



HAL
open science

Étude de l'évolution de l'ordre des gènes de vertébrés par simulation

Joseph Lucas

► **To cite this version:**

Joseph Lucas. Étude de l'évolution de l'ordre des gènes de vertébrés par simulation. Génomique, Transcriptomique et Protéomique [q-bio.GN]. Université Pierre et Marie Curie - Paris VI, 2016. Français. NNT : 2016PA066140 . tel-01398369

HAL Id: tel-01398369

<https://theses.hal.science/tel-01398369>

Submitted on 17 Nov 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Université Pierre et Marie Curie
École Doctorale 515 Complexité du Vivant
Institut de Biologie de l'École Normale Supérieure de Paris
Génomique Fonctionnelle
Dynamique et Organisation des Génomes

Étude de l'évolution de l'ordre des gènes de vertébrés par simulation

présentée par Joseph LUCAS

Thèse de doctorat en Biologie
pour une soutenance le 17 mai

Sous la direction de : **Mr Hugues Roest Crolius**
Directeur de recherche

Rapporteurs : **Mr David Sankoff**
Professeur des universités
Mr Nicolas Lartillot
Directeur de recherche

Membres du jury : **M^{me} Alessandra Carbone**
Professeur des universités
M^{me} Sèverine Bérard
Maître de conférences
M^{me} Claire Lemaitre
Chargée de recherche

Résumé

Durant les millions d'années qu'ont duré leurs évolutions, les génomes ont subi de nombreux réarrangements chromosomiques. Les plus fréquents sont les inversions, les translocations réciproques, les fissions et les fusions de chromosomes. Pour étudier ces événements nous avons premièrement développé une méthode, PhylDiag, qui, à partir de l'ordre des gènes dans les génomes modernes, identifie les segments de chromosomes qui ont été conservés sans être cassé par les réarrangements. Cette méthode tolère des gènes dupliqués, des clusters de duplications, des erreurs d'annotation, des duplications segmentales et elle identifie également les nombreux segments courts y compris ceux composés d'un gène unique. Dans un deuxième temps nous avons développé Magsimus, un simulateur réaliste qui fait évoluer *in silico* un génome ancestral artificiel en lui faisant subir des réarrangements chromosomiques ainsi que des événements géniques : des duplications, des naissances *de novo* et des délétions de gènes. Avec ce simulateur nous avons tenté de reproduire des évolutions équivalentes à l'évolution réelle qui, à partir d'un ancêtre commun vivant il y a environ 325 million d'années, a abouti à l'humain, la souris, le chien, l'opossum et le poulet. Nous avons utilisé les segments conservés entre les espèces réelles pour calculer une première estimation des nombres de réarrangements le long des branches de l'arbre des espèces. En comparant les génomes modernes simulés aux génomes modernes réels nous avons quantifié le réalisme de ce paramétrage initial puis nous l'avons affiné par une procédure d'optimisation, ce qui nous a simultanément permis d'estimer une distribution des tailles des inversions. Enfin nous montrons avec un exemple simple qu'il sera nécessaire de forcer la réutilisation de points de cassures pour améliorer le simulateur.

Abstract

In the course of evolution genomes have been restructured by massive chromosomal rearrangements. The most common rearrangements are inversions, reciprocal translocations, fissions and fusions of chromosomes. To study these events we first developed a tool, PhylDiag, that uses the conservation of gene order in extant genomes to uncover chromosomal segments that remained unbroken from rearrangements. This tool deals with gene duplications, clusters of duplications, annotation errors, segmental duplications and is able to identify small segments, even those that contain a unique gene. Subsequently, we developed Magsimus, a realistic simulator that evolves *in silico* an artificial genome through chromosomal rearrangements and genic events – gene duplications, *de novo* gene births and gene deletions. With this simulator we made an attempt to reproduce evolutions analogous to the real evolution that happened from a common ancestor, that lived 325 million years ago, to the human, the mouse, the dog, the opossum and the chicken. Conserved segments between these species were used to infer a first estimation of the number of rearrangements that occurred along the branches of the species tree. The realism of this initial parameterisation has been quantified by comparing our simulated extant genomes to the real ones. We then used an optimisation process to correct our estimation while we also estimated the shape of a probabilistic distribution of the lengths of inverted segments. At last, with a simple example, we describe why it will be necessary to enforce breakpoints reuses if we want to improve our simulator.

À mon père
ainsi qu'à toute ma chaleureuse famille,

Remerciements

Je remercie les rapporteurs, M^r David Sankoff et M^r Nicolas Lartillot, les examinateurs, M^{me} Alessandra Carbone, M^{me} Sèverine Bérard et M^{me} Claire Lemaitre d'avoir accepté d'être membres de mon jury. Je les remercie également de lire cette thèse malgré sa longueur. Je remercie plus particulièrement M^r David Sankoff de venir de si loin pour la soutenance. Enfin, je remercie M^{me} Alessandra Carbone d'avoir accepté d'être présidente du jury.

Hugues, merci pour ces 4 années de thèse lors desquelles tu m'as poussé à bout! Dans le bon sens du terme, tu as réussi à quotidiennement entretenir mon désir d'avoir de meilleurs résultats, plus vite et à toujours les expliquer plus clairement et plus pédagogiquement. La culture encyclopédique et l'esthétique *british*, la rigueur chaleureuse et l'humilité avec lesquels tu as animé les travaux sur lesquels nous avons coopéré m'ont profondément inspiré. J'espère que par rapport aux objectifs initiaux, les résultats de mon travail sont à la hauteur de la confiance avec laquelle tu m'as accueilli dans l'équipe. J'aurais aimé faire tellement plus et plus vite. Merci enfin pour tes relectures et tes nombreuses corrections salvatrices qui font de ce travail, plus qu'un travail personnel, une œuvre collective.

Alexandra, merci pour ton talent féminin au sujet des problématiques relationnelles. Merci d'avoir été là dans les moments d'allégresses ainsi que dans des moments plus difficiles. Merci Yves, pour ton énergie contagieuse, merci Christine pour nos déjeuners à cœurs ouverts, Lambert pour ta douceur et les gâteaux de ta mère, Leila pour ton accueil, Lilian pour tes bonbons malaisiens, Céline pour les parties de poker chez toi, Camille pour la précision de tes explications, Magali pour l'intérêt que tu portes au sens commun. Plus particulièrement, je vous remercie Nga, pour ton apport concernant le code de PhylDiag, et Lucas, « über-stagiaire », sans toi la paramétrisation de MagSimus aurait été impossible. Ce fut une grande chance de collaborer avec vous deux. Merci enfin, Matthieu, pour toutes tes fonctions informatiques bien utiles.

Je souhaite également remercier tous les « cousins » des autres familles de l'institut, notamment l'équipe de Denis, et l'équipe des bioinformaticiens. Merci Morgane et Denis d'insuffler toute la joie de vivre énergique dont l'activité scientifique a tant besoin pour s'animer. Merci Samuel, Pauline et tous les étudiants mexicains pour les nombreuses discussions le midi dans la salle de réunion. Merci aux autres équipes qui animent l'institut et qui concourent à le faire rayonner à l'international. Merci Bertrand pour l'initiation au jazz ainsi que pour les nombreux scientifiques que tu m'as fait rencontrer.

Merci Marc Jaekel de nous avoir partagé ta grande culture de la physique, de l'histoire, de l'économie et de la philosophie durant les pauses du midi. Merci à SPIBENS pour les *happy hours*. Merci à toute l'équipe informatique qui gère des quantités d'informations titanesques, de manière à ce que l'informagie puisse faire son effet. Merci Brigitte, tu as révolutionné mon point de vue sur l'administration publique, merci également Abdul et bon courage, la barre est haute. Merci enfin, Wassim, pour ton amitié fidèle, ton écoute, les animations musicales, ainsi que pour ta profonde générosité et ta patience.

Merci également aux membres de mon comité de thèse : Éric Tannier, Simonetta Gribaldo et Nicolas David pour leur écoute et leurs conseils. Merci plus particulièrement, Éric, pour nos nombreuses discussions scientifiques et pour tes critiques qui ont permis de consolider mon travail. Un grand merci Nicolas, pour ton tutorat humain et chaleureux.

Merci aux financements publiques : à l'ANR Ancestrome, au CNRS ainsi qu'au Labex MemoLife qui ont rendu ce travail possible. Merci aux scientifiques du groupe Ancestrome pour nos échanges fructueux sur la thématiques de la reconstruction des génomes ancestraux. Merci également à l'école doctorale 515 ainsi qu'à M^{me} Clément.

Merci très chers amis pour vos messages de soutien et votre confiance malgré ce travail accaparant. Merci pour vos invitations trop souvent refusées pour cause de rédaction. J'aimerais vous nommer et vous dire à quel point vous comptez pour moi, mais je ne saurais plus comment m'arrêter et ce serait indiscret. Merci de nouveau, mes colocataires, pour les blagues le soir (un grand merci), pour nos dîners de colocation et pour l'entraide dans nos tâches quotidiennes.

Merci très spécialement, D^r Quentin, pour ta passion quotidienne et communicative au sujet de la biologie, merci pour tes corrections de ce travail ainsi que pour tes suggestions issues de ta culture Gouldienne et de tes lectures innombrables.

Merci chère famille, pour votre confiance malgré mon éloignement et mes absences répétées aux réunions familiales. Merci chers parents, Marie, Paul, Pierre et Jean-Louis pour vos ondes positives et pour votre affection indéfectible. Merci également magnifiques belles-soeurs et puissant beau-frère pour l'élan nouveau que vous communiquez à la tribu LUCAS. Merci entre autres pour mes neveux et nièces pétillants et pour nos réunions familiales toujours vivantes.

Merci enfin à tous ceux que j'ai oublié et qui se reconnaîtront.

Table des Matières

Introduction	1
1 La structure chromosomique	1
2 Les réarrangements de la structure chromosomique	3
3 Variations phénotypiques associées à un réarrangement	4
4 Les réarrangements durant l'évolution	8
I Définitions	11
I.1 Les génomes	11
I.1.1 Origine du mot gène	11
I.1.2 Localisation des gènes sur les chromosomes	12
I.1.3 La double hélice d'ADN	12
I.1.4 Le code génétique	13
I.1.5 Structure d'un gène de cellule eucaryote	14
I.1.6 Séquençage	15
I.1.7 Annotation	16
I.1.8 Modélisation des génomes adaptée à l'étude des réarrange- ments chromosomiques	17
I.1.9 Conclusion concernant notre modélisation des génomes	24
I.2 Entités	24
I.2.1 Gène	24
I.2.2 Chromosome	26
I.2.3 Génome	28
I.2.4 Espèce	28
I.3 Évènements	28
I.3.1 Évènements géniques	29
I.3.2 Évènements chromosomiques	30
I.3.3 Évènement génomique	31
I.3.4 Évènement relatif à une espèce	32
I.4 Inférences et représentation des évènements relatifs à une espèce	33
I.4.1 Arbre des espèces	33

I.5	Inférence et représentation des évènements relatifs à un gène . . .	33
I.5.1	Arbre de gène	33
I.5.2	Famille d'un gène	35
I.5.3	Relations entre les gènes d'un même arbre	36
I.5.4	Conservation de l'identité d'un gène	38
I.5.5	Gène ancestral	39
I.6	Les points de cassures	40
I.6.1	Vestige d'un flanc de point de cassure	40
I.6.2	Réutilisation des points de cassures	40
I.7	Représentation d'un chromosome	40
I.7.1	Distances et gaps entre deux gènes	40
I.7.2	Clusters de gènes dupliqués en tandem	40
I.7.3	Réécriture d'un chromosome en tandem blocs	42
I.8	Comparaison de deux chromosomes	45
I.8.1	Comparaison des orientations des gènes	45
I.8.2	Matrice d'homologies	45
I.8.3	Matrice de packs d'homologies	46
I.8.4	Distances et gaps entre deux homologies dans une matrice .	46
I.8.5	Diagonales dans une matrice	48
I.9	Comparaison de deux génomes	50
I.9.1	Matrice d'homologies de deux génomes	50
I.9.2	Filtrages des gènes	51
I.10	Bases de données	52
I.10.1	Base de données pour l'arbre des espèces	52
I.10.2	Base de données pour nos génomes et pour les arbres de gènes	52
II	Les segments conservés	53
II.1	Vestiges de génomes ancestraux	53
II.2	Vestiges d'une co-localisation de deux gènes ancestraux	56
II.2.1	Vestiges de deux gènes synténiques	56
II.2.2	Vestiges d'autres co-localisations entre deux gènes ancestraux	57
II.3	Vestiges d'une co-localisation de plusieurs gènes ancestraux . . .	58
II.3.1	Vestiges d'un cluster de gènes ancestraux	59
II.3.2	Les segments conservés, vestiges des segments de chromo- somes non réarrangés	60
II.3.3	Les blocs de synténie, un moyen pour retrouver les segments conservés	72
II.4	Affiner les blocs de synténie pour révéler les segments conservés	79
II.4.1	Prétraitement	80
II.4.2	Post-traitement	81
II.4.3	Évaluation de nos segments conservés	85

II.4.4	Conclusion à propos de l'identification des segments conservés	88
--------	--	----

III	Simuler l'évolution de l'ordre des gènes de vertébrés	92
III.1	État de l'art des simulateurs de l'évolution d'un génome	94
III.1.1	Simulateurs <i>ad hoc</i>	94
III.1.2	Simulateurs dédiés	96
III.2	Inférence des paramètres d'une histoire évolutive à partir des génomés modernes	99
III.2.1	Nombre de gènes dans les génomes ancestraux	99
III.2.2	Nombre d'évènements géniques le long de chaque branche	99
III.2.3	Probabilité qu'un gène inséré par une duplication en tandem ait la même orientation que le gène dupliqué	101
III.2.4	Nombres de chromosomes dans les ancêtres	102
III.2.5	Distribution des gènes dans les chromosomes initiaux	103
III.2.6	Nombre de réarrangements	104
III.3	Le fonctionnement du simulateur MagSimus	107
III.3.1	Génome initial	107
III.3.2	L'évolution est décomposée en évolutions par branches	107
III.3.3	Évènements évolutifs modélisés	107
III.3.4	Liste ordonnée d'évènements à chaque branche	108
III.3.5	Simulations des scénarios les plus simples	109
III.3.6	Contraintes sur les nombres d'évènements	109
III.3.7	Modes de sélection des chromosomes réarrangés	110
III.3.8	Tailles limites d'un chromosome	110
III.3.9	Un réarrangement chromosomique implique au moins un gène ancestral	111
III.3.10	Le gestionnaire de la liste d'évènements	111
III.3.11	Tailles des segments réarrangés	113
III.3.12	Le gestionnaire de segments conservés et des points de cassures	116
III.4	Simulation de cinq amniotes	117
III.4.1	Le génome initial	117
III.4.2	Les nombres et les caractéristiques des évènements géniques	119
III.4.3	Mode de sélection des chromosomes hôtes d'une inversion	120
III.4.4	Mode de sélection des chromosomes dont la taille varie suite au réarrangement	120
III.4.5	Tailles limites des chromosomes	121
III.4.6	Affinage de l'inférence du nombre d'inversions et de translo- cations réciproques	123
III.4.7	Choix de la distribution des tailles d'inversions	125
III.5	Écart d'une simulation par rapport à la réalité	125
III.5.1	Statistiques pour un génome moderne	127

III.5.2	Statistiques pour une comparaison de deux génomes modernes	127
III.5.3	Erreur de réalisme pour une statistique	128
III.5.4	Erreur de réalisme moyenne pour un paramétrage du simulateur	129
III.5.5	Erreur de réalisme synthétique pour toutes les espèces modernes	131
III.5.6	Erreur générale de réalisme	132
III.6	Résultats de MagSimus	134
III.7	Comparaison de nos résultats avec la littérature	139
III.8	Les régions fragiles et la réutilisation des points de cassure	142
III.8.1	Mise en évidence des régions fragiles	143
III.8.2	Illustration des régions fragiles avec les inversions du chromosome X	144
III.8.3	Suggestions pour simuler les réutilisations de points de cassures	147
Discussion		148
D.1	Discussion à propos de notre modélisation des génomes	150
D.1.1	Gènes et paires de bases	150
D.1.2	Gènes non-codants et autres marqueurs	151
D.1.3	Modéliser les centromères	151
D.1.4	Des inversions avec un point de cassure	151
D.1.5	Considérer plus les transpositions	151
D.1.6	Influence des évènements relatifs à une espèce	152
D.2	Discussion à propos des segments conservés	152
D.2.1	Identification des duplications segmentales	152
D.2.2	Comparaison de N génomes au lieu de uniquement 2	153
D.2.3	Choix de la métrique de distance	154
D.2.4	Graphe et linéarisation	155
D.2.5	Définition d'un segment conservé indépendante de notre échelle d'étude	155
D.2.6	Identifier les segments conservés à l'échelle des paires de bases	156
D.2.7	Inversions courtes et erreurs d'assemblage	156
D.2.8	Optimisation du calcul de la matrice d'homologies	156
D.2.9	Matrice d'homologies surprenante	157
D.3	Discussion à propos de notre simulateur	157
D.3.1	Évènements et localisation des gènes	157
D.3.2	Duplications de gènes qui viennent de naître	157
D.3.3	Délétion des gènes précédemment dupliqués	159
D.3.4	Simuler les arbres de gènes réels	159

D.3.5	Insertion d'une copie au voisinage du gène copié	159
D.3.6	Évolution spécifique des chromosomes sexuels	160
D.3.7	Simuler les erreurs d'assemblage et les erreurs d'annotation	160
D.3.8	Paramétrage le plus réaliste	160
D.3.9	Simuler l'évolution des tailles des intergènes	160
D.3.10	Réutilisation des points de cassure et résolution des seg- ments conservés	161
D.3.11	Quantifier le réalisme des réutilisations des points de cassures	162
D.4	Ouverture	162
Annexes		164
A.1	Tailles des inversions attendues d'après le RBM	164
A.2	Implémentation informatique	167
Glossaire		168
Bibliographie		171

Introduction

Où nous avons plus de peine à suivre ces biologistes,
c'est quand ils tiennent les différences inhérentes au germe
pour purement accidentelles et individuelles.

HENRI BERGSON, *L'Évolution Créatrice*, 1907

Les gènes d'un génome sont reliés les uns aux autres par la structure chromosomique. Les réarrangements de cette structure peuvent avoir des conséquences, même si aucune séquence génique n'est altérée [Harewood et Fraser, 2014]. Les réarrangements sont parfois bénéfiques, comme par exemple lorsqu'ils diversifient les anticorps du système immunitaire [Aiden et Casellas, 2015]. Malgré cela, de nombreuses maladies congénitales ainsi que des cancers leurs sont souvent associés [Weischenfeldt *et al.*, 2013].

Avant d'entamer notre étude des réarrangements chromosomiques nous introduirons quelques autres enjeux majeurs qui leurs sont liés. Dans un premier temps nous ferons écho aux recherches qui donnent une importance nouvelle à la structure chromosomique. Nous décrirons ensuite les conséquences phénotypiques que peuvent avoir les réarrangements de cette structure. Puis nous expliquerons pourquoi nous avons choisi de simuler ces évènements pour évaluer le réalisme des théories concernant la logique des réarrangements durant l'évolution.

1 La structure chromosomique

Dans le noyau des cellules eucaryotes, la double hélice d'ADN est enroulée en nucléosomes. Chaque nucléosome est une nano-structure composée de huit histones qui interagissent entre elles et avec leur environnement par des modifications spécifiques de leurs extrémités (« queues ») qui s'étendent dans le nucléosome où à sa surface [Felsenfeld et Groudine, 2003].

Les chromosomes sont compactés en de multiples structures supérieures : à une échelle supérieure à 500 kilobases (kb), mis à part les régions

d'hétérochromatine et d'euchromatine, l'organisation des chromosomes s'apparente à celle d'un « globule fractal », une organisation dense, invariante par changement d'échelle et isotropique [Lieberman-Aiden *et al.*, 2009][Sanborn *et al.*, 2015]. De plus chaque chromosome a une location préférentielle, un « territoire » [Cremer et Cremer, 2010] à l'interface desquels les chromosomes s'entremêlent plus ou moins [Lanctôt *et al.*, 2007]. Il semble que les chromosomes denses en gènes (comme le chromosome humain 19) occupent préférentiellement le centre du noyau, alors que les chromosomes pauvres en gènes (comme le chromosome 18) occupent des régions plus périphériques [Lanctôt *et al.*, 2007]. De plus, des sous-structures du noyau sont identifiables comme la lamina, le nucléole, les corps de Cajal ainsi que les usines transcriptionnelles [Osborne *et al.*, 2004].

À l'échelle des dizaines de kb, intermédiaire entre le nucléosome (146 basepairs (bp)) et le demi mégabase (Mb), les chromosomes semblent extrudés en boucles [Rao *et al.*, 2014] dont les extrémités sont délimitées par deux protéines CCCTC-binding factors (CTCFs) et un anneau de cohésine [Sanborn *et al.*, 2015]. Ces boucles, d'une longueur moyenne de 185 kb, se touchent peu entre elles et délimitent des régions préférentielles d'interactions parfois nommées domaines de contacts ou Topologically Associated Domains (TADs), au sein desquels sont souvent regroupés des promoteurs et des activateurs (enhancers) associés à l'activation de gènes [Shlyueva *et al.*, 2014]. Ce dernier point vaut à ces domaines d'être comparés à des « aires de jeux » où activateurs, promoteurs et gènes coopèrent pour coordonner temporellement l'expression génique [Delpretti *et al.*, 2013], cette structure est également appelée un « hub » chromatinien [de Laat et Duboule, 2013]. De manière générale la régulation des gènes semble faire intervenir des séquences régulatrices situées à moins d'un mégabase du gène qu'elles régulent, souvent au sein d'un TAD, comme nous venons de le voir, ou en périphérie [de Laat et Duboule, 2013][Shlyueva *et al.*, 2014]. Une région qui contient de nombreuses interactions internes entre des séquences *cis*-régulatrices et des gènes, est parfois appelée un bloc de régulation génomique, Genomic Regulatory Block (GRB) [Kikuta *et al.*, 2007]. L'une des plus longues *cis*-régulations qui ait été confirmée expérimentalement relie un activateur et un promoteur de gène situés l'un de l'autre à 1,3 Mb d'écart; le dysfonctionnement de cette régulation provoque la maladie de la Séquence de Pierre Robin [Benko *et al.*, 2009]. Des interactions entre des séquences *trans*-régulatrices et des gènes cibles situés sur des chromosomes différents ont parfois lieu. Par exemple, dans le génome de la souris, une séquence activatrice unique interagit spécifiquement avec un des 1300 gènes codants pour les récepteurs olfactifs, bien que ces derniers soient pourtant disséminés sur plusieurs chromosomes. C'est pourquoi, au final, un unique gène est exprimé dans chaque neurone sensoriel de l'olfaction

[Lomvardas *et al.*, 2006]. D'autres exemples de régulations par « embrassades » inter-chromosomiques [Kioussis, 2005] ont été proposées [Spilianakis *et al.*, 2005][Williams *et al.*, 2010], néanmoins les confirmations expérimentales sont encore anecdotiques.

La conformation tridimensionnelle des chromosomes est dynamique. Par exemple le cluster de gènes *HoxD* est situé entre deux TADs et il migre d'un TAD à l'autre lors de la morphogénèse de l'avant bras des souris [Andrey *et al.*, 2013]. Dans ce cas les deux TADs qui flanquent le cluster sont des régions chromosomiques sans gènes (des « déserts de gènes ») et des séquences activatrices localisées dans ces deux TADs déclenchent l'expression des gènes *Hox* dès que ceux-ci s'en rapprochent. De nombreux déserts de gènes similaires, regroupés par paires autour d'un petit nombre de gènes, sont conservés entre l'humain et le poulet [Ovcharenko *et al.*, 2005] et de manière générale, les TADs sont fortement conservés d'une lignée cellulaire à une autre, ainsi qu'entre l'humain et la souris [Rao *et al.*, 2014].

2 Les réarrangements de la structure chromosomique

Parmi les modifications de la structure chromosomique, les variations *déséquilibrées* altèrent le contenu en gène. À l'inverse, les réarrangements *équilibrés* changent la localisation des gènes sans qu'aucun gène ne soit ajouté ou supprimé du génome. Par exemple, une duplication de gène (ou la délétion d'un gène) est une variation déséquilibrée, alors que l'inversion d'un segment de chromosome, dont les extrémités du segment inversé sont localisées en dehors des séquences géniques, est un réarrangement équilibré. De nombreuses maladies sont liées aux variations déséquilibrées (effet de dosage génique, Copy Number Variations (CNVs)) par contre les conséquences phénotypiques des réarrangements équilibrés sont encore peu connues. Par la suite nous considérons que les réarrangements sont équilibrés, sauf s'il est explicitement mentionné que les réarrangements altèrent une séquence génique.

D'après Motoo Kimura [Kimura, 1984], la première mutation identifiée en 1901 par Hugo de Vries lors de ses travaux sur l'herbe aux ânes (*Oenothera lamarckiana*), était manifestement un réarrangement chromosomique et non une mutation ponctuelle. Vingt ans plus tard, le premier réarrangement chromosomique est clairement mis en évidence par Sturtevant, alors qu'il établit les cartes génétiques de mouches drosophiles [Sturtevant, 1921]. Il s'agit de l'inversion d'un segment de chromosome. Mis à part les recombinaisons V(D)J du système immunitaire, d'autres réarrangements chromosomiques

ont été identifiés depuis : les translocations réciproques, les translocations Robertsoniennes, les transpositions, les fissions et les fusions de chromosomes [Griffiths, 2005] ainsi que les chromothripsis [Zhang *et al.*, 2015] (lors desquels un segment de chromosome est massivement réarrangé). Les gènes sont déplacés de toutes ces différentes manières le long des chromosomes. Savoir à quel point la relocalisation des gènes le long des chromosomes modifie le phénotype d'un organisme est un sujet sur lequel plusieurs points de vues existent dans la communauté scientifique. Certains affirment, qu'à part à des échelles locales, il n'y a pas de structuration des génomes [Koonin, 2009], ce qui tendrait à donner un rôle majoritairement neutre aux réarrangements. D'autres, au contraire, sont convaincus que les génomes sont organisés d'une manière fondamentalement non-aléatoire [Misteli, 2007][Misteli, 2009]. De nombreuses questions sont en suspens. Les localisations des gènes, mis à part les contraintes des structures locales, varient-elles le long des chromosomes, sans affecter l'organisme ? Les gènes bougent-ils souvent au cours de l'évolution ? Quels sont les réarrangements les plus fréquents ? Existe-t-il des régions où l'ordre des gènes est moins réarrangé que d'autres régions ? Des régions ont-elles été réarrangées plusieurs fois ? Y a-t-il des réarrangements réversibles ? Les segments réarrangés sont-ils majoritairement courts, ou longs ? Est-ce que l'évolution de la localisation des gènes diffère selon les organismes ? Quels rôles les réarrangements ont-ils eu dans l'évolution et dans l'origine des espèces ? À quelles fréquences ont-ils modifiés l'architecture des génomes modernes ? Pouvons-nous prédire les régions où ils ont le plus de chance d'avoir lieu ?

Pour illustrer l'intérêt de répondre à ces questions concernant les réarrangements nous exposons quelques exemples de variations phénotypiques qui leurs sont associés.

3 Variations phénotypiques associées à un réarrangement

Comme nous le mentionnions, les réarrangements, même équilibrés, peuvent avoir de multiples conséquences sur l'expression des gènes [Harewood et Fraser, 2014]. Nous décrivons quelques exemples qui nous paraissent remarquables.

Lorsqu'un réarrangement déplace un gène, les fonctions associées au gène déplacé sont parfois affectées. Ce phénomène est appelé « effet de position ». La première mise en évidence de cet effet est due à Muller. En 1930, il induit, par rayons X, l'inversion d'un segment du chromosome sexuel X d'une gonade de mouche drosophile. La mouche femelle qui hérite

de ce chromosome muté (ainsi que d'un chromosome X d'une mouche aux yeux blancs) manifeste une mosaïque de cellules visuelles rouges et blanches (phénomène appelé variégation), alors qu'en l'absence de l'inversion les cellules de l'œil sont toutes rouges [Muller, 1930]. Il est assez clair aujourd'hui que la variégation des cellules rouges et blanches est due à la relocalisation de l'allèle dominant w^1 à proximité d'une région de chromatine hautement compactée, l'hétérochromatine, ce qui réprime partiellement l'expression de ce gène responsable de la coloration rouge.

Autre exemple, le développement embryonnaire des vertébrés dépend de la coopération spatialement ordonnée des gènes du cluster *HoxD*. Le rôle de l'ordre de ces gènes dans la morphogénèse de l'avant bras est maintenant bien compris [Andrey *et al.*, 2013] et il a été montré, pour les souris, qu'une inversion, même locale, de l'ordre des gènes altère le développement de ce membre [Spitz *et al.*, 2005].

Les inversions semblent également avoir joué un grand rôle dans la différenciation des chromosomes sexuels X et Y chez les thériens. Il est actuellement admis que ces deux chromosomes étaient initialement des chromosomes autosomiques homologues et que des inversions successives ont causé leur divergence en limitant graduellement les recombinaisons méiotiques entre eux [Lemaitre *et al.*, 2010].

Autre conséquence probable des inversions, celles-ci semblent avoir facilité les spéciations sympatriques (resp. parapatriques) [Hoffmann et Rieseberg, 2008][Kirkpatrick, 2010], lorsqu'une nouvelle espèce émerge d'une communauté ancestrale alors que les individus ont continuellement habité la même région géographique (resp. une région géographique périphérique). Les inversions semblent également avoir participé aux adaptations locales, lorsque différents gènes sont positivement sélectionnés dans différents environnements. Une inversion qui englobe des gènes adaptés à un environnement local aura un avantage sélectif qui lui permettra de se répandre dans la population. L'avantage est conféré par la suppression de la recombinaison, car après quelques générations le segment inversé contient toujours les allèles co-adaptés à l'environnement alors que l'ancien arrangement contient, lui, un mélange d'allèles adaptés et non adaptés. Les inversions se propageront car elles empêchent les recombinaisons de séparer des ensembles d'allèles qui ont des effets synergétiques. L'analyse des génomes de deux espèces issues d'une spéciation sympatrique ou parapatrique, montre que les régions les plus différenciées des génomes sont regroupées en « îlots génomiques » [Via, 2009], ou Highly Divergent Regions (HDRs) [Malinsky *et al.*, 2015]. Les génomes de sous-

¹*white*, l'allèle récessif, qui, lorsqu'il est muté génère des mouches homozygotes aux cellules visuelles blanches.

populations d'une même espèce en cours de différenciation, ont également des îlots génomiques. Ces îlots sont moins recombinaisonnés que les autres régions du génome, ce qui semble bien s'expliquer par des inversions [Lowry et Willis, 2010][Yeaman, 2013]. Cependant le rôle des réarrangements dans la formation des îlots génomiques est encore débattu et l'existence de ces îlots pourrait être due à un effet d'« auto-stop génique » (divergence hitchhiking theory) autour d'allèles sous pression de sélection divergente [Via, 2012].

Entre deux génomes humains, de nombreuses inversions polymorphiques ont été décrites [Sudmant *et al.*, 2015]. En Europe une inversion de 900 kb sur le bras court du chromosome 17 est portée par 80% de la population européenne [Stefansson *et al.*, 2005] et il semble qu'une pression de sélection négative ait défavorisé l'ancien variant (H2) par rapport au nouveau (H1) [Zody *et al.*, 2008].

De grandes inversions sont associées à des polymorphismes phénotypiques stables. Les gènes associés spécifiquement à un phénotype sont parfois soudés les uns aux autres par les inversions et rassemblés dans des structures génomiques nommées « supergènes » [Thompson et Jiggins, 2014]. Par exemple, les motifs sur les ailes de certains papillons sont associés à une inversion [Joron *et al.*, 2011]. De même, deux organisations sociales des colonies de fourmis (*Solenopsis*) sont associées à une inversion de plus de 9 Mb [Bourke et Mank, 2013][Wang *et al.*, 2013]. Dans une organisation la reine est monandre (elle s'accouple avec un mâle) alors que dans l'autre la reine est polyandre (elle s'accouple avec plusieurs mâles). Les deux organisations diffèrent de plus par leurs agressivités vis à vis des autres colonies ainsi que sur la manière d'initier la fourmilière. Une paire de chromosomes « sociaux », comparables aux chromosomes sexuels X et Y, rassemblent les gènes co-exprimés spécifiquement dans chacun des deux cas. Comme précédemment l'inversion de 9 Mb, semble avoir permis à ces chromosomes de se différencier. Un troisième supergène, qui correspond là encore à une région chromosomique inversée, contrôle le tri-morphisme des combattants variés mâles (*Philomachus pugnax*) [Jiggins, 2015][Küpper *et al.*, 2015], des oiseaux échassiers migrateurs. Selon le statut de l'inversion les mâles arborent des collerettes nuptiales différentes et ils ont différentes manières de courtiser les femelles dans les aires de parade.

Outre les inversions, les translocations altèrent parfois l'expression des gènes [Harewood *et al.*, 2010]. Certaines translocations altèrent même des interactions intra-chromosomiques entre deux gènes alors que ceux-ci sont séparés par plus de 14.3 Mb [Levesque et Raj, 2013].

Les réarrangements qui modifient les GRB ont une probabilité *a priori* non négligeable d'avoir des effets délétères et la conservation de ces régions peut être mise à profit pour les identifier par une approche de génomique comparative [Naville *et al.*, 2015].

Enfin, il semble que de nombreuses translocations soient à l'origine de cancers [Mitelman *et al.*, 2007] et il y a une corrélation indiscutable entre la formation de tumeur cancéreuses et la présence de nombreux réarrangements chromosomiques [Stratton *et al.*, 2009][Wijchers et de Laat, 2011]. Dans les génomes de cellules cancéreuses, le réarrangement le plus surprenant qui ait été identifié est certainement le chromithripsis qui pulvérise un segment de chromosome avant de rabouter les fragments [Stephens *et al.*, 2011].

Nous arrêtons l'énumération des conséquences fonctionnelles des réarrangements et nous allons maintenant exposer les arguments qui tendent à minimiser leur rôle. Car, bien qu'il y ait de nombreuses modifications phénotypiques liées aux réarrangements, il se peut que ceux-ci soient majoritairement neutres.

La plupart des réarrangements équilibrés sont probablement neutres

Des segments de chromosomes de souris, longs de plus d'un mégabase, semblent neutres et « supprimables ». Les souris à qui l'on a retiré ces segments de chromosomes ne représentent pas d'altérations identifiables du phénotype par rapport aux autres et elles sont viables et fertiles [Nóbrega *et al.*, 2004]. Une inversion incluse dans un long segment de chromosome superflu aura donc *a priori* peu de conséquences phénotypiques également.

Alors que certains travaux considèrent que les gènes co-exprimés sont regroupés en clusters [Hurst *et al.*, 2004], d'autres travaux ont trouvé une faible corrélation entre co-localisation des gènes et co-expression des gènes [Sémon et Duret, 2006]. D'après ces derniers travaux il semble que les gènes, co-exprimés et proches les uns des autres, soient quasi-exclusivement les gènes qui partagent le même bi-promoteur, où ceux qui sont transcrits en « read-through ».

D'autres travaux répertorient de multiples arguments qui laissent penser qu'il n'y a pas d'architecture génomique à grande échelle mais que par contre des structures locales (blocs de régulation, les opérons des bactéries, les origines de réplication) jouent un rôle important dans la conservation de la localisation des gènes à petite échelle [Koonin, 2009].

De plus, lorsque des génomes de mouches sont comparés, la proportion de gènes en synténie partagée (sur le même chromosome dans les deux génomes comparés) est proportionnelle à la similarité moyenne des séquences protéiques [Zdobnov et Bork, 2007], ce qui semble indiquer une fois de plus que les réarrangements sont principalement neutres [Koonin, 2009].

Quelques estimations systématiques des réarrangements ont été effectuées

dans les génomes humains [Feuk et Carson, 2006][Warburton, 1991]. Elles ont révélé que de nombreuses inversions courtes et polymorphiques existent : récemment 786 inversions polymorphiques ont été répertoriées (d'une taille moyenne de 1,7 kb) [Sudmant *et al.*, 2015] et 617 inversions semblent confirmées sur la base de plusieurs études agglomérées dans la base de donnée InvFEST [Martínez-Fundichely *et al.*, 2014]. Mais, de manière générale, la fréquence à laquelle les réarrangements se sont fixés dans les espèces, durant l'évolution, est encore peu connue.

4 Les réarrangements durant l'évolution

Pour identifier les inversions polymorphiques que nous venons de mentionner les chercheurs ont comparé les génomes de différents humains. Sans rentrer dans les détails, deux segments chromosomiques similaires, l'un inversé dans un génome et l'autre non inversé dans un autre génome, correspondent intuitivement à une inversion de segment de chromosome durant l'évolution du génome de leur ancêtre commun jusqu'à un des génomes modernes comparés¹. Dans ce travail, nous suivrons cette démarche, de génomique comparative, pour identifier et étudier les vestiges des réarrangements dans les génomes d'individus de différentes espèces.

Durant l'évolution des espèces, mis à part les fusions, les réarrangements chromosomiques brisent les chromosomes en différents *points de cassures* dont les vestiges sont identifiables dans les génomes modernes. Étant donné qu'au début des recherches sur les réarrangements, ceux-ci étaient encore mal connus, les premières études à but exhaustif les ont abordé à travers le prisme de ces vestiges de points de cassures. Les nombres de cassures et les distributions spatiales de celles-ci le long des chromosomes furent d'un intérêt tout particulier.

En 1984, Nadeau et Taylor [Nadeau et Taylor, 1984] comparent les localisations de plus de 83 marqueurs (des séquences chromosomiques de mêmes origines ancestrales) le long des chromosomes de l'humain et de la souris. La conservation de l'ordre de ces marqueurs, dans 13 vestiges de segments chromosomiques ancestraux, semble correspondre à des réarrangements qui ont brisé les chromosomes en des localisations dont la densité de probabilité est uniforme le long des chromosomes. C'est le modèle aléatoire de répartition des points de cassures, aussi nommé Random Breakage Model (RBM). Au cours des années qui suivirent, les vestiges de segments conservés ont été

¹En toute rigueur il serait nécessaire de prendre des précautions pour s'assurer que cette inversion n'ait pas été causée par un autre scénario de réarrangement, de résultat équivalent et moins parcimonieux.

identifiés de plus en plus précisément, au fur et à mesure que le nombre de marqueurs augmentait. Durant presque 20 ans, le RBM a prédit les résultats de la communauté scientifique concernant la distribution spatiale des points de cassures ainsi que la distribution des tailles de segments de chromosomes conservés sans réarrangement. Jusqu'en 2003 la distribution des tailles de segments conservés, entre l'humain et la souris, s'ajuste à une exponentielle décroissante de paramètre égal à la taille moyenne des segments conservés, environ 8 centiMorgans (cM).

Fin 2002 le génome de la souris est séquencé et deux auteurs, Pevzner et Tesler, le comparent au génome humain [Pevzner et Tesler, 2003a][Pevzner et Tesler, 2003b], séquencé l'année précédente. Les scénarios évolutifs qui minimisent les nombres de réarrangements entre ces deux génomes suggèrent que, contrairement à ce qui était pensé, les cassures n'ont pas lieu au hasard, avec une distribution uniforme, mais qu'en réalité elles réutilisent fréquemment les mêmes régions. Les auteurs proposent donc un nouveau modèle de répartition des points de cassures, le modèle des régions fragiles, aussi nommé Fragile Breakage Model (FBM). Dans ce modèle les chromosomes sont composés de régions fragiles courtes qui alternent avec des régions solides longues. Une région *fragile* est ici une région qui durant l'évolution, a été brisée par de multiples réarrangements chromosomiques et dont les vestiges sont identifiables dans les espèces modernes. Dans les autres régions, *solides*, il est possible qu'il y ait eu de nombreuses cassures également, mais nous n'en voyons pas les vestiges aujourd'hui. Les cassures dans les régions solides sont donc : soit moins fréquentes que les cassures dans les régions fragiles, soit au moins aussi fréquentes mais moins sélectionnées. Par exemple, les régions solides abritent peut-être des réarrangements négativement sélectionnés et il se peut qu'une cassure dans ces régions perturbe les interactions entre des séquences régulatrices et les gènes que ces séquences régulent [Becker et Lenhard, 2007]. Sous cette dernière hypothèse, les régions solides seraient donc entre autres les GRBs. Les régions fragiles, relativement courtes (de l'ordre de 0,6 Mb en moyenne), sont probablement réparties uniformément le long des chromosomes [Pevzner et Tesler, 2003b] ce qui expliquerait que le RBM suffisait à prévoir la distribution des points de cassures lorsque la résolution des marqueurs était plus grossière que le mégabase [Pevzner et Tesler, 2003b].

Un modèle de distribution des points de cassures récent [Berthelot *et al.*, 2015] tend à rendre compte des différentes sensibilités des régions à être brisées en considérant que seules les régions de chromatine ouverte sont cassables. Dans ce dernier modèle cette propriété est couplée avec une modélisation probabiliste des inversions, les réarrangements qui sont de loin les plus fréquents. Les tailles des segments inversés sont estimées par la

probabilité de contact intra-chromosomique à l'échelle de l'ordre du mégabase. À cette échelle la probabilité de contact d'un premier locus avec un deuxième locus décroît comme l'inverse de la distance génomique, en paires de bases, qui sépare ces deux locus [Lieberman-Aiden *et al.*, 2009]. Ce modèle à la particularité de rendre compte des biais de distribution des points de cassures observés précédemment tout en proposant un mécanisme qui ne fait pas intervenir la sélection.

Enfin, les régions dupliquées semblent jouer un rôle crucial pour amorcer les recombinaisons homologues non alléliques, ce qui a été vérifié expérimentalement en induisant des cassures double brins artificielles par rayons X [Argueso *et al.*, 2008]. Des incertitudes subsistent néanmoins dans la communauté, pour savoir à quel point ces duplications segmentales causent les réarrangements et à quel point les réarrangements causent les duplications segmentales.

Outre les lois qui régissent les localisations spatiales des points de cassures, de nombreuses autres caractéristiques des réarrangements sont inconnues : les fréquences de chaque réarrangement dans les différentes espèces, les fréquences des tailles de segments inversés, *etc.* Pour répondre à toutes ces questions, nous avons premièrement développé PhylDiag, un logiciel qui détecte précisément les segments de chromosomes conservés durant l'évolution. Dans un deuxième temps, nous avons développé MagSimus, un simulateur de l'évolution de l'ordre des gènes de vertébrés, qui nous a permis de quantifier le réalisme de plusieurs modélisations des réarrangements. Avant de détailler ces deux logiciels et leurs résultats, nous commençons par définir clairement notre échelle d'étude.

Chapitre I

Définitions

Cooperation is needed for evolution to construct new levels of organization. Genomes, cells, multicellular organisms, social insects, and human society are all based on cooperation. Cooperation means that selfish replicators forgo some of their reproductive potential to help one another.

MARTIN ANDREAS NOWAK, *Five rules for the evolution of cooperation*, 2006

I.1 Les génomes

I.1.1 Origine du mot gène

En 1865 le moine Gregor Mendel étudie les plants de petits pois et découvre les lois fondamentales de l'hérédité particulaire, aussi connues sous le nom de lois de Mendel. La couleur, blanche ou violette, des fleurs de pois est héritée de leurs parents, via des particules d'hérédité, les allèles. C'est la fin de la théorie de l'héritage par mélange, les fleurs, issues du croisement d'un plant aux fleurs violettes et d'un autre aux fleurs blanches, ne sont pas d'un pâle violet, mais totalement violettes ou totalement blanches selon la présence ou non de l'allèle dominant violet. Les découvertes de Mendel furent longtemps ignorées avant d'être reconnues suite aux travaux de de Vries. En 1889 ce dernier formule une théorie alternative à la théorie de la pangénèse de Darwin dans laquelle les particules d'hérédité sont nommées pangènes. Vingt ans plus tard, en 1909, Johannsen les nommera plus simplement les gènes. L'ensemble des gènes s'appelle le génome et par extension le génome désigne aussi le support matériel des gènes, l'ensemble des chromosomes.

I.1.2 Localisation des gènes sur les chromosomes

En continuant ses travaux, de Vries découvre en 1901 que de nouvelles formes apparaissent soudainement dans la nature et peuvent persister sur plusieurs générations. Ces variations spectaculaires sont nommées mutations et elles seront étudiées largement chez la mouche drosophile par Morgan et Muller au début du vingtième siècle. Les déséquilibres de liaison entre les gènes permettent alors de les localiser le long de macromolécules appelées chromosomes. Les bases de la génétique moderne sont posées : les gènes sont des unités d'hérédité qui correspondent à une portion identifiable de chromosome.

I.1.3 La double hélice d'ADN

Pour cette sous-section je me suis inspiré du chapitre 1 du livre « Computational Genome Analysis » de Deonier, Tavaré et Waterman [Deonier et al., 2005]

La structure principale des chromosomes est découverte en 1953 par Crick et Watson. Un chromosome est majoritairement constitué d'une molécule d'ADN à deux brins qui forment une double hélice. Chaque brin est un polymère composé de millions de monomères, les nucléotides mono-phosphates A, T, C et G. À chaque nucléotide correspond une base azotée particulière liée à un sucre désoxyribose qui lui même est lié à un groupe phosphate. La position 5' du sucre de chaque nucléotide est connectée via un groupe phosphate à la position 3' du sucre du nucléotide qui lui précède directement, voir la figure I.1.

Chaque brin d'ADN a une extrémité 5', correspondant au groupe phosphate attaché à la position 5' du sucre du premier nucléotide, et une extrémité 3', correspondant au groupe -OH à la position 3' du sucre sur le dernier nucléotide. Les deux brins sont antiparallèles, c'est à dire que les deux chaînes polynucléotidiques ont des orientations 5'-3' opposées. On observe habituellement un appariement des bases : A s'apparie avec T et G s'apparie avec C. Deux brins d'ADN sont dits complémentaires s'ils ont des séquences qui leurs permettent de s'apparier et c'est ce qui arrive le plus souvent. Une paire de deux bases complémentaires s'appelle communément une paire de base, ou « base pair » en anglais ou plus simplement bp. C'est l'unité classique pour quantifier la longueur d'un chromosome. Ainsi la longueur du chromosome 1 de l'humain fait environ 285 000 000 bp, soit 285 Mb. Une molécule d'ADN peut donc être représentée par une séquence de lettres tirées dans l'ensemble {A,T,C,G}, avec l'orientation gauche-droite de la séquence correspondant à l'orientation 5'-3' d'un des deux brins d'ADN. L'autre brin se déduit par

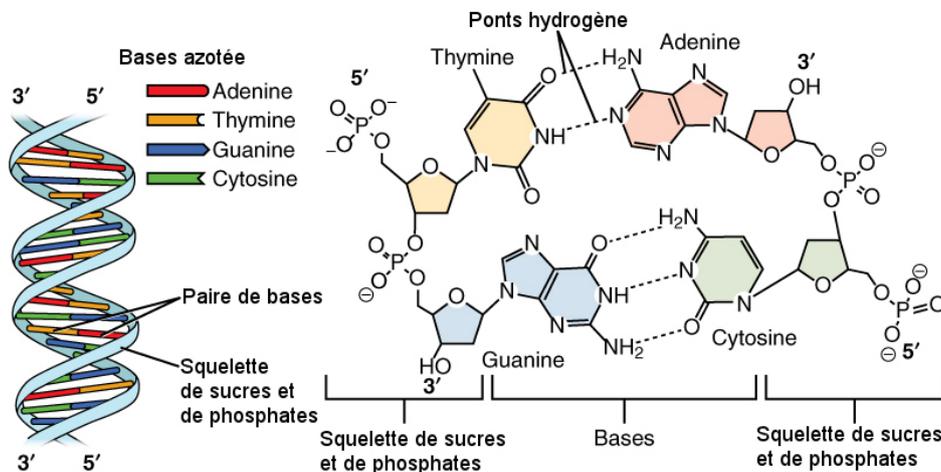


Figure I.1 – Structure de l’ADN. (traduit de <https://en.wikipedia.org/wiki/Nucleotide> ©). À gauche, la structure en double hélice. À droite, la structure chimique des nucléotides monophosphates ainsi que les ponts hydrogènes qui relient les nucléotides complémentaires.

complémentarité.

I.1.4 Le code génétique

Comme précédemment cette sous-section est inspirée du chapitre 1 du livre « *Computational Genome Analysis* » de Deonier, Tavaré et Waterman [Deonier et al., 2005].

Suite à la découverte des séquences de nucléotides une question reste en suspens : Comment les gènes composés de nucléotides peuvent-ils porter l’information nécessaire à la synthèse des molécules polypeptidiques, les protéines, qui constituent la majorité des molécules structurales des organismes vivants ? Cette question trouve une ébauche de réponse au début des années 1960 lorsque le code génétique, qui permet la traduction de l’information encodée par les 4 nucléotides de l’ADN jusqu’aux 20 acides aminés des protéines, est déchiffré¹. La traduction des nucléotides aux acides aminés se fait par triplets de nucléotides appelés codons et à chaque codon correspond un acide aminé. Il y a néanmoins trois exceptions à cela, les codons TAA, TAG et TGA ne codent pas pour des acides aminés mais ils correspondent à des signaux de fin de traduction chez la plupart des eucaryotes. Enfin, le codon

¹Le code génétique possède néanmoins quelques variations [Osawa *et al.*, 1992], notamment dans les mitochondries, mais les différences sont marginales et il est courant de considérer le code comme valide dans la plupart des espèces.

ATG, en plus de coder pour une méthionine, correspond souvent, mais pas tout le temps, au codon de début de la séquence traduite en protéines.

I.1.5 Structure d'un gène de cellule eucaryote

Une meilleure connaissance de la structure des gènes de cellules eucaryotes sera acquise en 1977 à la suite des expériences de Sharp et Roberts qui étudient indépendamment l'un de l'autre la réplication de l'adénovirus dans les cellules humaines [Cooper, 2000]. Ils découvrent chacun de leur côté que les séquences géniques sont partitionnées en introns et en exons. Les exons sont les séquences codant pour les protéines (nommées également Coding Sequence (CDS)) et les introns sont des séquences transcrites mais non traduites. Dans un premier temps, la séquence de nucléotides d'un brin d'ADN correspondant à un gène, est transcrite par l'ARN polymérase en une molécule simple brin appelée pre-ARNm. Cette molécule est une molécule d'ARN qui diffère de l'ADN par la présence d'un sucre ribose à la place du sucre désoxyribose de l'ADN. En plus de cela le nucléotide de base azotée Thymine (T) de l'ADN n'existe pas dans l'ARN et il y est substitué par le nucléotide de base azotée Uracile (U). La séquence de nucléotides de la molécule pre-ARNm est une copie de la séquence génique à ceci près que les nucléotides T de l'ADN sont transcrits en nucléotides U. Cet ARN est ensuite épissé (on dit qu'il mature), les séquences correspondant aux introns et à certains exons sont excisées à ce moment. L'épissage peut aboutir à différents ARN messagers (ARNm) selon les exons conservés. La figure I.2 illustre deux épissages alternatifs d'un gène, contenant 6 exons et 5 introns, qui aboutissent à deux ARNm. L'ARNm mature sort ensuite du noyau de la cellule où il est traduit en protéine par les ribosomes. À partir du codon AUG de début de transcription, le ribosome traduit les codons de l'ARNm trois par trois selon la direction 5'-3' en suivant le code génétique, et cela jusqu'au codon de fin de traduction UAA. Les protéines qui résultent de l'expression des gènes sont les constituants moléculaires fondamentaux du phénotype des organismes vivants.

Une courte séquence de l'ARNm à l'extrémité 5', en amont du codon de début de transcription, n'est pas traduite en protéine. C'est la 5' Untranslated region (UTR) qui contient des signaux pour l'amorçage de la traduction. De même à l'autre extrémité de l'ARNm, à l'extrémité 3', la région 3' UTR, qui débute par le codon de fin de transcription, est une région non traduite.

En amont de l'extrémité 5' du gène, en plus du promoteur, il y a parfois des régions qui contrôlent la transcription du gène, ce sont les régions régulatrices, aussi appelées Transcription Factor Binding Sites (TFBSs). Elles peuvent contenir des sites de fixation pour des protéines qui vont s'y lier pour faciliter la fixation de l'ARN polymérase au promoteur.

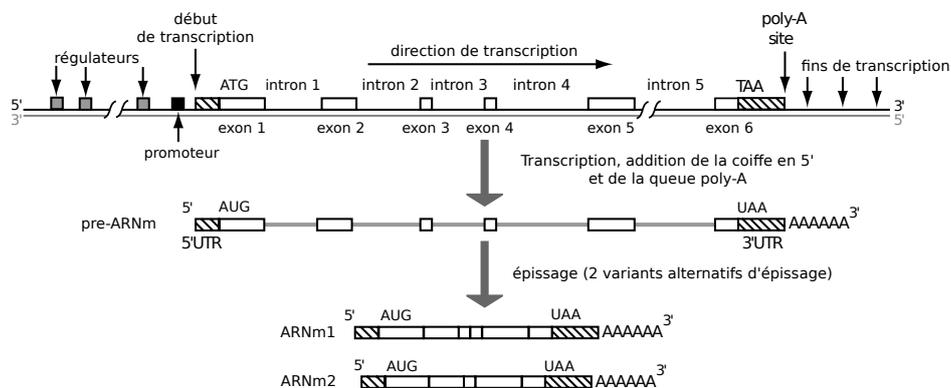


Figure I.2 – Transcription et épissages alternatifs d’un gène de cellule eucaryote. Image modifiée à partir d’une illustration du livre Computational Genome Analysis [Deonier *et al.*, 2005].

Comme nous venons de le voir certains codons permettent d’identifier le début et la fin de la traduction du gène. Dans notre exemple il s’agit respectivement des codons ATG et TAA sur l’ADN, AUG et UAA sur les molécules d’ARN. L’analyse de la distribution de ces codons le long de la molécule d’ADN par rapport à une répartition aléatoire des nucléotides permet d’identifier les séquences possiblement traduites en protéines, elles sont appelées les fenêtres ouvertes de lectures ou Open Reading Frames (ORFs).

I.1.6 Séquençage

La connaissance de la structure moléculaire des gènes s’accroissant, il devint utile de déterminer le plus exhaustivement possible la succession des nucléotides le long des chromosomes. Cela s’appelle aujourd’hui séquencer un génome. En 1977, Sanger séquence le premier génome, l’ADN d’un bactériophage d’une longueur d’environ 5 kb. Le premier séquençage d’un organisme auto-reproducteur fut celui du génome de la bactérie *Haemophilus influenzae* en 1995, celui-ci contient plus de 1,8 Mb. En 2001, le génome humain¹ fut le premier génome de vertébré à être séquencé, fin 2002 ce fut celui de la souris, en 2004 le poulet, *etc.* En 2015, la base de donnée Ensembl collecte 64 génomes de vertébrés. Par abus de langage nous écrivons que le génome d’une espèce a été séquencé, alors qu’il s’agit bien entendu, la plupart du temps, du génome d’un individu d’une espèce. Néanmoins, comme le polymorphisme intra-spécifique est relativement faible chez les vertébrés [Leffler *et al.*,

¹Plus de 70% du génome séquencé par le HGP (Human Genome Project) provient d’un humain anonyme de la ville de Buffalo à New York, nom de code RP11.

2012], ces génomes peuvent être considérés, dans une certaine mesure, comme représentatifs de tous les génomes de leur espèce. Ceci sera d'autant plus vrai que l'échelle d'étude sera grande et qu'elle négligera la plupart du polymorphisme, dû à des substitutions de nucléotides, à des insertions/délétions nucléotidiques, à de courtes micro-inversions ainsi qu'à des CNVs.

I.1.7 Annotation

Progressivement les milliards de paires de bases des génomes sont annotées par fonctions biologiques et autres caractéristiques remarquables.

Masquer les répétitions disperses

Plus de 47% [Lander *et al.*, 2001] du génome humain est composé de séquences répétées. Certaines de ces séquences sont des répétitions localisées en tandem. Ces séquences satellites sont particulièrement nombreuses autour des centromères et au voisinage des télomères. Les autres séquences répétées sont les répétitions dispersées correspondant à des éléments transposables (SINE, LINE, LTR et divers transposons tels que les séquences de virus endogènes et les transposons à ADN [Deonier *et al.*, 2005]). Analyser un génome débute la plupart du temps par l'identification et le masquage de ces séquences qui contiennent peu d'information et qui sont extrêmement redondantes. La deuxième étape est généralement dévolue à l'identification des séquences correspondant aux gènes.

Annotation des gènes

Après le masquage des séquences répétées, les localisations des gènes sont annotées. Les séquences d'acides aminés des protéines et les séquences nucléotidiques des transcrits trouvées dans l'organisme et dans le cytoplasme des cellules, par séquençage protéique, RNA-seq ou cDNA sont alignées sur les chromosomes en s'aidant du code génétique pour les séquences protéiques. Une validation croisée des alignements est effectuée avec une détection *ab initio* des ORFs. Il est ainsi possible de localiser les gènes codant pour les protéines ainsi que les gènes codant pour des transcrits fonctionnels, par exemple les gènes des ARN ribosomiques (ARNr)¹. L'annotation des gènes est très laborieuse et il y a différentes méthodes d'annotation c'est pourquoi

¹Pour plus d'informations sur l'annotation des gènes d'Ensembl, l'explication du pipeline « genebuild » est disponible sur la page internet : www.ensembl.org/info/genome/genebuild/genome_annotation.html. La revue de Yandell [Yandell et Ence, 2012] permet également de comprendre certains des concepts généraux associés à l'annotation.

des plateformes collaboratives, telle que la base de donnée Consensus Coding Sequences (CCDS), œuvrent à définir des localisations qui tendent à être les plus indépendantes des choix méthodologiques employés.

Autres annotations

L'annotation peut se poursuivre par l'identification des séquences régulatrices ainsi que d'autres éléments, tels que les séquences insulatrices, les CTCFs, les zones de chromatines ouvertes ou fermées, les modifications d'histones, *etc.*

I.1.8 Modélisation des génomes adaptée à l'étude des réarrangements chromosomiques

Dans cette partie nous définissons l'échelle d'étude que nous avons choisie pour étudier les réarrangements chromosomiques à partir de la comparaison de génomes modernes séquencés et annotés.

Marqueurs

Lorsque les bioinformaticiens comparent deux génomes ils identifient souvent dans ceux-ci des segments de chromosomes de mêmes origines ancestrales. Ces segments sont habituellement appelés des *marqueurs* ou segments homologues et ils correspondent à des segments hérités d'un segment de chromosome ancestral commun. L'ensemble des marqueurs qui descendent du même segment ancestral forme une *famille* de marqueurs. Notons toutefois que même si elles sont marginales, les comparaisons de génomes « family-free » [Braga *et al.*, 2013] sautent l'étape préliminaire de l'identification de l'origine commune des marqueurs. Cette identification est alors faite plus tard, simultanément à l'inférence des réarrangements par une procédure d'optimisation sur le nombre de réarrangements. Cette dernière méthode permet de récolter plus d'informations lors de la comparaison des génomes, ce qui réduit théoriquement les erreurs d'assignation d'un marqueur à une famille. Néanmoins les méthodes « family-free » sont complexes à mettre en oeuvre et elles sont trop récentes pour que leur apport puisse être estimé avec du recul. Nous faisons donc le choix de débiter, comme la plupart des études de génomique comparative, par l'étape préliminaire de définition des familles de marqueurs.

Critères qui qualifient un bon marqueur À priori n'importe quelle segment de chromosome peut être un marqueur si sa séquence a été conservée depuis un ancêtre commun jusqu'aux espèces modernes. Peu importe qu'il s'agisse d'un segment de chromosome qui corresponde à un gène, à un

régulateur, à une séquence répétée ou même à toute autre séquence, il suffit que sa séquence soit conservée depuis le chromosome ancestral jusqu'aux chromosomes modernes.

Néanmoins en pratique un marqueur de qualité correspond à un segment dont la séquence nucléotidique a évolué suffisamment peu pour que son identification par similarité de séquence soit assurée. Sans ce critère, nous risquerions d'affirmer à tort que le marqueur est absent dès lors que la séquence du segment a trop évolué et qu'elle n'est plus identifiable. De plus les marqueurs choisis pour étudier les réarrangements doivent être suffisamment nombreux pour assurer une bonne couverture des chromosomes. En outre l'on préférera des marqueurs avec de faibles chevauchements de séquences ce qui permettra d'assurer que chaque marqueur est distinct et que l'ordre des marqueurs est identifiable sans ambiguïté. Enfin, dans l'idéal, la distribution spatiale des marqueurs doit être la plus uniforme possible car sinon l'identification des petits réarrangements sera sous-estimée dans les zones pauvres en marqueurs par rapport aux zones denses en marqueurs, où les nombreux marqueurs permettront cette fois-ci de les identifier.

La conservation d'une séquence est généralement liée à l'existence d'une fonction biologique qui contraint l'évolution de la séquence [Boffelli *et al.*, 2004]. Bien qu'il y ait quelques exemples connus de séquences conservées qui ne semblent pas fonctionnelles [Ahituv *et al.*, 2007][Chen *et al.*, 2007], la fonctionnalité des séquences conservées semble être la norme dès lors que les séquences sont conservées dans des espèces éloignées. Pour que nos marqueurs soient identifiables par similarité de séquences il est donc utile de s'intéresser aux séquences fonctionnelles.

Le projet ENCODE [Gerstein *et al.*, 2007] a tenté récemment de dresser l'inventaire exhaustif des séquences fonctionnelles dans le génome humain mais cet inventaire a été fortement critiqué [Graur *et al.*, 2013]. Un résumé de ce débat ainsi que des propositions méthodologique pour tenter d'en sortir ont été formulés par M. Kellis en 2014 [Kellis *et al.*, 2014]. Un désaccord persiste sur l'estimation de la fraction du génome humain qui est fonctionnelle et selon les opinions celle-ci oscille encore entre 5% et 70% du génome. Un désaccord sur la définition de la fonctionnalité explique en grande partie cette différence [Graur *et al.*, 2013], ainsi l'activité biochimique de 70% du génome n'est pas un gage de fonctionnalité. Il y a aussi la difficulté que les scientifiques ont à distinguer un transcrit fonctionnel d'un transcrit non fonctionnel alors qu'il y a de très nombreux transcrits pervasifs. Une autre raison qui explique que le débat n'a pas été tranché concerne les séquences régulatrices. Ces dernières séquences, bien que non exprimées, sont toutefois fonctionnelles étant donné qu'elles modulent l'expression des gènes, et malheureusement elles ne peuvent pas être identifiées facilement. De plus, les séquences répétées semblent en

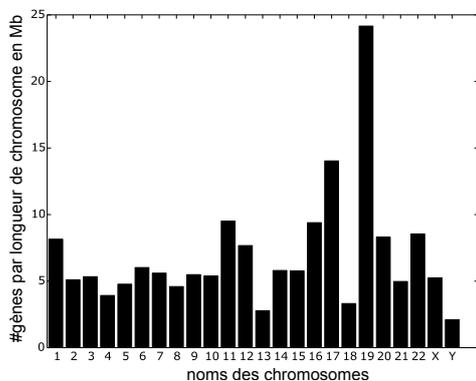
réalité avoir des rôles qui mettent en doute la justification de leur masquage habituel, ainsi certains transposons sont une source de matériel génétique et ils fournissent le substrat de nouveaux promoteurs et de nouveaux sites de régulation. Ces raisons sont loin d'être exhaustives mais nous arrêtons là l'analyse de ce débat qui dépasse largement le cadre de cette thèse.

Choix des gènes codant pour les protéines Même si le débat demeure, l'ensemble des scientifiques s'accorde sur le statut fonctionnel des gènes codants pour les protéines, et nous les choisissons donc comme marqueurs. Ce choix a de multiples avantages, car ces gènes sont généralement les entités les mieux annotées et les plus étudiées¹, nous pouvons ainsi nous appuyer sur les nombreuses études phylogénétiques de ces séquences. Pour limiter les chevauchements, nous ne considérons pas l'entièreté de la séquence d'un gène codant, mais uniquement la séquence correspondant à son transcrit le plus court. Nous ne choisissons pas les séquences correspondants à l'exon le plus court car ces séquences sont parfois extrêmement courtes.

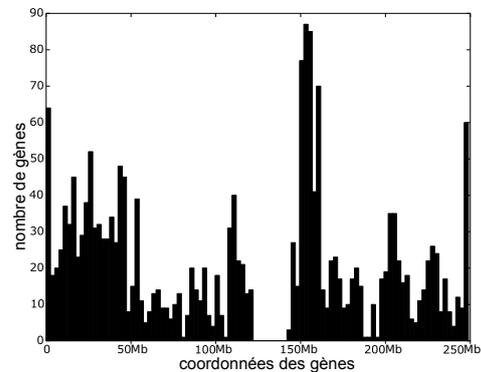
Nous reconnaissons que de nombreuses autres séquences pourraient être utiles, mais les reconstructions d'arbres phylogénétiques de séquences sont presque exhaustivement appliquées aux séquences géniques codant pour les protéines. L'absence d'arbres phylogénétiques pour les autres séquences est au final un argument majeur dans notre choix.

Discussion sur notre choix de marqueurs Les densités de nos marqueurs (les nombres de marqueurs par paires de bases) des chromosomes sont en réalité assez inégales dans le génome humain, voir la figure I.3a. Par exemple, le chromosome 19 a une densité de 24 marqueurs/Mb alors que le chromosome 13 a une densité de 2.5 marqueurs/Mb. De plus la distribution spatiale des marqueurs le long des chromosomes est elle aussi irrégulière, comme cela peut se constater sur l'histogramme du nombre de marqueurs le long du chromosome 1 visible sur la figure I.3b. Ainsi, au voisinage de 160 Mb, la densité locale en marqueurs approche de 34 marqueurs/Mb alors qu'à une dizaine de Mb en amont, il n'y a aucun gène dans la région autour du centromère. Notre distribution de marqueurs est au final non-idéale et des zones riches en marqueurs alternent avec des régions qui en contiennent peu et des régions qui n'en ont aucun. Les régions désertiques en gènes sont généralement les régions centromériques et les bras courts des chromosomes acrocentriques, chez l'humain il s'agit des bras 13p, 14p, 15p, 21p et 22p.

¹Les séquences géniques codants pour les ARNr ont elles aussi été largement exploitées comme marqueurs pour les classifications phylogénétiques.



(a) Densité en gènes codants par paires de bases de chaque chromosome de l'humain.



(b) Distribution des gènes codants le long du chromosome 1 de l'humain. La zone autour de 125 Mb correspond au voisinage du centromère du chromosome 1 et elle ne contient pas de gène codant.

Figure I.3

Quantifions maintenant, pour chaque chromosome, la longueur à partir de laquelle un segment réarrangé est à coup sûr identifiable avec nos marqueurs. Tous les réarrangements chromosomiques réarrangent des segments de chromosomes, une inversion inverse un segment, une translocation réciproque réarrange deux segments, enfin les fusions et les fissions de chromosomes réarrangent réciproquement deux chromosomes et deux moitiés de chromosomes. Si un segment de chromosome est réarrangé, il est identifiable à partir du moment où le segment contient au moins un marqueur et inversement le réarrangement n'est pas identifiable si le segment réarrangé ne contient pas de marqueur.

Pour un chromosome donné, la longueur maximale d'un segment réarrangé non identifiable correspond donc, télomères mis à part, à l'intervalle le plus long qui sépare deux marqueurs qui se font suite le long du chromosome. Si l'extension spatiale des marqueurs est négligée, et si les réarrangements dans les séquences géniques le sont aussi, il est assez simple de montrer que cette dernière longueur est la même que la longueur à partir de laquelle un segment réarrangé est visible.

Dans le cas du chromosome 1 cette longueur est celle du désert de gène qui ceinture le centromère, c'est à dire une longueur de 23 Mb, visible sur la figure I.3b entre 120 et 143 Mb. Une inversion péricentrique (dont le segment inversé inclut le centromère) sera invisible pour peu que les extrémités du segment soient dans la zone déserte. Les inversions invisibles peuvent donc

atteindre une longueur de 23 Mb sur le chromosome 1. De plus, sous les hypothèses précédentes, toute inversion plus longue que 23 Mb inclura au moins un marqueur et nous serons donc en mesure de la voir. Les longueurs à partir desquelles un segment réarrangé est à coup sûr identifiable sont données pour chaque chromosome humain par le graphique de la figure I.4.

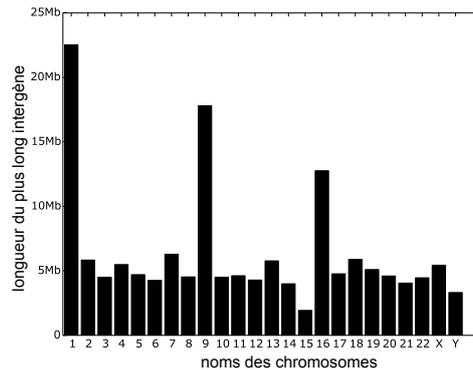
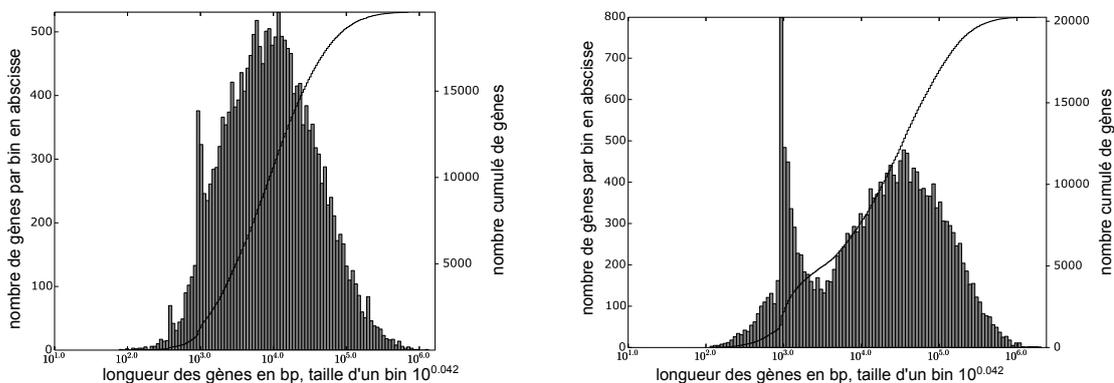


Figure I.4 – Longueurs des plus grands intervalles entre gènes codant pour chaque chromosome de l’humain. La longueur du plus grand intervalle correspond, à la longueur d’un gène près, à la longueur à partir de laquelle un segment réarrangé est identifiable. Les bras courts de chromosomes acrocentriques et les télomères sont en général totalement exempts de gènes codants, donc pour être exacte il faudrait également les inclure dans l’histogramme.

La distribution des tailles de nos marqueurs est visible, pour l’humain, sur la figure I.5a et, pour l’opossum, sur la figure I.5b. La longueur du marqueur le plus petit correspond à la taille du réarrangement le plus petit que nous puissions mettre en évidence. En effet, si une inversion, dont la taille du segment inversé fait exactement celle du marqueur, venait à inverser ce marqueur, étant donné que nos marqueurs sont orientés, nous serions en mesure de la voir grâce à ce marqueur et à ses voisins. La distribution des longueurs des marqueurs dans le génome humain est log-normale et la distribution des longueurs des gènes dans le génome de l’opossum est bimodale, elle a un mode principal à environ 300 kb ($10^{3.1}$ bp) et un mode secondaire plus étroit à environ 1 kb. Une brève analyse a montré que les gènes du mode secondaire sont enrichis en gènes de l’olfaction. La longueur des plus petits marqueurs de chaque chromosome est en moyenne de 200 bp. Cette longueur correspond grossièrement à la taille du plus petit segment réarrangé identifiable avec nos marqueurs.

Enfin, la longueur moyenne de nos marqueurs est de 24 kb dans le génome humain et en moyenne un de ces marqueurs chevauche un autre marqueur



(a) Distribution des longueurs des séquences correspondant aux transcrits les plus courts des gènes codants pour les protéines dans le génome humain, en échelle logarithmique.

(b) Distribution des longueurs des séquences correspondant aux transcrits les plus courts des gènes codants pour les protéines dans le génome de l'opossum en échelle logarithmique.

Figure I.5

sur au moins 19 bp, soit 0.07% de la longueur moyenne.

Marqueurs et duplications de séquences Nous terminons cette partie dédiée au choix de nos marqueurs en mentionnant une difficulté liée aux événements de duplications de séquences. Comme nous le disions, les marqueurs sont censés marquer la conservation locale de la région chromosomique ancestrale. Or si au cours de l'évolution la séquence du marqueur est dupliquée, comment déterminer laquelle des deux séquences post-duplication correspond le plus à la séquence ancestrale ? Pour répondre à cette question appelons séquence *dupliquée* la séquence pré-duplication et séquence *copie* la séquence nouvellement insérée. Juste après la duplication, la séquence dupliquée n'a pas bougé, elle est toujours à la même place, alors que la séquence copie est à une nouvelle localisation, loin ou juste à côté. Parmi les deux séquences post-duplication, la séquence dupliquée est donc la plus indiquée pour continuer à être le marqueur de *référence* de la séquence ancestrale. Néanmoins rien ne nous assure que la séquence dupliquée conserve mieux la séquence nucléotidique ancestrale que la séquence copie.

Prenons un exemple. Un chromosome ancestral est composé de trois séquences : A B C, dans cet ordre. Durant l'évolution la séquence B est dupliquée. Après duplication, continuons à appeler B la séquence dupliquée et appelons B.a la séquence copie. La séquence dupliquée est à la même position

que la séquence ancestrale et la séquence copie est insérée à la droite de C. Le génome post-duplication est donc A B C B.a. La séquence de *référence* est B, c'est elle qui reflète le plus la localisation de la séquence d'origine dans le chromosome ancestral car ses voisins sont toujours A et C. Juste après duplication les séquences B et B.a ont toutes les deux la même séquence mais l'évolution continue et il se peut que la séquence de B varie plus que celle de B.a. Ce sera le cas par exemple si B.a devient porteur de la fonction (si la séquence code pour une protéine par exemple), la pression de sélection conservatrice sera alors relâchée sur B et cette dernière séquence pourra évoluer par pseudogénéisation, sub-fonctionnalisation ou néo-fonctionnalisation. Dans notre exemple considérons que l'évolution se déroule selon l'un de ces trois scénarios. En fin d'évolution la séquence moderne B.a sera donc plus similaire à la séquence d'origine que ne le sera la séquence B moderne. En parallèle de cette évolution, nous considérons également qu'un deuxième génome, d'une autre espèce, a survécu en conservant intact le chromosome ancestral A B C ainsi que ses séquences.

Dans la réalité : nous avons accès aux génomes modernes, nous ne connaissons pas l'histoire évolutive des génomes et c'est par similarité de séquences que nous identifions les marqueurs. Suite à l'évolution décrite plus haut nous avons donc un premier génome moderne A B C B.a et un deuxième génome moderne A B C. Ajoutons maintenant un indice pour distinguer l'instance de la séquence dans le 1^{er} génome de l'instance de la séquence dans le deuxième génome. Ainsi le premier génome moderne s'écrit $A_1 B_1 C_1 B.a_1$ et le deuxième génome s'écrit $A_2 B_2 C_2$. La comparaison des séquences nous informera que les best-hits réciproques sont (A_1, A_2) , (C_1, C_2) et $(B.a_1, B_2)$ car, comme nous l'avons décrit plus haut, l'évolution a été telle que la séquence de B.a₁ est plus similaire à celle de B₂ que la séquence de B₁ ne l'est de la séquence de B₂. Nous en concluons à raison, que dans le deuxième génome, B₂ est le marqueur de la séquence ancestrale. Mais nous risquerions d'en conclure également que, dans le premier génome, le marqueur de la séquence ancestrale B est B.a₁, alors que nous aurions préféré que ce soit la séquence B₁ ! Car, d'après l'histoire évolutive, c'est cette dernière séquence qui est le véritable marqueur de référence.

La difficulté que nous avons évoquée nous incite à considérer des familles de marqueurs plus larges que celles construites par des relations de best-hits réciproques 1:1. Avec de telles familles, la famille issue de la séquence ancestrale B ne sera plus $(B.a_1, B_2)$, mais $(B.a_1, B_1, B_2)$ et à défaut de contenir exclusivement l'entièreté des marqueurs de référence, elle aura l'avantage de les inclure tous.

À partir des similarités de séquences et des familles qui en résultent il est impossible de retrouver sans ambiguïté le marqueur de référence B₁, nous

aurions besoin pour cela d'une information sur le voisinage des séquences. Nous reviendrons sur ce point plus tard (section I.5.4). En attendant, pour éviter d'exclure le marqueur de référence, nous utiliserons désormais des familles de marqueurs qui ne sont pas que des best-hits réciproques. Ce qui a l'avantage d'éviter d'exclure les marqueurs de référence qui ont moins conservé la séquence ancestrale que leurs copies. L'inconvénient de ce choix est que dans un même génome moderne plusieurs marqueurs correspondent maintenant à une séquence ancestrale.

I.1.9 Conclusion concernant notre modélisation des génomes

Les gènes que nous avons choisi comme marqueurs sont par la suite les constituants fondamentaux de nos génomes. Grâce à eux nous n'avons pas besoin de nous encombrer des milliards de paires de bases qui composent les génomes modernes. Nous sommes ainsi en mesure d'étudier les réarrangements chromosomiques à une échelle génomique adaptée.

Au final notre modélisation d'un génome est la suivante. Un génome est un ensemble de chromosomes. Chaque chromosome a un nom et nous considérons qu'il est une liste ordonnée de gènes orientés. Le long d'un chromosome les gènes sont ordonnés par leurs extrémités 5', c'est à dire par leurs débuts de transcriptions. L'orientation d'un gène correspond à son orientation de transcription, c'est à dire l'orientation 5'-3' du brin d'ADN sur lequel il est encodé. La figure I.6 résume la manière dont nous modélisons un chromosome.

Dans les prochaines sections de ce chapitre nous définissons les concepts qui nous seront utiles pour décrire et représenter la dynamique des génomes à cette échelle d'étude.

I.2 Entités

I.2.1 Gène

Nous considérons qu'un gène est une entité insécable et qu'il a une *orientation* définie par la direction de sa transcription par l'ADN polymérase, de l'extrémité 5' (start) à l'extrémité 3' (end). L'*extrémité d'un gène* nommé A sera notée (A, s) s'il s'agit de l'extrémité 5' et (A, e) s'il s'agit de l'extrémité 3'.

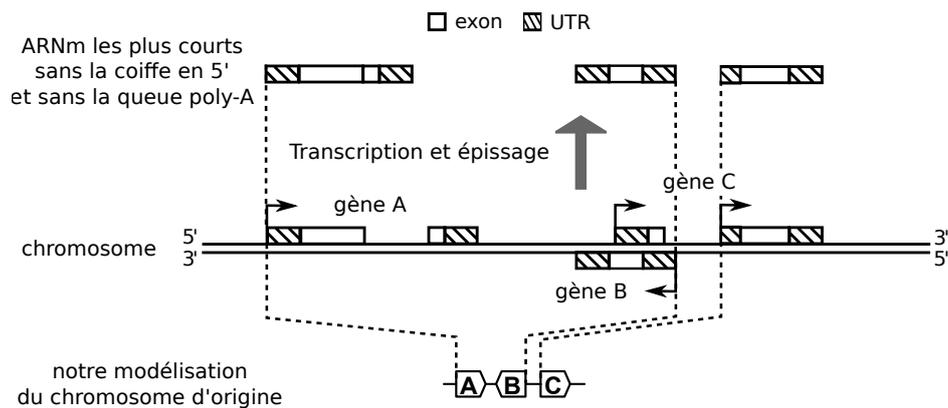


Figure I.6 – Modélisation des chromosomes avec nos gènes. À mi-hauteur, trois gènes codant pour des protéines sont représentés sur un chromosome commun. Les UTRs sont représentés par des rectangles quadrillés et les exons par des rectangles blancs. Les gènes A et C sont encodés sur le brin d'ADN du dessus, dont l'orientation 5'-3' va de la gauche vers la droite. Le gène B est encodé sur le brin d'ADN complémentaire. Le gène C contient deux sites de début de transcription représentés par des flèche coudées. Au dessus de chaque gène sont représentés leurs ARNm les plus courts après épissage. La coiffe 5' et la queue poly-A ne sont pas représentées. Enfin, tout en bas, il y a notre modélisation d'un chromosome en une liste ordonnée de trois gènes orientés. L'ordre des gènes correspond à l'ordre des extrémités 5' des transcripts les plus courts et leurs localisations sont matérialisées par les lignes en pointillées.

I.2.2 Chromosome

Un chromosome sera, à notre échelle, une liste ordonnée de gènes. Tous nos chromosomes sont linéaires. Nous ne considérerons pas les chromosomes circulaires spécifiques aux procaryotes et nous ne considérons pas non plus les plasmides ou les chromosomes circulaires des organites de cellules eucaryotes comme la mitochondrie ou le chloroplaste.

Quand nous représenterons un chromosome nous définirons, parfois explicitement parfois implicitement, une *origine* de la liste des gènes du chromosome. Cette origine est choisie arbitrairement. Si un chromosome est composé des gènes ABCDE, les uns à la suite des autres, la liste peut être ordonnée de A à E ou de E à A. Le choix de l'origine définit une *orientation de référence* (ou un ordre de référence) de l'origine jusqu'à l'extrémité opposée. Si le gène A est choisi comme origine, l'orientation de référence va de A à E : le 1^{er} gène est A, le 2^{ème} est B, ... et le dernier gène est E. Nous noterons parfois un chromosome, c , composé de N gènes, de la manière suivante

$$c = [g_1, \dots, g_N] = [g_k]_{k \in [1, N]}$$

avec g_k le $k^{\text{ème}}$ gène du chromosome c . Dans ce cas, l'origine du chromosome c est le gène g_1 et l'orientation de référence va de g_1 à g_N . L'orientation du gène g_k dans ce chromosome est notée $o(g_k)$; elle est égale à +1 si la transcription du gène g_k est réalisée dans la même direction que l'orientation de référence du chromosome c , dans le cas contraire $o(g_k) = -1$.

Comme nous l'écrivions précédemment, les gènes sont indexés par leurs extrémités 5' et par la suite nous considérerons qu'à notre échelle d'étude les gènes ne se chevauchent pas. Une explication graphique est donnée par la figure I.7.

Dans un chromosome, un *intergène* est une région chromosomique située entre deux gènes adjacents, voir figure I.8. Les deux régions aux extrémités d'un chromosome sont appelées *télomères*. Les deux extrémités de gènes qui flanquent un intergène sont les *flancs* de cet intergène et un télomère a un *flanc* unique. Par exemple, dans la figure I.8

- les extrémités (B, e) et (C, s) sont les deux flancs du 2^{ème} intergène
- et l'extrémité (L, e) est le flanc unique du télomère de droite.

Dans les espèces métazoaires que nous étudierons les chromosomes sont soit autosomiques, soit sexuels. Les chromosomes autosomiques sont souvent nommés par un chiffre (1, 2, 3, ...) alors que les chromosomes sexuels sont habituellement nommés par des lettres (X et Y pour l'humain, W et Z pour le poulet).

	un ordre de gènes	ordre des gènes inverse
chromosome avec gènes en échelle physique		
gènes ordonnés par les extrémités les plus à gauche		
gènes ordonnés par leurs extrémités 5'		

Figure I.7 – Indexage des gènes. La première colonne du tableau représente un chromosome avec origine le gène A. Trois représentations de ce chromosome sont visibles en dessous. De haut en bas, la première image montre les localisations des gènes et leurs extensions spatiales. Dans cet exemple le gène B chevauche entièrement les gènes C et D. La deuxième image représente l'ordre des gènes lorsqu'ils sont ordonnés par leur 1^{ère} extrémité (la plus à gauche, la plus proche de l'origine) et la troisième image donne l'ordre des gènes lorsqu'ils sont ordonnés par leur extrémité 5'. La deuxième colonne représente les mêmes informations lorsque l'ordre de référence du chromosome est inverse, de E vers A. L'origine est alors le gène E. Cette seconde colonne illustre bien que, parmi les deux indexages, l'indexage des gènes par leurs extrémités 5' est le seul à générer un ordre consistant d'une orientation du chromosome à l'autre.

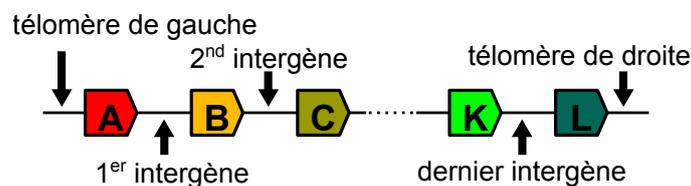


Figure I.8 – Intergènes et télomères d'un chromosome. L'extrémité de gauche du chromosome est par convention l'origine du chromosome et l'orientation de gauche à droite correspond à l'orientation de référence.

I.2.3 Génome

Un génome est un ensemble de chromosomes. Dans notre étude il s'agit du génome d'une cellule représentative d'un individu, lui-même représentatif d'une espèce.

I.2.4 Espèce

Il est courant d'approximer une espèce à un ensemble d'individus interféconds dont la descendance est féconde. Entre deux espèces différentes, des incompatibilités morphologiques et/ou génétiques empêchent systématiquement la fécondation, le développement de l'embryon ou alors rendent le descendant incapable de se reproduire. C'est le cas du mulet, hybride issu de l'accouplement d'un âne (*Equus asinus*) et d'une jument (*Equus caballus*). Comme nous l'avons expliqué précédemment nous considérerons que le génome d'un individu est représentatif des autres génomes de l'espèce car le polymorphisme intra-spécifique des espèces que nous étudierons semble suffisamment faible [Leffler *et al.*, 2012] pour le permettre. Une espèce sera donc caractérisée par un génome de référence.

I.3 Évènements

Au sein d'une espèce avec un dimorphisme sexuel mâle/femelle, les individus se reproduisent sexuellement par l'accouplement du mâle et de la femelle. Le génome des géniteurs est transmis à travers des divisions cellulaires mitotiques et méiotiques. Cette transmission s'achève par la fusion des gamètes mâles et femelles pour donner naissance à la cellule œuf d'un descendant qui pourra à son tour transmettre le génome dont il a hérité. Avant chacune de ces divisions cellulaires, les chromosomes sont répliqués. Des *mutations* ont parfois lieu durant la réplication et le génome après réplication diffère alors du génome avant réplication¹. Ces variations par rapport à l'exemplaire original se transmettent par la suite de génération en génération.

Selon que la mutation est avantageuse ou délétère, le phénotype de l'individu qui la porte dans son génome est sous une pression de sélection respectivement positive ou négative. Dans le premier cas la fréquence de la mutation aura tendance à augmenter jusqu'à parfois finir par se fixer défini-

¹Les mutations peuvent également être causées par des rayons X, des rayons UV, des substances mutagènes, la radioactivité, ... néanmoins il est communément admis que la majorité des mutations naturelles ont lieu durant la réplication. [Mani et Chinnaiyan, 2010]

tivement dans tous les génomes de l'espèce. Dans le deuxième cas la mutation disparaîtra plus ou moins rapidement de la population. D'autres mutations sont neutres et ne semblent pas sélectionnées. La dérive génétique peut néanmoins amener ces dernières mutations à de hautes fréquences jusqu'à les fixer elles aussi dans la population, surtout lorsque la population est composée de peu d'individus.

I.3.1 Évènements géniques

Dans cette sous-section nous énumérons l'ensemble des mutations qui modifient le contenu en gènes d'un génome.

Duplication

Il arrive souvent qu'une mutation duplique un gène et que celui-ci se retrouve en plusieurs copies dans le génome. En 1970 Susumu Ohno formula même l'hypothèse que les duplications jouent un rôle majeur dans l'évolution [Ohno, 1970]. Si la nouvelle copie d'un gène est insérée près du gène d'origine la duplication est qualifiée de *tandem* sinon la duplication est qualifiée de duplication *disperse*. Suite à un évènement de duplication nous distinguons le gène copié de la nouvelle copie du gène. D'autres auteurs utilisent eux aussi le concept de duplications dirigées « parent-enfant » ou « source-target » [Dewey, 2011][Han et Hahn, 2009]. Nous reviendrons sur ce point plus tard (section I.5.4).

Délétion

D'autres mutations suppriment des gènes. La délétion d'un gène se fait par exemple par un processus de délétion segmentale, ou de pseudogénéisation si, suite à une mutation dans sa séquence codante, le gène perd sa capacité à être transcrit ou traduit.

Naissances *de novo*

Les duplications ne sont pas les seuls évènements à générer de nouveaux gènes et il semble que certains nouveaux gènes proviennent de la transformation d'une séquence non-génique en une séquence génique [Carvunis *et al.*, 2012][Schlötterer, 2015].

Transfert horizontal

Un dernier évènement¹ par lequel un génome peut acquérir un nouveau gène est un transfert horizontal de gène. Bien que cet évènement soit répandu chez les procaryotes, cet évènement est considéré comme rare chez les eucaryotes [Schlötterer, 2015]. Un gène de rétro-virus, *synctin-A*, semble pourtant avoir été recruté par les souris et il est nécessaire à la formation du placenta [Dupressoir *et al.*, 2009]. Lorsqu'ils ont lieu, les transferts de gènes sont souvent considérés, à défaut de pouvoir les identifier, comme des naissances *de novo* de gènes. Ce sera le cas dans notre étude.

Évènement génique segmental

Nous finissons notre énumération des évènements géniques en mentionnant les évènements géniques segmentaux. Les duplications, les délétions et les transferts horizontaux peuvent s'effectuer par segment de plusieurs gènes. Comme dans le cas de la duplication en tandem d'un gène, une duplication segmentale en tandem insérera la copie d'un segment de chromosome, contenant plusieurs gènes, dans le voisinage proche du segment dupliqué. Enfin des segments de chromosomes peuvent être transférés horizontalement.

I.3.2 Évènements chromosomiques

Nous énumérons dans cette section l'ensemble des réarrangements équilibrés qui ne modifient pas le contenu en gènes mais qui modifient soit l'ordre des gènes, soit les orientations des gènes, soit le nombre de chromosomes. Nous avons donné des exemples de réarrangements dans notre Introduction donc nous ne décrivons que leurs caractéristiques les plus remarquables.

Inversion

Une inversion est caractérisée par le chromosome hôte du segment inversé, la localisation du point de cassure le plus près de l'origine du chromosome (figure I.8) et la longueur du segment inversé. Une inversion génère deux points de cassures, une à chaque extrémité du segment inversé.

Translocation réciproque

Une translocation réciproque est une mutation par laquelle deux chromosomes s'échangent des segments à leurs extrémités. Une translocation réciproque est caractérisée par les deux chromosomes qui échangent leurs extrémités, les

¹qui n'est pas une mutation à vrai dire

deux extrémités de chromosomes qui s'échangent un segment et les longueurs des segments échangés. Une translocation cause deux points de cassures.

Fusion

Une fusion de deux chromosomes a lieu quand deux chromosomes sont unis au niveau de leurs télomères. Une fusion est caractérisée par les deux chromosomes fusionnés l'un à l'autre ainsi que par leurs deux extrémités fusionnées.

Fission

À l'inverse, il s'agit d'une fission, si un chromosome est brisé en deux. Le chromosome dans lequel la cassure a lieu et la position du point de cassure par rapport à l'origine caractérisent cet évènement.

Transposition de segment de chromosome

Une transposition déplace un segment de chromosome à un autre endroit par un processus similaire à un coupé-collé [Coghlan, 2002][Zhao et Bourque, 2009]. Cet évènement est caractérisé par le chromosome hôte du segment transposé, la localisation du segment transposé par rapport à l'origine, le chromosome qui reçoit le segment transposé, la localisation du site d'insertion par rapport à l'origine du chromosome de destination et l'orientation du segment lorsqu'il est inséré. Une transposition produit trois points de cassures, deux aux extrémités du segment lors de son départ du chromosome d'origine et une lors de son insertion sur le chromosome d'arrivé.

I.3.3 Évènement génomique

Duplication complète de génome

Le seul évènement qui implique un génome entier est, à notre connaissance, la duplication complète de génome, Whole-Genome Duplication (WGD). Suite à cet évènement extraordinaire le génome est intégralement dupliqué et chaque chromosome pré-duplication se retrouve en deux copies. Le nombre de chromosomes d'un individu diploïde passe de $2n$ à $4n$ chromosomes et pour cette raison cet évènement est parfois appelé tétraploïdisation. Cet évènement rare est arrivé deux fois à la racine des vertébrés [Dehal et Boore, 2005] et de nombreuses fois durant l'évolution des poissons [Jaillon *et al.*, 2004].

I.3.4 Évènement relatif à une espèce

Après avoir décrit les différentes mutations qui altèrent les génomes nous énumérons maintenant le principal évènement associé à l'évolution d'une espèce.

Spéciation

Cette sous-section est en grande partie une traduction de l'article wikipedia : https://en.wikipedia.org/wiki/Patrick_Matthew

La naissance d'une nouvelle espèce à partir d'une ancienne est un évènement nommé *spéciation*. En 1831 Patrick Matthew publia un livre *On Naval Timber and Arboriculture* [Matthew, 1831], dont le but était de déterminer le meilleur moyen de faire pousser des arbres pour la construction des navires de guerre de la Royal Navy. Dans son ouvrage il explique les nombreux effets, néfastes sur le long terme, de l'abattage systématique des arbres fournissant un bois de qualité supérieure. Dans l'appendice de son ouvrage, en se basant sur sa connaissance de la sélection artificielle négative, il énonce des utilisations possibles de la sélection artificielle positive ayant pour but d'accroître la qualité des bois. Il évoque également la possibilité de créer de nouvelles espèces d'arbres grâce à cette sélection. En extrapolant son raisonnement il rédigea une description basique du processus de « sélection naturelle » et il expliqua comment ce processus pouvait engendrer de nouvelles espèces incapables d'inter-fécondation. En 1859 Darwin, et Wallace peu de temps après, énoncèrent à leur tour le principe de la sélection naturelle. De nombreux exemples biologiques à l'appui, ils montrèrent comment ce processus peut expliquer l'origine des espèces. En 1860 Patrick Matthew eut connaissance de *The Origin Of Species* de Darwin et il exprima son désir de faire prévaloir son droit d'antécédent pour avoir été le premier à penser la sélection naturelle [Matthew, 1860]. Darwin accepta sa requête et il écrivit « *I freely acknowledge that Mr. Matthew has anticipated by many years the explanation which I have offered of the origin of species, under the name of natural selection.* » [Darwin, 1860]. Depuis, notre compréhension de l'origine des espèces n'a cessé de s'accroître et de nombreux mécanismes évolutifs expliquent ce phénomène. La spéciation allopatrique est le plus simple d'entre eux : un sous-ensemble d'individus se sépare du reste de l'espèce et subit une isolation reproductrice due par exemple à un obstacle physique. À cause de l'isolation, les individus de ce sous-ensemble accumulent de nombreuses mutations, au point qu'ils ne sont plus, au bout du compte, féconds avec les individus de l'espèce d'origine. Par exemple des spéciations allopatriques furent certainement à l'origine des différentes espèces de pinsons découvertes par Darwin sur les îles Galapagos.

I.4 Inférences et représentation des évènements relatifs à une espèce

I.4.1 Arbre des espèces

Depuis Darwin, de nombreuses phylogénies d'espèces ont été inférées en se basant sur des strates géologiques, des fossiles, des homologues morphologiques entre espèces modernes et, plus récemment, en se basant sur les similarités de séquences génomiques. Ces phylogénies sont classiquement illustrées sous la forme d'un *arbre des espèces* dont la racine représente une espèce ancestrale, dont les nœuds internes représentent les évènements de spéciation et dont les feuilles sont les espèces modernes. L'extinction d'une espèce est parfois représentée par une feuille au bout d'une branche arrêtée de manière précoce. Quand les dates de spéciations ne sont pas bien résolues, plusieurs spéciations sont parfois attribuées à un même nœud mais généralement les arbres d'espèces sont des arbres à bifurcations dans lesquels chaque nœud représente une spéciation suite à laquelle deux branches succèdent à la branche d'origine. La longueur d'une branche de l'arbre des espèces est généralement proportionnelle au temps évolutif entre les deux évènements aux extrémités de cette branche. L'ancêtre commun le plus récent de deux espèces désigne l'espèce ancestrale qui précède, de justesse, l'évènement de spéciation commun aux deux espèces et le plus récent. Cette espèce ancestrale est également appelée Most Recent Common Ancestor (MRCA)¹ des deux espèces considérées.

I.5 Inférence et représentation des évènements relatifs à un gène

I.5.1 Arbre de gène

Comme pour une espèce, l'histoire évolutive d'un gène peut être représentée par un arbre phylogénétique. Dans sa forme la plus basique la racine de l'arbre d'un gène représente le gène ancestral² et les nœuds successifs représentent les évènements de duplications ou de spéciations qui donnent naissance aux gènes descendants. Si l'arbre des espèces, dans lequel ces gènes ont évolué, est connu, il est possible de concilier l'arbre de gène avec l'arbre des espèces

¹Elle est parfois nommée « dernier ancêtre commun » et, dans ce cas, elle est désignée par l'acronyme Last Common Ancestor (LCA).

²Quand cela est possible la racine représente plus précisément l'évènement de naissance *de novo* à l'origine de ce gène.

et d'annoter les noeuds plus précisément pour distinguer les évènements de duplication des évènements de spéciation. Un type de noeud représente alors les duplications (généralement un carré) et un autre type de noeud représente les évènements de spéciation (généralement un cercle). La figure I.9 représente un tel arbre. Si les transferts horizontaux sont connus, un autre type de noeud peut être ajouté pour représenter l'origine des transferts.

Un ensemble de plusieurs arbres de gènes est une forêt d'arbres de gènes. Un sujet de recherche majeur en génomique comparative est d'inférer la forêt d'arbres dont les feuilles sont les gènes de génomes modernes séquencés et annotés [Felsenstein, 2004].

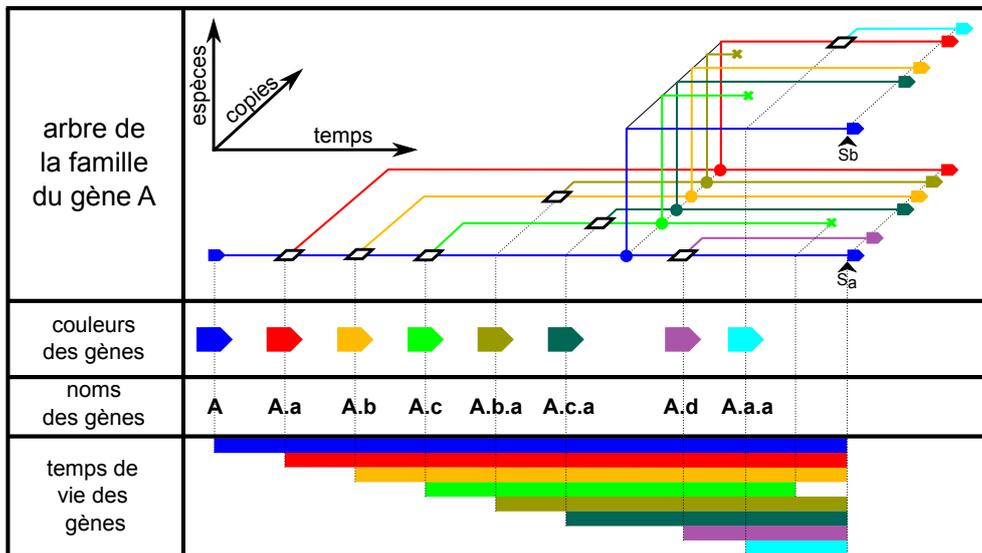


Figure I.9 – Phylogénie de la famille du gène ancestral A. En haut, une représentation en trois dimensions de l'arbre phylogénétique de la famille du gène ancestral A (en bleu tout à gauche). De gauche à droite il s'agit de l'axe du temps, de bas en haut sont représentées les différentes espèces suites aux spéciations et de devant à derrière sont représentées les copies du gène A. Les duplications de gènes sont représentées par des carrés, les spéciations sont représentées par des cercles et les délétions de gènes sont représentées par des croix. Après chaque duplication un nouveau gène est inséré. Les nouveaux gènes (copies) sont nommés en ajoutant un suffixe au nom du gène copié tout en prenant garde que ce suffixe n'ait pas déjà été utilisé pour une copie précédente.

Comme dans le cas des espèces, il est possible de définir le MRCA de deux gènes, il s'agit du gène ancestral le plus récent dont ces deux gènes descendent. Nous dirons qu'un gène *descend indirectement* d'un autre gène s'il y a eu des

événements de duplication durant l'histoire évolutive entre ces deux gènes. À l'inverse, le *descendant directe* d'un gène n'a subi que des événements de spéciation.

I.5.2 Famille d'un gène

Une famille de gène est un ensemble de gènes qui descendent d'un même gène d'origine (l'origine de la famille) par une succession d'évènements de duplications, de spéciations et de transferts horizontaux. La famille d'un gène est donc l'ensemble des gènes de l'arbre du gène d'origine.

Les familles peuvent être définies de différentes manières selon le choix des gènes à l'origine des familles. La figure I.10 montre quelques-unes de ces manières. Dans cette figure un arbre initial est édité, des branches sont coupées et de nouvelles racines sont définies. Ce processus est souvent appelé un *élagage*. Les élagages modifient le nombre d'arbres de gènes ainsi que leurs racines, ce qui génère des familles de gènes différentes.

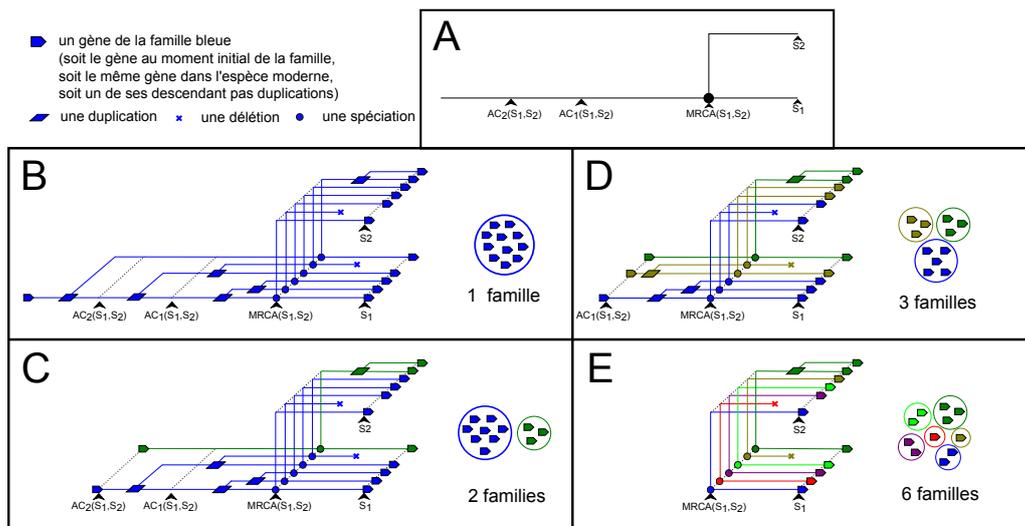


Figure I.10 – Différentes manières de définir les familles de gènes de deux génomes modernes. Le **panel A** représente un arbre des espèces avec deux espèces modernes S_1 et S_2 . $MRCA(S_1, S_2)$ est le dernier ancêtre commun de S_1 et S_2 . $AC_i(S_1, S_2)$ est le $i^{ème}$ ancêtre commun le plus récent après $MRCA(S_1, S_2)$. Les **panels B, C, D** et **E** représentent divers élagages du même arbre de gène (**panel B**) et les familles correspondantes. À chaque fois il y a autant de familles que d'arbres. Par conséquent plus l'ancêtre utilisé pour définir les gènes à la racine des arbres de gènes est récent plus il y a de familles.

Dans notre travail, lorsque les génomes de deux espèces sont comparés, les familles de gènes sont implicitement définies de manière à ce que les arbres correspondants prennent chacun racine dans un gène unique du MRCA des deux espèces. Par exemple, dans la figure I.10, si nous comparons les génomes des espèces S_1 et S_2 les familles seront celles du panel E.

I.5.3 Relations entre les gènes d'un même arbre

Homologie, paralogie et orthologie

Nous reprenons dans ce paragraphe et les suivants la présentation de l'homologie, de la paralogie, de l'orthologie ainsi que de la in-paralogie qui a été développée dans la thèse de C. Lemaitre [Lemaitre, 2008]. Nous nous sommes de plus fortement inspirés de la présentation de l'orthologie positionnelle qui y a été faite.

Deux gènes sont *homologues* s'ils font partie de la même famille. La relation d'homologie sera parfois notée \mathcal{H} . Avec cette notation, $g_1 \mathcal{H} g_2$ signifie que les gènes g_1 et g_2 sont homologues¹.

On distingue deux types d'homologies : l'orthologie et la paralogie. Deux gènes sont *orthologues* s'ils sont homologues et s'ils ont divergé à la suite d'une spéciation. Deux gènes orthologues appartiennent à deux espèces différentes. On dit aussi que deux gènes, l'un dans le génome d'une espèce moderne et l'autre dans le génome d'une seconde espèce, sont orthologues s'ils sont dérivés d'un unique gène ancestral du MRCA des deux espèces. Si deux gènes homologues ont divergé suite à une duplication ils sont *paralogues*. Si nous reprenons l'arbre du gène A de la figure I.9, tous les gènes de l'arbre sont homologues et ils constituent une famille. Dans l'espèce moderne S_1 les gènes bleu (A) et orange (A.b) sont paralogues car leur événement commun le plus récent est une duplication. Par contre le gène bleu dans S_1 (A) et le gène bleu dans S_2 (A) sont orthologues car leur événement commun le plus récent est une spéciation.

Relations 1:1, 1:n et n:m

La relation précédente relie un gène à un autre, c'est une relation de type 1:1. D'autres types de relations sont des relations entre un gène et plusieurs gènes (relations 1:n) ou entre plusieurs gènes et plusieurs autres gènes (relations n:m). Par exemple, toujours dans la figure I.9, le gène rouge dans S_1 (A.a) est orthologue au gène rouge dans S_2 (A.a) et il est également orthologue au gène cyan (A.a.a) dans S_2 . Une relation d'orthologie 1:n relie ici le gène rouge

¹ \mathcal{H} est une relation d'équivalence dont les classes d'équivalence sont les familles de gènes.

(A.a) de S_1 à ses orthologues. Deuxième exemple, dans le génome de l'espèce S_2 , le gène rouge (A.a) est en relation de paralogie avec le gène cyan (A.a.a) de S_2 ainsi qu'avec les gènes bleus (A) de S_1 et de S_2 . Là encore une relation de type 1:n relie le gène A.a de S_2 avec ses paralogues (A.a.a et A dans S_1 et S_2) et aucun autre gène ne lui est paralogue. Troisième exemple, dans S_1 , le gène bleu (A) est en relation de paralogie avec les gènes rouges (A.a) de S_1 et S_2 , mais pas uniquement, car il est aussi le paralogue de nombreux autres gènes (A.d dans S_1 et A.a, A.b, A.c, dans les génomes dans lesquels ils sont présents), il s'agit d'une autre relation 1:n reliant cette fois-ci le gène bleu (A) de S_2 à tous ses paralogues.

Enfin, dernier exemple, dans la famille du gène bleu d'origine (A à la racine), les 6 gènes de S_1 sont en relation d'homologie avec les 5 gènes de S_2 , nous avons ici affaire à une *relation n:m*, d'un ensemble de gènes vers un autre ensemble de gènes.

In-paralogues et orthologues positionnels

Nous avons vu des exemples de relations d'orthologie multiples et de manière générale n gènes d'une espèce peuvent être orthologues à m gènes d'une autre espèce. C'est le cas si les duplications à l'origine des n copies sont ultérieures au MRCA dans la première lignée et s'il en va de même des m duplications dans la deuxième lignée. Dans ce cas les n copies de la première espèce moderne sont appelées des *co-orthologues* ou des *in-paralogues*¹. Pour se rapprocher d'une relation d'orthologie 1:1 (la relation qui relie deux instances d'un même gène) il est courant d'avoir recours à la notion d'orthologie positionnelle. Deux gènes orthologues sont appelés *orthologues positionnels* s'ils ont le même contexte génomique [Swidan *et al.*, 2006]. Le but de cette relation est d'identifier deux gènes modernes qui descendent directement d'un même gène ancestral sans avoir été insérés suite à une copie. La plupart des méthodes de reconstructions phylogénétiques ne distinguent cependant pas le gène copié du gène inséré suite à la duplication et les deux gènes post-duplication ont un rôle symétrique. Pour identifier le gène copié, la relation d'orthologie positionnelle s'intéresse au contexte génomique, au voisinage des gènes. Ainsi, dans les figures I.11 et I.12, deux orthologues, M2 dans le génome de la souris et H4 dans le génome humain, sont orthologues positionnels car leurs contextes génomiques sont équivalents et par conséquent ils semblent descendre directement du même gène ancestral, sans avoir été dupliqués. Néanmoins l'orthologie positionnelle ne permet pas toujours de retrouver la localisation du gène ancestral. Par exemple H3 et H2 sont co-orthologues et ils sont orthologues avec le gène M1.

¹À l'inverse, deux gènes qui sont paralogues mais qui ne sont pas in-paralogues sont parfois appelés *out-paralogues*.

Nous aimerions déterminer quel gène entre H3 et H2 descend directement du même gène ancestral que M1. Néanmoins le contexte génomique n'est ici pas suffisant car H3 et H2 ont le même et par conséquent M1 a deux orthologues positionnels [Dewey, 2011].

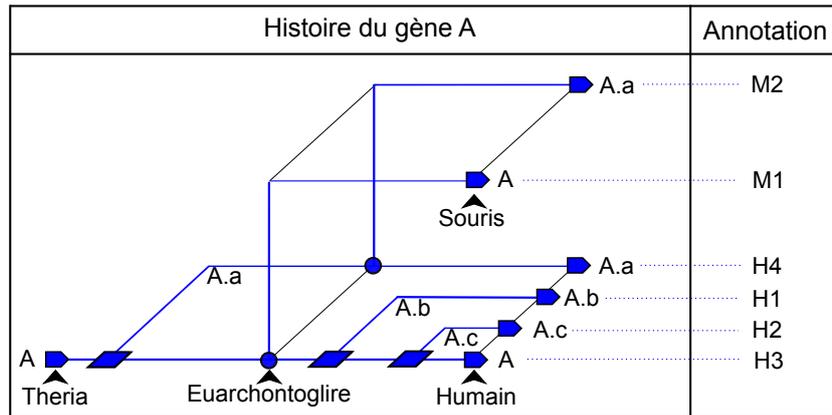


Figure I.11 – Arbre de gène de la famille du gène A depuis l'ancêtre Theria jusqu'à l'humain et la souris. Dans la colonne de droite les noms des gènes modernes sont attribués durant l'annotation sans connaître l'histoire évolutive.

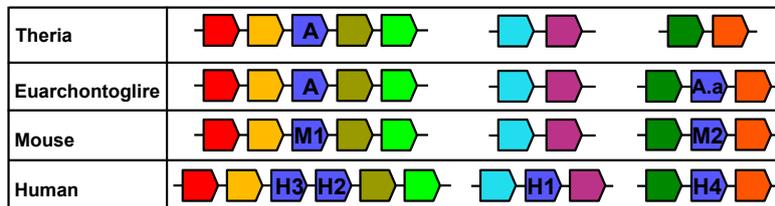


Figure I.12 – Contexte génomique des orthologues. Les co-orthologues H3 et H2 ont le même contexte génomique et il est donc impossible avec ces informations uniquement de distinguer lequel des deux est le gène ancestral et lequel est la copie. M1 a donc deux orthologues positionnels dans le génome de l'humain, H3 et H2.

I.5.4 Conservation de l'identité d'un gène

Dans ce travail nous avons utilisé et nous utiliserons la notion de gène de manière à ce que l'identité d'un gène soit conservée même s'il est dupliqué. Comme précédemment (section I.1.8), le nouveau gène inséré suite à une duplication est la *copie*. Par exemple si un gène est dupliqué de manière

disperse, la copie est le gène inséré loin du gène d'origine et le gène d'origine est toujours lui-même, au même endroit, après la duplication. Un gène défini de cette manière correspond à ce qui a été appelé le « true exemplar » ou le « descendant direct » [Sankoff, 1999]. Si un évènement de spéciation a lieu nous écrivons qu'un gène ancestral pré-spéciation est présent en deux *instances* dans les deux lignées. Dans la figure I.11 les instances du même gène dans différentes lignées sont nommées avec le même nom de gène. Pour distinguer les deux instances d'un même gène, nous spécifions la lignée dans laquelle est présente l'instance du gène. La durée de vie d'un gène peut être définie comme le temps entre sa naissance et la délétion de sa dernière instance.

Même si cette notion de gène dont l'identité est conservée à travers les évènements de duplication et les évènements de spéciation peut être source de confusion dans le cas d'une duplication en tandem, ce concept nous sera très utile pour définir les segments conservés. Par exemple, en l'absence d'information sur l'histoire évolutive, des co-orthologues positionnels, comme les gènes H3 et H2 dans la figure I.12, sont considérés comme d'équi-probables gènes pré-duplication. L'arbre de gène de la figure I.11 nous donne l'information manquante : le gène A (H3) a été copié et le gène A.c (H2) est la copie. Cette dernière information pourrait avoir été obtenue avec des informations plus fines sur le contexte génomique, sur les séquences des deux gènes et sur leurs voisinages nucléotidiques. Au final, malgré la duplication en tandem, avec ces informations, nous serions en mesure d'identifier H3 comme le gène pré-duplication. L'usage de duplications « dirigées » qui distinguent le gène copié du gène nouvellement inséré, même en présence de duplications en tandem nous semble ainsi justifiable.

I.5.5 Gène ancestral

Lorsqu'un ancêtre a été défini, un gène dans un génome descendant, qui était présent dans le génome de l'ancêtre, sera souvent qualifié par la suite de *gène ancestral*. Autrement dit, les gènes ancestraux sont également les descendants directs des gènes de l'ancêtre. Cependant, nous préférons considérer que l'identité de chaque gène du génome ancestral a été conservée à travers les évènements de duplication et de spéciation. Avec ces conventions, durant l'évolution, le gène peut être présent, dans les génomes de différentes espèces. Cette définition d'un gène ancestral dont l'identité est conservée à travers les évènements de duplications et de spéciations nous sera extrêmement pratique pour définir les blocs de synténie et les segments conservés (section II.3.2). Si nous considérons qu'après une duplication le gène ancestral (A) est remplacé par deux nouveaux gènes (A.a et A.b) il nous serait beaucoup plus difficile de

définir les blocs de synténie et les segments conservés.

I.6 Les points de cassures

I.6.1 Vestige d'un flanc de point de cassure

À notre échelle d'étude, nous appelons *point de cassure*, l'intergène (ou le télomère) brisé par une cassure double brin liée à un réarrangement chromosomique. *Les flancs d'une cassure* sont définis de la même manière que les flancs d'un intergène (ou d'un télomère). Un point de cassure a généralement deux flancs, l'un à gauche et l'autre à droite de l'intergène cassé (figure I.13). Si le point de cassure est situé à un télomère du chromosome, il n'y a qu'un flanc.

Par la suite nous qualifierons de *vestige d'un flanc de cassure* l'extrémité d'un gène ancestral qui a flanqué un point de cassure. À cause des délétions de gènes ancestraux il est parfois nécessaire de transférer le vestige d'un flanc de cassure à l'extrémité du gène ancestral le plus proche. Ceci est expliqué dans la figure I.14.

I.6.2 Réutilisation des points de cassures

Si une cassure brise un chromosome dans un intergène dont un des flancs est le vestige d'un flanc d'une cassure précédente, nous conviendrons qu'il s'agit là d'une *réutilisation de point de cassure*.

I.7 Représentation d'un chromosome

I.7.1 Distances et gaps entre deux gènes

Le *gap* entre deux gènes d'un chromosome est égal au nombre de gènes qu'il y a entre eux. La *distance* entre ces deux gènes est égale au gap qui les sépare plus un. Deux gènes adjacents sont séparés par un gap nul et ils sont à une distance de un gène.

I.7.2 Clusters de gènes dupliqués en tandem

Si un gène est dupliqué en tandem de nombreuses fois et si ses copies en tandem sont elles aussi copiées de nombreuses fois en tandem, l'ensemble de ces gènes paralogues sont proches les uns des autres autour du gène d'origine et nous écrirons qu'il forment un *cluster de gènes dupliqués en tandem*. De manière

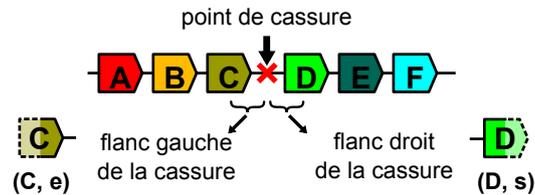


Figure I.13 – Point de cassure et ses deux flancs.

événements	vestige du flanc gauche de la cassure	évolution du génome	vestige du flanc droit de la cassure
génom initial			
duplication de gène en tandem			
fission		<p style="text-align: center;">cassure</p>	
après fission			
délétion de gène			
duplication de gène disperse			
naissance de gène de novo			

Figure I.14 – Évolution des vestiges de deux flancs d’une même cassure. Une cassure a lieu entre les extrémités de gènes ancestraux (C, e) et (D, s). Par la suite plusieurs évènements géniques altèrent les régions cassées. Après chaque évènement, dans la colonne de gauche l’extrémité du gène ancestral (C,e) reste le vestige du flanc gauche de la cassure. Dans la colonne de droite, l’extrémité du gène ancestral (D,s) est initialement le vestige du flanc droit de la cassure. Lorsque ce gène est supprimé, l’extrémité du gène ancestral le plus proche, (E,s), devient le vestige du flanc droit de la cassure. Les insertions de nouveaux gènes n’altèrent pas les vestiges de flancs de cassures.

générale, un *cluster de duplications en tandem de gaps* $\leq \text{tandemGapMax}$ (une valeur entière arbitraire) est un ensemble de gènes de la même famille dont chacun des gènes est séparé d'un autre gène du cluster par au plus tandemGapMax gènes qui ne sont pas de la famille. Nous ne considérons que les clusters maximaux, les gènes d'un cluster d'une famille ne sont pas inclus dans un autre cluster de cette famille. Dans les génomes modernes il y a de nombreux clusters de duplications en tandem et nous en déduisons donc qu'il y a eu de nombreuses duplications en tandem. La figure I.15 détaille la fréquence des duplications en tandem qui se sont produites depuis Amniota (il y a environ 325 Millions d'années) et que l'on retrouve encore en tandem dans les génomes de cinq espèces modernes descendantes, en fonction du paramètre *tandemGapMax* croissant.

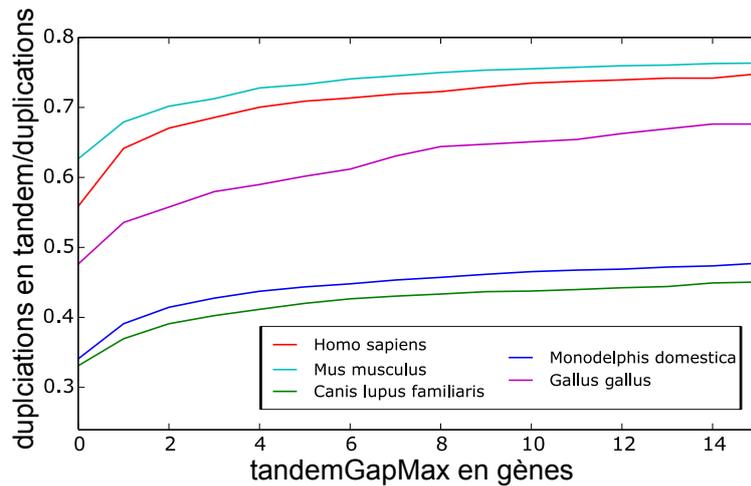


Figure I.15 – Fraction des duplications qui ont eu lieu en tandem depuis 325 millions d'années jusqu'à cinq espèces modernes. Le nombre de duplications en tandem qui ont donné naissance à un cluster de gènes de la même famille est estimé par le nombre de gènes dans le cluster moins un (le gène à l'origine du cluster). La fraction des gènes dupliqués en tandem varie substantiellement d'une lignée à l'autre (de environ 40% à 70%).

I.7.3 Réécriture d'un chromosome en tandem blocs

Dans ce paragraphe nous *réécrivons* les chromosomes de manière à ne conserver qu'un seul gène par cluster de gènes dupliqués en tandem, qui sera représentatif du gène à l'origine de chaque cluster. Dans un premier temps les clusters sont formés en regroupant tous les gènes d'une même famille séparés par au plus tandemGapMax gènes, un paramètre défini par l'utilisateur. Suite à cela,

chaque cluster est réduit à une unique entité, appelée tandem bloc [Lucas *et al.*, 2014]. Intuitivement un *tandem bloc* représente le gène à l'origine du cluster : il est situé à la localisation du gène ancestral et il a la même orientation que le gène ancestral. Néanmoins, dans les faits, il n'est pas toujours possible de déterminer la localisation du gène à l'origine du cluster ni l'orientation de ce gène. Sans informations suffisantes pour déterminer la véritable localisation, celle-ci sera donc choisie arbitrairement comme la localisation du premier gène du cluster, selon l'ordre du chromosome. Si tous les gènes du cluster ont la même orientation, le gène d'origine avait *a priori* la même orientation aussi et l'orientation du tandem bloc est choisie identique à cette orientation consensuelle. Par contre si au moins un des gènes du cluster a une orientation opposée à l'orientation d'un autre gène du cluster, nous considérons que l'orientation du gène ancestral n'est pas identifiable et nous attribuons une orientation « inconnue » (\emptyset) au tandem bloc. Les gènes d'un tandem bloc sont les gènes du cluster que le tandem bloc représente. La taille d'un tandem bloc est égale au nombre de gènes de celui-ci. Un tandem bloc est en relation d'homologie avec un deuxième tandem bloc si leurs gènes font partie de la même famille.

La figure I.16 illustre la réécriture d'un chromosome en tandem blocs avec un $tandemGapMax = 1$. Un gène qui ne faisait pas partie d'un cluster est considéré comme un tandem bloc de taille 1 après la réécriture. Nous verrons plus tard l'utilité de cette réécriture pour identifier les segments conservés à partir de comparaisons de génomes d'espèces modernes (section II.4.1).

En réalité cette étape de réécriture, bien qu'elle soit très utile, n'est pas non plus sans défaut. Durant la réécriture, le gène à l'origine du cluster est positionné arbitrairement là où se situe le premier gène du cluster et par conséquent le résultat de la réécriture peut varier selon l'orientation de référence du chromosome avant réécriture, figure I.16B et I.16C.

Dans notre précédent travail [Lucas *et al.*, 2014] un « tandem bloc » était un cluster de gènes dupliqués en tandem avec des gaps nuls, c'est à dire un segments de gènes adjacents qui sont tous de la même famille. Dans le travail présent un tandem bloc peut correspondre à un cluster qui contient des gaps non nuls si le $tandemGapMax$ utilisé lors de la réécriture est supérieur à zéro.

Une fois les chromosomes réécrits, les tandems blocs d'un chromosome peuvent être considérés comme des gènes. Comme pour ces derniers, nous écrirons que le *gap* entre deux tandems blocs est égal au nombre de tandems blocs qu'il y a entre eux et la *distance* entre ces deux tandems blocs sera égale au gap qui les sépare plus un.

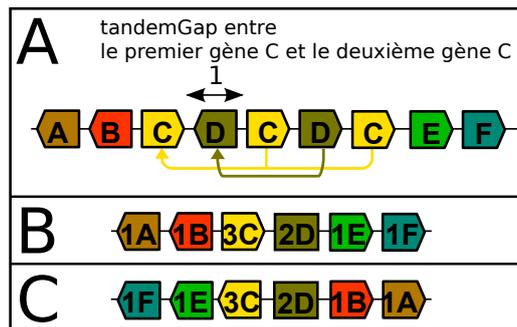


Figure I.16 – Réécriture en tandem blocs avec un $tandemGapMax = 1$. Le **panel A** détaille le processus de réécriture du chromosome humain en réduisant les clusters de gènes dupliqués en tandem. Si $tandemGapMax = 1$ il y a deux clusters, un de 3 gènes de la famille C et un autre cluster de 2 gènes de la famille D. Les trois gènes C forment un cluster car pas plus de $tandemGapMax$ autres gènes séparent chaque paire de gènes C voisins. Tous les gènes d'un cluster sont réduits à un unique gène positionné là où se trouve le premier gène du cluster, comme indiqué par la flèche jaune. Le cluster de gènes de la famille D subit le même traitement. Le **panel B** montre le résultat de la réécriture précédente. Le tandem bloc du cluster de gènes C contient 3 gènes et il a une orientation positive, comme tous les gènes qu'il contient. Le tandem bloc du cluster de gènes D n'a pas d'orientation car ses gènes ont des orientations opposées. Les autres gènes qui ne faisaient pas partie d'un cluster sont des tandems blocs de taille un. Le **panel C** donne l'exemple de la réécriture du chromosome si l'orientation de référence inverse du chromosome d'origine avait été choisie. Nous constatons que l'ordre relatif des tandem blocs n'est plus le même que précédemment.

I.8 Comparaison de deux chromosomes

Considérons deux chromosomes :

- le chromosome $c_a = [g_{a,k}]_{k \in [1, N_a]}$ du génome d'une espèce S_a et
- le chromosome $c_b = [g_{b,k}]_{k \in [1, N_b]}$ du génome d'une espèce S_b .

Conformément à la convention précédente (section I.5.2) les familles de gènes sont définies de manière à ce que les arbres correspondants prennent chacun racine dans un gène unique du MRCA de S_a et S_b .

I.8.1 Comparaison des orientations des gènes

Nous définissons le *signe* d'un couple de gènes, $g_{a,i}$ du chromosome c_a et $g_{b,j}$ du chromosome c_b , noté $g_{a,i} \bullet g_{b,j}$, de la manière suivante :

$$g_{a,i} \bullet g_{b,j} = \begin{cases} +1, & \text{si } o(g_{a,i}) = o(g_{b,j}) \\ -1, & \text{si } o(g_{a,i}) = -o(g_{b,j}) \end{cases} \quad (\text{I.1})$$

avec $o(g_{a,i})$ l'orientation du gène $g_{a,i}$ dans le chromosome c_a et $o(g_{b,j})$ l'orientation du gène $g_{b,j}$ dans le chromosome c_b . Si les gènes $g_{a,i}$ et $g_{b,j}$ sont homologues ($g_{a,i} \mathcal{H} g_{b,j}$), le signe de ce couple de gènes sera souvent appelé *signe de l'homologie* qui relie $g_{a,i}$ et $g_{b,j}$.

I.8.2 Matrice d'homologies

La Matrice d'Homologies classique $MH \in \mathfrak{M}_{N_a, N_b}$ de deux chromosomes $c_a = [g_{a,k}]_{k \in [1, N_a]}$ et $c_b = [g_{b,k}]_{k \in [1, N_b]}$, de respectivement N_a et N_b gènes, est définie telle que :

$$MH[i, j] = \begin{cases} g_{a,i} \bullet g_{b,j}, & \text{si } g_{a,i} \mathcal{H} g_{b,j} \\ 0, & \text{sinon} \end{cases} \quad \forall (i, j) \in [1, N_a] \times [1, N_b]. \quad (\text{I.2})$$

Une MH peut être représentée par un tableau de valeurs égales à $+1$, -1 ou 0 . Chaque valeur non nulle correspond à une homologie entre deux gènes homologues et plus particulièrement au signe de cette homologie, voir figure I.17A.

I.8.3 Matrice de packs d'homologies

Si c_a est réécrit en n_a tandem blocs $c_a = [tb_{a,1}, \dots, tb_{a,n_a}]$ et si c_b est réécrit en n_b tandems blocs $c_b = [tb_{b,1}, \dots, tb_{b,n_b}]$ nous introduisons une nouvelle matrice, la Matrice de Packs d'Homologies, $MHP \in \mathfrak{M}_{n_a, n_b}$ de deux chromosomes $c_a = [tb_{a,k}]_{k \in [1, n_a]}$ et $c_b = [tb_{b,k}]_{k \in [1, n_b]}$, que nous définissons de la manière suivante :

$$MHP[i, j] = \begin{cases} tb_{a,i} \bullet tb_{b,j}, & \text{si } tb_{a,i} \mathcal{H} tb_{b,j} \\ 0, & \text{sinon} \end{cases} \quad \forall (i, j) \in [1, n_a] \times [1, n_b] \quad (\text{I.3})$$

avec $tb_{a,i} \bullet tb_{b,j}$ le *signe* du *pack d'homologies* entre $tb_{a,i}$ et $tb_{b,j}$

$$tb_{a,i} \bullet tb_{b,j} = \begin{cases} +1, & \text{si } o(tb_{a,i}) = o(tb_{b,j}) \\ -1, & \text{si } o(tb_{a,i}) = -o(tb_{b,j}) \\ \emptyset, & \text{si } o(tb_{a,i}) = \emptyset \text{ ou } o(tb_{b,j}) = \emptyset \end{cases} . \quad (\text{I.4})$$

En d'autres termes la construction de cette matrice MHP est similaire à la matrice MH de c_a et c_b , avec des tandem blocs et des packs d'homologies à la place des gènes et des homologies. Néanmoins, alors qu'un gène a toujours une orientation connue, un tandem bloc peut avoir une orientation *inconnue* qui génère des packs d'homologies de signes inconnus (\emptyset). De manière similaire, une MHP peut être représentée par un tableau de valeurs égales à +1, -1, \emptyset ou 0. Les valeurs non-nulles correspondent à des packs d'homologies entre deux tandem blocs. Les figures I.17A et I.17B donnent une représentation graphique de la transition entre la MH et la MHP lorsque les chromosomes sont réécrits en tandem blocs.

I.8.4 Distances et gaps entre deux homologies dans une matrice

Dans une matrice d'homologies, nous définissons la *distance* entre deux homologies, comme la distance 2D entre les coordonnées 2D des deux homologies. À chaque relation d'homologie, de deux gènes homologues, correspond une *coordonnée* : une paire d'indices (x, y) . Le premier indice, x , est le rang du gène homologue dans le premier chromosome et le deuxième indice, y , est le rang du gène homologue dans le deuxième chromosome.

Si (x_0, y_0) et (x_1, y_1) sont les coordonnées de deux homologies dans une matrice, selon la métrique utilisée, la distance entre ces deux homologies est

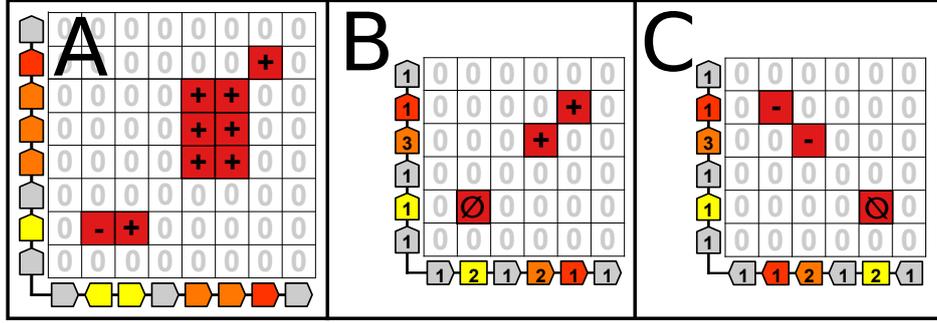


Figure I.17 – Matrice d’homologies et matrice de packs d’homologies après la réécriture des chromosomes en tandem blocs ainsi qu’un exemple de diagonale backslash. Le **panel A** représente une matrice d’homologies de deux chromosomes écrits en gènes. Le **panel B** représente la matrice de packs d’homologies après la réécriture des chromosomes précédents en tandems blocs. Le **panel C** représente la même matrice de packs d’homologies que dans le panel B, si l’ordre du chromosome en abscisse est inversé. Dans les illustrations de matrices d’homologies les signes des homologies égaux à +1 (resp. -1) sont représentées par des + (resp. -).

calculée par l’une des équations suivantes :

$$d_{CD}((x_0, y_0), (x_1, y_1)) = \max(|x_1 - x_0|, |y_1 - y_0|) \quad (\text{I.5})$$

$$d_{ED}((x_0, y_0), (x_1, y_1)) = \sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2} \quad (\text{I.6})$$

$$d_{MD}((x_0, y_0), (x_1, y_1)) = |x_1 - x_0| + |y_1 - y_0| \quad (\text{I.7})$$

$$d_{DPD}((x_0, y_0), (x_1, y_1)) = 2\max(|x_1 - x_0|, |y_1 - y_0|) - \min(|x_1 - x_0|, |y_1 - y_0|) \quad (\text{I.8})$$

où $[x]$ est l’entier le plus proche de x . Chebyshev Distance (CD), Euclidian Distance (ED), Manhattan Distance (MD) et Diagonal Pseudo-Distance (DPD) sont quatre métriques de distances différentes. Nous représentons les distances calculées à partir de ces différentes métriques dans la figure I.18.

Dans ce qui suit nous utiliserons exclusivement la métrique de Chebyshev. Par exemple, avec cette métrique, dans la figure I.17A, le 3^{ème} gène du 1^{er} chromosome (en abscisse) est homologue au 2^{ème} gène du deuxième chromosome (en ordonnée). La coordonnée de l’homologie est donc (3, 2) et la valeur de la matrice d’homologie à cette coordonnée est $MH[3, 2] = +1$. Le signe de l’homologie, égal à +1 étant donné que les deux gènes ont la même orientation. De même, la coordonnée (5, 4) correspond à l’homologie entre le 5^{ème} gène du 1^{er} chromosome et le 4^{ème} gène du deuxième chromosome. La distance qui

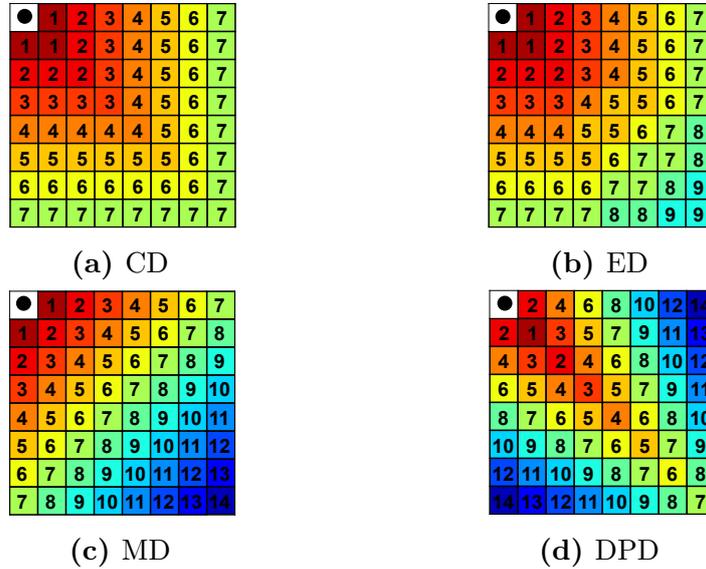


Figure I.18 – Différentes métriques pour mesurer des distances en deux dimensions. Les valeurs des distances sont calculées par rapport au point noir en haut à gauche de chaque matrice. Plus la couleur tire vers le rouge, plus le point est « près » du point noir.

sépare la première homologie (de coordonnée $(3, 2)$) de la deuxième homologie (de coordonnée $(5, 4)$) est égale à $d((3, 2), (5, 4)) = \max(|3 - 5|, |2 - 4|) = 2$.

Le *gap* entre deux homologies est égal à la distance entre les homologies moins un. Si nous reprenons l'exemple précédent le *gap* entre la première homologie $(3, 2)$ et la deuxième $(5, 4)$ est égal à $d((3, 2), (5, 4)) - 1 = 1$.

Les distances et les *gaps* qui séparent les packs d'homologies dans une matrice de packs d'homologies sont définis de la même manière en substituant les gènes par des tandems blocs dans les définitions¹.

I.8.5 Diagonales dans une matrice

Nous définissons une relation d'ordre sur l'ensemble des coordonnées 2D. Une coordonnée est strictement inférieure ($<$) à une autre : $(x_1, y_1) < (x_2, y_2)$, si $x_1 < x_2$ ou si $(x_1 = x_2 \text{ et } y_1 < y_2)$. Deux coordonnées sont égales si $x_1 = x_2$ et si $y_1 = y_2$. Enfin, si une première coordonnée n'est pas inférieure à une deuxième et si ces deux coordonnées diffèrent (ne sont pas égales), alors la première coordonnée est strictement supérieure ($>$) à la deuxième. Dans ce

¹Le lecteur pourra se référer à notre précédent travail s'il veut une explication plus développée [Lucas *et al.*, 2014].

dernier cas, soit $x_1 > x_2$, soit ($x_1 = x_2$ et $y_1 > y_2$). Les relations inférieur ou égal (\leq) et supérieur ou égal (\geq) se déduisent directement de ce qui précède. Dans une matrice d'homologies trier les coordonnées $(x_1, y_2), \dots, (x_N, y_N)$ d'un ensemble d'homologies de manière croissante revient à trier les coordonnées des homologies par coordonnées croissantes grâce à la relation d'ordre définie précédemment. Cela revient à trier dans un premier temps les homologies par composantes x croissantes et à trier dans un deuxième temps les homologies qui ont la même première composante (x) par composantes y croissantes.

Dans une matrice d'homologies une liste ordonnée de N homologies (de la 1^{ère} à N^{ème}) forme une diagonale

- *slash* si $\begin{cases} x_i \leq x_{i+1} \\ y_i \leq y_{i+1} \end{cases} \quad \forall i \in [1, N - 1]$
- et *backslash* si $\begin{cases} x_i \leq x_{i+1} \\ y_i \geq y_{i+1} \end{cases} \quad \forall i \in [1, N - 1]$,

avec $(x_1, y_1), \dots, (x_N, y_N)$ les coordonnées respectives des N homologies. Dans le premier cas la diagonale est qualifiée de slash (/) car dans la matrice d'homologies, les homologies qui la composent sont distribuées selon une direction qui s'élanche d'en bas à gauche vers en haut à droite. Dans le deuxième cas la diagonale est qualifiée de backslash (\) car cette fois-ci les homologies qui la composent sont distribuées selon un axe qui se dirige d'en haut à gauche à en bas à droite.

Nous définissons de la même manière les diagonales dans les matrices de packs d'homologies en substituant les tandems blocs aux gènes et les packs d'homologies aux homologies. La figure I.17B représente une diagonale slash composée des packs d'homologies correspondant aux tandems blocs homologues jaunes, oranges et rouges (dans cet ordre). Dans la figure I.17C les packs d'homologies correspondant aux tandems blocs homologues rouges, oranges et jaunes (dans cet ordre) forment cette fois-ci une diagonale backslash. Dans la figure I.17A la liste des 9 homologies, ordonnées de manière à ce que les coordonnées soient triées par ordre croissant, ne forme pas une diagonale car de la 5^{ème} homologie (de coordonnée (5,6)) à la 6^{ème} (de coordonnée (6,4)) la deuxième coordonnée (y) décroît. Alors que les coordonnées des homologies précédentes (de la 1^{ère} à la 5^{ème}) s'enchaînaient avec des y croissants, la succession de la 5^{ème} à la 6^{ème} homologie se fait via une décroissance des y . Par contre dans la matrice de packs d'homologies correspondante (figure I.17B), la liste ordonnée de 5 homologies de coordonnées respectives (2,2), (3,2), (5,4), (6,5) et (6,7) forme bien une diagonale de type slash.

Deux listes ordonnées de gènes dont les gènes, de même rang, sont homologues, sont deux listes de gènes *colinéaires* si les homologues correspondantes forment une diagonale (slash ou backslash) dans une matrice d'homologies.

Nous appelons *gaps* d'une diagonale les gaps qui séparent deux homologues successives d'une diagonale. Par exemple dans la diagonale de 3 packs d'homologies de la figure I.17B, le gap entre le premier pack d'homologies et le deuxième est égal à $d_{CD}((2, 2), (4, 4)) - 1 = \max(|2-4|, |2-4|) - 1 = 2 - 1 = 1$.

Une diagonale *stricte* est une diagonale dont tous les gaps sont nuls, nous dirons parfois qu'elle n'a pas de gap. Les coordonnées $(x_1, y_1), \dots, (x_N, y_N)$ d'une telle diagonale vérifient l'une des propriétés suivantes.

- La diagonale est slash et
$$\begin{cases} x_{i+1} = x_i + 1 \\ y_{i+1} = y_i + 1 \end{cases} \quad \forall i \in [1, N - 1].$$
- La diagonale est backslash et
$$\begin{cases} x_{i+1} = x_i + 1 \\ y_{i+1} = y_i - 1 \end{cases} \quad \forall i \in [1, N - 1].$$

Enfin, une diagonale est *consistante* si elle vérifie l'une des deux propriétés suivantes.

- La diagonale est « slash » et les signes de ses homologues sont positifs ou *inconnus* (\emptyset).
- La diagonale est « backslash » et les signes de ses homologues sont négatifs ou *inconnus* (\emptyset).

Les diagonales et les propriétés précédentes nous seront utiles pour identifier les segments conservés et les blocs de synténie (figure II.7 et section II.3.3).

I.9 Comparaison de deux génomes

I.9.1 Matrice d'homologies de deux génomes

La *matrice d'homologies de deux génomes* est une super-matrice composée de toutes les matrices d'homologies correspondant aux différentes combinaisons de chromosomes des deux génomes. Un exemple d'une telle matrice est donné par la figure III.4 pour la comparaison du génome du poulet avec le génome du diamant mandarin (*Taeniopygia guttata*). La matrice de packs d'homologies de deux génomes est définie de la même manière.

I.9.2 Filtrages des gènes

Lorsque nous comparons deux génomes, une première étape consiste souvent à filtrer les gènes des génomes. Nous présentons cinq filtrages :

- *GènesAncestraux* conserve uniquement les gènes ancestraux, hérités de manière directe, depuis le MRCA. Ce filtre supprime donc tous les gènes qui ont été insérés par duplications ainsi que les gènes insérés suite à des naissances *de novo*¹.
- *GènesAncestrauxDansLesDeuxGénomes* conserve uniquement les gènes ancestraux qui ont été hérités depuis le MRCA et qui sont présents dans les deux génomes comparés. Par rapport au filtre précédent ce filtrage supprime donc les gènes ancestraux qui ont été supprimés dans une lignée même s'ils ont été conservés dans l'autre.
- *FamillesAncestrales* conserve uniquement les gènes qui descendent (directement ou par duplications) d'un gène ancestral du génome du MRCA. Comme le filtre *GènesAncestraux*, ce filtre supprime les gènes qui sont nés *de novo* dans les lignées. Par contre, les gènes qui sont dans les familles (définie au niveau du MRCA) sont tous gardés, même si ce sont en réalité des copies de gènes ancestraux.
- *FamillesAncestralesDansLesDeuxGénomes* conserve uniquement les gènes dont les familles ancestrales sont représentées par au moins un gène dans les deux génomes. Ce filtrage supprime tous les gènes qui n'ont pas d'homologue dans l'autre génome.
- *BrhDansLesDeuxGénomes* conserve uniquement les gènes dont les séquences sont appariées par un best-hit réciproque de similarité de séquences, Best Reciprocal Hit (BRH).

Les deux premiers filtres *GènesAncestraux* et *GènesAncestrauxDansLesDeuxGénomes* sont des filtrages idéaux qui ne sont pas toujours réalisables. Pour cette raison les trois filtres suivant *FamillesAncestrales*, *FamillesAncestralesDansLesDeuxGénomes* et *BrhDansLesDeuxGénomes* sont souvent utilisés comme des substituts pratiques.

¹Les gènes qui ont été insérés par duplications ainsi que les gènes insérés suite à des naissances *de novo* sont parfois appelés des *gènes spécifiques à une lignée*

I.10 Bases de données

I.10.1 Base de données pour l'arbre des espèces

Dans la suite de ce travail nos arbres d'espèces sont ceux de la base de donnée Ensembl. Cette base de donnée incorpore les arbres des espèces du NCBI et les estimations des dates de spéciations du projet Timetree [Hedges *et al.*, 2015].

I.10.2 Base de données pour nos génomes et pour les arbres de gènes

Nous utiliserons par la suite les génomes de la base de donnée Ensembl ainsi que les arbres de gènes correspondants, reconstruits par l'équipe Ensembl Compara [Vilella *et al.*, 2009]. Les arbres téléchargés ont été édités de manière à déplacer chaque nœud de duplication vers les feuilles tant que le score de consistance du nœud est inférieur à un seuil. Le *score de consistance* (« consistency score ») d'un nœud de duplication est égal au nombre d'espèces modernes qui possèdent chacune des descendants des deux gènes post-duplications divisés par le nombre total d'espèces qui possèdent chacune au moins un descendant d'un des deux gènes post-duplication. Plusieurs seuils sont testés et le seuil retenu est celui qui édite les arbres de manière à maximiser le nombre d'adjacences conservées entre les espèces. Plus exactement il s'agit du seuil qui édite les arbres de manière à ce que la méthode de reconstruction de génomes ancestraux AGORA [Muffato, 2010] infère, à partir des arbres édités, les chromosomes ancestraux les plus longs (en gènes). Le seuil optimal est de 30% et par conséquent, avec nos arbres édités, parmi les espèces qui possèdent au moins un descendant d'un gène post-duplication, pas moins de 30% de celles-ci ont chacune des descendants des deux gènes post-duplications.

Grâce aux définitions fondamentales précédentes nous définirons dans le prochain chapitre les segments conservés, qui nous seront d'une grande aide dans notre étude des réarrangements.

Chapitre II

Les segments conservés

Of all natural systems, living matter is the one which, in the face of great transformations, preserves inscribed in its organization the largest amount of its own past history. Using Hegel's expression we may say that there is no other system that is better *aufgehoben* (constantly abolished and simultaneously preserved). We may ask the questions where in the now living systems the greatest amount of their past history has survived and how it can be extracted.

EMILE ZUCKERKANDL et LINUS PAULING,
Molecules as documents of evolutionary history, 1965

II.1 Vestiges de génomes ancestraux

Les génomes actuels sont issus d'une succession de réplifications et de mutations qui dure depuis des milliards d'années. Malgré l'ensemble des altérations qui ont eu lieu, ces génomes ont néanmoins conservé de nombreux vestiges des génomes ancestraux.

Prenons le cas d'une séquence dans le génome du MRCA de deux espèces modernes S_1 et S_2 . Convenons que cette séquence est *conservée* tant qu'elle ne subit que des mutations ponctuelles et s'il est continuellement possible de l'identifier par similarité de séquence avant et après chaque mutation¹. Ces conditions excluent que la séquence subisse de grands réarrangements, par exemple la délétion d'un long segment de nucléotides. Si la séquence est conservée durant son évolution jusqu'aux deux espèces modernes S_1 et

¹Les *éléments ultraconservés* [Bejerano *et al.*, 2004] sont des cas extrêmes où les séquences sont, comme leurs noms l'indique, parfaitement conservées.

S_2 , leurs deux génomes contiennent l'un et l'autre un vestige de la séquence ancestrale et, selon les taux de mutations des lignées, la séquence héritée dans le premier génome sera plus ou moins similaire à la séquence du deuxième génome. Ainsi, en prenant le cas d'une séquence conservée de 100 paires de bases, si dans chaque lignée il y a eu, en moyenne, 0,2 substitutions/bp, les séquences modernes auront chacune une séquence qui sera similaire à la séquence ancestrale avec au moins 80% de similarité¹. Par conséquent la similarité des deux séquences modernes sera au minimum égale à 60%.

À partir de maintenant oublions cette histoire évolutive. Dans la réalité, nous n'avons accès qu'aux génomes modernes S_1 et S_2 . Admettons que nous savons qu'il y a eu, en moyenne, 0,2 substitution/bp dans chaque lignée depuis le dernier ancêtre commun de S_1 et S_2 . Séquences dupliquées mises à part, deux séquences, l'une dans le génome de S_1 et l'autre dans le génome de S_2 , toutes les deux de 100 bp et similaires entre elles à 65%, semblent *a priori* être les vestiges d'une séquence commune qui existait dans le génome de leur ancêtre commun. Néanmoins, en toute rigueur, une deuxième explication pourrait également expliquer autrement la similarité des séquences modernes. Il se pourrait en effet que les séquences n'aient pas d'origine commune et qu'elles proviennent par exemple de deux séquences ancestrales dissemblables², qui ont, après substitutions, atteint une similarité de 65%. Si les séquences de départ ne sont similaires que de 25% cela requiert que les 20 substitutions de chaque lignée aient fait converger 40 nucléotides qui différaient à l'origine. En partant du principe que les mutations se font aléatoirement, cette dernière explication est très improbable, en tout cas moins probable que l'hypothèse d'une origine commune. En validant l'origine commune des deux séquences modernes nous prenons un faible risque et étant donné l'histoire que nous avons présentée, nous aurions même raison ici. Dans d'autres cas il est beaucoup plus difficile de trancher entre les deux explications. Par exemple, quelle conclusion tirer si les deux séquences modernes de 100 bp sont similaires à 40% ? Quel risque prenons nous d'affirmer qu'elles sont toutes les deux les vestiges d'une même séquence ancestrale ?

Sans nous étendre davantage sur les nombreuses difficultés qui pourraient entraver le raisonnement précédent, nous considérerons malgré tout qu'il constitue une méthode fiable pour retrouver les vestiges des génomes ancestraux à partir des génomes modernes. Ce pourquoi il a été au cœur de plus de 30 années de recherches en génomique comparative, lors desquels il a justifié de multiples usages de l'algorithme BLAST (Basic Local Alignment Search Tool)

¹La similarité est ici le pourcentage de paires de nucléotides de mêmes rangs qui sont identiques.

²La similarité attendue d'après une distribution aléatoire de nucléotides dans deux séquences de longueur 100 bp est de 25%.

[Altschul *et al.*, 1990] pour identifier des séquences d'ancestralité commune.

Généralisons la logique précédente à d'autres entités qu'aux séquences. Pour retrouver les vestiges de ces entités nous procéderons à chaque fois à un raisonnement en cinq étapes.

1. Premièrement nous définirons les *entités* dans un génome ancestral.
2. Deuxièmement nous définirons les conditions nécessaires pour qu'une entité soit *conservée* durant l'évolution du génome ancestral et malgré les mutations qui peuvent l'altérer. Une entité qui a été conservée depuis un ancêtre sera appelée un *vestige* de l'entité ancestrale. Un vestige d'une entité ancestrale sera parfois nommé l'*instance* de cette entité dans le génome qui la contient.
3. Troisièmement nous expliquerons ce qu'implique la conservation de chaque entité durant l'évolution, depuis un ancêtre jusqu'à plusieurs espèces modernes. Ceci permettra d'explicitier les conditions nécessaires pour que des entités modernes (possiblement dans différents génomes) soient toutes les vestiges d'une même entité ancestrale. Si des entités modernes satisfont ces conditions, nous dirons qu'elles *semblent* être les vestiges d'une entité ancestrale commune, ou qu'elles semblent descendre d'une même entité ancestrale.
4. Quatrièmement, en utilisant les conditions nécessaires de conservation, nous identifierons dans les génomes modernes les entités qui semblent être les vestiges d'une même entité ancestrale. À la fin de cette étape nous aurons donc plusieurs ensembles d'entités qui correspondent chacun à une entité ancestrale.
5. Enfin, cinquièmement, si cela est possible, nous validerons ou nous rejetterons par différents critères les semblants d'origines communes.

La validation peut se faire de différentes manières. Elle peut être statistique, elle peut être effectuée par parcimonie ou par optimisation d'un coût et elle peut aussi se faire par une linéarisation de graphe. Dans l'exemple précédent, les deux séquences modernes semblent être les vestiges d'une même entité ancestrale et pour valider cette hypothèse nous effectuons une comparaison de probabilités. Étant donné le modèle d'évolution des séquences, ici les taux de substitutions, la quantification qui sert de critère est la différence entre p_1 , la probabilité que les séquences modernes semblent conservées car elles descendent effectivement d'une même entité ancestrale (hypothèse H_1) moins p_0 , la probabilité que les séquences modernes semblent conservées alors

qu'elles ne descendent pas d'une même entité ancestrale (hypothèse H_0). Si la différence est supérieure à un seuil, la parenté commune est validée, sinon elle est rejetée ou considérée comme indécise. Le risque de se tromper est d'autant plus faible que le seuil est haut. Celui-ci quantifie en quelque sorte le risque de faux positifs. Au cours du chapitre nous donnerons d'autres exemples qui illustreront d'autres types de validations.

Nous dirons que la validation statistique est analytique si elle est effectuée à l'aide d'une formule mathématique. Dans le cas contraire la validation peut être basée sur des simulations de l'évolution *in silico*.

Nous avons vu que l'entité dont nous étudions les vestiges peut être une séquence dont les nucléotides mutent. Dans la suite nous nous intéresserons à des entités dans lesquelles la localisation des gènes peut varier.

II.2 Vestiges d'une co-localisation de deux gènes ancestraux

II.2.1 Vestiges de deux gènes synténiques

Parmi les relations de co-localisation de gènes ou de marqueurs, la relation de synténie fut la première à avoir été étudiée¹. Suivons le raisonnement en cinq étapes que nous avons établi précédemment.

1. Deux gènes ancestraux sont synténiques (en relation de synténie) dans le génome ancestral s'ils sont sur le même chromosome.
2. La synténie entre ces deux gènes est conservée tant qu'ils restent sur un même chromosome.
3. Si la synténie est conservée jusqu'aux espèces modernes, les deux gènes sont donc sur le même chromosome dans toutes ces espèces modernes. Deux gènes qui sont synténiques dans les espèces modernes semblent donc être les vestiges d'une synténie ancestrale.
4. Recensons tous les couples de gènes ancestraux qui sont synténiques dans toutes les espèces modernes. Chacune de ces paires de gènes semble, d'après ce qui précède, correspondre à un vestige d'une synténie ancestrale.

¹Les études étaient effectuées par coloration de marqueurs, plus précisément grâce à la méthode Fluorescent *In-Situ* Hybridization (FISH).

5. Pour valider cette hypothèse nous calculerons cette fois-ci la probabilité que deux gènes ancestraux soient par hasard en synténie dans toutes les espèces modernes. Admettons que nous étudions 2 espèces modernes qui ont toutes les deux 10 chromosomes. D'après un modèle d'évolution simpliste qui considère que les mutations ont été telles que les gènes de l'ancêtre ont été redistribués aléatoirement dans les chromosomes modernes, la probabilité que les deux synténies soient le fruit du hasard est $p = \left(\frac{1}{10}\right)^2 = 1\%$ ¹. Nous en concluons que quand deux gènes sont synténiques dans la première espèce et quand ils le sont également dans la deuxième, alors nous pouvons affirmer que ces synténies modernes sont les vestiges d'une même synténie ancestrale, avec un risque de nous tromper estimé à une chance sur 100.

Bien entendu le modèle d'évolution, qui considère que les gènes sont relocalisés aléatoirement durant l'évolution peut être complètement faux. Ainsi, dans certains calculs [Mazowita *et al.*, 2006] les translocations réciproques sont estimées être plus fréquentes entre grands chromosomes qu'entre petits chromosomes. Cet autre modèle d'évolution modifie la probabilité précédente. Supposons qu'un génome ancestral de deux très grands chromosomes et d'autres petits chromosomes donne, après spéciation et évolution, naissance à deux espèces modernes qui ont, elles aussi, deux très grands chromosomes et des petits chromosomes. Les nombreuses translocations réciproques qu'il y a probablement eu entre les deux grands chromosomes augmentent fortement le risque que des relations de synténie modernes soient dues au hasard.

Par la suite, pour les validations, nous ferons couramment l'hypothèse pratique que l'évolution est telle que les gènes ancestraux sont redistribués aléatoirement dans les génomes modernes. Car ce modèle d'évolution est souvent le seul pour lequel une validation analytique existe.

II.2.2 Vestiges d'autres co-localisations entre deux gènes ancestraux

Comme nous avons étudié les vestiges de synténies ancestrales entre deux gènes, nous pourrions également étudier les vestiges d'adjacences ancestrales [Bérard *et al.*, 2012], ou les vestiges d'adjacences relâchées (« gapped adjacencies ») [Gagnon *et al.*, 2012] avec des raisonnements similaires. Dans la première étude [Bérard *et al.*, 2012] la validation de l'ancestralité des adjacences de

¹Cette dernière probabilité est égale à la probabilité que le deuxième gène soit sur le même chromosome que le premier, dans le premier génome moderne, $p_1 = \frac{1}{10}$, multipliée par la probabilité analogue, $p_2 = \frac{1}{10}$, pour le deuxième génome.

gènes est basée sur une optimisation de « coûts ». La quantité qui est utilisée ici n'est plus une différence de probabilité, mais la différence entre

- le « coût » c_1 associé aux mutations qui expliquent que les adjacences modernes semblent conservées car elles descendent effectivement d'une même adjacence ancestrale (hypothèse H_1)
- moins le « coût » c_0 associé aux mutations qui expliquent que les adjacences modernes semblent conservées alors qu'elles ne descendent pas d'une même adjacence ancestrale (hypothèse H_0).

Si cette dernière différence est supérieure à zéro la parenté commune est validée, sinon elle est rejetée. Si le coût est égal au nombre de mutations nécessaires pour réaliser chaque hypothèse, le raisonnement précédent est équivalent à un raisonnement qui choisi le scénario le plus parcimonieux. Dans la deuxième étude [Gagnon *et al.*, 2012], la validation de l'ancestralité des adjacences relâchées est indirecte, elle est faite lors de la linéarisation du graphe d'adjacence. Ce graphe est construit grâce à l'agglomération des informations provenant des adjacences relâchées qui semblent ancestrales. Comme ce graphe n'est pas toujours linéaire, les auteurs le linéarisent pour s'assurer que le résultat est conforme à ce qui est attendu : un génome ancestral composé de chromosomes linéaires.

Dans tous les cas qui précèdent l'entité étudiée est une relation (synténie, adjacence ou adjacence relâchée) entre les localisations deux gènes ancestraux. Étudions maintenant une entité qui implique plusieurs gènes.

II.3 Vestiges d'une co-localisation de plusieurs gènes ancestraux

Comme nous avons défini une synténie entre deux gènes, nous définissons une synténie entre un ensemble de gènes. Un ensemble de gènes est synténique si tous les gènes se trouvent sur le même chromosome. À partir de cette définition, en appliquant notre raisonnement en 5 étapes avec plusieurs gènes au lieu de seulement deux, nous serions capables de généraliser les conclusions précédentes à propos des vestiges de relations de synténie. Au lieu de faire cela, nous nous proposons d'appliquer notre méthodologie à une autre entité : à un cluster de gènes.

II.3.1 Vestiges d'un cluster de gènes ancestraux

Dans un génome ancestral un *cluster de gènes* de gaps $\leq gapMax$ est un ensemble de gènes ancestraux qui a la propriété suivante : chaque gène du cluster est distant d'au plus $gapMax$ gènes ancestraux d'un autre gène du cluster¹. Durant l'évolution, le cluster de gaps $\leq gapMax$ est conservé si les écarts entre les gènes du cluster restent inférieurs ou égaux à $gapMax$ gènes ancestraux². Avec cette définition, idéalement, les duplications de gènes et les naissances *de novo* de gènes qui ont lieu durant l'évolution du génome ancestral n'ont aucune conséquence sur la conservation du cluster et elles sont donc négligées³. Il s'ensuit que, dans les espèces modernes, un cluster de gènes, de gaps $\leq gapMax$, qui a été conservé depuis un ancêtre commun, est un cluster de gènes dont chaque gène est distant d'au plus $gapMax$ gènes ancestraux d'un autre gène du cluster. Là encore, seuls les gènes ancestraux sont comptabilisés dans le calcul du gap entre deux gènes. Par la suite, les clusters que nous étudierons sont les clusters maximaux, c'est à dire les clusters qui ne sont pas inclus dans d'autres clusters.

Plusieurs algorithmes existent pour recenser les clusters de gènes ancestraux qui semblent avoir été conservés. Team [Bergeron *et al.*, 2002] est le plus basique, il permet de les retrouver si toutes les localisations des gènes ancestraux ont été identifiées dans les génomes modernes. HomologyTeams [He et Goldwasser, 2005] est plus flexible que Team car il permet de prendre en compte les localisations des paralogues des gènes ancestraux, ce qui est utile lorsqu'il n'a pas été possible d'identifier les localisations des gènes ancestraux parmi les localisations de ses paralogues. L'indécision sur la localisation du gène ancestral requiert alors de comptabiliser les paralogues des gènes ancestraux dans le calcul de gaps. Un dernier algorithme GTT (Gene Team Tree) [Zhang et Leong, 2009] permet d'identifier et de représenter les clusters de gènes ancestraux par une structure hiérarchique en arbre dans laquelle chaque nœud de l'arbre correspond à un cluster de gènes ancestraux de gaps $\leq g$ (un entier) et dont les nœuds enfants correspondent à des clusters de gènes ancestraux de gaps $\leq g - 1$. Enfin, de nombreux calculs de probabilités ont été menés pour valider l'ancestralité des clusters [Hoberman *et al.*, 2005][Raghupathy *et al.*, 2008][Raghupathy et Durand, 2009].

¹Nous rappelons qu'un gène ancestral est un gène qui descend directement (sans duplication) d'un gène du génome de l'ancêtre considéré (section I.5.5). Un tel gène existe dans le génome ancestral, il n'est pas né durant l'évolution du génome ancestral et il n'est pas non plus issu d'une duplication d'un gène de l'ancêtre.

²Un cas extrême de cluster de gènes conservé de gaps ≤ 0 correspond à la notion de « common interval » [Uno et Yagiura, 2000].

³Dans la pratique il n'est pas si simple de négliger les duplications de gènes, mais faisons comme si cela était possible pour le moment.

Les clusters de gènes conservés sont couramment utilisés pour étudier les génomes de levures et de procaryotes. Une raison possible est que ces organismes subissent principalement des inversions de segments chromosomiques courts, ce qui expliquerait que l'ordre des gènes n'est pas conservé mais que les gènes restent généralement proches les uns des autres [Sankoff, 2002]. Ces clusters ont également été utilisés pour étudier la dynamique des « domaines » géniques et protéiques [Pasek *et al.*, 2005]. À la différence des organismes unicellulaires mentionnés précédemment, les eucaryotes multicellulaires, comme la plupart des plantes et des animaux, sont caractérisés par une forte conservation de l'ordre des gènes. Ceci laisse penser que les inversions sont plus longues [Sankoff, 2002] ou qu'il y a peu d'inversions courtes. Notre objectif est d'étudier l'évolution de l'ordre des gènes de vertébrés, nous allons donc considérer une entité plus adaptée.

II.3.2 Les segments conservés, vestiges des segments de chromosomes non réarrangés

Définition d'un segment conservé

Le concept de segment conservé a été utilisé en 1984 [Nadeau et Taylor, 1984] pour étudier les réarrangements entre l'humain et la souris à partir de la localisation de plus de 83 marqueurs (de mêmes origines ancestrales) localisés le long des chromosomes de l'humain et de la souris. Cette notion de segments conservés est au cœur de toute la suite de notre étude.

En nous appuyant sur la méthodologie énoncée plus haut, nous procédons à notre raisonnement en 5 étapes pour les définir. Dans le génome ancestral un segment conservé est un segment de chromosome constitué d'une succession ininterrompue de gènes ancestraux. Durant l'évolution *un segment est conservé tant qu'aucune cassure liée à un réarrangement ne le fragmente*¹. De plus, le segment reste conservé même si un de ses gènes ancestraux est supprimé, à moins que le gène supprimé soit le dernier gène ancestral du segment, dans ce cas le segment est perdu. Enfin, les segments conservés depuis un ancêtre sont choisis maximaux, un segment conservé n'est pas inclus dans un autre segment conservé.

Nous précisons quelques points concernant la définition précédente. Les cassures considérées ici sont celles qui sont liées à un des réarrangements chromosomiques : inversion, translocation réciproque, fission et transposition

¹Cette dernière propriété vaut parfois aux segments conservés d'être appelés des « atomes ». C'est le cas dans l'article de Jian Ma [Ma *et al.*, 2008a]. Néanmoins, à la différence de cette étude, nous considérons qu'un segment issu d'une duplication d'un segment conservé n'est pas un segment conservé car il ne contient pas de gènes ancestraux.

(section I.3.2). Les cassures qui ne précèdent que des duplications et celles dues au fonctionnement normal des topo-isomérases de classe II ne sont pas considérées. En bref, un segment conservé n'a pas subi de réarrangement interne de l'ordre de ses gènes ancestraux et ceux-ci préservent donc leurs rangs et leurs orientations les uns par rapport aux autres. Si un réarrangement a lieu entre les gènes ancestraux d'un segment conservé (sans que ce réarrangement perturbe l'ordre ou les orientations des gènes ancestraux du segment) le segment conservé n'est pas fragmenté. Autrement dit les réarrangements qui n'impliquent que des gènes non ancestraux ne changent pas les segments conservés et un réarrangement fragmente un segment conservé uniquement s'il modifie l'ordre ou les orientations de transcription de ses gènes ancestraux. Quand un segment a été conservé, aucun gène ancestral extérieur au segment ne s'y est inséré car nous estimons qu'une telle transposition de gène nécessiterait des points de cassures pour délocaliser le gène ancestral et le repositionner. Comme chaque gène ancestral est présent en une instance unique par génome moderne il n'y a qu'un seul segment qui contient ces gènes ancestraux.

Nous mentionnons une exception à la définition précédente : avec notre modélisation des génomes il n'est pas possible de discerner une délétion de gène ancestral due à une pseudogénéisation (sans point de cassure) d'une délétion de gène qui serait due à l'excision spécifique de ce gène unique par une transposition mono-génique (avec deux points de cassure). Toutes les transpositions d'un unique gène ancestral seront donc traitées comme des pseudogénéisations et elles ne fragmenteront pas la conservation du segment.

Vestiges d'un segment conservé

Si un segment est conservé de l'ancêtre jusqu'à plusieurs espèces modernes, alors, dans chacune de ces espèces modernes il est présent dans un chromosome. Selon l'ordre de référence du chromosome moderne qui le contient, le segment sera soit dans le même ordre que dans le chromosome ancestral (1^{er} cas), soit dans l'ordre inverse (2^{ème} cas). Pour expliciter ce dernier point nous assignons une orientation au segment conservé, la même que l'orientation du chromosome ancestral. La figure II.1 donne un exemple de la manière dont pourrait se manifester la conservation d'un segment dans les génomes de deux espèces modernes. Dans la première espèce ce segment est présent sur un chromosome c_1 et, ici, l'orientation de référence du chromosome c_1 est la même que l'orientation du segment conservé. Le segment conservé apparaît dans le même ordre que dans le chromosome ancestral. Dans la deuxième espèce le segment est présent sur le chromosome c_2 et cette fois-ci il se trouve que l'orientation du chromosome c_2 est inverse par rapport à l'orientation du segment. Le segment apparaît donc dans l'ordre inverse. Par ailleurs,

conformément à notre définition, l'insertion du gène non-ancestral durant l'évolution de c_a à c_1 ne casse pas le segment conservé.

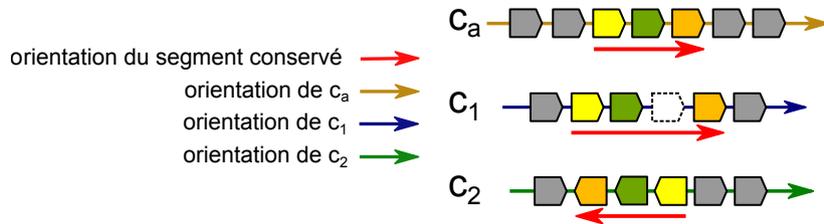


Figure II.1 – Conservation de l'ordre des gènes et de leurs orientations au sein d'un segment conservé. Dans le génome ancestral, le segment conservé est représenté, le long du chromosome c_a , par les trois gènes, jaune, vert et orange, dans cet ordre. Les gènes gris sont d'autres gènes ancestraux. Les flèches colorées correspondent aux orientations de références du segment conservé et des chromosomes. Dans le chromosome c_1 le gène blanc représente un gène non ancestral, comme par exemple un gène qui a été inséré dans le segment à la suite d'une duplication.

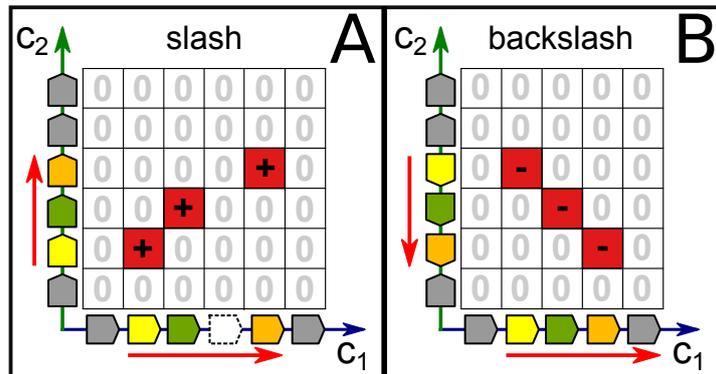


Figure II.2 – Diagonale représentant un segment conservé dans une matrice d'homologies. La matrice d'homologies correspond à la comparaison des deux chromosomes modernes c_1 et c_2 de la figure II.1. Dans le **panel A**, le chromosome c_1 est représenté avec son gène non ancestral (entouré d'une ligne en pointillés) et l'orientation de référence du chromosome c_2 est ici la même que celle du segment conservé. Le segment conservé forme alors une diagonale slash consistante. Dans le **panel B**, le chromosome c_1 est représenté sans son gène ancestral et cette fois-ci l'orientation de référence du chromosome c_2 a été fixée inverse à celle du segment conservé. Ce dernier forme donc une diagonale backslash consistante.

La figure II.2 montre que la conservation d'un segment ancestral jusqu'à deux espèces modernes S_1 et S_2 implique qu'il y a au moins une matrice d'homologies (correspondant à la comparaison d'un chromosome de S_1 et d'un chromosome de S_2) dans laquelle le segment conservé forme une diagonale (section I.8.5).

De plus, la conservation de l'ordre des gènes ancestraux (la colinéarité) au sein du segment implique que la diagonale est consistante. Enfin, si les génomes ont été filtrés avec le filtre *GènesAncestrauxDansLesDeuxGénomes*, la diagonale est stricte¹.

Nous utiliserons souvent la notion de segments conservés par la suite. Nous nous proposons donc de l'illustrer avec des exemples détaillés.

Illustration de l'évolution des segments conservés

La figure II.3 illustre l'évolution du génome de l'ancêtre *Anc* jusqu'à deux espèces modernes S_1 et S_2 . La connaissance des localisations des points de cassures est suffisante pour définir les segments conservés. Initialement le génome ancestral est composé d'autant de segments conservés que de chromosomes. Durant l'évolution de ce génome, selon les évolutions considérées, les segments initiaux sont progressivement fragmentés par les points de cassures. Considérons l'évolution des segments *de l'ancêtre Anc jusqu'à S_1* de la figure II.3B. Comme initialement il n'y a qu'un seul chromosome, il n'y a par conséquent qu'un segment à l'origine. L'inversion 1 avec les points de cassures 1 et 2 fragmente ce segment en trois segments et la fission ultérieure fragmente une nouvelle fois un des segments en deux segments. Au final il y a 4 segments conservés durant l'évolution de *Anc* jusqu'à S_1 . Considérons ensuite l'évolution des segments *de Anc jusqu'aux deux espèces modernes S_1 et S_2* de la figure II.3C. Là encore il n'y a qu'un seul chromosome à l'origine, il y a donc, de même, un unique segment initialement. Durant l'évolution qui aboutit aux deux espèces modernes, cette fois-ci il y a eu un total de 5 cassures qui ont progressivement fragmenté le segment ancestral en 6 segments conservés. Au cours du processus de fragmentation les segments du génome ancestral sont conservés tant qu'ils ne sont pas cassés. Dans les deux cas, les segments conservés durant une évolution peuvent s'obtenir en cassant le génome ancestral aux différents points de cassure qui ont eu lieu durant l'évolution considérée.

¹Si les affirmations précédentes ne sont pas évidentes, la section 4 de notre précédent travail [Lucas *et al.*, 2014] pourra les rendre plus compréhensibles. Les notions de colinéarité, de diagonale, stricte et/ou consistante, ont été définies dans le chapitre précédent (section I.8.5), tout comme le filtrage *GènesAncestrauxDansLesDeuxGénomes* (section I.9.2).

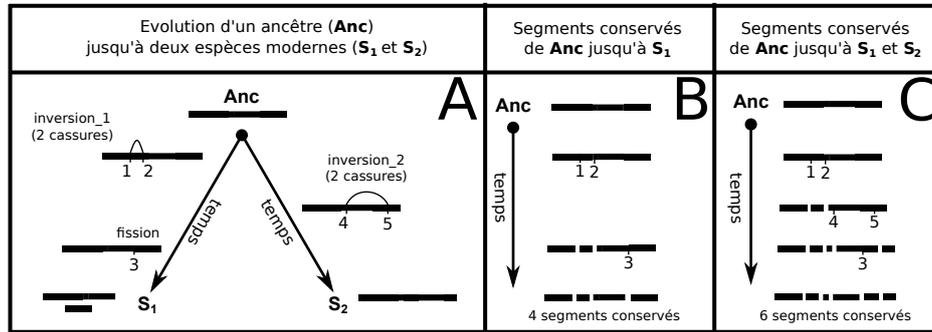


Figure II.3 – Évolution d'un génome ancestral jusqu'à deux espèces modernes et évolutions simultanées de ses segments conservés selon les lignées considérées. Panel A, évolution du génome ancestral le long de chaque branche de l'arbre qui relie les espèces modernes S_1 et S_2 à leur MRCA, Anc . **Panel B**, évolution des segments conservés de l'ancêtre jusqu'à l'espèce moderne S_1 . **Panel C**, évolution des segments conservés jusqu'aux deux espèces modernes S_1 et S_2 .

Considérons un exemple plus détaillé. La figure II.4 décrit graphiquement l'évolution d'un génome ancestral Anc jusqu'à une première espèce moderne S_1 ainsi que l'évolution des segments conservés correspondante. Au début de l'évolution, les segments conservés sont des copies parfaites des chromosomes ancestraux. Par la suite chaque cassure du génome fragmente un segment au niveau de l'adjacence entre les gènes ancestraux où cette cassure a lieu¹. Nous décrivons chacune des 8 étapes de cette évolution qui font suite aux conditions initiales².

1. Initialement le génome contient deux chromosomes, le chromosome 1 contient 15 gènes ancestraux (nommés de A à O) et le chromosome 2 en contient 8 (P à W). Les segments conservés sont pour le moment les copies exactes de ces chromosomes initiaux. Le segment conservé 1 contient les 15 gènes ancestraux du chromosome 1 et le segment conservé 2 contient les 8 gènes ancestraux du chromosome 2 et dans chaque segment conservé l'ordre et les orientations sont les mêmes que dans les chromosomes.
2. Au début de son évolution le génome initial subit une fission. Le point de

¹Plus exactement, c'est le cas si les cassures ne réutilisent pas de points de cassures (section I.6.2). Si une cassure réutilise un point de cassure aucun segment conservé n'est fragmenté.

²La cinquième et la 6^{ème} étape méritent l'attention, les autres peuvent être survolées dès que le raisonnement qui les sous-tend est compris.

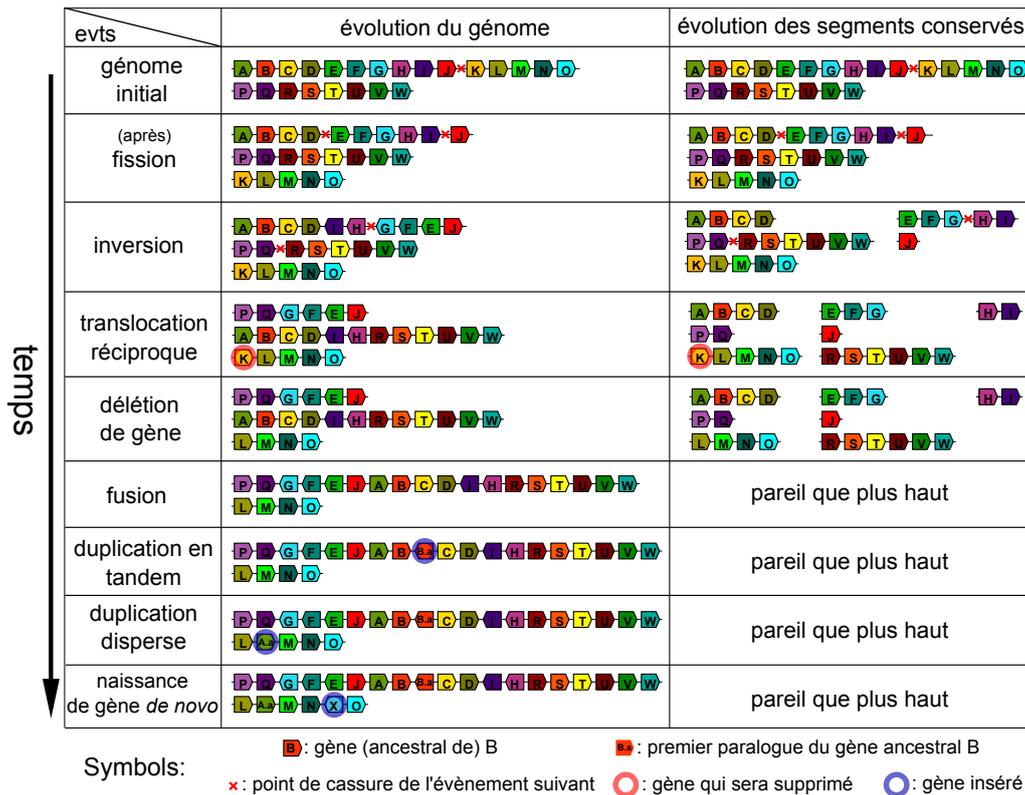


Figure II.4 – Évolution d'un génome d'un ancêtre *Anc* jusqu'à une espèce **modern S_1 .** Le génome initial contient 2 chromosomes, un de 15 gènes et le second de 8 gènes. Ce génome évolue du haut vers le bas selon la flèche du temps. Chaque ligne du tableau représente un moment. La première ligne correspond aux conditions initiales, la seconde ligne correspond à l'état du génome juste après l'évènement de fission, *etc.* La colonne de droite décrit l'évolution des segments conservés. Comme pour le génome, à chaque moment de l'évolution, les segments conservés depuis l'ancêtre sont représentés.

cassure tombe dans le 10^{ème} intergène du chromosome 1, entre la fin du gène J et le début du gène K. Ce point de cassure sépare le chromosome 1 en deux chromosomes. Après l'évènement de fission l'ancienne partie de gauche du chromosome 1 est le nouveau chromosome 1, tronqué, qui contient dorénavant 8 gènes et le second morceau forme un nouveau chromosome, le chromosome 3, qui contient les 7 gènes restants, de l'ancienne partie droite du chromosome 1. Les segments (qui ne sont plus conservés car l'évènement les a fragmenté) évoluent exactement de la même manière. Le segment 1 est fragmenté en deux segments, un petit segment conservé 1 qui contient 12 gènes et un nouveau segment conservé 3 de 8 gènes.

3. Puis une inversion a lieu dans le chromosome 1. Le segment de chromosome entre le 4^{ème} intergène et le 9^{ème} intergène est inversé. Les deux points de cassures de l'inversion brisent le chromosome 1 en trois morceaux. Dans les segments précédemment conservés, les points de cassures sont eux aussi localisés dans le segment 1 et ils fragmentent ce segment en trois segments. Le 1^{er} segment correspond au segment 1 précédent, tronqué de la partie à droite du point de cassure le plus à gauche. Le deuxième segment, le segment 4, correspond à l'intervalle entre les deux points de cassure et le troisième segment, segment 5, contient les gènes restant de l'ancien segment 1, situés à droite du point de cassure le plus à droite.
4. La translocation réciproque génère elle aussi deux points de cassures. Dans le génome qui précède l'évènement, ces deux points de cassures tombent dans le 6^{ème} intergène du chromosome 1 et dans le 2^{ème} intergène du chromosome 2. Chaque chromosome est brisé en deux. Du côté des segments, les deux points de cassures correspondants tombent dans les segments 2 et 4. Le segment 2 est fragmenté en deux nouveaux segments, l'un correspond à l'ancien segment 2 tronqué à gauche du point de cassure et l'autre segment est le nouveau segment 6 qui contient les gènes restants, de l'ancien segment 2, à droite du point de cassure. Le segment 4 évolue de la même manière et il est fragmenté en un segment 4 tronqué et un nouveau segment 7.
5. Un nouvel évènement a lieu ensuite, le gène ancestral K est supprimé du génome. Le contenu en gènes ancestraux des segments conservés suit cette évolution et le gène ancestral K est également supprimé dans le segment correspondant. Étant donné que l'extrémité (K,e) était le vestige d'un flanc de cassure, le nouveau vestige du flanc de cassure

dans le segment est, suite à la délétion, l'extrémité (L,e) car il s'agit de l'extrémité de gène ancestral la plus proche¹.

6. Les fusions, les duplications en tandem ou disperses ainsi que les naissances *de novo* ne modifient pas les segments conservés étant donné qu'il ne changent pas le contenu en gènes ancestraux et qu'il ne changent pas leur ordre ni leurs orientations de transcription.

La figure II.5 décrit l'évolution de l'ancêtre *Anc* (le même que précédemment) jusqu'à une autre espèce moderne, S_2 . Cette évolution se déroule parallèlement à l'évolution de la figure II.4. Nous ne décrivons pas la succession des événements car la logique est la même que dans la description précédente.

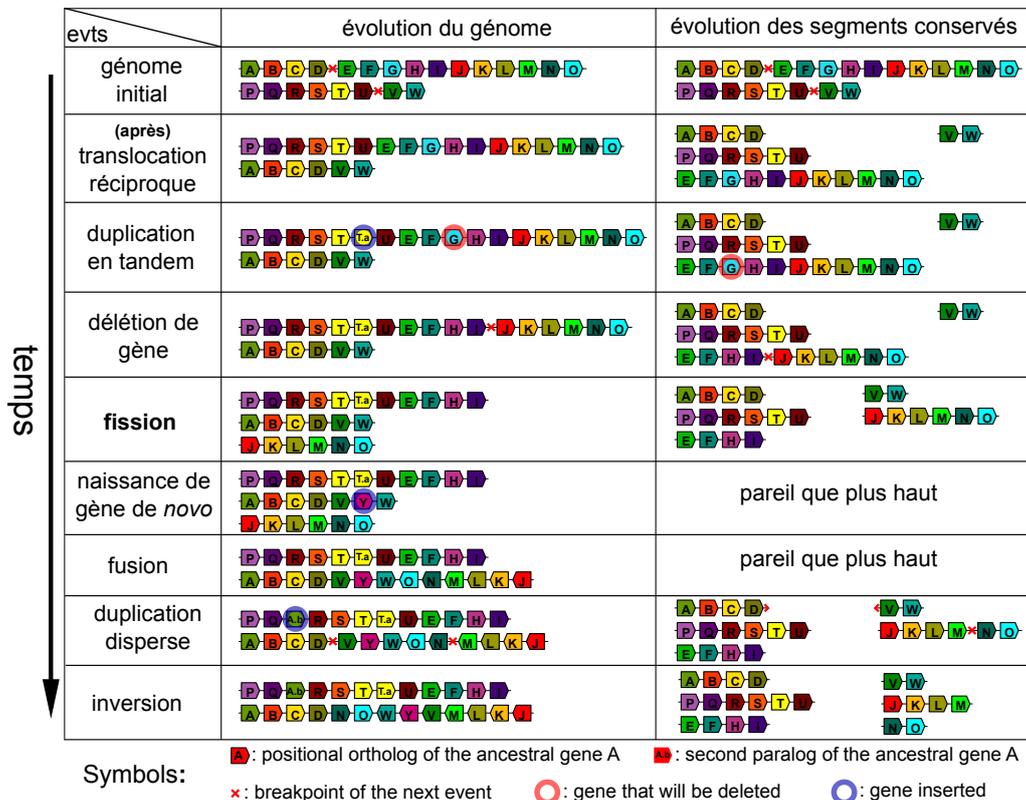


Figure II.5 – Évolution parallèle du génome ancestral de *Anc* jusqu'à l'espèce moderne S_2 et évolution correspondante des segments conservés.

¹Les extrémités de gènes ont été définies dans la section I.2.1 et les vestiges de flancs de cassures ont été définis dans la section I.6.1.

Les segments conservés le long de la lignée jusqu'à S_1 peuvent être comparés aux segments conservés le long de la lignée jusqu'à S_2 pour en déduire les segments conservés de *Anc* jusqu'aux deux espèces S_1 et S_2 . Dans le cas de l'évolution d'une lignée, pour obtenir ses segments conservés, il suffit de fragmenter le génome ancestral au niveau des points de cassures qui ont eu lieu le long de celle-ci, voir figure II.3B. Si les deux lignées sont considérées, il faut alors prendre en compte l'ensemble des points de cassures qui ont eu lieu dans les deux lignées, voir figure II.3C. Pour faire simple, la figure II.3 ne fait pas intervenir d'évènements de délétion de gènes. Ces évènements modifient pourtant les segments conservés. Par exemple la délétion décrite dans la figure II.4, supprime le gène ancestral K du segment qui le contenait et par la suite il n'y a que les gènes ancestraux conservés dans les segments conservés. Il en va de même pour le gène G dans la figure II.5. La variation différentielle du contenu en gène doit être prise en compte lorsque l'on étudie les segments conservés d'un ancêtre jusqu'à deux espèces modernes.

La figure II.6 explique comment, à partir des segments conservés dans chaque lignée, à un même moment¹, il est possible d'en déduire les segments conservés simultanément dans les deux lignées au même moment.

Les génomes modernes sont comparés à l'aide de leur matrice d'homologies (figure II.7). Comme précédemment, les relations d'homologies entre les gènes ancestraux sont représentées par des cellules non nulles. Néanmoins, dans la pratique il est souvent impossible de discerner la localisation du gène ancestral parmi les localisations de ses in-paralogismes (s'il en a). Nous sommes donc forcé de considérer les localisations des in-paralogues de gènes ancestraux comme de possibles gènes ancestraux. C'est pourquoi, ici, dans la matrice d'homologie nous représentons au final toutes les relations d'homologies entre les gènes de la famille d'un gène ancestral². Enfin, les gènes apparus à la suite d'une naissance *de novo* ne sont pas représentés, ni les gènes ancestraux supprimés dans au moins une lignée car ces gènes ne sont d'aucune utilité pour détecter les segments conservés.

Comme précédemment avec la figure II.2A, nous constatons dans la figure II.7 que, dans les matrices d'homologies, les segments conservés se manifestent par des diagonales consistantes, à l'exception de quelques lignes horizontales ou verticales. Ce que nous expliquons par la présence de duplications en tandem.

Effets des duplications en tandem Durant les évolutions précédentes, les duplications en tandem ont donné naissance aux gènes T.a et B.a et elles

¹dans ce cas à la fin de leurs évolutions

²filtrage *FamillesAncestralesDansLesDeuxGénomes*, voir section I.9.2

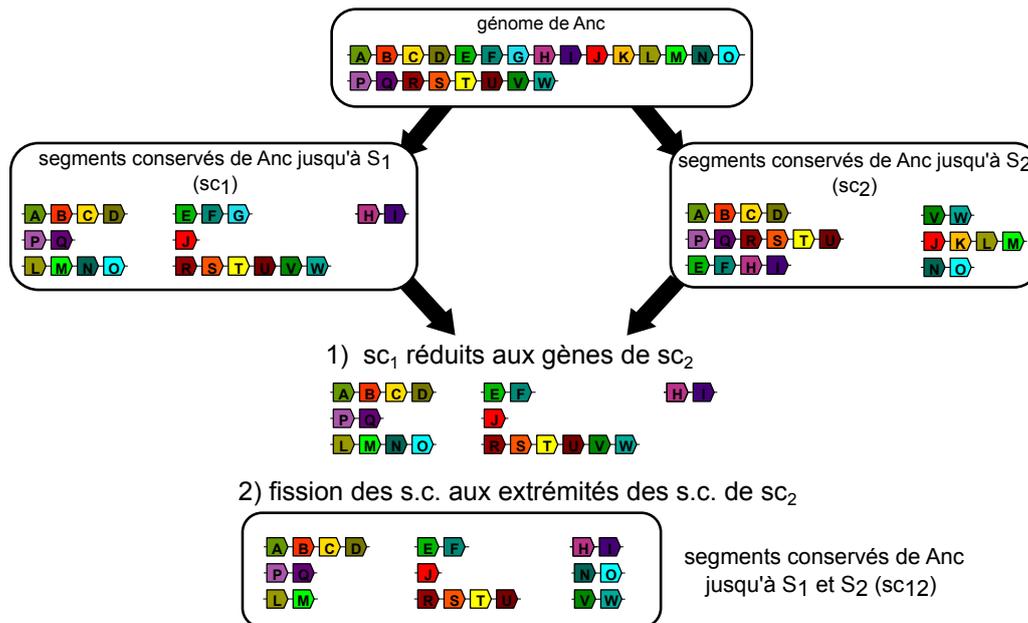


Figure II.6 – Dédudion des segments conservés le long de deux lignées à partir des segments conservés dans chaque lignée au même moment. En haut le génome de l'ancêtre. À gauche (resp. à droite), les segments conservés dans la lignée qui aboutit à l'espèce S_1 (resp. l'espèce S_2). Un des ensembles de segments conservés est choisi arbitrairement (ici c'est celui de S_1) et le contenu en gènes ancestraux est modifié pour qu'il ne reste que les gènes ancestraux conservés depuis l'ancêtre jusqu'aux espèces modernes S_1 et S_2 (étape 1). Cette étape supprime le gène ancestral G qui a été perdu dans la lignée de S_2 . Ensuite les segments conservés choisis sont fragmentés à toutes les localisations des extrémités des segments conservés non choisis (ici ceux de S_2), car ces extrémités correspondent aux points de cassures qu'il y a eu dans l'autre lignée (étape 2). Si à l'étape 1 nous avions choisis les segments de S_2 , le résultat aurait été le même.

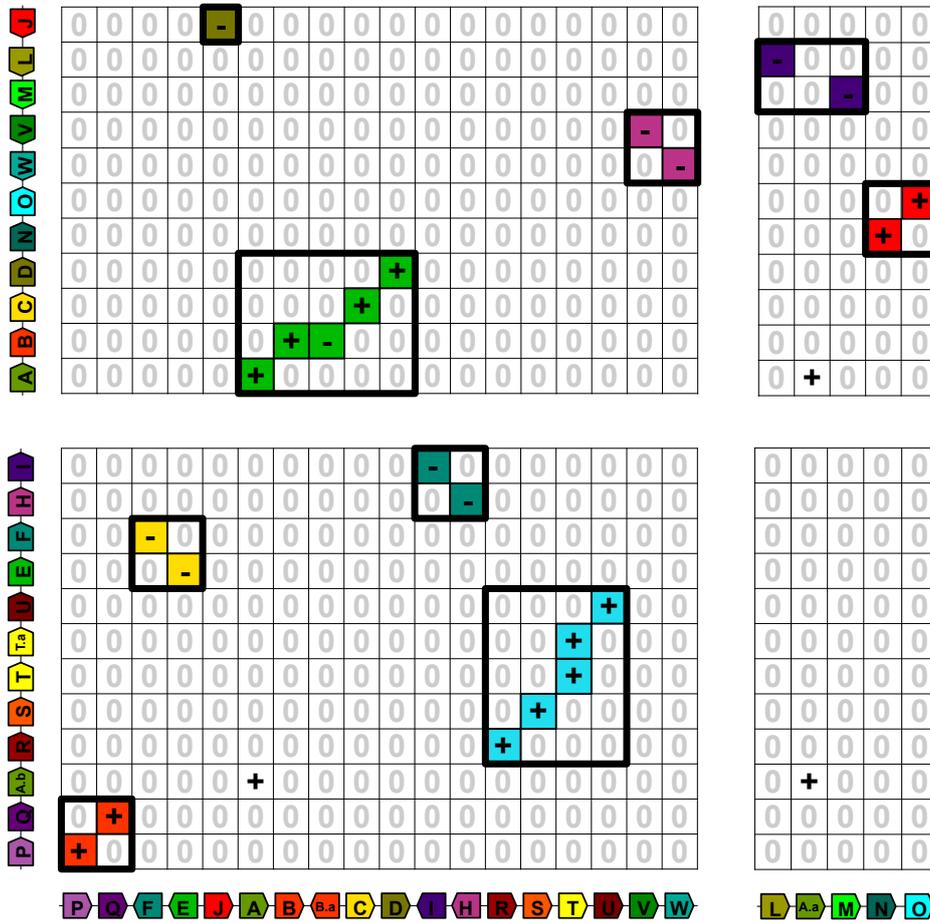


Figure II.7 – Matrice d’homologies de la comparaison des génomes de S_1 (en abscisse) et S_2 (en ordonnée). Dans cette matrice, les gènes issus de naissances *de novo* (X et Y) et ceux qui n’ont pas d’homologue dans l’autre génome (K et G) ne sont pas représentés. Les relations entre les gènes modernes d’une même famille de gène ancestral sont représentées par des cellules non-nulles. Pour les différencier des autres homologies, les homologies entre gènes ancestraux sont représentées par des cellules non-nulles coloriées. Chacune de ces homologies a une couleur qui correspond au segment ancestral auquel l’homologie appartient.

ont donc généré des clusters de duplications en tandem. Les in-paralogues T et T.a forment un tandem bloc dans le génome de S_1 et les in-paralogues B et B.a forment un deuxième tandem bloc dans le génome S_2 . Dans la figure II.7, la relation d'orthologie entre le gène B du génome de S_1 et le tandem bloc [B, B.a] du génome de S_2 se manifeste par une ligne horizontale composée de deux cellules non nulles. De même, la relation d'orthologie entre le gène T du génome S_1 et le tandem bloc [T, T.a] du génome de S_2 forme une autre ligne, verticale cette fois. Si un autre gène, T.b, fut inséré en tandem par une duplication durant l'évolution de *Anc* jusqu'à S_1 , la ligne verticale deviendrait un carré de quatre cellules non nulles représentant les relations d'homologie entre quatre gènes (T et T.b dans le génome de S_1 , T et T.a dans le génome de S_2).

Vestiges des segments conservés avec un filtrage idéal des génomes comparés Les diagonales sont par contre bien des diagonales consistantes et strictes dans la matrice d'homologies lorsque les génomes sont filtrés avec le filtre *GènesAncestrauxDansLesDeuxGénomes*. En plus de ne conserver que le véritable gène ancestral des clusters de duplications en tandem, ce filtre évite également d'avoir des gaps entre les homologies, causés par les délétions de gènes ancestraux qui ont eu lieu spécifiquement dans une lignée¹.

Vestiges des segments conservés avec un filtrage pragmatique des génomes comparés Le filtrage *GènesAncestrauxDansLesDeuxGénomes* est malheureusement difficile à mettre en pratique, comme nous le mentionnions (section I.9.2). Ce pourquoi nous lui substituons dans les faits le filtrage *FamillesAncestralesDansLesDeuxGénomes*, suite à quoi les génomes filtrés sont comparés avec la matrice de packs d'homologies (MHP) dans laquelle, les segments conservés se manifestent aussi par des diagonales consistantes; avec cette fois-ci, par contre, des gaps non nuls, causés par les duplications disperses.

Les segments conservés mono-géniques Nous profitons de l'exemple de la figure II.7 pour définir les *segments conservés mono-géniques* (composés d'un unique gène ancestral). Un tel segment est représenté par une cellule non nulle et isolée en haut de la matrice d'homologie de la figure II.7. Durant les évolutions de *Anc* jusqu'à S_1 et S_2 , le segment composé du gène ancestral J a été conservé sans subir de réarrangement interne. Un point de cassure dû à la fission dans la lignée de S_1 , un deuxième point de cassure dû à la fission dans la lignée de S_2 et un dernier point de cassure dû à l'inversion de

¹Nous aurions de tels gaps si nous utilisions le filtrage *GènesAncestraux*.

la lignée de S_2 sont tombés aux deux extrémités du gène. En fin d'évolution, dans le génome de S_1 , le gène J se trouve entre le segment conservé (FE) et le segment conservé (ABCD), et dans le génome S_2 , il se trouve entre le segment conservé (LM) et une extrémité de chromosome. Ce qui fait que cette cellule se trouve à la verticale de la bordure droite de la diagonale jaune, à la verticale de la bordure gauche de la diagonale verte, à l'horizontale de la bordure haute du segment violet et sur la bordure haute de sa matrice d'homologie.

La forte ressemblance de cette cellule isolée avec les cellules isolées dues à l'insertion d'une copie dispersée (comme les cellules représentant les homologies des gènes A.a ou A.b) ne doit pas cacher la différence qu'il y a entre une homologie isolée de deux instances modernes d'un gène ancestral (J) et une homologie isolée entre un gène non-ancestral, issu d'une duplication dispersée (A.a ou A.b), avec ses homologues.

Avec l'exemple précédent nous avons illustré ce que sont les segments conservés et comment ils se manifestent, à quelques corrections près, sous la forme de diagonales dans les matrices d'homologies des génomes modernes dans lesquels ils ont été conservés. Passons à la quatrième étape de notre raisonnement, l'identification des vestiges de segments conservés à partir des génomes modernes. La plupart des logiciels que nous avons étudiés identifient ces segments conservés grâce à l'identification préalable de blocs de synténie [Pevzner et Tesler, 2003a].

II.3.3 Les blocs de synténie, un moyen pour retrouver les segments conservés

Le terme « bloc de synténie » est d'ailleurs souvent utilisé comme un synonyme de segment conservé [Choi *et al.*, 2007]. La confusion entre bloc de synténie et segment conservé ainsi que les multiples définitions de l'un et de l'autre compliquent la comparaison des méthodes [Ghiurcuta et Moret, 2014]. De plus, de nombreux concepts utilisés dans la littérature représentent des entités extrêmement similaires aux blocs de synténie et aux segments conservés : les *chaînes de blocs* [ZHANG *et al.*, 1994][Kent *et al.*, 2003], les Homologous Synteny Blocks (HSBs) [Murphy *et al.*, 2005], les *strips* et *pre-strips* [Choi *et al.*, 2007], les *atoms* [Ma *et al.*, 2008a], les Locally Collinear Blocks (LCBs) [Darling *et al.*, 2010], les *multiplicons* et les *baseclusters* dans le logiciel i-ADHoRe 3.0 [Proost *et al.*, 2012], *etc.*

Nous avons déjà donné notre définition d'un segment conservé dans la section II.3.2. La définition d'un bloc de synténie, que nous retenons, se base sur la définition intuitive que Pevzner et Tesler [Pevzner et Tesler, 2003a] ont

formulé les premiers ainsi : « *synteny blocs are segments that can be converted into conserved segments by micro-rearrangements* »¹. Plus loin les auteurs ajoutent que les blocs de synténie dans les génomes modernes ne correspondent pas forcément à des régions de similarité continue entre deux génomes et qu'ils sont, au lieu de cela, souvent constitués de courtes régions similaires qui peuvent être interrompues par des régions dissemblables et des gaps. Ils affirment même que la plupart des blocs de synténie sont sujets à des micro-réarrangements en leur sein. À l'origine « bloc de synténie » n'était donc pas synonyme de segment conservé. Les blocs de synténie ont été inventés pour pouvoir s'abstraire des micro-réarrangements qui étaient souvent dus à des erreurs d'assemblage². Pour avoir une idée intuitive, il peut être pratique de se représenter un bloc de synténie comme un segment conservé dans lequel les micro-réarrangements sont négligés. Malheureusement, comme le concèdent Pevzner et Tesler, contourner ainsi les erreurs d'annotation a un prix. En rejetant tous les micro-réarrangements, les véritables micro-réarrangements sont eux aussi rejetés, et avec eux des points de cassures qui ont réellement brisés des chromosomes, entre deux extrémités d'un même bloc de synténie. Considérer qu'un bloc de synténie est un segment conservé sera souvent faux car il y a de nombreux et véritables micro-réarrangements qui n'ont rien à voir avec des erreurs d'assemblage, par exemple des micro-inversions [Lefebvre *et al.*, 2003][Chaisson *et al.*, 2006]. De plus, l'identification des blocs de synténie repose, par définition, sur la conservation de la relation de synténie entre au moins deux marqueurs. Par conséquent, les blocs de synténie ne pourront jamais être des segments conservés mono-géniques et ils ne pourront jamais rendre compte des inversions mono-géniques.

Pour avoir une définition formelle d'un bloc de synténie Pevzner et Tesler recommandent à leurs lecteurs de se référer à la description de leur algorithme GRIMM-synteny. Nous choisissons ici de définir les blocs de synténie, non pas en nous basant sur un algorithme, dont certains fonctionnements sont arbitraires, mais sur une entité conservée depuis un ancêtre. Nous montrerons comment notre définition rejoint la définition intuitive de Pevzner et Tesler.

Voici notre définition : *un bloc de synténie est un cluster conservé de gaps $\leq gapMax$, contenant au moins 2 gènes ancestraux, dont l'ordre des gènes et leurs orientations sont également conservés.*

Ici, en plus de contraindre la distance entre les gènes d'un cluster, nous

¹« Les blocs de synténie sont des segments qui peuvent être transformés en segments conservés par des micro-réarrangements ».

²C'était particulièrement le cas en 2003 avec les génomes tout juste séquencés de l'humain et de la souris.

ajoutons une contrainte sur leur ordre, leurs orientations ainsi que sur le nombre minimal de gènes ancestraux. Re commençons notre raisonnement en 5 étapes (section II.1) et de nouveau considérons un cluster de gènes de gaps $\leq gapMax$ dans le génome ancestral. Tant que les gaps¹ entre les gènes ancestraux du cluster ne dépassent pas $gapMax$, et tant que l'ordre de ces gènes ainsi que leurs orientations sont conservées, le bloc de synténie est conservé. À l'inverse un bloc de synténie est fragmenté dès que l'ordre ou les orientations de ses gènes ancestraux changent, ou lorsqu'un gap devient supérieur à $gapMax$. Les fragments sont alors les blocs de synténie composés des gènes ancestraux qui ont conservé leurs ordres et leurs orientations et des gaps $\leq gapMax$ malgré l'évènement. Les blocs de synténie sont de plus choisis maximaux, un bloc de synténie ne peut pas être composé d'un sous-ensemble des gènes ancestraux d'un autre bloc de synténie. Enfin, un bloc de synténie disparaît s'il était composé de deux gènes ancestraux et si, suite à une délétion, l'un d'entre eux est supprimé. Qu'implique la conservation d'un bloc de synténie dans les génomes modernes ? Les gènes ancestraux d'un bloc de synténie conservé seront nécessairement sur un même chromosome moderne. Si le bloc a été conservé jusqu'à deux espèces modernes, un chromosome de la première espèce contiendra l'ensemble des instances des gènes du bloc et un autre chromosome, de la deuxième espèce, contiendra toutes les autres instances également. Expliquons maintenant pourquoi notre définition d'un bloc de synténie rejoint la définition de Pevzner et Tesler.

Pour commencer, *un segment conservé d'au moins deux gènes ancestraux est un bloc de synténie sans gaps ($gapMax = 0$)*².

En effet, les seuls évènements susceptibles de perturber l'ordre des gènes ancestraux ou leurs orientations sont les réarrangements chromosomiques. Or, si un réarrangement chromosomique perturbe soit l'ordre, soit les orientations des gènes ancestraux du bloc, il y aura forcément une cassure à l'intérieur du bloc. De plus, les blocs de synténie de gaps $\leq gapMax$ avec n gaps non nuls sont constitués de $n + 1$ sous-blocs de synténie de gaps nuls espacés par les n gaps non nuls et au plus égaux à $gapMax$.

Nous en venons à notre deuxième conclusion, les *blocs de synténie de gaps $\leq gapMax$ avec n gaps non nuls sont constitués de $n + 1$ segments conservés espacés par ces n gaps non nuls*.

¹Là encore seuls les gènes ancestraux sont comptabilisés dans les gaps.

²La réciproque est également vraie, un bloc de synténie de gaps nuls est aussi un segment conservé. Autrement dit il n'y a pas de différence entre un bloc de synténie de gaps nuls et un segment conservé.

Nous appelons *micro-réarrangement* un réarrangement qui a eu lieu dans un gap de bloc de synténie. Il s’agit par exemple de l’inversion d’un segment de moins de *gapMax* gènes ancestraux. Nous nommons *micro-réarrangement artificiel* un micro-réarrangement dû à une erreur d’assemblage ou d’annotation. Et nous appelons *micro-segment-conservé* un segment conservé dans le gap d’un bloc de synténie. Un micro-segment conservé ne contient donc pas plus de *gapMax* gènes ancestraux. Ces dernières définitions dépendent bien entendu du contexte et de la valeur du *gapMax* considéré. Avec ces définitions, un bloc de synténie (défini de notre manière) peut être transformé en segment conservé si les micro-réarrangements qui ont eu lieu dans les gaps du bloc sont renversés de manière à restaurer la configuration ancestrale des micro-segments. Notre définition d’un bloc de synténie rejoint ainsi la définition intuitive de celle de Pevzner et Tesler.

Vestiges d’un bloc de synténie

Vestiges des blocs de synténie avec un filtrage idéal des génomes comparés Partant de cette définition d’un bloc de synténie, notre raisonnement en 5 étapes (section II.1), montrerait qu’un bloc de synténie de gaps $\leq gapMax$ se manifeste, dans une matrice d’homologies, par une diagonale consistante dont les gaps sont $\leq gapMax$. Pour cela il faudrait toutefois que les génomes soient filtrés avec le filtre *GènesAncestrauxDansLesDeuxGénomes* (section I.9.2) pour éviter, là encore, les gaps entre les homologies causés par les délétions de gènes ancestraux qui ont eu lieu spécifiquement dans une lignée.

Vestiges des blocs de synténie avec un filtrage pragmatique des génomes comparés Comme précédemment (section II.3.2), dans la pratique le filtrage *FamillesAncestralesDansLesDeuxGénomes* est utilisé, suite à quoi, les génomes filtrés sont comparés avec la matrice de packs d’homologies (MHP). Dans cette dernière les vestiges de blocs de synténie forment également des diagonales consistantes de gaps $\leq gapMax$, si nous négligeons les quelques cas lors desquels les duplications disperses élargissent les gaps au delà de *gapMax*.

Ce qui précède décrit comment se manifestent les vestiges d’un bloc de synténie dans des génomes modernes si l’assemblage, l’annotation et les familles sont parfaitement connues. Néanmoins, dans la réalité, des erreurs d’assemblages et d’annotations génèrent parfois des micro-réarrangements artificiels ainsi que des gaps artificiels qui rendent difficile l’identification directe des segments conservés. Heureusement grâce au concept de bloc de

synténie il est possible de s'abstraire de ces artéfacts et, dans un premier temps, de les considérer comme des gaps naturels, par exemple causés par des micro-réarrangements réels ou des duplications dispersées réelles. Une analyse ultérieure (section II.4.2) permettra de distinguer les micro-réarrangements naturels des micro-réarrangements artificiels et de distinguer également les gaps causés par de réelles inversions mono-géniques des gaps artificiels causés par des duplications dispersées ou des erreurs d'annotations.

La détection des vestiges de blocs de synténie

De nombreux logiciels ont été développés pour identifier ce qui semble être les vestiges des blocs de synténie ou ceux des segments conservés. Le logiciel le plus populaire est incontestablement GRIMM-synteny dont la meilleure explication se trouve dans [Bourque *et al.*, 2004]. Dans un travail précédent nous avons dressé un état de l'art de ces logiciels et nous avons développé le nôtre, PhylDiag [Lucas *et al.*, 2014]. Parmi les critères qui distinguent ces méthodes il y a : la prise en compte des in-paralogues, la gestion des duplications en tandem, la métrique de distance utilisée pour le « chaînage » des homologies, la comparaison simultanée de plusieurs génomes, l'heuristique utilisée (analyse d'une matrice d'homologie, optimisation, construction et édition d'un graphe), la prise en compte de l'orientation des gènes, la considération des micro-réarrangements, l'assurance que les blocs renvoyés ne se chevauchent pas, la méthode de validation statistique et le nombre de paramètres ainsi que leurs significations.

Depuis notre article nous avons découvert d'autres logiciels et d'autres travaux qui méritent d'être mentionnés.

Une revue à propos de la détection des points de cassures a été rédigée en 2008 [Lemaitre et Sagot, 2008] et un comparatif approfondi de nombreuses méthodes de détection de blocs de synténie a également été effectué [Lemaitre, 2008]. Dans un autre travail, une méthode visant à étudier finement les vestiges de flancs de cassures (section I.6.1) a été développée. Elle exploite les alignements de séquences nucléotidiques aux extrémités des blocs de synténie non-chevauchants [Lemaitre *et al.*, 2008][Lemaitre *et al.*, 2009]. L'ensemble de programmes, Cassis [Baudet *et al.*, 2010], utilisé pour ces études ne considère néanmoins que des homologies 1:1 (des BRHs en pratique) et, comme nous le mentionnions¹, cette limitation peut induire en erreur l'identification des gènes ancestraux dans les génomes modernes. De plus Cassis résout les conflits d'identification de blocs de synténie en les supprimant [Lemaitre, 2008]. Nous pensons que cette solution radicale empêche d'identifier tous les segments

¹voir le filtrage *BrhDansLesDeuxGénomes* (section I.9.2) et la difficulté d'identifier les gènes ancestraux parmi les in-paralogues (section I.5.3 et section I.1.8)

conservés inclus dans les gaps des blocs de synténie (le paramètre k de Cassis correspond à *gapMax*).

Le logiciel SynChro [Drillon *et al.*, 2014] vise à identifier les blocs de synténie d'eucaryotes. Synchro a l'avantage sur PhylDiag d'analyser directement les génomes en partant des séquences nucléotidiques des gènes en plus de leur ordre le long des chromosomes. Cependant la construction des blocs de synténie de Synchro repose sur une première étape qui ne considère là encore que des homologies 1:1 (ici aussi des BRHs), avec, nous le croyons, le même défaut que Cassis. Les in-paralogues sont ajoutés une fois que l'ossature des blocs de synténie a été fixée. Synchro identifie de plus des blocs de synténie qui se chevauchent et qui sont parfois inclus les uns dans les autres, ces blocs ne conviendront donc pas à de nombreuses études de scénarios de réarrangements. Les performances de SynChro ont été évaluées sur la base de données réelles en recoupant les résultats avec d'autres logiciels et nous pensons que les performances de Synchro gagneraient à être évaluées sur la base de simulations. Les outils de visualisation graphique de Synchro, sa rapidité (40 minutes pour comparer deux génomes de vertébrés) et sa capacité à définir des familles de gènes lui-même en font un outil polyvalent et directement utilisable sur des séquences protéiques. Cependant, malgré ses nombreux avantages, ce logiciel ne nous a pas semblé en mesure, dans son état actuel en tout cas, de détecter les segments conservés tels que nous les avons définis.

Le problème algorithmique Maximum Strip Recovery (MSR) défini par [Zheng *et al.*, 2007] a été largement étudié [Choi *et al.*, 2007] et il présente de nombreux points communs avec la recherche des segments conservés. Étant donné des génomes annotés par des gènes ancestraux, le but est ici d'extraire les segments de marqueurs les plus longs (*strips*) de manière à ce que les génomes comparés soient décomposables en segments qui ne se chevauchent pas. Néanmoins, la prise en compte de marqueurs répétés semble complexifier énormément cette approche [Bulteau *et al.*, 2013].

De nombreux logiciels d'alignement de génomes entiers commencent par une étape préalable d'identification des segments conservés, c'est le cas de MuMmer [Kurtz *et al.*, 2004], de SuperMap [Dubchak *et al.*, 2009], progressiveMauve [Darling *et al.*, 2010] et MUGSY [Angiuoli et Salzberg, 2011]. Parmi ceux-ci progressiveMauve est le plus connu¹. Les segments conservés de progressiveMauve sont appelés des LCBs [Darling *et al.*, 2004] et ils sont construits grâce à l'optimisation d'un coût lié aux points de cassures (« breakpoint analysis » [Blanchette *et al.*, 1997]). À la différence de progressiveMauve, MUGSY déduit ses segments conservés à partir d'un « graphe d'alignement »

¹de juin 2010 à janvier 2016 Google Scholar lui comptabilise plus de 962 citations

dans lequel les nœuds représentent des marqueurs conservés dans les différents génomes. Les pondérations des arêtes qui relient les nœuds du graphe sont égales au nombre d’adjacences des marqueurs (correspondants aux nœuds) dans les génomes. Ce graphe est analysé en trois étapes. Premièrement, les cycles du graphe sont divisés en segments acycliques. Deuxièmement, les extrémités de segments séparés par des micro-segments (de longueurs inférieures au paramètre L) sont fusionnées et les micro-segments sont retirés du graphe. Troisièmement, une procédure coupe le graphe pour s’assurer que les LCBs ne contiennent pas de gaps supérieurs au paramètre G . L’algorithme « Maximum Flow Min Cut » [Ford et Fulkerson, 1956] est utilisé pour découper un minimum d’arêtes tout en maximisant les longueurs et les poids cumulés des chemins linéaires dans le graphe après découpage. SuperMap est basé sur l’algorithme Shuffle-LAGAN. MuMmer est à l’origine basé sur le concept de Maximum Unique Matches (MUMs), des marqueurs orthologues 1:1, même si dans sa dernière version les marqueurs ne sont plus nécessairement des orthologues 1:1 [Kurtz *et al.*, 2004]. Il semblerait que le chaînage de ce dernier algorithme soit basé sur l’algorithme Longest Increasing Subsequence (LIS) [Delcher *et al.*, 1999]. Malgré leurs grandes popularités, ces logiciels d’alignement de génomes entiers semblent avoir pour but premier d’aligner le plus de séquences nucléotidiques possible et la qualité des segments conservés semble secondaire.

Une comparaison de plusieurs logiciels de détection de blocs de synténie a été effectuée en 2014 [Ghiurcuta et Moret, 2014]. Trois algorithmes de référence (DRIMM, i-ADHoRe 3.0 et Cyntenator) sont comparés sur un même jeu de 8 génomes de levures. Les auteurs soulignent que les algorithmes se basent sur des définitions différentes d’un « bloc de synténie » et par conséquent de substantielles spécificités se manifestent quand leurs blocs sont comparés sur la base de critères impartiaux. Les critères utilisés dans l’article représentent des caractéristiques d’un « bloc de synténie » consensuel : parmi ceux-ci il y a :

- 1) la densité en gènes de familles ancestrales (dont la racine est un gène du MRCA) entre les extrémités modernes des blocs de synténie (« $E(X)/E(X')$ »)
- 2) la similarité des blocs de synténie modernes homologues. Cette similarité est quantifiée par le nombre de relations d’homologie entre les gènes de blocs de synténie homologues (« Relaxed » and « Weighted » « Scoring »).

La comparaison place i-ADHoRe 3.0 en très bonne position :

- 1) ce dernier génère plus de blocs denses en gènes de familles ancestrales

- 2) et tous les gènes d'un bloc ont au moins un homologue dans un bloc homologue (« relaxed scoring » = 100%). Ce qui n'est pas le cas pour DRIMM et Cyntenator.

Même si la comparaison précédente est critiquable, elle nous conforte d'avoir choisi i-ADHoRe 3.0 comme algorithme de référence pour comparer les performances de notre algorithme PhylDiag.

Depuis la publication de notre article [Lucas *et al.*, 2014], nous avons élaboré de nombreuses améliorations qui ont permis d'accroître substantiellement la qualité de nos blocs de synténie. PhylDiag est de plus en mesure, aujourd'hui, d'analyser ces blocs de manière à obtenir de véritables segments conservés. Nous expliquons dans la prochaine section la procédure d'analyse que nous avons développée.

II.4 Affiner les blocs de synténie pour révéler les segments conservés

La plupart des algorithmes conçus pour recenser les segments conservés, y compris notre première version de PhylDiag [Lucas *et al.*, 2014], ne font trop souvent que de recenser des blocs de synténie qui se chevauchent, qui incluent des micro-réarrangements et qui sont à tort écourtés. Commençons par énumérer et décrire les enjeux de ces difficultés récurrentes qui compliquent l'identification précise des segments conservés :

- 1) Les nombreux clusters de gènes dupliqués en tandem (figure I.15) brouillent la conservation de l'ordre des gènes ancestraux.
- 2) L'absence de synténie conservée dans les segments conservés monogénique (section II.3.2) empêche généralement de les identifier. Dans les études précédentes, soit les blocs de synténie composés d'au moins deux gènes sont étudiés, soit les segments résultant d'inversions monogéniques sont étudiés, mais nous ne connaissons pas d'études où les deux sont étudiés simultanément; alors que toutes les tailles de segments conservés sont intéressantes pour les scénarios de réarrangements.
- 3) Il existe de nombreux véritables micro-réarrangements qui ne devraient pas être négligés. Les micro-segments conservés qui en résultent ont par exemple été au cœur du débat entre le RBM et le FBM [Peng *et al.*, 2006][Sankoff, 2006].

- 4) Les blocs de synténie se chevauchent souvent alors que les segments conservés ne se chevauchent pas. Les chevauchements doivent donc être résolus pour que les blocs puissent être utilisés dans le cadre d'études sur les scénarios de réarrangements [Tesler, 2002][Ma *et al.*, 2008a]. Car celles-ci utilisent principalement des segments conservés (ou des blocs de synténie) non chevauchant comme unités de base pour décrire les scénarios qui transforment un génome en un autre.
- 5) Des ambiguïtés sur l'ordre des gènes ancestraux dans les génomes modernes écourtent parfois prématurément les blocs de synténie.

La première difficulté est dénouée par la réduction des clusters de duplications en tandem à un unique gène qui les représente. Décrivons des étapes d'affinage qui résolvent les 4 difficultés restantes. La première étape identifie les segments conservés mono-géniques. La deuxième étape identifie les micro-réarrangements correspondant à des points de cassures dans les gaps des blocs de synténie. La troisième étape de l'affinage résout les chevauchements en minimisant les éditions des blocs de synténie. Pour finir, la dernière étape fusionne des blocs de synténie précédemment édités pour recouvrir l'intégralité des segments conservés à partir de blocs écourtés à tort. Toutes ces étapes seront évaluées en se basant sur des segments conservés simulés et nous donnerons les raisons qui expliquent pourquoi, avec notre modélisation des génomes, il reste des faux positifs et des faux négatifs. Toutes les étapes de pré-traitement et de post-traitement peuvent être appliquées en amont et en aval de n'importe quel logiciel d'inférence de blocs de synténie. Nous développons tout cela dans les sous-sections qui suivent.

II.4.1 Prétraitement

La réduction des clusters de gènes dupliqués en tandem

Cette étape de prétraitement ¹ réécrit les chromosomes de manière à ne conserver qu'une seule localisation représentative du gène à l'origine de chaque cluster. Nous l'avons décrite précédemment (section I.7.3) et nous l'illustrons ici dans avec la figure II.8B. Cette méthode remédie aux ruptures de colinéarité causées par les duplications segmentales en tandem dont les segments dupliqués ont des longueurs $\leq \text{tandemGapMax}$.

Comme nous le disions, cette étape de réécriture n'est pas sans défaut. Le gène à l'origine du cluster est positionné arbitrairement là où se situe le

¹déjà mise en place dans ADHoRe [Vandepoele *et al.*, 2002], MCScanX [Wang *et al.*, 2012] et SynChro [Drillon *et al.*, 2014]

premier gène du cluster, et par conséquent, le résultat de la réécriture peut varier selon l'orientation de référence du chromosome, qui ordonne les gènes dans un sens ou dans l'autre, voir figure II.9.

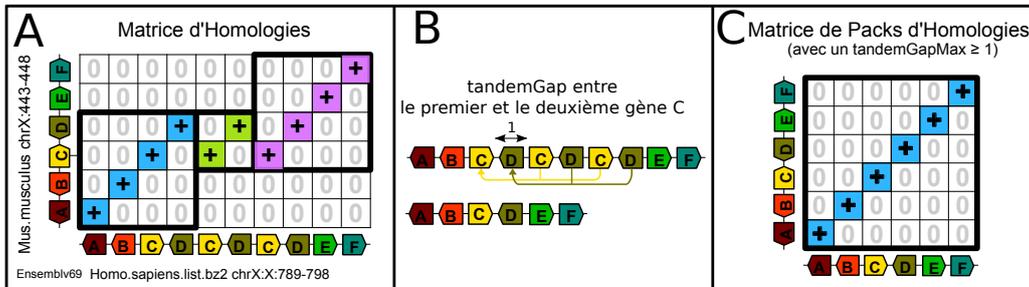


Figure II.8 – Réduction des clusters de gènes dupliqués en tandem pour remédier aux ruptures de colinéarité causées par les duplications segmentales. Le **panel A** représente une matrice d'homologies de la comparaison d'un segment de chromosome X humain avec un segment du chromosome X de la souris à partir de la version 69 d'Ensembl. Deux duplications segmentales des gènes C et D compliquent l'identification de la conservation de l'ordre des gènes ancestraux. Le **panel B** détaille le processus de réécriture du chromosome humain en réduisant les clusters de gènes dupliqués en tandem. Si $tandemGapMax \geq 1$ il y a deux clusters, un de 3 gènes de la famille C et un autre cluster de 3 gènes de la famille D. Les trois gènes C forment un cluster car moins de $tandemGapMax$ autres gènes ancestraux séparent chaque paire de gènes C voisins. Tous les gènes d'un cluster sont réduits à un unique gène positionné là où se trouve le premier gène du cluster, comme indiqué par la flèche jaune. Le cluster de gènes de la famille D subit le même traitement. Quand les gènes dupliqués en tandem sont regroupés, la conservation de l'ordre des gènes ancestraux se manifeste alors par une diagonale ininterrompue dans la matrice d'homologies, **panel C**. Dans le panel A les rectangles noirs représentent les contours des blocs de synténie. Dans le panel C, les rectangles noirs représentent cette fois-ci les contours des segments conservés.

II.4.2 Post-traitement

Dans ce qui suit, nous écrirons qu'un bloc de synténie A est *complètement logé* dans un bloc de synténie B si les gènes de A sont situés dans les gaps de B dans les deux génomes modernes (figure II.10A). Nous écrirons que A est *partiellement logé* dans B si les gènes de A sont localisés dans les gaps de B dans seulement un génome moderne, voir la figure II.10C).

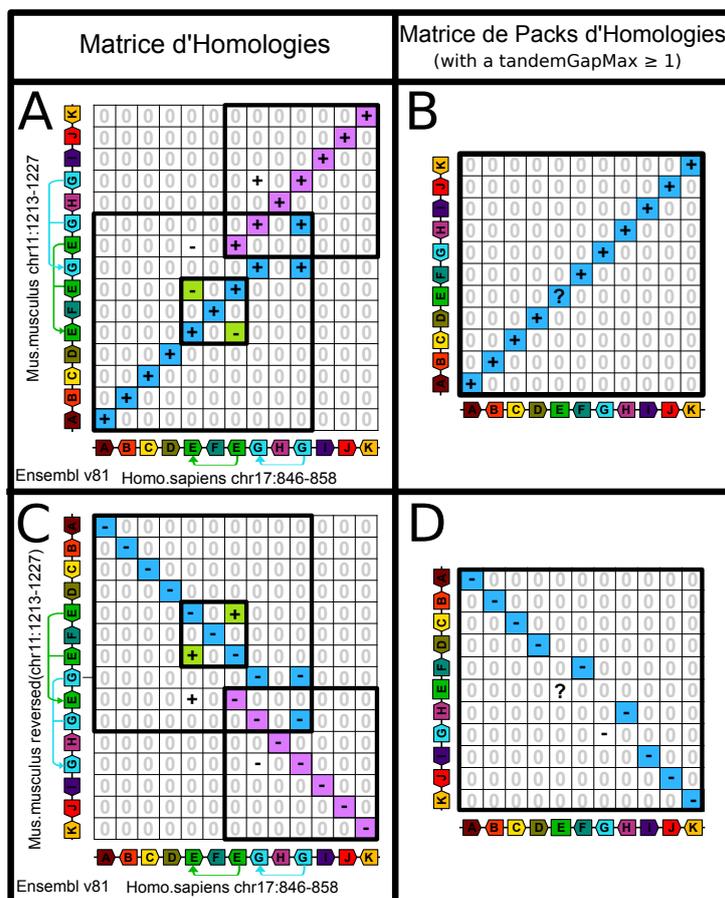


Figure II.9 – L'ordre de référence des chromosomes peut influencer la réécriture des chromosomes lors de la réduction des clusters de gènes dupliqués en tandem. Le **panel A** montre la matrice d'homologies de la comparaison d'un segment du chromosome humain 17 avec un segment du chromosome 11 de la souris, à partir de la version 81 d'Ensembl. Pour un $tandemGapMax \geq 1$, le segment humain contient deux clusters de gènes dupliqués en tandem, un avec deux gènes E et un avec deux gènes G, alors que le segment de la souris contient aussi deux clusters des familles E et G dupliqués, mais cette fois-ci en trois copies. La matrice d'homologies après avoir réduit les clusters est montrée dans le **panel B**. Le **panel C** montre les mêmes données que dans le panel A, sauf que cette fois-ci le segment de la souris est représenté inversé sur l'axe des ordonnées, de sorte que les gènes de la souris sont ordonnés dans l'ordre inverse. Au final le contenu en gènes du bloc dans le panel A (ABCDEFGHIJK) diffère de celui du panel D (ABCDFHIJK). Cependant dans les deux cas les blocs de synténie ont les mêmes extrémités : l'extrémité 5' du gène A et l'extrémité 5' du gène K.

Identification des segments conservés mono-géniques

L'identification des segments conservés mono-géniques commence par trouver les gènes ancestraux isolés et complètement logés dans les gaps de blocs de synténie. Dans les matrices d'homologies ces gènes génèrent la plupart du temps des cellules non-nulles incluses dans les gaps d'une diagonale consistante et les signes de chacune de leurs homologies sont opposés aux signes des homologies de la diagonale dans laquelle ils sont inclus (figures II.10A et II.10B). Des segments conservés mono-géniques supplémentaires sont identifiés aux extrémités des blocs de synténie. Dans une matrice d'homologie, un gène ancestral d'un de ces segments se manifestera par une cellule isolée placée le long des bordures de diagonales et/ou sur les bords de la matrice, comme c'est le cas avec le gène J de la figure II.7 ou avec le gène I dans les figures II.10A et II.10B.

Identification des micro-réarrangements dans les gaps des blocs de synténie

Les micro-réarrangements se manifestent par des blocs de synténie complètement ou partiellement logés dans les gaps d'autres blocs de synténie. Dès qu'un bloc est détecté dans un gap, le bloc hôte est fragmenté de manière à ce que ses fragments correspondent aux sous-blocs de part et d'autre du gap (figures II.10C et II.10D). L'identification des micro-réarrangements a pour conséquence d'augmenter le nombre total de blocs, de diminuer les longueurs des blocs et également d'identifier de nouvelles localisations de points de cassure.

Résoudre les chevauchements

Nous décrivons maintenant un traitement en deux étapes pour résoudre le problème des chevauchements de blocs de synténie. Premièrement, pour les chevauchements courts entre deux blocs de synténie, dans lesquels moins de *truncationMax* (un paramètre fixé par l'utilisateur) gènes se chevauchent, l'extrémité du bloc contenant le moins d'homologies est tronquée (figures II.11A et II.11B). Deuxièmement, pour les chevauchements restants, la procédure « t-opt » [Bafna *et al.*, 1996] sélectionne un sous-ensemble de blocs de synténie non chevauchant de manière à minimiser le nombre d'homologies jetées. Ici les « weighted rectangles » de [Bafna *et al.*, 1996] sont les rectangles encadrés en noir et pondérés par le nombre d'homologies qu'ils contiennent. Cette étape du post-traitement est très similaire à [Tang *et al.*, 2011].

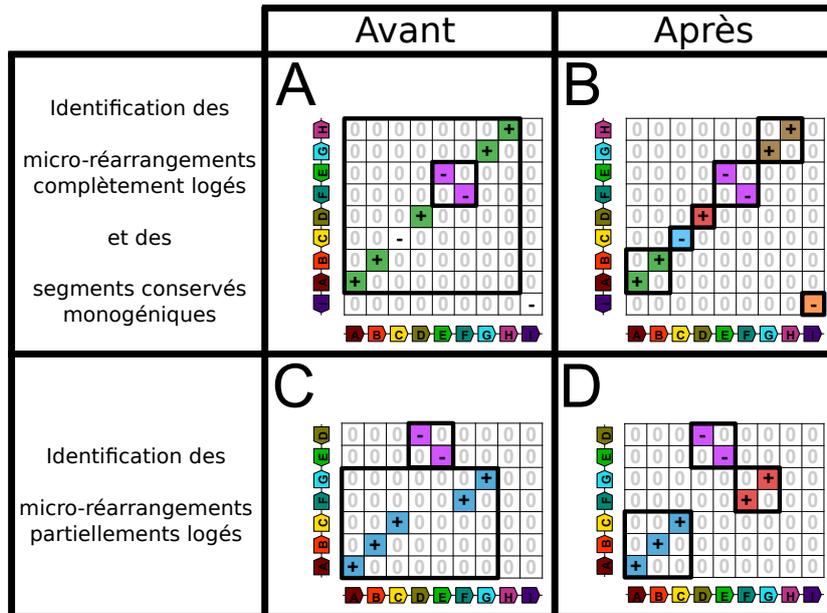


Figure II.10 – Identification des micro-réarrangements et des segments conservés mono-géniques. Dans le **panel A**, la matrice d'homologies révèle deux blocs de synténie. Le premier (violet) est complètement logé dans le gap d'un autre (vert). De plus, le long bloc de synténie vert héberge une homologie isolée (signe « - » au milieu) et est adjacent à une autre homologie solitaire (le signe « - » en bas à droite). Dans le **panel B**, après post-traitement, les deux blocs de synténie sont maintenant divisés en quatre segments conservés dont un est mono-génique. De plus les deux homologies isolées sont identifiées comme des segments conservés mono-géniques. Dans le **panel C**, il y a un bloc de synténie (violet) partiellement logé dans un autre bloc (bleu). Dans le **panel D**, l'identification du micro-réarrangement correspondant aboutit à trois segments conservés distincts.

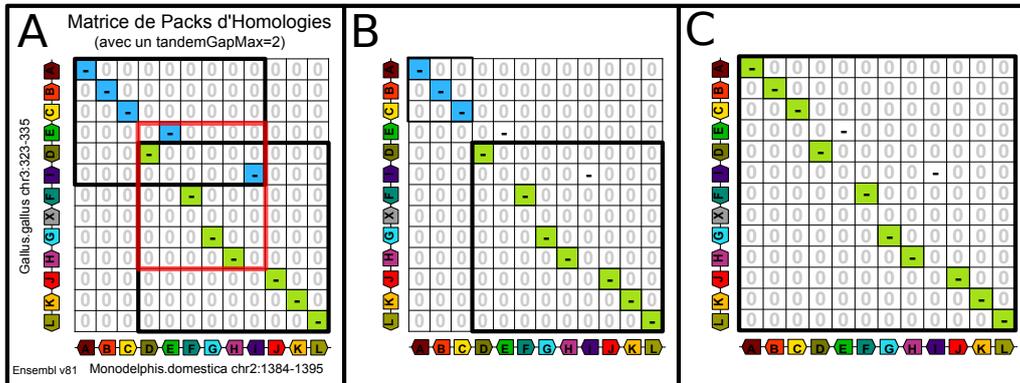


Figure II.11 – Résolution des chevauchement et des ruptures précoces de colinéarité. La matrice d’homologies dans le **panel A**, correspond à la comparaison d’un segment de chromosome de l’opossum et d’un segment de chromosome du poulet de la version 81 d’Ensembl. La matrice d’homologies du **panel B** représente les résultats de la troncation utilisée pour résoudre les chevauchements. Dans le **panel C**, la fusion de l’extrémité tronquée résout la rupture précoce de collinéarité précédente (délimitée par le rectangle rouge dans le panel A).

Corriger les interruptions précoces de blocs de synténie en fusionnant les extrémités qui ont été préalablement tronquées

Une ambiguïté sur l’ordre ancestral des gènes génère parfois deux blocs de synténie incomplets, un de chaque côté de la région ambiguë, voir figure II.11. Après la troncation expliquée plus haut, la dernière étape du post-traitement fusionne simplement les extrémités des blocs tronqués et les véritables extrémités de blocs de synténie sont retrouvées (figures II.11B et II.11C).

II.4.3 Évaluation de nos segments conservés

Analyse des améliorations en sensibilité et spécificité

Le prétraitement qui réduit les clusters de gènes dupliqués en tandem et les quatre étapes de post-traitement ont été implémentés dans PhylDiag [Lucas *et al.*, 2014]. Leur évaluation, basée sur des simulations, a révélé une amélioration substantielle de la détection des extrémités de segments conservés.

En utilisant une simulation réaliste de l’évolution de l’ordre des gènes (chapitre III) et les véritables segments conservés durant les simulations nous avons calculé la sensibilité et la spécificité de PhylDiag à retrouver les véritables

extrémités de segments conservés. Nous avons également calculé la sensibilité et la spécificité des « baseclusters » d'i-ADHoRe 3.0 [Proost *et al.*, 2012] par rapport aux segments simulés. Lors de l'exécution d'ADHoRe les paramètres ont été fixés de la manière suivante : `anchor_points=3`, `prob_cutoff=0.001` et `tandem_gap=5`. Les paramètres `gap_size` et `cluster_gap` sont variables et sont ici égaux à notre paramètre *gapMax*. Différentes valeurs de la probabilité seuil (`proba_cutoff`) ont été testées et seule a été gardée la valeur qui maximise les résultats d'ADHoRe. Pour ne pas désavantager ADHoRe, les extrémités des segments conservés sont les noms des gènes aux extrémités et les orientations de ces gènes ne sont pas comptabilisées.

Les résultats de PhylDiag et d'ADHoRe sont illustrés par les graphes de la figure II.12. Avec PhylDiag, l'identification des segments conservés mono-géniques et des micro-réarrangements génère une augmentation de 20% en sensibilité pour un *gapMax*=1, de 65% jusqu'à 85%. Cependant dans un même temps la spécificité décroît de 5%. Comme expliqué dans les figures II.13 et II.14, la perte de spécificité semble principalement imputable aux nombreuses duplications de gènes en tandem ainsi qu'aux nombreuses délétions qui ont eu lieu dans les lignées menant à la souris ou au poulet et qui ont élevées le nombre de faux positifs. Pour des espèces plus proches comme l'humain et la souris, la sensibilité atteint 95% et la spécificité reste autour de 98%, étant donné que moins de duplications et délétions de gènes ont eu lieu (données non présentées). La figure II.12 montre également qu'un *gapMax* de 1 est suffisant pour que la détection soit correcte quand il s'agit de retrouver des segments conservés à partir de données simulées. Dans la réalité, un plus grand *gapMax* est souvent nécessaire pour une détection optimale des segments conservés à cause des larges gaps causés par les micro-réarrangements artificiels et les erreurs d'annotations. Néanmoins le *gapMax* peut être augmenté sans crainte car l'identification des micro-réarrangements (l'option `imr`) stabilise la sensibilité même pour de très grands *gapMax*.

La même évaluation a été faite en comptabilisant les nombres d'adjacences de gènes dans les segments conservés et là encore PhylDiag obtenait de meilleurs résultats qu'ADHoRe¹.

Explication des derniers faux positifs et faux négatifs dans l'identification des extrémités de segments conservés

Grâce aux segments conservés simulés nous pouvons enquêter sur les causes des faux positifs et faux négatifs lors de la détection des extrémités de seg-

¹Pour plus de détails sur cette dernière comparaison le lecteur pourra se référer à notre précédent travail [Lucas *et al.*, 2014] même si les performances de PhylDiag ont largement évoluées depuis, grâce aux étapes de pré- et post-traitement que nous avons détaillées.

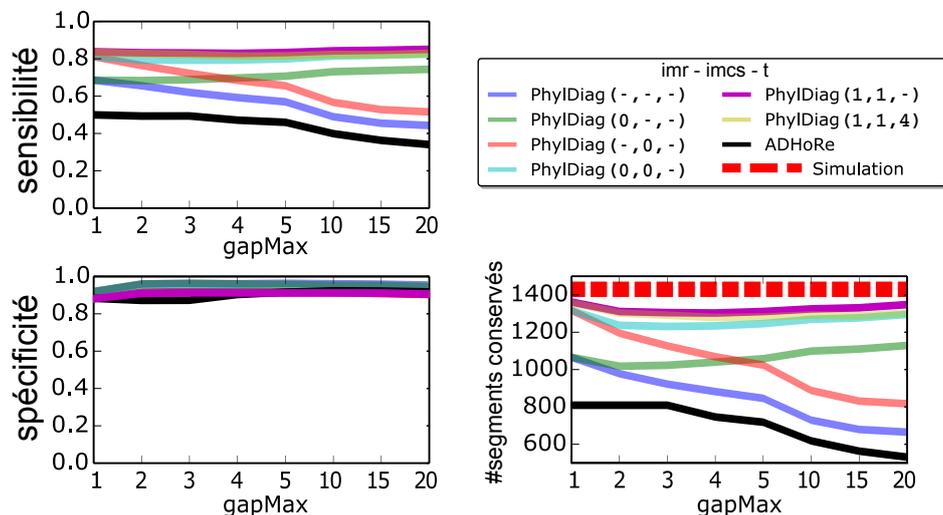


Figure II.12 – Analyse de la sensibilité et la spécificité basée sur des segments conservés simulés de deux espèces distantes, la souris et le poulet qui ont divergé il y a environ 325 Millions d’années. L’analyse est basée sur une simulation réaliste (chapitre III) de l’évolution de l’ordre des gènes, qui tend à reproduire les données de la version 81 Ensembl. Pour chaque paramétrisation, la sensibilité (en haut), la spécificité (au milieu) et le nombre de segments conservés (en bas) sont évalués pour différentes valeurs du paramètre *gapMax*. Les étapes de post-traitement décrites dans les paragraphes précédents sont Identify Micro-Rearrangements (imr), Identify Mono-genic Conserved Segments (imcs) et Truncation (t). Un signe «-» signifie que l’option est inactive et un entier, même 0, signifie que l’option est active. Pour l’option imr, la valeur entière spécifie le gap maximal autorisé entre : l’extrémité d’un micro-segment et l’homologie du bloc de synténie dans le gap duquel il est inclus. Pour l’option imcs, la valeur entière spécifie le gap maximal autorisé autour des cadres de blocs de synténie où peuvent être identifiés les segments conservés mono-géniques. Pour l’option t, la valeur entière correspond au paramètre *truncationMax*. Tronquer et résoudre les chevauchements restants avec *truncationMax* = 4 ne décroît pas perceptiblement la sensibilité et la spécificité alors que cela permet de s’assurer que les segments conservés ne se chevauchent pas. Sur le même graphique nous avons également reporté les résultats de l’algorithme i-ADHoRe 3.0.

ments conservés, qui empêchent la sensibilité et la spécificité d’atteindre 100% (figure II.12). Nous trouvons que les erreurs proviennent de deux types de scénarios. Le premier implique les extrémités de segments conservés mono-géniques, qui apparaissent comme des inversions mono-géniques mais qui sont en réalité le résultat d’une duplication en tandem inversée [Ma *et al.*, 2008a] suivie d’une délétion de gène (figure II.13). Les deux évolutions sont impossibles à distinguer à partir de notre modélisation des génomes modernes. Notre méthode d’identification des segments conservés mono-géniques sélectionnera toujours le scénario correspondant à une inversion mono-génique (figure II.13A) au lieu de la séquence d’évènements duplication-délétion (figure II.13B), qui est une évolution moins parcimonieuse et sans réarrangement. D’où une augmentation de la sensibilité quand une inversion mono-génique a eu lieu et une baisse de spécificité quand le scénario duplication-délétion a eu lieu.

Une seconde cause de faux positifs et de faux négatifs sont les points de cassures qui tombent entre deux gènes dupliqués en tandem (figure II.14).

Savoir quel gène d’un cluster d’une famille génique est le gène ancestral permettrait de résoudre les erreurs illustrées dans les figures II.13 et II.14 mais puisque ces gènes jouent des rôles symétriques dans nos données il est impossible de les distinguer. Les évolutions équivalentes des figures II.13 et II.14 génèrent donc des extrémités de segments conservés qui ne peuvent pas être identifiées précisément avec nos données ce qui empêche la sensibilité et la spécificité d’atteindre 100%.

II.4.4 Conclusion à propos de l’identification des segments conservés

Dans le domaine de la génomique comparative à l’échelle des génomes entiers, jusqu’à récemment, les petits réarrangements étaient écartés car ils ne pouvaient pas être dissociés des erreurs d’assemblage et d’annotation [Pevzner et Tesler, 2003a]. Cependant les points de cassures correspondant aux véritables petites inversions peuvent remodeler de fond en comble un ensemble de segments conservés qui auraient été identifiés sans les considérer. Par conséquent, ce que l’on croyait être un long segment conservé peut se révéler être en réalité plusieurs segments conservés de tailles modestes quand les micro-inversions sont prises en compte. Avec l’amélioration de la précision des assemblages et des annotations de génomes, ces petits réarrangements peuvent maintenant être étudiés. Dans ce contexte notre méthode est la première à avoir identifié des segments conservés allant jusqu’à des segments conservés mono-géniques; elle trouve les segments conservés avec un degré

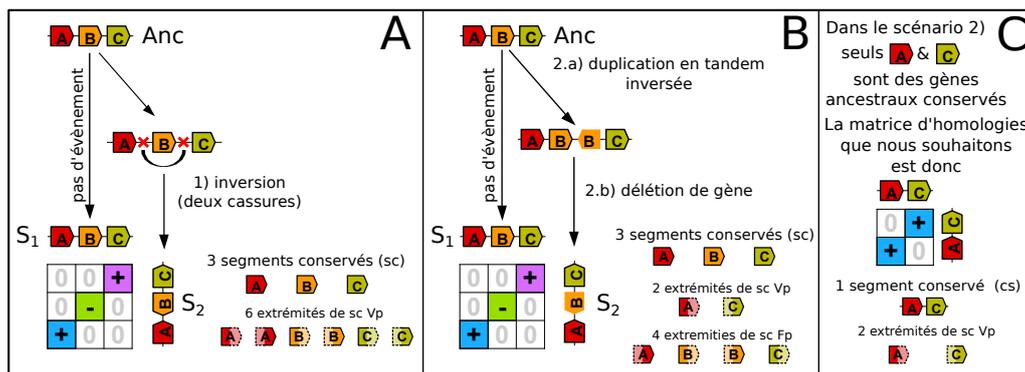


Figure II.13 – Deux scénarios, l'un avec des points de cassures et l'autre sans, qui ne peuvent pas être distingués l'un de l'autre avec nos données. Dans le **panel A**, le chromosome d'un ancêtre *Anc* évolue jusqu'à deux espèces modernes, S_1 et S_2 . Aucun évènement n'a lieu de *Anc* à S_1 mais le gène B est inversé entre *Anc* et S_2 , créant deux points de cassures. En fin d'évolution il y a donc trois segments conservés qui sont facilement identifiables dans la matrice d'homologies. Dans le **panel B**, de *Anc* à S_2 le gène B est dupliqué en tandem et sa copie est insérée avec une orientation inverse puis le gène ancestral est supprimé. Dans le **panel B**, la copie du gène B est considérée à tort comme le gène ancestral B, ce pourquoi la matrice d'homologies est identique à la précédente. Par conséquent il est impossible de distinguer les deux évolutions. Dans ce dernier cas, 3 segments conservés sont inférés alors qu'un seul segment conservé devrait être identifié car il n'y a pas eu de point de cassure. Le **panel C**, montre la matrice d'homologies correcte quand la copie du gène ancestral est négligée et qu'elle n'est pas confondue avec le gène ancestral.

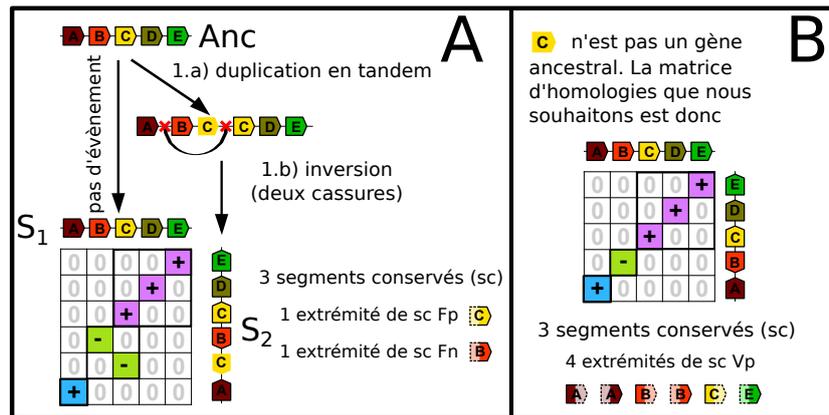


Figure II.14 – Un point de cassure entre deux gènes dupliqués en tandem génère un faux positif et un faux négatif lors de l'identification des extrémités de segments conservés. Dans le **panel A**, un chromosome de l'espèce ancestrale S_a évolue jusqu'à deux espèces modernes, S_1 et S_2 . Le long de la lignée de S_a à S_1 le chromosome est parfaitement conservé, et le long de la lignée de S_a à S_2 le gène C est dupliqué en tandem. Pour la différencier, la copie non-ancestrale de C n'est pas entourée d'une ligne noire. Par la suite, une inversion a lieu et un de ses deux points de cassures tombe entre les gènes dupliqués en tandem. À partir des génomes modernes de S_1 et S_2 , il est impossible de savoir lequel des deux gènes de la famille C est ancestral, par conséquent les deux sont considérés comme de possibles gènes ancestraux. L'analyse de la matrice d'homologie nous incite à identifier 3 segments conservés, comme attendu, mais ici il y a une extrémité de segment conservé qui est un faux positif et une qui est un faux négatif. Le **panel B** décrit la matrice d'homologies et les segments conservés obtenus quand la copie non-ancestrale du gène C est supprimée.

de précision qui profite pleinement de la connaissance des orientations de transcription des gènes (figure II.10). De plus elle résout les cas complexes de collinéarité qui impliquent les vestiges de nombreuses duplications en tandem (figures II.8 et II.9) ainsi que les ruptures précoces de colinéarité (figure II.11). Néanmoins, notre détection des segments conservés fait parfois des erreurs et elle ne peut pas être parfaite avec notre modélisation des génomes, car des scénarios d'évolution sans point de cassure produisent les mêmes génomes modernes que des scénarios d'évolution avec points de cassures, voir les figures II.13 et II.14.

Nous avons également montré qu'un prétraitement des génomes modernes et des post-traitements modulaires d'affinage des blocs de synténie avec peu de paramètres, principalement *tandemGapMax* et *troncationMax*, représentent une alternative intuitive et cependant rigoureuse à des heuristiques d'édition de graphe [Paten *et al.*, 2008][Pham et Pevzner, 2010] qui peuvent être plus difficiles à configurer.

Chapitre III

Simuler l'évolution de l'ordre des gènes de vertébrés

If you don't know history, then you don't know anything.
You are a leaf that doesn't know it is part of a tree.

MICHAEL CRICHTON, auteur de *Jurassic Park*

Avant de détailler notre simulateur, nous justifions l'usage de simulations pour reproduire l'évolution des génomes sur des échelles de temps de plusieurs millions d'années.

Un physicien peut facilement reproduire l'expérience des fentes d'Young pour étudier les interférences lumineuses car une fois que le laser, les fentes et l'écran sont installés sur le banc optique, l'expérience est quasi-instantanée. Elle se fait à la vitesse de la lumière. La durée d'une expérience pour constater l'évolution des génomes est bien plus lente. Une telle expérience a été initiée en 1988 par le chercheur Richard Lenski et elle est toujours en cours. L'expérience suit l'évolution de 12 lignées de *Escherichia coli*. Grâce aux données récoltées en 2011 sur plus de 40 000 générations, il a été estimé qu'à chaque nouvelle génération il y a en moyenne dix mutations ponctuelles tous les milliards de paires de bases, 8.9×10^{-11} mutations/bp/génération pour être plus précis [Wielgoss *et al.*, 2011]. Étant donné que le temps de génération des cellules d'*E. coli* est d'environ 7 générations par jours, cela fait à peu de choses près 70 mutations ponctuelles tous les gigabases (Gb) chaque jour. Il faut donc être patient pour étudier ces mutations et encore plus patient pour étudier les réarrangements chromosomiques qui sont encore moins fréquents. Les expériences d'évolution sur le long terme comme celle-ci, Long-Term Experimental Evolutions (LTEEs), sont rares. Par conséquent l'étude de l'évolution des génomes se fait aussi via le séquençage de génomes ancestraux.

Par exemple en 2009, une étude a séquencé des génomes mitochondriaux issus de cadavres humains d'Europe centrale, morts entre -13 400 et -2 300 ans [Bramanti *et al.*, 2009]. La comparaison des séquences entre elles ainsi que leur comparaison avec les séquences mitochondriales de populations modernes locales ont permis de mieux comprendre la transition de la vie de chasseur-cueilleur à celle de fermiers sédentaires dans cette région. De manière analogue, il est tentant d'essayer de séquencer les génomes de fossiles conservés dans des strates sédimentaires successives pour pouvoir reconstruire, génome après génome, l'histoire évolutive. L'assemblage de photographies sur une pellicule permet de reconstruire le film de l'évènement qui s'est déroulé devant la caméra, nous aimerions faire de même avec l'histoire des génomes ancestraux séquencés. Malheureusement, en raison de la dégradation de l'ADN au cours du temps, il semble qu'il ne soit pas possible d'obtenir des séquences génomiques anciennes de plus d'un million d'année [Dabney *et al.*, 2013]. À l'heure actuelle la séquence la plus ancienne provient d'un cheval du Pléistocène qui vivait il y a environ 700 000 ans [Orlando *et al.*, 2011]. Les évolutions étudiées par le séquençage de génomes ancestraux sont donc *a priori* limitées à des durées inférieures au million d'années, ce qui est court par rapport aux centaines de millions d'années qui caractérisent l'évolution des génomes d'amniotes que nous souhaitons étudier ici.

Nous sommes donc dans l'incapacité de confronter nos estimations de l'évolution du génome d'Amniota (le MRCA de la souris et du poulet et plus généralement l'ancêtre des amniotes) à ce qui s'est réellement passé. Ainsi, il ne nous est pas possible de vérifier la véracité de nos segments conservés car nous n'avons pas de données directes concernant les génomes des ancêtres le long des lignées qui relient le génome d'Amniota jusqu'aux espèces modernes.

Pour palier à ce manque une solution consiste à simuler l'évolution réelle. Cela revient dans notre cas à créer un génome artificiel de l'ancêtre d'Amniota et à le faire évoluer *in silico* de la manière la plus réaliste possible. Durant les évolutions informatiques les génomes intermédiaires sont sauvegardés et ils se substituent aux génomes ancestraux dont nous manquons. Dans le chapitre précédent, notre méthode d'inférence a été évaluée sur la base de ces évolutions simulées de manière réaliste (section II.4.3) et nous en avons conclu que la même méthode devait également être correcte pour des inférences sur l'histoire évolutive réelle.

En plus de garder en mémoire les génomes ancestraux, simuler l'évolution des génomes présente de nombreux avantages. Le simulateur concrétise un modèle d'évolution et l'intégration des différentes composantes de ce modèle permet de mieux comprendre les interactions entre des phénomènes qui sont trop souvent pensés de manières autonomes. Par exemple, les évènements géniques et les réarrangements sont souvent étudiés séparément, alors que notre

simulateur nous a montré que, dans les faits, les évènements géniques influent sur l'estimation du nombre de réarrangements. Nous verrons également que, couplé à une évaluation quantitative du réalisme, le simulateur devient un outil de choix pour tester différents modèles d'évolution ou différents paramètres. Pour ce dernier point il suffit en effet d'appliquer une procédure d'optimisation, par exemple la méthode Approximate Bayesian Computing (ABC) [Wegmann *et al.*, 2010], qui effectue de nombreuses simulations avec différents modèles et différents paramètres. Le réalisme de chaque simulation est ensuite quantifié et au bout du compte les paramètres et les modèles aboutissant aux simulations les plus réalistes sont sélectionnés. Pour terminer, le simulateur agglomère les connaissances actuelles de l'évolution et, par conséquent, s'il est utilisé comme un modèle nul, il permettra de distinguer le signal d'un processus évolutif inattendu et donc intéressant.

III.1 État de l'art des simulateurs de l'évolution d'un génome

De nombreux simulateurs existent déjà. Par exemple eVolver [Brylinski, 2013] est un simulateur de séquences codantes et GenPop [Rousset, 2008] est un simulateur de génétique des populations. Les évolutions informatiques d'organismes « digitaux » [Lenski *et al.*, 2003] peuvent aussi être considérées comme des simulations de l'évolution. Cependant dans ce dernier cas les auteurs cherchent surtout à démontrer qu'un programme informatique peut reproduire les caractéristiques fondamentales d'une entité évolutive. De plus, même si ces organismes digitaux ont permis d'acquérir une meilleure connaissance de l'évolution [Wilke *et al.*, 2001] ils ne sont pas adaptés pour reproduire l'évolution moléculaire du génome d'un organisme réel.

Nous effectuons maintenant un inventaire des principaux simulateurs qui incorporent les réarrangements chromosomiques.

III.1.1 Simulateurs *ad hoc*

De nombreux simulateurs *ad hoc* ont été développés pour valider des méthodes d'inférence de réarrangements ou des méthodes de reconstruction de génomes ancestraux. C'est le cas du simulateur de l'article de ChromEvol [Mayrose *et al.*, 2010] qui simule la variation du nombre de chromosomes à travers une succession de fissions, de fusions et d'évènements de polyploidisations. C'est également le cas du simulateur développé par Zhao et Bourque en 2009 [Zhao et Bourque, 2009] qui simule cette fois-ci l'évolution des réarrangements

chromosomiques de Boreoeutheriens tels que l'humain, la souris et le chien. Les évènements pris en compte sont les inversions, les translocations réciproques, les transpositions ainsi que des fissions/fusions de chromosomes. Estimées à partir de données réelles, les proportions relatives de ces évènements sont fixées à 10:2:2:0.1 respectivement. Néanmoins ce simulateur ne prend pas en compte les évènements géniques et aucune information ne nous est donnée sur la manière dont les chromosomes et les segments de chromosomes sont sélectionnés pour être réarrangés. Les auteurs de MSOAR 2.0 [Shi *et al.*, 2011] ont validé leur algorithme de détection de gènes orthologues en simulant des génomes à un seul chromosome qui subissent des évènements de duplications de gènes, des naissances de gènes, des délétions de gènes et des inversions de gènes et de segments de chromosomes. Ces 4 évènements sont effectués dans les proportions 4:1:1:4 respectivement. Durant l'évolution de ce génome, les séquences évoluent par des mutations ponctuelles effectuées grâce à l'utilisation du logiciel eVolver [Brylinski, 2013]. Dans leur modélisation il y a autant de duplications disperses que de duplications en tandem.

Jian Ma a développé trois logiciels de reconstruction de génomes ancestraux : inferCAR [Ma *et al.*, 2006], DUPCAR [Ma *et al.*, 2008b] et le « infinite site model » [Ma *et al.*, 2008a]. Dans les données additionnelles de ses algorithmes l'auteur décrit le simulateur qu'il a progressivement mis à jour pour effectuer les validations de ses méthodes. Au final son simulateur considère les génomes à la même échelle d'étude que la nôtre. Les génomes sont composés de plusieurs chromosomes et ces derniers sont des listes ordonnées de gènes orientés. Il a estimé sur la base des données modernes, sans néanmoins expliquer comment, que le génome d'Amniota était composé de 25 chromosomes et que les réarrangements arrivent généralement dans les proportions suivantes : 90% d'inversions, 5% de translocations, 3.75% de fusions et 1.25% de fissions. Quand un réarrangement a lieu les points de cassures sont choisis avec une distribution uniforme dans l'ensemble des régions précédemment utilisées ou non, selon le taux de réutilisation des points de cassures spécifié par l'utilisateur. Ce dernier taux est défini comme le pourcentage d'intergènes ancestraux qui ont été cassés plus d'une fois. Les longueurs des segments inversés sont tirées dans une distribution gamma (Γ) de paramètre de forme $\alpha = 0.7$ et d'échelle $\theta = 500$ gènes. La distribution est tronquée et re-normalisée pour que les inversions n'excèdent pas 50 gènes. Des ajustements par rapport à ces paramètres ont été effectués sur chaque branche de l'arbre des espèces. Les évènements géniques d'insertions de nouveau gènes, de duplications et de délétions sont tous les trois pris en compte, mais il n'y a pas de distinction entre duplication en tandem et duplication dispersée.

Un autre simulateur a été développé par [Gagnon *et al.*, 2012]. Il est néanmoins limité : son implémentation informatique est imparfaite et en

pratique les génomes simulés ne font pas plus de 200 gènes, car au-delà l'exécution prend beaucoup de temps. Malgré ce défaut ce simulateur considère des génomes multi-chromosomiques, les réarrangements majeurs (inversions, translocations, fissions et fusions) ainsi que les pertes de gènes. L'atout de ce simulateur par rapport aux autres est la possibilité de simuler des duplications complètes de génomes, WGDs. Après la duplication complète du génome, entre un huitième et un quart des gènes post-duplication sont supprimés entre l'évènement de WGD et le prochain nœud de spéciation de l'arbre des espèces. Cette perte de gènes reproduit artificiellement la forte déplétion génique qui a lieu à la suite d'une WGD [Jaillon *et al.*, 2004].

L'équipe de David Sankoff a développé de nombreux simulateurs spécifiques à chacune de leurs études. Les inversions ainsi que les translocations sont simulées dans deux de ces études [Sankoff et Mazowita, 2005][Mazowita *et al.*, 2006]. La première étude [Sankoff et Mazowita, 2005] simule des inversions le long de génomes circulaires de bactéries et compare les distributions de tailles d'inversions simulées aux tailles d'inversions trouvées grâce à l'algorithme HP¹ appliqué à 32 génomes de bactéries. Comme l'algorithme HP n'est précis que pour les tailles des petites inversions, se sont celles-ci qui sont étudiées lors de la comparaison. Différentes distributions de tailles d'inversions sont utilisées, deux distributions en exponentielles décroissantes de paramètres respectifs $\lambda = 0.002$ et $\lambda = 0.05$ bp ainsi qu'une distribution gamma de paramètre de forme $\alpha = 0.6$ et d'échelle $\beta = 1200$ bp. L'autre étude [Mazowita *et al.*, 2006] effectue des simulations pour valider un estimateur de nombre d'inversions et de translocations. Les tailles d'inversions sont tirées dans une distribution gamma de paramètres de forme $\alpha = 3$ et d'échelle $\beta = 1.127 \ln(\text{bp})$, en échelle logarithmique. Dans ce dernier simulateur, les translocations réciproques se font plus fréquemment entre grands chromosomes qu'entre petits chromosomes, la probabilité qu'un chromosome soit transloqué y est proportionnelle à la taille des chromosomes. Les auteurs justifient ce choix en se basant sur l'hypothèse d'une répartition uniforme des points de cassures le long des chromosomes. Les deux simulateurs précédents sont exclusivement destinés à étudier les inversions et les translocations. Ils ne prennent pas en compte les évènements géniques ni les fissions et les fusions de chromosomes.

III.1.2 Simulateurs dédiés

Nous allons maintenant présenter trois simulateurs qui, à la différence des précédents, peuvent être téléchargés et paramétrés.

¹L'algorithme HP [Hannenhalli et Pevzner, 1995] permet de trouver le scénario le plus parcimonieux pour passer d'un génome à un autre, avec des inversions, des translocations réciproques, des fissions et des fusions.

Le premier est Aevol [Batut *et al.*, 2013], il s'agit d'un simulateur d'évolution de bactéries. Les génomes simulés sont circulaires et au lieu d'être composés des nucléotides ATCG ils sont composés des deux bits informatiques, 0 et 1. Un grand soin a été apporté à la modélisation des entités génomiques. Des séquences binaires correspondent à des promoteurs ainsi qu'à des terminaisons de transcriptions et d'autres séquences déterminent les segments codants. La structure des gènes reproduit fidèlement celle des gènes de bactéries, voir la figure III.1. Aevol effectue des mutations ponctuelles, des insertions et des délétions de petites séquences. Des variations du génome font intervenir de plus grands segments, les duplications, les délétions, les inversions et les translocations. Dans ce simulateur les « translocations » correspondent à ce que nous avons nommé des transpositions dans la partie (section I.3.2). Les réarrangements suivent de plus le modèle de répartition aléatoire des points de cassure. Enfin, une méthode calculatoire permet de passer du génotype binaire à un phénotype, par une agglomération des composantes phénotypiques des gènes et la fitness d'une cellule est estimée par la plus ou moins grande adéquation de ce phénotype avec l'environnement.

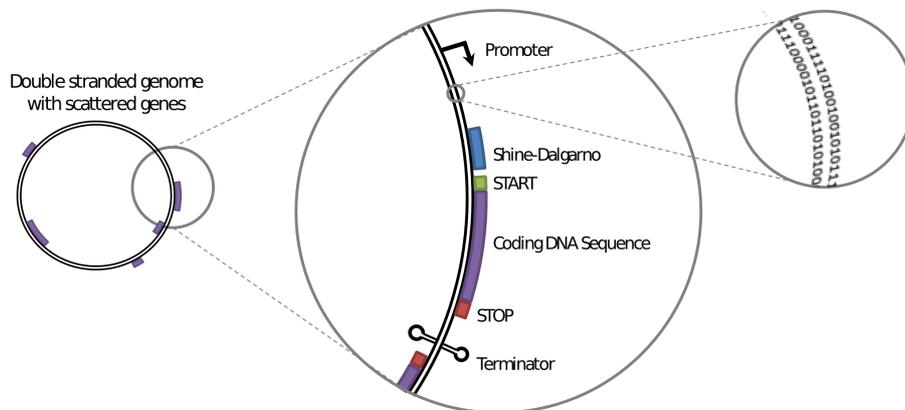


Figure III.1 – Modélisation des génomes dans Aevol. Les séquences codantes sont identifiées par des séquences prédéfinies : les promoteurs et les terminateurs marquent les limites des séquences transcrites, les séquences codantes sont comprises entre des codon start et des codons stop. Des nucléotides binaires remplacent les nucléotides ATCG.

Le deuxième simulateur est ALF (Artificial Life Framework) [Dalquen *et al.*, 2012]. En plus de modéliser les génomes à la même échelle que nous le faisons il modélise également l'évolution des séquences géniques par insertions, délétions ou substitutions de nucléotides, des insertions et des délétions de

segments de chromosome sont aussi prises en compte. Différentes classes de séquences sont disponibles (gènes, pseudogènes, non-codants, *etc*) et les séquences peuvent être divisées en domaines qui ont chacun leur propre vitesse d'évolution. L'évolution du contenu en GC, les évènements de spéciation, les duplications de gènes, les pertes de gènes et les transferts latéraux de gènes sont tous simulés. Dans ALF, les duplications se font en tandem ou de manière disperse. Elles dupliquent un gène ou un segment de gènes voisins et les évènements de transferts latéraux peuvent transférer plusieurs gènes lors du même évènement. Le réalisme de ce simulateur va jusqu'à la modélisation de la variation temporaire des taux d'évolution des séquences géniques suite à des évènements de néo-fonctionnalisation ou de sub-fonctionnalisation. Ce simulateur considère quatre réarrangements : les inversions de segments de gènes, les transpositions (appelées « translocations » dans leur article) de gènes, les fusions et fissions de domaines géniques. Les génomes sont constitués d'un unique chromosome, et ALF ne simule donc pas les fusions, les fissions et les translocations réciproques. De plus les inversions ont des longueurs arbitraires. Ainsi, les réarrangements d'ALF sont loin de s'approcher de la réalité.

Enfin, nous mentionnons l'existence d'un autre simulateur très complet qui modélise l'évolution du réseau de régulation en plus des duplications de gènes, des délétions de gènes et des inversions [ten Tusscher et Hogeweg, 2009]. Outre l'évolution du génome, le simulateur intègre également une modélisation des spéciations sympatriques, différents moyens de reproduction ainsi que des aires de compétitions et de reproduction. La richesse de ce simulateur se fait au coût d'une modélisation simpliste des réarrangements, car seules les inversions sont simulées et aucune information relative à la taille des inversions n'est fournie.

Les simulateurs que nous avons présentés sont peu adaptés à reproduire l'évolution réaliste de l'ordre des gènes dans les génomes. Le simulateur de Jian Ma se rapproche le plus de nos objectifs mais il ne modélise pas les duplications en tandem. De plus il a été développé de manière *ad hoc* pour valider ses méthodes de reconstruction de génomes ancestraux et il n'est pas téléchargeable. Nous avons donc décidé de développer notre propre simulateur.

Dans ce chapitre nous présentons notre simulateur, MagSimus (Modeling Ancestral Genomes by Simulations), qui a pour but premier de combler l'absence de génomes ancestraux liés les uns aux autres et liés à leur descendants modernes par une histoire évolutive commune. Pour palier à ce manque MagSimus simule une évolution qui reproduit l'évolution des génomes avec un réalisme quantifié.

Dans le chapitre précédent nous avons évoqué les différences substantielles

qui existent, au niveau des inversions chromosomiques, entre les métazoaires et des organismes moins complexes comme les procaryotes ou les levures (section II.3.1). Pour cette raison, et parce qu'il est encore prématuré de vouloir rendre compte de ces deux dynamiques évolutives dans un seul et même modèle, nous préférons nous concentrer sur la simulation des génomes de métazoaires, tout en gardant à l'esprit la nécessité de développer une méthode aussi générale que possible.

Dans un premier temps nous expliquerons les méthodes à partir desquelles nous inférons de nombreux paramètres concernant l'histoire évolutive de génomes modernes. Puis nous expliquerons le fonctionnement de notre simulateur MagSimus. Nous l'utiliserons ensuite pour simuler l'évolution de cinq espèces d'amniotes : l'humain, la souris, le chien, l'opossum et le poulet, à partir des paramètres précédents. Nous affinerons et nous estimerons les paramètres en comparant les simulations à la réalité. Enfin nous quantifierons le réalisme de nos simulations.

III.2 Inférence des paramètres d'une histoire évolutive à partir des génomes modernes

Étant donné un ensemble d'espèces modernes dont on souhaite reproduire l'évolution, le but est ici d'estimer les paramètres évolutifs : nombres de chromosomes dans les ancêtres communs, nombres de gènes, nombre des différents événements évolutifs, ... à partir de toutes les données modernes disponibles dans les bases de données.

III.2.1 Nombre de gènes dans les génomes ancestraux

La forêt d'arbres de gène d'Ensembl Compara (section I.10.2) est analysée de manière à estimer quels étaient les gènes présents dans les ancêtres des espèces modernes simulées. L'analyse de ces arbres nous donne accès aux nombres de gènes que nous inférons pour chaque espèce ancestrale.

III.2.2 Nombre d'événements géniques le long de chaque branche

Inférence des nombres d'événements géniques

Pour chaque branche de l'arbre des espèces simulées, nous comparons les gènes présents dans le génome parent avec les gènes présents dans le génome

enfant. Les gènes du génome enfant qui ne font pas partie d'une famille de gène du génome parent sont considérés comme des gènes issus de naissances *de novo*. Si un gène ancestral du parent est présent en plusieurs copies dans le génome enfant, nous considérons qu'il y a eu un nombre de duplications égal au nombre de descendants du gène ancestral moins 1 (le gène ancestral lui-même). Enfin, nous inférons autant de délétions qu'il y a de gènes du parent, sans descendant, dans le génome de l'enfant. Nous expliquons ces choix par ce qui suit.

Choix des scénarios les plus simples entre scénarios géniques équivalents

Les nombres de duplications, de délétions et de naissances *de novo* que nous avons inférés correspondent à une évolution idéale lors de laquelle, le long d'une branche, les gènes dupliqués ou issus d'une duplication n'ont pas été supprimés et dans laquelle les gènes nés *de novo* ne sont ni supprimés ni dupliqués. Il est certain que l'évolution réelle ne s'est pas déroulée comme cela et que le long d'une branche de nombreuses copies de gènes ont été insérées puis supprimées. Ce phénomène semble d'autant plus courant que suite à une duplication une des deux copies est souvent sujette à un relâchement de la pression de sélection et, par conséquent, l'une des deux copies peut facilement être perdue. De même, de nombreux gènes nés *de novo* au milieu d'une branche ont probablement été dupliqués ou supprimés le long de la même branche. Néanmoins ces événements précédents, s'ils ont eu lieu ne changeront pas nos résultats. Une analyse fine des arbres de gènes permettrait de les identifier mais ces successions d'évènements seraient par la suite difficiles à reproduire et il nous semble que les simuler complexifierait le simulateur démesurément par rapport à l'apport en réalisme qui en résulterait. En bref, dans notre étude, mis à part pour les nombres de naissances et de duplications par branche, cela ne change rien qu'un gène ait été inséré par une naissance *de novo* le long d'une branche ou qu'il ait été inséré suite à la duplication d'un gène qui venait de naître précédemment, lui aussi dans la même branche. Nous nous permettons donc de prendre en compte ces éventuelles duplications comme autant de naissances *de novo*. Nous devons néanmoins garder en mémoire que le nombre de naissances *de novo* que nous inférons le long d'une branche diffère probablement du nombre réel. Il correspond en fait au nombre de descendants de gènes nés *de novo* qui ont été conservés dans l'enfant. De même les nombres de délétions inférés correspondent aux gènes ancestraux du parent qui ont été supprimés et les suppressions de gènes, insérés à la suite d'une duplication ou d'une naissance *de novo*, ne sont pas comptabilisées si la duplication ou la naissance à eu lieu dans la même branche que la délétion.

Enfin, le nombre de duplications dans une branche correspond au nombre de descendants d'un gène ancestral moins un, pour ne pas comptabiliser le gène ancestral dans l'éventualité où il ait été conservé. Là encore il se pourrait que dans la réalité le gène ancestral ait été supprimé et que seules ses copies aient été conservées dans l'enfant. Enfin, comme nous le disions, les autres duplications de familles de gènes nés dans la branche, sont comptabilisées comme des naissances *de novo* si ces nouvelles copies de gènes sont conservées jusqu'à l'enfant.

Comme nous le verrons, notre simulateur a été conçu de manière à reproduire ces évolutions géniques simples.

Nombre de duplications en tandem sur chaque branche

En nous basant sur les arbres de gènes et sur les génomes réels des espèces que nous souhaitons simuler nous avons appliqué une méthode très simple pour estimer le nombre de duplications en tandem dans chaque branche de l'arbre des espèces. Dans une branche, le gène initial est copié lors d'une duplication et la copie est insérée dans le génome. Si la duplication est en tandem, la copie est insérée près (séparée par moins de 10 gènes) du gène initial, sinon elle est insérée loin sur le même chromosome ou sur un autre chromosome. Ainsi nous estimons que la duplication a eu lieu en tandem si au moins un descendant du gène initial est proche d'un descendant du gène inséré dans les espèces modernes, en aval de l'évènement.

III.2.3 Probabilité qu'un gène inséré par une duplication en tandem ait la même orientation que le gène dupliqué

Si une duplication en tandem a lieu, le gène inséré (la copie) peut avoir la même orientation que le gène copié ou une orientation différente. Dans les génomes modernes, il est souvent impossible de savoir lequel des deux gènes en tandem est le gène ancestral et les deux copies sont donc considérées toutes les deux comme l'éventuel gène ancestral. Si les deux copies ont la même orientation nous pouvons inférer l'orientation du gène ancestral par parcimonie. Dans le cas contraire, les deux gènes ont des orientations différentes et nous n'avons pas de raison d'assigner une orientation au gène ancestral plutôt qu'une autre. De manière générale, quelque soit le nombre de copies dans les clusters de gènes dupliqués en tandem; le nombre d'orientations des gènes ancestraux, à l'origine des clusters, qui peuvent être inférées, dépend de la fréquence à laquelle les copies, insérées par duplications en tandem, se sont

insérées avec la même orientation que le gène dupliqué. Plus particulièrement, simuler de manière réaliste les orientations de gènes dupliqués en tandem est déterminant pour évaluer la fréquence des erreurs d'identification d'inversions mono-géniques (figure II.13). Estimons donc la probabilité, $ptandem$, qu'un gène dupliqué en tandem soit inséré avec la même orientation que le gène d'origine.

Cette estimation est effectuée en comparant plusieurs simulations¹ à la réalité. Chaque simulation a une probabilité $ptandem$ différente que le gène dupliqué ait la même orientation. À la fin de chaque simulation les clusters de gènes dupliqués en tandem sont comparés aux clusters réels. La comparaison se fait via le calcul d'une statistique, s . Nous considérons que la simulation dont la statistique est la plus proche de la statistique réelle correspond à la simulation la plus réaliste. La probabilité $ptandem$ utilisée pour effectuer la simulation la plus réaliste sera notre estimation de la probabilité $ptandem$ réelle. Nous expliquons la statistique s qui a été utilisée pour comparer une simulation à la réalité.

Considérons un ensemble de clusters dont le nombre total de duplications est G , le nombre de gènes dans les clusters moins le nombre de clusters. Pour chaque cluster nous notons g le nombre de ses gènes et nous définissons, g_+ (resp. g_-) les nombres de gènes orientés positivement (resp. négativement), $p_+ = \frac{g_+}{g}$, la fraction de gènes orientés positivement et $p_- = \frac{g_-}{g}$, la fraction de gènes orientés négativement. La probabilité qu'une duplication de ce cluster ait inséré un nouveau gène avec la même orientation que le gène copié peut être estimée par

$$p = p_+ \times \frac{g_+ - 1}{g - 1} + p_- \times \frac{g_- - 1}{g - 1} \quad (\text{III.1})$$

La statistique s est une estimation de la probabilité $ptandem$,

$$s = \sum_i \frac{g_i - 1}{G} p_i \quad (\text{III.2})$$

avec chaque valeur relative à un cluster, indexée par la lettre i . $\frac{g_i - 1}{G}$ correspond à la probabilité qu'une des G duplications qui a eu lieu durant l'évolution soit une duplication du cluster i . Nous utiliserons la statistique s et l'estimation de $ptandem$ plus tard (section III.4.2).

III.2.4 Nombres de chromosomes dans les ancêtres

Pour inférer le nombre de chromosomes dans les différents ancêtres, un grand arbre phylogénétique est constitué de manière à ce que les feuilles incluent les

¹Ces simulations sont générées par un simulateur annexe à MagSimus.

espèces modernes dont on souhaite reproduire l'évolution. Pour avoir plus d'informations sur les états ancestraux, cet arbre contient d'autres espèces modernes non simulées. Les nombres de chromosomes à tous les nœuds internes de ce grand arbre sont inférés avec ChromEvol [Mayrose *et al.*, 2010]. Pour l'inférence, seuls les chromosomes autosomiques des espèces modernes sont utilisés au niveau des feuilles. Les chromosomes sexuels sont ajoutés par la suite car ils évoluent souvent indépendamment des chromosomes autosomiques¹. Les nombres de chromosomes modernes sont issus des bases de données Genome Size [Gregory, 2016].

III.2.5 Distribution des gènes dans les chromosomes initiaux

Dans le génome initial, la distribution des nombres de gènes par chromosome est obtenue en faisant la moyenne des distributions des tailles (en gènes) des chromosomes dans les génomes modernes. Pour faire la moyenne, nous associons une fonction de cumul des tailles de chromosomes à chacun des N génomes modernes. Dans chaque génome les c chromosomes sont triés de gauche à droite par tailles décroissantes. Nous normalisons les fonctions de cumul pour que la valeur finale soit égale à 1 et nous répartissons les chromosomes sur l'axe des abscisses de manière à ce que les rangs des chromosomes soient uniformément répartis dans l'intervalle $[0, 1]$. De cette manière le plus grand chromosome est à l'abscisse $x = 0$, le deuxième plus grand chromosome est à l'abscisse $x = \frac{1}{(c-1)}$, ... et le plus petit chromosome est à l'abscisse $x = 1$. Une interpolation linéaire nous permet d'associer la fonction de cumul, linéaire par morceaux

$$F_i: [0, 1] \rightarrow [0, 1] \tag{III.3}$$

$$x \mapsto y$$

à chaque génome moderne $i \in [1, N]$, avec $y = F_i(x)$, une interpolation de la somme des gènes des chromosomes de rangs inférieurs ou égaux à x (dans le génome i). Nous définissons la fonction de cumul, normalisée, des tailles des

¹Par exemple, la X Added Region (XAR) a été ajoutée à l'extrémité pseudo-autosomique du chromosome X avant la radiation des mammifères placentaires. Autre exemple, il semble que des réarrangements multiples, entre une paire ancestrale de chromosomes sexuels et quatre chromosomes autosomiques, aient eu lieu durant l'évolution des monotrèmes jusqu'à l'ornithorynque. Ce qui expliquerait pourquoi un ornithorynque possède 5 paires de chromosomes sexuels : les femelles ont cinq paires de chromosomes X, et les mâles ont cinq X et cinq Y [Graves, 2015]

chromosomes du génome initial par la fonction suivante

$$\begin{aligned} \bar{F}: [0, 1] &\rightarrow [0, 1] \\ x &\mapsto \frac{1}{N} \sum_{i=1}^N F_i(x). \end{aligned} \quad (\text{III.4})$$

Étant donné γ , le nombre de chromosomes dans le génome initial, le nombre de gènes $|c_j|$ dans le chromosome de rang $j \in [1, \gamma]$ est donné par les équations suivantes

$$\bar{f}(j) = \begin{cases} \bar{F}(0) & \text{si } j = 1 \\ \bar{F}\left(\frac{j-1}{\gamma-1}\right) - \bar{F}\left(\frac{(j-1)-1}{\gamma-1}\right) & \text{si } j \geq 2 \end{cases} \quad (\text{III.5})$$

$$|c_j| = \lfloor g \times \bar{f}(j) \rfloor \quad (\text{III.6})$$

avec $\bar{f}(j)$ l'histogramme discrétisé et normalisé des tailles de chromosomes du génome initial, g le nombre de gènes dans le génome initial, $x = \frac{k-1}{\gamma-1}$, le rang normalisé du chromosome de rang k et $\lfloor g \times \bar{f}(j) \rfloor$, la partie entière de $g \times \bar{f}(j)$.

Enfin, si la somme des gènes dans les chromosomes du génome initial, $\sum_{j=1}^{\gamma} |c_j|$ est inférieure à g , les gènes restant, $g - \sum_{j=1}^{\gamma} |c_j|$, sont distribués dans les γ chromosomes. Autant que possible, les gènes restant sont distribués uniformément dans les γ chromosomes et, si cela est nécessaire, quelques chromosomes de destination sont choisis au hasard.

III.2.6 Nombre de réarrangements

Nombre de fissions et de fusions

Dans un premier temps le long de chaque branche du grand arbre les nombres de fissions et de fusions sont estimés par parcimonie à partir des nombres de chromosomes à chaque noeud (section III.2.4). Si l'espèce parent a plus de chromosomes que l'espèce enfant, le nombre de fusions nécessaires, pour diminuer en conséquence le nombre de chromosomes, est attribué à la branche. De même si l'espèce parent a moins de chromosomes que l'espèce enfant, le nombre de fissions nécessaires, pour ajuster le nombre de chromosomes, est attribué à la branche. Les nombres de fissions et de fusions le long des branches de l'arbre simulé sont déduits de ceux du grand arbre en additionnant les événements de fusions et de fissions des multiples branches du grand arbre. Les nombres de fusions et de fissions peuvent donc tous les deux être non

nuls sur la même branche de l'arbre simulé. Par exemple, le long d'une lignée du grand arbre entre trois espèces consécutives $A \rightarrow B \rightarrow C$, admettons qu'il y ait 2 fusions sur la branche $A \rightarrow B$ et une fission sur la branche $B \rightarrow C$. L'arbre simulé, qui, par exemple, ne contient que les espèces A et C, cumulera alors 2 fusions et une fission sur la branche $A \rightarrow C$. Si nous avons calculé les fissions et les fusions par parcimonie directement sur l'arbre simulé, nous n'aurions inféré qu'une fusion sur la branche $A \rightarrow C$.

Nombre de translocations réciproques et nombre d'inversions

Dans un premier temps nous estimons le nombre de translocations et d'inversions qu'il y a probablement eu entre deux espèces modernes.

Les inversions et les translocations qui séparent deux espèces modernes sont estimées à partir des segments conservés entre ces deux espèces. Nous reprenons pour cela la démarche de Matthew Mazowita [Sankoff et Mazowita, 2005][Mazowita *et al.*, 2006]. À l'origine de son calcul, Mazowita fait une hypothèse issue du RBM. Dans son calcul, la sélection des deux chromosomes transloqués se fait proportionnellement à leurs tailles. Ceci garantit que toutes les localisations des chromosomes sont d'équiprobables points de cassures de translocations. Si les chromosomes transloqués sont sélectionnés de manière uniforme la probabilité $p_i(j)$ de Mazowita sera $\frac{1}{c}$, avec c le nombre de chromosome dans le 1^{er} génome (A), au lieu d'être $\frac{p(j)}{1-p(i)}$ comme c'est le cas dans le calcul de Mazowita. Une fois ce changement opéré, l'équation (9) devient

$$\sum_i c^{(i)} = cd - c(d-1) \left[\frac{d-1}{d} \right]^{\frac{2t}{c}}, \quad (\text{III.7})$$

avec d le nombre de chromosomes dans le deuxième génome (B), $c^{(i)}$ le nombre de chromosomes de B qui contiennent des segments conservés en commun avec le chromosome i de A et t le nombre de translocations que nous cherchons à estimer.

Nous proposons une interprétation intuitive de cette équation. Le terme de gauche correspond au nombre de paires de chromosomes, l'un dans A et l'autre dans B, qui partagent au moins un segment conservé. Ce terme est évalué à partir des données réelles. Le terme de droite correspond au nombre attendu de paires de chromosomes avec au moins un segment conservé en commun quand il y a t translocations. cd est le nombre total de paires de chromosomes, $\left[\frac{d-1}{d} \right]^{\frac{2t}{c}}$ est la probabilité qu'une paire de chromosomes ne partage pas de segment conservé malgré t translocations et $c(d-1)$ est égal au nombre de paires de chromosomes moins les c chromosomes de B qui

contiennent *a priori* un segment conservé (les vestiges des chromosomes de A) même s'il n'y a pas eu de translocations.

La formule précédente permet d'estimer le nombre de translocations qui séparent chaque combinaison de deux espèces modernes. Similairement à la suite du raisonnement de Mazowita, nous estimons le nombre d'inversions à partir du système d'équations suivant

$$s = c + k - (r + e + d) \quad (\text{III.8})$$

$$k = f + 2i + 2t. \quad (\text{III.9})$$

Dans la première équation s est le nombre de segments conservés, c le nombre de chromosomes de A, k le nombre de points de cassures, r le nombre de réutilisations de points de cassures, e le nombre de points de cassures qui sont tombés aux extrémités des chromosomes de A et d le nombre de segments conservés supprimés à cause des délétions de gènes ancestraux. Dans la deuxième équation f est le nombre de fissions, i le nombre d'inversions et t le nombre de translocations.

Comme Mazowita, pour mener à bien notre calcul nous faisons l'hypothèse que r , e et d sont suffisamment petits par rapport à s , c et k pour pouvoir les négliger. Nous reviendrons sur cette hypothèse plus tard (section III.8.2). Il s'ensuit que

$$i = \frac{1}{2}(s - c - 2t - f). \quad (\text{III.10})$$

Nous construisons deux matrices de distances, l'une pour les inversions et l'autre pour les translocations. Dans la première matrice, les distances correspondent aux nombres de translocations réciproques qui séparent deux espèces modernes et dans la deuxième les distances sont les inversions. Le nombre d'inversions et le nombre de translocations pour chaque branche de l'arbre des espèces, sont calculés tour à tour en appliquant la méthode d'inférence phylogénétique des moindres carrés [Felsenstein, 2004], avec la contrainte que les nombres de translocations et d'inversions soient non nuls le long de toutes les branches, Non-Negative Least Squares (NNLS).

Nous venons d'inférer de nombreux paramètres évolutifs des espèces que nous souhaitons simuler. Nous détaillons par la suite comment le simulateur effectue les simulations.

III.3 Le fonctionnement du simulateur MagSimus

III.3.1 Génome initial

Un génome initial sert de point de départ à la simulation. Pour commencer la simulation, le simulateur a besoin de connaître le nombre de chromosomes, le nombre de gènes et les tailles en gènes (au moins relatives) des chromosomes. Toutes ces valeurs sont des paramètres. Par la suite notre simulateur fait évoluer ce génome artificiel le long des branches de l'arbre des espèces selon les lois suivantes.

III.3.2 L'évolution est décomposée en évolutions par branches

Notre simulateur décompose l'évolution en une succession d'évolutions le long de chaque branche de l'arbre des espèces. Dans un premier temps, le génome initial est copié autant de fois qu'il y a de ramifications à la racine. Le simulateur fait ensuite évoluer indépendamment chaque copie du génome ancestral le long de chacune des branches qui descendent de la racine. Au bout de chaque évolution, le génome modifié correspond à l'espèce située à l'extrémité de la branche. Si cette espèce est une espèce ancestrale intermédiaire, de manière récursive, son génome est de nouveau copié et les copies évoluent le long des branches successives. Dès qu'une évolution débouche sur une espèce moderne le processus récursif précédent est arrêté. Durant le déroulement du processus récursif, les génomes ancestraux sont sauvegardés.

III.3.3 Évènements évolutifs modélisés

Parmi les évènements évolutifs présentés dans le premier chapitre, les évènements implémentés dans notre simulateur sont les suivants :

- Les évènements géniques : duplications mono-génique (*gDup*) (soit en tandem soit disperse), délétion mono-génique (*gDel*) et naissances de novo (*gBirth*).
- Les évènements de réarrangements chromosomiques : les inversions (*Inv*) de segments de chromosomes, les translocations réciproques (*Transl*), les fissions (*Fis*) et les fusions (*Fus*) de chromosomes.

Nous considérons ici les éventuels transferts horizontaux de gènes comme des évènements de naissances *de novo*. Notre simulateur ne modélise pas

actuellement les duplications de segments de chromosomes contenant plusieurs gènes, ni les délétions de segments de chromosomes. Il ne modélise pas les transpositions mono-géniques, ni les éventuelles transpositions de segments de chromosomes. Les duplications complètes de chromosomes ne sont pas simulées non plus. Tous ces évènements semblent cependant peu fréquents dans les lignées de vertébrés (section I.3). Pourvu que les lignées simulées ne contiennent pas de duplication complète de génome, même s’il ne reproduit pas les évènements que nous venons de mentionner, nous pensons néanmoins que, à notre échelle d’étude, notre simulateur nous permettra de reproduire les caractéristiques génomiques les plus remarquables¹. Nous décrirons ces caractéristiques remarquables ultérieurement (section III.5).

L’implémentation des évènements découle directement de leurs descriptions dans le chapitre 1 (section I.3). Dans le cas des translocations, nous autorisons qu’un des deux segments de chromosome échangé ne contienne pas de gène. Nous présentons par la suite les idées maîtresses qui font l’intérêt et la spécificité de notre simulateur.

III.3.4 Liste ordonnée d’évènements à chaque branche

Au début de chaque évolution le long d’une branche, une liste ordonnée d’évènements est créée. Cette liste indique la succession d’évènements qui vont modifier le génome parent (l’origine de la branche) en un génome enfant (la fin de la branche). Par exemple la liste $[gDup, Inv, gDel]$ indique qu’il y a trois évènements durant l’évolution du génome. Le premier évènement est une duplication de gène, le deuxième évènement est une inversion de segment de chromosome et enfin que le troisième et dernier évènement est une délétion génique.

Voici comment est construite la liste d’évènements d’une branche. Pour chaque type d’évènement ($e \in [gDup, gDel, gBirth, Inv, Transl, Fus, Fis]$) le nombre ($n_{e,b}$) d’évènements le long de la branche (b) est égal au taux de l’évènement le long de la branche ($r_{e,b}$ en nombre d’évènement par millions d’années) multiplié par la longueur de la branche (l_b en millions d’années). Les longueurs des branches, en millions d’années, sont fixées par l’arbre des espèces et les taux d’évènements spécifiques à chaque branche sont des variables de notre simulateur.

¹Les principes de fonctionnement du simulateur sont suffisamment flexibles pour que d’autres évènements évolutifs rares puissent être ajoutés par la suite.

III.3.5 Simulations des scénarios les plus simples

Pour être cohérent avec nos choix précédents (section III.2.2) nous nous assurons que les scénarios d'évènements géniques simulés correspondent aux scénarios sélectionnés pour l'inférence des nombres d'évènements géniques à partir des données réelles. Par conséquent dans une branche : seuls les gènes d'une famille ancestrale peuvent être dupliqués et seuls les gènes ancestraux conservés en une copie peuvent être supprimés. Les gènes ancestraux qui ont été dupliqués ou les nouveaux gènes qui ont été insérés suite à une duplication ou une naissance *de novo* ne peuvent pas être supprimés. Les gènes insérés puis supprimés dans la même branche ne sont pas simulés. Ceci nous assure que, dans les simulations, les nombres de duplications, délétions et naissances ont la même signification que précédemment. Les bilans de nombre de gènes suivants seront donc vérifiés :

$$|g_p| + n_{gBirth} + n_{gDup} - n_{gDel} = |g_e| \quad (\text{III.11})$$

$$|g_p| - n_{gDel} = |g_e \cap g_p| \quad (\text{III.12})$$

avec, g_p et g_e les ensembles de gènes dans le parent et dans l'enfant, n_{gBirth} le nombre de naissances de gènes, n_{gDup} le nombre de duplications, n_{gDel} le nombre de délétions et $|x|$ le nombre d'éléments dans l'ensemble x . $g_e \cap g_p$ est l'ensemble des gènes ancestraux conservés dans le génome de l'enfant. En plus de ces bilans, dans l'enfant, la somme des gènes des familles (de gènes du parent) contenant au moins deux gènes sera égal à n_{gDup} moins le nombre de ces familles (le nombre de gènes du parent à l'origine des familles). Enfin le nombre de gènes de l'enfant qui ne font pas partie d'une famille d'un gène ancestral est égal au nombre de naissances *de novo* tel que nous l'avons défini.

III.3.6 Contraintes sur les nombres d'évènements

Pour que l'évolution le long d'une branche soit possible, il est nécessaire que le nombre de délétions de gènes dans la liste soit inférieur au nombre de gènes du parent. Il est également nécessaire que le nombre de fusions dans la liste soit inférieur au nombre de chromosomes initiaux moins un. Si l'une des deux contraintes précédente n'est pas vérifiée la simulation ne peut pas aboutir et, dès le début de l'exécution, notre simulateur avertit l'utilisateur que les paramètres ne sont pas conformes à ce qui est attendu.

Les deux contraintes précédentes ne sont pas les seules conditions nécessaires pour qu'une simulation puisse aboutir. Par exemple si à un moment donné tous les chromosomes n'ont qu'un seul gène il ne peut pas y avoir d'inversion. Anticiper un tel cas de figure peut néanmoins être complexe étant donné que notre simulateur est non-déterministe. Reprenons l'exemple

précédent, bien qu'il soit volontairement caricatural. Si l'ordre arbitraire des évènements fait que de très nombreuses délétions sont faites au début d'une branche, à un moment, il se peut qu'il n'y ait plus qu'un gène ancestral par chromosome. Si l'évènement suivant est sensé être une inversion, elle ne pourra pas être effectuée. Par contre si l'ordre arbitraire des évènements fait que l'inversion est première par rapports aux nombreuses délétions, alors elle pourra avoir lieu. La faisabilité d'une liste d'évènements dépend donc en partie de l'ordre des évènements.

D'autres fonctionnements arbitraires du simulateur peuvent empêcher le déroulement successif des évènements d'une liste. Nous avons développé un gestionnaire de liste d'évènements pour contrôler, au cas par cas, ces exceptions de fonctionnement quand elles arrivent et pour tenter de continuer le déroulement des évènements, quitte à changer parfois l'ordre des évènements ultérieurs. Nous l'introduirons plus tard (section III.3.10).

III.3.7 Modes de sélection des chromosomes réarrangés

Lorsqu'un réarrangement a lieu, celui-ci implique un ou deux chromosomes. Une inversion et une fission impliquent l'une et l'autre un chromosome, une translocation réciproque et une fusion impliquent deux chromosomes. Dans chacun des cas, le (ou les) chromosome(s) sélectionnés peuvent être sélectionné(s) indépendamment de leur(s) taille(s) ou proportionnellement à leur(s) taille(s). Un paramètre de notre simulateur permet de choisir le mode de sélection des chromosomes réarrangés. Le premier mode, *uniforme*, correspond à un tirage équiprobable de tous les chromosomes alors que le deuxième mode, *proportionnel*, correspond à une sélection des chromosomes proportionnellement à leurs tailles (dans notre cas en gènes). Avec ce dernier mode de sélection, un chromosome deux fois plus grand qu'un autre a deux fois plus de chances d'être sélectionné. Ce dernier mode de sélection a parfois été utilisé pour des translocations [Ma *et al.*, 2006]. Nous n'avons pas considéré d'autres modes de sélection, mais il pourrait par exemple y avoir une sélection inversement proportionnelle à la taille pour les chromosomes fusionnés ou une sélection pondérée par une fonction de fitness [Arkendra *et al.*, 2001].

III.3.8 Tailles limites d'un chromosome

Certains auteurs ont supposé qu'il existe des limites inférieures ou supérieures aux tailles de chromosomes. Il se peut par exemple qu'un chromosome viable et fonctionnel requière un centromère, deux télomères et un gène [Sankoff et Ferretti, 1996]. De plus sa taille minimale est peut-être limitée pour qu'un crossover puisse avoir lieu [Sankoff et Ferretti, 1996]. À l'opposé il a également

été proposé de limiter la taille maximale d'un chromosome [Arkendra *et al.*, 2001]. L'argument biologique de cette limite est qu'un chromosome trop grand ne pourra pas ségréger normalement lors de la division cellulaire. À partir d'expériences sur des cellules de haricots, Schubert et Oud [Schubert et Oud, 1997] suggèrent que pour qu'un organisme puisse se développer « le plus long chromosome ne doit pas excéder la moitié de l'axe du fuseau mitotique à la télophase ». Dans MagSimus il y a donc deux paramètres, un pour spécifier les tailles minimales et l'autre pour spécifier les tailles maximales de chromosomes.

III.3.9 Un réarrangement chromosomique implique au moins un gène ancestral

Lors d'une évolution le long d'une branche qui relie un génome ancestral parent à un génome enfant, des gènes non-ancestraux sont parfois insérés suite à des duplications ou des naissances. Par la suite, le long de la même branche, il est probable que certains réarrangements modifient l'ordre ou les orientations de ces gènes non-ancestraux, sans modifier l'ordre ou les orientations des gènes ancestraux du parent. Néanmoins ces réarrangements ne pourront pas être détectés à partir des génomes modernes car aucune comparaison d'espèces modernes ne pourra les révéler. Le long de cette branche, seuls des réarrangements qui modifient l'ordre ou les orientations des gènes ancestraux du parent pourront être détectés à partir de comparaisons d'espèces modernes. Ceci est vrai pour l'évolution réelle qu'il y a eu le long des différentes branches de l'arbre des espèces et c'est également vrai durant la simulation.

Une particularité de notre simulateur est de pouvoir effectuer spécifiquement des réarrangements qui modifient l'ordre et les orientations d'au moins un gène ancestral. Grâce à cette particularité, lorsque nous spécifions un nombre de réarrangements, il s'agit directement du nombre de réarrangements qui impliquent des gènes ancestraux. Cette fonctionnalité est particulièrement utile pour les branches qui ont de nombreuses insertions de nouveaux gènes suite à des duplications ou des naissances.

III.3.10 Le gestionnaire de la liste d'évènements

Si les nombres de délétions de gènes et de fusions sont conformes à ce qui est attendu, une liste ordonnée d'évènements est constituée. Un chiffre est attribué à chaque évènement : par exemple 1 correspond à une inversion, 2 à une translocation, 3 à une fission, 4 à une fusion, 5 à une duplication de gènes,

6 à une délétion de gènes et 7 à une naissance *de novo*. La liste contiendra autant de 1 qu'il y a d'inversions le long de la branche, c'est à dire $n_{Inv,b}$ (ou de manière synonyme $n_{1,b}$). De même il y aura $n_{Transl,b}$ chiffres 2 dans la liste, *etc.* Au début de l'évolution l'ordre de ces évènements dans la liste est choisi aléatoirement. À partir de cette liste initiale, le génome parent est modifié successivement par le premier évènement, puis le deuxième, *etc.* Une fois que l'évènement a eu lieu, le chiffre correspondant est retiré de la liste.

Malheureusement lorsque les évènements sont effectués les uns à la suite des autres il arrive parfois (comme nous l'avons vu, section III.3.6) qu'il soit impossible d'effectuer un évènement. Nous détaillons l'ensemble de ces cas et pour chacun d'eux nous définissons une exception de fonctionnement du simulateur (en italique) :

- *MoinsDeDeuxChrs* : une translocation ou une fusion est sensée avoir lieu alors qu'il y a moins de deux chromosomes.
- *PasAssezDeGènesAncDansChrs* : un réarrangement est sensé être effectué (par exemple une translocation) et un des chromosomes réarrangés doit avoir au moins x gènes ancestraux, or aucun chromosome n'en contient suffisamment. Par exemple une inversion ou une fission est sensée avoir lieu alors que tous les chromosomes contiennent moins de deux gènes ancestraux.
- *ChrsTropCourts* : un évènement qui décroît la longueur des chromosomes (fission ou délétion de gène) est sensé avoir lieu mais, s'il se produit, il est certain, ou excessivement probable¹, que la taille d'au moins un chromosome sera inférieure à $minChrL$, la taille minimale autorisée.
- *ChrsTropLongs* : un évènement qui accroît la longueur des chromosomes (fusion, duplication ou naissance de gène) est sensé avoir lieu mais, s'il se produit, il est certain, ou excessivement probable, que la taille d'au moins un chromosome sera supérieure à $maxChrL$, la taille maximale autorisée.
- *TropImprobable* : un réarrangement est sensé être effectué, il y a au moins une combinaison de chromosomes qui convient, mais le tirage aléatoire d'une combinaison qui convient est improbable.

Pour chacune des exceptions, le gestionnaire d'évènement suit la logique suivante :

¹si plus de 100 tentatives ont échoué

- *MoinsDeDeuxChrs* : le gestionnaire reporte l'évènement après une fission s'il y en a.
- *PasAssezDeGènesAncDansChrs* : le gestionnaire reporte l'évènement après une fusion ou une translocation s'il y en a¹.
- *ChrsTropCourts* : le gestionnaire reporte l'évènement après un évènement qui peut accroître le nombre de gènes dans un chromosome (fusions ou duplications et naissances géniques, translocation réciproque).
- *ChrsTropLongs* : le gestionnaire reporte l'évènement après un évènement qui peut faire décroître le nombre de gènes dans un chromosome (fissions et délétions de gènes, translocation réciproque).
- *TropImprobable* : Si la raison pour laquelle l'évènement ne peut pas avoir lieu n'est pas identifiable, la simulation s'arrête et informe l'utilisateur qu'une exception de fonctionnement rare a eu lieu. Il est suggéré à l'utilisateur de vérifier les nombres d'évènements par branche pour éviter les cas limites de fonctionnement.

Dans le cas des exceptions *ChrsTropCourts* et *ChrsTropLongs*, le double rôle des translocations s'explique par le fait que les translocations peuvent à la fois accroître et décroître la longueur d'un des chromosomes du génome.

Nous reconnaissons enfin qu'en pratique, pour éviter de trop nombreux reports, les éditions de la liste reportent en réalité les évènements à la fin de la liste.

III.3.11 Tailles des segments réarrangés

Tailles des segments transloqués

Comme nous le mentionnions (section I.3.2), une translocation est caractérisée entre autre par les longueurs des deux segments transloqués. Nous avons choisi la solution la plus simple pour tirer ces longueurs. À chaque fois qu'un chromosome est sélectionné pour être réarrangé, la longueur du segment transloqué est tirée uniformément entre une longueur nulle et la longueur totale du chromosome. De plus, un des deux segments échangés doit contenir au moins un gène ancestral.

¹Nous rappelons que les naissances et les duplications géniques n'augmentent pas le nombre de gènes ancestraux.

Tailles des segments inversés

Une inversion est également caractérisée par une taille de segment, celle du segment inversé. De nombreux travaux ont estimé la distribution des tailles de segments inversés : dans des lignées de bactéries [Lefebvre *et al.*, 2003][Sankoff, 2005b], ou durant l'évolution des amniotes [Ma *et al.*, 2006]. Dans d'autres travaux, les auteurs font l'hypothèse qu'il existe une distribution probabiliste des longueurs de segments inversés qui est valable pour toutes les lignées dont ils simulent l'évolution [Sankoff et Mazowita, 2005][Ma, 2006]. Une telle distribution informe le simulateur sur la taille des segments à inverser pour reproduire le plus possible l'estimation des véritables tailles de segments inversés. Comme les études précédentes [Sankoff, 2005b], nous avons fait l'hypothèse d'une distribution unique et valable pour différentes lignées. Cette distribution probabiliste des tailles de segments inversés peut être définie de différentes manières : avec des tailles absolues ou avec des tailles relatives. Nous faisons plusieurs suggestions sur la forme de cette distribution.

- **Distribution uniforme**

C'est la distribution la plus simple. Une fois qu'un chromosome est sélectionné, toutes les tailles d'inversions sont équiprobables. Cette distribution est représentée sur la figure III.2a pour un chromosome d'une longueur $L = 1330$ gènes. Cette dernière longueur correspond à la longueur de chromosome moyenne sur les espèces modernes. Plus exactement, pour chaque espèce nous avons calculé une longueur de chromosome barycentrique : le barycentre des longueurs de chromosomes. Dans le calcul du barycentre, chaque pondération est égale à la proportion du nombre de gènes dans le chromosome. La longueur de 1330 gènes correspond à la moyenne arithmétique des longueurs barycentriques, sur l'ensemble des espèces modernes.

- **Distribution triangulaire décroissante du RBM**

D'après le RBM, chacun des deux points de cassure d'une inversion peut être considéré comme uniformément distribué le long du chromosome hôte de l'inversion. Avec notre simulateur nous modélisons ce cas de figure de la manière suivante. Dans un premier temps nous sélectionnons le chromosome hôte de l'inversion puis nous tirons au hasard les deux localisations des points de cassures avec une distribution uniforme bornée par la taille du chromosome. Il est possible de démontrer (section A.1) que d'après cette modélisation, la probabilité qu'un segment inversé soit

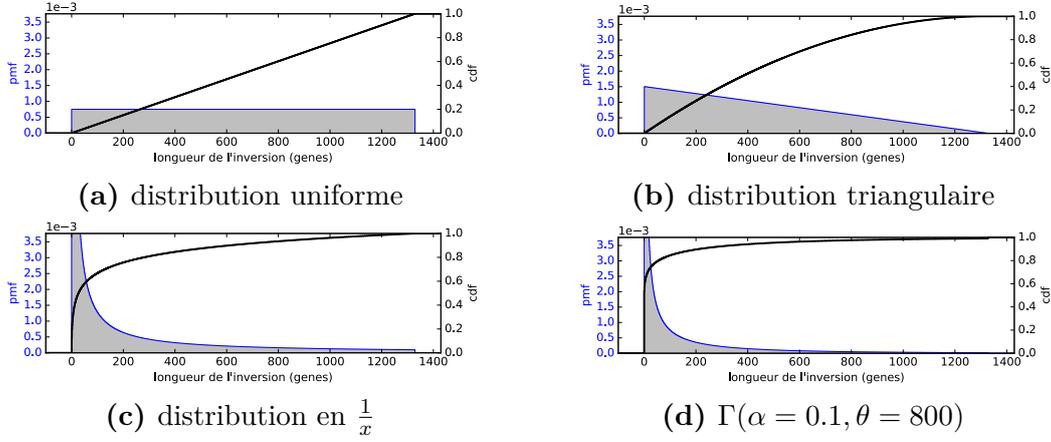


Figure III.2 – Différentes distributions probabilistes de tailles d’inversions, pour un chromosome de 1330 gènes. En bleu, la Probability Mass Function (pmf) est la fonction de masse de probabilité de chaque longueur d’inversion et, en noir, la Cumulative Distribution Function (cdf) est la fonction de répartition.

de taille x , dans un chromosome de L gènes est :

$$P(X = x) = \frac{2(L - x + 1)}{L(L + 1) - 2} \quad \forall x \in \llbracket 1, L - 1 \rrbracket. \quad (\text{III.13})$$

Il s’agit d’une distribution triangulaire décroissante qui débute à $\frac{2L}{L(L+1)-2}$ et qui décroît linéairement jusqu’à à atteindre $\frac{4}{L(L+1)-2}$ en $x = L - 1$. Cette distribution est représentée sur la figure III.2b pour un chromosome de taille, là encore, $L = 1330$ gènes.

- **Distribution attendue d’après la probabilité de contact entre deux locus d’un chromosome**

La densité de probabilité de contact entre deux locus d’un même chromosome a été estimée comme inversement proportionnelle à la distance qui les sépare [Lieberman-Aiden *et al.*, 2009]. Même si cette loi montre des limites pour des distances inférieures à 1 Mb et supérieures à 10 Mb, elle peut être une approximation intéressante de la distribution réelle des inversions. La distribution des tailles d’inversions qui en découle est représentée sur la figure III.2c, toujours pour un chromosome de 1330 gènes.

- **Distributions gamma**

Différentes distributions gamma ont été proposées (section III.1). Les distributions gamma (Γ) dont les paramètres de formes (α) sont proches

de 1 sont très similaires à des distributions exponentielles décroissantes de même paramètre d'échelle (θ). Un paramètre de forme $\alpha < 1$ permet d'accentuer la tête de la distribution Γ et de faire plus de petites inversions, alors que le paramètre d'échelle θ peut être augmenté de manière à accroître la densité de probabilité de la queue, dans le but de se rapprocher d'une distribution à « queue lourde » (*heavy tail*). La distribution $\Gamma(\alpha = 0.1, \theta = 800)$ est représentée sur la figure III.2c, tronquée et normalisée, pour correspondre à un chromosome de 1330 gènes.

- **Distributions écartées par les études précédentes**

Parmi les distributions précédentes, il a été montré que la distribution uniforme est incapable de reproduire les tailles d'inversions de bactéries [Lefebvre *et al.*, 2003]. La distribution exponentielle s'est montrée incapable de reproduire les inversions de petits segments de chromosomes de levures [Lefebvre *et al.*, 2003], néanmoins cela n'a pas été prouvé pour les métazoaires.

Nous terminons notre description du fonctionnement de MagSimus en présentant un outil fort pratique pour analyser les réarrangements simulés.

III.3.12 Le gestionnaire de segments conservés et des points de cassures

Durant la simulation il est utile d'enregistrer les extrémités des gènes ancestraux qui flanquent les points de cassures. Quand la simulation est achevée, un bon moyen de retrouver les segments qui ont été conservés est de fragmenter le génome ancestral aux niveaux de toutes les extrémités de gènes qui ont été enregistrées. Les segments conservés peuvent ensuite être filtrés de manière à enlever les gènes ancestraux qui ont été supprimés durant l'évolution. Grâce à ces deux étapes il est possible d'obtenir les segments conservés de chaque lignée. L'obtention des segments conservés simultanément le long de plusieurs lignées se fait selon la méthode illustrée précédemment (figure II.6).

Pour aller encore plus loin, nous avons développé un gestionnaire de segments conservés et de points de cassures. Ce gestionnaire nous informe étape par étape sur l'état des segments conservés et il enregistre également le nombre de réutilisations de points de cassures, le nombre de cassures aux extrémités de chromosomes ainsi que le nombre de segments conservés supprimés à cause des délétions de gènes ancestraux; les variables r , e et d que nous avons négligées dans l'équation précédente (section III.2.6). Nous renvoyons le lecteur aux définitions du chapitre 1 (section I.6) pour la définition

des réutilisations des points de cassures ainsi que pour notre définition des vestiges de flancs de points de cassures.

Nous avons fini de présenter notre méthode pour simuler l'évolution des génomes avec MagSimus. Nous détaillons par la suite l'application de cette méthode à cinq espèces modernes et nous expliquons dans ce cas particulier les choix que nous avons fait pour les paramètres qu'il reste à fixer.

III.4 Simulation de cinq amniotes

Nous avons choisi de simuler l'évolution des génomes de cinq amniotes modernes : l'humain, la souris, le chien, l'opossum et le poulet. Toutes les valeurs suivantes sont calculées sur la base des génomes et des arbres de gènes de la version 81 d'Ensembl. Le poulet nous a semblé être un excellent choix d'espèce out-group étant donné son faible taux de réarrangements [Bourque *et al.*, 2005]. L'opossum est un génome original car il contient peu de chromosomes et ceux-ci sont tous très grands par rapport aux chromosomes des autres espèces. Pour illustrer ce dernier point il suffit de constater que le plus petit chromosome autosomique de l'opossum est le chromosome 7 et qu'il contient environ 260 Mb. Ce dernier est donc plus long que le plus grand chromosome autosomique de l'humain, le chromosome 1 qui contient environ 250 Mb. Enfin, nous avons choisi le génome du chien. La branche qui relie cette espèce moderne à l'arbre se connecte très près de l'ancêtre commun de l'humain et de la souris, voir la figure III.3. Nous pensons ainsi pouvoir mettre en évidence des singularités liées à la branche courte entre Boreoeutheria (l'ancêtre commun du chien et de la souris) et Euarchontoglires (l'ancêtre commun de l'humain et de la souris).

III.4.1 Le génome initial

L'ancêtre commun des cinq espèces modernes que nous avons choisies est Amniota. Comme nous l'avons expliqué, nous utilisons ChromEvol sur un grand arbre qui contient l'arbre des espèces simulées. Le grand arbre phylogénétique constitué à partir des bases de données contient 21 espèces d'amniotes. Nous forçons les taux de fissions et de fusions de ChromEvol à être égaux, le Akaike Information Criterion (AIC) le plus faible est obtenu pour des taux de fissions et de fusions égaux tous deux à environ 0.5 événements par millions d'années. Au final le nombre de chromosomes dans le génome d'amniote est estimé à 21. L'analyse de la forêt d'arbres de gènes aboutit à estimer que le génome de cet ancêtre initial contenait 20252 gènes. Enfin, la distribution des tailles de chromosomes est obtenue en faisant la moyenne des distributions des tailles

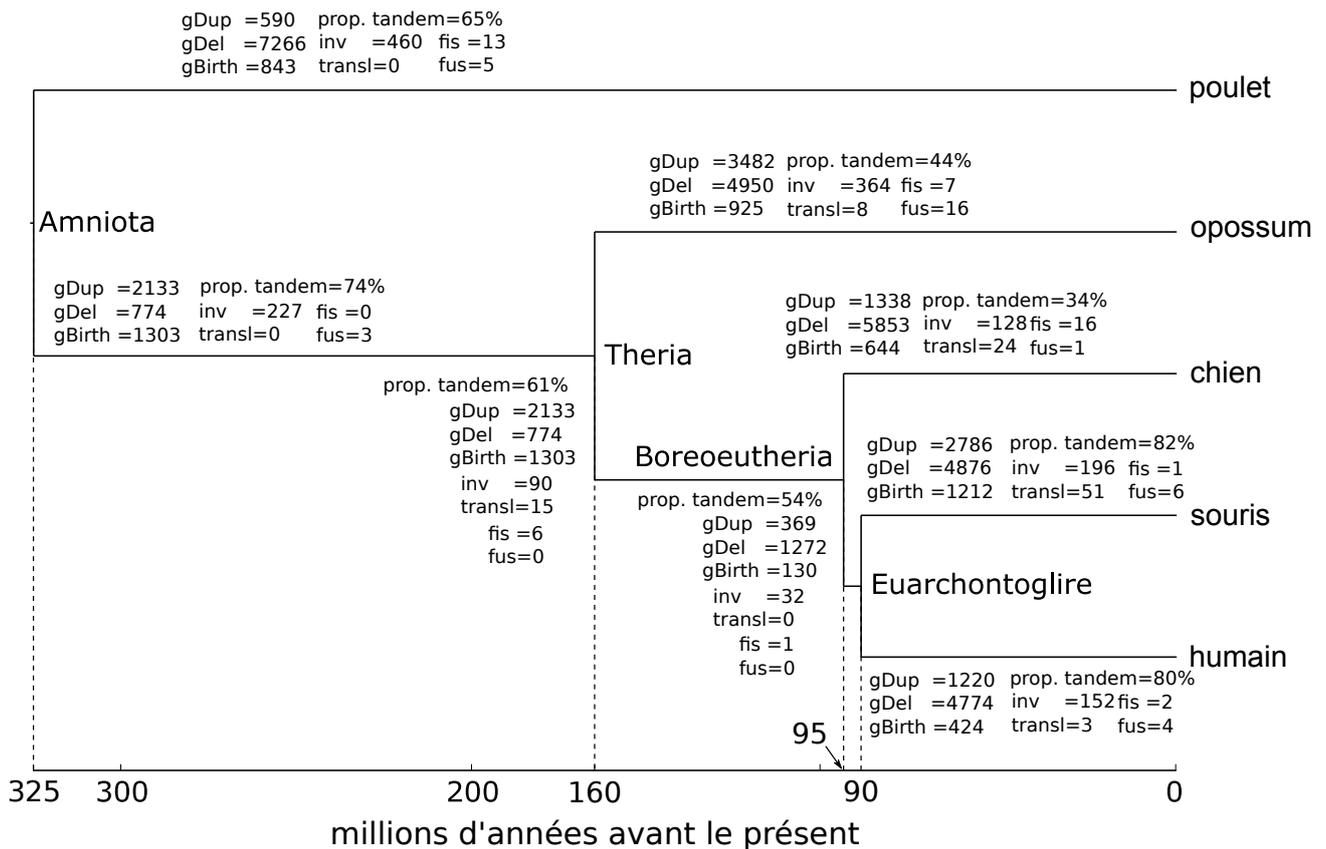


Figure III.3 – Arbre phylogénétique reliant les espèces modernes choisies (humain, souris, chien, opossum et poulet), nombre d'événements et proportions de duplications en tandem sur chaque branche. Les valeurs de cette figure ont été obtenues après une optimisation du simulateur (section III.6) paramétré avec la distribution $\Gamma(\alpha = 0.1, \theta = 800)$ et une taille maximale d'inversion de 1330 gènes.

de chromosomes modernes, conformément au calcul que nous avons expliqué (section III.2.5).

III.4.2 Les nombres et les caractéristiques des évènements géniques

Les nombres d'évènements géniques, les nombres de réarrangements chromosomiques et les proportions de duplications en tandem trouvés par les méthodes précédentes sont représentés sur l'arbre de la figure III.3.

Concernant les duplications en tandem, dans la réalité la statistique s_{reel} ¹ est égale à 69% pour les clusters de duplications en tandem, formés dans le génome humain depuis Amniota, avec un *tandemGapMax* de 15 gènes. La simulation dont la statistique s_{sim} (calculée sur la base des clusters issus de la simulation) est la plus proche de s_{reel} a une probabilité $ptandem_{sim}$ fixée à 75% et la statistique s_{sim} est égale à 68%. D'après nos simulations, la statistique s sous-estime donc la probabilité $ptandem$, c'est probablement le cas dans la réalité et il est raisonnable de croire que $ptandem_{reel} > s_{reel}$. Néanmoins, grâce à l'optimisation, nous pensons que la probabilité $ptandem_{sim} = 75%$, doit être proche de la probabilité $ptandem_{reel}$. De la même manière nous avons trouvé des probabilités $ptandem_{reel}$ de 75% pour les autres lignées donc nous avons considéré que cette probabilité est valable pour toutes les branches. La probabilité la plus éloignée de 75% est celle du poulet (pour laquelle nous avons estimé $ptandem_{reel}$ à 80%) néanmoins, pour simplifier le paramétrage du simulateur, l'écart de 5% a été négligé.

Les branches où il y a le plus d'insertions de gènes sont celles de Theria jusqu'à l'opossum (400-500 insertions), de Theria jusqu'à Boreoeutheria (~3500 insertions) et la branche terminale de la souris (~4000 insertions). Ces nombres élevés d'insertions de nouveaux gènes justifient l'importance de forcer les réarrangements à réarranger des segments de chromosomes qui contiennent au moins un gène ancestral (section III.3.9). De manière assez frappante la branche du poulet, bien qu'elle soit la plus longue contient peu de duplications et peu de naissances de gènes alors qu'il y a de nombreux gènes supprimés.

¹une approximation de $ptandem_{reel}$, la probabilité qu'une copie de gène soit insérée avec la même orientation que le gène copié, voir section III.2.3 pour plus de détails

III.4.3 Mode de sélection des chromosomes hôtes d'une inversion

Quand une inversion a lieu, nous choisissons le chromosome hôte de l'inversion proportionnellement à sa taille. Ainsi la densité des cassures¹ liées aux inversions sera la même entre les petits et les grands chromosomes.

Nous reconnaissons néanmoins que la fréquence des inversions dans un chromosome pourrait être indépendante de la taille du chromosome. Dans nos données nous avons en effet observé sans le quantifier que les petits chromosomes contiennent souvent une densité en inversions plus élevée que les grands. Une analyse quantitative du nombre des segments conservés par tailles de chromosomes modernes pourrait être utile pour choisir moins arbitrairement le mode de sélection des chromosomes hôtes d'une inversion.

III.4.4 Mode de sélection des chromosomes dont la taille varie suite au réarrangement

Reproduire la distribution uniforme des longueurs de chromosomes réels

Lorsque les réarrangements de chromosomes sont simulés, les distributions des tailles des chromosomes simulés sont souvent irréalistes [Sankoff et Ferretti, 1996][Arkendra *et al.*, 2001]. Par rapport aux tailles des chromosomes réels, les grands chromosomes simulés sont trop grands et/ou les petits chromosomes simulés sont trop petits. Nous avons eu les mêmes soucis avec le paramétrage de notre simulateur. Par conséquent, à la suite des deux études que nous avons mentionnées, il nous a semblé judicieux de choisir systématiquement les modes de sélections de chromosomes réarrangés qui tendent à égaliser les tailles des chromosomes.

Sélection des chromosomes transloqués

Si les translocations réciproques se font entre deux chromosomes choisis proportionnellement à leurs tailles [Mazowita *et al.*, 2006], les grands chromosomes s'échangeront souvent des segments de chromosomes. Après de nombreuses translocations, la distribution asymptotique des tailles de chromosomes attendue sera par conséquent une distribution où les grands chromosomes seront très grands et les petits chromosomes très petits. Si au contraire la sélection des chromosomes transloqués est indépendante de leurs tailles, tous les chromosomes ont autant de chance d'échanger un de leur segment par une

¹le nombre de cassures/longueur du chromosome en gène

translocation réciproque. Dans ce deuxième cas la distribution asymptotique sera plus uniforme [Sankoff et Ferretti, 1996] et nous avons donc choisi une sélection uniforme des chromosomes transloqués.

Le génome du poulet a de très grands chromosomes (macrochromosomes) et de nombreux très petits chromosomes (microchromosomes), sa distribution n'a donc rien d'uniforme. Ceci semble contredire notre argumentation précédente, ce point a déjà été soulevé [Sankoff et Ferretti, 1996]. Néanmoins il semble qu'il y ait eu peu de translocations dans la lignée du poulet. Même si ces dernières sélectionnent les chromosomes uniformément, notre choix ne semble donc pas incohérent avec les grands chromosomes et les petits chromosomes du poulet. Il est possible que les petits chromosomes soient majoritairement dus aux nombreuses délétions de gènes (figure III.3). Le faible nombre de translocations peut être confirmé partiellement, par la visualisation de la matrice d'homologies entre le génome du poulet et celui du mandarin (*Taeniopygia guttata*), qui a divergé il y a approximativement 104 millions d'années de la lignée du poulet (figure III.4). À quelques exceptions, dans les grands chromosomes, les gènes sont restés synténiques et il semble qu'il n'y ait eu presque que des inversions. D'autres travaux trouvent également que la synténie est bien conservée entre les oiseaux [Green *et al.*, 2014].

Mode de sélection des chromosomes fusionnés et fissionnés

Enfin les modes de sélection des chromosomes lors des fusions et des fissions sont eux aussi choisis dans le but d'uniformiser les distributions de tailles de chromosomes. Les fissions sélectionnent les chromosomes brisés proportionnellement à leurs tailles et les chromosomes fusionnés sont sélectionnés uniformément.

III.4.5 Tailles limites des chromosomes

Dans les travaux précédents [Sankoff et Ferretti, 1996][Arkendra *et al.*, 2001] la sélection uniforme des chromosomes transloqués n'expliquait pas entièrement la tendance uniforme des distributions de tailles de chromosomes modernes. Pour mieux l'expliquer différentes hypothèses ont été faites : imposer une taille minimale [Sankoff et Ferretti, 1996], imposer une taille maximale et définir une fitness pour chaque taille de chromosome [Arkendra *et al.*, 2001].

L'opossum a de très grands chromosomes (son chromosome 1 contient plus 748 Mb et 4441 gènes) et le poulet a de très petits chromosomes (son chromosome 32 est long de quelques Mb, pas plus). Les micro-chromosomes du poulet contredisent quelque peu l'hypothèse d'une taille minimale de chromosome viable [Sankoff et Ferretti, 1996]. Ou alors si cette limite existe il

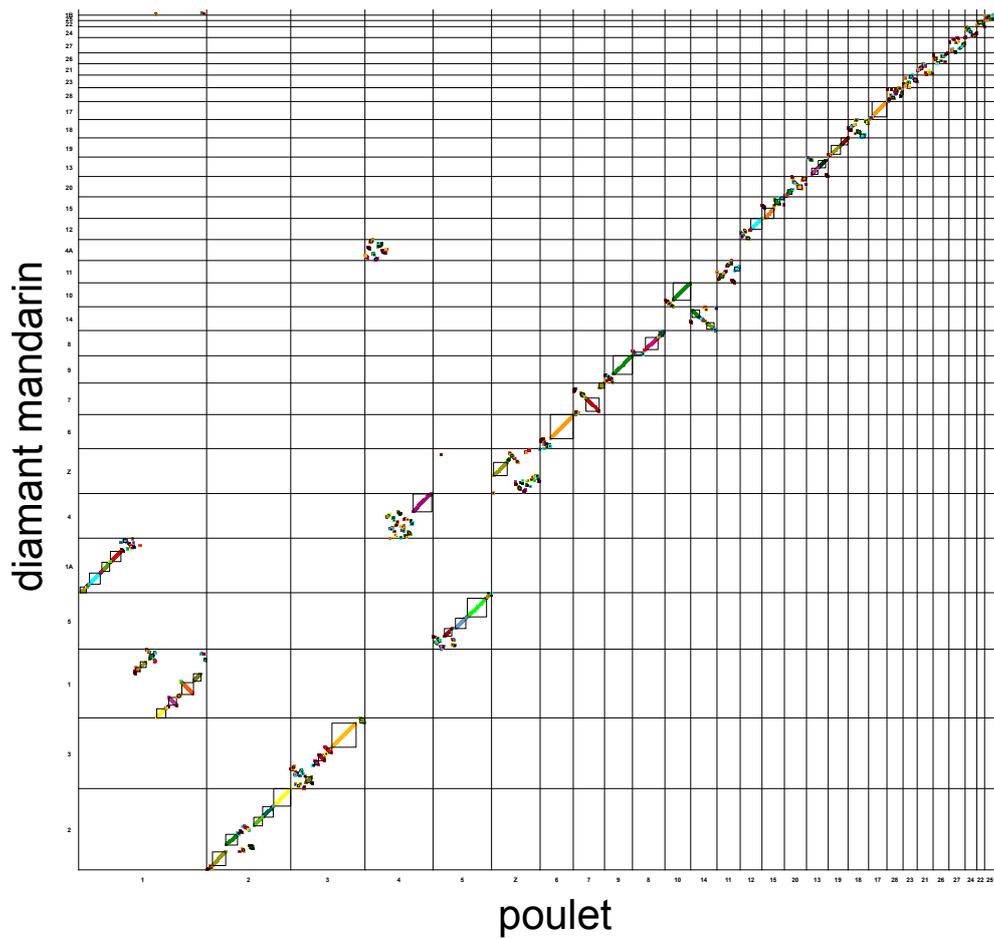


Figure III.4 – Matrice d’homologies des génomes du poulet et du diamant mandarin (zebra finch).

semble qu'elle soit très petite. De manière analogue les longs chromosomes de l'opossum, bien plus longs que les chromosomes des autres espèces, montrent qu'il ne semble pas non plus y avoir de taille maximale de chromosome [Arkendra *et al.*, 2001]. Ce qui semble aller à l'encontre du raisonnement de [Schubert et Oud, 1997] basé sur la longueur maximale de l'axe du fuseau mitotique à la télophase, à moins que les cellules d'opossum soient plus larges que les cellules des autres espèces. Nous avons donc décidé de ne pas limiter la taille maximale d'un chromosome. Dans nos simulations, seule la taille minimale d'un chromosome a été fixée à un gène pour éviter que des chromosomes disparaissent suite à des délétions de gènes. Ce dernier point permet de s'assurer que seules les fusions et les fissions font varier les nombres de chromosomes.

III.4.6 Affinage de l'inférence du nombre d'inversions et de translocations réciproques

À partir des génomes modernes simulés (et non plus des génomes modernes réels) nous avons effectué à nouveau les inférences des réarrangements simulés. Les nombres de chromosomes ancestraux et les nombres de fusions et fissions sont inférés avec ChromEvol sans aucune différence de résultat par rapport à précédemment, car les nombres de chromosomes modernes sont exactement égaux aux nombres de chromosomes réels. Par contre lorsque nous utilisons les génomes simulés pour inférer les translocations et les inversions (section III.2.6) nous trouvons des nombres de réarrangements par branches parfois très éloignés de ceux que nous avons fixés comme paramètres. De manière générale nous inférons substantiellement moins d'inversions à partir des génomes simulés que nous n'inférons d'inversions à partir des génomes réels. Nous expliquons cette différence par les délétions de gènes ancestraux. Dans l'équation (III.8), nous avons sous-estimé le nombre de réutilisations de points de cassure (r), le nombre de points de cassures aux extrémités de chromosomes (e) ainsi que le nombre de segments conservés perdus à cause des délétions des gènes ancestraux (d)¹. Or, bien que r et e soient effectivement négligeables dans nos simulations, le gestionnaire de segments conservés et de points de cassures (section III.3.12) nous informe que d ne l'est pas dans de nombreuses lignées. Ce qui était prévisible étant donné les nombreuses délétions dans les branches, figure III.3.

Pour affiner l'estimation des nombres d'inversions et de translocations en

¹Nous rappelons qu'un segment conservé ne contenant qu'un gène ancestral est perdu, si son unique gène ancestral est supprimé. Par conséquent, les délétions de gènes influent sur le nombre de segments conservés.

présence de nombreuses délétions de gènes ancestraux nous avons effectué d'autres simulations et d'autres estimations avec une boucle de rétroaction sur le nombre d'inversions et de translocations simulées. Le schéma de la figure III.5 représente cette boucle. Pour accélérer la convergence et limiter les

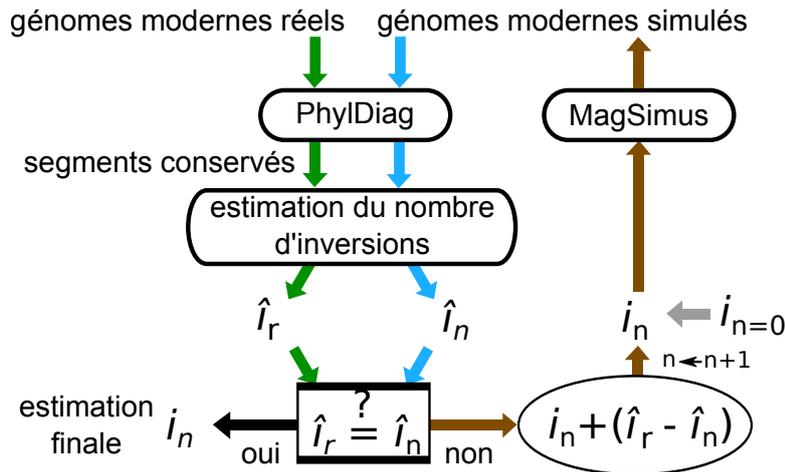


Figure III.5 – Boucle de rétroaction pour affiner l’estimation des nombres d’inversions sur une branche. Ici, seuls les inversions sont montrées pour ne pas surcharger la figure. En réalité l’affinage des translocations est fait simultanément. Le processus commence par la flèche grise avec l’initialisation du nombre d’inversions simulées, i_0 , itération $n = 0$. Les génomes simulés sont analysés par PhylDiag et à partir des segments conservés nous estimons, \hat{i}_n , le nombre d’inversions qui ont eu lieu (flèches bleues). Ce nombre est comparé à, \hat{i}_r , l’estimation du nombre d’inversions réelles (flèches vertes). Si ces deux estimations diffèrent de plus de 1% un nouveau nombre d’inversions simulées est calculé, $i_{n+1} = i_n + (\hat{i}_r - \hat{i}_n)$, pour réduire l’écart (la boucle de rétro-action en marron). Une nouvelle itération peut commencer, avec ce nouveau nombre en paramètre. Quand l’écart est suffisamment faible ou quand le nombre d’itérations dépasse une limite fixée par l’utilisateur, le processus d’affinage s’arrête et le nombre d’inversions simulées lors de la dernière itération correspond à l’estimation finale du nombre d’inversions.

itérations, le nouveau paramètre spécifiant le nombre d’inversions simulées est plus précisément calculé de la manière suivante : $i_n + (\hat{i}_r - \hat{i}_n)\alpha$ avec $\alpha = \frac{\hat{i}_r}{\hat{i}_n}$ la proportion des inversions effectivement réalisées par rapport à l’estimation des inversions simulées. Les translocations sont calculées simultanément aux inversions. Lors de l’estimation, de la comparaison et du calcul des nombres d’évènements simulés, les nombres d’inversions et de translocations de toutes les branches sont traités indépendamment et de la même manière. À chaque

fois les calculs précédents sont exécutés en faisant des moyennes sur 100 simulations pour limiter les fluctuations dues au fonctionnement stochastique du simulateur.

III.4.7 Choix de la distribution des tailles d'inversions

De nombreuses distributions de tailles de segments inversés sont disponibles (section III.3.11) : la distribution inverse, la distribution gamma avec différents paramètres de forme et d'échelle, ... et nous en avons testé plusieurs. Pour chaque distribution de tailles d'inversions, les nombres d'inversions et de translocations par branches ont été affinés par la boucle de rétroaction décrite ci-dessus. Il est nécessaire de faire l'affinage pour chaque distribution de taille d'inversions pour la raison suivante. Par exemple, si nous étudions deux distributions de tailles d'inversions et si, dans la première, les inversions courtes sont plus probables que dans la deuxième distribution, les petites inversions généreront plus de petits segments conservés et ceux-ci seront donc plus facilement supprimés par les délétions de gènes. Il y aura au final plus de segments conservés supprimés avec la première distribution qu'avec la deuxième. Par conséquent, avec la première distribution, le calcul d'estimation du nombre d'inversions (section III.2.6) sous-estimera plus le nombre d'inversions simulées. L'estimation finale du nombre d'inversions réelles, après optimisation, sera donc plus élevée si la première distribution est choisie.

Il s'ensuit qu'il existe un paramétrage (nombres d'inversions simulées, nombre de translocations simulées) de MagSimus pour chaque distribution des tailles d'inversions. Nous calculerons un score de réalisme pour chacun de ces paramétrages et la distribution des tailles d'inversions la plus réaliste correspondra au paramétrage qui générera les simulations les plus réalistes. La section suivante explique comment nous quantifions le réalisme d'un paramétrage de MagSimus.

III.5 Écart d'une simulation par rapport à la réalité

Notre simulateur est conçu pour reproduire une évolution qui se rapproche le plus possible de la réalité. Néanmoins l'évolution des génomes étant complexe, des différences sont attendues. La quantification de cet écart restant par rapport à la réalité est le sujet de ce chapitre.

Rappelons que nous n'ambitionnons pas de reproduire à l'identique l'évolution réelle. Il serait donc absurde de quantifier l'écart à la réalité en

comparant directement l'ordre des gènes dans un génome moderne simulé avec l'ordre des gènes dans un génome moderne réel. Par exemple cela n'a pas de sens de comparer les adjacences de gènes d'un génome simulé avec les adjacences de gènes du génome réel correspondant. Cela n'a pas de sens car un gène simulé ne correspond pas à un gène réel et aucun réarrangement chromosomique simulé ne correspond à un réarrangement qui a réellement eu lieu, les points de cassures simulés ne sont pas localisés aux endroits précis où les points de cassures réels ont eu lieu.

Ceci étant dit, l'écart d'une simulation à la réalité peut se baser sur d'autres critères. Par exemple un critère pourrait être le nombre de gènes, ou le nombre de chromosomes, la distribution des tailles de chromosomes, le nombre de segments conservés, la distribution des tailles de segments conservés, *etc.* Plus ces statistiques issues d'une simulation seront proches des statistiques réelles, plus nous dirons que la simulation est réaliste. L'écart de ces statistiques avec les valeurs réelles constitue une erreur, une imperfection de la simulation. Nous allons expliciter la signification de cette erreur.

De manière générale, nous nommerons *erreur* d'une simulation, l'écart d'une simulation par rapport à la réalité. Nous considérerons différents types d'erreurs selon le contexte, comme cela sera expliqué plus tard. Néanmoins toute erreur sera à chaque fois issue de la comparaison d'une donnée simulée avec une donnée réelle. Ainsi, l'erreur sur les nombres de chromosomes comparera le nombre de chromosomes simulés avec le nombre de chromosomes réels. À la fin de ce chapitre nous serons en mesure de définir une erreur générale quantifiée précisément. Grâce à cette quantification nous aurons un critère pour classer les simulations des plus réalistes aux moins réalistes, la simulation la plus réaliste étant celle dont l'erreur générale sera la plus faible.

Considérons un ensemble de N espèces modernes, numérotées de 1 à N , dont nous connaissons les génomes. Ces espèces ont un MRCA, nommé *Anc*. L'évolution réelle est une boîte noire et seuls sont connus:

- **l'arbre des espèces** dont la racine est *Anc* et les feuilles sont les N espèces modernes. Les longueurs des branches de l'arbre des espèces correspondent aux temps évolutifs entre les différents événements de spéciations ou feuilles. Nous considérons que les longueurs de branches correspondent aux véritables durées évolutives et qu'il n'y a pas d'erreur dans la datation des événements de spéciation.
- **la forêt d'arbres de gènes** pour les gènes qui descendent d'un gène qui était présent dans le génome *Anc* ainsi que pour les gènes qui sont nés après *Anc* le long d'une branche de l'arbre des espèces. En réalité la forêt d'arbres de gènes est inférée par *Ensembl* et contient certainement

de nombreuses erreurs. Néanmoins, nous partons du principe que les arbres sont suffisamment réalistes pour pouvoir nous appuyer dessus avec confiance comme s'il s'agissait de la véritable forêt d'arbres de gènes.

- **les génomes annotés des N espèces modernes.** Chaque génome est un ensemble de chromosomes et chaque chromosome est une liste de gènes orientés.

L'arbre des espèces est fixé pour toutes les simulations. Chaque simulation considère la même topologie et les mêmes longueurs de branches en millions d'années. Le nombre d'évènements géniques est lui aussi fixé dans toutes les simulations à partir des nombres d'évènements dans les arbres de gènes réels. Nous considérons donc qu'il n'y a aucune erreur en ce qui concerne l'arbre des espèces dans les simulations, et qu'il n'y a également aucune erreur en ce qui concerne le nombre d'évènements géniques par rapport à la forêt d'arbres de gènes que nous considérons comme réelle.

Il reste à mettre en place une méthode de quantification de l'erreur en ce qui concerne les autres statistiques, par exemple le nombre de chromosomes dans les espèces modernes, les tailles des segments conservés, *etc.* C'est l'objet des prochaines sous-sections.

III.5.1 Statistiques pour un génome moderne

Des statistiques caractéristiques d'un génome moderne, par exemple le génome de l'espèce 1, sont

- c^1 , le **nombre de chromosomes** dans le génome 1
- et γ^1 la **distribution des tailles de chromosomes** dans le génome 1.

Ces statistiques sont calculées de la même manière pour les génomes modernes réels et les génomes modernes simulés.

III.5.2 Statistiques pour une comparaison de deux génomes modernes

Des statistiques caractéristiques d'une comparaison de deux génomes modernes, par exemple le génome de l'espèce 1 et le génome de l'espèce 2, sont

- $b^{1,2}$, le **nombre de segments conservés** entre le génome 1 et le génome 2

- et $\beta^{1,2}$, la **distribution des tailles des segments conservés** entre le génome 1 et le génome 2.

Encore une fois les statistiques sont calculées de la même manière pour les comparaisons de génomes réels et pour les comparaisons de génomes simulés.

III.5.3 Erreur de réalisme pour une statistique

Comparons les statistiques précédentes calculées sur les génomes simulés avec les mêmes statistiques calculées sur les génomes réels.

Considérons la statistique σ , par exemple c^1 , et deux valeurs de cette statistique; l'une calculée à partir de la réalité et l'autre calculée à partir d'une simulation. Nous les noterons σ_r et σ_s . Il pourrait s'agir, par exemple, de c_r^1 et c_s^1 .

Erreur de réalisme pour une statistique scalaire

Si la statistique σ correspond à une valeur scalaire (par exemple un entier), alors σ_r et σ_s sont des scalaires et l'erreur de réalisme absolue est notée

$$\Delta(\sigma_s, \sigma_r) = \sigma_s - \sigma_r. \quad (\text{III.14})$$

Par la suite nous la nommerons simplement *erreur absolue* de la statistique σ . Par exemple, supposons qu'une simulation ait été effectuée et que le nombre de segments conservés dans la comparaison des génomes simulés 1 et 2 soit égal à 473; avec la notation précédente nous écrirons $b_s^{1,2} = 473$. Nous voulons estimer l'erreur de cette statistique par rapport à la réalité. Pour cela nous la comparons à la statistique réelle, $b_r^{1,2}$, le nombre de segments conservés dans la comparaison des génomes réels 1 et 2. Si $b_r^{1,2} = 483$, l'erreur absolue de la statistique simulée sera tout simplement $\Delta(b_s^{1,2}, b_r^{1,2}) = 473 - 483 = -10$. Autrement dit, dans la simulation il y a dix segments conservés de moins que dans la réalité.

L'erreur de réalisme peut aussi être calculée de manière relative, sa valeur est alors

$$\rho(\sigma_s, \sigma_r) = \frac{\sigma_s}{\sigma_r} \quad (\text{III.15})$$

si σ_r est non nul. Si $\sigma_r = 0$ il y a deux cas : soit $\sigma_s = 0$ et $\rho(\sigma_s, \sigma_r) = 1$, soit $\sigma_s \neq 0$ et alors $\rho(\sigma_s, \sigma_r) = \sigma_s$. Cette fois ci nous nommons cette erreur, l'*erreur relative* de la statistique σ . En reprenant l'exemple précédent, $\rho(\sigma_s, \sigma_r) = \frac{\sigma_s}{\sigma_r} = \frac{473}{483} = 98\%$. Nous verrons que lorsque nous rassemblerons les différentes erreurs pour n'en avoir au final plus qu'une, l'erreur relative sera très pratique pour sélectionner la simulation la plus réaliste [Fleming et Wallace, 1986].

Erreur de réalisme lorsque la statistique est une distribution

Si la statistique σ est une distribution, alors σ_r et σ_s ne sont plus des entiers mais des distributions et l'erreur de réalisme ne peut pas être calculée aussi simplement que précédemment.

Cette fois-ci, nous calculons l'erreur absolue d'une manière analogue au calcul de la statistique de Kolmogorov-Smirnov (KS) :

$$\Delta(\sigma_s, \sigma_r) = cdf_s(X) - cdf_r(X) \quad (\text{III.16})$$

avec cdf_s une cdf qui cumule la distribution σ_s et cdf_r la cdf de la distribution σ_r et

$$X = \operatorname{argmax}_x |cdf_s(x) - cdf_r(x)|. \quad (\text{III.17})$$

Par exemple, dans le cas de la distribution des tailles de chromosomes de l'humain ($\sigma = \gamma^h$), $cdf_s(x)$ est égal au nombre de chromosomes simulés de tailles inférieures ou égales à x , $cdf_r(x)$ est égal au nombre de chromosomes réels de tailles inférieures ou égales à x et X est la taille de chromosome x pour laquelle l'écart entre $cdf_s(x)$ et $cdf_r(x)$ est maximal.

L'erreur relative est calculée de la manière suivante

$$\rho(\sigma_s, \sigma_r) = \frac{cdf_s(W)}{cdf_r(W)} \quad (\text{III.18})$$

avec

$$W = \operatorname{argmax}_w \left\langle \frac{cdf_s(w)}{cdf_r(w)} \right\rangle \quad (\text{III.19})$$

et

$$\langle \bullet \rangle :]0, \infty[\rightarrow [1, \infty[\quad (\text{III.20})$$

$$w \mapsto \begin{cases} w & \text{si } w \geq 1 \\ \frac{1}{w} & \text{sinon} \end{cases}$$

la fonction équivalente à la valeur absolue pour les ratios. Pour les valeurs de w où $cdf_r(w) = 0$, le dénominateur de l'équation III.19 est choisi égal à 1. De même, pour les valeurs de w où $cdf_s(w) = 0$, le numérateur est choisi égal à 1.

Avec la définition de l'erreur absolue et de l'erreur relative il est maintenant possible de quantifier, de deux manières, l'erreur de réalisme d'une statistique.

III.5.4 Erreur de réalisme moyenne pour un paramétrage du simulateur

Notre simulateur comporte une part d'aléatoire, comme, par exemple, le choix des positions des points de cassures. Par conséquent le même paramétrage

peut générer deux simulations avec des valeurs différentes pour la même statistique. Ainsi, pour un paramétrage donné, une première simulation peut avoir une statistique $b_1^{1,2} = 500$ alors qu'une deuxième simulation peut avoir $b_2^{1,2} = 530$. Aucune simulation n'est plus légitime qu'une autre pour le calcul des statistiques et au final pour le calcul de l'erreur. En bref, un même paramétrage peut générer des simulations avec différentes erreurs de réalisme et il n'y a pas de raison d'en choisir une plutôt qu'une autre. Pour éviter de choisir arbitrairement une simulation lors de l'estimation de l'erreur d'un jeu de paramètres, nous moyennons les erreurs de 100 simulations, chacune paramétrée avec le paramétrage dont nous souhaitons évaluer l'erreur.

Fixons un paramétrage, effectuons 100 simulations et calculons l'erreur pour la statistique σ . Si $\Delta_1, \dots, \Delta_{100}$ sont les erreurs absolues de cette statistique pour les 100 simulations, alors l'erreur absolue du paramétrage sera la moyenne arithmétique

$$\Delta = \overline{\Delta_i} = \frac{\sum_{i=1}^{100} \Delta_i}{100}. \quad (\text{III.21})$$

De même, si $\rho_1, \dots, \rho_{100}$ sont les erreurs relatives de cette statistique, le paramétrage aura une erreur relative

$$\rho = GM(\rho_i) = \sqrt[100]{\prod_{i=1}^{100} \rho_i}, \quad (\text{III.22})$$

la moyenne géométrique des ρ_i [Fleming et Wallace, 1986]. Le calcul de l'erreur de réalisme du paramétrage est effectué de la même manière pour toutes les statistiques. Dans la suite nous noterons simplement $\Delta(\sigma)$ (resp. $\rho(\sigma)$), l'erreur absolue (resp. relative), du paramétrage pour la statistique σ . Il est évident que lorsque la statistique σ est une valeur, $\Delta(\sigma) = \overline{\sigma_{s,i}} - \sigma_r$ et $\rho(\sigma) = \frac{GM(\sigma_{s,i})}{\sigma_r}$.

Pour rendre ces valeurs plus concrètes, étudions le nombre de chromosomes dans le génome humain (h), en d'autres termes considérons la statistique c^h . Dans la réalité le génome humain a 23 chromosomes, $c_r^h = 23$. Effectuons 100 simulations avec le même paramétrage et notons $c_{s,i}^h$ le nombre de chromosomes de la $i^{\text{ème}}$ simulation. L'erreur absolue de la $i^{\text{ème}}$ simulation est alors $\Delta_i = \Delta(c_{s,i}^h, c_r^h) = c_{s,i}^h - 23$ et, au final, l'erreur absolue du paramétrage est $\Delta(c^h) = \frac{\sum_{i=1}^{100} \Delta_i}{100} = \frac{\sum_{i=1}^{100} c_{s,i}^h}{100} - 23$. Parallèlement, l'erreur relative de la $i^{\text{ème}}$ simulation est $\rho_i = \rho(c_{s,i}^h, c_r^h) = \frac{c_{s,i}^h}{23}$ et l'erreur relative du paramétrage est égal à $\rho(c^h) = \sqrt[100]{\prod_{i=1}^{100} \rho_i} = \frac{\sqrt[100]{\prod_{i=1}^{100} c_{s,i}^h}}{23}$.

III.5.5 Erreur de réalisme synthétique pour toutes les espèces modernes

Nous avons calculé les erreurs du paramétrage pour une espèce moderne donnée (resp. pour une comparaison de deux espèces modernes) et nous aimerions avoir une erreur de réalisme qui agglomère les scores de toutes les espèces (resp. de toutes les comparaisons d'espèces). Ainsi, dans la section précédente, nous avons calculé $\Delta(c^1)$, l'erreur de réalisme absolu du nombre de chromosomes dans l'espèce 1 et nous aimerions avoir une unique valeur pour quantifier l'erreur absolue des nombres de chromosomes dans toutes les espèces modernes, Δ_c . Nous calculerons pour cela la moyenne arithmétique des $|\Delta(c^i)|$, avec chaque i correspondant à une espèce moderne. Nous faisons ici la moyenne des valeurs absolues pour éviter qu'un écart à la réalité positif soit compensé par un écart négatif. D'où, s'il y a N espèces modernes,

$$\Delta_c = \overline{|\Delta(c^i)|} = \frac{1}{N} \sum_{i=1}^N |\Delta(c^i)|. \quad (\text{III.23})$$

De la même manière l'erreur de réalisme absolue des distributions des tailles de chromosomes sera $\Delta_\gamma = \overline{|\Delta(\gamma^i)|}$.

Le calcul diffère un peu pour les statistiques qui ont trait aux comparaisons de deux espèces modernes, la moyenne arithmétique ne portera plus sur les N espèces modernes, mais sur le total des $\binom{N}{2}$ comparaisons de deux espèces. Ainsi l'erreur de réalisme absolue des nombres de segments conservés dans toutes les comparaisons de deux espèces modernes Δ_b est la moyenne arithmétique des $|\Delta(b^{i,j})|$ pour toutes les comparaisons de deux espèces modernes (i, j) . Le calcul est le même pour Δ_β et les distributions des tailles de segments conservés.

Dans le cas des statistiques c et γ , faire la moyenne sur les espèces revient à dire que chaque espèce a autant de poids qu'une autre espèce dans le calcul l'erreur finale. De même, dans le cas des statistiques b et β , faire la moyenne sur les comparaisons donne à chaque comparaison de deux espèces autant de poids qu'une autre comparaison de deux espèces.

Encore une fois, le calcul analogue est effectué pour les erreurs synthétiques relatives. Par exemple

$$\rho_c = GM(\langle \rho(c^i) \rangle) = \sqrt[N]{\prod_{i=1}^N \langle \rho(c^i) \rangle}. \quad (\text{III.24})$$

Ici encore, nous faisons la moyenne des écarts édités avec la fonction $\langle \bullet \rangle$ (équation III.19) pour éviter qu'un écart relatif supérieur à 1 soit compensé par un écart relatif inférieur à 1.

Δ_c ne correspond plus à une statistique précise, comme c'était le cas avec $\Delta(c^i)$. Néanmoins Δ_c peut s'interpréter facilement. Si $\Delta_c = 2$, le paramétrage du simulateur génère des génomes modernes dont les nombres de chromosomes diffèrent, en moyenne, de 2 chromosomes, en plus ou en moins, par rapport aux nombres réels de chromosomes. De même, si $\rho_c = 10\%$, il y a dans les génomes modernes simulés, des nombres de chromosomes qui diffèrent, en moyenne, de 10% (là encore, en plus ou en moins) par rapport aux nombres de chromosomes réels correspondants.

III.5.6 Erreur générale de réalisme

Nous avons calculé plusieurs erreurs de réalisme pour un paramétrage de notre simulateur. Au début les erreurs de réalisme correspondaient à une statistique focalisée sur une espèce ou bien à une statistique focalisée sur une comparaison de deux espèces. Dans la section précédente, nous avons regroupé ces erreurs en erreurs synthétiques qui, bien qu'elles ne correspondent plus à une statistique, conservent néanmoins une signification précise. Ainsi, pour un paramétrage donné, en changeant un peu la notation, nous avons calculé

- Δ_c l'erreur absolue en nombre de chromosomes,
- Δ_γ l'erreur absolue des distributions des tailles de chromosomes,
- Δ_b l'erreur absolue en nombre de segments conservés,
- et Δ_β l'erreur absolue des distributions des tailles de segments conservés.

Nous aimerions une erreur générale qui soit une fonction de ces 4 erreurs. Comme il y a en général beaucoup plus de segments conservés que de chromosomes, nous nous attendons à avoir $\Delta_c \ll \Delta_b$ et $\Delta_\gamma \ll \Delta_\beta$. Si l'erreur générale est la moyenne des valeurs absolues de ces 4 erreurs synthétiques, c'est à dire si

$$\Delta = \frac{1}{4} (|\Delta_c| + |\Delta_\gamma| + |\Delta_b| + |\Delta_\beta|), \quad (\text{III.25})$$

il faut s'attendre à ce que la majeure partie de l'erreur soit due à $|\Delta_b|$ et $|\Delta_\beta|$. $|\Delta_c|$ et $|\Delta_\gamma|$ pourront varier du simple au double sans que cela soit visible dans l'erreur générale, alors que nous aimerions que cela ait un effet significatif, pour discriminer une simulation par rapport à une autre. En effet, si pour un premier paramétrage $|\Delta_c| = 1$ et $|\Delta_b| = 30$ et si pour un deuxième paramétrage $|\Delta_c| = 4$ et $|\Delta_b| = 27$; toutes autres erreurs égales par ailleurs, faire la moyenne des erreurs, ne permettra pas de distinguer les deux paramétrages. Or nous aimerions que l'erreur nous indique que le deuxième paramétrage est moins réaliste que le premier, car, du premier au deuxième

paramétrage, $|\Delta_c|$ a quadruplé alors que $|\Delta_b|$ est resté dans le même ordre de grandeur. C'est pourquoi nous éviterons d'utiliser les erreurs absolues pour calculer l'erreur générale¹. Nous utiliserons de préférence les erreurs relatives, comme nous allons l'expliquer.

Le calcul avec les erreurs relatives est très analogue. Comme précédemment, nous avons, pour un paramétrage donné

- ρ_c l'erreur relative en nombre de chromosomes,
- ρ_γ l'erreur relative des distributions des tailles de chromosomes,
- ρ_b l'erreur relative en nombre de segments conservés,
- et ρ_β l'erreur relative des distributions des tailles de segments conservés.

Cette fois-ci, au lieu de calculer une moyenne arithmétique, nous calculerons la moyenne géométrique. L'erreur générale sera donc

$$\rho = \sqrt[4]{\langle \rho_c \rangle \langle \rho_\gamma \rangle \langle \rho_b \rangle \langle \rho_\beta \rangle}. \quad (\text{III.26})$$

Si un premier paramétrage du simulateur commet une erreur générale deux fois plus importante qu'un deuxième paramétrage, c'est qu'en moyenne (géométrique) les erreurs synthétiques relatives, et éditées, les $\langle \rho_\bullet \rangle$, sont deux fois plus importantes. Par exemple, un paramétrage dont toutes les erreurs synthétiques éditées sont deux fois plus importantes aura une erreur générale deux fois supérieure, $\rho' = \sqrt[4]{2\langle \rho_c \rangle 2\langle \rho_\gamma \rangle 2\langle \rho_b \rangle 2\langle \rho_\beta \rangle} = 2\rho$. Si nous voulons donner plus d'importance à une erreur relative qu'à une autre il suffira de modifier la formule de l'erreur générale de la manière suivante :

$$\rho = \sqrt[4]{\langle \rho_c \rangle^{w_c} \langle \rho_\gamma \rangle^{w_\gamma} \langle \rho_b \rangle^{w_b} \langle \rho_\beta \rangle^{w_\beta}} \quad (\text{III.27})$$

avec w_x le poids de l'erreur correspondante, tel que : $w_c + w_\gamma + w_b + w_\beta = 4$. Ainsi, si nous souhaitons donner deux fois plus de poids à $\langle \rho_c \rangle$ qu'aux 3 autres erreurs, nous utiliserons la formule $\rho = \sqrt[4]{\langle \rho_c \rangle^{\frac{8}{5}} \langle \rho_\gamma \rangle^{\frac{4}{5}} \langle \rho_b \rangle^{\frac{4}{5}} \langle \rho_\beta \rangle^{\frac{4}{5}}}$.

Avec une définition claire du calcul du réalisme d'une simulation, via l'erreur générale de réalisme, nous pouvons aborder la mise en œuvre du simulateur en ayant continuellement à cœur le but final : générer des simulations les plus réalistes possibles.

¹Les incohérences liées à l'usage de la moyenne arithmétique dans un cas analogue sont explicitées plus en détails dans [Fleming et Wallace, 1986].

III.6 Résultats de MagSimus

Nous avons testé de nombreuses distributions de tailles de segments inversés et nous détaillons les résultats pour la distribution avec laquelle nous avons obtenu nos meilleurs résultats. Il s'agit une distribution gamma avec un paramètre de forme $\alpha = 0.1$, un paramètre d'échelle $\theta = 800$ gènes et une longueur maximale d'inversion égale à 1330 gènes (figure III.2d). Les nombres d'évènements que nous avons obtenus sont visibles sur l'arbre de la figure III.3.

Les écarts à la réalité ont été quantifiés grâce aux calculs précédents (section III.5) et quelques exemples d'écarts sont donnés dans le tableau III.1. À titre de comparaison, les écarts à la réalité obtenus avec les mêmes paramètres et une distribution uniforme des tailles d'inversions sont également fournis. Les écarts types sont les écarts types classiques lorsque le calcul de la moyenne est arithmétique (c'est le cas pour les écarts absolus, les Δ), ce sont les écarts types géométriques lorsque les moyenne correspondantes sont calculée de manière géométrique [Kirkwood, 1979] (pour les écarts relatifs, les ρ).

Les écarts en nombres de chromosomes (statistiques c) sont nuls. Ce résultat était attendu car nous avons fixé les nombres de fusions et de fissions de manière à obtenir le même nombre de chromosomes que dans la réalité. De plus la taille minimale des chromosomes a été fixée à 1 gène pour qu'aucun chromosome ne soit supprimé à cause des délétions géniques. Les écarts en nombres de segments conservés (b) sont quasi-nuls car l'optimisation sur le nombre d'inversions et de translocation nous assure que les nombres de segments correspondent à la réalité. Par contre l'écart des distributions des tailles de segments conservés (β) varie fortement selon la distribution des inversions. Comme attendu, la distribution uniforme ne fait pas assez de petites inversions et par conséquent ρ_β est très petit. La distribution Γ génère beaucoup plus de petites inversions (III.2d) et par conséquent la distribution des tailles de ses segments conservés est plus proche de la distribution réelle.

Le tableau III.1 nous montre que quantitativement, le défaut majeur de notre simulateur concerne la distribution des tailles de chromosomes simulés, car dans les deux cas, ρ_γ est au moins égal à 3.8. Prenons le cas de la distribution des tailles de chromosomes de l'humain (γ^h). L'égalité $\rho(\gamma^h) = 4.19$ peut s'interpréter par le fait qu'il y a, en moyenne, 4.19 fois plus de petits chromosomes (de tailles $\leq W$) dans les simulations que dans la réalité. Par conséquent, malgré nos choix de sélections de chromosomes réarrangés, de manière à uniformiser la distribution des tailles de chromosomes (section III.4.4), nos simulations ont encore des petits chromosomes trop petits par rapport à la réalité. Par ailleurs, les grands chromosomes simulés sont la plupart du temps trop grands par rapport aux grands chromosomes réels

	ρ	ρ_c	ρ_γ	ρ_b	ρ_β	$\Delta(c^h)$	$\rho(\gamma^h)$	$\Delta(b^{h,s})$	$\rho(\beta^{h,s})$
moyenne	2.37	1.01	3.84	1.01	8	0	3.26	-0.28	0.02
écart type	2.44	1.01	1.96	1.01	2.59	0	1.78	7.46	1.83

(a) Écarts à la réalité avec une distribution uniforme des tailles d'inversions.

	ρ	ρ_c	ρ_γ	ρ_b	ρ_β	$\Delta(c^h)$	$\rho(\gamma^h)$	$\Delta(b^{h,s})$	$\rho(\beta^{h,s})$
moyenne	1.52	1.01	4.14	1.01	1.27	0	4.19	-0.63	0.8
écart type	1.8	1.01	1.89	1.01	1.11	0	1.45	13.82	1.05

(b) Écarts à la réalité avec des tailles d'inversions données par la loi de probabilité $\Gamma(\alpha = 0.1, \theta = 800)$ tronquée pour que la taille maximale des inversions soit de 1330 gènes.

Tableau III.1 – Écarts à la réalité de deux paramétrages du simulateur. La nomenclature des écarts correspond à celle que nous avons définie précédemment (section III.5). Les exposants h et s font référence à l'humain et à la souris. ρ , l'écart relatif général, est plus proche de 1 avec la distribution Γ qu'avec la distribution uniforme; ce qui signifie, qu'avec les statistiques que nous considérons, ce paramétrage est plus réaliste. $\Delta(c^h) = 0$ signifie qu'il y a exactement les mêmes nombres de chromosomes modernes dans les simulations que dans la réalité. Dans le cas de la distribution uniforme, $\rho(\beta^{h,s}) = 0.02$ signifie que, dans les simulations, le nombre de segments conservés mono-géniques (entre l'humain et la souris) représente, en moyenne, 2% du nombre de segment conservés mono-géniques dans la réalité. Il s'agit des segments conservés mono-géniques car dans ce cas $W = 1$ (voir section III.5.3). Dans le cas de la distribution Γ , $\rho(\beta^{h,s}) = 0.8$ signifie que cette fois-ci il n'y a, en moyenne, que 20% de segments conservés humain-souris (de tailles $\leq W$ gènes) en moins dans les simulations que dans la réalité. Là encore le maximum de l'écart relatif est souvent atteint pour une taille de 1 gène, $W = 1$.

(analyse non développée ici).

La figure III.6 représente une distribution $\beta^{h,s}$ obtenue avec la distribution Γ . La distribution des tailles de segments conservés dans la réalité y est juxtaposée.

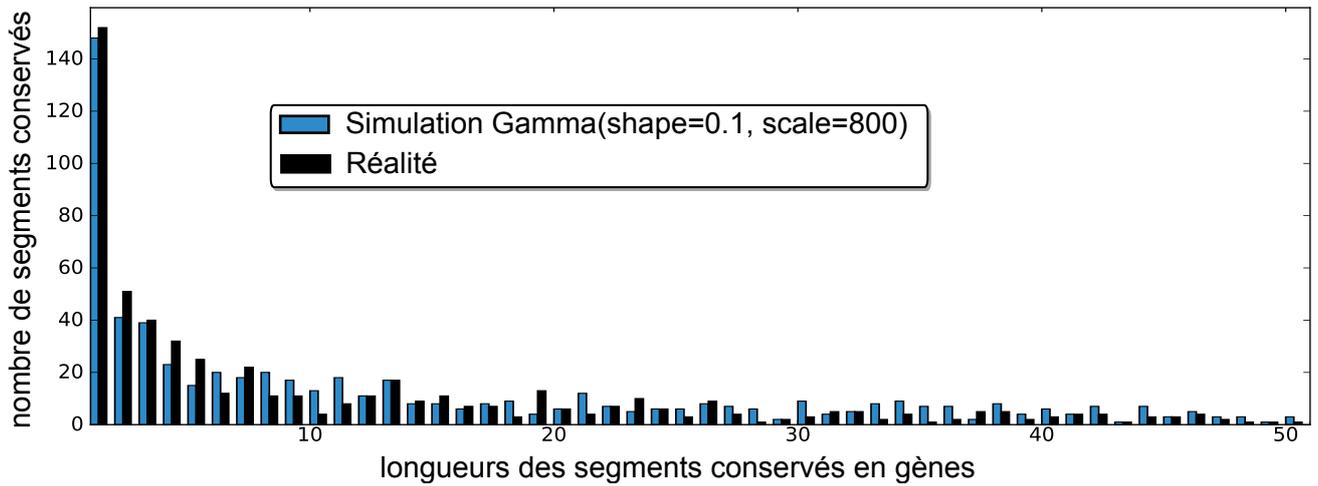


Figure III.6 – Distributions des tailles de segments conservés entre l’humain et la souris $\beta^{h,s}$. En noir la distribution des segments conservés entre les génomes réels de l’humain et de la souris ($\beta_r^{h,s}$) et en bleu la même distribution entre les génomes simulés ($\beta_{i,s}^{h,s}$) avec la distribution $\Gamma(\alpha = 0.1, \theta = 800)$, tronquée à 1330 gènes.

Les figures III.7 et III.8 représentent respectivement les matrices d’homologies, réelle et simulée (avec la distribution Γ), de la comparaison des génomes de l’humain et de la souris.

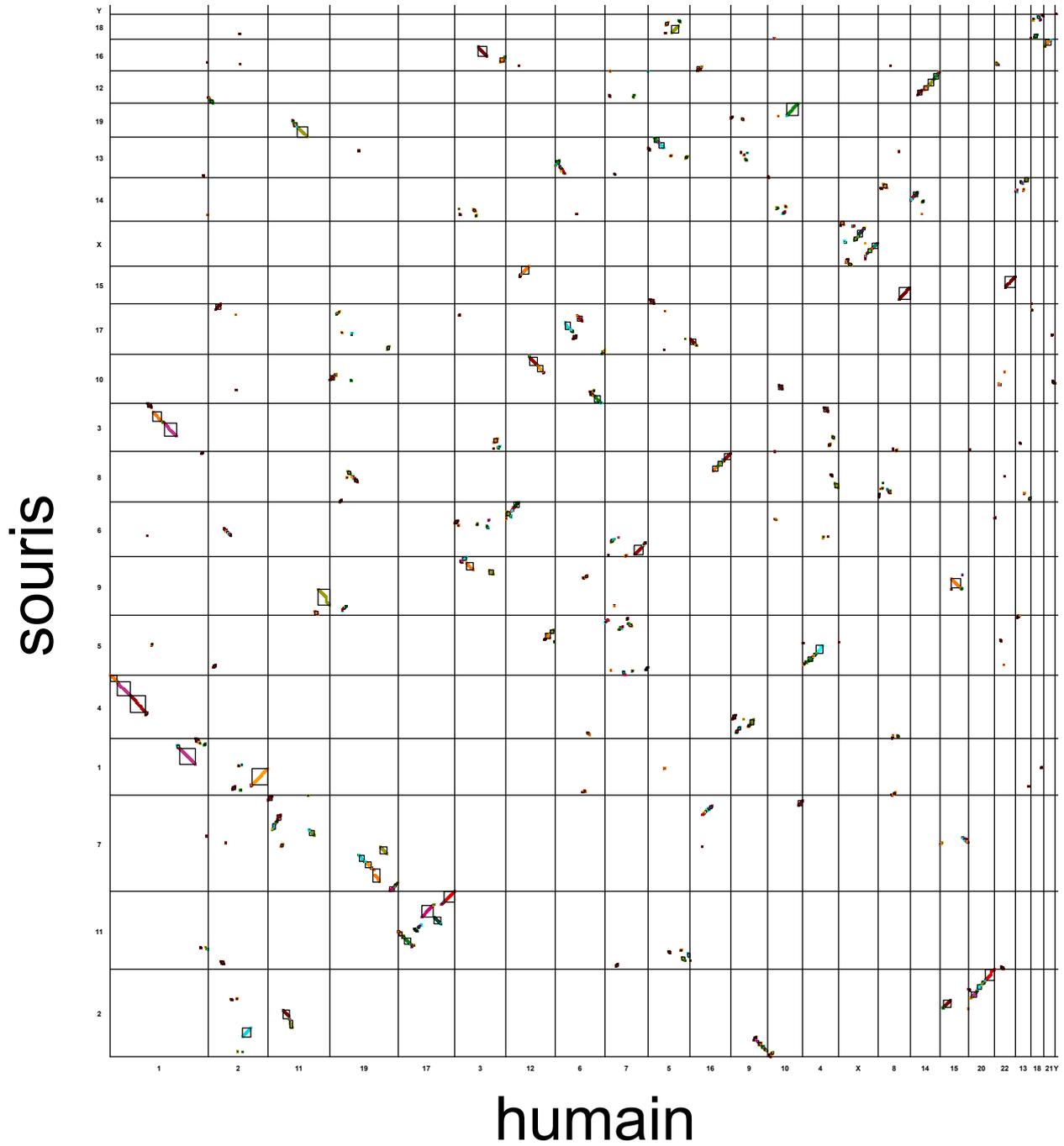


Figure III.7 – Matrice d’homologies de la comparaison du génome réel de l’humain et du génome réel de la souris.

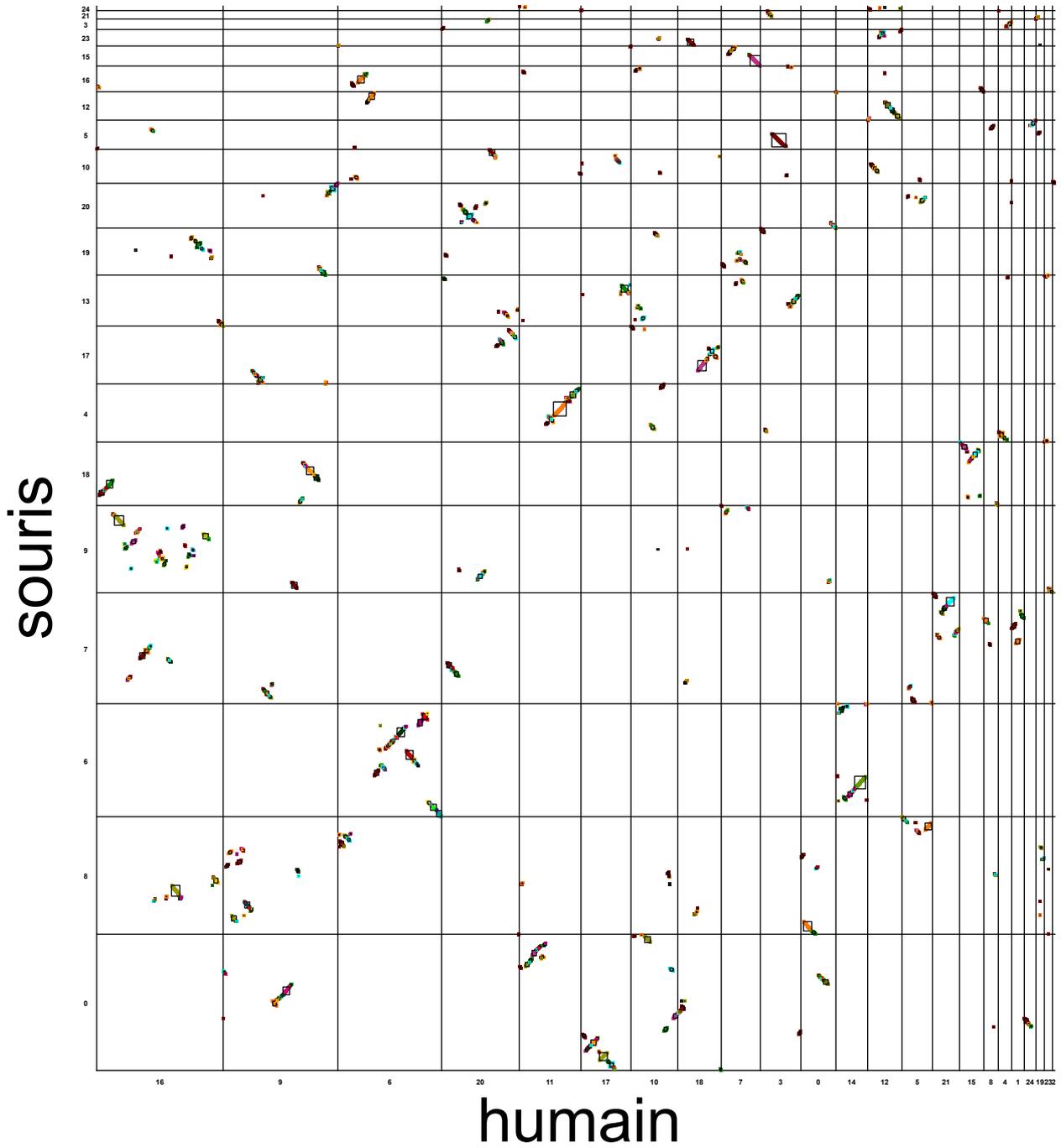


Figure III.8 – Matrice d’homologies de la comparaison du génome simulé de l’humain et du génome simulé de la souris.

Dans la figure III.8, mis à part les tailles des chromosomes aux bords de la matrice (les petits et les grands chromosomes), nous constatons que la dispersion des segments conservés et que les inversions semblent qualitativement proches de la réalité. Nous constatons néanmoins ce qui semble être un excès de petites inversions, ou, peut-être, une distribution trop uniforme des petites inversions le long des chromosomes. Il se pourrait que, dans la réalité, les petites inversions soient statistiquement localisées plus proches les unes des autres que dans nos simulations. Nous reviendrons plus en détail sur ce point par la suite (section III.8 et section D.3).

III.7 Comparaison de nos résultats avec la littérature

Hao Zhao et Guillaume Bourque ont publié en 2009 une extension de la méthode EMRAE (Efficient Method to Recover Ancestral Events) qui, à partir de segments conservés, permet de retrouver les réarrangements qui ont eu lieu [Zhao et Bourque, 2009]. Ils ont appliqué leur méthode à 6 génomes modernes de Boreoeutheriens : l'humain, le chimpanzé, le macaque rhésus, la souris, le rat et le chien. Les nombres de réarrangements qu'ils ont trouvés pour chacune des branches de l'arbre des espèces sont présentés au dessus des branches de l'arbre phylogénétique de la figure III.9.

De Euarchontoglire, l'ancêtre commun de l'humain et de la souris, jusqu'à la souris nous avons trouvé 196 inversions, 51 translocations réciproques, une fission et 6 fusions (figure III.3). Ces nombres sont à comparer aux 230 inversions, 5 translocations, 19 transpositions et 5 fissions/fusions, obtenus en cumulant le nombre de réarrangements de la figure III.9, depuis Euarchontoglire jusqu'à la souris. La dernière valeur (fissions/fusions) semble être égale à la variation du nombre de chromosomes, en valeur absolue, entre les deux extrémités de la branche. Cette valeur doit donc être comparée à la valeur absolue de la différence entre les nombres de fusions et les fissions que nous avons trouvés. Comme notre simulateur n'effectue pas de transposition, nous regroupons les translocations et les transpositions trouvées par Zhao et Bourque. La somme de ces deux derniers nombres correspond au nombre de *réarrangements inter-chromosomiques*, ils en trouvent $5 + 19 = 24$ dans la lignée de la souris, ce chiffre peut être comparée à nos 51 translocations réciproques. Le tableau III.2 regroupe les comparaisons des triplets (inversions, réarrangements de segments inter-chromosomiques et variation du nombre de chromosomes par fissions/fusions).

La différence la plus notable est celle des réarrangements inter-chromoso-

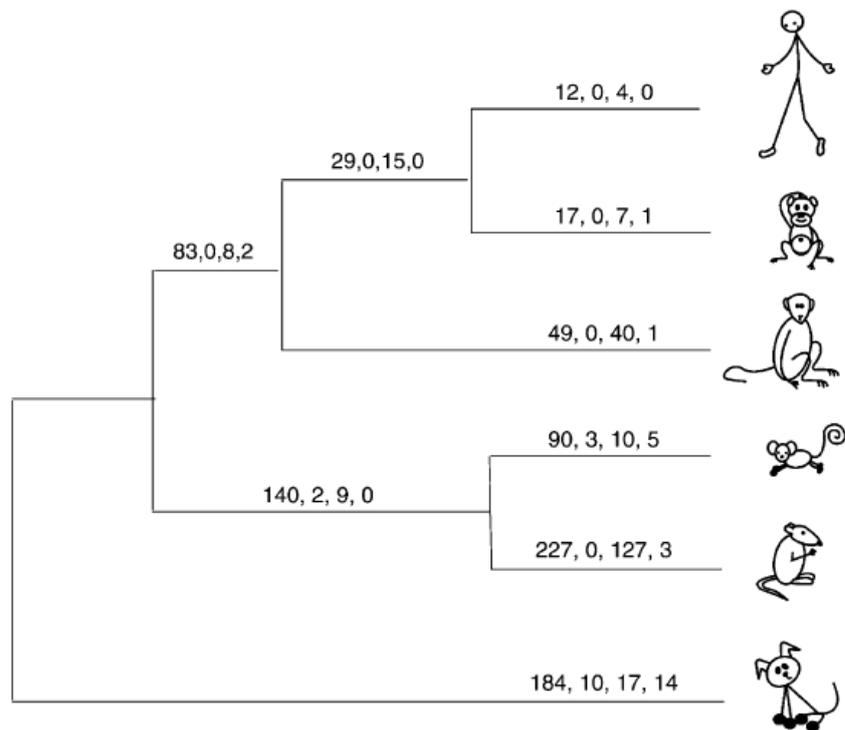


Figure III.9 – Nombres d'évènements de réarrangements chromosomiques prédits par EMRAE dans la phylogénie de 6 espèces (humain, chimpanzé, macaque rhésus, souris, rat et chien) avec une résolution des segments conservés de 10 kb. Les quatre nombres au dessus de chaque branche sont respectivement les nombres d'inversions, de translocations, de transpositions, et de fissions/fusions prédits. Figure tirée de [Zhao et Bourque, 2009].

	Inversions	Réar. inter-chrom.	Fission/Fusion
	Euarchontoglire → souris		
EMRAE	230	24	5
MagSimus	196	51	5
	Euarchontoglire → humain		
EMRAE	124	27	2
MagSimus	152	3	2
	Chien → Boreoeutheria → Euarchontoglire		
EMRAE	184	27	14
MagSimus	160	24	14

Tableau III.2 – Comparaison des nombres de réarrangements chromosomiques trouvés avec EMRAE et MagSimus. Pour les dernières valeurs, nous avons additionné les nombres de réarrangements dans la branche de Boreoeutheria jusqu’au chien avec ceux de Boreoeutheria jusqu’à Euarchontoglire en comptant la fission comme une fusion.

miques dans la lignée de l’humain. Alors que nous n’avons inféré que 3 translocations réciproques, Zhao et Bourque trouvent 27 transpositions. Avec notre calcul précédent (section III.2.6), nous devrions inférer ces réarrangements inter-chromosomiques, qui sont sensés être perçus comme des translocations (mis à part leurs 3 points de cassures). Nous pensons que nous n’avons pas vu ces transpositions car notre résolution est plus grossière que celle de Zhao et Bourque. Leurs segments conservés ont une résolution qui descend jusqu’à 10 kb, alors que la taille moyenne de nos gènes est de 24 kb (section I.1.8). De plus les segments transposés trouvés par Zhao sont généralement très courts (< 50 kb). Ce qui conforte l’hypothèse que nous n’avons pas identifié ces segments conservés car nos gènes sont trop grands et peut-être aussi par ce qu’ils ne couvrent pas de manière uniforme les chromosomes : les segments conservés de Zhao sont possiblement dans des régions pauvres en gènes. Il n’est pas étonnant que les nombres de fissions/fusions soient les mêmes étant donné que ces réarrangements sont assez simples à inférer via la comparaison des nombres de chromosomes modernes. Les nombres d’inversions sont assez proches des nombres trouvés par Zhao et Bourque, mis à part dans la lignée de la souris, où nous avons trouvé 15% moins d’inversions qu’EMRAE. Encore une fois nous pensons que la résolution des segments conservés explique probablement cette différence, car la lignée de la souris est la lignée dans laquelle ils ont trouvé le plus d’inversions courtes (< 50 kb). Enfin, avec MagSimus nous inférons plus d’inversions dans la lignée humaine qu’EMRAE. Nous n’avons pas vraiment d’explication à fournir pour cette différence et

une étude plus approfondie serait nécessaire pour la comprendre.

De manière générale il peut sembler surprenant que, étant donné les faibles valeurs de sensibilités et spécificités d'EMRAE [Zhao et Bourque, 2009], nous aboutissions à des résultats très similaires. Nous l'expliquons en rappelant que le calcul de sensibilité et de spécificité d'EMRAE semble avoir été conçu non pas pour refléter la sensibilité et la spécificité d'EMRAE à retrouver les réarrangements réels sur la base de 3356 segments conservés (« blocs de synténie »), mais pour démarquer EMRAE de MGR¹, ce pourquoi seuls 100 segments conservés sont utilisés dans les simulations. De plus, la sensibilité et la spécificité d'EMRAE sont calculées par rapport aux réarrangements effectués lors de leurs simulations, alors que notre calcul et celui d'EMRAE concerne les réarrangements dont il reste des traces en fin d'évolution (section III.3.5).

Même si quelques écarts existent entre nos résultats et ceux d'EMRAE, la majorité des nombres d'évènements sont très proches et la comparaison permet de confirmer quelques tendances générales de l'évolution des Boreoeutheriens. Il s'avère y avoir environ dix fois plus d'inversions que de réarrangements inter-chromosomiques. Il semble qu'il y a peu de fissions et de fusions de chromosomes. De plus la lignée de la souris donne l'impression d'avoir expérimenté 60% plus de réarrangements que la lignée humaine, lors du même temps évolutif. Enfin, les nombres de réarrangements inter-chromosomiques sont manifestement très variables d'une lignée à l'autre.

Ces résultats similaires n'enlèvent rien à l'intérêt de MagSimus. EMARE n'est pas un simulateur et, par exemple, il est incapable de fournir des génomes ancestraux de référence à partir desquels évaluer des logiciels de génomique comparative, similairement à ce qui a été fait dans la section II.4.3.

Nos résultats sont par contre assez éloignés de ceux obtenus dans un autre travail [Miklós et Tannier, 2010]. Car les auteurs trouvent peu d'inversions (34) et de nombreuses translocations réciproques (112) dans la lignée de la souris. Les auteurs considèrent également les translocations non réciproques (télomériques) et ils en trouvent 53.

III.8 Les régions fragiles et la réutilisation des points de cassure

Nous décrivons ici une étude complémentaire qui a mis en évidence une limite importante de notre simulateur dans sa capacité à reproduire une évolution réaliste de l'ordre des gènes.

¹un algorithme qui a le même but qu'EMRAE

MagSimus fait l'hypothèse que les régions réarrangées sont réparties majoritairement aléatoirement. Or, comme nous l'évoquions dans l'Introduction (section 4), depuis 2003 la communauté scientifique a mis en évidence des régions chromosomiques plus réarrangées que d'autres, des régions « fragiles » [Pevzner et Tesler, 2003b].

III.8.1 Mise en évidence des régions fragiles

En juin 2003 Pevzner et Tesler [Pevzner et Tesler, 2003b] ont inféré 281 blocs de synténie le long de l'évolution qui sépare l'humain de la souris avec le logiciel GRIMM-synteny. Ces blocs de synténie négligent les micro-réarrangements qui font moins de 1 Mb, pour s'abstraire de certains micro-réarrangements qui étaient causés par des erreurs d'assemblages. Les scénarios de réarrangements parcimonieux trouvés par le logiciel GRIMM [Tesler, 2002] transforment le génome de l'humain en celui de la souris à travers un nombre minimum de réarrangements¹. « *Les scénarios trouvés dévoilent qu'au moins 245 réarrangements des 281 blocs de synténie ont eu lieu entre l'humain et la souris. Ce résultat, combiné à des formules pour calculer le nombre de réutilisations de points de cassure implique que tout scénario de réarrangement entre l'humain et la souris requiert au moins 190 réutilisations de points de cassure* »² [Pevzner et Tesler, 2003b]. À titre de comparaison, 43 réutilisations sont attendues d'après le RBM, ce dernier modèle est donc fortement remis en question par les auteurs. Pour expliquer ces nombreuses réutilisations les auteurs font l'hypothèse que les intervalles courts entre les blocs de synténie, les « breakpoint regions », sont les hôtes de nombreuses cassures voisines. Chacune des 190 cassures précédentes, lorsqu'elle a eu lieu dans un intervalle, entre deux blocs de synténie, a créé un nouveau micro-segment conservé d'une taille inférieure à l'intervalle. Comme les intervalles sont courts (en moyenne ≤ 0.668 Mb) ces segments sont invisibles à une résolution de 1 Mb. Alors que la distribution des tailles des 281 blocs de synténie s'ajustait correctement à une distribution en exponentielle décroissante prédite par le RBM, l'ajout de ces 190 segments, invisibles sans les scénarios de réarrangements, change la donne. Encore une fois, ces nombreux segments conservés courts ne peuvent pas s'expliquer par le RBM, cette modélisation théorique des cassures ne peut pas rendre compte des nombreuses cassures voisines qui ont été trouvées par l'étude des scénarios de réarrangements. Comme nous l'écrivions (section 4 de l'Introduction) Pevzner et Tesler introduisent un nouveau modèle de

¹Les réarrangements considérés sont les inversions, les translocations réciproques, les fusions et les fissions. Les transpositions sont négligées en s'appuyant sur le raisonnement de Nadeau et Taylor [Nadeau et Taylor, 1984]

²traduit de l'anglais

répartition des points de cassure, le FBM. Dans ce modèle, des régions solides longues alternent avec des régions fragiles courtes et les régions fragiles sont distribuées avec une probabilité uniforme le long des chromosomes [Pevzner et Tesler, 2003b]. De plus ces dernières régions représenteraient une faible proportion des génomes, 5%, du génome humain [Pevzner et Tesler, 2003b].

Ce modèle a été débattu par les tenants du RBM et comme nous l'évoquions (section II.4), l'identification des micro-segments conservés, dans les régions fragiles, fut un sujet central du débat [Peng *et al.*, 2006][Sankoff, 2006][Sankoff, 2009].

Pour tenir compte des petits segments conservés, nous avons refait l'analyse des scénarios d'inversions parcimonieux du chromosome X, effectuée en 2003 par Pevzner et Tesler, avec notre méthode de détection des segments conservés, dont la précision a été soigneusement évaluée (section II.4.3).

III.8.2 Illustration des régions fragiles avec les inversions du chromosome X

En comparant les chromosomes X de l'humain et de la souris, nous avons trouvé 31 segments conservés dont 13 segments conservés mono-géniques. La validité des segments conservés mono-géniques a été vérifiée manuellement en s'aidant de la matrice d'homologies de la figure III.10. Ces segments sont numérotés de 19 à 31 et les homologies correspondantes sont toutes localisées entre deux extrémités de segments conservés multi-géniques ou aux extrémités des chromosomes, comme attendu¹. Un exemple est donné avec le segment conservé mono-génique 25, qui contient un unique gène, *PLS3* (ENSG00000102024). Celui-ci fait partie des gènes CCDS et il n'a pas d'inparalogue dans le génome humain. Nous pensons donc que la position de son homologie dans la matrice d'homologies n'est pas un artéfact.

Sur la base de ces 31 segments conservés, nous avons utilisé le logiciel GRIMM [Tesler, 2002] pour inférer le scénario de réarrangement qui minimise le nombre d'inversions. La figure III.11 illustre le scénario trouvé. Le code couleur des segments conservés de cette figure ne correspond pas à la figure III.10, mais la numérotation correspond bien. Ce scénario révèle que sur un total de 52 cassures, un minimum de 20 réutilisations de points de cassure est nécessaire, ce qui fait une statistique de réutilisation de points de cassures $r = \frac{52}{(31-1)} = 1,73$ [Sankoff, 2005a].

Dans notre étude, contrairement à ce qui est parfois attendu [Sankoff, 2009], avoir des segments conservés de petites tailles, avec une meilleure résolution que ceux de Pevzner et Tesler, n'atténue pas le phénomène de

¹nous renvoyons le lecteur à l'exemple précédent, section II.3.2

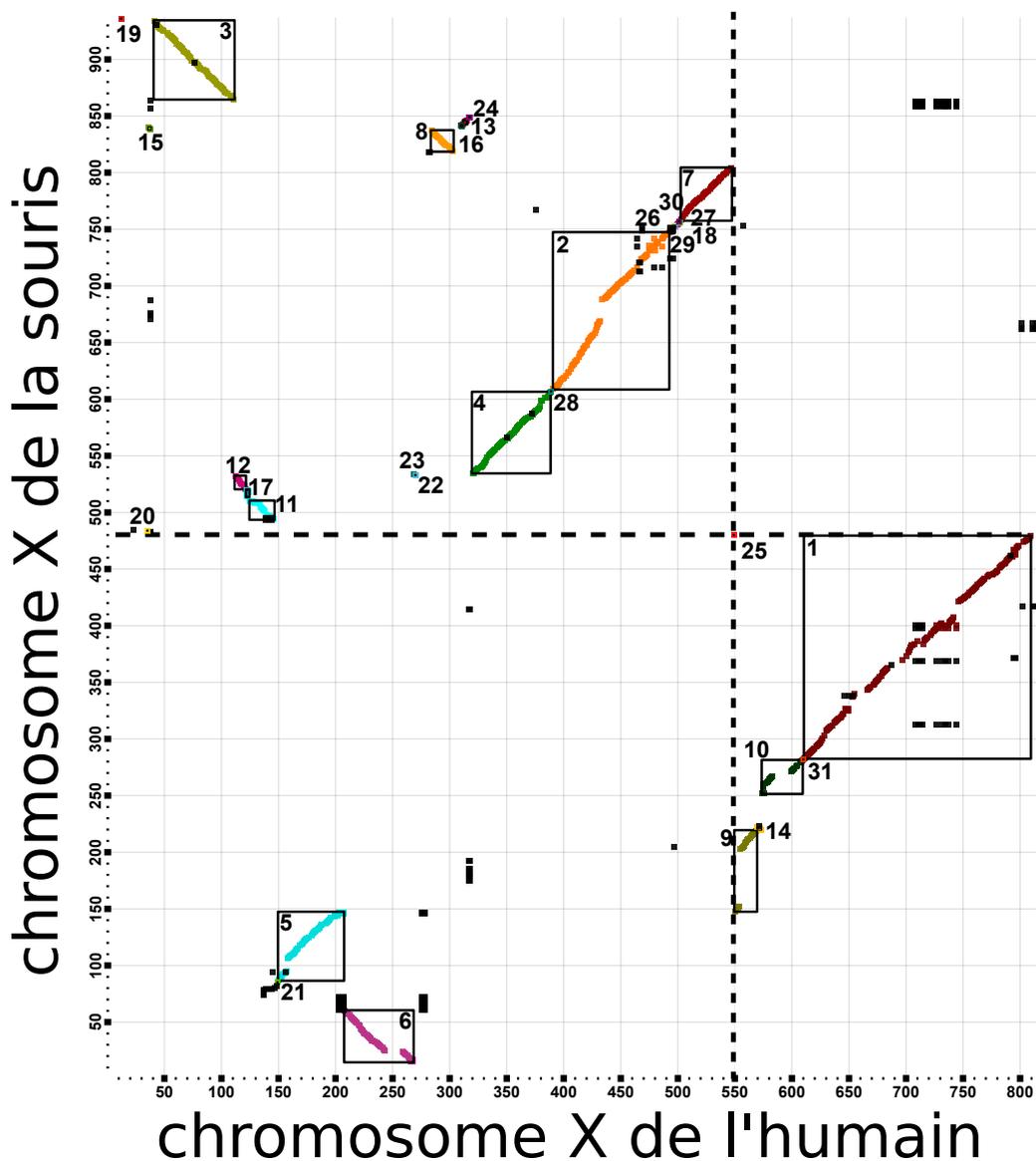


Figure III.10 – Matrice d’homologies de la comparaison des chromosomes réels X de l’humain et de la souris. Les segments conservés sont numérotés de 1 à 31, du segment qui contient le plus d’homologies au segment qui en contient le moins. Les segments 19 à 31 ne contiennent qu’une seule homologie et ils correspondent à des segments conservés mono-géniques. Nous vérifions graphiquement que ceux-ci sont localisés aux voisinages des extrémités de segments conservés multigéniques, ou à une extrémité de chromosome. Par exemple le segment conservé mono-génique 25 est situé entre les segments conservés 7 et 9 sur le chromosome X humain (ligne pointillée verticale) et entre les segments conservés 1 et 20 sur le chromosome de la souris (ligne pointillée horizontale).

réutilisation des points de cassures. Le scénario d'inversions révèle également que des grandes inversions sont parfois nécessaires pour relocaliser les petits segments conservés d'une zone fragile à une autre. De plus il semble qu'il y ait eu de nombreuses micro-inversions dans la région entre les segments conservés 2 et 7, impliquant les segments conservés 26, 29, 18, 30 et 27. De manière assez surprenante aucune grande inversion ne semble avoir brisé cette région. Comme Pevzner et Telser, la région voisine du centromère du chromosome X humain, vers 55 Mb, nous apparaît également comme fragile et de nombreuses cassures semblent y avoir eu lieu.

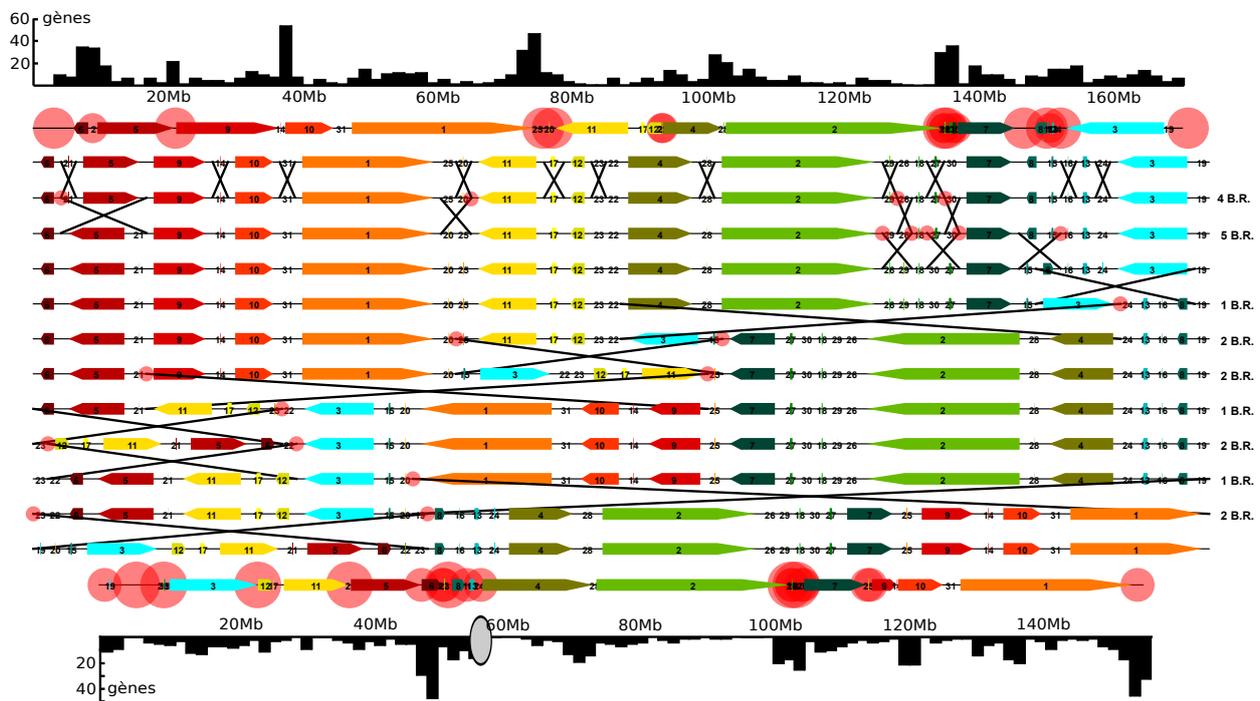


Figure III.11 – Scénario de réarrangements avec un minimum d'inversions du chromosome X de la souris jusqu'au chromosome X de l'humain, de haut en bas. Le choix de placer l'humain en bas, au bout du scénario, permet de reproduire la figure 2 de [Pevzner et Tesler, 2003b] ainsi que la figure 2C de [Bourque *et al.*, 2004]. Les cercles rouges symbolisent les réutilisations de points de cassures. L'ovale gris sur le chromosome humain représente la localisation du centromère. En haut et en bas, à côté de chaque chromosome, l'histogramme représente le nombre de gènes dans chaque région du chromosome.

En nous basant sur cette analyse, qui confirme et affine les résultats précédents, obtenus depuis 2003, nous croyons qu'il existe bel et bien des

régions fragiles et solides le long des chromosomes, au moins le long du chromosome X. La présence de ces régions n'a pas encore été modélisée dans notre simulateur. Par conséquent le nombre de réarrangements que nous avons inférés et la distribution des tailles d'inversions que nous avons proposée (III.2d) sont à relativiser. De plus, dans l'équation III.8, négliger r (le nombre de réutilisations de points de cassures), aboutit à sous-estimer le nombre d'inversions qu'il y a probablement eu.

Pour illustrer simplement la limite de notre simulateur MagSimus, la simulation de 26 inversions d'un chromosome équivalent au chromosome X (composé des 600 gènes ancestraux conservés depuis Euarchontoglires) ne génère en moyenne que deux ou trois réutilisations de points de cassures; alors que 20 réutilisations de points de cassures sont inférés avec GRIMM pour les 26 inversions qu'il trouve entre l'humain et la souris.

III.8.3 Suggestions pour simuler les réutilisations de points de cassures

De manière analogue aux simulations de Jian Ma [Ma *et al.*, 2008b] :

- définir les régions fragiles dans le génome ancestral
- garder en mémoire les extrémités des gènes dans ces régions
- et, de temps en temps (selon le taux de réutilisation des points de cassure) forcer les points de cassures à avoir lieu à proximité de ces extrémités

serait une solution simple pour rendre compte dans notre simulateur des régions fragiles.

Une autre solution consisterait à modéliser les tailles (en bp) et/ou les régions de chromatine ouverte des intergènes. Nous y reviendrons dans la discussion qui suit (section D.3.9).

Discussion

Une énorme tuile arrachée par le vent, tombe, et assomme un passant. Nous disons que c'est le hasard. Le dirions-nous, si la tuile s'était simplement brisée sur le sol ? Peut-être, mais c'est que nous penserions vaguement alors à un homme qui aurait pu se trouver là, ou parce que, pour une raison ou pour une autre, ce point spécial du trottoir nous intéressait particulièrement, de telle sorte que la tuile semble l'avoir choisi pour y tomber. [...] Ne pensez qu'au vent arrachant la tuile, à la tuile tombant sur le trottoir, au choc de la tuile sur le sol : vous ne voyez plus que du mécanisme, le hasard s'évanouit. Le hasard est donc le mécanisme se comportant comme s'il avait une intention.

HENRI BERGSON, *Les Deux Sources de la morale et de la religion*, 1932

Synthèse du travail accompli

Nous avons modélisé les génomes à une échelle d'étude pratique pour étudier les réarrangements chromosomiques, nous avons présenté les principaux événements connus qui font évoluer ces génomes. Nous avons brièvement introduit un arbre de gène en trois dimensions qui facilite la visualisation de l'intrication entre l'histoire des gènes et l'histoire des espèces dans lesquels ils évoluent. Nous avons défini méthodologiquement un segment conservé durant l'évolution d'un génome ancestral jusqu'à deux génomes descendants. Nous avons expliqué qu'un tel segment génère une diagonale dans la matrice d'homologies des deux chromosomes descendants où il a été conservé, si toutefois les localisations des gènes conservés depuis l'ancêtre sont clairement identifiées. Nous avons expliqué qu'en pratique cette identification est ardue car un gène ancestral est souvent inextricablement confondu avec ses in-paralogues voisins. Pour palier à cette difficulté, nous avons réécrit les chromosomes en tandem blocs¹, pour assigner à tous les clusters de gènes dupliqués en tandem un

¹procédure développée avant nous dans les logiciels ADHoRe [Vandepoele *et al.*, 2002], MCSanX [Wang *et al.*, 2012] et SynChro [Drillon *et al.*, 2014]

unique représentant du gène ancestral. Nous avons également rappelé que les erreurs d’annotations et d’assemblages qui étaient légions dans les premières séquences ont mené à la notion de bloc de synténie [Pevzner et Tesler, 2003a]. Nous avons proposé une définition méthodologique d’un bloc de synténie et nous avons montré comment notre définition correspond à sa définition intuitive initiale. Nous espérons que cette définition aidera la communauté à se rapprocher d’une définition qui fasse consensus [Ghiurcuta et Moret, 2014]. Nous avons montré qu’un bloc de synténie, défini de notre manière et conservé avec des gaps $\leq gapMax$ gènes, génère une diagonale avec des gaps $\leq gapMax$ dans la matrice d’homologies des deux chromosomes où le bloc est présent, pour peu que les chromosomes soient correctement filtrés (section I.9.2). Puis nous avons explicité le lien qui existe entre notre définition d’un bloc de synténie et notre définition d’un segment conservé.

À partir de ces définitions, nous avons développé PhylDiag [Lucas *et al.*, 2014], un logiciel qui identifie les vestiges de segments conservés. PhylDiag commence par identifier les vestiges de blocs de synténie de gaps $\leq gapMax$ comme des diagonales avec des gaps $\leq gapMax$ dans les matrices d’homologies de génomes modernes filtrés. Ces blocs de synténie sont par la suite traités de manière à retrouver les segments conservés. À la différence des autres méthodes d’inférences de segments conservés, lors du traitement, les segments conservés sont identifiés même si ceux-ci ne sont composés que d’un gène, ce qui ne semble pas anecdotique étant donné les nombreuses inversions courtes [Lefebvre *et al.*, 2003][Chaisson *et al.*, 2006] et les régions fragiles [Pevzner et Tesler, 2003b] dans les espèces de vertébrés. La sensibilité et la spécificité de PhylDiag ont été estimées sur la base de segments conservés idéaux issus de simulations réalistes. Nous avons montré que, lorsque deux génomes simulés de manière réaliste sont comparés, PhylDiag identifie avec une meilleure précision les segments conservés que le logiciel i-DAHoRe 3.0 [Proost *et al.*, 2012], un des logiciels de référence [Ghiurcuta et Moret, 2014].

Nous avons également développé MagSimus, un simulateur de l’évolution de l’ordre des gènes de vertébrés qui prend en compte les principaux réarrangements de la littérature : les inversions, les translocations réciproques, les fissions et les fusions de chromosomes. Les principaux événements géniques (à notre échelle d’étude) sont eux-aussi simulés : les gènes sont dupliqués en tandem ou de manière disperse, nous simulons également des délétions de gènes et des naissances *de novo* de gènes. MagSimus permet à l’utilisateur de configurer les principaux paramètres qui influent sur l’ordre des gènes : les chromosomes réarrangés peuvent être sélectionnés indépendamment de leurs tailles ou proportionnellement à leurs tailles, une taille minimale et une taille maximale des chromosomes peuvent être fixées, les taux d’évènements de réarrangements et d’évènements géniques peuvent être assignés à chaque

branche ainsi que la proportion des duplications en tandem et une distribution des tailles de segments inversés peut être définie. Magsimus a été utilisé pour simuler l'évolution de cinq amniotes. Les simulations font évoluer un génome artificiel d'Amniota jusqu'à cinq espèces modernes : l'humain, la souris, le chien, l'opossum et le poulet. Tous les paramètres d'évolution ont été ajustés de manière à maximiser le réalisme de la simulation. Ce réalisme a été quantifié par le calcul d'un écart général à la réalité qui est une moyenne géométrique de plusieurs écarts relatifs. Ces derniers quantifient la différence entre les chromosomes simulés et les chromosomes réels ainsi que les différences entre les segments conservés simulés et les segments conservés réels. L'usage des simulations a mis en lumière de nombreux biais dans nos estimations des paramètres initiaux, dus principalement au bruit introduit par les événements géniques. Ces biais ont été corrigés par une optimisation sur les nombres d'inversions et de translocations réciproques. Enfin, les segments conservés que nous avons identifiés et les limites de notre simulateur, mettent en relief l'importance de la réutilisation des points de cassures durant l'évolution et nous avons suggéré une amélioration simple qui permettra à MagSimus de les reproduire.

Les logiciels PhylDiag et MagSimus ont été programmés en python et sont disponibles via GitHub (<https://github.com/DyogenIBENS>), sous licences GPL 3 et CeCiLL 2. Pour plus de détails sur l'implémentation informatique de ce travail, voir section A.2.

D.1 Discussion à propos de notre modélisation des génomes

D.1.1 Gènes et paires de bases

Utiliser les séquences protéiques comme marqueurs (section I.1.8) plutôt que des séquences quelconques a des avantages [Bourque *et al.*, 2005]. Par exemple, le génome du poulet est assez éloigné du génome des mammifères. Utiliser tous types de séquences homologues, identifiées avec un seuil de similarité faible (à cause de la distance) génèrera des paires d'homologies douteuses et difficiles à analyser. Il serait tout de même intéressant de vérifier que notre choix de marqueurs n'influe pas sur les conclusions générales de notre étude. Par exemple, il serait intéressant d'utiliser les séquences homologues non-codantes pour pouvoir identifier les nombreux micro-réarrangements qui ont lieu dans des intergènes. Une telle étude comparative a déjà été menée avec des espèces similaires (humain, souris, rat et poulet), dans laquelle les blocs de synténie et des génomes ancestraux sont inférés avec d'un côté uniquement des séquences

génomiques et de l'autre des alignements de séquences de toutes sortes [Bourque *et al.*, 2005]. Après avoir énuméré les avantages et les inconvénients de chaque type de données les auteurs concluent néanmoins que les inférences sont en grande partie constantes d'un type de séquence à l'autre.

D.1.2 Gènes non-codants et autres marqueurs

Les séquences conservées que nous avons utilisées comme marqueurs pour modéliser nos génomes sont des séquences qui correspondent aux transcrits les plus courts de gènes codant pour les protéines. La possibilité de télécharger les arbres phylogénétiques de ces séquences fut un argument majeur dans notre choix. Au début de ce travail la base de donnée Ensembl ne mettait pas à notre disposition de reconstruction d'arbres phylogénétiques d'autres séquences. Entre temps des arbres phylogénétiques pour des gènes transcrits en ARN non codant (ARNnc) ont été partagés [Pignatelli en préparation] et il serait intéressant de les utiliser pour accroître la densité de nos marqueurs le long des chromosomes. D'autres séquences conservées pourraient être utilisées, comme les séquences des éléments ultraconservés [Bejerano *et al.*, 2004], pour peu que leurs arbres phylogénétiques soient calculés. Avec plus de marqueurs, nous parviendrions à identifier plus finement les réarrangements.

D.1.3 Modéliser les centromères

Dans notre modélisation des génomes nous n'avons pas considéré les centromères or il est fort probable que ceux-ci jouent un rôle non négligeable dans les réarrangements chromosomiques. Par exemple certains auteurs font l'hypothèse que les translocations se font entre bras chromosomiques [Arkendra *et al.*, 2001] et non entre chromosomes entiers. Modéliser les centromères nous permettra d'étudier cette caractéristique évolutive.

D.1.4 Des inversions avec un point de cassure

Nous avons systématiquement associé deux points de cassures aux inversions, néanmoins il se peut qu'une inversion n'implique qu'un point de cassure si une des extrémités du segment réarrangé correspond à une extrémité de chromosome.

D.1.5 Considérer plus les transpositions

Depuis l'article de Nadeau et Taylor en 1984 [Nadeau et Taylor, 1984] les transpositions de segments de chromosomes sont souvent négligées dans les

scénarios de réarrangements chromosomiques. Des travaux ont néanmoins mis en évidence de nombreuses transpositions de segments [Zhao et Bourque, 2009][Kent *et al.*, 2003][Coghlan, 2002] de plus les transpositions sont de plus en plus souvent considérées dans les calculs de distances en réarrangements chromosomiques [Alekseyev et Pevzner, 2008]. Il pourrait donc être intéressant d’intégrer les transpositions dans notre simulateur. Dans ce but, il sera probablement nécessaire d’accroître la résolution de nos marqueurs car il semble que les segments chromosomiques transposés soient majoritairement plus courts que 50 kb [Zhao et Bourque, 2009].

D.1.6 Influence des évènements relatifs à une espèce

Dans notre étude nous ne considérerons que les évènements de spéciation, néanmoins les goulots d’étranglement démographiques et les périodes d’hybridations ont probablement joué un grand rôle dans l’évolution de nos génomes. Ce problème a été soulevé dans la conclusion du premier article de ChromEvol [Mayrose *et al.*, 2010]. Les auteurs y expliquent qu’ils ont fait l’hypothèse que les variations des nombres de chromosomes dans les génomes ont lieu graduellement et proportionnellement au temps écoulé. De cette manière, des variations au moment des spéciations¹ ont autant de chance d’avoir lieu que des variations à tout autre moment, ce qu’ils reconnaissent être critiquable. Associer une plus grande probabilité de réarrangement le long des lignées qui ont subi de nombreuses spéciations [Mooers *et al.*, 1999] serait une solution possible pour ne pas sous-estimer l’influence des spéciations sur l’évolution des génomes.

D.2 Discussion à propos des segments conservés

D.2.1 Identification des duplications segmentales

PhylDiag a été développé dans le but de retrouver les segments conservés en comparant deux génomes différents. Cependant au lieu de comparer deux génomes différents il peut être intéressant de comparer un génome avec lui-même. Si les familles de gènes sont choisies de manière à ce que les gènes à l’origine des familles soient les gènes d’un ancêtre suffisamment lointain (section I.5.2), cette comparaison permettrait de mettre en évidence

¹Les auteurs de ChromEvol semblent supposer implicitement que les spéciations sont liées à des goulots d’étranglement démographiques.

les duplications segmentales. Pour que PhylDiag identifie les duplications segmentales de cette manière, il suffit de supprimer les homologies situées sur la diagonale principale des matrices d’homologies, dont les coordonnées sont (i, i) avec $i \in [1, n]$ et n le nombre de gènes dans le chromosome qui est comparé à lui-même.

D.2.2 Comparaison de N génomes au lieu de uniquement 2

Cette sous-section est très spéculative. Nous évoquons ici des idées que nous aimerions développer. Nous ne pouvons donc pas assurer qu’elles puissent être mises en pratique.

Nous pensons que le fonctionnement de PhylDiag pourrait servir de base à une méthode de comparaison plus générale qui comparerait N chromosomes au lieu de uniquement 2. Les matrices d’homologies seraient des matrices à N dimensions. Par exemple la matrice de packs d’homologies serait telle que chaque cellule puisse s’écrire $\text{MHP}[i_1, i_2, \dots, i_N]$, avec i_k , le rang du tandem bloc du $k^{\text{ème}}$ chromosome de la comparaison. Les signes des packs d’homologies seraient des vecteurs de $N - 1$ valeurs (toutes dans $\{+1, -1, \emptyset\}$). La $i^{\text{ème}}$ valeur représenterait l’orientation du tandem bloc homologue dans le $i^{\text{ème}}$ chromosome par rapport à l’orientation du tandem bloc homologue dans le premier chromosome. Le type d’une diagonale ne serait plus slash (/), backslash (\) ou *unknown* (\emptyset), mais un vecteur de $N - 1$ valeurs (toutes dans $\{/, \backslash, \emptyset\}$). La $i^{\text{ème}}$ valeur de ce vecteur représenterait ainsi l’orientation du segment dans le $i^{\text{ème}}$ chromosome par rapport à son orientation (choisie positive) dans le premier chromosome.

Dans l’algorithme 1 de notre précédent travail [Lucas *et al.*, 2014], lors de l’identification des diagonales strictes, il serait nécessaire d’effectuer N boucles « for » au lieu de uniquement 2. La matrice à N dimensions peut, comme précédemment, être enregistrée dans une structure de données adaptée à une matrice creuse¹) pour éviter une explosion du temps de parcours de la matrice, qui serait sinon de l’ordre de n^N avec n la taille moyenne des N chromosomes comparés. Après que les diagonales strictes (à N dimensions) soient identifiées, le chaînage de celle-ci² devrait lui aussi être adapté pour que les distances entre diagonales soient explorées dans des directions qui correspondent au type de la diagonale courante. Ce dernier point correspondrait à une généralisation en N dimensions des patrons³ de calcul des distances, illustrés dans les panels

¹ « sparse matrix » dans l’article [Lucas *et al.*, 2014]

² « *mergeDiags* » à la fin de l’algorithme 1 de l’article

³ « framework » dans la figure 2 de l’article

E et F de la figure 2 de notre précédent travail [Lucas *et al.*, 2014].

Les équations des métriques des distances (section I.8.4) et les calculs de probabilités auxquels nous avons fait référence [Lucas *et al.*, 2014] sont généralisables à N -dimensions. D’ailleurs l’équation (2) de notre précédent travail [Lucas *et al.*, 2014] a déjà été présentée pour un cluster d’une comparaison multi-génomiques [Raghupathy *et al.*, 2008].

Pour obtenir les segments conservés dans N lignées il serait également utile de procéder de manière analogue à ce que nous avons fait dans la figure II.6. Par exemple, les segments conservés durant l’évolution d’un génome ancestral de l’espèce *Anc* jusqu’à deux espèces modernes S_1 et S_2 , pourraient être combinés aux segments conservés durant l’évolution de *Anc* jusqu’à S_1 et une troisième espèce modernes, S_3 . Dans ce cas les segments conservés de *Anc* jusqu’à S_1 et S_2 se substituerait à « sc1 » dans la figure II.6 et les segments de *Anc* jusqu’à S_1 et S_2 se substituerait à « sc2 ».

D.2.3 Choix de la métrique de distance

Dans notre précédent travail [Lucas *et al.*, 2014] nous avons ouvert une discussion sur le choix de la métrique de distance 2D pour mesurer les gaps des diagonales dans les matrices d’homologies¹. Nous avons énuméré les principales métriques utilisées dans la littérature : la MD est utilisé par GRIMM-synténie [Pevzner et Tesler, 2003a], Fish [Calabrese *et al.*, 2003], Cinteny [Sinha et Meller, 2007] et SyMap [Soderlund *et al.*, 2006], la DPD est utilisée par ADHoRe [Simillion *et al.*, 2004] et DiagHunter [Cannon *et al.*, 2003]. La CD semble rarement utilisée alors qu’il s’agit de la métrique qui génère les distances les plus courtes et alors que c’est la seule métrique qui assure qu’au moins une des distances en gènes dans les génomes, la plus grande, soit égale à la distance qui sépare les homologies.

Par exemple considérons la matrice d’homologies de deux chromosomes. Deux gènes A et B sont distants de 5 gènes dans le premier chromosome et leurs homologues respectifs A’ et B’ sont distants de 4 gènes dans le deuxième chromosome. Ces gènes génèreront deux homologies et la distance entre celles-ci peut se calculer de différentes manières en fonction de la métrique de distance choisie. La métrique MD calculera une distance égale à $4 + 5 = 9$, la métrique DPD calculera une distance de $2 \times \max(4, 5) - \min(4, 5) = 6$ et la métrique CD calculera une distance de $\max(4, 5) = 5$.

Nous pensons au final que la métrique CD est la plus représentative des distances entre les gènes des chromosomes. Pour finir, l’usage des autres métriques de distance perd de son intérêt lorsqu’un affinage est effectué pour

¹les métriques de distances ont été définies dans la section I.8.4

identifier les micro-réarrangements dans les gaps des diagonales (section II.4.2). Nous pensons donc que la métrique de distance CD peut être utilisée sans craindre que le chaînage des diagonales ne cache des micro-réarrangements, car ceux-ci sont identifiables après coup.

D.2.4 Graphe et linéarisation

Lors du chaînage des diagonales, il serait utile de créer un graphe qui répertorie les multiples façons de chaîner les diagonales. Ce graphe serait par la suite linéarisé en maximisant les longueurs des chemins linéaires. La linéarisation de graphes est une thématique récurrente en génomique comparative [Muffato, 2010][Bérard *et al.*, 2012][Gagnon *et al.*, 2012]. Durant notre étude bibliographique nous avons eu connaissance du travail du mathématicien J. Edmonds qui a étudié le couplage (aussi appelé appariement ou « matching ») maximal d'un graphe [Edmonds, 1963]. Ce développement mathématique a été utilisé par Jian Ma dans son travail sur le *infinite site model* [Ma *et al.*, 2008a] et nous pensons qu'il pourrait être mis à profit dans notre cas.

D.2.5 Définition d'un segment conservé indépendante de notre échelle d'étude

Nous avons défini un segment conservé comme : un segment *de gènes*¹ ancestraux successifs qui n'a été fragmenté par aucune cassure liée à un réarrangement. À notre échelle d'étude nous avons ajouté que si un réarrangement a lieu dans l'intergène d'un segment conservé (sans que ce réarrangement perturbe l'ordre ou les orientations des gènes ancestraux du segment) le segment conservé n'est pas fragmenté. Ceci nous a permis de nous affranchir des nombreux micro-réarrangements que nous ne pouvons pas voir avec nos gènes. Cependant notre définition s'écartait ici de la définition biologique d'un segment conservé qui est indépendante de notre échelle d'étude et qui définit rigoureusement un segment conservé comme un segment de chromosome (et non pas une succession de gènes) qui n'a pas été fragmenté par une cassure liée à un réarrangement. Les logiciels d'alignements de génomes (section II.3.3), visent à retrouver de tels segments conservés, définis au nucléotide près, et ces segments idéaux se rapprochent donc plus de la définition biologique d'un segment conservé. Néanmoins, comme nous l'indiquons, ces logiciels semblent porter une importance secondaire à la qualité des segments conservés par rapport à la quantité de séquences nucléotidiques alignées.

¹et plus généralement de marqueurs

D.2.6 Identifier les segments conservés à l'échelle des paires de bases

Nous avons considéré les chromosomes comme des listes ordonnées de gènes orientés, mais nous reconnaissons que les considérer à l'échelle des nucléotides pourrait nous aider à résoudre les difficultés qui persistent. À cette échelle il serait plus facile de distinguer le gène ancestral parmi l'ensemble des gènes dupliqués en tandem ce qui permettrait d'identifier le véritable scénario d'évolution dans les figures II.13 et II.14. Les séquences nucléotidiques pourraient également être utilisées pour étendre les extrémités des segments conservés dans le but d'investiguer plus en détail les vestiges de flancs de cassures [Lemaître *et al.*, 2008].

Il serait de plus utile d'implémenter une mesure des gaps en paires de bases dans PhylDiag, pour chaîner les diagonales dont les gaps qui les séparent sont inférieurs à un *gapMax* en paires de bases. Ceci aurait l'avantage de ne pas défavoriser le chaînage des zones denses en gènes par rapport aux zones pauvres en gènes.

D.2.7 Inversions courtes et erreurs d'assemblage

Comme nous l'avons rappelé, le concept de bloc de synténie a été introduit de manière à s'abstraire des erreurs d'assemblages qui créaient l'impression de nombreux micro-réarrangements et de nombreuses petites inversions [Pevzner et Tesler, 2003b]. Nous soulevons l'hypothèse qu'il s'agissait possiblement de moins d'erreurs d'assemblage que nous le pensions et qu'il s'agissait peut-être de véritables inversions de courts segments de chromosomes.

D.2.8 Optimisation du calcul de la matrice d'homologies

Le calcul des matrices d'homologies est la partie la plus longue lors de l'exécution de PhylDiag et nous avons essayé de nombreux algorithmes pour l'accélérer. Nous avons aussi utilisé Cython pour compiler un exécutable en langage C qui soit plus rapide que le code écrit en python. Cette partie du code de PhylDiag pourrait encore être optimisée pour diminuer le temps d'exécution qui est actuellement d'un total de légèrement plus de 7 secondes sur un processeur de 1.7GHz, lorsque le génome humain est comparé à celui de la souris.

D.2.9 Matrice d’homologies surprenante

Durant notre travail sur PhylDiag, nous avons été surpris par certains arrangements rares de gènes. Par exemple dans la figure D.1, le grand chevauchement des extrémités de deux blocs de synténie, dans la comparaison des chromosomes X de l’humain et de la souris, semble être le vestige d’un réarrangement qui aurait séparé deux segments précédemment dupliqués en tandem.

D.3 Discussion à propos de notre simulateur

D.3.1 Évènements et localisation des gènes

Notre simulateur fait une hypothèse implicite d’indépendance entre les évènements géniques et la localisation des gènes mutés : les gènes dupliqués et les gènes supprimés sont choisis aléatoirement lorsque l’évènement a lieu. Or notre simulateur gagnerait certainement en réalisme s’il prenait en compte des cas de coévolutions et de co-localisations. Prenons l’exemple d’une région bi-promotrice [Hurst *et al.*, 2004] qui co-régule deux gènes d’orientations de transcriptions opposées. Dans la réalité si une mutation inactive la séquence bi-promotrice les deux gènes seront probablement supprimés en même temps et notre simulateur ne reproduit pas ce phénomène.

Autre exemple, il est assez clair aujourd’hui qu’il y a une corrélation entre le nombre de duplications segmentales et la proximité d’une région cassée, même si la relation de causalité entre cassure et duplications segmentales est moins évidente [Lemaitre *et al.*, 2009][Sankoff, 2009][Berthelot *et al.*, 2015]. Durant nos simulations nous pourrions modéliser cette corrélation par une propension accrue de réarrangements dans les régions chromosomiques avec des duplications en tandem. Enfin, il semble qu’il existe une corrélation entre les zones denses en gènes et les fortes probabilités de cassure [Lemaitre *et al.*, 2009].

D.3.2 Duplications de gènes qui viennent de naître

Nous avons expliqué précédemment que dans MagSimus, les duplications de gènes précédemment insérés par naissances *de novo* dans la même branche sont considérées comme des naissances *de novo*. Nous pourrions corriger cela en introduisant un nouvel évènement génique qui correspond à une duplication de gène qui vient de naître dans la branche courante.

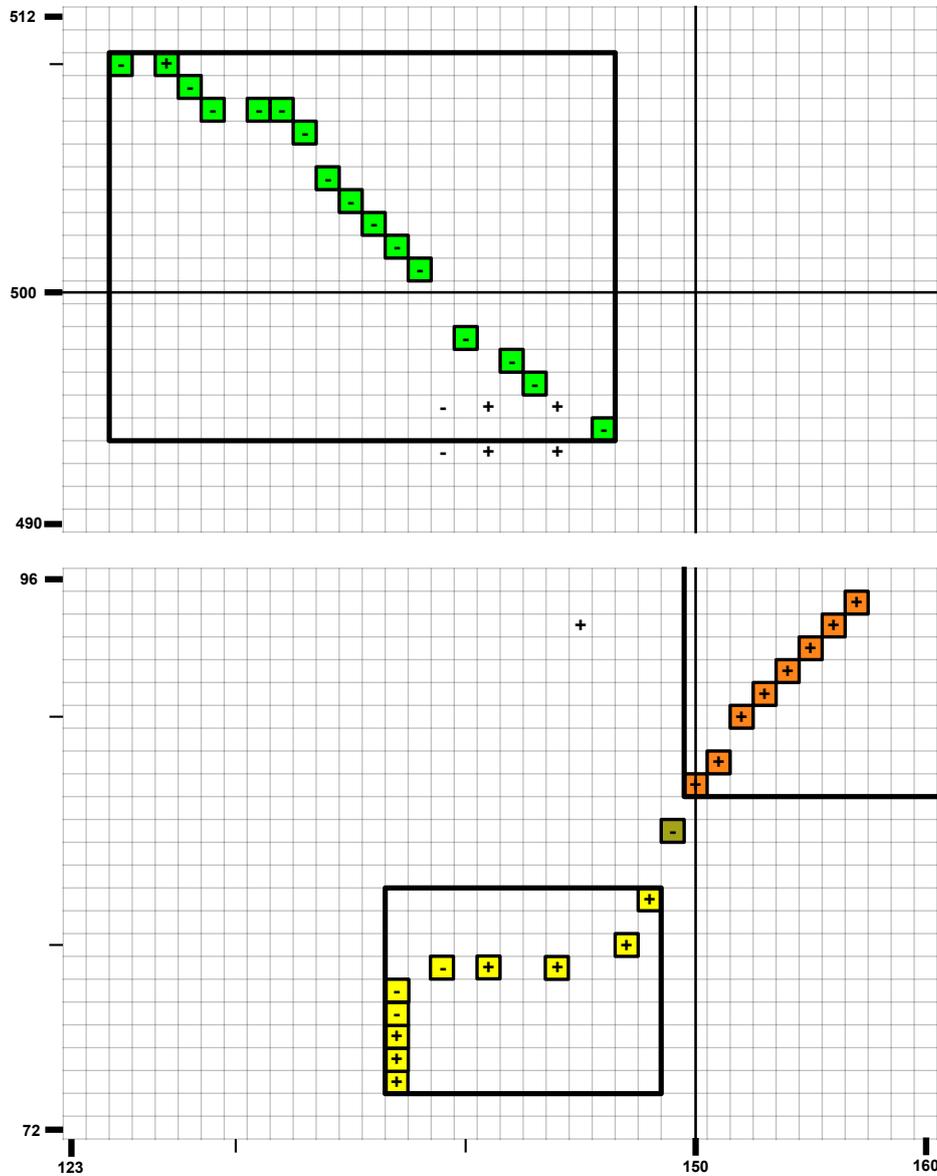


Figure D.1 – Matrice d’homologies entre le chromosome X de l’humain (en bas) et le chromosome X de la souris (à gauche). Les indices des gènes de chaque chromosome sont renseignés sur l’axe des abscisses pour le chromosome humain et sur l’axe des ordonnées pour le chromosome de la souris. Le chevauchement des extrémités des blocs de synténie vert et jaune est assez surprenant et il semble être le vestige d’une inversion qui aurait séparé deux segments précédemment dupliqués en tandem. Cette image est un agrandissement de deux régions de la matrice de la figure III.10.

D.3.3 D l tion des g nes pr c demment dupliqu s

Nous avons fait attention   ne simuler que des  volutions simples de mani re   ce que les bilans en g nes qui incluent les  v nements g niques soient exacts ( quations III.12 et III.12). De cette mani re les nombres de g nes dans les g nomes sont parfaitement pr dictibles   partir des nombres d' v nements simul s le long des branches (par exemple nous avons  vit  de supprimer un g ne qui venait d' tre ins r  suite   une duplication le long de la branche courante). Malgr  cela nous constatons un exc s de vestiges de duplications dans les g nomes simul s aux feuilles de l'arbre des esp ces, par rapport   ce que nous constatons dans la r alit . Nous pensons que cette diff rence s'explique peut- tre par une forte propension des g nes dupliqu s, ou ins r s suite   une duplication sur une branche,    tre supprim s dans la branche suivante. Par exemple, dans l' volution r elle, il est probable qu'un g ne dupliqu , ou issu d'une duplication, dans la branche d'Amniota jusqu'  Theria aura en moyenne plus de chance d' tre supprim  le long de la branche de Theria jusqu'  Boreoeutheria qu'un g ne d'Amniota conserv  en une copie jusqu'  Theria. Ceci peut  tre d    un rel chement de la pression de conservation des g nes en plusieurs copies qui aboutit, au long terme (sur une branche suivante),   la disparition d'une des copies.   la diff rence de ce qui semble se passer dans la r alit , notre simulateur consid re que, quand un  v nement de sp ciation a eu lieu, tous les g nes ont la m me probabilit  d' tre supprim s dans la nouvelle branche. Une solution pour mieux reproduire la r alit  serait d'utiliser directement les histoires  volutives de chaque famille de g ne dans notre simulateur.

D.3.4 Simuler les arbres de g nes r els

Dans ce but, nous avons commenc    d velopper une version de MagSimus dans laquelle les arbres de g nes sont utilis s directement pour d terminer les  v nements g niques. Ceci nous permet de reproduire les  volutions g niques sp cifiques   chaque famille de g ne. Par exemple certaines familles de g nes sont conserv es en une copie (certains g nes constitutifs aussi appel s « g nes de m nage » ou « house keeping genes ») dans tous les g nomes alors que d'autres familles sont sujettes   de nombreuses duplications (h moglobine, les g nes de l'olfaction, *etc*).

D.3.5 Insertion d'une copie au voisinage du g ne copi 

Actuellement les duplications se font de deux mani res, soit exactement en tandem et la copie est adjacente au g ne copi , soit les duplications sont

dispersés et la copie est insérée aléatoirement dans un intergène quelconque du génome. Une amélioration de MagSimus serait d'autoriser une duplication en tandem moins draconienne et d'insérer la copie dans le voisinage du gène copié avec une probabilité graduellement décroissante de se trouver éloigné du gène d'origine. Il nous semble que ce type de duplication en tandem reproduirait mieux la réalité, nous invitons le lecteur à se référer à la figure I.15.

D.3.6 Évolution spécifique des chromosomes sexuels

Les chromosomes sexuels évoluent différemment des chromosomes autosomiques et il serait utile de tenir compte de cette spécificité dans MagSimus.

D.3.7 Simuler les erreurs d'assemblage et les erreurs d'annotation

En fin d'évolution MagSimus pourrait comporter une étape supplémentaire qui fragmenterait les génomes et générerait des micro-réarrangements de manière à reproduire les erreurs d'assemblages. Les erreurs d'annotations pourraient quant à elles être simulées en échangeant des gènes d'une famille à une autre et en supprimant quelques gènes. Ceci permettrait par exemple de mettre en évidence la sensibilité des méthodes d'analyse de génomes modernes vis à vis des erreurs d'assemblages et d'annotations.

D.3.8 Paramétrage le plus réaliste

Il est probable que d'autres paramétrages de la distribution Γ aient un écart général à la réalité (ρ) encore plus faible (plus proche de 1) que le paramétrage que nous avons retenu (tableau III.1b) car nous n'avons évalué qu'un nombre restreint de valeurs des paramètres. Néanmoins, étant donné que notre simulateur ne reproduit pas encore la réutilisation des points de cassure, nous n'avons pas souhaité pousser plus loin l'optimisation des paramètres.

D.3.9 Simuler l'évolution des tailles des intergènes

À notre échelle d'étude les longueurs des intergènes ne sont pas considérées. Or nous avons vu (section I.1.8) que certains intergènes sont beaucoup plus grands que d'autres. Par exemple nous avons vu que dans le chromosome 1, le plus grand intergène fait 23 Mb. Cet intergène a très probablement plus de chances d'être brisé qu'un autre intergène plus petit. Ce phénomène peut être la cause de nombreuses réutilisations de points de cassures (section III.8.2).

Dans le but d'estimer à quel point les tailles des intergènes influent sur la dynamique des réarrangements et sur les probabilités de cassures nous avons débuté l'implémentation d'une version de MagSimus qui considère les chromosomes, non plus comme des listes qui ne contiennent que des gènes orientés mais, cette fois-ci, comme des listes avec une alternance de gènes orientés et d'intergènes¹, dont les longueurs en bp sont variables. Dans cette nouvelle modélisation les gènes ont eux aussi des longueurs, mais sont insécables. Lorsqu'un réarrangement a lieu, il modifie les longueurs des intergènes : les intergènes sont édités en fonction des localisations des cassures et des nouvelles jonctions. Nous espérons également qu'en attribuant de manière réaliste des quantités de chromatine ouverte aux intergènes, cette modélisation nous permettra d'estimer le rôle de celle-ci dans les réutilisations de points de cassures [Berthelot *et al.*, 2015].

D.3.10 Réutilisation des points de cassure et résolution des segments conservés

Employer le terme de « réutilisation des points de cassures » pourrait faire croire que les cassures ont lieu exactement aux mêmes endroits durant l'évolution. Pevzner & Tesler ont déjà énoncé une clarification à ce sujet dans le but de contrer une mauvaise interprétation de leurs résultats. « *We emphasize that by reusing breakpoints we do not mean multiple use of exactly the same genomic position as an endpoint of rearrangements, but rather the fact that the breakpoint regions host endpoints for multiple rearrangement events* » [Pevzner et Tesler, 2003b]. Les « breakpoint regions » sont ici les intervalles qui séparent deux blocs de synténie voisins. Il n'est donc pas étonnant que le nombre de réutilisations de points de cassure varie selon la résolution des segments conservés utilisés dans les scénarios de réarrangements. Nous nous attendons à ce que ce nombre soit principalement dépendant de l'identification des petits segments conservés dans les régions fragiles, ceux que Pevzner & Tesler appellent des « hidden blocks ». Certaines études ont fait l'hypothèse que le nombre de réutilisations de points de cassures diminuerait dès que la résolution des segments conservés augmenterait [Sankoff, 2009], néanmoins dans notre cas, même après avoir identifié ce qui nous semble être quelques-uns des « hidden blocks » de Pevzner et Tesler, (par exemple le segment conservé 25 de la figure III.10) nous constatons un nombre de réutilisations toujours élevé pour le chromosome X. Nous pensons tout de même qu'avec des marqueurs encore plus petits le nombre de réutilisations pourrait bien tendre vers zéro.

¹nouvelle entité à ajouter à celles de la section I.2

D.3.11 Quantifier le réalisme des réutilisations des points de cassures

Nous pensons, qu’avec notre quantification de l’écart à la réalité, un simulateur pourrait avoir un écart à la réalité nul ($\rho = 1$), sans pour autant simuler de manière réaliste la réutilisation des points de cassures. Il serait par exemple bienvenu de mettre en place une nouvelle statistique pour quantifier le nombre de réutilisations de points de cassures qui semblent avoir eu lieu entre deux espèces modernes. La statistique réelle serait le nombre de réutilisations de points de cassures (section I.6.2) inféré sur les données réelles, d’après les scénarios de GRIMM et les segments conservés de PhylDiag. Par exemple, $r_r^{1,2}$ serait le nombre de réutilisations de points de cassures (inféré par GRIMM) entre le génome moderne 1 et le génome moderne 2. La statistique simulée, $r_s^{1,2}$, serait le même nombre, mais inféré à partir des génomes correspondants simulés.

D.4 Ouverture

Nous espérons que notre logiciel PhylDiag aidera tous types d’applications qui requièrent l’identification préalable des segments conservés ou l’identification des duplications segmentales. Nous pensons que les segments conservés de PhylDiag pourront aider à mieux reconstruire les génomes ancestraux, notamment grâce à l’inférence de l’orientation des gènes ancestraux dans les segments conservés grâce à l’équation (1) de notre précédent travail [Lucas *et al.*, 2014].

La paramétrisation de MagSimus nous a permis d’estimer les nombres de réarrangements durant l’évolution d’Amniota jusqu’à cinq espèces modernes : l’humain, la souris, l’opossum, le chien et le poulet. Nous pensons que cette méthode d’inférence peut *a priori* être appliquée à d’autres espèces métazoaires et qu’elle pourrait être améliorée en effectuant la recherche des paramètres qui maximisent le réalisme à l’aide d’un algorithme de type ABC [Wegmann *et al.*, 2010].

La paramétrisation réaliste (mis à part les réutilisations des points de cassures) que nous avons trouvée, et qui tend à reproduire une évolution réaliste, des cinq espèces précédentes nous sera d’une grande aide pour comparer, sur un pied d’égalité, les algorithmes de génomique comparative qui étudient les génomes comme des listes ordonnées de gènes orientées. Par exemple, les qualités des reconstructions de l’ordre des gènes dans les génomes ancestraux d’AGORA [Muffato, 2010] et de DeCo [Bérard *et al.*, 2012][Semeria *et al.*, 2015] pourront être comparées. Nous suggérons également d’utiliser

les segments conservés simulés par MagSimus pour évaluer les nombreux algorithmes qui visent à les retrouver [Ghiurcuta et Moret, 2014].

Nous avons implémenté notre compréhension globale des réarrangements chromosomiques dans un simulateur. Les écarts des simulations par rapport à la réalité ont dévoilé des dynamiques dont nous sous-estimions l'importance. Intégrer dans le simulateur les réutilisations de points de cassures nous permettra probablement de dévoiler d'autres dynamiques évolutives majeures, négligées à tort ou inconnues, qui continueront à échapper à la simulation.

Réarranger les chromosomes en les brisant au hasard, avec une répartition relativement uniforme des cassures, nous semblait, *a priori*, suffire à reproduire les vestiges de l'évolution de l'ordre des gènes dans les génomes de vertébrés. Grâce à la simulation et à une identification fine des segments conservés, nous sommes maintenant convaincus que les points de cassures sont voisins les uns des autres plus souvent qu'attendu par une répartition aléatoire (uniforme) des cassures. Deux mécanismes nous semblent pouvoir expliquer simplement ce phénomène : des inversions de segments courts (de environ la taille d'un gène, 24 kb) et/ou des inversions de segments plus longs, mais dont les extrémités seraient localisées préférentiellement dans des régions étroites. Quantifier la fréquence d'un type d'inversions par rapport à l'autre reste à déterminer. Ces deux types d'inversions pourraient être liés à la structure des chromosomes.

Annexes

A.1 Tailles des inversions attendues d'après le RBM

Nous démontrons ici que si les points de cassures des inversions, sont modélisés d'après le RBM, alors la distribution de probabilité des tailles des segments inversés d'un chromosome est triangulaire décroissante.

Le RBM correspond à une répartition aléatoire des points de cassures le long du chromosome. À notre échelle d'étude, cela revient à dire que chaque intergène, ou télomère, a autant de chance d'être brisé par une cassure qu'un autre¹.

Considérons une inversion d'un segment d'un chromosome contenant L gènes. Les cassures ont $L + 1$ localisations possibles le long du chromosome : $x = 0$ pour le télomère de gauche, $x = 1, \dots, x = L - 1$ pour les intergènes ordonnés de gauche à droite, et $x = L$ pour le télomère de droite. Introduisons X_1 (resp. X_2), une variable aléatoire égale à la localisation de la première cassure (resp. de la deuxième). Si chaque cassure est modélisée d'après le RBM, chaque télomère et chaque intergène a la même probabilité d'être cassé qu'un autre intergène ou qu'un autre télomère. La probabilité que la première cassure ait lieu dans l'intergène situé à la localisation x_1 (un entier entre 0 et L) est donc égale à $\frac{1}{L+1}$. Le même raisonnement s'applique également à la deuxième cassure. Par conséquent

$$P(X_1 = x_1) = \frac{1}{L+1} \quad \forall x_1 \in \llbracket 0, L \rrbracket$$

et

$$P(X_2 = x_2) = \frac{1}{L+1} \quad \forall x_2 \in \llbracket 0, L \rrbracket.$$

Nous introduisons également $Y = |X_1 - X_2|$, une variable aléatoire égale à la longueur d'un segment inversé attendue par le RBM. Pour trouver

¹Pour appliquer les résultats de ce chapitre à l'échelle des paires de bases il suffit de remplacer les unités indivisibles, ici les gènes, par des paires de bases.

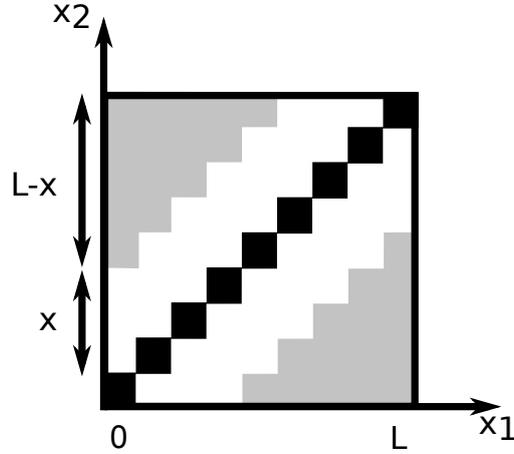


Figure A.1 – La surface Ω_x est en gris et les points tels que $x_1 = x_2$ sont en noir. Pour $x = 3$ et $L = 8$.

la densité de probabilité de Y nous calculons premièrement sa probabilité cumulée $P(Y \leq x)$. Pour tout x dans l'intervalle $\llbracket 0, L \rrbracket$

$$\begin{aligned} P(Y \leq x) &= P(|X_1 - X_2| \leq x) \\ &= 1 - P(|X_1 - X_2| > x) \\ &= 1 - \frac{|\Omega_x|}{(L+1)^2} \end{aligned}$$

avec $|\Omega_x|$ la surface de l'ensemble bidimensionnel Ω_x , et

$$\Omega_x = \{(x_1, x_2) \in \llbracket 0, L \rrbracket^2, |x_1 - x_2| > x\}.$$

La figure A.1 montre graphiquement que cette surface se calcule de la manière suivante :

$$|\Omega_x| = (L-x)(L-x+1). \quad (\text{A.28})$$

D'où

$$P(Y \leq x) = 1 - \frac{(L-x)(L-x+1)}{(L+1)^2}.$$

Nous vérifions que $P(Y \leq L) = 1$ et nous trouvons

$$P(Y \leq 0) = P(Y = 0) = \frac{1}{L+1}.$$

En d'autres termes, la seule manière d'avoir $Y = 0$ est que $X_1 = X_2$, ce qui a $\frac{1}{L+1}$ chances d'arriver. La probabilité d'avoir un segment plus long ($x \geq 1$)

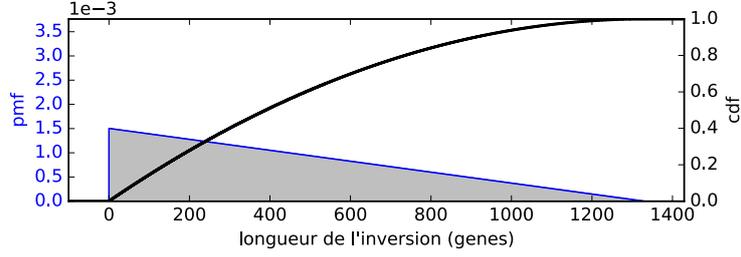


Figure A.2 – Probabilité des inversions (attendue d’après le RBM) en fonction de leurs longueurs. Les probabilités ont été calculées le long d’un chromosome de 1330 gènes.

est

$$\begin{aligned}
 P(Y = x) &= P(Y \leq x) - P(Y \leq x - 1) \\
 &= -\frac{(L - x)(L - x + 1)}{(L + 1)^2} + \frac{(L - x + 1)(L - x + 2)}{(L + 1)^2} \\
 &= \frac{2(L - x + 1)}{(L + 1)^2}.
 \end{aligned} \tag{A.29}$$

Même si cela est anecdotique pour les grands chromosomes, nous mentionnons qu’il est possible de tronquer la loi de probabilité précédente pour exclure les inversions de tailles nulles et l’inversion de taille L , car ces deux derniers types d’inversions ne modifient pas l’ordre des gènes. Nous tronquons donc la loi de probabilité de la manière suivante

$$\begin{aligned}
 P(Y' = x) &= P(Y = x | 1 \leq x \leq L - 1) \quad \forall x \in \llbracket 1, L - 1 \rrbracket \\
 &= \frac{P(Y = x)}{P(Y \leq L - 1) - P(Y = 0)},
 \end{aligned}$$

ce qui nous donne au final la probabilité

$$P(Y' = x) = \frac{2(L - x + 1)}{L(L + 1) - 2} \quad \forall x \in \llbracket 1, L - 1 \rrbracket. \tag{A.30}$$

La figure A.2 représente la variation de cette probabilité pour des longueurs d’inversions entre 1 et $L - 1 = 1329$ gènes.

Cette dernière équation est définie $\forall L \geq 2$, elle est moins élégante que la précédente mais elle peut néanmoins s’avérer utile pour ne considérer que les inversions qui changent l’ordre des gènes, c’est à dire les seules inversions que nous pouvons observer à notre échelle d’étude. Nous vérifions que si $L = 2$, $P(Y' = 1) = 1$ et bien entendu

$$\sum_{x=1}^{L-1} P(Y' = x) = 1$$

A.2 Implémentation informatique

Les logiciels PhylDiag et MagSimus ont été programmés en python à l'aide d'une librairie informatique dédiée. Celle-ci est née il y a un peu moins de 10 ans sous l'impulsion de Matthieu Muffato lors de la programmation du logiciel AGORA [Muffato, 2010] et de l'interface web de l'équipe, GENOMICUS¹ [Louis *et al.*, 2015]. Elle a depuis bénéficié des mises à jours suscitées par les développements de PhylDiag et MagSimus, ainsi que des développements ultérieurs d'AGORA et GENOMICUS. Cette librairie contient de nombreuses classes informatiques conçues sur-mesure pour manipuler les génomes à l'échelle d'étude qui a été définie dans le chapitre I. Nous l'avons quotidiennement utilisée pour uniformiser les travaux informatiques de notre équipe, son développement a été collaboratif et les différentes améliorations ont été organisées avec le logiciel de gestion de versions GIT.

Nous avons de plus développé un *wrapper* en python, pour l'interface HTCondor du cluster de calculs de l'IBENS. D'autres wrappers en python ont également été développés pour i-ADHoRe 3.0, GRIMM et HomologyTeams de manière à pouvoir exploiter ces logiciels avec nos formats de fichiers. Nous les mettons eux-aussi à la disposition du publique dans notre dépôt GitHub.

Nous avons optimisé le temps d'exécution de PhylDiag en utilisant Cython pour traduire en langage C la fonction de calcul des matrices d'homologies. Nous avons utilisé principalement les librairies scipy et numpy pour nos calculs. Les inférences d'évènements le long des branches de l'arbre des espèces, à partir des matrices de distances (section III.2.6), utilisent l'implémentation de la méthode d'optimisation NNLS du package python scipy.optimize. Les graphiques ont été dessinés grâce aux librairies matplotlib et biopython. Les matrices d'homologies et les chromosomes ont été dessinés grâce aux fonctions de dessins vectoriels de notre librairie.

Les codes sources de PhylDiag, de MagSimus, ainsi que celui de notre librairie et de nos programmes d'analyses sont tous disponibles via GitHub (<https://github.com/DyogenIBENS>) et sont sous licences GPL 3 et CeCiLL 2.

¹accessible par le lien www.genomicus.biologie.ens.fr

Glossaire

ABC Approximate Bayesian Computing. 94, 162

AIC Akaike Information Criterion. 117

ARNm ARN messenger. 14, 25

ARNnc ARN non codant. 151

ARNr ARN ribosomique. 16, 19

bp basepair. 2, 12, 21, 22, 54, 92, 96, 147, 161

BRH Best Reciprocal Hit. 51, 76, 77

CCDS Consensus Coding Sequences. 17, 144

CD Chebyshev Distance. 47, 48, 154, 155

cdf Cumulative Distribution Function. 115, 129

CDS Coding Sequence. 14

cM centimorgan. 9

CNV Copy Number Variation. 3, 16

CTCF CCCTC-binding factor. 2, 17

DPD Diagonal Pseudo-Distance. 47, 48, 154

ED Euclidian Distance. 47, 48

FBM Fragile Breakage Model. 9, 79, 144

FISH Fluorescent *In-Situ* Hybridization. 56

Gb gigabase. 92

GRB Genomic Regulatory Block. 2, 6, 9

HDR Highly Divergent Region. 5

HSB Homologous Synteny Block. 72

imcs Identify Mono-genic Conserved Segments. 87

imr Identify Micro-Rearrangements. 86, 87

kb kilobase. 1, 2, 6, 8, 15, 21, 140, 141, 152, 163

KS Statistique de Kolmogorov-Smirnov. 129

LCA Last Common Ancestor. 33

LCB Locally Collinear Block. 72, 77, 78

LIS Longest Increasing Subsequence. 78

LTEE Long-Term Experimental Evolution. 92

Mb mégabase. 2, 6, 9, 12, 15, 19–21, 115, 117, 121, 143, 146, 160

MD Manhattan Distance. 47, 48, 154

MH Matrice d’Homologies. 45, 46

MHP Matrice de Packs d’Homologies. 46, 71, 75

MRC Most Recent Common Ancestor. 33–37, 45, 51, 53, 64, 78, 93, 126

MSR Maximum Strip Recovery. 77

MUM Maximum Unique Match. 78

NNLS Non-Negative Least Squares. 106, 167

ORF Open Reading Frame. 15, 16

pmf Probability Mass Function. 115

RBM Random Breakage Model. 8, 9, 79, 105, 114, 143, 144, 164, 166

t Truncation. 87

TAD Topologically Associated Domain. 2, 3

TFBS Transcription Factor Binding Site. 14

UTR Untranslated region. 14, 25

V(D)J mécanisme de recombinaison de l'ADN site-spécifique qui permet de diversifier les anticorps. 3

WGD Whole-Genome Duplication. 31, 96

XAR X Added Region. 103

Bibliographie

- [Ahituv *et al.*, 2007] AHITUV, N., ZHU, Y., VISEL, A., HOLT, A., AFZAL, V., PENNACCHIO, L. A. et RUBIN, E. M. (2007). Deletion of Ultraconserved Elements Yields Viable Mice. *PLoS Biology*, 5(9):e234.
- [Aiden et Casellas, 2015] AIDEN, E. L. et CASELLAS, R. (2015). Somatic Rearrangement in B Cells: It's (Mostly) Nuclear Physics. *Cell*, 162(4):708–11.
- [Alekseyev et Pevzner, 2008] ALEKSEYEV, M. A. et PEVZNER, P. A. (2008). Multi-break rearrangements and chromosomal evolution. *Theoretical Computer Science*, 395(2-3):193–202.
- [Altschul *et al.*, 1990] ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W. et LIPMAN, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–10.
- [Andrey *et al.*, 2013] ANDREY, G., MONTAVON, T., MASCREZ, B., GONZALEZ, F., NOORDERMEER, D., LELEU, M., TRONO, D., SPITZ, F. et DUBOULE, D. (2013). A Switch Between Topological Domains Underlies HoxD Genes Collinearity in Mouse Limbs. *Science*, 340(6137):1234167–1234167.
- [Angiuoli et Salzberg, 2011] ANGIUOLI, S. V. et SALZBERG, S. L. (2011). Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics*, 27(3):334–42.
- [Argueso *et al.*, 2008] ARGUESO, J. L., WESTMORELAND, J., MIECZKOWSKI, P. A., GAWEL, M., PETES, T. D. et RESNICK, M. A. (2008). Double-strand breaks associated with repetitive DNA can reshape the genome. *Proceedings of the National Academy of Sciences*, 105(33):11845–11850.
- [Arkendra *et al.*, 2001] ARKENDRA, D. E., FERGUSON, M., SINDI, S. et DURRETT, R. (2001). The equilibrium distribution for a generalized

- Sankoff-Ferretti model accurately predicts chromosome size distributions in a wide variety of species. *Journal of Applied Probability*, 38(2):324–334.
- [Bafna *et al.*, 1996] BAFNA, V., NARAYANAN, B. et RAVI, R. (1996). Nonoverlapping local alignments (weighted independent sets of axis-parallel rectangles). *Discrete Applied Mathematics*, 71(1-3):41–53.
- [Batut *et al.*, 2013] BATUT, B., PARSONS, D. P., FISCHER, S., BESLON, G. et KNIBBE, C. (2013). In silico experimental evolution: a tool to test evolutionary scenarios. *BMC bioinformatics*, 14 Suppl 1(Suppl 15):S11.
- [Baudet *et al.*, 2010] BAUDET, C., LEMAITRE, C., DIAS, Z., GAUTIER, C., TANNIER, E. et SAGOT, M.-F. (2010). Cassis: detection of genomic rearrangement breakpoints. *Bioinformatics*, 26(15):1897–8.
- [Becker et Lenhard, 2007] BECKER, T. S. et LENHARD, B. (2007). The random versus fragile breakage models of chromosome evolution: A matter of resolution. *Molecular Genetics and Genomics*, 278(5):487–491.
- [Bejerano *et al.*, 2004] BEJERANO, G., PHEASANT, M., MAKUNIN, I., STEPHEN, S., KENT, W. J., MATTICK, J. S. et HAUSSLER, D. (2004). Ultraconserved elements in the human genome. *Science*, 304(5675):1321–1325.
- [Benko *et al.*, 2009] BENKO, S., FANTES, J. a., AMIEL, J., KLEINJAN, D.-J., THOMAS, S., RAMSAY, J., JAMSHIDI, N., ESSAFI, A., HEANEY, S., GORDON, C. T., MCBRIDE, D., GOLZIO, C., FISHER, M., PERRY, P., ABADIE, V., AYUSO, C., HOLDER-ESPINASSE, M., KILPATRICK, N., LEES, M. M., PICARD, A., TEMPLE, I. K., THOMAS, P., VAZQUEZ, M.-P., VEKEMANS, M., ROEST CROLLIUS, H., HASTIE, N. D., MUNNICH, A., ETCHEVERS, H. C., PELET, A., FARLIE, P. G., FITZPATRICK, D. R. et LYONNET, S. (2009). Highly conserved non-coding elements on either side of SOX9 associated with Pierre Robin sequence. *Nature genetics*, 41(3):359–364.
- [Bérard *et al.*, 2012] BÉRARD, S., GALLIEN, C., BOUSSAU, B., SZOLLOSI, G. J., DAUBIN, V. et TANNIER, E. (2012). Evolution of gene neighborhoods within reconciled phylogenies. *Bioinformatics*, 28(18):i382–i388.
- [Bergeron *et al.*, 2002] BERGERON, A., CORTEEL, S. et RAFFINOT, M. (2002). The algorithmic of gene teams. *Algorithms in Bioinformatics*, 2452:464–476.

- [Berthelot *et al.*, 2015] BERTHELOT, C., MUFFATO, M., ABECASSIS, J. et ROEST CROLLIUS, H. (2015). The 3D Organization of Chromatin Explains Evolutionary Fragile Genomic Regions. *Cell Reports*, 10(11):1913–1924.
- [Blanchette *et al.*, 1997] BLANCHETTE, M., BOURQUE, G. et SANKOFF, D. (1997). Breakpoint phylogenies. *Genome Informatics*, 5(3):555–570.
- [Boffelli *et al.*, 2004] BOFFELLI, D., NOBREGA, M. a. et RUBIN, E. M. (2004). Comparative genomics at the vertebrate extremes. *Nature reviews Genetics*, 5(6):456–465.
- [Bourke et Mank, 2013] BOURKE, A. F. G. et MANK, J. E. (2013). Genetics: A social rearrangement. *Nature*, 493(7434):612–613.
- [Bourque *et al.*, 2004] BOURQUE, G., PEVZNER, P. A. et TESLER, G. (2004). Reconstructing the genomic architecture of ancestral mammals: lessons from human, mouse, and rat genomes. *Genome research*, 14(4):507–16.
- [Bourque *et al.*, 2005] BOURQUE, G., ZDOBNOV, E. M., BORK, P., PEVZNER, P. A. et TESLER, G. (2005). Comparative architectures of mammalian and chicken genomes reveal highly variable rates of genomic rearrangements across different lineages. *Genome research*, 15(1):98–110.
- [Braga *et al.*, 2013] BRAGA, M. D. V., CHAUVE, C., DOERR, D., JAHN, K., STOYE, J., THÉVENIN, A. et WITTLER, R. (2013). The Potential of Family-Free Genome Comparison. In *Models and Algorithms for Genome Evolution SE - 13*, volume 19, pages 287–307. Springer London.
- [Bramanti *et al.*, 2009] BRAMANTI, B., THOMAS, M. G., HAAK, W., UNTERLAENDER, M., JORES, P., TAMBETS, K., ANTANAITIS-JACOBS, I., HAIDLE, M. N., JANKAUSKAS, R., KIND, C.-J., LUETH, F., TERBERGER, T., HILLER, J., MATSUMURA, S., FORSTER, P. et BURGER, J. (2009). Genetic discontinuity between local hunter-gatherers and central Europe’s first farmers. *Science*, 326(5949):137–140.
- [Brylinski, 2013] BRYLINSKI, M. (2013). eVolver: an optimization engine for evolving protein sequences to stabilize the respective structures. *BMC research notes*, 6(1):303.
- [Bulteau *et al.*, 2013] BULTEAU, L., FERTIN, G. et RUSU, I. (2013). Maximal strip recovery problem with gaps: Hardness and approximation algorithms. *Journal of Discrete Algorithms*, 19:1–22.

- [Calabrese *et al.*, 2003] CALABRESE, P. P., CHAKRAVARTY, S. et VISION, T. J. (2003). Fast identification and statistical evaluation of segmental homologies in comparative maps. *Bioinformatics*, 19(suppl 1):i74–i80.
- [Cannon *et al.*, 2003] CANNON, S. B., KOZIK, A., CHAN, B., MICHELMORE, R. et YOUNG, N. D. (2003). DiagHunter and GenoPix2D: programs for genomic comparisons, large-scale homology discovery and visualization. *Genome biology*, 4(10):R68.
- [Carvunis *et al.*, 2012] CARVUNIS, A.-R., ROLLAND, T., WAPINSKI, I., CALDERWOOD, M. a., YILDIRIM, M. a., SIMONIS, N., CHARLOTEAUX, B., HIDALGO, C. a., BARBETTE, J., SANTHANAM, B., BRAR, G. a., WEISSMAN, J. S., REGEV, A., THIERRY-MIEG, N., CUSICK, M. E. et VIDAL, M. (2012). Proto-genes and de novo gene birth. *Nature*, pages 3–7.
- [Chaisson *et al.*, 2006] CHAISSON, M. J., RAPHAEL, B. J. et PEVZNER, P. a. (2006). Microinversions in mammalian evolution. *Proceedings of the National Academy of Sciences*, 103(52):19824–19829.
- [Chen *et al.*, 2007] CHEN, C. T. L., WANG, J. C. et COHEN, B. a. (2007). The strength of selection on ultraconserved elements in the human genome. *American journal of human genetics*, 80(4):692–704.
- [Choi *et al.*, 2007] CHOI, V., ZHENG, C., ZHU, Q. et SANKOFF, D. (2007). Algorithms for the extraction of synteny blocks from comparative maps. *Algorithms in Bioinformatics*, pages 277–288.
- [Coghlan, 2002] COGHLAN, A. (2002). Fourfold Faster Rate of Genome Rearrangement in Nematodes Than in Drosophila. *Genome Research*, 12(6):857–867.
- [Cooper, 2000] COOPER, G. M. (2000). The Complexity of Eukaryotic Genomes. *In The Cell: A Molecular Approach. 2nd edition.* Sinauer Associates, sunderland édition.
- [Cremer et Cremer, 2010] CREMER, T. et CREMER, M. (2010). Chromosome territories. *Cold Spring Harbor perspectives in biology*, 2(3):1–22.
- [Dabney *et al.*, 2013] DABNEY, J., MEYER, M. et PAABO, S. (2013). Ancient DNA Damage. *Cold Spring Harbor Perspectives in Biology*, 5(7):a012567–a012567.
- [Dalquen *et al.*, 2012] DALQUEN, D. a., ANISIMOVA, M., GONNET, G. H. et DESSIMOZ, C. (2012). ALF—a simulation framework for genome evolution. *Molecular biology and evolution*, 29(4):1115–23.

- [Darling *et al.*, 2004] DARLING, A. E., MAU, B., BLATTNER, F. R. et PERNA, N. T. (2004). GRIL: Genome rearrangement and inversion locator. *Bioinformatics*, 20(1):122–124.
- [Darling *et al.*, 2010] DARLING, A. E., MAU, B. et PERNA, N. T. (2010). Progressivemauve: Multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE*, 5(6).
- [Darwin, 1860] DARWIN, C. (1860). Letter. *Gardeners' Chronicle and Agricultural Gazette*, 362(21 april):3.
- [de Laat et Duboule, 2013] de LAAT, W. et DUBOULE, D. (2013). Topology of mammalian developmental enhancers and their regulatory landscapes. *Nature*, 502(7472):499–506.
- [Dehal et Boore, 2005] DEHAL, P. et BOORE, J. L. (2005). Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biology*, 3(10).
- [Delcher *et al.*, 1999] DELCHER, A. L., KASIF, S., FLEISCHMANN, R. D., PETERSON, J., WHITE, O. et SALZBERG, S. L. (1999). Alignment of whole genomes. *Nucleic Acids Research*, 27(11):2369–2376.
- [Delpretti *et al.*, 2013] DELPRETTI, S., MONTAVON, T., LELEU, M., JOYE, E., TZIKA, A., MILINKOVITCH, M. et DUBOULE, D. (2013). Multiple Enhancers Regulate Hoxd Genes and the Hotdog LncRNA during Cecum Budding. *Cell Reports*, 5(1):137–150.
- [Deonier *et al.*, 2005] DEONIER, R. C., TAVARÉ, S. et WATERMAN, M. S. (2005). *Computational Genome Analysis*. Springer New York, New York, NY, springer édition.
- [Dewey, 2011] DEWEY, C. N. (2011). Positional orthology: Putting genomic evolutionary relationships into context. *Briefings in Bioinformatics*, 12(5): 401–412.
- [Drillon *et al.*, 2014] DRILLON, G., CARBONE, A. et FISCHER, G. (2014). SynChro: A Fast and Easy Tool to Reconstruct and Visualize Synteny Blocks along Eukaryotic Chromosomes. *PloS one*, 9(3):e92621.
- [Dubchak *et al.*, 2009] DUBCHAK, I., POLIAKOV, A., KISLYUK, A. et BRUDNO, M. (2009). Multiple whole-genome alignments without a reference organism. *Genome research*, 19(4):682–689.

- [Dupressoir *et al.*, 2009] DUPRESSOIR, A., VERNOCHET, C., BAWA, O., HARPER, F., PIERRON, G., OPOLON, P. et HEIDMANN, T. (2009). Syncytin-A knockout mice demonstrate the critical role in placentation of a fusogenic, endogenous retrovirus-derived, envelope gene. *Proceedings of the National Academy of Sciences*, 106(29):12127–12132.
- [Edmonds, 1963] EDMONDS, J. (1963). Paths, trees, and flowers. *Canad J Math*.
- [Felsenfeld et Groudine, 2003] FELSENFELD, G. et GROUDINE, M. (2003). Controlling the double helix. *Nature*, 421(6921):448–453.
- [Felsenstein, 2004] FELSENSTEIN, J. (2004). Ch1 Parsimony methods. In *Infering Phylogenies*, pages 1–10. Sinauer Associates; 2 edition.
- [Feuk et Carson, 2006] FEUK, L. et CARSON, A. (2006). Structural variation in the human genome. *Nature Reviews Genetics*, 7(February):1169–1171.
- [Fleming et Wallace, 1986] FLEMING, P. J. et WALLACE, J. J. (1986). How not to lie with statistics: the correct way to summarize benchmark results. *Communications of the ACM*, 29(3):218–221.
- [Ford et Fulkerson, 1956] FORD, L. R. et FULKERSON, D. R. (1956). Maximal flow through a network. *Canadian Journal of Mathematics*, 8:399–404.
- [Gagnon *et al.*, 2012] GAGNON, Y., BLANCHETTE, M. et EL-MABROUK, N. (2012). A flexible ancestral genome reconstruction method based on gapped adjacencies. *BMC Bioinformatics*, 13(Suppl 19):S4.
- [Gerstein *et al.*, 2007] GERSTEIN, M. B., BRUCE, C., ROZOWSKY, J. S., ZHENG, D., DU, J., KORBEL, J. O., EMANUELSSON, O., ZHANG, Z. D., WEISSMAN, S. et SNYDER, M. (2007). What is a gene, post-ENCODE? History and updated definition. *Genome Research*, 17(6):669–681.
- [Ghiurcuta et Moret, 2014] GHIURCUTA, C. G. et MORET, B. M. E. (2014). Evaluating synteny for improved comparative studies. *Bioinformatics*, 30(12):i9–i18.
- [Graur *et al.*, 2013] GRAUR, D., ZHENG, Y., PRICE, N., AZEVEDO, R. B. R., ZUFALL, R. a. et ELHAIK, E. (2013). On the immortality of television sets: "Function" in the human genome according to the evolution-free gospel of encode. *Genome Biology and Evolution*, 5(3):578–590.

- [Graves, 2015] GRAVES, J. A. M. (2015). Evolution of vertebrate sex chromosomes and dosage compensation. *Nature reviews Genetics*, 17(1):33–46.
- [Green *et al.*, 2014] GREEN, R. E., BRAUN, E. L., ARMSTRONG, J., EARL, D., NGUYEN, N., HICKEY, G., VANDEWEGE, M. W., JOHN, J. a. S., CAPELLA-GUTIÉRREZ, S., CASTOE, T. a., KERN, C., FUJITA, M. K., OPAZO, J. C., JURKA, J., KOJIMA, K. K., CABALLERO, J., HUBLEY, R. M., SMIT, A. F., PLATT, R. N., LAVOIE, C. a., RAMAKODI, M. P., JR, J. W. F., SUH, A., ISBERG, S. R., MILES, L., CHONG, A. Y., JARATLERDSIRI, W., GONGORA, J., MORAN, C., IRIARTE, A., MCCORMACK, J., BURGESS, S. C., EDWARDS, S. V., LYONS, E., WILLIAMS, C., BREEN, M., HOWARD, J. T., GRESHAM, C. R., PETERSON, D. G., SCHMITZ, J., POLLOCK, D. D., HAUSSLER, D., TRIPLETT, E. W., ZHANG, G., IRIE, N., JARVIS, E. D., BROCHU, C. a., SCHMIDT, C. J., MCCARTHY, F. M., FAIRCLOTH, B. C., HOFFMANN, F. G., GLENN, T. C., GABALDÓN, T., PATEN, B., RAY, D. a., ST. JOHN, J. a., CAPELLA-GUTIERREZ, S., CASTOE, T. a., KERN, C., FUJITA, M. K., OPAZO, J. C., JURKA, J., KOJIMA, K. K., CABALLERO, J., HUBLEY, R. M., SMIT, A. F., PLATT, R. N., LAVOIE, C. a., RAMAKODI, M. P., FINGER, J. W., SUH, A., ISBERG, S. R., MILES, L., CHONG, A. Y., JARATLERDSIRI, W., GONGORA, J., MORAN, C., IRIARTE, A., MCCORMACK, J., BURGESS, S. C., EDWARDS, S. V., LYONS, E., WILLIAMS, C., BREEN, M., HOWARD, J. T., GRESHAM, C. R., PETERSON, D. G., SCHMITZ, J., POLLOCK, D. D., HAUSSLER, D., TRIPLETT, E. W., ZHANG, G., IRIE, N., JARVIS, E. D., BROCHU, C. a., SCHMIDT, C. J., MCCARTHY, F. M., FAIRCLOTH, B. C., HOFFMANN, F. G., GLENN, T. C., GABALDON, T., PATEN, B. et RAY, D. a. (2014). Three crocodylian genomes reveal ancestral patterns of evolution among archosaurs. *Science*, 346(6215):1254449–1254449.
- [Gregory, 2016] GREGORY, T. (2016). Animal Genome Size Database.
- [Griffiths, 2005] GRIFFITHS, A. J. (2005). *An Introduction to Genetic Analysis (8th edition)*. W.H. Freeman and Company.
- [Han et Hahn, 2009] HAN, M. V. et HAHN, M. W. (2009). Identifying parent-daughter relationships among duplicated genes. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 125:114–125.
- [Hannenhalli et Pevzner, 1995] HANNENHALLI, S. et PEVZNER, P. a. (1995). Transforming cabbage into turnip. *Proceedings of the twenty-seventh annual ACM symposium on Theory of computing (STOC)*, 46(1):178–189.

- [Harewood et Fraser, 2014] HAREWOOD, L. et FRASER, P. (2014). The impact of chromosomal rearrangements on regulation of gene expression. *Human Molecular Genetics*, 23(R1):76–82.
- [Harewood et al., 2010] HAREWOOD, L., SCHÜTZ, F., BOYLE, S., PERRY, P., DELORENZI, M., BICKMORE, W. A. et REYMOND, A. (2010). The effect of translocation-induced nuclear reorganization on gene expression. *Genome Research*, 20(5):554–564.
- [He et Goldwasser, 2005] HE, X. et GOLDWASSER, M. M. H. (2005). Identifying conserved gene clusters in the presence of homology families. *Journal of computational biology*, 12(6):638–656.
- [Hedges et al., 2015] HEDGES, S. B., MARIN, J., SULESKI, M., PAYMER, M. et KUMAR, S. (2015). Tree of life reveals clock-like speciation and diversification. *Molecular Biology and Evolution*, 32(4):835–845.
- [Hoberman et al., 2005] HOBERMAN, R., SANKOFF, D. et DURAND, D. (2005). The statistical analysis of spatially clustered genes under the maximum gap criterion. *Journal of computational biology : a journal of computational molecular cell biology*, 12(8):1083–102.
- [Hoffmann et Rieseberg, 2008] HOFFMANN, A. a. et RIESEBERG, L. H. (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.
- [Hurst et al., 2004] HURST, L. D. L., PÁL, C. et LERCHER, M. J. M. (2004). The evolutionary dynamics of eukaryotic gene order. *Nature reviews Genetics*, 5(4):299–310.
- [Jaillon et al., 2004] JAILLON, O., AURY, J.-M., BRUNET, F., PETIT, J.-L., STANGE-THOMANN, N., MAUCELI, E., BOUNEAU, L., FISCHER, C., OZOUF-COSTAZ, C., BERNOT, A., NICAUD, S., JAFFE, D., FISHER, S., LUTFALLA, G., DOSSAT, C., SEGURENS, B., DASILVA, C., SALANOUBAT, M., LEVY, M., BOUDET, N., CASTELLANO, S., ANTHOUARD, V., JUBIN, C., CASTELLI, V., KATINKA, M., VACHERIE, B., BIÉMONT, C., SKALLI, Z., CATTOLICO, L., POULAIN, J., DE BERARDINIS, V., CRUAUD, C., DUPRAT, S., BROTTIER, P., COUTANCEAU, J.-P., GOUZY, J., PARRA, G., LARDIER, G., CHAPPLE, C., MCKERNAN, K. J., MCEWAN, P., BOSAK, S., KELLIS, M., VOLFF, J.-N., GUIGÓ, R., ZODY, M. C., MESIROV, J., LINDBLAD-TOH, K., BIRREN, B., NUSBAUM, C., KAHN, D., ROBINSON-RECHAVI, M., LAUDET, V., SCHACHTER, V., QUÉTIER, F., SAURIN, W., SCARPELLI, C.,

- WINCKER, P., LANDER, E. S., WEISSENBACH, J. et ROEST CROLLIUS, H. (2004). Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature*, 431(7011):946–957.
- [Jiggins, 2015] JIGGINS, C. D. (2015). A flamboyant behavioral polymorphism is controlled by a lethal supergene. *Nature Genetics*, 48(1):7–8.
- [Joron *et al.*, 2011] JORON, M., FREZAL, L., JONES, R. T., CHAMBERLAIN, N. L., LEE, S. F., HAAG, C. R., WHIBLEY, A., BECUWE, M., BAXTER, S. W., FERGUSON, L., WILKINSON, P. a., SALAZAR, C., DAVIDSON, C., CLARK, R., QUAIL, M. a., BEASLEY, H., GLITHERO, R., LLOYD, C., SIMS, S., JONES, M. C., ROGERS, J., JIGGINS, C. D. et FRENCH-CONSTANT, R. H. (2011). Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature*, 477(7363):203–206.
- [Kellis *et al.*, 2014] KELLIS, M., WOLD, B., SNYDER, M. P., BERNSTEIN, B. E., KUNDAJE, A., MARINOV, G. K., WARD, L. D., BIRNEY, E., CRAWFORD, G. E., DEKKER, J., DUNHAM, I., ELNITSKI, L. L., FARNHAM, P. J., FEINGOLD, E. A., GERSTEIN, M., GIDDINGS, M. C., GILBERT, D. M., GINGERAS, T. R., GREEN, E. D., GUIGO, R., HUBBARD, T., KENT, J., LIEB, J. D., MYERS, R. M., PAZIN, M. J., REN, B., STAMATOYANNOPOULOS, J. A., WENG, Z., WHITE, K. P. et HARDISON, R. C. (2014). Defining functional DNA elements in the human genome. *Proceedings of the National Academy of Sciences*, 111(17):6131–8.
- [Kent *et al.*, 2003] KENT, W. J., BAERTSCH, R., HINRICHS, A., MILLER, W. et HAUSSLER, D. (2003). Evolution’s cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. *Proceedings of the National Academy of Sciences*, 100(20):11484–11489.
- [Kikuta *et al.*, 2007] KIKUTA, H., LAPLANTE, M., NAVRATILOVA, P., KOMISARCZUK, A. Z., ENGSTRÖM, P. G., FREDMAN, D., AKALIN, A., CACCAMO, M., SEALY, I., HOWE, K., GHISLAIN, J., PEZERON, G., MOURRAIN, P., ELLINGSEN, S., OATES, A. C., THISSE, C., THISSE, B., FOUCHER, I., ADOLF, B., GELING, A., LENHARD, B. et BECKER, T. S. (2007). Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome research*, 17(5):545–55.
- [Kimura, 1984] KIMURA, M. (1984). *The Neutral Theory of Molecular Evolution*. Cambridge University Press.
- [Kioussis, 2005] KIOUSSIS, D. (2005). Gene regulation: Kissing chromosomes. *Nature*, 435(7042):579–580.

- [Kirkpatrick, 2010] KIRKPATRICK, M. (2010). How and why chromosome inversions evolve. *PLoS Biology*, 8(9).
- [Kirkwood, 1979] KIRKWOOD (1979). Geometric means and measures of dispersion. *Biometrics*, 35:908–909.
- [Koonin, 2009] KOONIN, E. V. (2009). Evolution of genome architecture. *International Journal of Biochemistry and Cell Biology*, 41(2):298–306.
- [Küpper *et al.*, 2015] KÜPPER, C., STOCKS, M., RISSE, J. E., REMEDIOS, N., FARRELL, L. L., MCRAE, B., MORGAN, T. C., KARLIONOVA, N., PINCHUK, P., VERKUIL, Y. I., KITAYSKY, A. S., WINGFIELD, J. C., PIERSMA, T., ZENG, K., SLATE, J., BLAXTER, M., LANK, D. B. et BURKE, T. (2015). A supergene determines highly divergent male reproductive morphs in the ruff. *Nature Genetics*, 48(1):79–83.
- [Kurtz *et al.*, 2004] KURTZ, S., PHILLIPPY, A., DELCHER, A. L., SMOOT, M., SHUMWAY, M., ANTONESCU, C. et SALZBERG, S. L. (2004). Versatile and open software for comparing large genomes. *Genome biology*, 5(2):R12.
- [Lanctôt *et al.*, 2007] LANCTÔT, C., CHEUTIN, T., CREMER, M., CAVALLI, G. et CREMER, T. (2007). Dynamic genome architecture in the nuclear space: regulation of gene expression in three dimensions. *Nature reviews Genetics*, 8(2):104–115.
- [Lander *et al.*, 2001] LANDER, E. S., HEAFORD, a., SHERIDAN, a., LINTON, L. M., BIRREN, B., SUBRAMANIAN, a., COULSON, a., NUSBAUM, C., ZODY, M. C., DUNHAM, a., BALDWIN, J., HUNT, a., DEVON, K., McMURRAY, a., MUNGALL, a., DEWAR, K., DOYLE, M., MARRA, M. a., FULTON, L. a., FITZHUGH, W., CHINWALLA, a. T., FUNKE, R., DELEHAUNTY, a., GAGE, D., HARRIS, K., OLSEN, a., GIBBS, R. a., FUJIYAMA, a., HOWLAND, J., TOYODA, a., KANN, L., ROSENTHAL, a., LEHOCZKY, J., LEVINE, R., RUMP, a., MCEWAN, P., MADAN, a., MCKERNAN, K., FEDERSPIEL, N. a., ABOLA, a. P., MELDRIM, J., MESIROV, J. P., EVANS, G. a., ROE, B. a., MIRANDA, C., DE LA BASTIDE, M., MORRIS, W., BAILEY, J. a., NAYLOR, J., BATEMAN, a., RAYMOND, C., ROSETTI, M., JONES, T. a., KASPRYZK, a., SANTOS, R., MCLYSAGHT, a., SOUGNEZ, C., SMIT, a. F., WILLIAMS, a., STANGE-THOMANN, N., STOJANOVIC, N., FELSENFELD, a., WETTERSTRAND, K. a., WYMAN, D., PATRINOS, a., ROGERS, J., SULSTON, J., AINSCOUGH, R., BECK, S., BENTLEY, D., BURTON, J., CLEE, C., CARTER, N., DEADMAN, R., DELOUKAS, P., DUNHAM, I., DURBIN, R., FRENCH, L., GRAFHAM, D., GREGORY, S., HUBBARD, T., HUMPHRAY, S., JONES, M., LLOYD, C., MATTHEWS, L., MERCER, S.,

MILNE, S., MULLIKIN, J. C., PLUMB, R., ROSS, M., SHOWNKEEN, R., SIMS, S., WATERSTON, R. H., WILSON, R. K., HILLIER, L. W., MCPHERSON, J. D., MARDIS, E. R., PEPIN, K. H., GISH, W. R., CHISSOE, S. L., WENDL, M. C., DELEHAUNTY, K. D., MINER, T. L., KRAMER, J. B., COOK, L. L., FULTON, R. S., JOHNSON, D. L., MINX, P. J., CLIFTON, S. W., HAWKINS, T., BRANSCOMB, E., PREDKI, P., RICHARDSON, P., WENNING, S., SLEZAK, T., DOGGETT, N., CHENG, J. F., LUCAS, S., ELKIN, C., UBERBACHER, E., FRAZIER, M., MUZNY, D. M., SCHERER, S. E., BOUCK, J. B., SODERGREN, E. J., WORLEY, K. C., RIVES, C. M., GORRELL, J. H., METZKER, M. L., NAYLOR, S. L., KUCHERLAPATI, R. S., NELSON, D. L., WEINSTOCK, G. M., SAKAKI, Y., HATTORI, M., YADA, T., ITOH, T., KAWAGOE, C., WATANABE, H., TOTOKI, Y., TAYLOR, T., WEISSENBACH, J., HEILIG, R., SAURIN, W., ARTIGUENAVE, F., BROTTIER, P., BRULS, T., PELLETIER, E., ROBERT, C., WINCKER, P., SMITH, D. R., DOUCETTE-STAMM, L., RUBENFIELD, M., WEINSTOCK, K., LEE, H. M., DUBOIS, J., PLATZER, M., NYAKATURA, G., TAUDIEN, S., YANG, H., YU, J., WANG, J., HUANG, G., GU, J., HOOD, L., ROWEN, L., QIN, S., DAVIS, R. W., PROCTOR, M. J., MYERS, R. M., SCHMUTZ, J., DICKSON, M., GRIMWOOD, J., COX, D. R., OLSON, M. V., KAUL, R., SHIMIZU, N., KAWASAKI, K., MINOSHIMA, S., ATHANASIOU, M., SCHULTZ, R., CHEN, F., PAN, H., RAMSER, J., LEHRACH, H., REINHARDT, R., MCCOMBIE, W. R., DEDHIA, N., BLÖCKER, H., HORNISCHER, K., NORDSIEK, G., AGARWALA, R., ARAVIND, L., BATZOGLOU, S., BIRNEY, E., BORK, P., BROWN, D. G., BURGE, C. B., CERUTTI, L., CHEN, H. C., CHURCH, D., CLAMP, M., COPLEY, R. R., DOERKS, T., EDDY, S. R., EICHLER, E. E., FUREY, T. S., GALAGAN, J., GILBERT, J. G., HARMON, C., HAYASHIZAKI, Y., HAUSSLER, D., HERMJAKOB, H., HOKAMP, K., JANG, W., JOHNSON, L. S., KASIF, S., KENNEDY, S., KENT, W. J., KITTS, P., KOONIN, E. V., KORF, I., KULP, D., LANCET, D., LOWE, T. M., MIKKELSEN, T., MORAN, J. V., MULDER, N., POLLARA, V. J., PONTING, C. P., SCHULER, G., SCHULTZ, J., SLATER, G., STUPKA, E., SZUSTAKOWSKI, J., THIERRY-MIEG, D., THIERRY-MIEG, J., WAGNER, L., WALLIS, J., WHEELER, R., WOLF, Y. I., WOLFE, K. H., YANG, S. P., YEH, R. F., COLLINS, F., GUYER, M. S., PETERSON, J., MORGAN, M. J., de JONG, P., CATANESE, J. J., OSOEGAWA, K., SHIZUYA, H., CHOI, S., CHEN, Y. J. et SZUSTAKOWKI, J. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

[Lefebvre *et al.*, 2003] LEFEBVRE, J. F., EL-MABROUK, N., TILLIER, E. et SANKOFF, D. (2003). Detection and validation of single gene inversions. *Bioinformatics*, 19(SUPPL. 1):190–196.

- [Leffler *et al.*, 2012] LEFFLER, E. M., BULLAUGHEY, K., MATUTE, D. R., MEYER, W. K., SÉGUREL, L., VENKAT, A., ANDOLFATTO, P. et PRZEWSKI, M. (2012). Revisiting an old riddle: what determines genetic diversity levels within species? *PLoS biology*, 10(9):e1001388.
- [Lemaitre, 2008] LEMAITRE, C. (2008). *Réarrangements chromosomiques dans les génomes de mammifères: caractérisation des points de cassure*. Thèse de doctorat, Université Claude Bernard Lyon 1.
- [Lemaitre *et al.*, 2010] LEMAITRE, C., BRAGA, M. D. V., GAUTIER, C., SAGOT, M. F., TANNIER, E. et MARAIS, G. A. B. (2010). Footprints of Inversions at Present and Past Pseudoautosomal Boundaries in Human Sex Chromosomes. *Genome Biology and Evolution*, 1(0):56–66.
- [Lemaitre et Sagot, 2008] LEMAITRE, C. et SAGOT, M.-F. (2008). A small trip in the untranquil world of genomes. *Theoretical Computer Science*, 395(2-3):171–192.
- [Lemaitre *et al.*, 2008] LEMAITRE, C., TANNIER, E., GAUTIER, C. et SAGOT, M.-F. (2008). Precise detection of rearrangement breakpoints in mammalian chromosomes. *BMC bioinformatics*, 9:286.
- [Lemaitre *et al.*, 2009] LEMAITRE, C., ZAGHLOUL, L., SAGOT, M.-F., GAUTIER, C., ARNEODO, A., TANNIER, E. et AUDIT, B. (2009). Analysis of fine-scale mammalian evolutionary breakpoints provides new insight into their relation to genome organisation. *BMC genomics*, 10:335.
- [Lenski *et al.*, 2003] LENSKI, R. E., OFRIA, C., PENNOCK, R. T. et ADAMI, C. (2003). The evolutionary origin of complex features. *Nature*, 423(6936):139–44.
- [Levesque et Raj, 2013] LEVESQUE, M. J. et RAJ, A. (2013). Single-chromosome transcriptional profiling reveals chromosomal gene expression regulation. *Nature Methods*, 10(3):246–248.
- [Lieberman-Aiden *et al.*, 2009] LIEBERMAN-AIDEN, E., van BERKUM, N. L., WILLIAMS, L., IMAKAEV, M., RAGOCZY, T., TELLING, A., AMIT, I., LAJOIE, B. R., SABO, P. J., DORSCHNER, M. O., SANDSTROM, R., BERNSTEIN, B., BENDER, M. A., GROUDINE, M., GNIRKE, A., STAMATOYANNOPOULOS, J., MIRNY, L. a., LANDER, E. S., DEKKER, J., BERKUM, N. L. V., WILLIAMS, L., IMAKAEV, M., RAGOCZY, T., TELLING, A., AMIT, I., LAJOIE, B. R., SABO, P. J., DORSCHNER, M. O., SANDSTROM, R., BERNSTEIN, B., BENDER, M. A., GROUDINE, M., GNIRKE, A., STAMATOYANNOPOULOS, J. et MIRNY, L. a. (2009). Comprehensive mapping

- of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293.
- [Lomvardas *et al.*, 2006] LOMVARDAS, S., BARNEA, G., PISAPIA, D. J., MENDELSON, M., KIRKLAND, J. et AXEL, R. (2006). Interchromosomal Interactions and Olfactory Receptor Choice. *Cell*, 126(2):403–413.
- [Louis *et al.*, 2015] LOUIS, A., NGUYEN, N. T. T., MUFFATO, M. et CROLLIUS, H. R. (2015). Genomic update 2015: KaryoView and MatrixView provide a genome-wide perspective to multispecies comparative genomics. *Nucleic Acids Research*, 43(D1):D682–D689.
- [Lowry et Willis, 2010] LOWRY, D. B. et WILLIS, J. H. (2010). A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation. *PLoS Biology*, 8(9).
- [Lucas *et al.*, 2014] LUCAS, J. M., MUFFATO, M. et CROLLIUS, H. R. (2014). PhylDiag : identifying complex synteny blocks that include tandem duplications using phylogenetic gene trees. *BMC Bioinformatics*, 15(268):1–15.
- [Ma, 2006] MA, J. (2006). *Reconstructing contiguous regions of an ancestral genome*. Thèse de doctorat, The Pennsylvania State University.
- [Ma *et al.*, 2008a] MA, J., RATAN, A., RANEY, B. J., SUH, B. B., MILLER, W. et HAUSSLER, D. (2008a). The infinite sites model of genome evolution. *Proceedings of the National Academy of Sciences*, 105(38):14254–14261.
- [Ma *et al.*, 2008b] MA, J., RATAN, A., RANEY, B. J., SUH, B. B., ZHANG, L., MILLER, W. et HAUSSLER, D. (2008b). DUPCAR: reconstructing contiguous ancestral regions with duplications. *Journal of computational biology : a journal of computational molecular cell biology*, 15(8):1007–27.
- [Ma *et al.*, 2006] MA, J., ZHANG, L., SUH, B. B., RANEY, B. J., BURHANS, R. C., KENT, W. J., BLANCHETTE, M., HAUSSLER, D. et MILLER, W. (2006). Reconstructing contiguous regions of an ancestral genome. *Genome Research*, 16(12):1557–1565.
- [Malinsky *et al.*, 2015] MALINSKY, M., CHALLIS, R. J., TYERS, A. M., SCHIFFELS, S., TERAJ, Y., NGATUNGA, B. P., MISKA, E. A., DURBIN, R., GENNER, M. J. et TURNER, G. F. (2015). Genomic islands of speciation separate cichlid ecomorphs in an East African crater lake. *Science*, 350(6267):1493–8.

- [Mani et Chinnaiyan, 2010] MANI, R.-s. et CHINNAIYAN, A. M. (2010). Triggers for genomic rearrangements :. *Nature Reviews Genetics*, 11(12):819–829.
- [Martínez-Fundichely *et al.*, 2014] MARTÍNEZ-FUNDICHELY, A., CASILLAS, S., EGEA, R., RÀMIA, M., BARBADILLA, A., PANTANO, L., PUIG, M. et CÁCERES, M. (2014). InvFEST, a database integrating information of polymorphic inversions in the human genome. *Nucleic Acids Research*, 42(D1):1027–1032.
- [Matthew, 1831] MATTHEW, P. (1831). *On Naval Timber and Arboriculture*. Neill & Co. Printers.
- [Matthew, 1860] MATTHEW, P. (1860). Nature’s law of selection. *Gardeners’ Chronicle and Agricultural Gazette*, 312–13(7 april).
- [Mayrose *et al.*, 2010] MAYROSE, I., BARKER, M. S. et OTTO, S. P. (2010). Probabilistic models of chromosome number evolution and the inference of polyploidy. *Systematic biology*, 59(2):132–44.
- [Mazowita *et al.*, 2006] MAZOWITA, M., HAQUE, L. et SANKOFF, D. (2006). Stability of rearrangement measures in the comparison of genome sequences. *Journal of computational biology : a journal of computational molecular cell biology*, 13(2):554–566.
- [Miklós et Tannier, 2010] MIKLÓS, I. et TANNIER, E. (2010). Bayesian sampling of genomic rearrangement scenarios via double cut and join. *Bioinformatics*, 26(24):3012–9.
- [Misteli, 2007] MISTELI, T. (2007). Beyond the Sequence: Cellular Organization of Genome Function. *Cell*, 128(4):787–800.
- [Misteli, 2009] MISTELI, T. (2009). Self-organization in the genome. *Proceedings of the National Academy of Sciences*, 106(17):6885–6886.
- [Mitelman *et al.*, 2007] MITELMAN, F., JOHANSSON, B. et MERTENS, F. (2007). The impact of translocations and gene fusions on cancer causation. *Nature reviews Cancer*, 7(4):233–45.
- [Mooers *et al.*, 1999] MOOERS, A. Ø., VAMOSI, S. M. et SCHLUTER, D. (1999). Using Phylogenies to Test Macroevolutionary Hypotheses of Trait Evolution in Cranes (Gruinae). *The American Naturalist*, 154(2):249–259.
- [Muffato, 2010] MUFFATO, M. (2010). *Reconstruction de génomes ancestraux chez les vertébrés*. Thèse de doctorat, Université d’Évry Val d’Essonne.

- [Muller, 1930] MULLER, H. J. (1930). Types of visible variations induced by X-rays in *Drosophila*. *Journal of Genetics*, 22(3):299–334.
- [Murphy *et al.*, 2005] MURPHY, W. J., LARKIN, D. M., EVERTS-VAN DER WIND, A., BOURQUE, G., TESLER, G., AUVIL, L., BEEVER, J. E., CHOWDHARY, B. P., GALIBERT, F., GATZKE, L., HITTE, C., MEYERS, S. N., MILAN, D., OSTRANDER, E. a., PAPE, G., PARKER, H. G., RAUDSEPP, T., ROGATCHEVA, M. B., SCHOOK, L. B., SKOW, L. C., WELGE, M., WOMACK, J. E., O'BRIEN, S. J., PEVZNER, P. a. et LEWIN, H. a. (2005). Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science*, 309(5734):613–7.
- [Nadeau et Taylor, 1984] NADEAU, J. H. et TAYLOR, B. A. (1984). Lengths of chromosomal segments conserved since divergence of man and mouse. *Proceedings of the National Academy of Sciences*, 81(3):814–818.
- [Naville *et al.*, 2015] NAVILLE, M., ISHIBASHI, M., FERG, M., BENGANI, H., RINKWITZ, S., KRECSMARIK, M., HAWKINS, T. A., WILSON, S. W., MANNING, E., CHILAMAKURI, C. S. R., WILSON, D. I., LOUIS, A., LUCY RAYMOND, F., RASTEGAR, S., STRÄHLE, U., LENHARD, B., BALLY-CUIF, L., van HEYNINGEN, V., FITZPATRICK, D. R., BECKER, T. S. et ROEST CROLLIUS, H. (2015). Long-range evolutionary constraints reveal cis-regulatory interactions on the human X chromosome. *Nature communications*, 6(May 2014):6904.
- [Nóbrega *et al.*, 2004] NÓBREGA, M. A., ZHU, Y., PLAJZER-FRICK, I., AFZAL, V. et RUBIN, E. M. (2004). Megabase deletions of gene deserts result in viable mice. *Nature*, 431(7011):988–93.
- [Ohno, 1970] OHNO, S. (1970). *Evolution by Gene Duplication*. Wiley-Blackwell.
- [Orlando *et al.*, 2011] ORLANDO, L., GINOLHAC, A., RAGHAVAN, M., VILSTRUP, J., RASMUSSEN, M., MAGNUSSEN, K., STEINMANN, K. E., KAPRANOV, P., THOMPSON, J. F., ZAZULA, G., FROESE, D., MOLTKE, I., SHAPIRO, B., HOFREITER, M., AL-RASHEID, K. a. S., GILBERT, M. T. P. et WILLERSLEV, E. (2011). True single-molecule DNA sequencing of a pleistocene horse bone. *Genome research*, 21(10):1705–19.
- [Osawa *et al.*, 1992] OSAWA, S., JUKES, T. H., WATANABE, K. et MUTO, a. (1992). Recent-Evidence for Evolution of the Genetic-Code. *Microbiological Reviews*, 56(1):229–264.

- [Osborne *et al.*, 2004] OSBORNE, C. S., CHAKALOVA, L., BROWN, K. E., CARTER, D., HORTON, A., DEBRAND, E., GOYENECHEA, B., MITCHELL, J. a., LOPES, S., REIK, W. et FRASER, P. (2004). Active genes dynamically colocalize to shared sites of ongoing transcription. *Nature genetics*, 36(10): 1065–1071.
- [Ovcharenko *et al.*, 2005] OVCHARENKO, I., LOOTS, G. G., NOBREGA, M. A., HARDISON, R. C., MILLER, W. et STUBBS, L. (2005). Evolution and functional classification of vertebrate gene deserts. *Genome Research*, 15(1):137–145.
- [Pasek *et al.*, 2005] PASEK, S., BERGERON, A., RISLER, J.-L., LOUIS, A., OLLIVIER, E. et RAFFINOT, M. (2005). Identification of genomic features using microsynteny of domains: domain teams. *Genome research*, 15(6): 867–74.
- [Paten *et al.*, 2008] PATEN, B., HERRERO, J., BEAL, K., FITZGERALD, S. et BIRNEY, E. (2008). Enredo and Pecan: Genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Research*, 18(11):1814–1828.
- [Peng *et al.*, 2006] PENG, Q., PEVZNER, P. a. et TESLER, G. (2006). The fragile breakage versus random breakage models of chromosome evolution. *PLoS computational biology*, 2(2):e14.
- [Pevzner et Tesler, 2003a] PEVZNER, P. et TESLER, G. (2003a). Genome rearrangements in mammalian evolution: lessons from human and mouse genomes. *Genome research*, 13(1):37–45.
- [Pevzner et Tesler, 2003b] PEVZNER, P. et TESLER, G. (2003b). Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proceedings of the National Academy of Sciences*, 100(13):7672–7677.
- [Pham et Pevzner, 2010] PHAM, S. K. et PEVZNER, P. A. (2010). DRIMM-Synteny: decomposing genomes into evolutionary conserved segments. *Bioinformatics*, 26(20):2509–16.
- [Proost *et al.*, 2012] PROOST, S., FOSTIER, J., WITTE, D. D., DHOEDT, B., DEMEESTER, P., de PEER, Y. V., VANDEPOELE, K., DE WITTE, D., DHOEDT, B., DEMEESTER, P., VAN DE PEER, Y. et VANDEPOELE, K. (2012). i-ADHoRe 3.0—fast and sensitive detection of genomic homology in extremely large data sets. *Nucleic acids research*, 40(2):e11.

- [Raghupathy et Durand, 2009] RAGHUPATHY, N. et DURAND, D. (2009). Gene cluster statistics with gene families. *Molecular biology and evolution*, 26(5):957–68.
- [Raghupathy et al., 2008] RAGHUPATHY, N., HOBERMAN, R. et DURAND, D. (2008). Two plus two does not equal three: statistical tests for multiple genome comparison. *Journal of bioinformatics and computational biology*, 6(1):1–22.
- [Rao et al., 2014] RAO, S. S. P., HUNTLEY, M. H., DURAND, N. C., STAMENOVA, E. K., BOCHKOV, I. D., ROBINSON, J. T., SANBORN, A. L., MACHOL, I., OMER, A. D., LANDER, E. S. et AIDEN, E. L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680.
- [Rousset, 2008] ROUSSET, F. (2008). genpop’007: a complete re-implementation of the genepop software for Windows and Linux. *Molecular ecology resources*, 8(1):103–6.
- [Sanborn et al., 2015] SANBORN, A. L., RAO, S. S. P., HUANG, S.-C., DURAND, N. C., HUNTLEY, M. H., JEWETT, A. I., BOCHKOV, I. D., CHINNAPPAN, D., CUTKOSKY, A., LI, J., GEETING, K. P., GNIRKE, A., MELNIKOV, A., MCKENNA, D., STAMENOVA, E. K., LANDER, E. S. et AIDEN, E. L. (2015). Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proceedings of the National Academy of Sciences*, 112(47):201518552.
- [Sankoff, 1999] SANKOFF, D. (1999). Genome rearrangement with gene families. *Bioinformatics*, 15(11):909–917.
- [Sankoff, 2002] SANKOFF, D. (2002). Short inversions and conserved gene clusters. *Proceedings of the 2002 {ACM} symposium on Applied computing*, 18(10):164–167.
- [Sankoff, 2005a] SANKOFF, D. (2005a). Conserved segment statistics and rearrangement inferences in comparative genomics. *In Mathematics of Evolution and Phylogeny*, pages 236–261. OUP Oxford.
- [Sankoff, 2005b] SANKOFF, D. (2005b). *The Distribution of Inversion Lengths in Bacteria*, volume 53. Springer.
- [Sankoff, 2006] SANKOFF, D. (2006). The signal in the genomes.

- [Sankoff, 2009] SANKOFF, D. (2009). The where and wherefore of evolutionary breakpoints. *Journal of biology*, 8(7):66.
- [Sankoff et Ferretti, 1996] SANKOFF, D. et FERRETTI, V. (1996). Karyotype distributions in a stochastic model of reciprocal translocation. *Genome Research*, 6(1):1–9.
- [Sankoff et Mazowita, 2005] SANKOFF, D. et MAZOWITA, M. (2005). Estimators of Translocations and Inversions. *Comparative and General Pharmacology*, pages 109–122.
- [Schlötterer, 2015] SCHLÖTTERER, C. (2015). Genes from scratch—the evolutionary fate of de novo genes. *Trends in genetics : TIG*, 31(4):215–9.
- [Schubert et Oud, 1997] SCHUBERT, I. et OUD, J. (1997). There Is an Upper Limit of Chromosome Size for Normal Development of an Organism. *Cell*, 88(4):515–520.
- [Semeria et al., 2015] SEMERIA, M., TANNIER, E. et GUÉGUEN, L. (2015). Probabilistic modeling of the evolution of gene synteny within reconciled phylogenies. *BMC Bioinformatics*, 16(Suppl 14):S5.
- [Sémon et Duret, 2006] SÉMON, M. et DURET, L. (2006). Evolutionary origin and maintenance of coexpressed gene clusters in mammals. *Molecular biology and evolution*, 23(9):1715–23.
- [Shi et al., 2011] SHI, G., PENG, M. C. et JIANG, T. (2011). MultiMSOAR 2.0: An accurate tool to identify ortholog groups among multiple genomes. *PLoS ONE*, 6(6):1–2.
- [Shlyueva et al., 2014] SHLYUEVA, D., STAMPFEL, G. et STARK, A. (2014). Transcriptional enhancers: from properties to genome-wide predictions. *Nature reviews Genetics*, 15(4):272–86.
- [Simillion et al., 2004] SIMILLION, C., VANDEPOELE, K., SAEYS, Y., VAN DE PEER, Y. et de PEER, Y. V. (2004). Building genomic profiles for uncovering segmental homology in the twilight zone. *Genome research*, 14(6):1095–1106.
- [Sinha et Meller, 2007] SINHA, A. U. et MELLER, J. (2007). Cinteny: flexible analysis and visualization of synteny and genome rearrangements in multiple organisms. *BMC Bioinformatics*, 8(1):82.

- [Soderlund *et al.*, 2006] SODERLUND, C., NELSON, W., SHOEMAKER, A. et PATERSON, A. (2006). SyMAP: A system for discovering and viewing syntenic regions of FPC maps. *Genome Research*, 16(9):1159–1168.
- [Spilianakis *et al.*, 2005] SPILIANAKIS, C. G., LALIOU, M. D., TOWN, T., LEE, G. R. et FLAVELL, R. a. (2005). Interchromosomal associations between alternatively expressed loci. *Nature*, 435(7042):637–645.
- [Spitz *et al.*, 2005] SPITZ, F., HERKENNE, C., MORRIS, M. a. et DUBOULE, D. (2005). Inversion-induced disruption of the Hoxd cluster leads to the partition of regulatory landscapes. *Nature genetics*, 37(8):889–93.
- [Stefansson *et al.*, 2005] STEFANSSON, H., HELGASON, A., THORLEIFSSON, G., STEINTHORSDOTTIR, V., MASSON, G., BARNARD, J., BAKER, A., JONASDOTTIR, A., INGASON, A., GUDNADOTTIR, V. G., DESNICA, N., HICKS, A., GYLFASON, A., GUDBJARTSSON, D. F., JONSDOTTIR, G. M., SAINZ, J., AGNARSSON, K., BIRGISDOTTIR, B., GHOSH, S., OLAFSDOTTIR, A., CAZIER, J.-B., KRISTJANSSON, K., FRIGGE, M. L., THORGEIRSSON, T. E., GULCHER, J. R., KONG, A. et STEFANSSON, K. (2005). A common inversion under selection in Europeans. *Nature genetics*, 37(2):129–137.
- [Stephens *et al.*, 2011] STEPHENS, P. J., GREENMAN, C. D., FU, B., YANG, F., BIGNELL, G. R., MUDIE, L. J., PLEASANCE, E. D., LAU, K. W., BEARE, D., STEBBINGS, L. A., MCLAREN, S., LIN, M. L., MCBRIDE, D. J., VARELA, I., NIK-ZAINAL, S., LEROY, C., JIA, M., MENZIES, A., BUTLER, A. P., TEAGUE, J. W., QUAIL, M. A., BURTON, J., SWERDLOW, H., CARTER, N. P., MORSBERGER, L. A., IACOBUZIO-DONAHUE, C., FOLLOWS, G. A., GREEN, A. R., FLANAGAN, A. M., STRATTON, M. R., FUTREAL, P. A. et CAMPBELL, P. J. (2011). Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell*, 144(1):27–40.
- [Stratton *et al.*, 2009] STRATTON, M. R., CAMPBELL, P. J. et FUTREAL, P. A. (2009). The cancer genome. *Nature*, 458(7239):719–724.
- [Sturtevant, 1921] STURTEVANT, A. (1921). A Case of Rearrangement of Genes in *Drosophila*. *Genetics*, 7:235–237.
- [Sudmant *et al.*, 2015] SUDMANT, P. H., RAUSCH, T., GARDNER, E. J., HANDSAKER, R. E., ABYZOV, A., HUDDLESTON, J., ZHANG, Y., YE, K., JUN, G., HSI-YANG FRITZ, M., KONKEL, M. K., MALHOTRA, A., STÜTZ, A. M., SHI, X., PAOLO CASALE, F., CHEN, J., HORMOZDIARI, F., DAYAMA, G., CHEN, K., MALIG, M., CHAISSON, M. J. P., WALTER, K., MEIERS, S.,

- KASHIN, S., GARRISON, E., AUTON, A., LAM, H. Y. K., JASMINE MU, X., ALKAN, C., ANTAKI, D., BAE, T., CERVEIRA, E., CHINES, P., CHONG, Z., CLARKE, L., DAL, E., DING, L., EMERY, S., FAN, X., GUJRAL, M., KAHVECI, F., KIDD, J. M., KONG, Y., LAMEIJER, E.-W., MCCARTHY, S., FLICEK, P., GIBBS, R. A., MARTH, G., MASON, C. E., MENELAOU, A., MUZNY, D. M., NELSON, B. J., NOOR, A., PARRISH, N. F., PENDLETON, M., QUITADAMO, A., RAEDER, B., SCHADT, E. E., ROMANOVITCH, M., SCHLATTL, A., SEBRA, R., SHABALIN, A. A., UNTERGASSER, A., WALKER, J. A., WANG, M., YU, F., ZHANG, C., ZHANG, J., ZHENG-BRADLEY, X., ZHOU, W., ZICHNER, T., SEBAT, J., BATZER, M. A., MCCARROLL, S. A., MILLS, R. E., GERSTEIN, M. B., BASHIR, A., STEGLE, O., DEVINE, S. E., LEE, C., EICHLER, E. E. et KORBEL, J. O. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, 526(7571):75–81.
- [Swidan *et al.*, 2006] SWIDAN, F., ROCHA, E. P. C., SHMOISH, M. et PINTER, R. Y. (2006). An integrative method for accurate comparative genome mapping. *PLoS computational biology*, 2(8):e75.
- [Tang *et al.*, 2011] TANG, H., LYONS, E., PEDERSEN, B., SCHNABLE, J. C., PATERSON, A. H. et FREELING, M. (2011). Screening synteny blocks in pairwise genome comparisons through integer programming. *BMC bioinformatics*, 12(1):102.
- [ten Tusscher et Hogeweg, 2009] ten TUSSCHER, K. H. W. J. et HOGEWEG, P. (2009). The role of genome and gene regulatory network canalization in the evolution of multi-trait polymorphisms and sympatric speciation. *BMC evolutionary biology*, 9:159.
- [Tesler, 2002] TESLER, G. (2002). GRIMM: genome rearrangements web server. *Bioinformatics*, 18(3):492–493.
- [Thompson et Jiggins, 2014] THOMPSON, M. J. et JIGGINS, C. D. (2014). Supergenes and their role in evolution. *Heredity*, 113(1):1–8.
- [Uno et Yagiura, 2000] UNO, T. et YAGIURA, M. (2000). Fast Algorithms to Enumerate All Common Intervals of Two Permutations. *Algorithmica*, 26(2):290–309.
- [Vandepoele *et al.*, 2002] VANDEPOELE, K., SAEYS, Y., SIMILLION, C., RAES, J. et VAN DE PEER, Y. (2002). The automatic detection of homologous regions (ADHoRe) and its application to microcolinearity between Arabidopsis and rice. *Genome research*, 12(11):1792–1801.

- [Via, 2009] VIA, S. (2009). Natural selection in action during speciation. *Proceedings of the National Academy of Sciences*, 106 Suppl:9939–9946.
- [Via, 2012] VIA, S. (2012). Divergence hitchhiking and the spread of genomic isolation during ecological speciation-with-gene-flow. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 367(1587):451–460.
- [Vilella *et al.*, 2009] VILELLA, A. J., SEVERIN, J., URETA-VIDAL, A., HENG, L., DURBIN, R. et BIRNEY, E. (2009). EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome research*, 19(2):327–35.
- [Wang *et al.*, 2013] WANG, J., WURM, Y., NIPITWATTANAPHON, M., RIBAGROGNOZ, O., HUANG, Y.-C., SHOEMAKER, D. et KELLER, L. (2013). A Y-like social chromosome causes alternative colony organization in fire ants. *Nature*, 493(7434):664–8.
- [Wang *et al.*, 2012] WANG, Y., TANG, H., DEBARRY, J. D., TAN, X., LI, J., WANG, X., LEE, T.-h., JIN, H., MARLER, B., GUO, H., KISSINGER, J. C. et PATERSON, A. H. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic acids research*, 40(7):e49.
- [Warburton, 1991] 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.
- [Wegmann *et al.*, 2010] WEGMANN, D., LEUENBERGER, C., NEUENSCHWANDER, S. et EXCOFFIER, L. (2010). ABCtoolbox: a versatile toolkit for approximate Bayesian computations. *BMC bioinformatics*, 11:116.
- [Weischenfeldt *et al.*, 2013] WEISCHENFELDT, J., SYMMONS, O., SPITZ, F. et KORBEL, J. O. (2013). Phenotypic impact of genomic structural variation: insights from and for human disease. *Nature Reviews Genetics*, 14(2):125–138.
- [Wielgoss *et al.*, 2011] WIELGOSS, S., BARRICK, J. E., TENAILLON, O., CRUVEILLER, S., CHANE-WOON-MING, B., MEDIGUE, C., LENSKI, R. E., SCHNEIDER, D. et ANDREWS, B. J. (2011). Mutation Rate Inferred From Synonymous Substitutions in a Long-Term Evolution Experiment With *Escherichia coli*. *Genes/Genomes/Genetics*, 1(3):183–186.

- [Wijchers et de Laat, 2011] WIJCHERS, P. J. et de LAAT, W. (2011). Genome organization influences partner selection for chromosomal rearrangements. *Trends in Genetics*, 27(2):63–71.
- [Wilke et al., 2001] WILKE, C. O., WANG, J. L., OFRIA, C., LENSKI, R. E. et ADAMI, C. (2001). Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature*, 412(6844):331–3.
- [Williams et al., 2010] WILLIAMS, A., SPILIANAKIS, C. G. et FLAVELL, R. A. (2010). Interchromosomal association and gene regulation in trans. *Trends in Genetics*, 26(4):188–197.
- [Yandell et Ence, 2012] YANDELL, M. et ENCE, D. (2012). A beginner’s guide to eukaryotic genome annotation. *Nature Reviews Genetics*, 13(5):329–342.
- [Yeaman, 2013] YEAMAN, S. (2013). Genomic rearrangements and the evolution of clusters of locally adaptive loci. *Proceedings of the National Academy of Sciences*, 110(19):E1743–51.
- [Zdobnov et Bork, 2007] ZDOBNOV, E. M. et BORK, P. (2007). Quantification of insect genome divergence.
- [Zhang et al., 2015] ZHANG, C.-Z., SPEKTOR, A., CORNILS, H., FRANCIS, J. M., JACKSON, E. K., LIU, S., MEYERSON, M. et PELLMAN, D. (2015). Chromothripsis from DNA damage in micronuclei. *Nature*, 522(7555):179–184.
- [Zhang et Leong, 2009] ZHANG, M. et LEONG, H. W. (2009). Gene team tree: a hierarchical representation of gene teams for all gap lengths. *Journal of computational biology : a journal of computational molecular cell biology*, 16(10):1383–98.
- [ZHANG et al., 1994] ZHANG, Z., RAGHAVACHARI, B., HARDISON, R. C. et MILLER, W. (1994). Chaining Multiple-Alignment Blocks. *Journal of Computational Biology*, 1(3):217–226.
- [Zhao et Bourque, 2009] ZHAO, H. et BOURQUE, G. (2009). Recovering genome rearrangements in the mammalian phylogeny. *Genome research*, 19(5):934–42.
- [Zheng et al., 2007] ZHENG, C., ZHU, Q. et SANKOFF, D. (2007). Removing noise and ambiguities from comparative maps in rearrangement analysis. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, 4(4):515–22.

[Zody *et al.*, 2008] ZODY, M. C., JIANG, Z., FUNG, H.-C., ANTONACCI, F., HILLIER, L. W., CARDONE, M. F., GRAVES, T. a., KIDD, J. M., CHENG, Z., ABOUELLEIL, A., CHEN, L., WALLIS, J., GLASSCOCK, J., WILSON, R. K., REILY, A. D., DUCKWORTH, J., VENTURA, M., HARDY, J., WARREN, W. C. et EICHLER, E. E. (2008). Evolutionary toggling of the MAPT 17q21.31 inversion region. *Nature genetics*, 40(9):1076–1083.