65-69.

Udall JA, Swanson JM, Nettleton D, Percifield RJ, Wendel JF. 2006. A novel approach for characterizing expression levels of genes duplicated by polyploidy. *Genetics* **173**(3): 1823-1827.

Van de Peer Y. 2004. Computational approaches to unveiling ancient genome duplications. *Nature reviews Genetics* **5**(10): 752-763.

Van de Peer Y, Fawcett JA, Proost S, Sterck L, Vandepoele K. 2009a. The flowering world: a tale of duplications. *Trends in plant science* **14**(12): 680-688.

Van de Peer Y, Maere S, Meyer A. 2009b. The evolutionary significance of ancient genome duplications. *Nature reviews Genetics* **10**(10): 725-732.

van Hoof A. 2005. Conserved functions of yeast genes support the duplication, degeneration and complementation model for gene duplication. *Genetics* **171**(4): 1455-1461.

van Someren HV, Beyersbergen van H, de Wit J. 1974. Proceedings: Evidence for synteny between the human loci for fumarate hydratase, UDP glucose pyrophosphorylase, 6-phosphogluconate dehydrogenase, phosphoglucomutase1, and peptidase-C in man-Chinese hamster somatic cell hybrids. *Cytogenetics and cell genetics* **13**(1): 150-152.

Vandepoele K, De Vos W, Taylor JS, Meyer A, Van de Peer Y. 2004. Major events in the genome evolution of vertebrates: paranome age and size differ considerably between ray-finned fishes and land vertebrates. *Proceedings of the National Academy of Sciences of the United States of America* **101**(6): 1638-1643.

Veitia RA. 2010. A generalized model of gene dosage and dominant negative effects in macromolecular complexes. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* **24**(4): 994-1002.

Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. 2009. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome research* **19**(2): 327-335.

Volker M, Backstrom N, Skinner BM, Langley EJ, Bunzey SK, Ellegren H, Griffin DK. 2010. Copy number variation, chromosome rearrangement, and their association with recombination during avian evolution. *Genome research* **20**(4): 503-511.

Wallace IM, O'Sullivan O, Higgins DG, Notredame C. 2006. M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic acids research* **34**(6): 1692-1699.

Wapinski I, Pfeffer A, Friedman N, Regev A. 2007a. Automatic genome-wide reconstruction of phylogenetic gene trees. *Bioinformatics* **23**(13): i549-558.

-. 2007b. Natural history and evolutionary principles of gene duplication in fungi. *Nature* **449**(7158): 54-61.

Warburton D. 1991. De novo balanced chromosome rearrangements and extra marker chromosomes identified at prenatal diagnosis: clinical significance and distribution of breakpoints. *American journal of human genetics* **49**(5): 995-1013.

Webber C, Ponting CP. 2005. Hotspots of mutation and breakage in dog and human chromosomes. *Genome research* **15**(12): 1787-1797.

Wendel JF. 2000. Genome evolution in polyploids. *Plant molecular biology* **42**(1): 225-249.

Wienberg J. 2004. The evolution of eutherian chromosomes. *Current opinion in genetics & development* **14**(6): 657-666.

Wienberg J, Stanyon R. 1995. Chromosome painting in mammals as an approach to comparative genomics. *Current opinion in genetics & development* **5**(6): 792-797.

-. 1997. Comparative painting of mammalian chromosomes. *Current opinion in genetics & development* **7**(6): 784-791.

Wilming LG, Gilbert JG, Howe K, Trevanion S, Hubbard T, Harrow JL. 2008. The vertebrate genome annotation (Vega) database. *Nucleic acids research* **36**(Database issue): D753-760.

Wittbrodt J, Meyer A, Schartl M. 1998. More genes in fish? *BioEssays : news and reviews in molecular, cellular and developmental biology* **20**(6): 511-515.

Wolfe KH. 2001. Yesterday's polyploids and the mystery of diploidization. *Nature reviews Genetics* **2**(5): 333-341.

Wolfe KH, Shields DC. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387**(6634): 708-713.

Woods IG, Kelly PD, Chu F, Ngo-Hazelett P, Yan YL, Huang H, Postlethwait JH, Talbot WS. 2000. A comparative map of the zebrafish genome. *Genome research* **10**(12): 1903-1914.

Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, Smith SF, North P, Callaway H, Kelly K et al. 2005. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS biology* **3**(1): e7.

Wurster-Hill DH, Gray CW. 1975. The interrelationships of chromosome banding patterns in procyonids, viverrids, and felids. *Cytogenetics and cell genetics* **15**(5): 306-331.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer applications in the biosciences : CABIOS* **13**(5): 555-556.

Yannic G, Basset P, Hausser J. 2009. Chromosomal rearrangements and gene flow over time in an inter-specific hybrid zone of the Sorex araneus group. *Heredity* **102**(6): 616-625.

Zhang J, Nei M. 1997. Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. *Journal of molecular evolution* **44 Suppl 1**: S139-146.

Zhang J, Wang X, Podlaha O. 2004. Testing the chromosomal speciation hypothesis for humans and chimpanzees. *Genome research* **14**(5): 845-851.

Zhao H, Bourque G. 2009. Recovering genome rearrangements in the mammalian phylogeny. *Genome research* **19**(5): 934-942.

Zhao S, Shetty J, Hou L, Delcher A, Zhu B, Osoegawa K, de Jong P, Nierman WC, Strausberg RL, Fraser CM. 2004. Human, mouse, and rat genome large-scale rearrangements: stability versus speciation. *Genome research* **14**(10A): 1851-1860.

Zierhut C, Diffley JF. 2008. Break dosage, cell cycle stage and DNA replication influence DNA double strand break response. *The EMBO journal* **27**(13): 1875-1885.

Annexe

Inter-gene distance determines evolutionary breakpoints occurrence in eukaryotic genomes

Camille BERTHELOT[1,2,3], Matthieu MUFFATO[1,2,3,4], Judith ABECASSIS[1,2,3] and Hugues ROEST CROLLIUS[1,2,3]

1. Ecole Normale Supérieure, Institut de Biologie de l'ENS, IBENS, Paris, F-75005 France.

2. CNRS, UMR 8197, Paris, F-75005 France.

3. Inserm, U1024, Paris, F-75005 France.

4. New address: European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK.

Corresponding author:

Hugues ROEST CROLLIUS

Institut de Biologie de l'ENS,

46, rue d'Ulm
75005 Paris, France

email: hrc@ens.fr
tel: +33144322370

Running title: A neutral evolutionary breakpoint model

Key words: Genome, Evolution, Rearrangements, Poisson model

Abstract

Evolutionary rearrangements that modify the order of genes in the genomes of different species are the result of a complex interplay between molecular events and evolutionary processes, from their initial occurrence in the genome of an individual to their fixation in a species. The distribution of evolutionary rearrangement breakpoints is debated, because the original "random" model proposed in the late 1980's does not account for finer scale observations including potential fragile regions, regions under negative selection, and observations that breakpoints are correlated to genomic features such as gene density, repeated elements or base composition. Here we propose a novel approach based on the reconstruction of gene order and local genomic characteristics in ancestral genomes of mammals and yeasts. We use Poisson regression to establish a simple mathematical model that accurately describes the pattern of breakpoints observed in mammalian and yeast genomes. The distribution of breakpoints can be exclusively explained as a log-linear function of a single parameter, the original length of intergenic spacers where breakpoints occurred. When applied to simulate breakpoints in the human genome, this model recapitulates remarkably well previously observed correlations between breakpoints and specific genomic features. We show that rearrangement occurrence and fixation can be explained as a simple, neutral process where natural selection to preserve local genome organisation plays only a marginal role. The log-linear relationship between breakage rate and intergene length could reflect the impact of local chromatin organisation on DNA susceptibility to breakage.

Introduction

Chromosome rearrangements are thought to occur through one of several mechanisms including non-allelic homologous recombination during meiosis (NAHR), non-homologous end joining in double-strand break repair (NHEJ), and fork stalling and template switching during replication (FoSTeS) (Shaw and Lupski 2004; Lee et al. 2007; Kidd et al. 2010). Some may be benign, such as the human chromosome 9 pericentric inversion present in 0.8 to 2.0 % of the population with no apparent functional effects (Tawn and Earl 1992). Others may disrupt functional sequences, either directly by interrupting their DNA sequence or indirectly by physically unlinking elements that normally function in *cis*. While the former will be subject to genetic drift, the latter may be affected by negative selection. Therefore phenotypically neutral rearrangements are more likely to reach fixation, and thus more likely to be observed in present-day genomic comparisons. Genomic rearrangements can be diagnosed when the order of orthologous sequences along the chromosomes of two species is interrupted, by so-called breakpoints. The distribution pattern of breakpoints in eukaryote genomes has

been the subject of much debate. In 1984, Nadeau and Taylor showed that the distribution of segment lengths between consecutive breaks of marker collinearity between the human and mouse genomes was consistent with a pure Poisson process, i.e. that occurrence and fixation resulted in a random distribution of breakpoints (Nadeau and Taylor 1984), a conclusion further supported by subsequent studies (Nadeau and Sankoff 1998; Sankoff and Trinh 2005). However, comparisons of whole genome sequences between species later provided increased resolution and revealed many closely located breakpoints that had previously been overlooked. In addition, when computational approaches were developed to identify the most likely scenario of rearrangements required to theoretically transform one extant genome into another, a higher incidence of closely located and sometimes indistinguishable breakpoints had to be inferred compared to random expectations, a phenomenon referred to as "breakpoint reuse" (Kent et al. 2003; Pevzner and Tesler 2003; Bourque et al. 2004; Murphy et al. 2005; Alekseyev and Pevzner 2007). This excess of clustered breakpoints was interpreted as evidence for fragile genomic regions recurrently broken during evolution, or with a higher likelihood of breakpoint fixation.

The existence of "fragile regions" has been interpreted in several ways: they might either be structural, resulting from a physical fragility in some genomic regions, or they might arise by contrast with functional regions where breakpoints are highly deleterious and are eliminated by selection. Previous evidence has shown that both do occur in genomes: genes, for instance, are very rarely disrupted by breakpoints (Peng et al. 2006), and even consecutive genes are thought to remain in conserved order throughout evolution because of interdigitated regulatory sequences (for example, the *shh* gene locus) (Goode et al. 2005; Engstrom et al. 2007; Kikuta et al. 2007; Hufton et al. 2009). On the other hand, modern genomes are known to harbour regions that have a higher propensity to breakage, as observed in cancer genomes (Darai-Ramqvist et al. 2008) and recurrent rearrangements causing genetic diseases (Shaw and Lupski 2004), and the statistical association of rearrangement breakpoints with features such as high GC content and short intergenes has prompted the hypothesis that isochores and domains of open and closed chromatin may play a substantial role in the occurrence of breakage events (Ma et al. 2006; Larkin et al. 2009; Lemaitre et al. 2009). How much each of these processes contributes to the final pattern of evolutionary breakpoints, however, is not understood (Becker and Lenhard 2007). In any case, it is now generally admitted that random processes alone cannot explain the observed distribution of evolutionary rearrangements breakpoints in genomes.

It is tempting to use the recent abundance of comparative genomic data to identify regions of lower evolutionary rearrangement frequency between genomes, and interpret them as evidence of selective pressure and, from there, of functional organization (Becker and Lenhard 2007; Sun et al. 2008; Mongin et al. 2011). Yet such approaches overlook the possibility that genome structure itself may condition breakage probability, resulting in variations in rearrangement frequency conserved over

long evolutionary times without any selective constraints against rearrangements. Understanding whether changes in gene order between different species are mainly governed by the physical structure of the genome or by selective constraints on functionally relevant regions would provide a framework for the use and interpretation of such data, and is important for our global understanding of genome evolution.

Here we use ancestral genomes reconstructions in mammals and yeasts to model how a genome is affected by rearrangements during evolution. The multivariate statistical approach we propose identifies the main parameters affecting the distribution of breakpoints along the genome. Surprisingly, we not only find that a small number of parameters is sufficient to accurately model the breakpoint pattern in eukaryotic genomes, but that this pattern is mainly governed by variations in intergene length between genes, although in a non-linear fashion. Our results suggest that chromatin compaction and DNA accessibility may be mostly responsible for the non-random distribution of breakpoints in eukaryotic genomes, and that selective pressure to preserve the gene organization from rearrangements is only playing a marginal role.

Results

Ancestral genome reconstruction

Studying the distribution pattern of breakpoints has traditionally relied on comparisons between two or more genomes to identify chromosomal regions that show a discontinuity in one of the lineages. These regions are rearrangement breakpoints. Numerous previous studies have uncovered specificities around such breakpoints, by examining the non-rearranged version in a different lineage and determining whether these regions are enriched or depleted in particular features compared to the genome average (Murphy et al. 2005; Ma et al. 2006; Gordon et al. 2007; Larkin et al. 2009; Lemaitre et al. 2009). Such comparisons usually involve genomes that are evolutionarily distant enough to provide a sufficient number of breakpoints to warrant statistical analysis. However, a distant reference genome may have changed substantially and lost at least partially the original features that have promoted breakage in another lineage. Here, we take a different approach by estimating the ancestral states of rearranged and non-rearranged regions in order to find which characteristics have influenced breakage. Since breakpoints are almost exclusively intergenic (Peng et al. 2006), we focus our study specifically on ancestral gene-to-gene intervals as the primary rearranged units, under the assumption that the precise details of the ancestral genomic sequence may not be necessary to understand breakpoint occurrences. The rationale of the method is explained in Figure 1.

We reconstructed ancestral gene adjacencies in the 95 million years old ancestral Boreoeutheria genome, the last common ancestor of Primates, Rodents and Laurasiatherians (e.g. cow, dog, horse). This ancestor has been the target of previous studies aiming at ancestral genome reconstructions, which either used less extant genomes (Ma et al. 2006; Chauve and Tannier 2008) or focused on genomic DNA sequence rather than the long range order of genes (Blanchette et al. 2004; Paten et al. 2008). With 28 sequenced descendant genomes (in Ensembl version 57) and several closely branching outgroups, it is ideally placed in the mammalian tree for ancestral genome reconstruction and breakpoint analysis over many lineages (Blanchette et al. 2004). We documented the gene content of the Boreoeutheria genome using Ensembl phylogenetic gene trees (Vilella et al. 2009). We then determined which genes were adjacent (directly consecutive to each other) in the ancestral genome, and thus define the borders of ancestral intergenes (continuous non-coding spacers between consecutive genes). Briefly, the intergenes in the ancestral Boreoeutheria genome were reconstructed by systematic comparisons of the order and orientation of genes for all pairs of genomes that diverged earlier than or at the Boreoeutheria node in the tree (Figure 1, methods, Supplementary Text S1). Under a parsimonious reasoning, genes that are adjacent and in the same orientation in such two genomes have probably retained their ancestral configuration relative to each other. They define an orthologous intergene that existed in the last common ancestor of two species, and in all intermediate ancestral genomes between them, including Boreoeutheria.

Using this method, we reconstructed 18,436 gene adjacencies (and therefore, intergenes) in the Boreoeutheria ancestor. A typical mammalian genome sequenced with high coverage contains 17,000 to 23,000 gene adjacencies, depending on total gene count and assembly fragmentation: this ancestral reconstruction of the ancestral Boreoeutheria genome is highly comprehensive and compares to a deep-sequenced modern genome. In order to assess the robustness of the reconstruction, we counted the number of modern genomes where a pair of adjacent genes can be found in the same configuration as the inferred ancestral one: a gene configuration that exists in multiple genomes in different lineages is very likely to be the inherited ancestral configuration. On average, ancestral intergenes are observed in 13.7 descendant genomes out of 28 (SD = 6.2; Figure 2A), 73.5% of intergenes are supported by more than 10 descendant species, and more than 90% of the intergenes can be observed in at least 5 descendant genomes, including at least one Laurasiatherian and one Euarchontoglire. Of note, 16 out of 28 sequenced Boreoeutherian genomes used in this study are highly fragmented assemblies, often preventing the observation of potentially conserved gene adjacencies (Figure 2B).

Identification of rearranged intergenes in five independent lineages

In order to identify intergenes that have been affected by a rearrangement breakpoint, we

compared ancestral Boreoeutheria intergenes to five extant Boreoeutherian genomes (human, mouse, dog, cow and horse) chosen for the quality of their assembly and annotation. The five lineages have radiated at short time intervals after the Boreoeutheria ancestor (Nery et al. 2012), so that the vast majority of breakpoints are expected to be independent events. We compared the orders and orientations of ancestral genes to their modern copies in each genome to identify lost adjacencies, which correspond to ancestral intergenes that have been affected by a rearrangement breakpoint in subsequent evolution. Some breakpoints cannot be precisely assigned to one ancestral intergene, but to a range of consecutive ancestral intergenes. This situation arises when a breakpoint occurred in a region where intergenes have also been modified by gene gains (creation of new intergenes) or losses (fusion of intergenes). As we are not able to decide whether these changes in gene content took place before, during or after the rearrangement event, it is not possible to know whether the ancestral state was still relevant at the time when the breakpoint occurred (see Figure S1). Breakpoints that cannot be mapped to a single ancestral intergene were therefore not considered in the analysis. In addition, manual curation revealed that many apparent breakpoints are most likely due to local assembly or annotation errors in one of the five genomes. We thus applied a filtering step to distinguish between true breakpoints and such false positives (see methods and Supplemental Text S2).

In the five lineages, we identify a total of 779 breakpoints that can be mapped to a precise ancestral intergene: 100 in the human lineage, 176 in the mouse, 116 in the dog, 305 in the cow and 82 in the horse lineage (Figure 3). Interestingly, 20 intergenes have been broken at least twice in independent lineages and correspond to breakpoint reoccurrence while 24 intergenes are disrupted in two or three species that share a short branch in the tree (human/mouse or cow/horse/dog). The latter probably reflect unique breakage events that occurred before speciation, and they are considered as such in our analysis, resulting in a total of 751 breakpoints. The breakpoint counts along the different branches of the tree are of the same order of magnitude as previously reported in the literature (Larkin et al. 2009; Zhao and Bourque 2009). Ancestral intergenes rearranged in one lineage can still be observed in a mean of 11.2 modern genomes, compared to 13.8 genomes for non-rearranged intervals. This difference is expected because rearranged intervals are, by definition, present less often in modern genomes than non-rearranged intervals. More than 87% of rearranged intervals are have retained the ancestral configuration in at least 5 modern boreoeutherian species, a similar proportion to the non-rearranged ones (proportion test, $P = 0.34$). We compared our list of breakpoints with an independent dataset (Larkin et al. 2009), which identified 433 breakpoint regions between the human, mouse, dog and cow genomes. Sixty percent of these are included in our set (see Table S1). The remaining 40% were manually inspected and correspond mostly to false positives in the Larkin dataset due to assembly errors in the genome releases available at the time, and regions of complex history where the exact breakpoint location could not be determined, which we had eliminated from our analysis.

Compared to previous studies, our collection of breakpoints follows a chronological timeline in independent lineages, and thus includes cases where breakpoints re-occur in one given ancestral intergene but in different lineages. Therefore comparisons between rearranged and non-rearranged intergenes, conventionally used to identify genomic characteristics that correlate with one or the other category, are not adapted. Instead, we propose to model the breakage rate of ancestral intergenes using a generalized linear model by Poisson regression (Figure 1). This approach enables us to describe how the breakage rate varies with one or more genomic parameters, and allows us to distinguish the effects and contribution of each parameter to the model.

Characteristics of ancestral intergenes

The simplest null hypothesis to describe the distribution of breakpoints in a genome is that both their occurrence and evolutionary fixation are a random process, except in regions transcribed as protein coding genes that we shall consider as "rearrangement-free", because of the known deleterious effects of perturbing the structure of genes. Intergene length is the only parameter of this model, since breakpoint occurrence will follow a Poisson law: the probability of a breakpoint to occur between two adjacent genes increases proportionally to intergene length. To test this null hypothesis on our data, and any potential deviation from it, we must first estimate the length of ancestral intergenes. In addition to intergene length, a number of genomic properties are statistically correlated to the presence or absence of breakpoints in mammalian genomes. These include gene density, GC content, long-range regulation by distant non-coding elements or repeated sequences, CpG islands density, recombination rate, etc. Each of these properties is a potential parameter for the model, but not all properties can be reconstructed in the Boreoeutherian ancestor. For example, the half-life of transposable elements in a mammalian genome is too short to reliably estimate their content in Boreoeutheria. We thus adopted a practical approach consisting in first testing as many parameters as possible from an ancestral perspective, thus leading to a potentially incomplete model. In a second step, we tested the model on modern genomes where additional parameters can be measured, in order to determine whether the model is able to reproduce the specific genomic characteristics that are associated with breakpoint occurrences in real data.

To estimate the length of the ancestral intergenes, we first examined whether adjacent genes conserved in their ancestral configuration in multiple modern genomes are typically separated by intergenes of similar length, which would suggest that the intergene has little changed in size since the ancestor. The correlation of orthologous intergene lengths in modern genomes is in fact remarkably high: for most ancestral intergenes, modern lengths do not deviate much from a median value (Figure 4A). This is consistent with an evolutionary process that randomly inserts or deletes DNA elements

from an ancestral intergene, independently in each lineage, thus leading to a length distribution of modern lengths centred on the ancestral value. Alternatively, the result is also consistent with global intergene expansion (more insertions than deletions) or contraction (more deletions than insertions) on a genome scale independently in each lineage, where the distribution of modern values has a median that has remained proportional to the ancestral length. In any case, the R^2 coefficient of 0.86 for the intergene lengths distribution (in a log-log scale) shown in Figure 4A is most parsimoniously explained by a median value that is directly informative about the ancestral state. Indeed, if modern intergene lengths are instead randomly shuffled within their respective genomes so that orthologous intergenes have no more evolutionary link, then the same distribution shows a R^2 coefficient of 0.04 (Figure S2A). We conclude from this analysis that in most cases, the median value of modern orthologous intergene lengths is a reliable estimate of the ancestral length. Using this approach, we obtained an ancestral length estimate for 16,115 intergenes out of 18,436 (87.4%). The remaining 12.6% of intergenes showed too much variability between modern versions to reliably infer the ancestral length (see methods). Remarkably, the distribution of these ancestral intergene lengths is log-normal and very similar to that of a high-quality modern genome (Figure 4B). This shows that both the reconstruction of ancestral intergenes and the estimation of their lengths do not exclude specific categories of intergenes based on their length (for example, the longest ones, which may have been rearranged beyond reconstruction). The 16,115 intergenes with an estimated length include 682 intergenes with a future breakpoint (representing 90.8% of the breakpoints set), which are thus accessible to further investigation. We applied the same strategy to estimate the GC content of ancestral intergenes based on the global GC content of modern intergenes, excluding repeats. We could thus compute the ancestral GC content for 15,856 ancestral intergenes (86.0%; Figure S3).

The third ancestral genomic property for our analysis is the likelihood that negative selection acts against rearrangements to preserve gene organisation, especially with regard to their regulatory sequences. Evidence for such constraints has been provided between a number of highly regulated genes and their long-distance enhancers, resulting in so-called “genomic regulatory blocks” (GRBs) (Engstrom et al. 2007; Kikuta et al. 2007). However, beyond a few precise examples, whether such negative selection has a significant impact on the distribution of breakpoints at the scale of whole genomes has not yet been tested. Identifying regulatory sequences and deciphering their target genes is a difficult task and to date, no genome-wide map exists that links non-coding regulatory elements and genes. In order to include this parameter in our model, we used the density of conserved non-coding elements as a proxy for the probability that regulatory elements/gene relationships constrain a given intergene. Conserved non-coding elements often correspond to enhancers and more generally to transcription factor binding sites (Woolfe et al. 2005; Pennacchio et al. 2006; Kikuta et al. 2007), and targets of GRBs are often found in arrays of conserved non-coding elements (Nobrega et al. 2003; Sandelin et al. 2004; Becker and Lenhard 2007; Dong et al. 2009; Hufton et al. 2009). Conserved non-

coding elements are not the only type of regulatory elements that exist in mammalian genomes: in fact, the large majority of transcription factor binding sites are species-specific, redundant and have a high turn-over in genome evolution (Schmidt et al. 2010). However, regulatory elements with a short life span, high redundancy and plasticity are very unlikely to constrain gene order over long evolutionary times, unlike elements that have remained conserved in location and sequence. This suggests that the density of conserved non-coding elements may be used as a proxy for the presence of evolutionarily constrained regulatory relationships in a region of the genome. We used conserved non-coding elements (CNEs) identified by GERP (Cooper et al. 2005) from the alignment of 33 eutherian genomes available from Ensembl. Elements that are well conserved between boreoeutherian genomes are most likely ancestral. We inferred the ancestral CNE content in an ancestral intergene as the median of the total CNE length in modern intergenes, similarly to intergene length and GC content (see methods). All ancestral intergenes with an estimated length have a CNE length estimate, including 34% that do not contain conserved non-coding elements and an additional 12.5% that show high disparities in total CNE length amongst species, suggesting that the multiple alignment may be poor in some of these intergenic regions. However, we retained all estimates for the analysis, based on the observations that (a) the correlation between modern orthologous CNE lengths and their median value is very high over the entire set ($R^2 = 0.82$; Figure S4A), (b) removing intergenes with uncertain total CNE estimates tend to bias the dataset towards long intergenes, presumably by removing many short intergenes that truly do not contain CNEs, and (c) in any case, analyses carried out on the remaining 53.5% of intergenes where the estimate is most reliable gave very similar results to those on the entire set (data not shown).

Intergene length mostly determines breakage probability

In order to test if the three parameters estimated in the ancestor (intergene length, GC content, CNE content) are significantly correlated with the breakage process, we use Poisson regression, a generalized linear regression method that models the distribution of rare events (here, breakpoints) in a set of intervals (here, intergenes) according to characteristics of these intervals. Other multivariate regression methods, such as logistic regression, have been used in a similar context (Poyatos and Hurst 2007) but the classical Poisson regression method is more adapted to the present study (see discussion). Importantly, Poisson regression relies on the fact that, after a logarithmic transformation, most rates can be appropriately modelled as a weighted sum of explicative parameters with Poisson-distributed errors. Our null hypothesis is that breakpoints are distributed randomly, in direct proportion to the size of intergenes, with a proportionality coefficient equal to the average number of breakpoints per intergenic base pair (total number of breakpoints divided by the total intergenic length). To test

this, let r be the breakage rate (mean number of breakpoints per intergene), and L be the mean length of a class of intergenes. If breakage is random, we expect:

$$r = a \,.\, L \quad \leftrightarrow \quad log(r) = log(L) + a$$

Therefore, in a log-log representation, we expect the breakage rate per intergene to be a linear function of intergene length, with $x=y$. If other factors influence breakage, under the Poisson regression assumptions, $log(r)$ will typically be a linear function of both $log(L)$ and these parameters.

We performed the Poisson regression with breakage rate as a function of ancestral intergene length. We find that breakage rate is highly correlated with intergene length (Figure 5A), i.e. larger intergenes are more frequently rearranged than smaller ones. But the regression slope is not 1, as expected for a random distribution of same average breakage rate, but 0.28 so that the regression equation is:

$$log(r) = 0.28\ log(L) - 6.04 \quad \leftrightarrow \quad r = 2.4 \,.\, 10^{-3} \,.\, L^{\,0.28}$$

The striking linear correlation between intergene length and breakage rate is sufficient to explain most of the variation observed in breakage rates; the remaining variation may be attributed to statistical noise according to a Chi^2 test ($P = 0.19$; McFadden's pseudo $R^2 = 0.93$; Table 1). Despite the observed linearity, the distribution is far from random expectations: large Boreoeutheria intergenes are broken less frequently during evolution than expected at random, and smaller ones are broken more often. The increase in breakage rates thus appears slower than expected under a random distribution when considering the increase in intergene lengths. We tested if introducing the GC content of intergenes into the regression modifies the model. The regression shows that the GC content is not significantly affecting breakage probability (Table 1): the correlation between breakpoints and regions of high GC content is entirely imputable to the fact that short intergenes tend to be GC-rich while long intergenes tend to be AT-rich. This confirms a previous report, which showed that the increased frequency of breakpoints in regions of high GC content in mammalian genomes was a secondary consequence of the relative increase in GC content in short intergenes (Lemaitre et al. 2009). We ruled out multicollinearity issues due to the high correlation between GC content and intergene lengths: when divided in classes of similar GC content, the average length of Boreoeutheria intergenes affected by breakpoints is significantly longer than those without breakpoints (Figure S5A). In contrast, when dividing intergenes by classes of similar length, we find no difference in GC content between intergenes with or without breakpoints (Figure S5B).

Lastly, we tested whether the proportion of CNEs has significantly influenced breakage rates since Boreoeutheria. Each class of intergenes based on length was further divided into two groups: we

ranked intergenes based on their proportion of CNE (in percentage) and separated them into top 50% and lower 50%. Intergenes with a higher proportion of CNE consistently display a lower breakage rate then those with less conservation across all the range of intergene length classes, although none of the comparisons is statistically significant when taken individually (Figure 5B). When introduced in the regression, the proportion of CNE is found to have a very small but statistically significant effect on the breakage rate, increasing the McFadden's pseudo R^2 by 3% (Table 1). This effect, however, is almost negligible compared to the very strong correlation between breakage rate and intergene length.

Together, our results show that, after the Boreoeutheria ancestor, breakpoints eventually became fixed in the genome of modern species with a probability that depends almost exclusively on the length of the intergene where it occurred. Interestingly, breakpoints are behaving as a Poisson random variable from a mathematical point of view, suggesting that their distribution reflects a neutral, random process despite the fact that they do not follow the pattern expected under the classical random model. This simple result may appear to contradict many previous observations from the literature, where the distribution of breakpoints has convincingly been shown to correlate with parameters such as transposable elements density, CpG island density or segmental duplications. Yet we show next that these observations can be explained by the simple dependency on intergene length and proportion of CNE, while the contrary is not true.

Potential confounding factors

While intergene length is a strong predictor of breakage, the relationship between breakage and intergene length is not linear, as it would be in a random Poisson process. Instead, breakage probability depends on a root of intergene length, so that the correlation is linear after a logarithmic transformation. From a biological perspective this is not intuitive, and it is thus reasonable to ask whether the true biological determinant of breakpoints might be another genomic characteristic for which intergene length L would act as a proxy in our model. We therefore searched for candidate confounding variables that may promote breakage, but which might have been overlooked in our model because they could not be reconstructed and tested in the ancestral state. Since we find that breakage rates are proportional to $L^{0.28}$, appropriate candidates are features that behave similarly to $L^{0.28}$: they increase in absolute numbers when intergene lengths do, and yet their density within intergenes decreases. We used these simple characteristics as a screen for a number of genomic properties suggested to promote breakpoints in the literature. Any genomic feature that fails to satisfy either of these two conditions can be ruled out as an explicative variable.

Repeated sequences have often been invoked as a promoters of rearrangements through

illegitimate recombination (Gray 2000; Shaw and Lupski 2004; Liu et al. 2012), and breakpoints have especially been associated with high densities of SINE elements (Ma et al. 2006; Schibler et al. 2006; Carbone et al. 2009). However, in the human genome, the average interspersed repeat content in intergenes increases sharply for intergenes smaller than 10 kb and remains remarkably constant for intergenes larger than 10 kb (56% of intergenes, covering 98.9% of total intergene length; Figure 6A). In contrast, interspersed repeats density would be expected to decrease regularly if transposable elements were to explain the "slower than expected" increase in breakage rate. The local density in transposable elements is therefore not a candidate variable that could underlie the variation of breakage rate with intergene length. We repeated the same analysis with SINEs, LINEs, LTR and DNA transposons taken separately (Figure S6); none of them display the decrease in density expected if these features were to promote breakage and explain the relationship we observe. Apart from transposable elements, segmental duplications are another type of repeated sequences that have been positively associated with breakpoints (Bailey et al. 2004; Ma et al. 2006; Zhao and Bourque 2009). The proportion of sequence in a segmental duplication varies with intergene length, but in different ways across the range of lengths found in the human genome (Figure 6B): it increases in intergenes up to 100 kb and decreases afterwards. Breakage rates on the other hand do not change in behaviour in intergenes smaller or larger than 100 kb, so segmental duplications are not an appropriate candidate either to explain the breakpoints distribution.

Secondly, recombination rate might affect the probability of rearrangements, with regions of high recombination being more prone to errors during meiosis than those of low recombination (Volker et al. 2010). Previous evidence has shown that the recombination rate is only weakly correlated with local gene density, and mostly linked to sequence composition and especially GC content (Kong et al. 2002), which we ruled out as an explanative parameter for the breakage rate. We nevertheless investigated the possibility that the recombination rate might be the underlying parameter linking breakage rate and intergene length; as expected, we found no correlation between the mean recombination rate of intergenes and their lengths to support this hypothesis (Figure 6C).

Thirdly, we examined if replication origins, which are frequently found in gene-dense regions and are associated with marks of open chromatin (Mechali 2010), might promote breakage and thus be associated with breakpoints, as suggested in a previous study (Gordon et al. 2009). Using 874 predicted replication origins in human chromosomes (Huvet et al. 2007), we show that their distribution displays a remarkable linear correlation with intergene length, reminiscent of the breakpoints distribution (Figure 6D). Replication origins and timing are a conserved feature in mammalian genomes (Ryba et al. 2010), allowing us to map replication origins to the ancestral Boreoeutheria intergenes (although their relatively small number precludes their use in the multivariate regression analysis). The overlap between breakpoints and replication origins is higher

than expected at random (Fisher's exact test: $P = 0.01$), as a likely consequence of their common enrichment in short intergenes. But this excess disappears when intergenes of homogeneous length are considered (intergenes < 20 kb: $P = 0.29$; intergenes > 20 kb: $P = 0.13$). There is therefore no evidence that breakpoints and replication origins are dependent of each other.

Lastly, recent findings have suggested that breakage probability might be linked to chromatin state in the germline (Lemaitre et al. 2009; Li et al. 2012). DNA in open chromatin regions may be more accessible and susceptible to double-strand breaks that in turn can give rise to rearrangements if misrepaired, while compacted chromatin might shield DNA from breakage. To investigate chromatin compaction, we used experimental evidence of open chromatin regions (see Material and Methods) obtained in human embryonic stem cells, within the framework of the ENCODE Consortium. Interestingly, the proportion of open chromatin per intergene decreases with intergene length, and could potentially explain the distribution of breakpoints since Boreoeutheria (Figure 6E). When studied in more details (Figure S7), the total length of DNA in an open chromatin state is increasing with intergene length L, but not as fast. This property is similar to breakage rate. The average increase is also linear in a logarithmic representation, with a slope smaller than 1. However, the overall correlation is rather poor, and when linear regression is performed, the amount of open chromatin per intergene is found to be proportional to $L^{0.59}$, rather than $L^{0.29}$ like the breakage rate. Whether this is because the open chromatin patterns in embryonic stem cells are different from those of germline cells (which are not available to date for investigation), or because open chromatin is not the sole or direct cause of the variations of breakage rate, remains an open question. However, chromatin compaction is the only genomic parameter with a pattern similar to that of breakage rate, suggesting that the variations in breakpoints density observed across the genome might in fact reflect DNA accessibility linked to genome organization in the nucleus.

Although we have not been exhaustive in testing every genomic parameter and its potential correlation with intergene length, we show here that three main features proposed to explain the non-random occurrence of breakpoints (repeated sequences, recombination rates, replication origins) do not satisfy the necessary conditions to account for our results. On the other hand, we show next that our model based on intergene length is sufficient to reproduce *in silico* a variety of correlations between breakpoints and genomic features reported in the literature.

Model-based simulated breakpoints behave as real breakpoints

Previous studies have conventionally compared the properties of "breakpoint regions", i.e. genomic windows of fixed size around breakpoints, to randomly sampled windows in the genome to

assess whether breakpoints occur in statistically biased regions. We simulated evolution by distributing "breakpoints" in the human genome according to our regression model, and tested if we can reproduce such observations. We collected 100 kb windows centred on the simulated breakpoints and compared them to randomly sampled 100 kb windows (controls). One set of control windows is entirely random: windows can be centred on any base of the genome. In a second control set, windows are centred on a random intergenic base (excluding protein coding transcribed regions), to ensure that any differences seen between the first control and the simulations are not due to the fact that simulated breakpoints can only be intergenic. Simulations were run a hundred times each.

We find that simulated breakpoints are associated with an average 2-fold increase in gene density compared to random controls (Wilcoxon's test: all simulations with $P < 1.10^{-13}$; Figure 7A), consistent with their higher occurrence in small intergenes. Simulated breakpoints are also associated with a 2-fold increase in CpG islands density (all simulations with $P < 2.10^{-8}$; Figure 7B). All simulations show that breakpoints are significantly associated with higher SINEs density, with an average of 16.7% compared to 12.8% and 11.0% in controls (random and random-intergenic respectively; Figure 7C). Ninety six per cent of simulations report a significant association with segmental duplications, with an average density of 6.9% versus 4.4% and 4.9% in each control respectively (Figure 7D). Simulations therefore recapitulate observations that appear recurrently in the literature as characteristics of breakpoint regions. LINE elements, however, are depleted in the vicinity of non-random breakpoints compared to random expectations (Figure 7E), as reported by (Carbone et al. 2009) but in disagreement with (Zhao and Bourque 2009). More interestingly, the order of magnitude of the enrichments and depletions in simulations are similar to those observed previously with breakpoint data in the literature (Ma et al. 2006; Gordon et al. 2007; Larkin et al. 2009), except for segmental duplications, which are less enriched than in several studies (Ma et al. 2006; Zhao and Bourque 2009). This is expected because segmental duplications may be generated during rearrangement events rather than causing them (Bailey et al. 2004; Bailey and Eichler 2006; Ranz et al. 2007; Girirajan et al. 2009). As our simulations do not model the formation of duplications at breakpoints, simulated breakpoints appear thus more weakly associated with segmental duplications than real data. Finally, the simulated breakpoints occur more clustered than expected at random, and define a higher number of short synteny blocks than in either of the two random controls (Figure 7F). The excess of short synteny blocks, either theoretically predicted (Pevzner and Tesler 2003) or observed in genome comparisons (Kent et al. 2003; Bourque et al. 2004; Zhao et al. 2004), is one of the major arguments sustaining the fragile breakage model. Together, our simulations demonstrate that the distribution of breakpoints based on intergene length and, marginally, on CNE proportion, is sufficient to reproduce the observations published in previous studies.

Deviations from random expectations are not due to paired inversion breakpoints

While our model accounts for the association of breakpoints with various genomic features, the reason why the intergenic breakage rate does not increase as fast as intergene length remains largely unexplained. The majority of genomic rearrangements in mammalian genomes are intra-chromosome inversions, where two breakpoints occur on the same chromosome (Bourque et al. 2004; Zhao et al. 2004). The dependency, if any, between the two breakpoints is not accounted for in our model. In fact, since our breakpoints can only be identified if they modify the order of genes, it is plausible that the observed excess of breakpoints in short intergenes may be the consequence of our ability to better identify them in gene dense regions. Indeed, even small inversions will be detected in a gene dense region of the genome, where intergenes are short, while inversions occurring within a large intergene will be missed. The bias may be particularly severe if inversions of short genomic segments are much more frequent than longer ones. Until now, most studies aimed at understanding the properties of breakpoint regions have ignored this, instead considering breakpoints as independent events in part because the distribution of inversion sizes is unknown.

Here, we test this dependency by simulating inversions as pairs of breakpoints in the human genome, and we measure how the resulting breakage rate correlates with intergene length. To estimate the distance between breakpoints (i.e. the length of inverted fragments) we make two assumptions. First, we posit that a rearrangement requires a physical contact between two distant regions of the same or different chromosomes in the nucleus. Second, repeated sequences such as transposable elements may in some cases facilitate the pairing of non-homologous DNA strands, which may in turn promote rearrangements. We showed previously that TEs alone cannot explain the distribution of breakpoints when they are considered independently (Figure 6A), but we test this hypothesis in the context of paired breakpoints. To estimate the length of inverted regions, we use results from a recent genome-wide study that applied the Hi-C technique to map contact points between intra and inter-chromosome DNA molecules during interphase (Lieberman-Aiden et al. 2009). The map shows that the probability of contact between loci on a chromosome decreases regularly as the genomic distance between the loci increases. Thus if contacts are necessary to rearrangements, inversions will typically be short.

We carried out simulations of inversions on the human genome by selecting a first breakpoint at random (outside genes) and a second from the distribution of contact probability according to distance in the Hi-C map of the human genome (Lieberman-Aiden et al. 2009). Two cases may arise: if the interval between the two breakpoints encompasses one gene or more (and may thus be detected) the rearrangement is retained, but if the two breakpoints occur within the same intergene (and thus cannot be identified) the inversion is not recorded. Simulations proceed until we obtain the same number of

visible simulated breakpoints as observed in our real dataset in the five lineages after Boreoeutheria. If the simulations recapitulate well true rearrangements and our ability to detect them, then a log-log correlation of breakage rate and intergene length should be linear with a slope close to 0.30 (Figure 5A). Results show that breakage rates does remain a linear function of intergene length in a logarithmic representation, but the slope of 0.74 (Figure 8A) is far from explaining the correlation observed between intergene length and breakage rate in the observed data, where breakage in short intergenes is much more frequent. We integrated the influence of non-homologous recombination due to transposable elements (TEs) in the simulations by requesting that a pair of breakpoints is accepted only when both occur in elements of the same class or type (see methods). The new linear correlation between breakage rates and intergene length displays a similar slope that remains much higher than in the observed data (slope = 0.75 with TEs of the same class; slope = 0.88 for TEs of the same type; Figure 8B).

Altogether, results show that chromosome-wide contacts in the nucleus and the pairing of breakpoints alone are not sufficient to explain the distribution of breakpoints in intergenes in a simple manner, although they may contribute (see discussion).

Distribution of breakpoints in an ancestral yeast genome

An interesting consequence of the model is that if the distribution of breakpoints follows a single distribution law across the genome, it may be due to a genome-wide form of organization of genes, chromatin or chromosome, and thus may be a general feature of eukaryote genomes. We tested this prediction in yeasts, an eukaryotic group very distant from mammals for which a wealth of genomic data is available. In fact, previous evidence on conservation of gene order in yeasts showed that the probability for two genes to be retained as neighbours in two species was weakly negatively correlated with intergene length (Poyatos and Hurst 2007). Using the same method as in mammals, we reconstructed the gene order for the last common ancestor of *Kluyveromyces* and *Lachancea* yeasts based on the genome annotations available from the Genolevures database (Sherman et al. 2009), and identified 505 rearrangement breakpoints in the lineages of *Lachancea kluyveri, Lachancea waltii,* and *Kluyveromyces lactis*, a finding consistent with (Gordon et al. 2009). The ancestral intergene lengths were estimated (Figure S8) and Poisson regression was carried out as described for mammals.

Strikingly, like in mammals, a logarithmic representation shows that the breakage rate is linearly correlated with intergene length (Figure 9). The regression equation obtained is:

$$log(r) = 0.43 \,.\, log(L) - 4.90 \quad \leftrightarrow \quad r = 7.5 \,.\, 10^{-3} \,.\, L^{0.43}$$

The model fits the data remarkably well, since intergenic length alone is sufficient to explain the distribution of breakpoints across the genome (Chi^2 test, $P = 0.14$; McFadden's pseudo-$R^2 = 0.81$). Of note, this model is not directly consistent with a previous modelling of breakage occurrence in yeast genomes, which concluded that intergene length is only weakly correlated to breakage probability (Poyatos and Hurst 2007) (see discussion). We do not find evidence that an additional parameter, such as cis-regulation, is necessary to explain the distribution of breakpoints in yeasts. However, intergene lengths in yeasts span a much narrower range compared to mammals, leading to a regression with less resolving power and thus to wider error ranges in the model, which hinders our ability to identify small-effect explanatory variables. In both cases, we do not exclude that additional parameters affect breakage, but their effect would be so subtle that it cannot be detected at this resolution and evolutionary depth.

Discussion

We developed here the first mathematical model to accurately describe the distribution of rearrangement breakpoints in eukaryotic genomes. Previous studies aimed at characterizing breakpoint regions have recurrently reported that breakpoints associate with regions of high gene density, high GC content, high segmental duplication content, etc., without being able to distinguish between true determinants of breakage and secondary correlations. Here we show that all these characteristics previously found to be associated with breakpoints can be explained by a simple relationship between breakage rate and intergene length. The model describes the probability of breakage fixation during evolution as a Poisson process, proportional to a root of intergene length. The model is likely to be universal in eukaryotes, because it applies in mammals as well as in single-celled yeasts.

The model is based on Poisson regression, a method unused until now for the analysis of evolutionary breakpoints. Gene-to-gene adjacencies have been commonly used in combinatorial studies that reconstruct the most likely rearrangement scenario to transform one genome into another (Bourque et al. 2005; Peng et al. 2006). However, such intervals have usually been disregarded when studying the features of breakpoint regions, in favour of either windows spanning larger genomic regions (Ma et al. 2006; Zhao and Bourque 2009), or the narrowest sequence frame that cannot be aligned between species (Lemaitre et al. 2009), none of which correspond to a biological reality. In the light of our results, gene-to-gene intervals are a biologically relevant scale to study the distribution of breakpoints. Poisson regression is particularly suitable to handle such data and disentangle the contribution of different genomic features to the breakage probability. Logistic regression, another generalized linear model for multivariate regression, has previously been proposed to study breakpoints in a similar context (Poyatos and Hurst 2007). This regression method deals with binary

variables and was used to model the presence or absence of orthologous intergenes in pairwise genome comparisons. In contrast, Poisson regression models the rate of occurrence of rare events in intervals, and is more suited to large datasets of rearrangement breakpoints in multiple lineages. Firstly, it can take into account independent breakpoint reuses in different lineages. Secondly, and unlike logistic regression, it proposes a straightforward expectation for the classical random model that provides clues on how to handle the genomic features data. Under a Poisson model, the logarithmic transformation on intergene length – necessary to highlight its link with breakage – is rational in the regression framework. However, this transformation is not intuitive from a biological perspective, which perhaps explains why previous studies missed the near-perfect correlation between breakage rates and intergene lengths uncovered in this work. Poisson models based on the gene/intergene structure of the genome seem especially powerful to model rare, punctual events from a different perspective, and the approach we propose here could be suitable for a wide range of applications, from the study of breakpoints distribution in other contexts (cancer, genetic variation, etc) to transposable elements insertions events and other mutational processes.

Importantly, the model demonstrates that selection to preserve gene organization plays a minor part in breakage fixation in mammals and is negligible in yeasts. Indeed, breakage probability is mainly governed by the amount of non-coding DNA between genes that can accommodate for breakage. The lengths of intergenic spacers, and more generally the amount of non-coding DNA in genomes, are thought to evolve neutrally and result from a balance between random insertions (duplications, transposable elements) and deletions (mainly due to recombination)(Petrov et al. 2000; Nam and Ellegren 2012). Breakage probability is thus mostly dependent of a neutrally evolving parameter linked to the genome structure. On the other hand, it is only marginally affected by the presence of conserved regulatory elements, which have been previously hypothesized to exert important selective pressure to preserve synteny (Kikuta et al. 2007; Hufton et al. 2009; Mongin et al. 2009). Invoking such selective constraints on specific gene arrangements and cis-regulatory interactions is in fact not necessary to explain the vast majority of breakpoints (or absence of breakpoints) in any specific genomic region. Furthermore, the model is not compatible with the widely held view that transposable elements are the main cause for genomic rearrangements. Non-homologous recombination between transposable elements may punctually facilitate rearrangements, but they cannot explain alone the length-dependent distribution of breakpoints we see here. Instead, the non-homogeneous abundance of transposable elements in mammalian gene-dense and gene-poor regions results in secondary correlations between repeats distributions and breakpoints. Finally, the model does not require the existence of so-called “fragile regions” in the genome. Short intergenes do indeed break more often than expected at random, and long intergenes less often, but in the context of a uniform distribution across all intergenes, not of specifically “fragile” or “robust” intergenes.

Instead of a specific causative agent for breakpoint occurrence or fixation, our results are consistent with the emerging view that the occurrence of breakage events is linked to the structural organization of the DNA in the nucleus, rather than to its sequence characteristics. Amongst the tested parameters to explain the intergene-length dependent variations of the breakage rate, only chromatin state displays a similar behaviour to breakage rate, as the intergenic amount of open chromatin increases on average with a root of intergene length. The higher-order organisation of the genome is poorly known to date, but it is thought to be in relation with transcription, to repress transcription of non-coding DNA and transposable elements or to accommodate the clustering of active genes in transcription factories (Osborne et al. 2004). The idea that chromatin compaction states influence mutational processes such as double-strand break susceptibility or the efficiency of base mutation repair has recently been put forward in other contexts by several reports (Lin et al. 2009; Mathas and Misteli 2009; Li et al. 2012; Schuster-Bockler and Lehner 2012). Higher-order organisation of chromatin or chromosomes in the nucleus would be a plausible explanation for our results, as non-coding regions in a functional accessible (i.e. breakable) configuration might be dense around genes, and might become rarer when distance from a gene increases. Interestingly, we show that the distribution of replication origins, while not correlated with breakpoint occurrence, displays a relationship with intergene length similar to that of breakpoints. It is thus conceivable that both processes (rearrangements and replication initiation) are dependent on the same organisational context in the genome. This is consistent with recent evidence showing that the distribution of replication origins is linked to the 3D architecture of the genome (Duan et al. 2010; Ryba et al. 2010). If so, and except for the strong negative selection across genic regions of the genome, the occurrence and fixation of breaks remains an evolutionary neutral process. Of note, the exact value of the exponential factor of the intergene length involved in the regression equation is different between yeasts (0.43) and mammals (0.30). The reason for this is unclear, but is consistent with the observation that intergenes are on average much shorter in yeasts than in mammals, which might result in a slightly different higher-order organization or level of compaction of the chromatin.

An interesting aspect of the model that we propose here resides in its predictive value. A tempting strategy to exploit the recent accumulation of genomic data is to use rearrangement breakpoints, or lack thereof, to map regions where the gene organisation is under selective constraint. Our results show that the breakage probability of a given intergene can be deduced directly from its length, with small-scale deviations that may be attributed to selection. These small-scale deviations, however, will become more prominent when a large number of genomes (and, therefore, a large number of breakpoints) is considered, so that it will become possible to assess whether each individual intergene has been broken significantly less often than expected according to its length. Considering that the breakage probability of large intergenes (> 100 kb) is about 10% when breakpoints from five independent boreoeutherian lineages are taken into account, we estimate that one hundred amniote

genomes well chosen into the phylogenetic tree should give sufficient statistical power to provide information on a large part of the genome. This is less than twice as many as have been sequenced to date, and it is thus likely that genomes appropriate for such a study will be available in the near future. By providing new insight into the background distribution of evolutionary breakpoints and elucidating its relationship with genome organisation, these results provide a statistical framework to map regions in mammalian genomes where the order of genes is under functional constraints.

The study that we present here has been carried out on evolutionary rearrangement breakpoints. Previous studies have shown that common fragile sites and breakpoints recurrently seen in cancer cells tend to correlate with regions of evolutionary rearrangement breakpoints (Murphy et al. 2005; Darai-Ramqvist et al. 2008; Fungtammasan et al. 2012). It would be of great interest to see whether the breakpoints that occur during the lifetime of organisms are distributed similarly to evolutionary breakpoints, which would explain why evolutionary and cancer breakpoints tend to overlap more often than expected. This is not obvious, however, because evolutionary breakpoints occur in germinal cells while the breakpoints occurring in healthy and cancerous tissue are somatic; if breakage probability is indeed linked to the 3D structure of the genome, this structure may be different in germinal cells, for reasons of compaction (spermatozoa) or cell physiology (ovocytes and first-stages embryo). Breakpoint datasets from sequenced somatic genomes available today are very biased towards either cancer genomes or genetically transmitted diseases. Healthy genomes, when they are sequenced, are usually not assembled de novo but mapped to the reference genome, which hinders the detection of breakpoints that do not affect cell physiology. Recent advances in sequencing technologies, however, are starting to provide a wealth of personal genome data that will allow the identification of unbiased neutral breakpoints in human genomes. It will soon become possible to test whether evolutionary breakpoints and somatic breakpoints occur in a similar fashion and, were this true, if deviations from the expectations can be used in somatic cells as well to map genome regions where breakpoints result in changes in cell physiology.

Materials and Methods

Reconstruction of ancestral gene adjacencies and estimation of ancestral characteristics

Information on gene trees and gene order were downloaded from Ensembl v.57 for all available genomes (51 species) and used to reconstruct the gene order and orientation at the Boreoeutheria node as described below. For yeasts, the gene order information was obtained from Genolevures for 11 species (Sherman et al. 2009); gene trees were built using TreeBest (Vilella et al. 2009). The reconstruction method used for the ancestral genomes computes all pairwise comparisons of genomes

that are informative for the ancestor of interest (the ancestor is on the pathway between both species in the tree). Pairs of genes that are next to each other and in the same orientation in two genomes or more are considered as potentially inherited from the ancestral genome. When ancestral genes are involved in several conflicting adjacencies so that the ancestral version is not directly deducible, a simple algorithm is applied to select the most likely candidate based on the highest frequency of retention in modern genomes. The method to identify ancestral gene adjacencies is formally described in Supplementary Text S1. We obtained 18,436 gene adjacencies for the Boreoeutheria ancestor, and 4,608 adjacencies for the ancestor of *Kluyveromyces* and *Lachancea* yeasts..

The length, GC content and total conserved non-coding sequence of ancestral intergenes were estimated by computing a table of their values in all sequenced modern descendants of the ancestor of interest (28 species for mammals, 5 for yeasts). The median modern value was used as an estimate of the ancestral value. For the Boreoeutheria ancestor, we retained ancestral estimates only when they are supported by at least two genomes sequenced with a coverage of 6x or more in different clades (primates, rodents and laurasiatherians), to ensure that the estimate is truly ancestral. Intergenes for which the scatter of modern values was too high were not considered in the analysis, as the ancestral estimate was then deemed unreliable; we chose as a cutoff to retain only intergenes for which the interquartile range (range spanned by the 50% of values closest to the median) is no larger than 1.5 times the median value for the intergene length, and within 10% of the median value for GC content. It should be noted that these cutoffs are not very stringent: their purpose is to remove intergenes for which modern values are very inconsistent, most of the time due to assembly errors or missing data in the genomes. However, as reported in the results, the scatter of the modern values of length and GC content in intergenes retained after filtering is generally small; more stringent cutoffs therefore affect little which intergenes are filtered out.

Identification of evolutionary rearrangement breakpoints

Using the reconstructed ancestral gene adjacencies, we concatenated the ancestral genes into blocks of consecutive genes. The ancestral gene order was then compared to the gene order in each of the five modern genomes under study (human mouse, dog, cow and horse) to identify regions in the ancestral genome where the gene order has been modified in subsequent evolution. Such regions are identified where genes have different neighbours in the modern genome compared to their ancestral counterparts, so that the ancestral intergenes have been lost. The regions may be individual ancestral intergenes, or groups of consecutive ancestral intergenes that no longer exist in the modern genome. In order to exclude regions where the perturbation of gene order is due solely to gain or loss of genes rather than a chromosomal rearrangement, we reduced the ancestral and modern genomes to the order

of the genes present in 1-to-1 copies in both genomes (excluding new genes, lost genes and duplications). Regions where the gene adjacencies remain different between the ancestral and modern genomes contain a breakpoint. Breakpoints, however, could not be mapped to a precise intergene when the region was affected both by a breakpoint and gene gain/loss events (see Figure S1). As it is not possible to date the gain/loss event relatively to the breakpoint, we cannot estimate which was the relevant ancestral state when the rearrangement occurred. Such breakpoints, although probably real, were rejected as they cannot be processed by the method we propose here. The complete list of breakpoints is provided as Supplementary Material.

Dubious rearrangement events were removed from the set when they meet one of the following criteria: (a) breakpoints occurring in ancestral intergenes that have no outgroup support (dubious ancestral state); (b) breakpoints occurring in ancestral intergenes with poor modern support (no support from a panel of reference high-quality modern genomes across Boreoeutherian clades); and (c) breakpoints in two consecutive intergenes corresponding to a number of single-gene events, as manual inspection of randomly chosen breakpoints showed that in the majority of cases, these are probably misplaced genes due to annotation errors (short monoexonic gene, spurious exon annotated several tens of kb upstream compared to other species, etc) or assembly errors, especially in the cow genome (e.g. a solitary gene on a contig bordered by two large gaps, etc). The filtering process is fully described in Supplementary Text0 S2.

The breakpoints set was compared to the breakpoint regions in (Larkin et al. 2009). The Larkin breakpoint regions (LBR) correspond to coordinates in the human genome (hg18) showing a discontinuity with other genomes. We performed a liftover to the hg19 version of the human genome using Galaxy (Giardine et al. 2005) for all LBR found in the human, cow, mouse and dog genomes (see Table S1 for numbers). For each LBR, we collected the human genes found within the region (plus two genes up- and downstream, to accommodate for differences in gene limits between versions of the human genome), and tested whether the ancestral copies of these genes include the borders of a breakpoint from our set in the lineage where the LBR was found. When such an overlap between breakpoints was found, we considered that the breakpoint regions are the same. Of note, the LBR are relatively large regions of the human genome (mean: 700 kb, 9.7 genes), and several LBR overlap more than one breakpoint in our finer-scale set.

Poisson regression analysis

Throughout the study, and for the sake of simplicity, we refer to the expected mean breakage rate R for a given class of intergenes x (rigorously, $E(R|x)$) as the "breakage rate", noted r. The multivariate

regression analysis was carried out in R (http://www.R-project.org/) using the generalized linear models implemented in the *glm()* function. Intergenes were divided into classes of similar length (bins of width 0.5 in log scale), then further into classes of GC content (bins of 0.2) or into top 50% and lower 50% according to the proportion of conserved non-coding elements. The mean value of each parameter was used as the predictor value for the entire class of intergenes in the regression. A stepwise regression procedure was carried out to progressively add new variables in the model, in an order based on their initial performance in explaining the data (intergene length, then GC content or proportion of CNEs). The goodness of fit of each step of the model was estimated by performing a Chi^2 test on the residual deviance and degrees of freedom of the model, which corresponds to comparing the model to a saturated model with no deviance and no degrees of freedom (perfect fit). When the test is not significant ($P > 0.05$), then variations between the model and the data can be attributed to statistical noise. A new parameter was retained in the model when a Chi^2 test on the difference of residual deviances with and without the parameter (with one degree of freedom) was significant. Of note, this was always in agreement with Akaike's Information Criterion, so no issues of over fitting arose.

Simulations of independent and dependent breakage

All simulations were performed using custom-made Python scripts. Independent breakpoint distributions were performed by calculating the expected breakage probability of each intergene in the human genome according to its length and proportion of CNEs, using the regression equation of the model. A set of intergenes was then drawn according to this probability, with replacement. Windows of size 100 kb centred on these intergenes were sampled for statistical analysis and compared to (a) a control set of 100 kb windows randomly sampled in the human genome (same probability of breakage for each base of the genome), and (b) a control set of 100 kb windows centred on intergenes drawn with a probability proportional to their length (same probability of breakage for each intergenic base of the genome, excluding genes). These steps were repeated a hundred times to obtain a hundred comparisons of three datasets. Statistical enrichment and depletion analysis for particular features were performed in R, using sequence feature tables downloaded from the UCSC Genome Browser (Fujita et al. 2011), which were also used to measure statistical correlations between intergene lengths and contents. The replication origins location data is an updated version of the dataset published in (Huvet et al. 2007) communicated by the authors.

Dependent breakpoints were drawn in two steps: an intergenic base is drawn randomly in the human genome and constitutes the first breakpoint. Then the distance d to the second breakpoint is drawn from the probability distribution derived from (Lieberman-Aiden et al. 2009) that describes the

probability of contact of two loci according to their distance. If the second breakpoint is in the same intergene as the first, the breakpoints are discarded. If the space between both breakpoints encompasses at least one gene, the breakpoints are recorded, and so on until the same number of breakpoints is obtained as observed between Boreoeutheria and the five lineages under study (682 breakpoints with an estimated ancestral intergenic length). This simulation was performed a hundred times, and the average breakage rate per class of intergene length was calculated and compared to the observations in the data. Alternatively, to test for a possible effect of non-homologous recombination due to transposable elements, a condition was applied during the simulation to record breakpoints only when the two breakpoints were drawn in TEs of the same class (SINEs, LINEs, LTR, DNA) or when they were drawn in TEs strictly of the same type (AluY, MIRb, L1M4 and so forth).

Acknowledgements

We thank Alexandra Louis and Pierre Vincens for help with computing resources. We thank Claude Thermes for sharing updated data on predicted human replication origins location.

References

Alekseyev MA, Pevzner PA. 2007. Are there rearrangement hotspots in the human genome? *PLoS computational biology* **3**(11): e209.

Bailey JA, Baertsch R, Kent WJ, Haussler D, Eichler EE. 2004. Hotspots of mammalian chromosomal evolution. *Genome biology* **5**(4): R23.

Bailey JA, Eichler EE. 2006. Primate segmental duplications: crucibles of evolution, diversity and disease. *Nature reviews Genetics* **7**(7): 552-564.

Becker TS, Lenhard B. 2007. The random vs. fragile breakage model of chromosome evolution: a matter of resolution. *Molecular Genetics and Genomics* **in press.** .

Blanchette M, Green ED, Miller W, Haussler D. 2004. Reconstructing large regions of an ancestral mammalian genome in silico. *Genome research* **14**(12): 2412-2423.

Bourque G, Pevzner PA, Tesler G. 2004. Reconstructing the genomic architecture of ancestral mammals: lessons from human, mouse, and rat genomes. *Genome research* **14**(4): 507-516.

Bourque G, Zdobnov EM, Bork P, Pevzner PA, Tesler G. 2005. Comparative architectures of mammalian and chicken genomes reveal highly variable rates of genomic rearrangements across different lineages. *Genome research* **15**(1): 98-110.

Carbone L, Harris RA, Vessere GM, Mootnick AR, Humphray S, Rogers J, Kim SK, Wall JD, Martin D, Jurka J et al. 2009. Evolutionary breakpoints in the gibbon suggest association between cytosine methylation and karyotype evolution. *PLoS genetics* **5**(6): e1000538.

Chauve C, Tannier E. 2008. A methodological framework for the reconstruction of contiguous regions of ancestral genomes and its application to Mammalian genomes. *PLoS Comp Biol* **4**(11): e1000234.

Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A. 2005. Distribution and intensity of constraint in mammalian genomic sequence. *Genome research* **15**(7): 901-913.

Darai-Ramqvist E, Sandlund A, Muller S, Klein G, Imreh S, Kost-Alimova M. 2008. Segmental duplications and evolutionary plasticity at tumor chromosome break-prone regions. *Genome*

research **18**(3): 370-379.
Dong X, Fredman D, Lenhard B. 2009. Synorth: exploring the evolution of synteny and long-range regulatory interactions in vertebrate genomes. *Genome biology* **10**(8): R86.
Duan Z, Andronescu M, Schutz K, McIlwain S, Kim YJ, Lee C, Shendure J, Fields S, Blau CA, Noble WS. 2010. A three-dimensional model of the yeast genome. *Nature* **465**(7296): 363-367.
Engstrom PG, Ho Sui SJ, Drivenes O, Becker TS, Lenhard B. 2007. Genomic regulatory blocks underlie extensive microsynteny conservation in insects. *Genome research* **17**(12): 1898-1908.
Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, Cline MS, Goldman M, Barber GP, Clawson H, Coelho A et al. 2011. The UCSC Genome Browser database: update 2011. *Nucleic acids research* **39**(Database issue): D876-882.
Fungtammasan A, Walsh E, Chiaromonte F, Eckert KA, Makova KD. 2012. A genome-wide analysis of common fragile sites: What features determine chromosomal instability in the human genome? *Genome research*.
Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J et al. 2005. Galaxy: a platform for interactive large-scale genome analysis. *Genome research* **15**(10): 1451-1455.
Girirajan S, Chen L, Graves T, Marques-Bonet T, Ventura M, Fronick C, Fulton L, Rocchi M, Fulton RS, Wilson RK et al. 2009. Sequencing human-gibbon breakpoints of synteny reveals mosaic new insertions at rearrangement sites. *Genome research* **19**(2): 178-190.
Goode DK, Snell P, Smith SF, Cooke JE, Elgar G. 2005. Highly conserved regulatory elements around the SHH gene may contribute to the maintenance of conserved synteny across human chromosome 7q36.3. *Genomics* **86**(2): 172-181.
Gordon JL, Byrne KP, Wolfe KH. 2009. Additions, losses, and rearrangements on the evolutionary route from a reconstructed ancestor to the modern Saccharomyces cerevisiae genome. *PLoS genetics* **5**(5): e1000485.
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 research* **17**(11): 1603-1613.
Gray YH. 2000. It takes two transposons to tango: transposable-element-mediated chromosomal rearrangements. *Trends in genetics : TIG* **16**(10): 461-468.
Hufton AL, Mathia S, Braun H, Georgi U, Lehrach H, Vingron M, Poustka AJ, Panopoulou G. 2009. Deeply conserved chordate noncoding sequences preserve genome synteny but do not drive gene duplicate retention. *Genome research* **19**(11): 2036-2051.
Huvet M, Nicolay S, Touchon M, Audit B, d'Aubenton-Carafa Y, Arneodo A, Thermes C. 2007. Human gene organization driven by the coordination of replication and transcription. *Genome research* **17**(9): 1278-1285.
Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D. 2003. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proceedings of the National Academy of Sciences of the United States of America* **100**(20): 11484-11489.
Kidd JM, Graves T, Newman TL, Fulton R, Hayden HS, Malig M, Kallicki J, Kaul R, Wilson RK, Eichler EE. 2010. A human genome structural variation sequencing resource reveals insights into mutational mechanisms. *Cell* **143**(5): 837-847.
Kikuta H, Laplante M, Navratilova P, Komisarczuk AZ, Engstrom PG, Fredman D, Akalin A, Caccamo M, Sealy I, Howe K et al. 2007. Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome research*.
Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G et al. 2002. A high-resolution recombination map of the human genome. *Nature genetics* **31**(3): 241-247.
Larkin DM, Pape G, Donthu R, Auvil L, Welge M, Lewin HA. 2009. Breakpoint regions and homologous synteny blocks in chromosomes have different evolutionary histories. *Genome research* **19**(5): 770-777.
Lee JA, Carvalho CM, Lupski JR. 2007. A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell* **131**(7): 1235-1247.

Lemaitre C, Zaghloul L, Sagot MF, Gautier C, Arneodo A, Tannier E, Audit B. 2009. Analysis of fine-scale mammalian evolutionary breakpoints provides new insight into their relation to genome organisation. *BMC genomics* **10**: 335.

Li J, Harris RA, Cheung SW, Coarfa C, Jeong M, Goodell MA, White LD, Patel A, Kang SH, Shaw C et al. 2012. Genomic hypomethylation in the human germline associates with selective structural mutability in the human genome. *PLoS genetics* **8**(5): e1002692.

Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**(5950): 289-293.

Lin C, Yang L, Tanasa B, Hutt K, Ju BG, Ohgi K, Zhang J, Rose DW, Fu XD, Glass CK et al. 2009. Nuclear receptor-induced chromosomal proximity and DNA breaks underlie specific translocations in cancer. *Cell* **139**(6): 1069-1083.

Liu P, Carvalho CM, Hastings P, Lupski JR. 2012. Mechanisms for recurrent and complex human genomic rearrangements. *Current opinion in genetics & development*.

Ma J, Zhang L, Suh BB, Raney BJ, Burhans RC, Kent WJ, Blanchette M, Haussler D, Miller W. 2006. Reconstructing contiguous regions of an ancestral genome. *Genome research* **16**(12): 1557-1565.

Mathas S, Misteli T. 2009. The dangers of transcription. *Cell* **139**(6): 1047-1049.

Mechali M. 2010. Eukaryotic DNA replication origins: many choices for appropriate answers. *Nature reviews Molecular cell biology* **11**(10): 728-738.

Molaro A, Hodges E, Fang F, Song Q, McCombie WR, Hannon GJ, Smith AD. 2011. Sperm methylation profiles reveal features of epigenetic inheritance and evolution in primates. *Cell* **146**(6): 1029-1041.

Mongin E, Dewar K, Blanchette M. 2009. Long-range regulation is a major driving force in maintaining genome integrity. *BMC evolutionary biology* **9**: 203.

-. 2011. Mapping association between long-range cis-regulatory regions and their target genes using synteny. *Journal of computational biology : a journal of computational molecular cell biology* **18**(9): 1115-1130.

Murphy WJ, Larkin DM, Everts-van der Wind A, Bourque G, Tesler G, Auvil L, Beever JE, Chowdhary BP, Galibert F, Gatzke L et al. 2005. Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science* **309**(5734): 613-617.

Nadeau JH, Sankoff D. 1998. Counting on comparative maps. *Trends in genetics : TIG* **14**(12): 495-501.

Nadeau JH, Taylor BA. 1984. Lengths of chromosomal segments conserved since divergence of man and mouse. *Proceedings of the National Academy of Sciences of the United States of America* **81**(3): 814-818.

Nam K, Ellegren H. 2012. Recombination drives vertebrate genome contraction. *PLoS genetics* **8**(5): e1002680.

Nery MF, Gonzalez DJ, Hoffmann FG, Opazo JC. 2012. Resolution of the laurasiatherian phylogeny: Evidence from genomic data. *Molecular phylogenetics and evolution*.

Nobrega MA, Ovcharenko I, Afzal V, Rubin EM. 2003. Scanning human gene deserts for long-range enhancers. *Science* **302**(5644): 413.

Osborne CS, Chakalova L, Brown KE, Carter D, Horton A, Debrand E, Goyenechea B, Mitchell JA, Lopes S, Reik W et al. 2004. Active genes dynamically colocalize to shared sites of ongoing transcription. *Nature genetics* **36**(10): 1065-1071.

Paten B, Herrero J, Fitzgerald S, Beal K, Flicek P, Holmes I, Birney E. 2008. Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome research* **18**(11): 1829-1843.

Peng Q, Pevzner PA, Tesler G. 2006. The fragile breakage versus random breakage models of chromosome evolution. *PLoS computational biology* **2**(2): e14.

Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD et al. 2006. In vivo enhancer analysis of human conserved non-coding sequences. *Nature* **444**(7118): 499-502.

Petrov DA, Sangster TA, Johnston JS, Hartl DL, Shaw KL. 2000. Evidence for DNA loss as a determinant of genome size. *Science* **287**(5455): 1060-1062.

Pevzner P, Tesler G. 2003. Human and mouse genomic sequences reveal extensive breakpoint reuse in

mammalian evolution. *Proceedings of the National Academy of Sciences of the United States of America* **100**(13): 7672-7677.
Poyatos JF, Hurst LD. 2007. The determinants of gene order conservation in yeasts. *Genome biology* **8**(11): R233.
Ranz JM, Maurin D, Chan YS, von Grotthuss M, Hillier LW, Roote J, Ashburner M, Bergman CM. 2007. Principles of genome evolution in the Drosophila melanogaster species group. *PLoS biology* **5**(6): e152.
Ryba T, Hiratani I, Lu J, Itoh M, Kulik M, Zhang J, Schulz TC, Robins AJ, Dalton S, Gilbert DM. 2010. Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome research* **20**(6): 761-770.
Sandelin A, Bailey P, Bruce S, Engstrom PG, Klos JM, Wasserman WW, Ericson J, Lenhard B. 2004. Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. *BMC genomics* **5**(1): 99.
Sankoff D, Trinh P. 2005. Chromosomal breakpoint reuse in genome sequence rearrangement. *Journal of computational biology : a journal of computational molecular cell biology* **12**(6): 812-821.
Schibler L, Roig A, Mahe MF, Laurent P, Hayes H, Rodolphe F, Cribiu EP. 2006. High-resolution comparative mapping among man, cattle and mouse suggests a role for repeat sequences in mammalian genome evolution. *BMC genomics* **7**: 194.
Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, Kutter C, Watt S, Martinez-Jimenez CP, Mackay S et al. 2010. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* **328**(5981): 1036-1040.
Schuster-Bockler B, Lehner B. 2012. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* **488**(7412): 504-507.
Shaw CJ, Lupski JR. 2004. Implications of human genome architecture for rearrangement-based disorders: the genomic basis of disease. *Human molecular genetics* **13 Spec No 1**: R57-64.
Sherman DJ, Martin T, Nikolski M, Cayla C, Souciet JL, Durrens P. 2009. Genolevures: protein families and synteny among complete hemiascomycetous yeast proteomes and genomes. *Nucleic acids research* **37**(Database issue): D550-554.
Sun H, Skogerbo G, Wang Z, Liu W, Li Y. 2008. Structural relationships between highly conserved elements and genes in vertebrate genomes. *PloS one* **3**(11): e3727.
Tawn EJ, Earl R. 1992. The frequencies of constitutional chromosome abnormalities in an apparently normal adult population. *Mutation research* **283**(1): 69-73.
Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. 2009. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome research* **19**(2): 327-335.
Volker M, Backstrom N, Skinner BM, Langley EJ, Bunzey SK, Ellegren H, Griffin DK. 2010. Copy number variation, chromosome rearrangement, and their association with recombination during avian evolution. *Genome research* **20**(4): 503-511.
Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, Smith SF, North P, Callaway H, Kelly K et al. 2005. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS biology* **3**(1): e7.
Zhao H, Bourque G. 2009. Recovering genome rearrangements in the mammalian phylogeny. *Genome research* **19**(5): 934-942.
Zhao S, Shetty J, Hou L, Delcher A, Zhu B, Osoegawa K, de Jong P, Nierman WC, Strausberg RL, Fraser CM. 2004. Human, mouse, and rat genome large-scale rearrangements: stability versus speciation. *Genome research* **14**(10A): 1851-1860.

Table and Figures legends

Table 1. Coefficients and statistics of Poisson regression models. Bold text corresponds to values indicative of an improvement in the model. A parameter affects the breakage rate when its regression coefficient is statistically different from 0 ($P(>|z|) < 0.05$). The goodness of fit of each model is assessed by a Chi^2 test on the residual deviance and degrees of freedom: a non-significant p-value means that the residual deviance may be attributed to statistical noise. The effect of an additional parameter on the fit is assessed by a Chi^2 test on the difference in residual deviances and degrees of freedom with and without the parameter: a significant p-value means that the fit is significantly better with the additional parameter. The pseudo R^2 corresponds to McFadden's pseudo R^2 (proportion of null deviance explained by the model).

Figure 1. Outline of the analysis. (1) Genome comparisons are performed for all pairs of species that are informative for the ancestral gene order (i.e. their common ancestor is the same as, or predates, the ancestor targeted in the reconstruction), to detect gene adjacencies that are identical in both genomes (ancestral). (2) All gene adjacencies identified in modern genomes are collected and weighted by the number of pairwise comparisons that report their conservation. (3) Ancestral adjacencies are selected from this collection under the reasoning that the best supported adjacencies are ancestral. (4a) The ancestral gene adjacencies define ancestral intergenes, with characteristics (length, %GC, %CNE) that can be robustly estimated by parsimony from modern values. (4b) Meanwhile, the reconstructed ancestral gene order is compared to independent descendant genomes to count the number of times each intergene has been affected by a rearrangement breakpoint later on during evolution. (5) The ancestral characteristics of intergenes and their breakage rate are then correlated using multivariate Poisson regression to quantify their contribution, if any, to breakage probability.

Figure 2. Support for Boreoeutheria gene adjacencies in modern genomes. (A) Distribution of the number of modern Boreoeutherian genomes supporting ancestral intergenes. (B) Contribution of each modern Boreoeutherian genome to the ancestral reconstruction support. The histogram represents the number of ancestral gene adjacencies found in a similar state in each modern genome. Genomes sequenced at low coverage (< 3x) are available as highly fragmented assemblies, and are less informative than high-coverage genomes in the ancestral reconstruction.

Figure 3. Phylogenetic tree of the five Boreoeutherian genomes used in the rearrangement analysis. Branch lengths correspond to breakpoint counts.

Figure 4. Estimation of ancestral intergene lengths. (A) Correlation of modern orthologous intergene lengths. For each ancestral intergene, the lengths of the modern intergenes are plotted (y axis) against their median value (x axis), which serves as an estimate for the ancestral value. Datapoints were grouped in bins of width 0.01 in log scale on both axis, and the density of datapoints is accounted for by the color scale on the right. (B) Distribution of ancestral intergene lengths estimates (in grey) and modern intergene lengths in five high-quality descendant genomes.

Figure 5. Breakage rate is a function of intergene length and CNE proportion. (A) After a logarithmic transformation, breakage rate increases linearly with intergene length. The regression model (red line: regression equation; shaded red area: 95% confidence interval) is different from the expectations of the classical "random model" (green line). (B) Intergenes without rearrangements contain a larger proportion of conserved non-coding sequence than those with rearrangement in all classes of intergene length.

Figure 6. Variations of four genomic parameters with intergene length in the human genome. (A) Transposable elements proportion increases linearly with log(intergene length) in intergenes between 500 bp and 10 kb but not in the rest of the genome. Dots correspond to individual datapoints, and filled circles to the average in 0.3 bins (log(bp)). (B) Segmental duplications increase in intergenes up to 100 kb and decrease afterwards. The correlation with intergene length is very low. Symbols as in A. (C) Recombination rates are constant in intergenes up to 10 kb, increase between 10 kb and 500 kb and decrease afterwards. The correlation with intergene length is very low. Symbols as in A. (D) The density of replication origins in intergenes increases linearly with intergene length after a logarithmic transformation (black dots: observed values; black line: regression equation). The distribution of replication origins does not follow random expectations (green line), but behaves similarly to the distribution of breakpoints in the ancestral genome (red line), with an increased density in small intergenes and a decreased density in large ones compared to random expectations. (E) The proportion of chromatin in an open configuration decreases with intergene length over the entire range of values. Symbols as in A.

Figure 7. Breakpoints simulated with the regression model based on intergene length and %CNE reproduce the characteristics of mammalian evolutionary breakpoints. In each case, the local environment of simulated breakpoints (Model) is compared to that of randomly chosen locations (Random) and randomly chosen intergenic locations (Random intergenic) in the human genome. Boxplots represent the distributions of average values obtained over 100 simulations. (A) Breakpoints occur in more gene-dense regions than expected at random. (B) Breakpoints are associated with higher densities of CpG islands. (C) Breakpoints occur in regions that are richer in SINEs. (D) Breakpoints are associated with segmental duplications. (D) Breakpoints occur in regions than are poorer in LINEs. (E) Breakpoints define an excess of short synteny blocks (red) compared to the Random (dark green) and Random intergenic (light green) controls. The distribution represents the average frequencies obtained over 100 simulations.

Figure 8. Simulations of paired dependent breakpoints are not sufficient to explain the correlation between breakage rate and intergene length. (A) Inversion breakpoints are simulated in the human genome so that the first breakpoint is chosen randomly, and the second is picked at a distance *d* with a probability derived from the contact probability in 3D in HiC experiments (Lieberman-Aiden et al., 2009). Breakpoints are considered as "visible" only if they alter the gene order, i.e. they do not occur in the same intergene. The simulated breakage rates according to intergene length are represented as black dots, the expectations of the classical random model as a green line, and the regression from the data as a red line. (B) Paired breakpoints are retained only if they occur in transposable elements of the same class (black dots) or the same type (white dots).

Figure 9. Breakage rate is a function of intergene length in yeasts. After a logarithmic transformation, breakage rate increases linearly with intergene length. The regression model (red line: regression equation; shaded red area: 95% confidence interval) is different from the expectations of the classical "random model" (green line).

Table 1.

	Coefficients					Goodness of fit		
	Simple regression	Stepwise regression	P(>|z|)	Null deviance (df)	Residual Deviance (df)	χ^2 P-value	Stepwise χ^2 P-value	Pseudo R^2
Model 1 : length only								
Intergene length	0.28	-	**< 2.10^{-16}**	167.3 (10)	12.4 (9)	**0.19**	**-**	**0.93**
Model 2 : length + %GC								
Intergene length	0.26	0.27	**< 2.10^{-16}**	137.8 (28)	25.7 (27)	**0.53**	**-**	**0.81**
%GC	-	0.003	0.44	137.8 (28)	25.1 (26)	0.52	0.42	0.82
Model 3 : length + %CNE								
Intergene length	0.28	0.30	**< 2.10^{-16}**	179.2 (19)	26.3 (18)	**0.09**	**-**	**0.85**
%CNE	-	- 4.55	**0.01**	179.2 (19)	20.7 (17)	**0.24**	**0.02**	**0.88**

Figure 1.

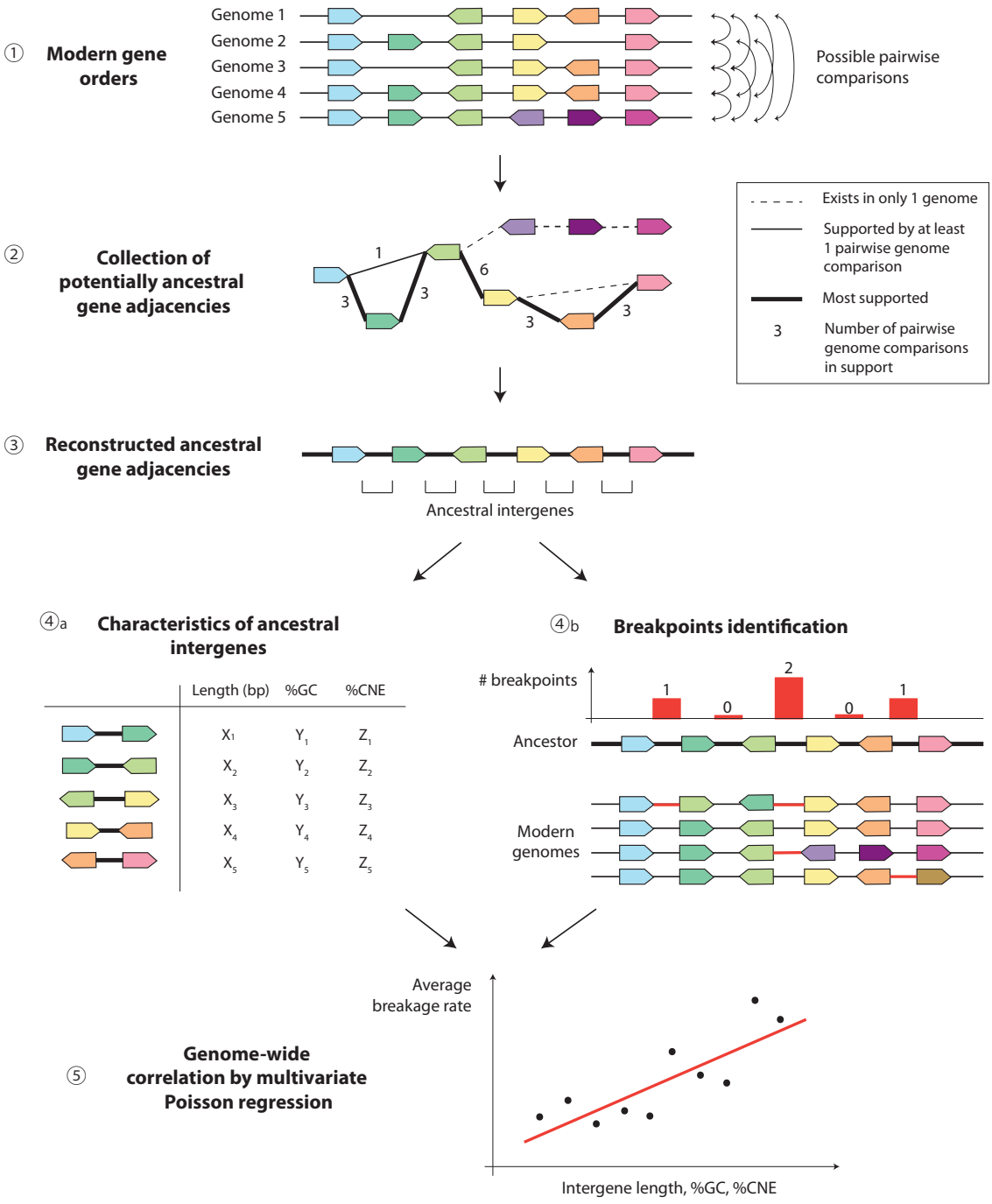

① Modern gene orders
Genome 1
Genome 2
Genome 3
Genome 4
Genome 5
Possible pairwise comparisons
② Collection of potentially ancestral gene adjacencies
1
3
3
6
3
3
Exists in only 1 genome
Supported by at least 1 pairwise genome comparison
Most supported
3
Number of pairwise genome comparisons in support
③ Reconstructed ancestral gene adjacencies
Ancestral intergenes
④a Characteristics of ancestral intergenes
Length (bp)
%GC
%CNE
X1
Y1
Z1
X2
Y2
Z2
X3
Y3
Z3
X4
Y4
Z4
X5
Y5
Z5
④b Breakpoints identification
breakpoints
1
0
2
0
1
Ancestor
Modern genomes
⑤ Genome-wide correlation by multivariate Poisson regression
Average breakage rate
Intergene length, %GC, %CNE

Figure 2.

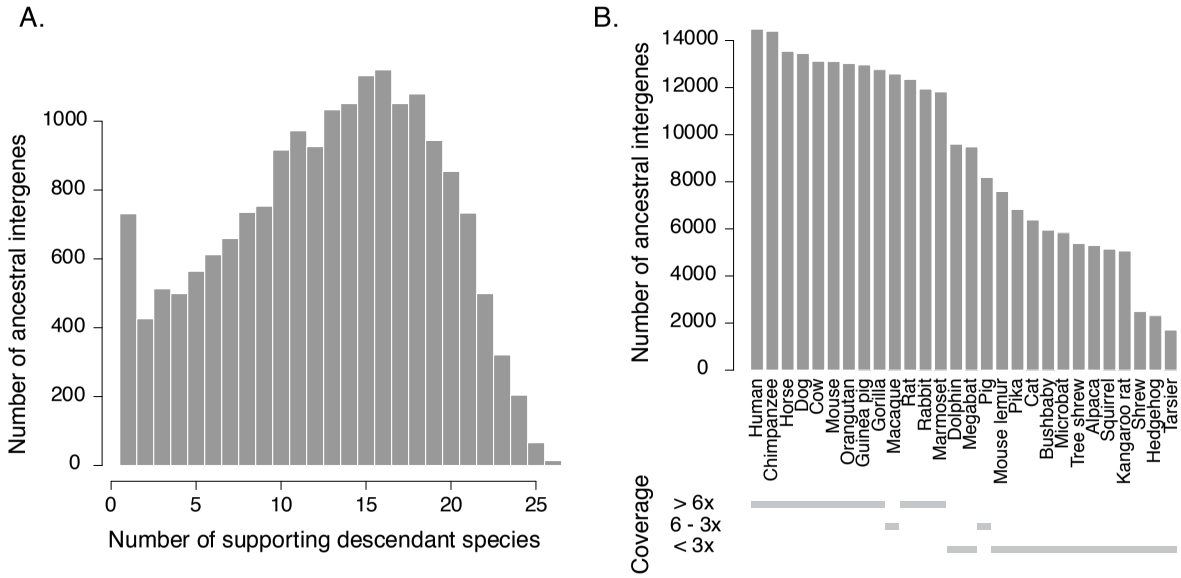


Figure 3.

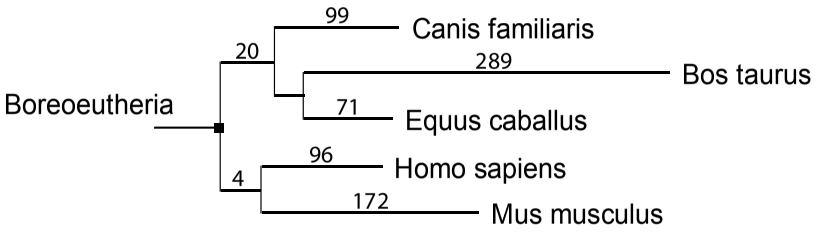


Figure 4.

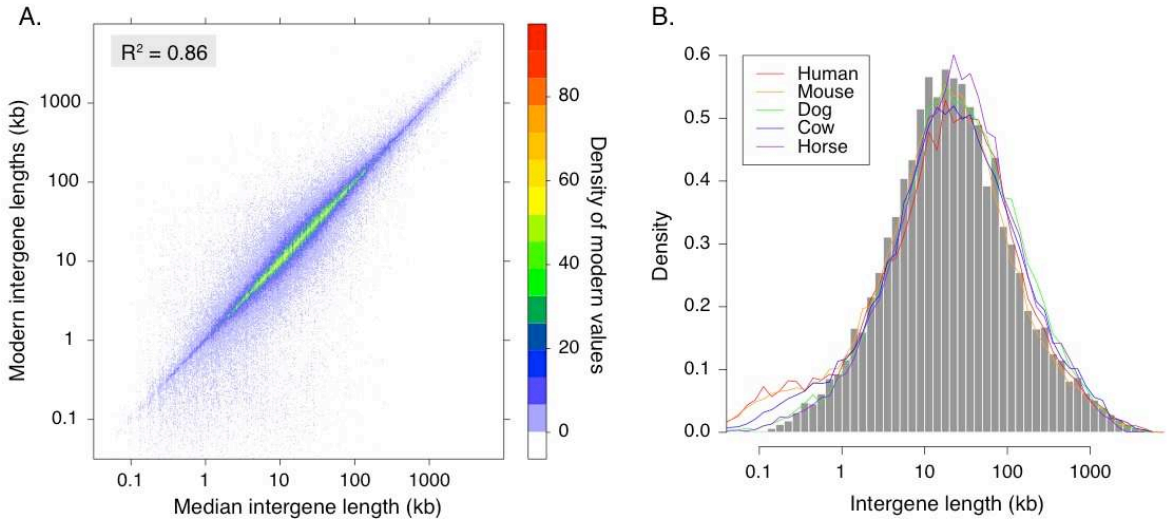


Figure 5.

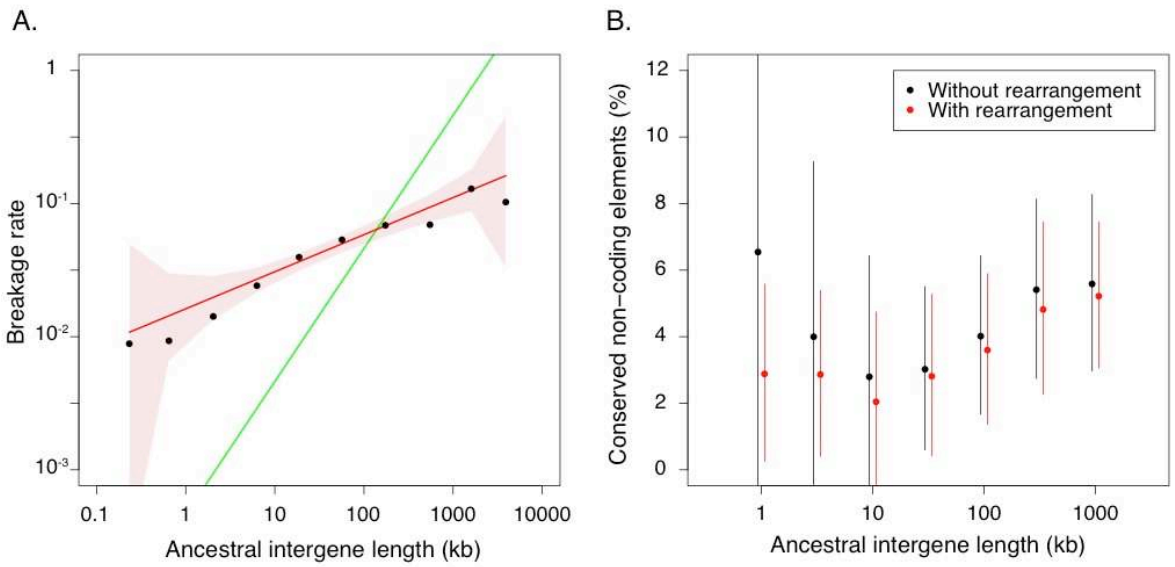


Figure 6.

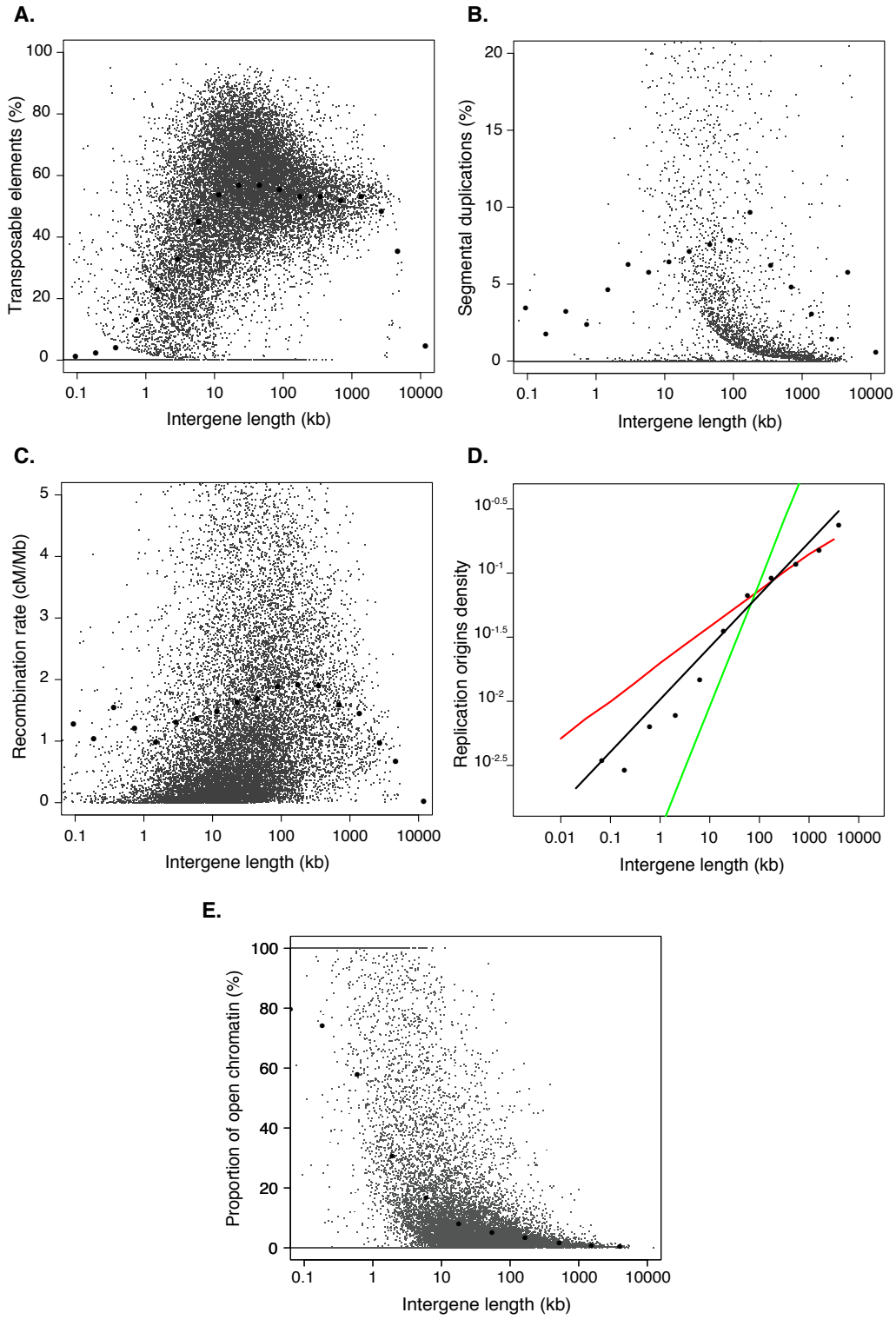

A.
Transposable elements (%)
Intergene length (kb)
B.
Segmental duplications (%)
Intergene length (kb)
C.
Recombination rate (cM/Mb)
Intergene length (kb)
D.
Replication origins density
Intergene length (kb)
E.
Proportion of open chromatin (%)
Intergene length (kb)

Figure 7.

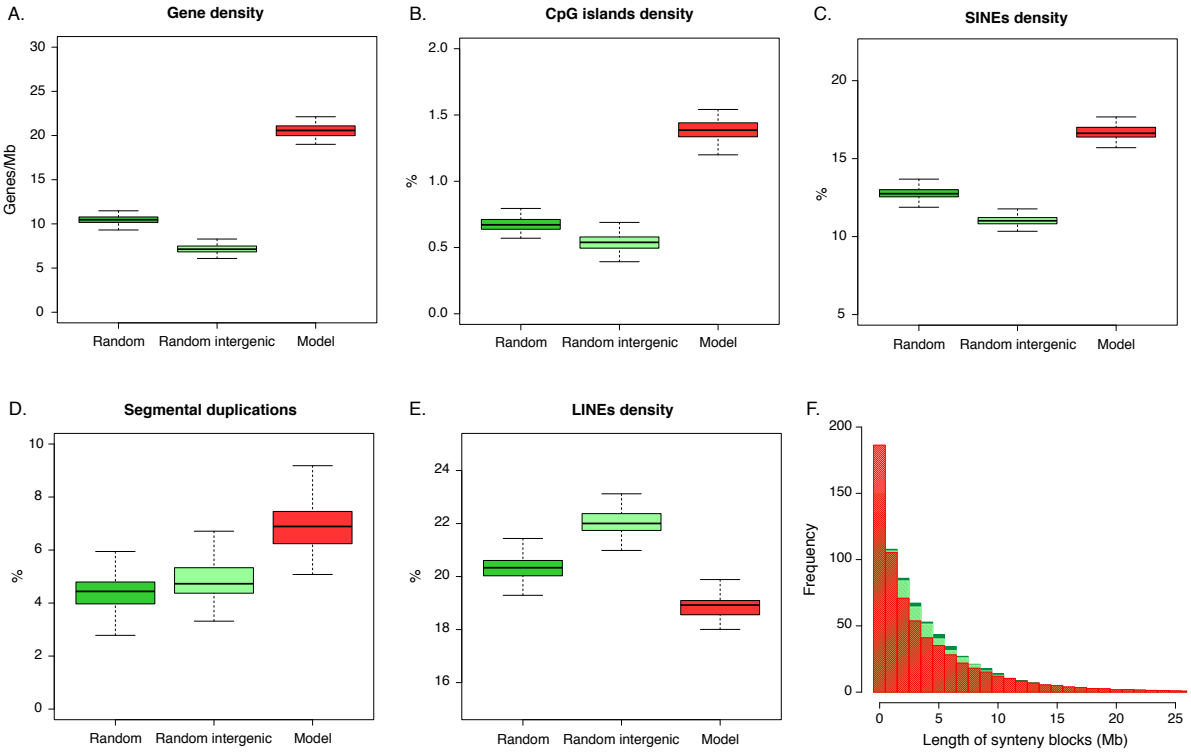

A.
Gene density
Genes/Mb
Random
Random intergenic
Model
B.
CpG islands density
%
Random
Random intergenic
Model
C.
SINEs density
%
Random
Random intergenic
Model
D.
Segmental duplications
%
Random
Random intergenic
Model
E.
LINEs density
%
Random
Random intergenic
Model
F.
Frequency
Length of synteny blocks (Mb)

Figure 8.

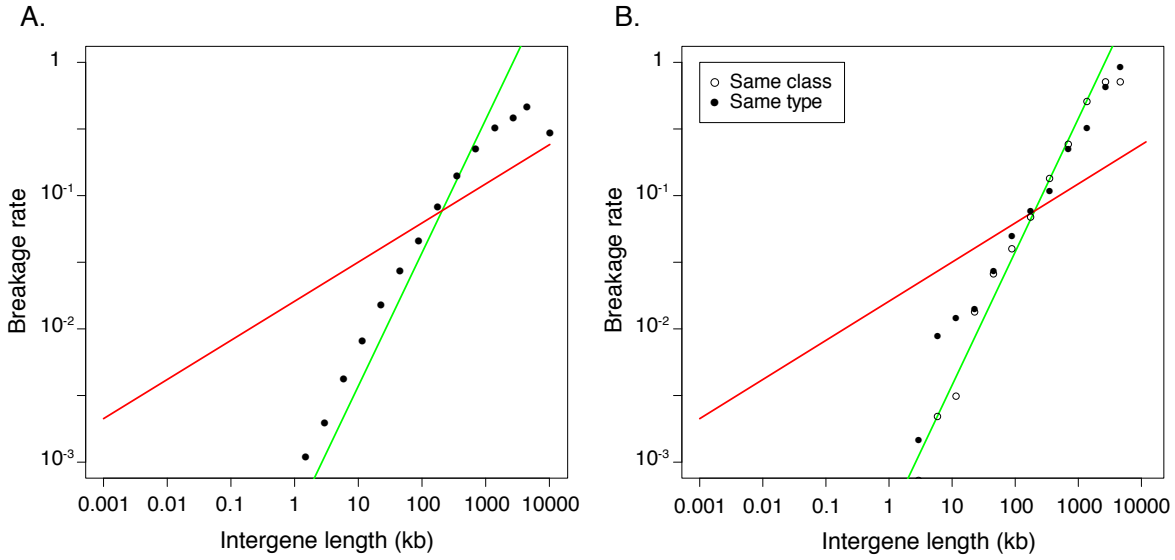
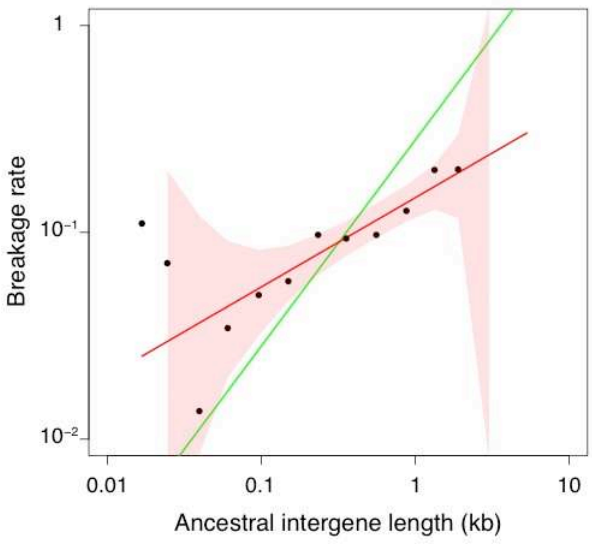


Figure 9.



The Zebrafish Reference Genome Sequence and its Relationship to the Human Genome

Kerstin Howe[1], Matthew D. Clark[1,2], Carlos F. Torroja[1,3], Camille Berthelot[4,5,6], James Torrance[1], Matthieu Muffato[7], John E. Collins[1], Sean Humphray[1], Karen McLaren[1], Lucy Matthews[1] , Stuart McLaren[1], Ian Sealy[1], Mario Caccamo[2], Romke Koch[8], Gerd-Jörg Rauch[9], Simon White[1], William Chow[1], Britt Kilian[1], Yi Zhou[10], Yong Gu[1], Jennifer Yen[1], Jan Vogel[1], Tina Eyre[1], Ruby Banerjee[1], Jianxiang Chi[1], Beiyuan Fu[1], Elizabeth Langley[1], Sean F. Maguire[1], Gavin Laird[1], David Lloyd[1], Emma Kenyon[1], Sarah Donaldson[1], Harminder Sehra[1], Jeff Almeida-King[1], Jane Loveland[1], Stephen Trevanion[1], Jonathan Bailey[1], Matt Jones[1], Mike Quail[1], Dave Willey[1], Adrienne Hunt[1], John Burton[1], Sarah Sims[1], Kirsten McLay[1], Suzanne Clarke[1], Adrian Clarke[1], Joy Davies[1], Melanie Robinson[1], Chris Clee[1], Sarah Holmes[1], Karen Oliver[1], Sami Bertrand[1], Clare Riddle[1], David Eliott[1], Glen Threadgold[1], Glenn Harden[1], Darren Ware[1], Beverly Mortimer[1], Giselle Kerry[1], Paul Heath[1], Benjamin Phillimore[1], Alan Tracey[1], Nicole Corby[1], Matthew Dunn[1], Christopher Johnson[1], Jonathan Wood[1], Susan Clark[1], Sarah Pelan[1], Guy Griffiths[1], Michelle Smith[1], Rebecca Glithero[1], Philip Howden[1], Nicholas Barker[1], Christopher Stevens[1], Joanna Harley[1], Karen Holt[1], Georgios Panagiotidis[1], Jamieson Lovell[1], Helen Beasley[1], Carl Henderson[1], Daria Gordon[1], Katherine Auger[1], Deborah Wright[1], Joanna Collins[1], Claire Raisen[1], Sarah Donaldson[1], Lauren Dyer[1], Kenric Leung[1], Lauren Robertson[1], Kirsty Ambridge[1], Daniel Leongamornlert[1], Sarah McGuire[1], Ruth Gilderthorp[1], Coline Griffiths[1], Deepa Manthravadi[1], Sarah Nichol[1], Gary Barker[1], Siobhan Whitehead[1], Michael Kay[1], Jacqueline Brown[1], Clare Murnane[1], Emma Gray[1], Matthew Humphries[1], Neil Sycamore[1], Darren Barker[1], David Saunders[1], Justene Wallis[1], Anne Babbage[1], Sian Hammond[1], Karen Oliver[1], Maryam Mashreghi-Mohammadi[1], Lucy Barr[1], Harminder Sehra[1], Sancha Martin[1], Paul Wray[1], Andrew Ellington[1], Nicholas Matthews[1], Matthew Ellwood[1], Rebecca Woodmansey[1], Graham Clark[1], James Cooper[1], Anthony Tromans[1], Darren Grafham[1], Carl Skuce[1], Richard Pandian[1], Robert Andrews[1], Elliot Harrison[1], Andrew Kimberley[1], Jane Garnett[1], Nigel Fosker[1], Rebekah Hall[1], Patrick Garner[1], Daniel Kelly[1], Christine Bird[1], Sophie Palmer[1], Ines Gehring[9], Andrea Berger[9], Christopher Dooley[9], Zübeyde Ersan-Ürün[9], Cigdem Eser[9], Horst Geiger[9], Maria Geisler[9], Lena Karotki [9], Anette Kirn[9], Judith Konantz[9], Martina Konantz[9], Martina Oberländer[9], Silke Rudolph-Geiger[9], Mathias Teucke[9], Fengtang Yang[1], Nigel P. Carter[1], Jennifer Harrow[1], Zemin Ning[1], Javier Herrero[7], Steve M. J. Searle[1], Anton Enright[7], Robert Geisler[9,11], Ronald H. A. Plasterk[8], Pieter J. de Jong[12], Leonard I. Zon[10], John H. Postlethwait[13], Christiane Nüsslein-Volhard[9], Tim J. P. Hubbard[1], Hugues Roest Crollius[4,5,6], Jane Rogers[1,2] and Derek L. Stemple*[1]

1 Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, United Kingdom

2 The Genome Analysis Centre, Norwich Research Park, Norwich, NR4 7UH, United Kingdom

3 Bioinformatics Unit, Centro Nacional de Investigaciones Cardiovasculares, 28029 Madrid, Spain

4 Ecole Normale Supérieure, Institut de Biologie de l'ENS, IBENS, Paris, F-75005 France

5 Inserm, U1024, Paris, F-75005 France

6 CNRS, UMR 8197, Paris, F-75005 France

7 EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom

8 Hubrecht Laboratory, Uppsalalaan 8, 3584 CT Utrecht, The Netherlands

9 Max Planck Institute for Developmental Biology, Spemannstr. 35, 72076 Tübingen, Germany

10 Stem Cell Program and Division of Hematology/Oncology, Children's Hospital and Dana Farber Cancer Institute, 1 Blackfan Cir., Karp 7, Boston, MA 02115, USA

11 Karlsruhe Institute of Technology (KIT), Campus North, Institute of Toxicology and Gentics (ITG), Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany

12 Children's Hospital Oakland, 747 52nd St. Oakland, Ca. 94609, USA

13 Institute of Neuroscience, University of Oregon, 1254 University of Oregon, 222 Huestis Hall, Eugene, OR 97403-1254

* Author for correspondence: Derek L. Stemple ds4@sanger.ac.uk

Summary
Zebrafish have become a popular organism for the study of vertebrate gene function. The ease of genetic studies accelerated by antisense oligonucleotide knockdown and mRNA over-expression combined with the virtual transparency of developing embryos have made zebrafish useful for the study of human genetic disease as well as the detailed study of vertebrate gene function. For effective modelling of human genetic disease, however, it is important to know how zebrafish genes and gene structures compare with orthologous human genes. To this end we have generated a high-quality sequence assembly of the zebrafish genome based on an overlapping set of large-insert clones, completely sequenced, ordered and oriented with a high-resolution high-density meiotic map. Moreover, we have combined manual genome annotation with the automated Ensembl pipeline, using high-coverage genome-wide sequenced cDNA data, resulting in more than 26,000 protein-coding gene models. We have compared the zebrafish reference genome with the human reference genome, focusing on overall genome structure and protein-coding genes.

Introduction

Although zebrafish had been used in embryological and toxicological studies for decades previously [1-3], it was the work of George Streisinger and colleagues during the 1980's that realised zebrafish as a genetically tractable organism. His group established, for zebrafish, a variety of mutagenesis methods and the direct screening of haploid or double-haploid embryos to isolate mutations affecting various aspects of development [4-6]. The systematic application of such genetic screens lead to the phenotypic characterisation of a large collection of mutations [7,8]. These mutations, when driven to homozygosity, can produce defects in a variety of organ systems with pathologies similar to human disease. Indeed, zebrafish have been used to study a variety of human diseases, such as cancers, muscular dystrophies and cilliopathies [9-12]. Zebrafish have also been used to contribute to our understanding of basic vertebrate biology. Notable examples including specification of the primary vasculature [13,14], control of hematopoietic stem cell (HSC) development [15,16], a paradigm for cardiac regeneration [17], acquisition of gut biota [18,19], host responses to pathogens [20,21], as well as providing critical evidence for the mechanism of lumen formation in vascular endothelial cells [22]. For basic vertebrate development, aside from systematically defining a large range of early developmental phenotypes, screens in zebrafish have contributed more generally to our understanding of the factors controlling specification of the cell types, organ systems and the vertebrate body axes [23-28].

Though the contributions have been significant, zebrafish research promises further to enhance our understanding of the detailed roles of specific genes in human diseases both rare and common. Increasingly, zebrafish experiments are included in studies of human genetic disease, often providing independent verification of the activity of a gene implicated in a human disease [10,29-31].

As the need for greater understanding of gene function increases in the context of human genetics, we will require systems to reach a more detailed understanding of the underlying biology. In many cases zebrafish research will provide a very good experimental situation to understand the gene function. Essential to this enterprise is a high quality genome sequence and complete annotation of zebrafish protein-coding genes with robust comparison to their human orthologues.

We present here a complete reference sequence for the zebrafish genome and a detailed annotation of protein coding genes in comparison with the orthologous genes in humans. Among the headline features of the zebrafish genome are: 1) The majority of human protein-coding genes, 69.2%, are orthologous to at least one zebrafish gene; 2) Nearly 40% of the zebrafish genome comprises Type II DNA repeats, transposable elements of >400 classes, whereas ~40% of the human genome comprises retrotransposon-derived LINE and SINE elements; 3) There are very few pseudogenes in the zebrafish genome; and 4) The zebrafish genome is a typical teleost genome showing, relative to the human genome, a similar degree of fragmentation in comparison with other teleost fish genomes.

Results

Assembly and Sequencing

The zebrafish genome-sequencing project started at the Wellcome Trust Sanger Institute in 2001. We chose the Tübingen strain as the reference strain as this strain had been used extensively to identify mutations affecting embryogenesis [8]. Our strategy resembled the clone-by-clone sequencing approach previously adopted for both the human and mouse genome projects. This involved generating genomic libraries, assembling genomic clones into a basic physical map of the genome, identifying a minimally overlapping set of clones, and then sequencing each clone to a 'gold standard' high quality. The resulting clone path for zebrafish assembly Zv9 shows similar overlap qualities to that of the human genome (99.3% Zv9 versus 99.9% GRCh37.p6 aligned clone overlap sequence) but differs from the human genome by the presence of transposon insertions and repeat extensions between clones.

Ultimately this physical assembly was tied to a high-density, high-resolution meiotic map, called the SATmap [32] (Supplementary Information), which provides genomic clone-size genetic resolution. The resolution and marker density of the SATmap allowed physical contigs and even individual clones to be ordered and oriented. This now provides a reliable scaffold to anchor the Zv9 reference genome sequence and is much more stable than previous assemblies. Through careful inspection of annotated gene order and application of the SATmap, many of the originally sequenced clones were found to represent haplotypic variants and were set aside. To provide the best possible representation of the genome sequence, the resulting clone path has been complemented by insertions from a whole genome shotgun assembly (WGS31, CABZ00000000). The Zv9 assembly is thus a hybrid of high-quality finished clone sequence (83%) and whole genome shotgun sequence (17%) with a total size of 1.412 Gb (Table 1).

A previous study, using chromosome flow measurements in comparison with human chromosomes, estimated the size of the zebrafish genome to be 1.454 Gb [33]. The Zv9 assembly thus accounts for at least 97.1% of the genome (Table 1) but might reside fully within the error margin of the chromosome flow estimate.

An aspect of the quality of the reference genome assembly can be obtained from the degree of alignment between previously described cDNAs and the Zv9 assembly. By comparing cDNA sequences available in public databases, we identified 21,471 candidate genes, which were aligned to Zv9. Considering a contiguous coverage of at least 90% of any given cDNA with at least 97% sequence identity, 90% of all candidate genes mapped to Zv9. Only 120 candidate genes could not be placed with at least 10% coverage at more than 90% identity. Thus only 0.6% of candidate genes remain to be incorporated into the genome assembly.

Protein-coding gene annotation

To obtain evidence for a more complete description of protein-coding genes we employed high-throughput short-read cDNA sequencing (RNA-seq) and

obtained a deep-coverage data set for mRNAs expressed in zebrafish at various stages of development and in adult tissues [34]. In the Ensembl annotation process, first, RNA-seq data are assembled into gene models. Separately obtained 3' end capture data are used to predict the 3' ends of transcripts and refine the initial models. Independently, a standard Ensembl gene build, incorporating filtered elements from the RNA-seq gene build, followed by a merge with the manually curated gene models [35] produced a comprehensive annotation of the 26,206 protein coding genes in Ensembl version 67.

With 26,206 protein-coding genes, zebrafish possess more protein-coding genes than any previously sequenced vertebrate and the largest number of species-specific genes when compared with the human, mouse or chicken genome. A likely explanation is the teleost specific whole genome duplication. All teleost fish are thought to have arisen from a common ancestor ~340 million years ago [36], which had undergone a whole genome duplication relative to other vertebrates. The zebrafish genome contains 8,083 gene duplicates that originate from the teleost genome duplication. The zebrafish gene count is the highest absolute number of protein-coding genes in comparison with any of the other sequenced fish genomes.

A direct comparison of the zebrafish protein-coding genes with those of human reveals a number of interesting features. Firstly, 69% of human genes have at least one zebrafish orthologue as defined by Ensembl Compara [37] (Table 2). Reciprocally, 83% of zebrafish genes have at least one human orthologue. Among the orthologous genes, 53% of human genes have a one-to-one relationship with a zebrafish orthologue. Probably reflecting the teleost specific whole genome duplication, the second largest orthology class is the one-human-to-many-zebrafish class (Table 3), resulting in an average of 2.32 zebrafish genes for each human gene. There are a few notable genes where there is no clear zebrafish orthologue identifiable, for example the LIF, OSM or IL6 genes, although the receptors *lifra, lifrb, osmr and il6r* are clearly present in the zebrafish genome. Similarly there is no zebrafish BRCA1 gene, but there is a zebrafish orthologue of the BRCA1-associated BARD1 gene, which encodes an associated and functionally similar protein, and a zebrafish *brca2* gene, which plays an important role in oocyte development, probably reflecting its role in DNA damage repair [38].

Zebrafish have been successfully used to understand in greater detail the biological activity of genes orthologous to human disease related genes. To understand the number of potential disease related genes, we compared the list of human genes possessing at least one zebrafish orthologue with OMIM listed gene bearing morbidity descriptions. In OMIM there are currently 3406 human genes associated with morbidity descriptions. Of those genes, 2601 (82%) possess at least one zebrafish orthologue. Similarly a comparison to human genes identified in genome wide association studies (GWAS), shows that of the 4255 human genes implicated, there is at least one zebrafish orthologue for 3075 (76%) of these genes.

One human genomic region of particular disease interest is the major histocompatibility complex (MHC) or HLA (human leukocyte antigen) system. In humans, the MHC comprises a super locus that encodes, among other proteins, a large number of proteins involved in immune system function. This super locus extends over a large region of human chromosome 6 and, despite a high degree of polymorphism [39], its content and synteny are conserved among mammals, thus a comparable genomic super locus exists on mouse chromosome 17 [40]. Using the Ensembl Compara data to identify zebrafish orthologues of the human HLA genes, we find the great majority of genes are situated on four main chromosomes (3, 8, 16 and 19), while several singleton HLA orthologues are scattered throughout the zebrafish genome (Figure 1).

Genomic Landscape

The Structure of Repeated Sequences

The zebrafish genome assembly Zv9 shows an overall repeat content of 52.2%, the highest so far reported in a vertebrate. All of the other teleost fish that have been sequenced so far, even the closest relative, the common carp, exhibits a much lower repeat content, ranging from below 20% to about 30%. This might suggest that the evolutionary path leading to the species *Danio rerio* has experienced a repeat expansion, possibly facilitated through a population bottleneck. Genomic sequences of other *Danio* species may clarify this.

Whereas the majority of transposed elements (TEs) found in the human genome are of Type I (retrotransposable elements), with more than 4.3 million placements covering 43.9% of the sequence, only 10.7% of the zebrafish genome sequence is covered by Type I elements in less than 500k instances. In contrast, the zebrafish genome contains a striking excess of Type II DNA TEs. Indeed, 2.3M instances of Type II DNA TEs cover 38.7% of the genome sequence (Table 4), whereas Type II repeats cover only 3.2% of the human genome.

This pronounced Type II TE abundance is unique among the sequenced vertebrate genomes and there is evidence in the genome sequence of recently active Type II TEs. The closest vertebrate species in terms of abundance is *Xenopus tropicalis* (25% Type II TEs), whereas the sequenced and annotated teleost fish (*Takifugu*, *Tetraodon*, Stickleback and Medaka) each possess Type II TE coverage of < 10%. Zebrafish Type II TEs are divided into 14 superfamilies with 401 repeat families in total (Supplementary Table 8). The DNA and hAT superfamilies are the most abundant and diverse in the zebrafish genome, together covering 28% of the sequenced genome.

Pseudogenes

The zebrafish reference genome contains comparatively few pseudogenes, a total of 154 manually annotated pseudogenes compared to 13,340 pseudogenes in the human genome (Table 5). This surfeit of zebrafish pseudogenes may be related to the balance of Type II TEs relative to Type I retrotransposable elements. In the human genome nearly 40% of the genome comprises LINE and SINE elements, which are retrotransposon derived. The majority of processed pseudogenes, i.e. those with no apparent

intronic sequence, are thought to arise from retrotransposition. Consistent with this notion, in the human reference genome 74.9% of all pseudogenes processed and 22.1% are unprocessed. By contrast in zebrafish 14.3% of pseudogenes are unprocessed and 77.9% are not. Several zebrafish processed-pseudogenes are flanked by Type I, retrotransposon elements, indicating that retrotransposon activity has modified the zebrafish genome, but has not led to the expansions seen in mammalian genomes.

Remarkable structure of chromosome 4
The long arm of chromosome 4 has long attracted interest due to is lack of protein coding genes and an apparent heterochromatic state, indicated by unusual uniform low-level C- and CMA3-staining. Chromosome 4 is known to be a late-replicating chromosome and hybridisation studies suggested that genomic copies of 5S rDNA are scattered along the long arm at high redundancy, but not significantly on any other chromosome [41].

Its sequence landscape (Figure 2) shows a remarkable increase in repeat content immediately after the presumed centromere (~ 24 Mb), which is continued through to the end of the sequence. At ~ 27 Mb, the otherwise uniform presence of the satellite repeat SAT-2 on the long arm ends abruptly. This location is also the starting point of uniform MOSAT-2 distribution, a satellite repeat which is nearly absent from all other chromosomes, but highly enriched on the long arm of chromosome 4. The sub-telomeric region of the long arm shows a distinct distribution of repeat elements, with relatively fewer interspersed elements and an increased content of satellite, simple and tandem repeats that do not harbour 5S rDNA sequence. The gene content is reduced on the long arm and the GC content slightly increased.

The long arm of chromosome 4 also has a special structure with regard to gene orthology and synteny. About 80% of the genes present have no identified orthologue in human. In fact, 110 genes (out of 663) have no identified orthologue in teleost fish genomes, and indeed appear to be zebrafish-specific genes. The genes in this region are highly duplicated (Figure SI8A), with 31 ancestral gene families alone providing 77.5% of the genes, the largest of which contains no less than 109 duplicates in this region. The largest of these families correspond to NLR-like proteins with putative roles in innate immunity and zinc finger proteins, a feature that also holds true for many genes that have no teleost orthologue, suggesting that they might be other duplicates from the same families that have not been correctly inserted in the gene trees.

We also observed a very high density of small nuclear RNAs on chromosome 4, especially from the spliceosome (Figure SI8B). The snRNAs carried on the long arm of chromosome 4 account for 53.2% of all snRNAs in the zebrafish genome.

Variation
Zebrafish used in experimental research are generally outbred. This is manifest in a fairly high degree of single nucleotide polymorphism (SNP). We measured SNP variation between the AB and Tübingen strains by comparing

two individuals used to generate the SATmap to each other and to the Zv9 reference. For intra-strain Tübingen variation, we used the haplotypic sequence derived from the genome project [42,43]. We found that the degree of variation within Tübingen is high with a rate of 29 +/- 44 SNPs per 10kb of genomic sequence, while the inter strain variation between the two individual fish was considerably higher at 84 +/- 37 SNPs per 10kb (Table 6).

Evolution

Orthologues between zebrafish and amniotes

The protein-coding gene content of vertebrates is relatively stable in numbers (from 17,000 to 23,000 in tetrapods, and up to 26,000 in fish approximately), although even closely related species may show great disparities in the nature of their protein gene content. We performed a 4-way comparison between the proteome of two mammals (human and mouse), one avian (chicken) and zebrafish to quantify the fraction of shared and species-specific genes present in each genome (Figure 3). A core group of 10,660 genes is found in all four species and likely approximates an essential set of vertebrate protein coding genes. This is notably less than the core set of 11,809 vertebrate genes previously identified in common between three fish genomes (*Tetraodon*, Medaka, zebrafish) and three amniotes (human, mouse, chicken) [44] but the discrepancy probably reflects the improved annotation of these genomes that often results in fusing fragmented gene structures. Each species has between 2,596 and 3,634 specific genes. The notable excess observed in zebrafish may be a consequence of the WGD, because pairs of duplicated genes that arose from the WGD, but with no orthologue in amniotes, will be counted as two specific genes. Also, 2,059 genes are found in human, mouse and zebrafish but not in chicken, which is twice higher than the number of genes found in all amniotes but not in zebrafish (892). It is unclear whether these genes have indeed been lost along the chicken branch, or whether this is due to annotation or orthology assignation errors in the chicken genome.

Double-conserved synteny (DCS) and identification of ohnologs

Relative to other vertebrate species teleosts are thought to have arisen from a common ancestor that underwent an additional round of whole genome duplication (WGD) called 3R [46]. We identified double-conserved synteny blocks between all sequenced tetrapod and four fish genomes (zebrafish, Medaka, stickleback and *Tetraodon*). The DCS blocks are defined as runs of genes in the non-duplicated species that are found on two alternating chromosomes in the species that underwent a WGD [47], although the genes may not be adjacent in the duplicated species [48]. The DCS between fish and human are represented on either side of each human chromosome (Figure 4).

Using DCS blocks, we identified zebrafish paralogous genes that are part of DCS blocks and consistent with the locally alternating chromosomes, hence with an origin at the WGD. We identified 3,440 pairs of such genes, called ohnologs [45], for a total of 8,083 genes when subsequent duplications are taken into account. This number of ancestral genes retained as duplicates in zebrafish is higher, both in absolute number and in proportion, than in other fish genomes (Chi^2 test, all *p*-values < 3.10^{-5}).

We compared the 8,083 zebrafish 3R-ohnologs with human ohnologs originating from the two rounds of WGD common to all vertebrates (ref Dehal), called 2R ohnologs and find that the two sets strongly overlap (pval?). In general, zebrafish ohnologous pairs are enriched in specific functions (neural activity, transcription factors) and are orthologous to mammalian genes under stronger evolutionary constraint than genes that have lost their second copy.

A circular representation of the pairs of ohnologs (Figure 5) highlights chromosomes or part of chromosomes that descend from the same pre-duplication ancestral chromosome (e.g. chromosomes 3 and 12, 17 and 20, 16 and 19). Among the zebrafish chromosome, chromosome 16 and chromosome 19 are unique in their one-to-one conservation of synteny. Consistent with the conservation of synteny chromosome 16 and chromosome 19 possess clusters of MHC orthologues as well as the HoxAb and HoxAa clusters, respectively.

DNA transposons

The abundance of DNA transposons is of particular interest since they are reported to cause chromosome rearrangements through alternative transposition and recombination [49], which is consistent with earlier reports of zebrafish showing a lack of long-range synteny with human genes compared to Medaka, *Tetraodon* and *Takifugu* [44,50]. With assembly Zv9, however, we find a comparable degree of synteny conservation in zebrafish and the other sequenced teleost fish.

The zebrafish genome is no more rearranged than other fish genomes when considering various measures of intra-chromosomal gene-order or gene-linkage conservation (Figure SI9 and SI10). This comes in contradiction to previous studies, but is ascribed to the lesser quality of previous versions of the assembly [50]. In contrast, the distribution of ohnologs among chromosomes shows that inter-chromosomal rearrangements are more frequent in the zebrafish lineage (Figure SI11).

Discussion

Since the earliest whole-genome shotgun-only assembly was made available to the public in 2002, the zebrafish reference genome sequence has enabled a wealth of new discoveries, especially the positional cloning of genes from mutations affecting embryogenesis. With assembly Zv9, more than 83% of the genome is represented by gold-standard finished overlapping large insert clones and the correct order and orientation of > 90% of these clones is given by the high-density, high-resolution meiotic SATmap [32]. With an integrated Ensembl gene-building pipeline using high-coverage next generation RNA sequencing [34], we are able to identify and annotate more than 26,000 protein-coding genes. Moreover, the annotated reference genome has enabled the generation of accurate whole exome enrichment reagents, which are accelerating both positional cloning projects and new genome-wide mutation discovery efforts.

While the zebrafish reference genome sequencing is complete, there are a few poorly assembled regions, which are being resolved by the Genome Reference Consortium (genomereference.org). With the next assembly release, we expect to have resolved the vast majority of the remaining assembly problems and to have finished the sequencing of the remaining gap-filling clones, ultimately removing the need to use WGS. Putting technical difficulties aside, the zebrafish reference genome provides a wealth of information for comparison with the human genome.

Overall, we find that the vast majority of human genes have at least one zebrafish orthologue and that genes associated with human disease and other traits show an even greater degree of conservation with zebrafish genes. The modelling of human disease genes in zebrafish is thus likely to yield extremely valuable biological information regarding the detailed function of these genes.

One remarkable feature of the zebrafish genome is the structure of repeats, which is unique among the sequenced vertebrates. There is a very high DNA transposon content and a correspondingly low coverage by retrotransposable elements, which may explain the marked lack of pseudogenes in comparison with the human genome.

The conserved synteny of ohnologs between chromosomes 16 and 19 is further correlated with two other major genome features. Firstly, the presence of two of the four major MHC gene and secondly, that of two of the major Hox gene clusters HoxAb and HoxAa[51,52]. The SATmap data and a previous publication[53] show a major signal for sex determination on chromosome 16, which taken together suggest an epi-genetic mechanism for sex determination in zebrafish [32].

Literature Cited

1 Weis, J. S. Analysis of the development of nervous system of the zebrafish, Brachydanio rerio. I. The normal morphology and development of the spinal cord and ganglia of the zebrafish. *J Embryol Exp Morphol* **19**, 109-119 (1968).

2 Weis, J. S. Analysis of the development of the nervous system of the zebrafish, Brachydanio rerio. II. The effect of nerve growth factor and its antiserum on the nervous system of the zebrafish. *J Embryol Exp Morphol* **19**, 121-135 (1968).

3 Battle, H. I. & Hisaoka, K. K. Effects of ethyl carbamate (urethan) on the early development of the teleost Brachydanio rerio. *Cancer Res* **12**, 334-340 (1952).

4 Streisinger, G., Walker, C., Dower, N., Knauber, D. & Singer, F. Production of clones of homozygous diploid zebra fish (Brachydanio rerio). *Nature* **291**, 293-296 (1981).

5 Walker, C. & Streisinger, G. Induction of Mutations by gamma-Rays in Pregonial Germ Cells of Zebrafish Embryos. *Genetics* **103**, 125-136 (1983).

6 Golling, G. *et al.* Insertional mutagenesis in zebrafish rapidly identifies genes essential for early vertebrate development. *Nat Genet* **31**, 135-140, doi:10.1038/ng896 (2002).

7 Driever, W. *et al.* A genetic screen for mutations affecting embryogenesis in zebrafish. *Development* **123**, 37-46 (1996).

8 Haffter, P. *et al.* The identification of genes with unique and essential functions in the development of the zebrafish, Danio rerio. *Development* **123**, 1-36 (1996).

9 Ceol, C. J. *et al.* The histone methyltransferase SETDB1 is recurrently amplified in melanoma and accelerates its onset. *Nature* **471**, 513-517, doi:10.1038/nature09806 (2011).

10 Roscioli, T. *et al.* Mutations in ISPD cause Walker-Warburg syndrome and defective glycosylation of alpha-dystroglycan. *Nat Genet* **44**, 581-585, doi:10.1038/ng.2253 (2012).

11 Gorden, N. T. *et al.* CC2D2A is mutated in Joubert syndrome and interacts with the ciliopathy-associated basal body protein CEP290. *Am J Hum Genet* **83**, 559-571, doi:10.1016/j.ajhg.2008.10.002 (2008).

12 Davis, E. E. *et al.* TTC21B contributes both causal and modifying alleles across the ciliopathy spectrum. *Nat Genet* **43**, 189-196, doi:10.1038/ng.756 (2011).

13 Zhong, T. P., Childs, S., Leu, J. P. & Fishman, M. C. Gridlock signalling pathway fashions the first embryonic artery. *Nature* **414**, 216-220, doi:10.1038/35102599 (2001).

14 Weinstein, B. M., Stemple, D. L., Driever, W. & Fishman, M. C. Gridlock, a localized heritable vascular patterning defect in the zebrafish. *Nat Med* **1**, 1143-1147 (1995).

15 Goessling, W. *et al.* Genetic interaction of PGE2 and Wnt signaling regulates developmental specification of stem cells and regeneration. *Cell* **136**, 1136-1147, doi:10.1016/j.cell.2009.01.015 (2009).

16 North, T. E. *et al.* Hematopoietic stem cell development is dependent on blood flow. *Cell* **137**, 736-748, doi:10.1016/j.cell.2009.04.023 (2009).

17 Lepilina, A. *et al.* A dynamic epicardial injury response supports progenitor cell activity during zebrafish heart regeneration. *Cell* **127**, 607-619, doi:10.1016/j.cell.2006.08.052 (2006).

18 Rawls, J. F., Mahowald, M. A., Ley, R. E. & Gordon, J. I. Reciprocal gut microbiota transplants from zebrafish and mice to germ-free recipients reveal host habitat selection. *Cell* **127**, 423-433, doi:10.1016/j.cell.2006.08.043 (2006).

19 Bates, J. M. *et al.* Distinct signals from the microbiota promote different aspects of zebrafish gut differentiation. *Developmental biology* **297**, 374-386, doi:10.1016/j.ydbio.2006.05.006 (2006).

20 Hegedus, Z. *et al.* Deep sequencing of the zebrafish transcriptome response to mycobacterium infection. *Molecular immunology* **46**, 2918-2930, doi:10.1016/j.molimm.2009.07.002 (2009).

21 Volkman, H. E. *et al.* Tuberculous granuloma induction via interaction of a bacterial secreted protein with host epithelium. *Science* **327**, 466-469, doi:10.1126/science.1179663 (2010).

22 Kamei, M. *et al.* Endothelial tubes assemble from intracellular vacuoles in vivo. *Nature* **442**, 453-456, doi:10.1038/nature04923 (2006).

23 Lewis, K. E. *et al.* Control of muscle cell-type specification in the zebrafish embryo by Hedgehog signalling. *Developmental biology* **216**, 469-480, doi:10.1006/dbio.1999.9519 (1999).

24 Ober, E. A., Verkade, H., Field, H. A. & Stainier, D. Y. Mesodermal Wnt2b signalling positively regulates liver specification. *Nature* **442**, 688-691, doi:10.1038/nature04888 (2006).

25 Gritsman, K. *et al.* The EGF-CFC protein one-eyed pinhead is essential for nodal signaling. *Cell* **97**, 121-132 (1999).

26 Zhang, J., Talbot, W. S. & Schier, A. F. Positional cloning identifies zebrafish one-eyed pinhead as a permissive EGF-related ligand required during gastrulation. *Cell* **92**, 241-251 (1998).

27 Halpern, M. E., Ho, R. K., Walker, C. & Kimmel, C. B. Induction of muscle pioneers and floor plate is distinguished by the zebrafish no tail mutation. *Cell* **75**, 99-111 (1993).

28 Talbot, W. S. *et al.* A homeobox gene essential for zebrafish notochord development. *Nature* **378**, 150-157, doi:10.1038/378150a0 (1995).

29 Panizzi, J. R. *et al.* CCDC103 mutations cause primary ciliary dyskinesia by disrupting assembly of ciliary dynein arms. *Nat Genet* **44**, 714-719, doi:10.1038/ng.2277 (2012).

30 Golzio, C. *et al.* KCTD13 is a major driver of mirrored neuroanatomical phenotypes of the 16p11.2 copy number variant. *Nature* **485**, 363-367, doi:10.1038/nature11091 (2012).

31 Tobin, D. M. *et al.* Host genotype-specific therapies can optimize the inflammatory response to mycobacterial infections. *Cell* **148**, 434-446, doi:10.1016/j.cell.2011.12.023 (2012).

32 Clark, M. D. *et al.* A high-resolution, high-density meiotic map for zebrafish. *In Preparation* (2012).

33 Freeman, J. L. *et al.* Definition of the zebrafish genome using flow cytometry and cytogenetic mapping. *BMC Genomics* **8**, 195, doi:10.1186/1471-2164-8-195 (2007).

34 Collins, J. E., White, S., Searle, S. M. & Stemple, D. L. Incorporating RNA-seq data into the Zebrafish Ensembl Gene Build. *Genome Res*, doi:10.1101/gr.137901.112 (2012).

35 Wilming, L. G. *et al.* The vertebrate genome annotation (Vega) database. *Nucleic Acids Res* **36**, D753-760, doi:10.1093/nar/gkm987 (2008).

36 Amores, A., Catchen, J., Ferrara, A., Fontenot, Q. & Postlethwait, J. H. Genome evolution and meiotic maps by massively parallel DNA sequencing: spotted gar, an outgroup for the teleost genome duplication. *Genetics* **188**, 799-808, doi:10.1534/genetics.111.127324 (2011).

37 Vilella, A. J. *et al.* EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res* **19**, 327-335, doi:10.1101/gr.073585.107 (2009).

38 Rodriguez-Mari, A. *et al.* Roles of brca2 (fancd1) in oocyte nuclear architecture, gametogenesis, gonad tumors, and genome stability in zebrafish. *PLoS genetics* **7**, e1001357, doi:10.1371/journal.pgen.1001357 (2011).

39 Robinson, J. *et al.* The IMGT/HLA database. *Nucleic Acids Res* **39**, D1171-1176, doi:10.1093/nar/gkq998 (2011).

40 Kulski, J. K., Shiina, T., Anzai, T., Kohara, S. & Inoko, H. Comparative genomic analysis of the MHC: the evolution of class I duplication blocks, diversity and complexity from shark to man. *Immunol Rev* **190**, 95-122 (2002).

41 Sola, L. & Gornung, E. Classical and molecular cytogenetics of the zebrafish, Danio rerio (Cyprinidae, Cypriniformes): an overview. *Genetica* **111**, 397-412 (2001).

42 Bradley, K. M. *et al.* A major zebrafish polymorphism resource for genetic mapping. *Genome Biology* **8**, R55, doi:10.1186/gb-2007-8-4-r55 (2007).

43 Guryev, V. *et al.* Genetic variation in the zebrafish. *Genome Res* **16**, 491-497, doi:10.1101/gr.4791006 (2006).

44 Kasahara, M. *et al.* The medaka draft genome and insights into vertebrate genome evolution. *Nature* **447**, 714-719, doi:10.1038/nature05846 (2007).

45 Wolfe, K. Robustness--it's not where you think it is. *Nat Genet* **25**, 3-4, doi:10.1038/75560 (2000).

46 Meyer, A. & Schartl, M. Gene and genome duplications in vertebrates: the one-to-four (-to-eight in fish) rule and the evolution of novel gene functions. *Current Opinion in Cell Biology* **11**, 699-704, doi:10.1016/s0955-0674(99)00039-3 (1999).

47 Kellis, M., Birren, B. W. & Lander, E. S. Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. *Nature* **428**, 617-624, doi:10.1038/nature02424 (2004).

48 Jaillon, O. *et al.* Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature* **431**, 946-957, doi:10.1038/nature03025 (2004).

49 Feschotte, C. & Pritham, E. J. DNA transposons and the evolution of eukaryotic genomes. *Annu Rev Genet* **41**, 331-368, doi:10.1146/annurev.genet.40.110405.090448 (2007).

50 Semon, M. & Wolfe, K. H. Rearrangement rate following the whole-genome duplication in teleosts. *Mol Biol Evol* **24**, 860-867, doi:10.1093/molbev/msm003 (2007).

51 Amores, A. *et al.* Developmental roles of pufferfish Hox clusters and genome evolution in ray-fin fish. *Genome Res* **14**, 1-10, doi:10.1101/gr.1717804 (2004).

52 Henkel, C. V. *et al.* Comparison of the Exomes of Common Carp (Cyprinus carpio) and Zebrafish (Danio rerio). *Zebrafish* **9**, 59-67, doi:10.1089/zeb.2012.0773 (2012).

53 Bradley, K. M. *et al.* An SNP-Based Linkage Map for Zebrafish Reveals Sex Determination Loci. *G3 (Bethesda)* **1**, 3-9, doi:10.1534/g3.111.000190 (2011).

Tables and Figures

Assembly		Annotation	
Total length	1,412,464,843 bp	Protein coding genes	26,039
Total clone length	1,175,673,296 bp	Pseudogenes	192
Total WGS31 contig length	234,099,447 bp	RNA genes	4,444
Placed scaffold length	1,357,051,643 bp	Immunoglobulin/T-cell receptor gene segments	56
Unplaced scaffold length	55,413,200 bp		
		Gene exons	320,161
Max scaffold length	12,372,269 bp	Gene transcripts	52,873
Scaffold N50	1,551,602 bp		
Clone number	11,100		
WGS31 contig number	26,199		
Placed scaffold number	3,452		
Unplaced scaffold number	1,107		

Table 1. Assembly and annotation statistics for Zv9 (Bioproject: PRJNA11776), based on Ensembl version 63.

Human genes with at least one fish orthologue	#	%
Total number of human protein coding genes	21976	
Human genes with at least one zebrafish orthologue	15216	69.2%
Human genes with at least one stickleback orthologue	14271	64.9%
Human genes with at least one Medaka orthologue	13788	62.7%
Human genes with at least one *Takifugu* orthologue	14006	63.7%
Human genes with MIM disease ID	#	%
Total number	3176	
MIM disease genes with at least one zebrafish orthologue	2601	81.9%
GWAS catalog* reported genes	#	%
Total with HGNC ID	4023	
Human GWAS catalog genes with at least one zebrafish orthologue	3075	76.4%

Table 2. Human versus fish orthologue comparison using Ensembl BioMart version 67. *GWAS Catalog: www.genome.gov/gwastudies/

Relation	Human	vs	Zebrafish	Ratio
One to one		9563		
One to many	3386		7855	1 to 2.32
Many to one	1315		488	2.69 to 1
Many to many	973	297	1545	1 to 1.59
Orthologous Total	15237	13734	19451	1 to 1.28
Unique	4758		6588	
Coding Gene Total	19995		26039	

Table 3: A comparison of human and zebrafish protein coding genes and their homology relationships using Ensembl Compara version 63 according to (www.ensembl.org/info/docs/compara/homology_method.html). For the 'many to many' and total orthologous gene counts, the middle figure denotes the number of core relationships.

Repeat type	Members (class)	Members (families)	Occurrence	Sequence coverage bp	Sequence coverage %
Simple repeats	1	1	2,072,975	90,907,462	6.4%
Tandem repeats	1	460,296	1,179,751	150,883,666	10.7%
Satellite repeats	2	5	64,230	12,683,635	0.9%
Type I LINE	4	66	114,662	37,050,077	2.6%
Type I SINE	2	4	132,378	31,330,142	2.2%
Type I LTR	9	478	151,963	50,079,393	3.5%
Type I non-LTR	6	34	72,922	35,495,847	2.5%
Type II DNA	14	401	2,256,483	546,740,746	38.7%
Transposons	1	8	167,078	28,908,390	2.0%
unclassified	1	1	679	677,694	0.05%

Table 4. Overview of repeat elements found in the zebrafish genome assembly Zv9. The repeat annotation was performed as described in the supplementary information. Note that the coverage is not additive, i.e. repeats can overlap each other.

	Human Pseudogenes		Zebrafish Pseudogenes	
Biotype	**Count**	**%**	**Count**	**%**
IG_pseudogene	161	1.2	9	5.8
Polymorphic_pseudogene	27	0.2	4	0.3
Processed_pseudogene	9992	74.9	21	13.6
Pseudogene	7	0.1	5	3.3
TR_pseudogene	44	0.3		
Unitary_pseudogene	159	1.2		
Unprocessed_pseudogene	2950	22.1	115	74.7
Total	13,340		154	

Table 5. Human versus Zebrafish pseudogene comparison, based on zebrafish Vega version 47. The biotype classification is described at (vega.sanger.ac.uk/info/about/gene_and_transcript_types.html).

Comparison		SNPs	Genome Length	Golden Path length	Density (%)	SNPs per 10kb
Intra strain	Tü/Ref	3,924,070	1,505,581,940	1,412,464,843	0.2778	29 +/- 44
Inter strain	Tü/AB*	6,995,534	1,505,581,940	1,307,864,843	0.5349	84 +/- 37
	AB/Ref	7,755,823	1,505,581,940	1,412,464,843	0.5491	

Table 6. Intra strain (Tübingen) and Inter strain (Tübingen versus AB) SNP density comparisons
* We removed 104,600,000 bp (7.4% of genome) where there were fewer than 5 SNPs due to monomorphic genomic regions shared between the individual AB and Tü double haploid fish used for this analysis.

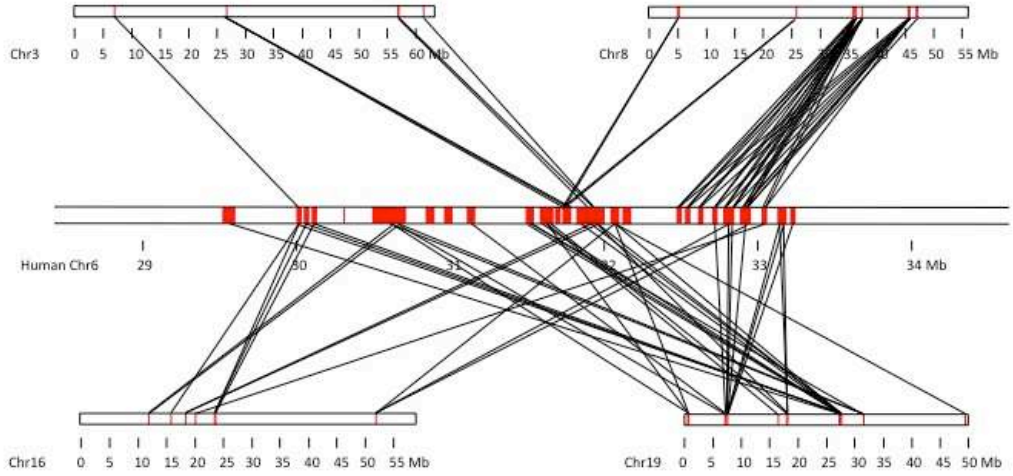


Figure 1. Orthologous relationship between human MHC genes clustered on chromosome 6, and the zebrafish chromosomes 3, 8, 16 and 19, carrying the majority of orthologous MHC genes (Ensembl Compara 67).

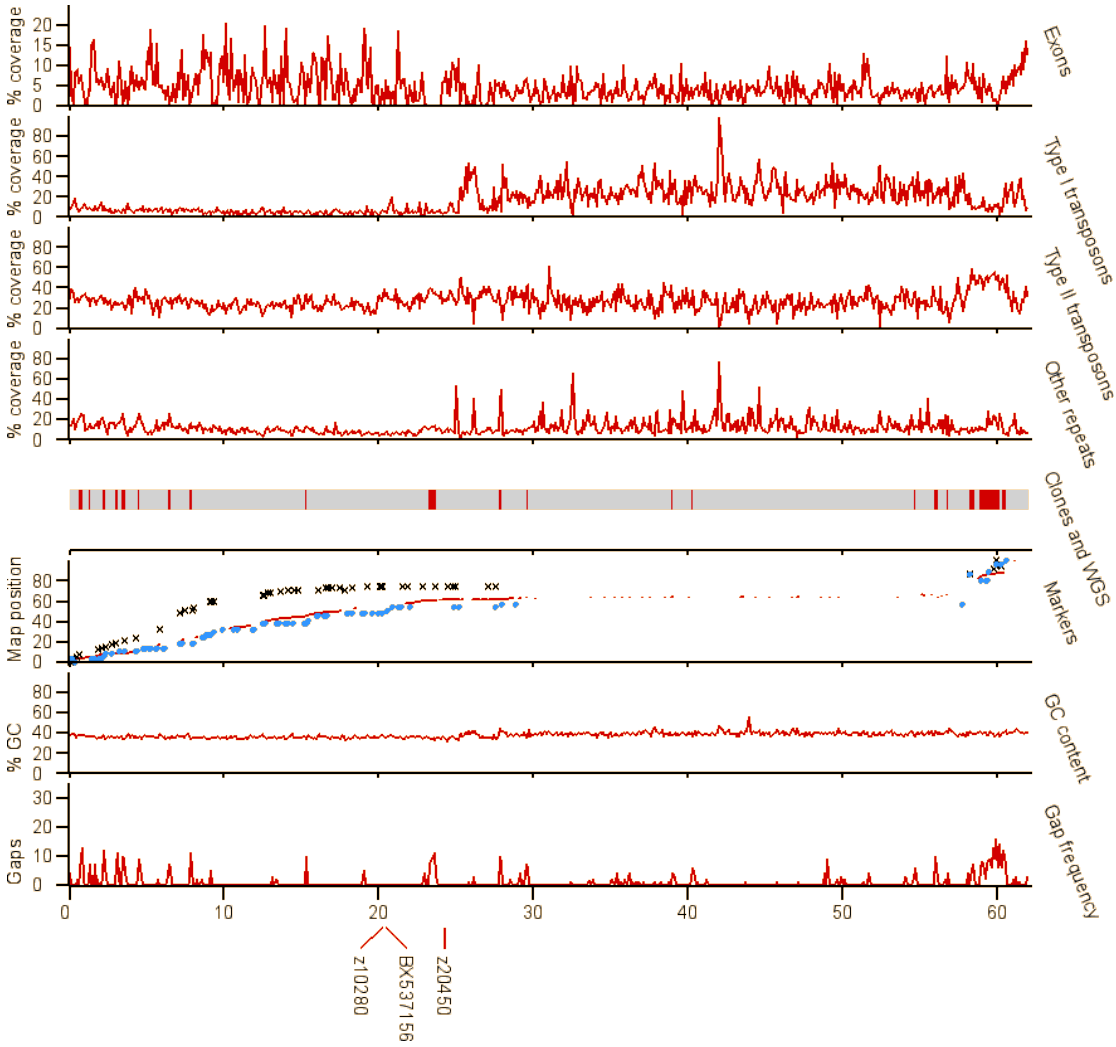


Figure 2. Landscape of LG4
The graphs depict exon and repeat distribution, sequence composition (grey bars = clones, red bars whole genome shotgun contigs), genetic map marker placement (SATmap, red spots; MGH map, blue spots; HS, black X's), GC content and gap density. The near centromeric markers are positioned at 20 Mb (BX537156), 20.2 Mb (Z10280) and 24.4 Mb (Z20450). More details and similar graphs for each chromosome are in the supplementary information.

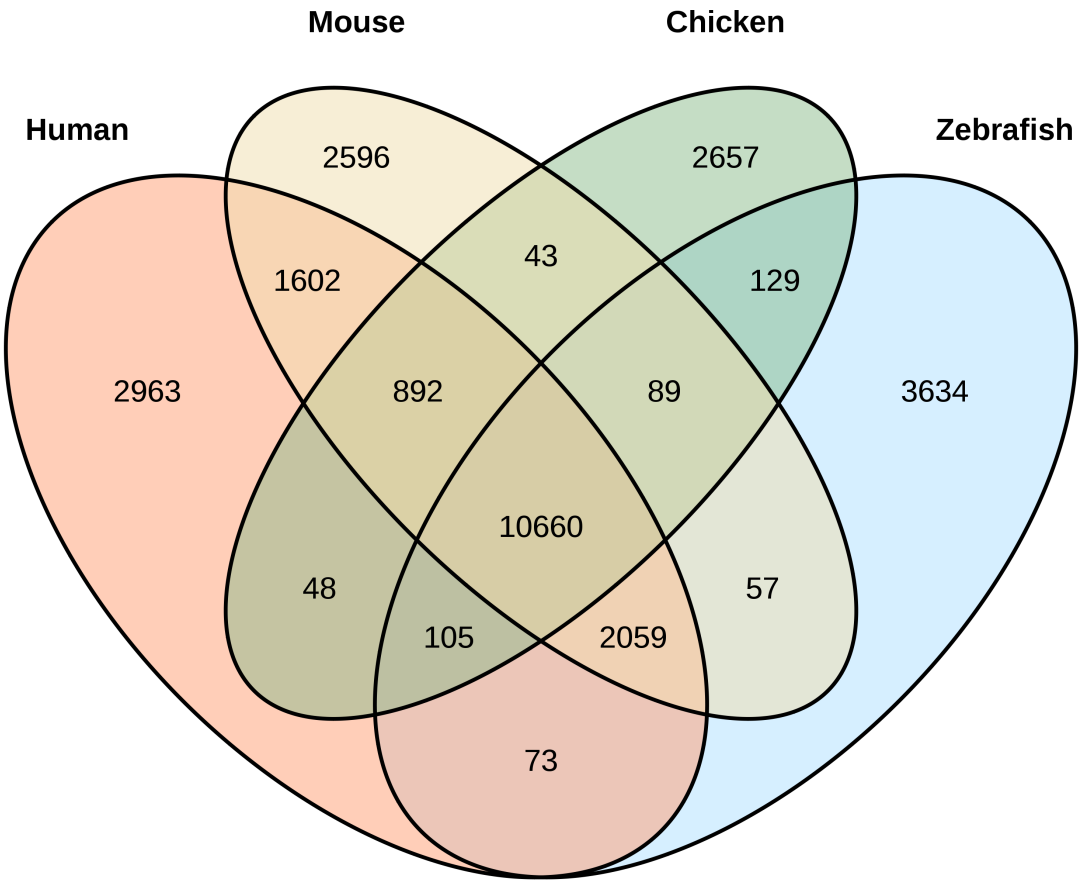


Figure 3. Shared orthologue genes between the zebrafish, human, mouse and chicken genomes using orthology relationships from Ensembl Compara 63. Genes shared across species are considered in terms of copies at the time of the split. For example, a gene that exists in one copy in zebrafish but has been duplicated in the human lineage will be counted as only one shared gene in the overlap.

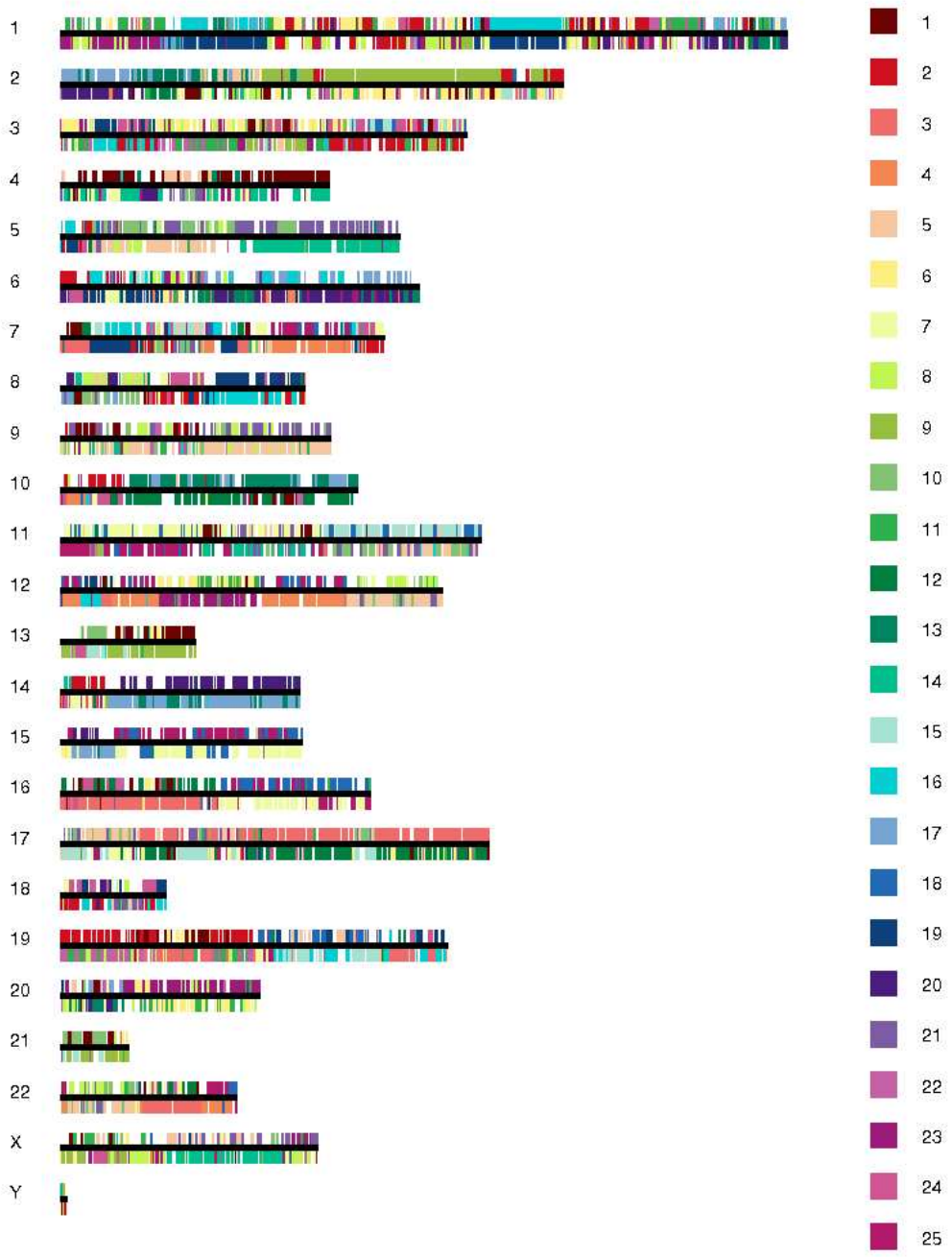


Figure 4. Double-conserved Synteny Between Human and Zebrafish
Each human chromosome is represented as an horizontal black line. The chromosomes carrying the orthologous copies of each human gene in the zebrafish genome are represented by colors on either side of the human chromosome. Runs of genes in conserved linkage between human and zebrafish appear as colored blocks.

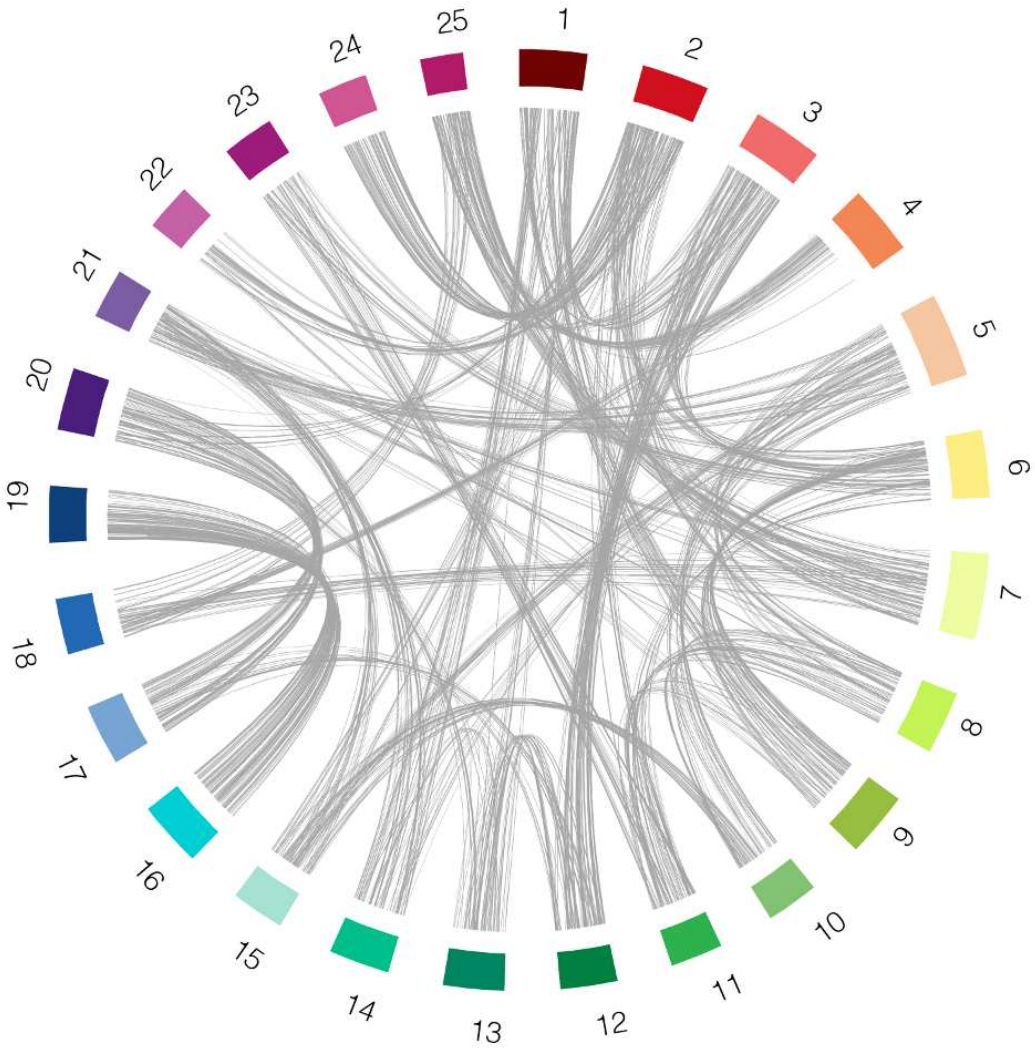


Figure 5. Ohnology Relationships Between Zebrafish Chromosomes
Chromosomes are represented as colored blocks. The position of ohnologous genes between chromosomes are linked in grey (for clarity purposes, links between chromosomes that share less than 20 ohnologs have been omitted). Figure produced with Circos (Krzywinski et al., 2009).